**Zeitschrift:** Jahrbuch der Schweizerischen Naturforschenden Gesellschaft. Wissenschaftlicher und administrativer Teil = Annuaire de la Société Helvétique des Sciences Naturelles. Partie scientifique et administrative

**Herausgeber:** Schweizerische Naturforschende Gesellschaft

**Band:** 161 (1981)

**Teilband:** Wissenschaftlicher Teil : Vom Ursprung der Dinge = Partie scientifique : de l'origine des choses

Die 1815 gegründete Schweizerische Naturforschende Gesellschaft ist die älteste wissenschaftliche Dachgesellschaft der Schweiz. Ihr Ziel ist die Förderung und Entwicklung der exakten und Naturwissenschaften und deren Vertretung in der Öffentlichkeit. Den intensiven Gedankenaustausch zwischen Wissenschaftern verschiedener Fachrichtungen fördert die SNG, indem sie Symposien durchführt oder unterstützt und publiziert. Das Jahrbuch, wissenschaftlicher Teil, ist die Fortsetzung der seit 1819 erschienenen Verhandlungen der Schweizerischen Naturforschenden Gesellschaft.

Fondée en 1815, la Société helvétique des sciences naturelles est la plus ancienne organisation faîtière scientifique du pays. Elle a pour but l'encouragement et le développement des sciences exactes et naturelles, leur compréhension auprès du public et l'intensification des échanges entre scientifiques de diverses disciplines. Elle organise et soutient des symposia et en publie les actes. L'Annuaire, partie scientifique, remplace les Actes de la Société helvétique des sciences naturelles, publiés depuis 1819.

# Vom Ursprung der Dinge
# De l'origine des choses

Cat

für die SNG herausgegeben von
édité pour la SHSN par

Claus Fröhlich

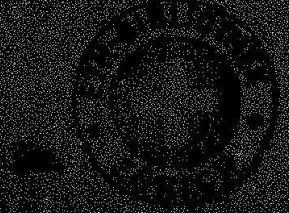
---

1981
Birkhäuser Verlag
Basel · Boston · Stuttgart

Printed in Switzerland by Birkhäuser AG, Graphisches Unternehmen, Basel
ISSN 0080/7362

# VOM URSPRUNG DER DINGE

# DE L'ORIGINE DES CHOSES



# Inhalt

# Naturforschung in Davos über zwei Jahrhunderte
# Vortrag des Jahrespräsidenten

Marcel de Quervain

Ursprüngliche Natur als Objekt der Erforschung ungestörter Naturvorgänge wird man in abgelegenen Gebieten suchen, je abgelegener desto besser. Gewiss hat die Gegend von Davos von jeher Möglichkeiten in dieser Beziehung geboten und bietet sie auch heute noch. Die in Davos anzutreffende Konzentration naturwissenschaftlich interessierter Menschen dürfte aber noch andere Gründe haben; es muss hier «etwas in der Luft liegen». Dies war jedenfalls in wörtlichem Sinn die Auffassung aufmerksamer Beobachter über Jahrhunderte zurück.

Der 1962 verstorbene hervorragende Davoser Historiker Jules Ferdmann ist dieser Frage in seinen Büchern «Die Anfänge des Kurortes Davos» (1938) und «Der Aufstieg von Davos» (1947) nachgegangen. Die nachfolgenden Hinweise auf Vergangenes sind bis ins erste Jahrzehnt dieses Jahrhunderts grösstenteils den genannten Werken entnommen.

Ulrich Campell, Pfarrer, Geschichtsschreiber und Geograph von Klosters (um 1550), schreibt Davos «eine äusserst heilkräftige Luft» zu – «wenn auch reichlich kalt», wie er beifügt. Die Idee von der besonderen Davoser Luft zieht sich wie ein roter Faden bis in die Gegenwart und hat die Entwicklung von Davos geprägt. Ritter Johannes Guler von Wyneck, ein aus Davos stammender Chronist, bestätigt anfangs des 17. Jahrhunderts die Heilwirkung von Sonne und Luft und empfiehlt bereits Kuren für Lungenkranke. Fast zwei Jahrhunderte später, 1806, führt der Davoser Landammann Jakob Valär als Begründung des Heilklimas an, es gebe in Davos keine feuchten Nebel und sei oft mild. Von den Davosern sagt er, sie seien «sehr gesund, stark und kropffrei» – nebenbei auch, sie hätten «viel Mutterwitz, sehr schnelle Begriffe und unglaubliche Schlauheit». Von ihm stammt das Wort: «An unserem Klima haben wir den besten Arzt».

Um 1820 wird das Davoser Klima vom Berner Forstmann Karl Kasthofer mit demjenigen voralpiner Lagen verglichen. Er stellt fest, dass in den südwestlichen Teilen der Landschaft auf 1300 m Höhe unter der Wirkung höherer Temperaturen Birnen und Äpfel reifen und erklärt das wärmere Klima richtig als Folge der alpinen Massenerhebung. Als weitere Gründe nennt er das sanftere Relief und die starke Bewaldung.

Ab 1828 treten in Davos in laufender Folge die Ärzte in Erscheinung, die den Ort zum Klimakurort gemacht haben, zunächst Luzius Rüedi, ein offenbar sehr eigenwilliger Charakter, der sich mit den Behörden immer wieder überwarf. Er teilte die früher schon von Horace-Bénédicte de Saussure und Ignaz Troxler vertretene Auffassung, dass in Hochlagen Kretinismus und Skrophulose aus klimatischen Gründen ausgeschlossen seien, und gründete 1841 in Davos eine private Anstalt für skrophulöse und kretine Kinder und damit quasi das erste Davoser Sanatorium. Die Schweizerische Naturforschende Gesellschaft befasste sich damals gesamtschweizerisch mit diesem Problem und veranstaltete darüber eine Umfrage, die Rüedi bündig mit dem Satz beantwortete: «In Davos Kretinismus unbekannt».

Auf Rüedi folgte 1849 J.G. Amstein, ein Arzt mit vielseitigen naturwissenschaftlichen Interessen. Einerseits untersuchte er die Wasserzusammensetzung der Mineralquellen von Davos, und anderseits widmete er sich der Erforschung der Schnecken, von denen er in Graubünden 130 Spezies fand. Amstein war einer der Gründer und Förderer der Naturforschenden Gesellschaft Graubünden.

Als bedeutendster Promotor des Lungenkurortes Davos gilt der 1853 nach Davos gewählte Arzt Alexander Spengler, ein politischer Flüchtling von 1848 aus Deutschland. Die nun einsetzende stürmische Kurortent-

*Abb. 1.* Davos-Platz 1865, zu Beginn der Kurortentwicklung (Foto Bibliothek Davos).

wicklung, die das Bergdorf auch baulich innerhalb weniger Jahrzehnte umgestaltet hat, soll hier nicht weiter verfolgt werden, wohl aber ihre Konsequenz für die Naturforschung (Abb. 1). Sowohl unter den Ärzten als auch unter den Patienten, die sich hier in zunehmender Zahl einfanden, gab es nämlich naturwissenschaftlich interessierte und zum Teil sehr aktive Persönlichkeiten. Doch vor dem Ende des letzten Jahrhunderts stand noch einmal ein Pfarrer im Mittelpunkt des naturwissenschaftlichen Lebens von Davos: Pfarrer Johannes Hauri, der sich nicht nur für Engel, sondern ebensosehr für Schmetterlinge interessierte und darüber eine Schrift verfasste. Hauri war Jahrespräsident der ersten in Davos abgehaltenen Jahresversammlung von 1890. Dem damaligen poetischen Zeitgeist verpflichtet, hat er der Versammlung eine humorvolle Dichtung, betitelt «St. Petrus und die Naturforscher», gewidmet, die in nachsichtiger Weise die auch in ihm schwelende Spannung zwischen Naturforschung und Religion glossiert. Eine Synthese findet er im Satz: «Und der die Natur gerufen ins Leben, der will auch: Es soll Naturforscher geben.»

Dann trat wieder ein Arzt in den Vordergrund der Davoser Naturforschung, nämlich Wilhelm Schibler. Dieser beliebte Landarzt und Alpinist war in seiner zweiten Berufung Botaniker und erforschte eingehend die Flora des Landwassertales. Besondere Aufmerksamkeit schenkte er den in Höhen über 2600 m noch anzutreffenden Pflanzenarten. Gesamthaft fand er deren 253 und verfolgte ihre Abnahme bis zur Schneegrenze an 66 Gipfeln und Pässen. Schibler war einer der Gründer der Naturforschenden Gesellschaft Davos (1916) und ihr erster Präsident. Im Jahr 1929 leitete er als Jahrespräsident die zweite in Davos abgehaltene Jahresversammlung der SNG. Erst sechs Jahre nach seinem 1931 eingetretenen Tod erschien sein Übersichtswerk «Die Flora von Davos».

Unter den Gründern der Naturforschenden Gesellschaft Davos befanden sich zwei weitere passionierte Naturforscher sehr ver-

schiedener Prägung, beide aus gesundheitlichen Gründen von Deutschland zugewandert. Der eine, Carl Dorno (1865–1942), war eine von strengen Prinzipien geleitete autoritäre Persönlichkeit, der andere, Otto Suchlandt (1873–1947), von sehr liebenswürdiger eher zurückhaltender Natur.
Suchlandt, von Beruf Apotheker, erforschte nach seiner Genesung das Plankton von acht Seen in verschiedenen Höhenlagen der Umgebung von Davos. Eine eingehende Studie galt der Veränderung des Planktons im Davosersee nach dessen ab 1923 zum Zweck der Energiegewinnung vorgenommenen Absenkung. Suchlandts ehemalige hydrobiologische Station am See ist heute noch zu sehen, umschwärmt vom heutigen oberflächlichen Plankton, den Seglern und Surfern.
Carl Dorno, der für seine Tochter in Davos Heilung suchte, setzte sich zum Ziel, das Davoser Klima und seine Heilwirkung messend zu ergründen. Im Jahr 1907 richtete er sich ein privates physikalisch-meteorologisches Observatorium ein, weiterhin kurz «Observatorium» genannt, und betrieb es bis zum Jahr 1922 mit eigenen Mitteln. Sein Hauptinteresse war auf die Sonnenstrahlung gerichtet. Täglich und stündlich mass er mit einem zum Teil selbst entwickelten Instrumentarium Strahlungswerte, berechnete Energieumsätze und untersuchte Schwankungen. Um den Heilfaktoren des Klimas auf die Spur zu kommen, überprüfte er kombinierte Einwirkungen der Umwelt auf den menschlichen Körper, vor allem dessen Abkühlung oder Erwärmung unter dem Einfluss von Strahlung, Lufttemperatur, Wind und Niederschlag. Hiezu baute er ein alle Einflüsse integrierendes thermostatiertes Instrument, gleichsam einen allem Wetter ausgesetzten menschlichen Kopf simulierend, dessen Wärmeumsatz direkt gemessen werden konnte. Mit diesem als «Frigorimeter» bezeichneten Instrument wurde in den dreissiger Jahren die sogenannte «Abkühlungsgrösse» an verschiedenen Orten der Schweiz im Jahreszyklus gemessen. Bemerkenswerterweise lag die Winterkurve von Davos in der Nähe derjenigen von Locarno und weit unter den Werten von Zürich. Dorno gilt als Begründer der Strahlungsklimatologie und der Bioklimatologie. Seine ehernen Prinzipien werden illustriert durch einen Ausspruch, den er einem Kollegen gegenüber geäussert haben soll, als dieser erwähnte, sein Assistent habe eine Temperaturmessreihe aufgenommen: «Was, Sie wagen es, ein Thermometer von einem Assistenten ablesen zu lassen!» (Mitgeteilt von W. Mörikofer).
Noch bevor Dorno sein privates Observatorium nach rund 20 Jahren aus finanziellen Gründen aufgeben musste – die Inflation hatte sein Vermögen dahingerafft –, hatte sich die Forschung in Davos auf der medizinischen Ebene institutionalisiert. Unter dem Patronat gesamtschweizerischer Ärztekreise wurde im Jahr 1922 ein bereits durch den Davoser Arzt Karl Turban unterbreiteter Vorschlag verwirklicht und ein Institut für Hochgebirgsphysiologie und Tuberkuloseforschung ins Leben gerufen. Aufgrund eines Landsgemeindebeschlusses konnte hiefür eine finanzielle Basis geschaffen werden, indem von jedem Davoser Übernachtungsgast eine bescheidene Stiftungstaxe erhoben wurde. Vier Jahre später, also 1926, wurde das Observatorium in das Forschungsinstitut eingegliedert, das fortan den Namen «Schweizerisches Forschungsinstitut für Hochgebirgsklima und Tuberkulose» trug, im weiteren kurz «Forschungsinstitut» genannt. Bevor wir die Entwicklung dieses Instituts, das nun eine Medizinische Abteilung und das Observatorium umfasste, weiterverfolgen, sei eine andere Begebenheit aus den Zwischenkriegsjahren eingeblendet:
Es gehört zum guten Ton eines Ortes von akademischem Rang, Albert Einstein beherbergt zu haben. Davos kann tatsächlich damit aufwarten. Er hat zwar hier nicht Forschung getrieben, hat aber 1928 die Eröffnungsansprache des ersten Davoser Hochschulkurses gehalten und versucht, den versammelten Geisteswissenschaftern – darunter Namen wie Häberlin, Piaget, Rappard, Tillich – die Wandlungen der modernen Physik nahezubringen. Es ging ihm aber auch um andere Fragen, z. B. um den bereits wieder bedrohten Frieden. Ein in dieser Richtung weisender Ausspruch aus Einsteins Davoser Rede lautete: «Man nützt der internationalen Verständigung am meisten dadurch, dass man an einem lebensfördernden Werk arbeitet.» Einstein hat in Davos nicht nur gesprochen, er hat auch öffentlich musiziert. Angesichts der prekären Finanzen des Hochschulkurses hat er spontan ein Trio zusammengerufen und nach wenigen Proben

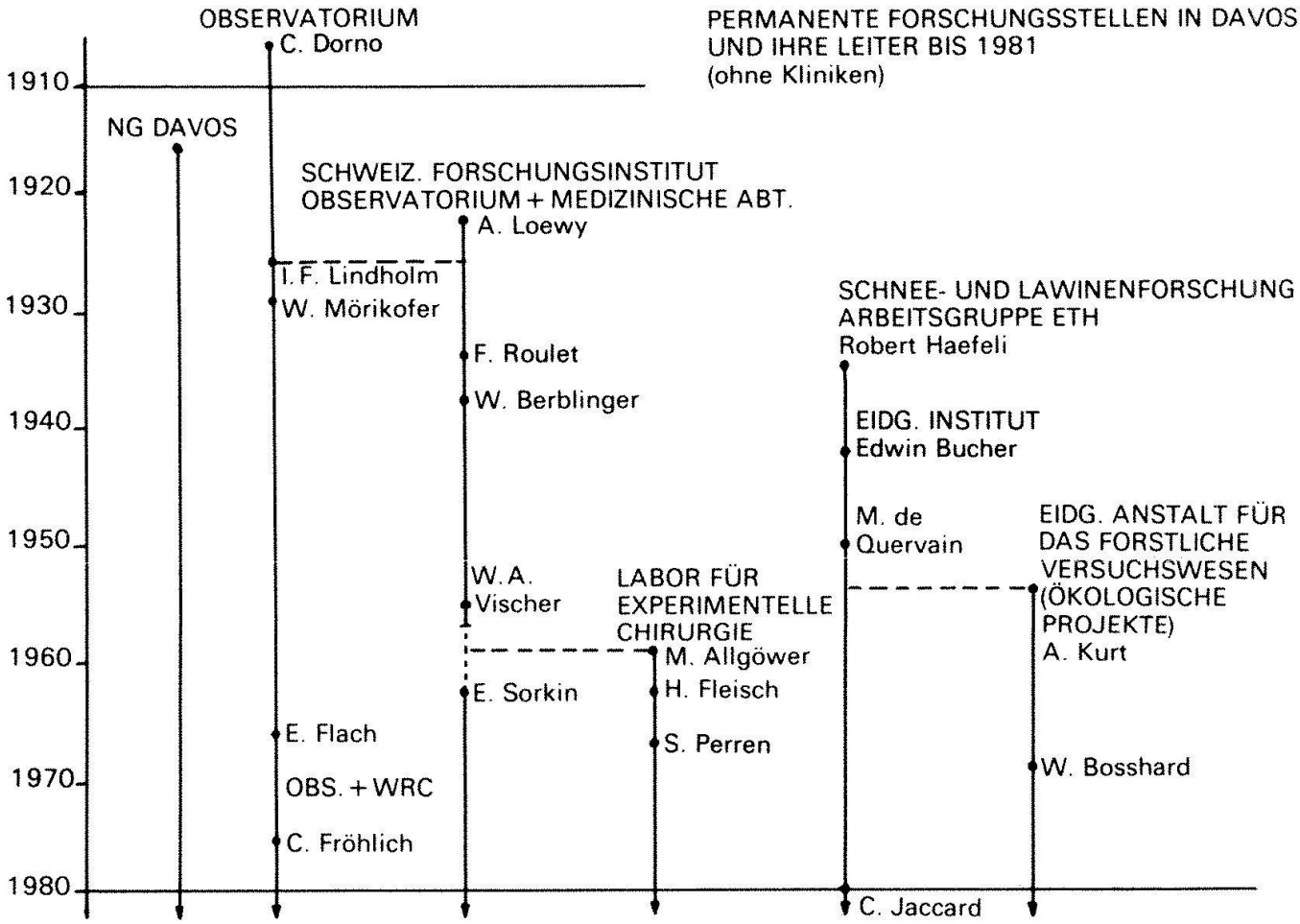



*Abb. 2.* Permanente Forschungsstellen in Davos und ihre Leiter 1907 bis 1981 (ohne Kliniken).

als dessen Violinist ein öffentliches Benefiz-Konzert gegeben.

Die im laufenden Jahrhundert in Davos entstandenen naturwissenschaftlichen Forschungsstellen (Abb. 2) sollen nun skizzenhaft bis in die Gegenwart begleitet werden. Hiezu muss wiederholt im Zeitablauf um einige Jahrzehnte zurückgeschaltet werden. Es sei auch auf die Darstellung der gegenwärtigen Forschungsaktivität in der Davoser Revue (Nr. 3, 1981) verwiesen, die den Teilnehmern der Jahresversammlung als Festgabe überreicht worden ist.

Im medizinischen Bereich des Forschungsinstituts stand vor der Fusion mit dem Observatorium die Höhenphysiologie im Vordergrund, d. h. die Wirkung der Höhe auf den gesunden und kranken menschlichen Körper. Der erste Leiter, Adolf Loewy, bearbeitete vor allem den Gasaustausch der Lunge in verschiedenen Höhenlagen, wobei als Höhenstandorte neben Davos Laboratorien auf Muottas Muragl (Engadin) und Gornergrat (Wallis) zur Verfügung standen. Mit jeder Änderung in der Leitung der Medizinischen Abteilung verlagerte sich jeweils auch das Forschungsgebiet. So wandten sich ab 1934 F. Roulet und ab 1938 Walther Berblinger dem Verlauf der Tuberkuloseerkrankung, ihrer Auswirkung auf den Kreislauf und ihrem pathologischen Bild zu. Diese Forschungsrichtungen dominierten bis Mitte der fünfziger Jahre und kulminierten 1951 im Bezug einer neuen Forschungsstätte, die dank einer gesamtschweizerischen Gönnerschaft in der Villa Fontana, dem ehemaligen Spenglerhaus, eingerichtet werden konnte. Doch bereits zeichnete sich der chemotherapeutische Erfolg im Kampf gegen die Tuberkulose ab, und damit auch ein Patientenschwund in Davos. Mit Arbeiten von W. A. Vischer über die Resistenzentwicklung von Tuberkulosebakterien fand die Tuberkuloseforschung im Jahr 1957 am Institut einen Abschluss, und die Medizinische Abteilung wurde, wie man sagt, eingemottet.

Dank der unentwegten Bemühungen des Davoser Arztes Felix Suter, der 1960 das

Präsidium der Stiftung übernahm, konnte zwei Jahre später die Abteilung mit neuen Zielen wieder eröffnet werden. Da Davos zunehmend von Patienten mit Allergien und anderen immunologischen Störungen aufgesucht wurde, die hier offensichtlich Milderung oder Heilung fanden, verlagerte sich das Interesse der Forschung in Richtung dieser Krankheiten.
Ein immunologisches Forschungsprogramm, wie es der neue bis heute aktive Leiter der Medizinischen Abteilung, Ernst Sorkin, seither verfolgt, war also naheliegend. Die bis dahin realisierten zahlreichen immunologischen Arbeiten beschäftigen sich vor allem mit Fragen der Immunregulation, da allergische Reaktionen zweifellos Fehlleistungen des Immunsystems darstellen. Zur Zeit werden die Wechselwirkungen zwischen dem Immunsystem und dem neuroendokrinen System untersucht. Die Davoser Arbeitsgruppe hat den ersten eindeutigen Beweis erbracht, dass das Immunsystem durch das Gehirn kontrolliert wird und seinerseits ebenfalls die Hirnfunktion erheblich beeinflusst. Es ist hier nicht möglich, in wenigen Worten darüber zu referieren, doch wurden diese Arbeiten kürzlich mit dem Otto-Nägeli-Preis und dem Wissenschaftspreis der Stadt Basel ausgezeichnet, was für ihre Bedeutung und Qualität spricht.
Im Hinblick auf die Wandlungen in der medizinischen Forschungsrichtung ist der Name des Forschungsinstituts in erweiterndem Sinn abgeändert worden in «Schweizerisches Forschungsinstitut für Hochgebirgsklima und Medizin».
Was geschah inzwischen mit der anderen Abteilung des Forschungsinstituts, mit Dornos Observatorium? Nach einer dreijährigen Betreuung durch den Schweden I.F. Lindholm übernahm 1929 der Basler Physiker Walter Mörikofer die Leitung und blieb dieser Aufgabe während 37 Jahren treu. Mit ständig verbessertem Instrumentarium wurden die Strahlungsmessungen von Dorno fortgesetzt und zur längsten existierenden Strahlungsmessreihe erweitert. Grosses Gewicht wurde auf die Erhöhung der Messgenauigkeit und der Eichkonstanz gelegt, wodurch die Davoser Instrumente einen weltweiten Ruf erlangten. An diesen Entwicklungen waren vor allem die Mitarbeiter P. Courvoisier und H. Wierzejewski beteiligt. Strahlungsspezialisten aus zahlreichen Ländern versammelten sich periodisch im Observatorium, um ihre Instrumente zu vergleichen und zu eichen.
Die von Dorno gestellte Aufgabe, das Davoser Klima im Hinblick auf seine Heilwirkung zu definieren, hat Mörikofer ganz allgemein auf die Klassierung des Klimas von Kurorten der Schweiz ausgedehnt und dabei die folgenden Reizstufen unterschieden:

| | |
|---|---|
| Reizstufe 0: | Klimakurorte mit Schonklima (6 Orte, 200–600 m ü. M.) |
| Reizstufe 1: | Klimakurorte mit leichten Reizfaktoren (17 Orte, 400–1100 m ü. M.) |
| Reizstufe 2: | Klimakurorte mit mässigen bis kräftigen Reizfaktoren, jedoch mit gutem Windschutz als Schonfaktor (13 Orte, 1200–1900 m ü. M., darunter Davos) |
| Reizstufe 3: | Klimakurorte mit intensiven Reizfaktoren und häufig kräftiger Luftbewegung (6 Orte, 1500–1900 m ü. M.) |

Walter Mörikofer war Jahrespräsident der Schweizerischen Naturforschenden Gesellschaft, als diese im Jahr 1950 zum dritten Mal Davos als Tagungsort erkor.
Nach Mörikofers Rücktritt 1966 setzte das Observatorium die klimatologischen Arbeiten unter der Leitung von Emil Flach mit bioklimatischem Schwergewicht fort. Es sind Parallelitäten oder Synergismen festgestellt worden zwischen dem Ablauf von Klimafaktoren und verschiedenen Krankheiten. Aber es muss hier gesagt werden, dass das äusserst komplexe und schwer zugängliche bioklimatologische Problem trotz grosser Anstrengungen von verschiedener Seite noch keineswegs entschlüsselt ist.
Der internationale Vergleich von Strahlungsmessgeräten und die Standardisierung von Eichskalen sind dem Observatorium durch die Meteorologische Weltorganisation (WMO) vor rund 10 Jahren als permanente Aufgaben übertragen worden, wobei die Schweizerische Meteorologische Anstalt als Treuhandstelle der Eidgenossenschaft mitwirkt. In dieser Funktion trägt das Observatorium den Titel eines «Weltstrahlungszentrums», abgekürzt «WRC». Observatorium und WRC werden seit 1975 als Einheit von Claus Fröhlich geleitet. In Fortführung der

Tradition hat sich das WRC wieder mit eigenen Instrumententwicklungen – jetzt ganz auf die moderne Elektronik und Computertechnik ausgerichtet – in grundlegende Aufgaben der Strahlungsmessung eingeschaltet und führt zur Zeit mit Hilfe von Stratosphärenballonen und Raketen neue Bestimmungen der Solarkonstanten, d.h. der extraterrestrisch einfallenden Strahlungsintensität, durch, mit dem Ziel, langfristige Änderungen dieser Grösse festzustellen. Hiezu ist eine Genauigkeit von mindestens 0,1% erforderlich. Im Sinn der Kontinuität des Observatoriums sind auch die sich über 70 Jahre zurück erstreckenden Messungen der direkten Sonnenstrahlung von Davos aufgearbeitet worden. Sie zeigen, dass sich die mittlere Strahlungsdurchlässigkeit der über 1600 m liegenden Atmosphäre seit 1909 nicht klimatisch relevant geändert hat.

Um 1958, als die Medizinische Abteilung des Forschungsinstituts stillgelegt war, schloss sich eine Gruppe von Schweizer Ärzten unter der Bezeichnung Arbeitsgemeinschaft für Osteosynthesefragen zusammen, und ihr Exponent, Martin Allgöwer, damals Chirurg in Chur, baute in unbenützten Räumen des Forschungsinstituts ein Laboratorium für Experimentelle Chirurgie auf. Forschungsthemen waren die Wundheilung, der Schock und vor allem die Osteosynthese, d.h. die operative Behandlung gebrochener Knochen unter Fixierung am Knochen selbst. Der Bedarf für die letztgenannte Schadenbehebung war in Davos mit der Zunahme der Skiunfälle immer vordringlicher geworden, so dass Davos als Standort für diese Forschungsstelle gut ausgewiesen war und es weiterhin ist. Stephan Perren, seit 1967 Leiter des Labors, hat an der letztjährigen Jahresversammlung der SNG in Winterthur ein umfassendes Symposium über Osteosynthese durchgeführt und damit die Arbeiten im Kreis unserer Gesellschaft bekannt gemacht. Die Methode wird in Davos nicht nur erforscht und ständig verbessert, sondern auch in jährlichen internationalen Kursen gelehrt. Sie gewinnt ständig an Boden. Ihre Vorteile liegen gegenüber einer konventionellen Behandlung darin, dass die Reposition der gebrochenen Knochen sehr genau vorgenommen und fixiert werden kann und dass die Bruchstelle bald wieder belastbar ist. Die Entwicklung des Materials und der Instrumente nimmt in dieser Forschung einen wichtigen Platz ein.

In den dreissiger Jahren entfalteten sich in Davos zwei Forschungsrichtungen, die, aus Ingenieurproblemen hervorgehend, die Naturwissenschaft zu Hilfe rufen mussten. Die eine hatte mit dem Wasser zu tun, die andere mit dem Schnee.

Die winterliche Absenkung des Davosersees ab 1923 hatte nicht nur für das Plankton deutliche Folgen, sie führte auch zu einem Rechtsfall zwischen der Gemeinde Davos und den Bündner Kraftwerken als Pächterin des Sees. Es ging um die Auswirkung der Absenkung auf die Wasserführung des Landwassers, auf den Grundwasserspiegel und die Kanalisation. Der Hydrologe Otto Lütschg, ein Pionier seines Fachs, lieferte als Expertise eine monumentale Studie über den gesamten Wasserhaushalt des Landwassertals ab. Vom Niederschlag bis zum Abfluss oder zur Verdunstung wurde, bildlich gesprochen, jedes Wassermolekül in seinem Kreislauf verfolgt. Es dürfte kaum ein alpines Siedlungsgebiet mit einer derart gründlichen hydrologischen Bearbeitung geben. Lütschg schliesst seinen Bericht mit handfesten Empfehlungen, z.B. zum dringenden Bau einer Kläranlage – einer inzwischen erfüllten Forderung.

Das mit dem Schnee verbundene Problem war von alters her in Davos beheimatet: Der Kampf gegen die Lawinen. Dank der historischen Arbeiten des einstigen Bündner Ständerates Andreas Laely besitzt Davos eine Lawinenchronik, die bis ins 15. Jahrhundert zurückreicht und auch als Beitrag an die Naturforschung zu werten ist.

Die heutige Schnee- und Lawinenforschung, die zu einem Aushängeschild von Davos geworden ist, wurde von aussen hierher gebracht. Warum gerade nach Davos? Einige Gründe sind rasch angeführt: Davos hat reichlich Schnee zu bieten, leicht zugängliches Lawinengelände und eine für die Forschung aufgeschlossene Atmosphäre. In den Augen der Einheimischen erschien es gleichwohl etwas befremdlich, als um 1935 junge Leute begannen, in der Umgebung Löcher in den Schnee zu graben und Schichten herauszupräparieren. Es handelte sich um eine aus verschiedenen Wissenschaftsdisziplinen zusammengesetzte Arbeitsgruppe der ETH Zürich, die im Auftrag der 1930 ge-

gründeten Eidgenössischen Schnee- und Lawinenforschungskommission und unter der Leitung des Bauingenieurs und Glaziologen Robert Haefeli grundlegende Studien über die Schneedeckenentwicklung und die Lawinenbildung einleitete. Im Hintergrund stand das praktische Ziel, die Lawinenschutztechnik auf einen wissenschaftlichen Boden zu stellen.

Zu den Förderern dieser Arbeiten gehörten Persönlichkeiten wie die ETH-Dozenten Paul Niggli und Ernst Meyer-Peter sowie der Eidgenössische Oberforstinspektor Marius Petitmermet. Um diese Zeit hat auch der gegenwärtige Zentralpräsident der SNG, Ernst Niggli, als junger Studienabsolvent während eines Winters auf Weissfluhjoch Schnee gesiebt. In den Jahren bis zum Krieg wurden in regelmässigen, mehrmonatigen Winterkampagnen wesentliche Erkenntnisse über die Entwicklung der Schneedecke gewonnen und im besonderen die Schneemechanik geschaffen. Es zeigte sich bald, dass das «Saisonnierstatut» der Schneeforscher nicht genügte, um den Anfall an Problemen zu meistern. Daher wurde 1942 auf Weissfluhjoch als Voraussetzung für eine Ganzjahrestätigkeit das Eidgenössische Institut für Schnee- und Lawinenforschung gebaut und mit permanentem Personal dotiert. Es hat baulich und personell inzwischen wesentliche Erweiterungen erfahren. Während der ersten sieben Jahre amtete Edwin Bucher als Leiter, anschliessend durfte der Autor diese Funktion während dreissig Jahren ausüben, und heute steht das Institut unter der Leitung von Claude Jaccard. Auf der Pionierarbeit von Robert Haefeli und Henri Bader aufbauend, sind verschiedene Forschungsrichtungen verfolgt worden, wobei stets die Verbindung zur Praxis gesucht wurde. Im Zentrum steht die Erforschung der physikalischen Grundlagen von Schnee und Eis. Darum herum gruppieren sich die spezifischen

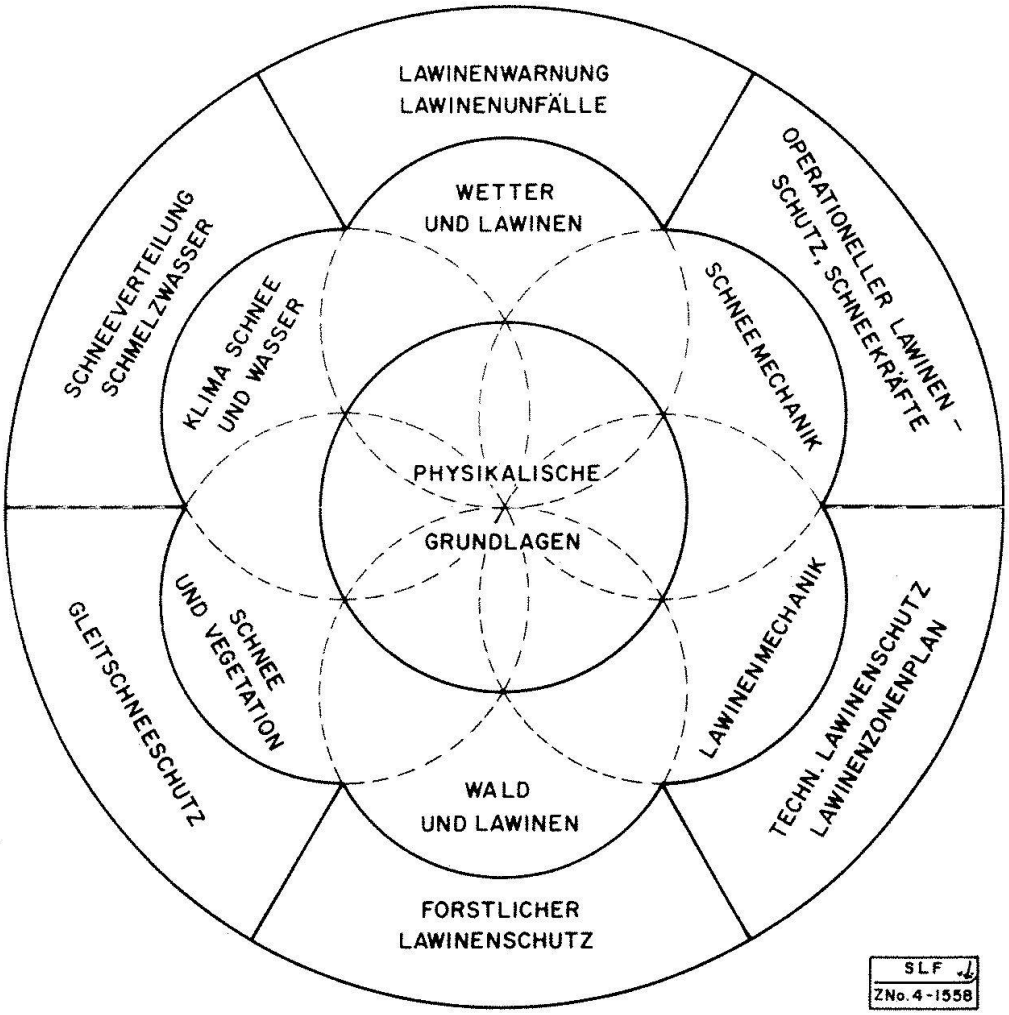



*Abb. 3.* Arbeitsgebiete des Eidgenössischen Instituts für Schnee- und Lawinenforschung in symbolischer Darstellung.

Forschungsrichtungen mit ihren Ausstrahlungen in die Praxis der Lawinenwarnung, des operationellen, technischen und forstlichen Lawinenschutzes und der Schneehydrologie (Abb. 3). Auch hier haben Elektronik und Computer Einzug gehalten. Als Beispiel eines aktuellen Forschungsprojektes sei die Messung von Lawinengeschwindigkeiten mit dem Dopplerradar genannt.
Im Bereich der forstlichen Schnee- und Lawinenforschung hat das Institut in der Eidgenössischen Anstalt für das Forstliche Versuchswesen (Birmensdorf) einen in Davos aktiven Partner. Die beiden Institute bearbeiten am Stillberg im Davoser Dischmatal seit ungefähr 25 Jahren ein langfristiges Versuchsprojekt zur Aufforstung von Lawinengelände an der oberen Waldgrenze. Die Birmensdorfer Gruppe untersucht vorwiegend die ökologischen Grenzbedingungen für die rund 100000 Testpflanzen, während das Institut Weissfluhjoch die Schnee- und Lawinenwirkungen verfolgt. Diese Anlage darf als eines der ökologisch am besten erforschten Biotope gelten. Das Dischmatal ist auch in anderer Hinsicht zu einem Forschungsobjekt geworden, so als schneehydrologisches Testgebiet und in Verbindung mit dem internationalen Programm «Man and Biosphere» (MAB).
Als naturwissenschaftliche Attraktion von Davos darf im weiteren eines der Exkursionsziele der Jahresversammlung genannt werden: Die ehemalige Bergbauanlage am Silberberg und das Bündner Bergbaumuseum im dazugehörigen Knappenhaus. Durch den Davoser Architekten Hans Krähenbühl ist unter Mitwirkung des Zürcher Geologen Kurt Bächtiger und weiterer Kreise eine historisch-geologische Forschung über den Bergbau in Graubünden aufgebaut und eindrücklich dargestellt worden.
Schliesslich wären noch Einzelarbeiten anzuführen, wie sie in der klassischen Zeit im Vordergrund standen, beispielsweise die eingehenden Beobachtungen der letzten Jahre über die Lebensbedingungen der Eulen und Adler von Heinrich Haller.
Nicht behandelt ist hier die systematische geologische Erforschung der Landschaft Davos, die vorwiegend von Bern und Basel ausging und die mit den Namen Joos Cadisch und Albert Streckeisen verknüpft ist, und nicht erläutert wurde die sehr aktive medizinische Forschung in verschiedenen Kliniken.
Eine jüngere Institution besonderer Art, auch mit gesamtschweizerischer Basis, ist das «Forum Davos». Es erinnert an die Hochschulkurse der zwanziger Jahre und dient der interdisziplinären Diskussion und Verbreitung von Forschungsresultaten unter besonderer Berücksichtigung sozialer Aspekte. «Grenzen der Medizin» war ein unlängst behandeltes Thema, «Skifahren und Sicherheit» ein anderes.
Die Frage drängt sich auf: Wovon lebt denn all diese Forschung in Davos? Es sind hier die verschiedensten Träger vertreten. Bund und Nationalfonds sind wichtige Finanzquellen. Hochschulinstitute mit eigenen Projekten in der Gegend kommen selbstverständlich für ihren Bedarf selbst auf. Die Industrie, vorab die chemische aus dem Raum Basel, spendet substantielle Beiträge an biologische Arbeiten, und der Kanton Graubünden und die Gemeinde Davos steuern erhebliche Leistungen bei. Einzig die Experimentelle Chirurgie ist mit ihrer weltweiten Kurstätigkeit privatwirtschaftlich selbsttragend. Das Wort «weltweit» darf übrigens für die Beziehungen aller Davoser Institute angewandt werden. Davos ist längst nicht mehr das «da hinten» von einst. Das erscheint alles recht euphorisch und wohl geregelt, doch fliessen die Mittel, gemessen am Bedarf, recht dosiert, und der Kampf um ein ausgeglichenes Budget zieht sich mehr oder weniger scharf durch alle Forschungsstellen. Diese liegen denn auch in der personellen Dotation durchwegs nahe der kritischen Grenze. Umsomehr darf sich ihre Produktivität sehen lassen. Jährlich werden hier von rund 30 Autoren etwa 50 bis 60 wissenschaftliche Publikationen verfasst.
Abschliessend sei ein persönliches Nachwort gestattet. Es ist nicht der Zweck dieser Übersicht, Davos lokalpatriotisch zu verklären. Die meisten Aktiven in der Forschung sind zugezogene Unterländer, und wir alle – der Autor gehört auch dazu – haben selbst einmal mit Erstaunen zur Kenntnis genommen, was in den beiden letzten Jahrhunderten hier oben alles geschah. Als kleine Gemeinschaft von Naturforschern sind wir eingebaut in die eigenartige Symbiose zwischen einem alteingesessenen Bergbauerntum, einer Genesung und Erholung suchenden Schar von Patien-

ten und ihren Betreuern, einem stets wachsenden Heer von Touristen und Sportbeflissenen und der Geschäftswelt. Es darf hier einmal den unsere Wissenschaft fördernden Kreisen der Schweiz gedankt werden, darunter der Schweizerischen Naturforschenden Gesellschaft, dass sie uns nicht vergessen haben, und anderseits den Davosern, dass sie uns in ihre Lebensgemeinschaft aufgenommen haben.

*Anschrift des Verfassers:*

Prof. Dr. Marcel de Quervain
Tschuggenstrasse 12
CH-7260 Davos-Dorf (Schweiz)

# Vom Ursprung der Dinge
# Symposium vom 24.–27. September 1981 in Davos, Schweiz

# Vorwort

Das Thema «Vom Ursprung der Dinge», von E. Sorkin angeregt, wurde vom Jahresvorstand nicht nur wegen seiner grundlegenden Aktualität, sondern auch als Ansporn zur Interdisziplinarität – im Sinne einer naturwissenschaftlichen Akademie – gewählt. Das hat auch die Spannweite der einzelnen Themen bestimmt: vom Ursprung des Universums zu den Grundlagen des kognitiven Denkens. Die fünf Vorträge im Rahmen des Symposiums können von diesem weiten Gebiet natürlich nur einige Punkte herausgreifen und die Lage der Punkte ist allein bestimmt durch die Vortragenden. Nicht nur die Kompetenz der gewonnenen Vortragenden ist hervorragend, auch die Verteilung der Punkte im grossen Gebiet scheint gleichmässig und ist in jedem Fall repräsentativ: Victor F. Weisskopf: Über den Ursprung des Universums, Hubert Reeves: Origine des éléments chimiques et naissance du système solaire, Manfred Eigen: Evolution des Lebens auf molekularer Ebene, David Hubel: On the Origin of the Brain, Günther S. Stent: Ursprung, Grenzen und Zukunft der Wissenschaften. Werner Arber leitete das Symposium und die Podiumsdiskussion.

Im folgenden sind die Einleitung von W. Arber, die fünf Vorträge und eine gekürzte Fassung der Podiumsdiskussion wiedergegeben. Die Diskussion wurde ab Tonband vom Herausgeber verfasst und von W. Arber, den Teilnehmern, M. de Quervain und E. Sorkin kritisch überarbeitet, was hiermit bestens verdankt wird. Dank gilt auch U. Bühler für das Schreiben und kritische Durchlesen der Manuskripte und U. Wyss für das Anfertigen der Zeichnungen.

Claus Fröhlich

# Einleitung

Werner Arber

Es hat mich gefreut zu vernehmen, dass der Jahresvorstand den «Ursprung der Dinge» als Tagungsthema gewählt hat, und ich fühle mich geehrt, mit dem Vorsitz dieses Symposiums betraut zu sein. Das Wort «Dinge» kann vieles oder nichts aussagen. Mir wäre es wahrscheinlich nie in den Sinn gekommen, dieses Wort in den Titel einer wissenschaftlichen Publikation, eines wissenschaftlichen Vortrages oder eines wissenschaftlichen Symposiums aufzunehmen, doch finde ich den gewählten Titel seiner Einfachheit halber nun sehr elegant. Es gehört zu den Zielen der Naturforscher, sich nicht nur einer Bestandesaufnahme der Natur zu widmen, sondern sich auch Gedanken darüber zu machen, welche speziellen Mechanismen die Natur in ihrem oft komplexen Wirken anwendet. Ich beziehe das auf die belebte und auf die unbelebte Natur. Dabei kommt man zwangsläufig zur Frage, wie alles entstanden sei – d.h. zur Frage nach dem Ursprung, dem Anfang. Eine Antwort auf diese Frage zu finden war schon immer von grossem Interesse und hat die Leute angeregt nachzudenken. Es ist vielleicht eines der wenigen Themen, die nie veralten; es ist so aktuell wie vor hundert oder vielleicht tausend Jahren. Und ich möchte behaupten, dass wir wahrscheinlich in den folgenden Vorträgen auch nicht eine abschliessende Stellungnahme und Diskussion vernehmen werden, sondern dass dieses gleiche Thema noch unsere Nachfahren beschäftigen wird.

Verbunden mit der Frage «Wo kommt alles her?» ist natürlich auch die andere Frage «Wo gehen wir hin?», «Wo geht die Natur hin?». Vielleicht wählt die Naturforschende Gesellschaft ein andermal letzteres Thema. Heute aber fragen wir uns nach dem Anfang. Im Vergleich vielleicht zu allein sachbezogenen Diskussionen ist meiner Ansicht nach in dieser Thematik ein Freiraum für Ideen und noch unbewiesene Spekulationen. Und wir sollten uns nicht stossen, wenn vielleicht aufbauend auf konkreten, soliden Tatsachen, der eine oder andere Redner Visionen bringen wird, die man nur mehr oder weniger glaubt. Das gehört zur Thematik.

In Abbildung 1 zeige ich Ihnen eine kurze Vorschau auf die zu erwartenden Vorträge. Wir starten mit einem Vortrag über das Universum – das Weltall. Wir kennen alle die riesigen Dimensionen des Weltalls, d.h. wir haben eine Idee davon, aber können es uns eigentlich nicht vorstellen. Der zweite Vortrag wird sich mit dem Sonnensystem befassen, das für uns schon etwas überblickbarer, abgegrenzter ist. Dazu gehört ja auch unsere Erde. Auf dieser Erde gibt es Lebewesen, und der dritte Vortrag wird sich mit der Fragestellung des Anfangs des Lebens befassen. Ein wichtiger Aspekt des Lebens, vor allem des Lebens der höheren Tiere und der

| Redner | Behandelt Ursprung von | «Dimension» | wird voraussichtlich sprechen über | «Dimension» |
|---|---|---|---|---|
| V. Weisskopf | Universum | sehr gross | Elementarteilchen | sehr klein |
| H. Reeves | Sonnensystem | ↑ | Atome, chemische Elemente | ↓ |
| M. Eigen | Lebewesen | | Moleküle | |
| D. Hubel | Gehirn | | Zellen | |
| G. Stent | intellektuelle Tätigkeit | klein | Individuum | gross |

*Abb. 1.* Vorschau auf Hauptvorträge.

Menschen, betrifft die Aktivitäten spezialisierter Gewebe, z.B. des Gehirns, welchem sich der vierte Vortrag widmet. Die Thematik des fünften Vortrages schliesslich wird sich mit einem Produkt der im Gehirn verankerten Fähigkeiten, mit den intellektuellen Fähigkeiten und mit der Wissenschaft, einem Teilbereich der menschlichen Kultur, befassen. Die Reihe der Vorträge beginnt somit beim unvorstellbar Grossen und führt zur Erde und zu dem, was darauf geschieht und was von diesem Geschehen ausstrahlt. Wenn man sich nun fragt, was die einzelnen Redner im Besondern behandeln werden, dann vermute ich folgendes: Herr Weisskopf wird höchstwahrscheinlich über Elementarteilchen, Herr Reeves über die chemischen Elemente sprechen. Ich denke, dass Herr Eigen Moleküle und Herr Hubel Zellen behandeln werden. Herr Stent schliesslich wird zur Behandlung seines Themas auf die intellektuelle Tätigkeit einzelner Individuen, Menschen, Bezug nehmen. Wenn wir nun in der einfachen Darstellung von Abbildung 1 für die räumlichen Grössen Werte einsetzen, so fällt uns auf, dass diese Werte in der ersten Kolonne von sehr gross nach klein abnehmen. Im Gegensatz dazu nehmen die Werte der zweiten Kolonne von sehr klein nach gross zu. Vielleicht ergäbe die Multiplikation der Werte der beiden Kolonnen für jede Zeile eine Konstante. Das Thema ist also auch in dieser Hinsicht sehr elegant und abgerundet.

*Anschrift des Verfassers:*

Prof. Dr. W. Arber
Biozentrum der Universität Basel
Abteilung Mikrobiologie
Klingelbergstrasse 70
CH-4056 Basel (Schweiz)

# Vom Ursprung des Universums

Victor F. Weisskopf

Als ich die Einladung der Schweizerischen Naturforschenden Gesellschaft bekam, einen Vortrag über den Ursprung des Universums zu halten, habe ich mich sehr gefreut. Nicht nur wegen der grossen Ehre, die mir damit bewiesen ist, sondern weil dieses Thema so ungeheuer interessant und gross ist, dass es sehr weit über die rein wissenschaftliche Bedeutung hinausgeht. Es hat Beziehungen zur menschlichen Existenz, zur Mythologie, zur Religion und ist als solches eines der aufregendsten Themen, mit denen sich ein Wissenschaftler befassen kann. Ich muss aber gestehen, dass ich kein Experte der Kosmologie bin. Es gibt Leute, die viel mehr davon verstehen, und ich kann meine Wahl durch den Rat nur dadurch begründen, dass Experten oft nicht so klar vortragen wie Nicht-Experten, weil sie nämlich durch die Menge der Einzelheiten daran gehindert sind.

Weiterhin ist das Faszinierende an diesem Thema, dass man automatisch zu Fragen kommt, die mit der fundamentalen Struktur der Materie zu tun haben, nämlich mit den Elementarteilchen. Heute trifft die Wissenschaft des ganz Grossen, die Kosmologie, mit der Wissenschaft des ganz Kleinen, der Elementarteilchen-Forschung, zusammen. In dieser zweiten Richtung muss ich mich leider als Experte betrachten und hoffe, dass ich dann nicht in den Fehler falle, Ihnen darüber zu viel zu erzählen.

Nun muss ich in der Tat mit den Elementarteilchen anfangen, denn nur so können wir verstehen, wie die heutigen Ideen über den Ursprung des Universums zustandegekommen sind.

## Die Struktur der Materie

In der Abbildung 1 habe ich versucht, in ganz elementarer Weise die Stufen der Materie darzustellen. Wir fangen mit einem Stück Metall als Beispiel an: Das besteht aus Atomen. Die Atome werden durch die chemische Kraft zusammengehalten, die an sich keine sehr grossen Energien enthält. Das Elektronenvolt (eV) ist ein Mass der Energie. Die Kräfte, die die Atome zusammenhalten, sind von dieser Grössenordnung. Jetzt nehmen wir ein Atom heraus. Wir wissen dann, dass dieses Atom aus einem Kern und mehreren Elektronen besteht. Sie werden durch die elektrische Kraft zusammengehalten, die bereits etwas stärker ist, in der Grössenordnung von 10 eV. Die Grösse der Atome ist ungefähr $10^{-8}$ cm. Der Kern selbst - vergrössert betrachtet - besteht aus Protonen und Neutronen, die durch die Kernkraft zusammengehalten werden; eine Kraft, die bedeutend stärker ist, nämlich von der Grössenordnung 10 Millionen Elektronenvolt (MeV). Der Kern selbst ist auch sehr viel kleiner, nämlich $10^{-12}$ cm.

Heute weiss man, dass die Protonen und Neutronen, die sogenannten Nukleonen, selbst auch wieder nicht elementar sind, sondern wahrscheinlich aus drei Elementarteilchen bestehen, die den unglücklichen Namen «Quark» führen. Man hätte einen besseren Namen wie «Parton» wählen sollen, aber nun blieb der Term «Quark» stecken.

Hier finden wir wieder viel grössere Kräfte, die die Quarks zusammenhalten. Sie sind nämlich von der Grössenordnung GeV, eine Milliarde eV, und die Grösse dieser Nukleonen ist ungefähr $10^{-13}$ cm. Die Kernkraft selbst, die die Protonen und Neutronen zusammenhält, wird heute als eine Folge dieser starken Kraft verstanden, die die Quarks zusammenhält. Diese Kräfte binden in schwächerem Masse auch die Nukleonen im Kerne zusammen, so ungefähr wie die chemischen Kräfte die Atome zusammenbinden. Die chemischen Kräfte sind ja auch schwächer als die Kräfte im Atom.

*Tab. 1.* Die Quantenleiter.

| Reich | Energie | Ort in der Natur |
|---|---|---|
| 1. Atom-Reich | | |
| Atome, Moleküle, Chemie, feste Körper, Flüssigkeiten, Gase, Plasma, Optik | ~eV | Erde, Planeten, Oberfläche der Sterne |
| 2. Kern-Reich | | |
| Protonen-Neutronen, Kernreaktionen, Fission, Fusion, Radioaktivität | ~MeV | Inneres der Sterne |
| 3. Subnukleares Reich | | |
| Angeregte Protonen, Mesonen, Anti-Materie, Schwere Elektronen, Quarks, Gluonen, kurzlebige Teilchen | ~GeV | kosmische Kataklysmen, Neutronen-Sterne, Urknall |

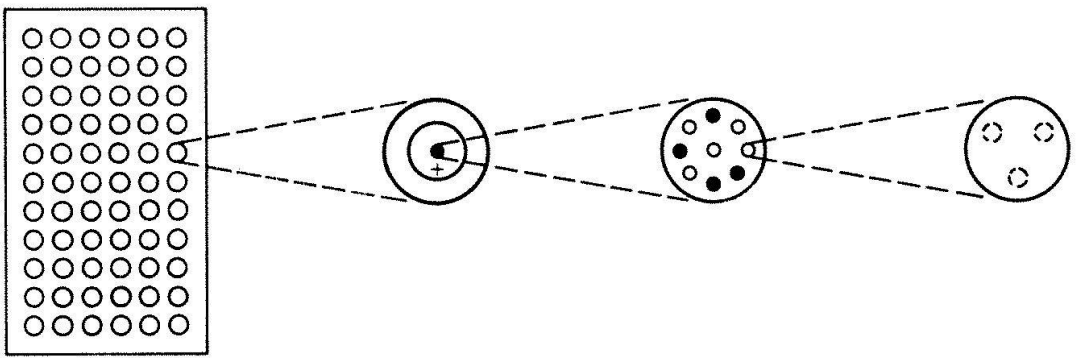
Dann möchte ich noch die Existenz des Phänomens der Radioaktivität erwähnen, das von einiger Wichtigkeit sein wird. Es besteht darin, dass die Quarks ihre Natur ändern. Es gibt Quarks mit verschiedener Ladung und Natur. Der Effekt der Radioaktivität besteht darin, dass ein Quark in ein anderes übergeht und dabei andere Teilchen, Elektronen und Neutrinos emittiert.

Die Sachlage, die in Abbildung 1 angedeutet ist, führt uns zum Begriff der Quantenleiter, die eine wichtige Rolle in der Geschichte des Universums spielen wird (siehe Tabelle 1). Man kann gewissermassen drei Reiche – bis heute wenigstens – unterscheiden. Erstens das atomare Reich, bei dem es sich um die Atome und ihre Wechselwirkungen handelt: Moleküle, Chemie, feste Körper, Flüssigkeiten, Gase, Plasma, Optik und eigentlich auch die Biologie. Die Wechselwirkungen sind von der Grössenordnung von einigen eV. In der Natur finden sich die Prozesse der ersten Stufe der Quantenleiter auf der Erde, auf den Planeten und auf der Oberfläche der Sterne.

Das Kernreich, die nächste Stufe der Leiter, hat viel höhere Energien. Im atomaren Reich sind die Kerne fest, sie verändern sich nicht, sie bleiben wie sie sind. Wenn man aber Energien von Millionen Elektronenvolt anwendet, dann gibt es Kernreaktionen, es gibt Fission und Fusion, und der Prozess der Radioaktivität tritt ein. In der Natur findet man diese Erscheinungen im Innern der Sterne, sie produzieren die Sternenergie. Das bisschen Radioaktivität, das man auf der Erde findet, ist der Überrest aus der Zeit, als die Erdmaterie von irgendeiner Supernova in den Weltraum geschleudert wurde.

Die dritte Stufe ist das subnukleare Reich, dem z.B. das CERN-Laboratorium gewidmet ist, und da handelt es sich um Energien von vielen GeV. Dann findet man angeregte Protonen, neuartige Teilchen wie Mesonen, Antimaterie in grossen Mengen, neben den gewöhnlichen Elektronen noch andere schwere Elektronen, Quarks und Gluonen. Die Gluonen sind die Quanten der starken Kraft, die die Quarks zusammenhält. Und zudem findet man viele, viele kurzlebige Einheiten, die erscheinen und bald wieder verschwinden und sich in andere Energieformen verwandeln. Ich erinnere Sie hier an die Tatsache, dass Materie und Antimaterie, Teilchen und Antiteilchen, wenn sie zusammentreffen, sich gegenseitig vernichten und ihre Massen-Energie in Strahlung verwandeln. Abgesehen von unseren Laboratorien, sind diese Phänomene wahrscheinlich in der Natur nur in kosmischen Kataklysmen zu finden, wie z.B. in Neutronen-Sternen und vor allem im Urknall, am Anfang des Universums, als, wie wir sehen werden, diese Prozesse eine ganz bedeutende Rolle spielten.

«Was sind nun die Elementarteilchen aufgrund unseres heutigen Wissens?»Wir kennen heute fünf Quarks mit verschiedenen Eigenschaften, auf die wir nicht eingehen. Sie tragen elektrische Ladungen, die ein Bruchteil der Einheitsladung sind, ein Drittel oder zwei Drittel. Sie haben verschiedene Massen, was wir heute überhaupt noch nicht verstehen. Das Interessante ist, dass sie sich immer nur in Verbindungen von drei Quarks oder einem Quark mit einem Anti-Quark befinden und niemals – bis heute wenigstens – allein isoliert gefunden wurden. Scheinbar



| Objekt: | Materie | Atom | Kern | Proton oder Neutron |
|---|---|---|---|---|
| Bestandteile: | Atome | Kern<br>Elektronen | Protonen<br>Neutronen | Quarks |
| Grösse (cm): | beliebig | $10^{-8}$ | $10^{-12}$ | $10^{-13}$ |
| Verbindende Kraft: | chemische Kraft | elektrische Kraft | Kernkraft | starke Kraft |
| Energie (eV): | 1 | 10 | $10^{6}$ | $10^{9}$ |

*Abb. 1.* Stufen der Materie.

ist die Kraft zwischen ihnen so stark, dass man sie nicht trennen kann. Das ist aber noch nicht ganz sicher.
Neben diesen Quarks haben wir die Leptonen, deren Hauptrepräsentant das Elektron ist. Nun gibt es, wie ich vorhin andeutete, völlig unverständlicherweise, schwere Elektronen, die sogenannten Myonen und Tauonen, die sich vom Elektron durch nichts unterscheiden als durch ihre Masse. Ausserdem haben sie eine kurze Lebensdauer, denn sie können unter Emission von Neutrinos in gewöhnliche Elektronen übergehen. Sie sind also kurzlebige Dinge, während das gewöhnliche Elektron – wie wir glauben – eine unendliche Lebensdauer hat. Neben diesen Elektronen gibt es dann noch die Neutrinos, die im Universum auch eine wichtige Rolle spielen. Es gibt verschiedenartige Neutrinos, für jedes Elektron ein spezielles.
Die weiteren Elementarteilchen sind die Träger der Wechselwirkungen. Die bis jetzt genannten Teilchen haben Wechselwirkungen, d.h. es gibt Kräfte zwischen ihnen: Wir unterscheiden elektromagnetische Wechselwirkungen, z.B. die Kraft, die die Elektronen an den Atomkern bindet, die schwachen Wechselwirkungen, die die Radioaktivität hervorrufen, und die starken Wechselwirkungen: die Kräfte zwischen den Quarks, die die Quarks zusammenhalten. Und dann gibt es natürlich auch die Gravitation. Solche Wechselwirkungen, solche Kraftfelder, haben auf Grund der Quantenfeldtheorie immer die Eigenschaft, dass sie durch gewisse Teilchen übertragen werden: die Feldquanten. Bei der elektromagnetischen sind es natürlich die bekannten Photonen, bei den schwachen Wechselwirkungen hypothetische Teilchen mit hoher Masse, die sogenannten W- und Z-Bosonen, die bis heute noch nicht gefunden sind; die starke Wechselwirkung zwischen den Quarks ist durch Gluonen übertragen. Bei der Gravitation sind die Feldquanten die sogenannten Gravitronen. Wir werden wenig über diese sprechen. Letzthin hat sich herausgestellt, dass die elektromagnetischen und die schwachen Wechselwirkungen miteinander verknüpft sind.
Die früher beschriebene Quantenleiter führt zu einem Begriff, der wichtig für die Entwicklung des Universums sein wird: bedingte Elementarität. Was ist damit gemeint? Zur Beurteilung, wie elementar ein Teilchen ist, kommt es darauf an, was für Energieaustausche stattfinden. Wenn man z.B. Energieaustausche hat, die unter einem eV liegen, wie z.B. in der Luft zwischen den Luftmolekülen, dann sind sogar die Atome und selbst

die Moleküle Elementarteilchen, weil sie sich ja nicht verändern. Das Wesentliche am Elementarteilchen ist, dass es so bleibt, wie es ist. Und dann gibt es natürlich das Licht, die Photonen. Wenn man dann aber zu einem höheren Energieaustausch geht, z.B. 10000 eV, dann gibt es noch immer die Photonen, aber dann werden die Atome in Kerne und Elektronen auseinandergerissen. Auf dieser Stufe sind also die Kerne und die Elektronen die Elementarteilchen. Wenn wir weitergehen in die Gegend von 100 MeV, dann haben wir wieder die Photonen und Elektronen, aber bei dieser Energie sind die Kerne in Neutronen und Protonen zerlegt, weil die Energie, die sie zusammenhält, kleiner ist als 100 MeV. Es gibt auch Elektronen, und wegen der schwachen Wechselwirkung kommen in diesem Gebiet auch die Neutrinos zur Wirkung. Und wenn man dann noch weitergeht in die Region der GeV, dann kommen eben die erwähnten neuen Erscheinungen heraus, nämlich schwere Elektronen, Neutrinos, Mesonen aller Arten. Auch sollten dann die Träger der schwachen Wechselwirkung aufscheinen. Sie sind bis heute zwar – wie gesagt – noch nicht entdeckt, aber es ist sehr wahrscheinlich, dass sie existieren. Ausserdem gibt es dann noch die Quarks und die Gluonen, die die Quarks zusammenhalten. Es ist natürlich eine interessante Frage, ob wir damit schon bei den wirklichen Elementarteilchen angelangt sind. Vielleicht geht diese Liste weiter und weiter. Darüber möchte ich dann am Ende noch sprechen.

## Das heutige Universum

Bevor wir uns mit der Entwicklung des Weltalls beschäftigen, möchte ich zwei Bemerkungen machen: Alles, was ich hier sagen werde, steht nicht ganz fest. Die Beobachtungen sind natürlich sehr schwierig, und daher wissen wir nicht genau, ob die Dinge wirklich den Tatsachen entsprechen. Die zweite Bemerkung ist die, dass ich die Sachlage vereinfachen werde. Ich tue es aus zwei Gründen: weil ich es nicht besser weiss und um die Sache ein bisschen leichter verständlich zu machen.
Um die Geschichte des Weltalls zu verstehen, muss man einige wichtige Tatsachen über das heutige Universum festhalten. Ich teile sie in vier Abschnitte, über die ich dann im einzelnen sprechen werde.
Erste Tatsache: das heutige Universum besteht eigentlich nur aus Wasserstoff und Helium. Ungefähr 92 Prozent aller Atome im ganzen Weltall sind Wasserstoffatome, ungefähr 8 Prozent Helium, und der Rest, nur 0,1 Prozent, sind Atome anderer Art wie etwa Sauerstoff, Kohlenstoff, Eisen usw.
Die zweite Tatsache ist vielleicht eines der überraschendsten Phänomene, das in den letzten 80 Jahren entdeckt worden ist; es ist die Ausdehnung, die Expansion des Universums. Heute kann man mit ziemlicher Sicherheit behaupten, dass sich das Universum tatsächlich ausdehnt, und zwar so, dass sich die Dinge um so rascher von uns weg bewegen, je weiter sie von uns entfernt sind.
Der dritte Punkt ist die heutige Verteilung der Materie im Weltraum. Das ist ein sehr schwieriger Punkt, denn es kann ja immer Materie existieren, die man nicht beobachtet hat. Man sieht nur leuchtende Materie, aber man kann Schlüsse ziehen über weitere Materie durch Gravitations- und andere Effekte. Wir interessieren uns hier weniger für die detaillierte Verteilung wie Sternsysteme oder Ansammlungen von Galaxien. Wichtig für uns ist die Dichte der Materie, gemittelt über ganz grosse Gebiete, die sehr viele Sternsysteme enthalten. Jedenfalls weiss man heute ungefähr, sehr ungefähr, dass es durchschnittlich $10^{-6}$ Protonen oder Neutronen pro Kubikzentimeter gibt, wenn man über sehr grosse Distanzen mittelt. Die Dichte ist natürlich viel grösser innerhalb eines Milchstrassensystems und viel kleiner in den Zwischenräumen. Wir geben hier den Durchschnittswert an.
Die vierte Tatsache ist von grosser Wichtigkeit: Der ganze Raum ist von einer Wärmestrahlung erfüllt, deren Temperatur ungefähr drei Grad über dem absoluten Nullpunkt (3 °K) liegt. Aller Wahrscheinlichkeit nach ist sie eine Strahlung, die zu einer sehr frühen Periode des Weltalls ausgesandt wurde. In diesem Sinne kann sie als der optische Widerhall des Urknalls betrachtet werden, der vor vielen Milliarden Jahren stattgefunden hat. Dass man also sozusagen den Urknall heute noch «sehen» kann, ist wohl eine der überraschendsten Entdeckungen der modernen Wissenschaft.

Ich möchte jetzt über einige dieser Punkte näher im Detail sprechen, und zwar fangen wir mit der Expansion an. Ich werde in meinen Ausführungen durchwegs annehmen, dass das Universum unendlich ist. Das ist nicht ganz sicher; wenn die Dichte der Materie hoch genug ist, könnte es nach Einstein ja auch ein endliches Universum sein, das aber keine Grenzen hat, indem es geschlossen ist wie die Oberfläche einer Kugel im Zweidimensionalen. Aber es genügt, um die Prinzipien zu verstehen, wenn wir das vergessen und annehmen, was durchaus den Tatsachen entsprechen könnte, dass das Universum wirklich unendlich ist.

Die Expansion des Weltalls wurde entdeckt, als man fand, dass sich kosmische Objekte von uns weg bewegen, und zwar umso rascher, je weiter sie von uns entfernt sind. Die Geschwindigkeit dieser Wegbewegung beträgt ungefähr 15 km in der Sekunde pro Million Lichtjahre Entfernung. Diesen Geschwindigkeitszuwachs per Distanz nennt man «Hubble-Konstante». Man muss sich diese Expansion als eine Verdünnung vorstellen. Überall im ganzen unendlichen Raum verdünnt sich die Materie und jeder Punkt könnte dann die Mitte dieser Verdünnung sein. Abbildung 2 ist ein eindimensionales Beispiel. Wir haben da Punkte Z, A, B, C, D. Wenn wir uns in A befinden, dann sehen wir Z sich nach links und B sich nach rechts bewegen und C, weil es weiter weg ist, sich stärker nach rechts und D sich noch stärker nach rechts bewegen. Aber wenn jemand sich in B oder C befindet, würde er natürlich genau dasselbe sehen, relativ gesprochen, so dass diese Ausdehnung keinen Punkt des Universums auszeichnet, also im ganzen unendlichen Universum überall dieselbe ist.

Aus dieser Ausdehnung oder Verdünnung könnte man sofort einen merkwürdigen

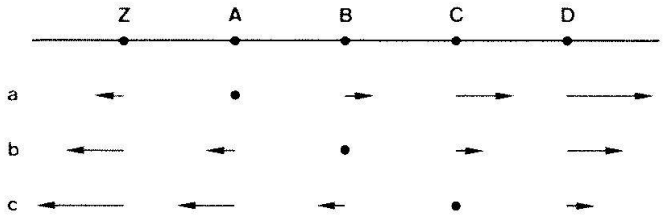



*Abb. 2.* Eindimensionales Beispiel der Expansion des Weltalls. Die Zeilen a, b, c zeigen die Geschwindigkeit von A, B, C aus gesehen.

Schluss ziehen: wenn die Hubble-Konstante wirklich konstant ist, so kann man ja auch zurück extrapolieren. Dann kann man sich leicht ausrechnen, dass vor 20 Milliarden Jahren die Dichte unendlich gewesen sein müsste. Allerdings nur dann, wenn die Expansion immer mit demselben Tempo gegangen wäre. Nun kann man aber leicht schliessen, dass sich die Expansion verlangsamt, weil sich die Massen im Universum ja durch die Gravitation anziehen, was gegen das Auseinanderstreben wirkt. Die Verlangsamung kann man mit Hilfe der halbwegs bekannten Massendichte berechnen. Man kommt zum Resultat, dass tatsächlich der Zeitpunkt der unendlichen Dichte, diese Singularität, vor kürzerer Zeit als 20 Milliarden Jahren stattgefunden hat, weil es ja am Anfang rascher gegangen ist. Diese Rechnung zeigt, dass es eher vor 18 Milliarden Jahren war. Jedenfalls kann man ungefähr das Datum des Urknalls bestimmen. Dieser Zeitpunkt ist der natürliche Nullpunkt der Zeit. Das ist die moderne Zeitrechnung: Sie fängt mit dem Urknall an.

Nun möchte ich eine wichtige Bemerkung machen: In vielen populären Darstellungen ist der Urknall nicht richtig beschrieben. Der Urknall ist kein lokales Phänomen. Es ist nicht so, dass der Urknall an einer Stelle des Universums stattgefunden hat und die Materie dann von diesem Punkt nach allen Richtungen weggeschleudert wurde. Nein, der Urknall ist der Anfang einer Dekompression des unendlichen Universums. Das heisst, zur Zeit «Null» war das ganze Universum, der unendliche Raum, mit einer unendlichen Dichte erfüllt, was schwer vorstellbar ist. Gleich danach wurde sie endlich, aber sehr hoch. Sie war aber überall dieselbe; mit der Zeit dehnte sich das Universum aus, und die Dichte nahm überall gleichmässig ab. Man könnte einwenden, das geht doch nicht, wohin dehnt es sich denn aus? Nun, es war ja unendlich, zehnmal unendlich ist ja auch unendlich. Also hat der Urknall überall stattgefunden, und es ist nicht so, dass er an einem Punkt stattfand, und dass die Materie sich von diesem Punkt aus in eine leere Umgebung verbreitet hat. Falls das Weltall heute endlich ist, aber ohne Grenzen, so wäre es allerdings zur Zeit des Urknalls zu einem Punkt zusammengeschrumpft. Dann aber gäbe es kein «ausserhalb» dieses Punk-

tes. In diesem Sinne hätte dann der Urknall auch überall stattgefunden.
Nun möchte ich einen neuen Begriff einführen: den Begriff des Kommunikationsradius. Heute sehen wir sehr ferne Sterne. Aber natürlich kann uns Nachricht nur von Distanzen kommen, die kleiner sind als etwas, das ich $R_K$ nenne, eben den Kommunikationsradius. Es kann uns nur Nachricht kommen von Punkten, von denen die Zeit, die das Licht braucht, herzukommen, kürzer ist als die Zeit vom Urknall bis heute. Das heisst, wenn die heutige Zeit, vom Urknall gerechnet, ungefähr 18 Milliarden Jahre zählt, so ist der heutige Kommunikationsradius 18 Milliarden Lichtjahre, und nicht mehr. Von weiter her können wir keine Nachricht bekommen. Früher natürlich, sagen wir vor 10 Milliarden Jahren, war der Kommunikationsradius kleiner, weil weniger Zeit seit dem Urknall vergangen war. Man beachte, dass das, was wir heute an der Grenze von 18 Milliarden Lichtjahren sehen, noch nie vorher in Kontakt mit uns war. Es konnte nicht, weil das Licht nicht genügend Zeit hatte, herzukommen. Zum Beispiel die Drei-Grad-Strahlung, die wir heute von allen Richtungen sehen, kommt von so weit weg, dass wir heute zum ersten Mal in einer Kommunikationsverbindung mit den Dingen sind, die sie ausgestrahlt haben. Das hat eine wichtige Konsequenz, wie wir bald sehen werden.
Nun komme ich zur Drei-Grad-Strahlung. Ich sagte, der Raum sei erfüllt mit einer Wärmestrahlung von 3 °K, die von allen Richtungen kommt. Es ist der optische Nachhall des Urknalls. Die Intensität dieser Strahlung entspricht ziemlich genau dem, was wir von einer solchen Wärmestrahlung erwarten: ungefähr 1000 Photonen pro Kubikzentimeter. Das ist eine Milliarde mal mehr als Protonen. Aber energetisch ist das nicht sehr viel, denn die Photonen einer Drei-Grad-Strahlung haben eine sehr kleine Energie. Tatsächlich ist die Energie der Wärmestrahlung ungefähr ein Viertausendstel der Massenenergie $mc^2$ der Materie.
Die Temperatur von 3 °K kann man auch als die heutige Temperatur des Weltalls betrachten, denn die Temperatur der heissen Sterne, die es da überall gibt, macht gar nichts aus, wenn man über den ganzen Raum mittelt. Es ist die Lichtstrahlung, die die Temperatur bestimmt. Nun komme ich zu einer wichtigen Feststellung: Die Drei-Grad-Strahlung scheint ganz uniform zu sein, d.h. sie hat gleiche Intensität von allen Richtungen. Sie scheint also den Raum völlig gleichmässig zu erfüllen. Daher ist es möglich durch den Dopplereffekt die Bewegung der Erde, des Sonnensystems und sogar des Milchstrassensystems zu beobachten. In der Tat findet man in einer Richtung die Frequenzen der Drei-Grad-Strahlung ganz wenig höher und in der umgekehrten Richtung etwas tiefer, was man nur so deuten kann, dass unser Sonnensystem sich in Richtung der höheren Frequenz bewegt. Wenn man die Geschwindigkeit dieser Bewegung dann ausrechnet, so sind es ungefähr 300 km/Sekunden. Bei Analyse der Richtung stellt sich heraus, dass es teilweise die Bewegung des Sonnensystems in der Milchstrasse ist, aber teilweise auch die Bewegung der Milchstrasse gegen den sog. Virgo-Cluster. Es ist interessant, dass es damit sinnvoll wird, von einer absoluten Bewegung zu sprechen: Der grosse Traum von Michelson und Morley ist in Erfüllung gegangen. Sie wollten doch die absolute Bewegung der Erde messen und nahmen an, dass die Lichtgeschwindigkeit in der Richtung dieser Bewegung verschieden sei von derjenigen in anderen Richtungen. Nach Einstein ist die Lichtgeschwindigkeit aber in jedem System unabhängig vom Bewegungszustand immer die gleiche. Aber die Drei-Grad-Strahlung gibt uns ein fixes Koordinatensystem. Es ist an jeder Stelle des Universums sinnvoll, zu sagen, was der Ruhezustand ist: Es ist jener Zustand, in dem die Drei-Grad-Strahlung uniform erscheint. Die tiefere Bedeutung dieser Erkenntnis ist noch nicht klar.

## Die Entwicklung des Universums

Was können wir aus den heutigen Gegebenheiten des Weltalls schliessen? Offensichtlich ist das Universum nicht statisch; es verändert sich. Wegen der Expansion nimmt die Dichte der Materie mit der Zeit ab, während im grossen Durchschnitt die Materiendichte überall im Raum dieselbe zu sein scheint. Der Zeitablauf bestimmt auch den Dichteablauf, nur in der umgekehrten Richtung.

Mehr Dichte, frühe Zeit; weniger Dichte, späte Zeit.
Nun wenden wir das Prinzip der Erhaltung der Energie an. Die grössere Dichte verringert die potentielle Energie der Gravitation, weil wegen der höheren Dichte die Anziehung grösser wird. Die kinetische Energie muss daher wachsen, um die Gesamtenergie gleich zu erhalten.
Die kinetische Energie ist hauptsächlich die Energie der wegeilenden Materie in der Expansion. Daher muss die Geschwindigkeit zunehmen, wenn man in der Zeit zurückgeht. Damit erklärt sich die Verlangsamung der Expansion in der positiven Zeitrichtung. Das kann man genau berechnen: wie schon früher erwähnt, ist die Zeit vom Urknall bis heute etwas kleiner, als man direkt von der heutigen Expansion und Dichte folgern würde.
Nun möchte ich Ihnen ein paar Kurven zeigen, die im wesentlichen das Ergebnis dieser Rechnung sind. Um diese Resultate quantitativ darzustellen, wollen wir das Schicksal der Materie betrachten, die sich heute innerhalb des Kommunikationsradius befindet; das ist jene Materie, mit der wir heute kommunizieren.
In früherer Zeit war dieselbe Materie innerhalb eines kleineren Radius enthalten, denn sie hatte sich ja inzwischen ausgedehnt. Nennen wir R den Radius, der jene Materie umschliesst. Heute ist R gleich dem heutigen Kommunikationsradius, nämlich 18 Milliarden Lichtjahre oder $10^{28}$ cm. «Heute» ist $3 \times 10^{17}$ Sekunden in unserer Zeitrechnung. In der halben Zeit – $1{,}5 \times 10^{17}$ Sekunden nach dem Urknall – war R nur $0{,}63 \times 10^{28}$ cm, etwas mehr als die Hälfte. Abbildung 3 zeigt die Zeit-Abhängigkeit von R.
Nun wenden wir uns der Temperatur im Weltall zu. Heute ist sie 3 °K. Die Temperatur ist immer umgekehrt proportional zum Radius R. Das sieht man so: Wenn sich die Materie ausdehnt, so dehnt sich auch das Licht aus, d.h. die Wellenlängen werden grösser und ihre Frequenz dementsprechend kleiner. Nun ist aber die Temperatur der Wärmestrahlung proportional der Frequenz des Lichtes. Also bedeutet eine Verdopplung der Ausdehnung eine Halbierung der Temperatur. Umgekehrt, als der Radius R halb so gross war, war sie zweimal so gross (6 °K). Es war heisser in früheren Zeiten. Die gebrochene Linie in Abbildung 3 stellt die Zeitabhängigkeit der Temperatur dar. Zur Zeit Null (Urknall) war sie natürlich unendlich hoch.
Abbildung 3 enthält auch die Werte des Kommunikationsradius $R_K$. Er ist ja nichts anderes als die Distanz, die das Licht seit dem Urknall zurückgelegt hat. Die gestrichelte Linie gibt seinen Wert an. Er wird auch kleiner, wenn man in der Zeit zurück geht, und zwar stärker als R. Man sieht in der Figur, dass er unterhalb R liegt. Das heisst, die äusseren Teile der Materie, die heute innerhalb des jetzigen Kommunikationsradius liegen, haben in der Vergangenheit noch nicht mit uns kommuniziert. Wir haben darauf schon früher hingewiesen.
Jetzt wenden wir uns den Ereignissen zu, die zu sehr kurzen Zeiten stattgefunden haben, so kurz, dass sie in Abbildung 3 nicht sichtbar sein können. Um zu sehen, was kurz nach dem Urknall stattfand, verwenden wir in Abbildung 4 eine logarithmische Zeitskala, die die Zehnerpotenzen der Zeit angibt. Das rechte Ende der Abszisse entspricht «heute», da heute etwa $10^{17}$ Sekunden seit dem Urknall vergangen sind; das linke Ende der Zeit einer Millionstelsekunde nach dem Urknall. Nachdem die Temperatur sich reziprok zum Radius R verhält, können in diesem logarithmischen Diagramm beide durch die gleiche Kurve dargestellt werden, wenn die Skala des Radius nach oben zunimmt, die Skala der Temperatur aber nach oben abnimmt.
Wir sehen in Abbildung 4, dass am Zeitpunkt von $10^{13}$ Sekunden (ungefähr 300000

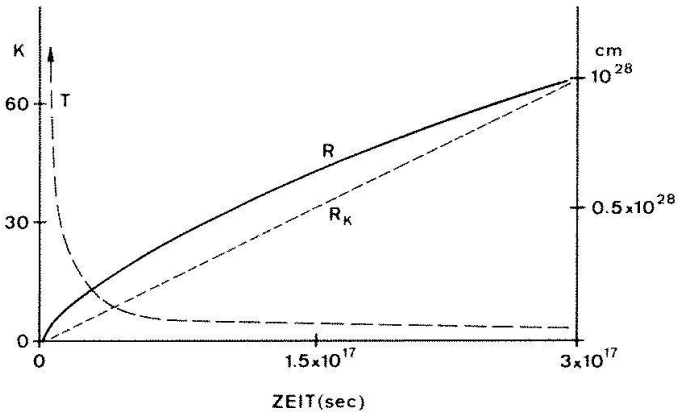



*Abb. 3.* Massenradius R, Kommunikationsradius $R_K$ (Skala rechts) und Temperatur (Skala links) als Funktion der Zeit seit dem Urknall.

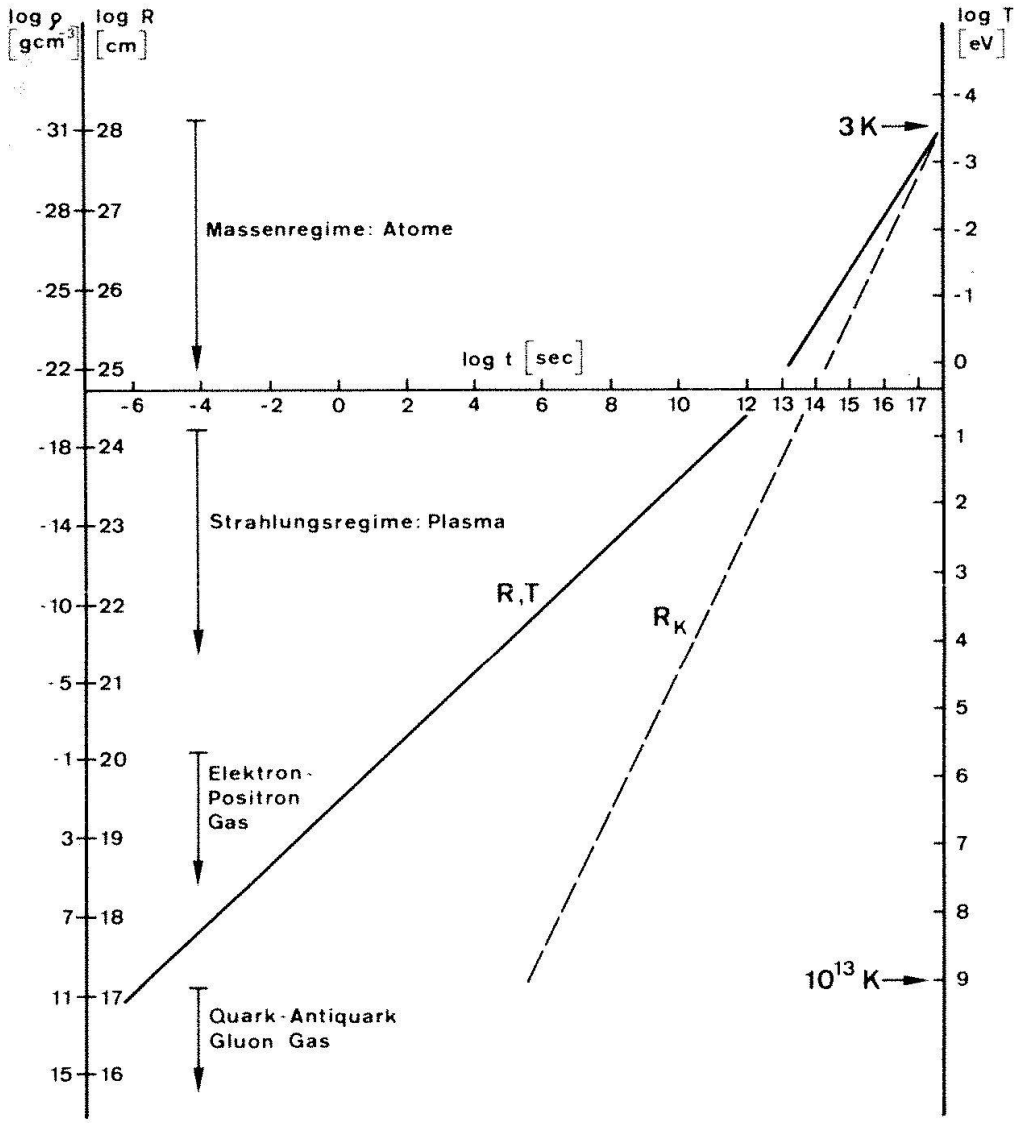



*Abb. 4.* Massenradius, Dichte, Temperatur und Kommunikationsradius als Funktion der Zeit in logarithmischer Darstellung. Die Ordinate links gibt den Logarithmus von R an und auch den Logarithmus der Dichte (g/cm³) einschliesslich des «Gewichtes» der Strahlung. Die Ordinate rechts zeigt den Logarithmus der Temperatur. Hier ist die Temperatur in Energie-Einheiten (Elektronenvolt) gemessen. Es ist die Energie der Wärmebewegung pro Teilchen. Ein Grad Kelvin (1 K) entspricht etwa $10^{-4}$ Elektronenvolt.

Jahre nach dem Urknall) eine Änderung eintrat. Um diese Zeit war die Temperatur etwa 4000 °K; das ist fast ein Elektronenvolt. Bei dieser und bei höheren Temperaturen kann es keine Wasserstoff- oder Heliumatome geben, sondern nur ein Plasma von Kernen und Elektronen. Die Wärmeenergie ist nämlich gross genug, um die Elektronen von den Kernen abzulösen. Ein Plasma ist für Licht undurchdringlich, es ist undurchsichtig, während Atome durchsichtig sind, wenn sie nicht in zu grossen Konzentrationen vorhanden sind. Vor diesem Zeitpunkt herrschte ein anderes Regime, das sogenannte Strahlungsregime. Nach diesem Zeitpunkt war die Massenenergie viel grösser als die Strahlungsenergie. Daher nennen wir die Periode von damals bis heute das Massenregime. Vor dem Zeitpunkt von $10^{13}$ Sekunden befand sich die meiste Energie in der Lichtstrahlung, daher der Name Strahlungsregime.

Eine interessante Folge der Undurchsichtigkeit des Universums vor diesem Zeitpunkt besteht im folgenden: Was wir heute als Drei-Grad-Strahlung beobachten, ist, genau genommen, nicht die Strahlung des Urknalls selber, sondern die Strahlung, die um diesen Zeitpunkt – 300000 Jahre nach dem Urknall – geherrscht hat. Denn erst nachher konnte sie frei und ungehindert durch den Raum dringen. Die Expansion des Raumes hat sie dann von einer Wärmestrahlung von 4000 °K auf 3 °K herabgesetzt. Was wir heute als diese Strahlung beobachten, wurde also 300000 Jahre nach dem Urknall, d.h. vor fast zwanzig Milliarden Jahren, von entfernten Regionen ausgesandt. Das zeigt übrigens, dass der Urknall und die darauffolgende Entwicklung überall im ganzen Raum stattgefunden hat. Sonst könnte man ja nicht heute «Nachricht» von dieser Entwicklung aus allen Richtungen bekommen.

Ein weiterer wichtiger Zeitpunkt war 10 Sekunden. Damals war die Temperatur ungefähr eine Million Elektronenvolt (MeV). Vor dieser Zeit, also bei noch höheren Temperaturen, war der Raum nicht nur mit Lichtstrahlung erfüllt, sondern auch mit Elektronenpaaren, positiven und negativen. Die Erzeugung eines Paares bedarf gerade ein MeV und wenn die Temperatur ungefähr so hoch oder höher ist, dann füllt sich das Vakuum mit einem dichten Gas von Elektronen und Positronen. Wir haben dann ein erweitertes Strahlungsregime; nicht nur Licht, sondern auch, wenn wir das Strahlung nennen dürfen, Elektronen und Positronen im thermischen Gleichgewicht. Übrigens war zu jener Zeit die Gesamtdichte schon mehr als ein Kilogramm per Kubikzentimeter.

Der nächste wichtige Zeitpunkt ist eine Millionstelsekunde nach dem Urknall. Vor diesem Zeitpunkt war die Temperatur höher als eine Milliarde Elektronenvolt (GeV). Schon als die Temperatur etwa 100000 eV und höher war (weniger als 1000 Sekunden), waren die Heliumkerne in Protonen und Neutronen aufgelöst. Aber bei Temperaturen höher als ein GeV werden die Quarks nicht mehr innerhalb der Protonen und Neutronen festgehalten. Übrigens war damals die Dichte so gross, dass sich die Protonen und Neutronen überlappt hätten. Ausserdem war die Energie so hoch, dass viele neue Quark-Antiquark-Paare erzeugt wurden und ein dichtes, heisses Gas von Quarks und Antiquarks bildeten. Dieses Gas enthielt auch sehr viele Gluonen, die Quanten des Kraftfeldes, das zwischen den Quarks wirkt. Das Weltall war also vor diesem Zeitpunkt erfüllt von einem Quark-Antiquark-Gas hoher Dichte, einem Gluongas, einem Elektron-Positron-Gas und von der optischen Wärmestrahlung (Photongas).

Als nach ein paar Millionstelsekunden die Temperatur unter ein GeV sank, verschwanden die Gluonen und die Quark-Antiquark-Paare, indem sie sich in Strahlung verwandelten; was übrig blieb waren die ganz wenigen überzähligen Quarks, die sich dann zu Protonen und Neutronen verbanden. Erinnern wir uns nämlich daran, dass es heute im Raum viel weniger Protonen oder Neutronen gibt als Photonen in der Drei-Grad-Strahlung. Die Zahl der Photonen war immer ungefähr dieselbe, nur waren in früheren Zeiten die Photonen viel «heisser», d.h. energiereicher. Die Anzahl der Gluonen und der Quark-Antiquark-Paare, die in der Zeit vor einer Millionstelsekunde existierten, war etwa ebenso gross. Es war also nur ein ganz kleiner Quarküberschuss, der sich dann zu Protonen oder Neutronen verband.

Hier möchte ich einen Punkt hinzufügen, der bisher vernachlässigt wurde. Bei Temperaturen höher als etwa 1 MeV werden auch viele Neutrinos erzeugt. Daher müssen wir noch ein weiteres Gas hinzufügen; es ist das Neutrino-Antineutrino-Gas. Da die Neutrinos nur sehr schwach mit der Materie wechselwirken, wurde das Universum schon früher für Neutrinos «durchsichtig» als für Licht. Daher ist das Weltall heute wahrscheinlich von einem Neutrinogas erfüllt, das sich etwa 100000 Jahre nach dem Knall freigemacht hat. Die Expansion müsste dann die Temperatur dieser bis jetzt zwar noch nicht beobachtbaren Neutrinostrahlung auf etwa 2 °K herabgedrückt haben.

Was an noch früheren Zeitpunkten als $10^{-6}$ Sekunden stattfand, ist recht hypothetisch. Wenn wir noch näher zum Anfangspunkt rücken, werden die Temperaturen viel höher als 1 GeV. Wir wissen wenig über das Verhalten der Materie unter solchen Bedingungen.

Zusammenfassend möchte ich nochmals einen kurzen Überblick über die Geschichte des Weltalls geben, diesmal in der richtigen Reihenfolge, vom Anfang bis heute. Am Anfang so wie jetzt: immer war das Universum mit Licht erfüllt. Aber am Anfang war es ein Licht von ungeheurer Intensität und Temperatur, während es heute kaltes Licht ist. Wir wissen wenig, was zwischen 0 und $10^{-6}$ Sekunden geschah. Dann aber finden wir – neben dem Licht – ein dichtes Gas von Quarks, Antiquarks, Gluonen, Elektronen und Positronen und Neutrinos. Nach zirka $10^{-5}$ Sekunden hat sich das Gas so weit abgekühlt, dass sich die Quark-Antiquark-Paare vernichtet haben, die Gluonen verschwunden sind und dass sich die wenigen übriggebliebenen Quarks zu Protonen und Neutronen verbunden haben. Nach weiterer Ausdehnung und Abkühlung, etwa 10 Sekunden nach dem Anfang, haben sich die Elektron-Positron-Paare vernichtet, und nur wenige Elektronen sind übriggeblieben. Nach ein paar Minuten vereinigten sich die

Neutronen mit einem Teil der Protonen zu Helium-Kernen. So verblieb es relativ lange, bis ungefähr zum Jahr 300000, als die Temperatur und die Dichte so tief waren, dass sich Atome (Wasserstoff und Helium) bilden konnten. So verblieb es im wesentlichen bis heute: die Temperatur der Strahlung fiel auf 3 °K, und in der Zwischenzeit haben sich Sterne und Galaxien gebildet. Hier endet unsere Skizze des Ursprungs des Weltalls.

**Die ungelösten Probleme**

Bisher haben wir die Geschichte des Universums nach rückwärts in der Zeit verfolgt bis etwa $10^{-6}$ Sekunden. Es ist wohl am Platze, nochmals zu betonen, wie unsicher und hypothetisch unsere Schlussfolgerungen waren. Ausserdem stehen wir vor grundlegenden Problemen, die bisher noch gar nicht beantwortet sind. Diese Probleme lassen sich in vier Gruppen einteilen: A: Das Horizontproblem; B: Was geschah vor $10^{-6}$ Sekunden? C: Der Ursprung des Protonenüberschusses; D: Die kritische Materiedichte; E: Was war vor dem Urknall?
Wir beginnen mit dem ersten Problem: Es wurde bereits betont, dass die Drei-Grad-Strahlung uniform und homogen ist; sie ist von allen Richtungen die gleiche. Es ist heute das erste Mal, dass jene Stellen, die uns die Drei-Grad-Strahlung senden, überhaupt mit uns in Kontakt sind, dass wir von ihnen oder sie von uns etwas «wissen» konnten. Noch weniger konnten die Stellen je in Verbindung gewesen sein, die uns die Drei-Grad-Strahlung aus entgegengesetzter Richtung zustrahlen. Wieso haben sie dieselbe Dichte und Temperatur, wieso ist das alles uniform? Es sieht so aus, als ob am Anfang jemand – ich will den Namen nicht nennen – bestimmt hat, dass die Dichte überall dieselbe sei. Es kann auch sein, dass die Naturgesetze so beschaffen sind, dass am Anfang nur ein und dieselbe Dichte möglich war. Das wäre eine wissenschaftlich befriedigende Lösung. Aber nachdem wir nicht wissen, was vor $10^{-6}$ Sekunden geschehen ist, können wir eben auf diese Frage keine Antwort geben, denn um die Zeit «Null» herum muss eine Physik gültig sein, die Energieaustausche behandeln kann, welche weit über denen liegen, die wir bis heute studiert haben. Und daher wissen wir nicht, wie die Gesetze sind, und wir können weder bejahen noch verneinen, dass es nur eine mögliche Dichte gegeben hat.
Diese Homogenität des Weltalls ist eine ganz grosse Frage. Es ist eigentlich dieselbe Frage, warum alles gleichzeitig angefangen hat, warum also der Urknall auf der ganzen unendlichen Welt im selben Moment geschah. Wenn dem nicht so wäre, würde man von einigen Regionen eine andere Strahlung bekommen, als von anderen. Dies ist noch ein grosses Geheimnis.
Nun zum zweiten Problem: Was geschah vor $10^{-6}$ Sekunden? Die ehrlichste Antwort wäre: Wir wissen gar nichts. Aber die Wissenschafter lieben es, zu spekulieren und Hypothesen aufzubauen. Es handelt sich um Temperaturen höher als 100 GeV. Es könnte ja sein, dass das Quark und das Elektron nicht wirklich elementar sind. Erinnern Sie sich an die bedingte Elementarität. Die Quarks und die Elektronen könnten ja aus Infraquarks und Infraelektronen bestehen. Bei genügend hoher Temperatur würde dann ein Gas aus diesen Teilchen und deren Antipartnern existiert haben.
Es gibt jetzt Hypothesen, die alle Naturkräfte auf eine einzige (neben der Schwerkraft) reduzieren wollen. Die heute geläufige unter diesen Theorien ist die sogenannte Grand Unification-Theory. In dieser Theorie gibt es keine Infra-Quarks oder Infra-Elektronen. Sie gibt dafür recht überzeugende Gründe an, nämlich: die jetzt bekannten Teilchen scheinen sich in eine schöne Ordnung einzureihen, in welcher kein Platz für die Infra-Quarks oder Infra-Elektronen wäre. Aber wer weiss, ob diese Einteilung die richtige ist. Wenn also keine Substruktur vorhanden ist, dann erzählt uns diese Theorie, dass bei $10^{15}$ GeV, das ist $10^{-36}$ Sekunden nach dem Urknall, alle Wechselwirkungen in eine einzige übergehen. Ausserdem sollte es auch noch Feldquanta von ungeheuer hoher Masse geben, die einer Energie von zirka $10^{15}$ GeV entsprechen. Falls das wahr sein sollte, wären diese Teilchen schon im Weltall vorhanden gewesen, als die Temperatur solche Werte erreichte. Ausserdem kommt die Theorie zu einem weiteren überraschenden Schluss: das Proton ist nicht unveränderlich. Es zerfällt in Mesonen und Elektronen, aber im Durchschnitt erst nach $10^{32}$ Jahren; das ist

$10^{22}$ mal länger als das Alter des Weltalls. Das ist zwar eine lange Zeit, aber wenn man sehr viele Protonen hat, dann zerfallen doch ein paar innerhalb der Lebensdauer eines Menschen. Man versucht heute diesen Zerfall zu messen, und es wird sich bald herausstellen, ob es wahr ist oder nicht.

Nun kommen wir zum dritten ungelösten Problem: der Protonenüberschuss. Er beruht auf der Beobachtung, dass es heute im Universum zwar Materie, aber wenig oder gar keine Antimaterie gibt. Erinnern wir uns, dass es vor $10^{-6}$ Sekunden ein dichtes Gas von Quarks und Antiquarks gab, das sich dann bei Abkühlung durch gegenseitige Vernichtung von Teilchen und Antiteilchen in Strahlung verwandelte. Aber es blieben Quarks übrig, um die Protonen und Neutronen zu bilden. Es waren also ein bisschen mehr Quarks im Gas als Antiquarks. Nur ein ganz klein bisschen, denn, wie wir hörten, ist die Anzahl der Nukleonen nur ungefähr ein Milliardstel der Anzahl der Quarkpaare im heissen Gas. Woher kommt aber dieser Überschuss?

Man würde erwarten, dass der Uranfang völlig symmetrisch ist. Warum sollte am Anfang Materie vor Antimaterie bevorzugt worden sein und warum um nur so wenig? Es könnte ja sein, dass «jemand» damals ein bisschen mehr Materie als Antimaterie in die Suppe hineingeworfen hat. Dagegen nimmt man doch gerne an, dass am Anfang völlige Symmetrie geherrscht hat. Aber wie kann ein kleiner Überschuss zustande kommen, wenn am Anfang keiner da war? Eine der Bedingungen muss wohl sein, dass die sogenannte «Baryonenzahl» nicht konstant ist. Was ist diese Zahl? Im wesentlichen die Zahl der Quarks minus die Zahl der Antiquarks. Wenn das Universum in völliger Symmetrie begonnen hat, d.h. mit ebensoviel Quarks wie Antiquarks, so war die Zahl am Anfang gleich Null. Diese Zahl ändert sich natürlich nicht, wenn Quarks und Antiquarks sich gegenseitig vernichten. Wenn aber Quarks übrigbleiben, so ist die Zahl nicht mehr Null. Also muss sie sich verändert haben, falls am Anfang völlige Symmetrie geherrscht hat. Interessanterweise hat das etwas mit dem Zerfall der Protonen zu tun. So ein Zerfall ist ja auch eine Änderung der Baryonenzahl. Nehmen wir an, das Proton würde z. B. in ein Positron und in Photonen zerfallen. Die drei Quarks, aus denen es besteht, würden dann verschwinden und die Baryonenzahl hätte sich vermindert. Wenn es also wirklich einen Protonenzerfall gibt, so müsste man nicht mehr an der exakten Konstanz der Baryonenzahl festhalten. Aber, um einen kleinen Überschuss der Materie von einem symmetrischen Uranfang zu erhalten, müssen Prozesse existieren, die die Materie vor der Antimaterie etwas bevorzugt. Nur ganz wenig, denn es handelt sich ja nur um $10^{-9}$. Tatsächlich gibt es gewisse Anzeichen, dass die Naturgesetze in bezug auf Materie und Antimaterie nicht ganz symmetrisch sind. Es besteht also eine gewisse Hoffnung, den Überschuss erklären zu können. Aber man hat heute noch keine Methode, die erlaubt, das wirklich zu rechnen.

Nun kommen wir zum vierten Problem, der kritischen Materiendichte. Wie wir früher erwähnten, ist die heutige mittlere Dichte der Materie im Weltall nicht sehr gut bekannt. Sie liegt aber nahe an einem Wert, den man die «kritische Dichte» nennt. Wenn die tatsächliche Dichte über diesem Wert läge, so würde die Expansion des Universums ein Ende nehmen. Die Massen wären dann so dicht, dass die Schwerkraft die weitere Expansion zu einem gewissen zukünftigen Zeitpunkt zum Stillstand bringen und dann die Bewegung umkehren und sich das Universum wieder zusammenziehen würde. Die Dichte würde dann wieder anwachsen, bis nach einer gewissen Zeit der hochkonzentrierte Anfangszustand wieder erreicht wäre und der ganze Prozess von vorne beginnen und sich vielleicht endlos wiederholen könnte. Wenn aber die tatsächliche Dichte unterhalb der kritischen Dichte liegt, so gibt es kein Ende der Ausdehnung. Selbst im kritischen Fall geht die Ausdehnung asymptotisch auf Null, aber sie kehrt sich nicht um.

Es sieht fast so aus, als läge die Wirklichkeit nahe an dieser Grenze. Die Dichte scheint immer noch geringer als kritisch zu sein, aber es besteht ja neuerdings die Möglichkeit, dass die Neutrinos eine kleine Masse haben. Dann würde das Neutrinogas, das wahrscheinlich den Weltraum füllt, schon ganz beträchtlich zur Massendichte beitragen. Man muss sich nun fragen, warum die Natur gerade so eine Massendichte gewählt hat, dass die Ausdehnung sich nicht umdreht, sondern nach unendlicher Zeit zum Still-

stand kommt; eine Frage, die heute noch keine Antwort hat. Vielleicht liegt die Antwort wieder in der Form der noch unbekannten Naturgesetze, die am Anfang bei extrem hohen Temperaturen geherrscht haben.

Nun komme ich zur letzten, mehr philosophischen Frage: Was war vor dem Urknall? Falls die Massendichte höher lag als die kritische Dichte, ist das Problem etwas anderes. Man könnte sagen, dass sich das Universum periodisch ausdehnt und zusammenzieht. Vor dem letzten Urknall, aus dem unser Universum stammt, hatte sich also ein früheres zu einer hohen Konzentration zusammengeballt. Die Hypothesen des periodischen Universums begegnen allerdings einigen Problemen, die mit der Frage verknüpft sind, ob die Zusammenballung jedesmal dieselbe Ursituation herstellt oder ob die Entropie des Universums nicht doch jedesmal etwas grösser wird. Dann kommt man wieder zum Problem eines wirklichen Anfangs, des ersten Urknalls. Wenn die Massendichte unter oder bei der kritischen Dichte liegt, dann war der Urknall vor zirka $2 \times 10^{10}$ Jahren der wirkliche Anfang der Welt. Die Frage, was vorher war, hat keinen konkreten Inhalt. Alles, was wir wissenschaftlich beschreiben können, fand nachher statt. Dies möge manchem als eine unzulängliche Antwort erscheinen, aber es ist die einzige wissenschaftliche Antwort, die man auf diese Frage geben kann.

Am Anfang dieses Symposiums wurde bemerkt, wie stark unser Thema mit anderen menschlichen Geistesrichtungen zusammenhängt, mit philosophischen, mythologischen und religiösen Belangen. Es trifft uns sozusagen ins Herz. Daher kann man über dieses Thema nicht nur wissenschaftlich reden, sondern, in komplementärer Weise, auch in poetischer und emotioneller Sprache. Tatsächlich hat ja die judäo-christliche Tradition die Entstehung des Weltalls in einer Art beschrieben, die nicht ganz verschieden ist von dem, was hier erzählt wurde. So sagt doch die Bibel über den Anfang der Dinge: «Und Gott sprach, es werde Licht, und es ward *Licht*. Und Gott sah, dass das Licht gut war.»

*Anschrift des Verfassers:*

Prof. Dr. V. F. Weisskopf
Massachusetts Institute of Technology
Cambridge, MA 02139 (USA)

# Origine des éléments chimiques et naissance du système solaire

Hubert Reeves

Le thème de ce symposium sur «l'Origine des choses» laisse entendre que les choses – et par «choses» ici on comprend aussi bien les atomes que les galaxies, les bactéries que les êtres humains – ont une origine. Cette idée émerge de la juxtaposition des résultats obtenus par l'ensemble des sciences modernes: physique, chimie, biologie, ainsi qu'astronomie et cosmologie.

Darwin, le premier, a introduit l'historique dans le domaine scientifique. Aujourd'hui ses disciples nous dessinent un schéma d'évolution biologique qui nous mène, en quatre milliards d'années des algues bleues et des bactéries à l'ensemble extrêmement varié de tous les organismes vivants.

Les biochimistes font évoluer ces cellules primitives – ou, plus exactement, leur code génétique – à partir d'un ensemble de molécules simples de l'océan et de l'atmosphère des débuts de notre planète. Ce chapitre s'appelle l'évolution prébiotique. On peut aussi le considérer comme la seconde phase, la phase planétaire, de l'évolution chimique.

Les astrophysiciens voient dans ces molécules simples le résultat d'une longue évolution nucléaire au long de laquelle les nucléons, issus du Big Bang, se sont combinés dans la chaleur des intérieurs stellaires pour engendrer l'ensemble de tous les noyaux atomiques jusqu'à l'uranium. Libérés dans l'espace par les explosions stellaires, ces noyaux s'habillent d'électrons, deviennent des atomes et commencent à se joindre les uns aux autres. Ces jonctions ont pour résultat, d'une part, la formation des grains de poussières d'où naîtront plus tard les planètes, et, d'autre part, les molécules simples qui constitueront les atmosphères et les océans de nos planètes. Ces derniers événements forment la première phase, la phase interstellaire, de l'évolution chimique, qui se poursuit ensuite à la surface des planètes.

Dans les schémas présents de la physique, les nucléons sont eux-mêmes le résultat d'une phase plus primitive encore, qui les fait naître de la combinaison des quarks aux premières microsecondes de l'univers. On pourrait parler ici d'une évolution leptoquarkienne.

## Evolution nucléaire

Ma tâche ici est d'illustrer pour vous la phase nucléaire – responsable de l'origine des éléments chimiques – et la phase chimique interstellaire responsable de la formation des systèmes planétaires et de leurs éventuelles biosphères.

Je décrirai cette histoire comme une pièce de théâtre, avec une scène et des acteurs. Je vais tenter de démêler les intrigues enchevêtrées des quatre grandes forces de la physique sur différentes scènes cosmiques.

Prenons comme premier exemple celui de la formation de noyaux complexes par la combinaison de noyaux plus simples (figure 1). Deux forces interviennent au premier plan. La force nucléaire qui voudrait les joindre et la force électromagnétique qui, parce que tous les noyaux sont chargés positivement, tend à les tenir éloignés. La force nucléaire est très puissante, mais sa portée est courte. La force électromagnétique est bien moins puissante mais, en revanche, sa portée est beaucoup plus grande. Résultat: quand les noyaux sont loin – à plus de $10^{-13}$ cm – ils se repoussent, mais quand ils sont près, ils s'attirent violemment et se joignent pour former des noyaux plus lourds.

Comment amener les noyaux à vaincre la répulsion électrostatique qui les sépare et à se rapprocher pour entrer en combinaison? Il faut leur donner de grandes vitesses. On y arrive soit en les accélérant avec des accélérateurs ou en les réchauffant (énergie thermique).

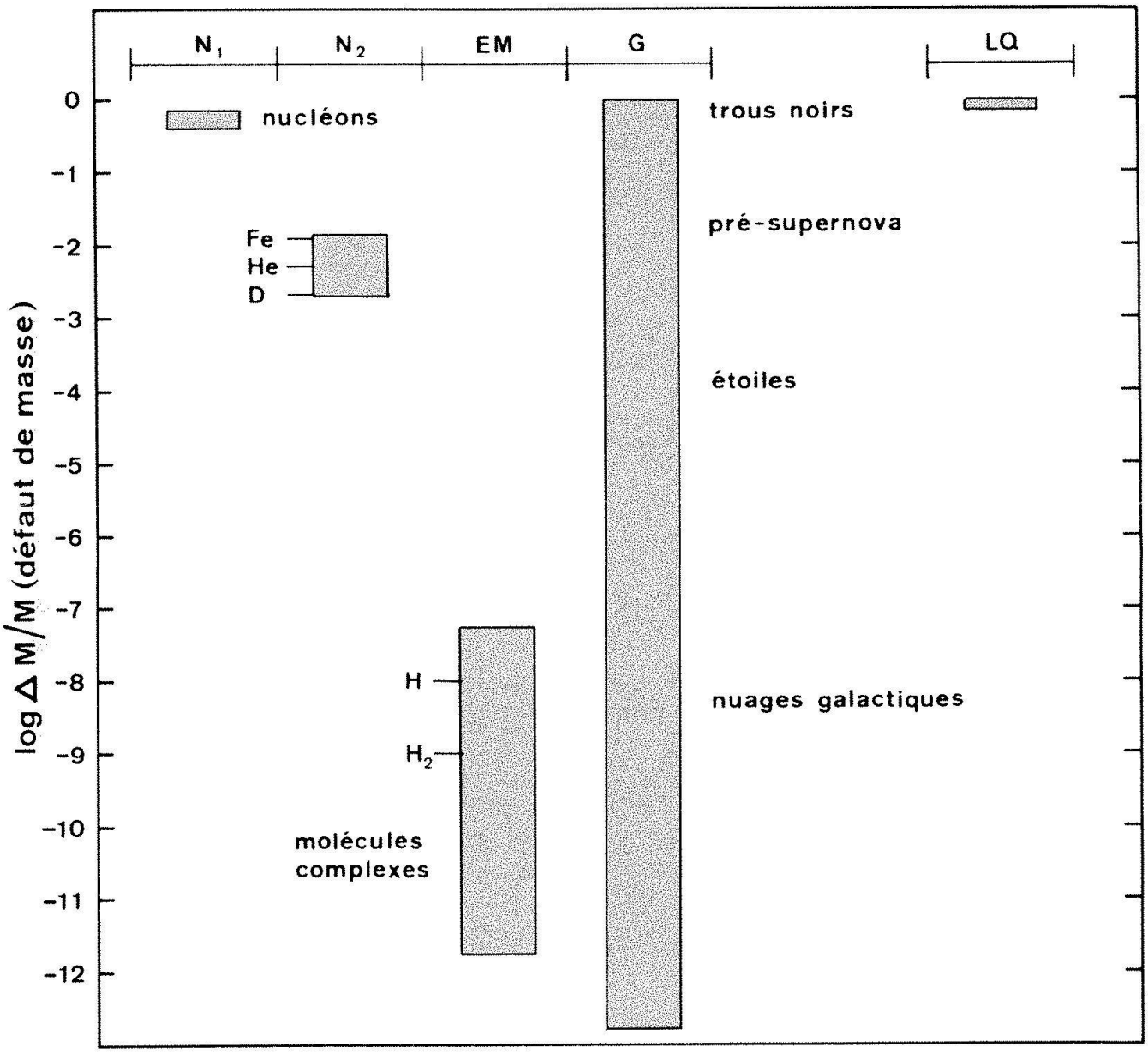
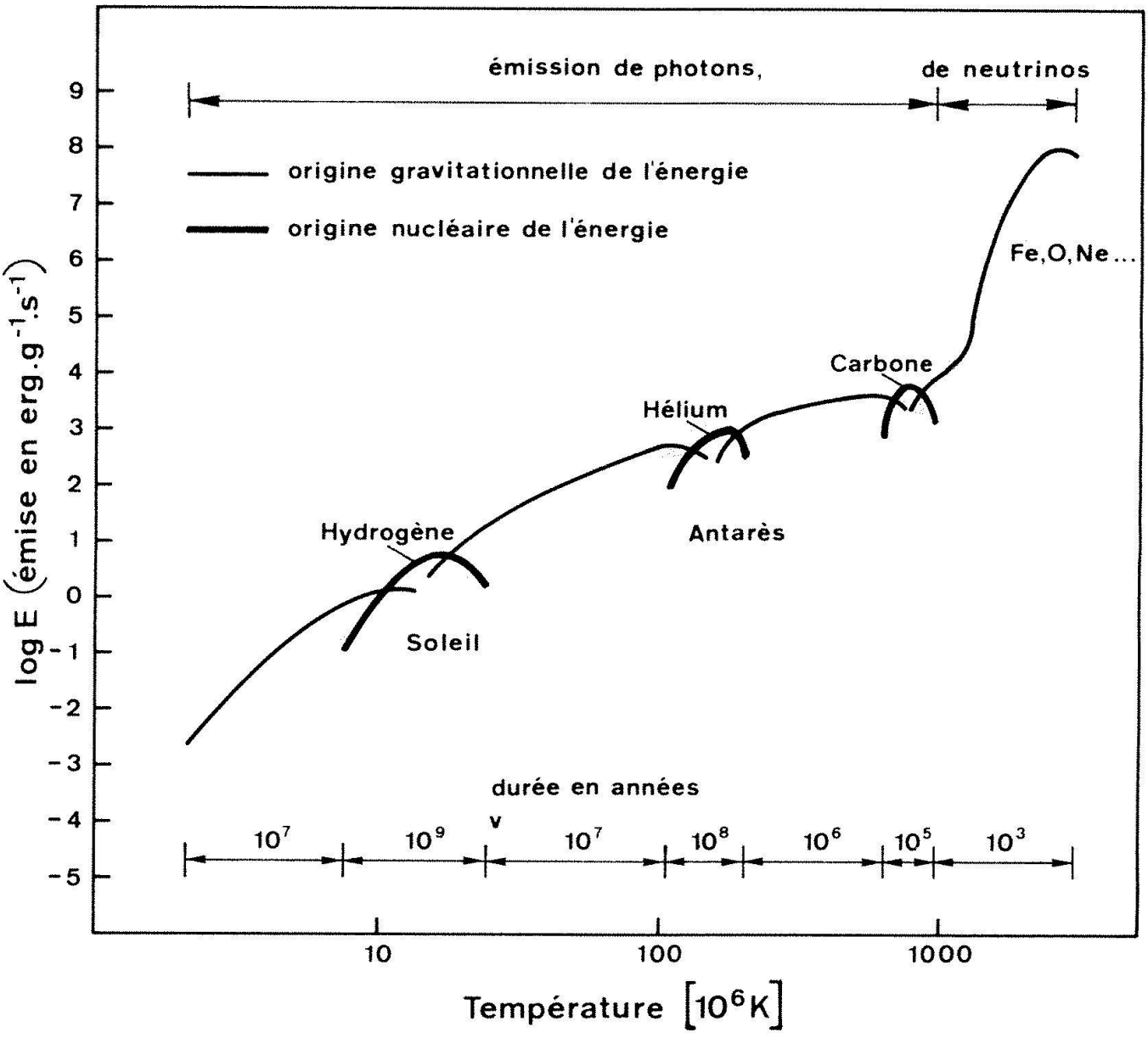



*Fig. 1.* Efficacité relative des différentes forces naturelles a engendré des états liés. L'ordonnée donne la fraction de masse qui est rejetée au loin au moment de la combinaison, appelée aussi le «défaut de masse» ou «énergie de stabilité». La force faible ne crée pas d'états liés. La force électromagnétique EM engendre des liaisons voisines de un électron volt. C'est-à-dire une fraction $10^{-9}$ de la masse totale du système ($10^{-8}$ pour l'atome d'hydrogène). La force nucléaire $N_2$ (au sens classique) engendre des systèmes (les noyaux) qui se situent entre $10^{-3}$ (le deuterium) et $10^{-2}$ (le fer). La force entre les quarks $N_1$ (la vraie force nucléaire) forme des systèmes encore plus liés. Le terme de «défaut de masse» ne s'applique plus très bien ici mais qualitativement on peut mettre la barre un peu en dessous de l'unité.
Contrairement aux autres forces, la gravité G n'est pas assignée à un domaine restreint dans l'échelle des défauts de masse fractionnels. Parce qu'elle est proportionnelle à la masse elle peut prendre toutes les valeurs jusqu'à l'unité (trou noir). J'ai indiqué ici les structures correspondantes.
La barre nommée «leptoquark» LQ indique ici le fait que, dans le cadre des théories de grandes unifications, les nucléons peuvent se désintégrer en lumière. Il ne s'agit naturellement pas ici d'un système lié. Le terme «défaut de masse» s'entend dans un sens élargi ...

*Fig. 2.* Histoire thermique d'une étoile. En abscisse, la température au centre, en ordonnée, le flux d'énergie émis. La source d'énergie est, alternativement, gravitationnelle et nucléaire. Les périodes correspondantes sont données au bas de l'échelle. A la fin de sa vie l'étoile émet surtout des neutrinos.




La gravité est l'une des principales sources des hautes températures qui existent dans l'univers. Ici, il faut changer de cadre, passer du microscopique au macroscopique. Comme la force électromagnétique la force de gravité est à longue portée; mais elle est incomparablement plus faible. Cette faiblesse est compensée par le fait qu'elle est toujours attractive (alors que la force électromagnétique est à deux composantes). Sur de grandes quantités de matière la force de gravité finit toujours par dominer sur toutes les autres (la force faible est à courte portée).
Considérons maintenant un nuage interstellaire; une masse de matière très diffuse qui contient environ $10^{60}$ atomes. Ici, la gravité domine. Si le nuage est froid, il s'effondre sous son propre poids. Cet effondrement crée de la chaleur. En d'autres termes, il y a libération d'énergie d'origine gravitationnelle (la chute de matière) et transformation de cette énergie en énergie thermique. La pression thermique augmente jusqu'au moment où elle équilibre la force de gravité. A ce moment, la contraction s'arrête et nous avons ... une étoile.

### Gravité, chaleur et nucléosynthèse

Ici la force électromagnétique va jouer un nouveau rôle. Elle va transformer l'énergie de l'agitation thermique en photons. L'étoile, chaude, brille. C'est-à-dire qu'elle envoie de l'énergie lumineuse dans l'espace. Pour compenser cette perte d'énergie (qui l'amènerait

à se refroidir) et pour rééquilibrer la force de gravité, l'étoile va continuer lentement à se contracter, à un rythme suffisant pour équilibrer la perte. Cette condensation va libérer encore de l'énergie gravitationnelle qui va se transformer, partiellement, en lumière émise, mais aussi, partiellement, en chaleur interne. Ainsi, loin d'amener l'étoile à se refroidir, l'émission de lumière va au contraire la porter à une température de plus en plus élevée et à un éclat de plus en plus grand.
La force électromagnétique joue ici deux rôles opposés vis-à-vis de la construction des noyaux. D'une part, par le biais de la répulsion électrostatique (c'est-à-dire par l'échange de photons dits «virtuels») elle tient les particules chargées à distance; mais, d'autre part, par le moyen des photons «réels», elle amène ces particules chargées à émettre vers l'espace, elle provoque la contraction progressive de l'étoile et donc son échauffement et, en conséquence, l'accroissement continu des vitesses entre les particules chargées. De ce fait, ces particules arrivent de plus en plus souvent à vaincre la répulsion électrostatique et à entrer en contact nucléaire, c'est-à-dire à se joindre pour former des noyaux nouveaux.
Le premier effet, répulsion électrostatique, ne change pas avec les conditions physiques; mais le second, vitesse des particules et fusion nucléaire, augmente avec la température. On atteindra donc nécessairement un point où il dominera sur le premier.
Prenons, par exemple, le cas du soleil (figure 2). Initialement, il est composé à 90% de protons (et d'électrons). Vers dix millions de degrés, les protons arrivent à se toucher. Mais ici il y a une complication supplémentaire: deux protons ne forment pas un système stable. Par contre, si au moment de la rencontre, un des protons se transforme en un neutron, ça marche: le proton et le neutron s'unissent pour former un deutéron (hydrogène lourd). Or cette transformation fait intervenir notre quatrième force, la force faible (cette même force provoque l'émission d'un neutrino qui s'échappe aussitôt du soleil).

**Fusion de l'hydrogène**

Le deutéron se combinera lui-même avec un proton. Puis, après deux ou trois autres réactions, nous voyons apparaître comme produit final de ces combinaisons, des noyaux d'hélium-4. Bilan net: les actions conjuguées des forces nucléaire et faible ont transformé quatre protons en un hélium. Or, la différence de masse entre ces systèmes est d'environ un pour cent de la masse totale. La fusion dégage une énergie de sept millions d'électrons volts par proton initial (dont la masse est d'un milliard d'électrons volts). Ce sont la force électromagnétique et la force faible qui se chargent de l'évacuation de ce surplus de masse. Des photons gammas et des neutrinos sont successivement émis tout au long de ces réactions.
Ces événements microscopiques vont, à leur tour, avoir une influence macroscopique. Les neutrinos s'échappent de l'étoile mais les photons gammas, non. Ils sont absorbés par la matière stellaire et leur énergie est transformée en chaleur. L'étoile, ainsi, s'est trouvée une nouvelle source d'énergie. Cette énergie va servir à compenser les pertes que provoque l'émission de sa lumière. Elle n'aura plus besoin de se contracter, c'est-à-dire de faire appel à ses ressources d'énergie gravitationnelle. Elle va se stabiliser. Désormais, son rayon, sa densité et sa température vont rester constants.
Grâce à la très grande puissance de l'énergie nucléaire, ce temps sera très long. Notre soleil vit de ces transformations nucléaires depuis près de cinq milliards d'années. Cette stabilité prolongée est vraisemblablement requise par le développement de la vie, en tout cas, si on en juge par l'exemple de notre biosphère.
Les étoiles qui vivent de la transmutation de l'hydrogène (protons) en hélium forment la grande famille dite de la «Série Principale». Quatre vingt dix pour cent des étoiles de notre galaxie appartiennent à cette famille. Dans notre ciel nocturne, notons en particulier Sirius, Véga, Arcturus et l'Etoile Polaire.
Pour le Soleil, cette phase va durer encore cinq milliards d'années. A ce moment-là, il aura épuisé l'hydrogène en son centre. La perte d'énergie stellaire par émission de lumière ne sera plus compensée par la libération d'énergie nucléaire. L'étoile aura de nouveau recours à ses «réserves» d'énergie gravitationnelle. Elle va recommencer à se contracter. Finie la stabilité. Sa densité et sa température vont croître à nouveau.

Paradoxalement, cette contraction va accroître le rayon stellaire. En fait, le noyau central de l'étoile - qui comprend presque toute la masse - va effectivement décroître en volume, mais les couche extérieures - une atmosphère raréfiée - vont prendre des proportions gigantesques et se refroidir, c'est-à-dire passer au rouge. L'étoile deviendra une «Géante Rouge» comme, dans notre ciel, Antarès, Bételgeuse ou Aldébaran.
Notons ici une analogie entre ces deux modes de génération d'énergie. Le gravitationnel implique la contraction d'une grande quantité de matière. Beaucoup de nucléons passent d'un volume grand à un volume plus petit. Cet événement est accompagné par l'émission d'énergie à la fois à l'intérieur du volume: les particules vont plus vite (la température s'accroît) et aussi à l'extérieur (l'étoile brille de plus en plus).
De la même façon, à l'échelle microscopique, le nucléaire implique que quatre nucléons, auparavant libres, se trouvent maintenant confinés dans le volume d'un noyau. Leur distance relative moyenne a considérablement décru. En même temps, leur énergie interne s'est accrue (mouvements orbitaux des nucléons dans le noyau) et ils ont émis vers l'extérieur des photons gammas.
Ici intervient un événement important en relation avec la quantification des charges électriques. Elles apparaissent toujours par nombre entier: l'hydrogène a une charge positive, l'hélium en a deux, etc. La répulsion est elle-même quantifiée. En conséquence nous aurons une série de phases du fusion nucléaire bien distinctes. Je m'explique.

**Fusion de l'hélium**

Nous avons vu l'étoile, sous l'effet de la force gravitationnelle, augmenter sa température jusqu'au moment où elle pouvait vaincre la répulsion entre les protons (une charge positive) et fusionner l'hydrogène en hélium à température constante. Après l'épuisement de l'hydrogène, le centre de l'étoile est composé d'hélium (deux charges électriques positives). La répulsion électrostatique entre ces particules est plus grande que celle des protons. En conséquence, il faut atteindre des températures plus élevées pour amener les héliums à se joindre et engendrer des noyaux nouveaux. Au lieu de dix ou vingt millions de degrés il faudra cent ou deux cents millions de degrés. Ce sont là les températures caractéristiques des géantes rouges. En leur centre elles brûlent de l'hélium et engendrent du carbone et de l'oxygène. Sur des couches plus éloignées du centre, elles transforment du carbone et de l'oxygène, préexistant, en azote. Carbone, azote, oxygène, on ne peut trop insister sur l'importance des géantes rouges vis-à-vis de l'évolution prébiotique et biologique ...
Le même scénario se produit après l'épuisement de l'hélium. L'étoile, à nouveau, reprend sa contraction et son ascension thermique. Le carbone possède six charges électriques. Pour en amorcer la fusion, il faut atteindre pratiquement le millard de degrés. Peu après démarre la fusion de l'hydrogène (huit charges électriques). Ces deux fusions sont responsables de la production du sodium, magnésium, aluminium, silicium et soufre. Voilà maintenant les principaux constituants de notre planète (avec l'oxygène et le fer) préparés par des réactions nucléaires à l'intérieur d'étoiles évoluées.
D'autres phases de fusion vont ensuite se succéder rapidement et engendrer, dans la foulée, tous les autres éléments de la table de Mendeleieff jusqu'au plomb et à l'uranium. L'abondance des éléments dans l'univers est montrées à la figure 3.
Vers cette période s'amorce une série de processus physiques qui vont amener l'étoile vers sa fin. Ils ont en commun de soustraire à l'étoile une partie de sa chaleur et donc de l'obliger à accélérer sa contraction pour compenser ces pertes. C'est cette accélération qui va l'amener à la catastrophe.

**Ejection des noyaux dans l'espace**

Il y a d'abord un accroissement prodigieux du rôle de la force faible. Des quantités de réactions se produisent qui ont pour résultat l'émission d'un flux intense de neutrinos. Ces particules deviennent maintenant le principal agent de l'émission de l'énergie stellaire. Elles transportent beaucoup plus d'énergie que les photons. Or, contrairement à ces derniers, les neutrinos ne sont pas absorbés par la matière stellaire. Ils sortent rapidement et, de ce fait, exigent une réaction

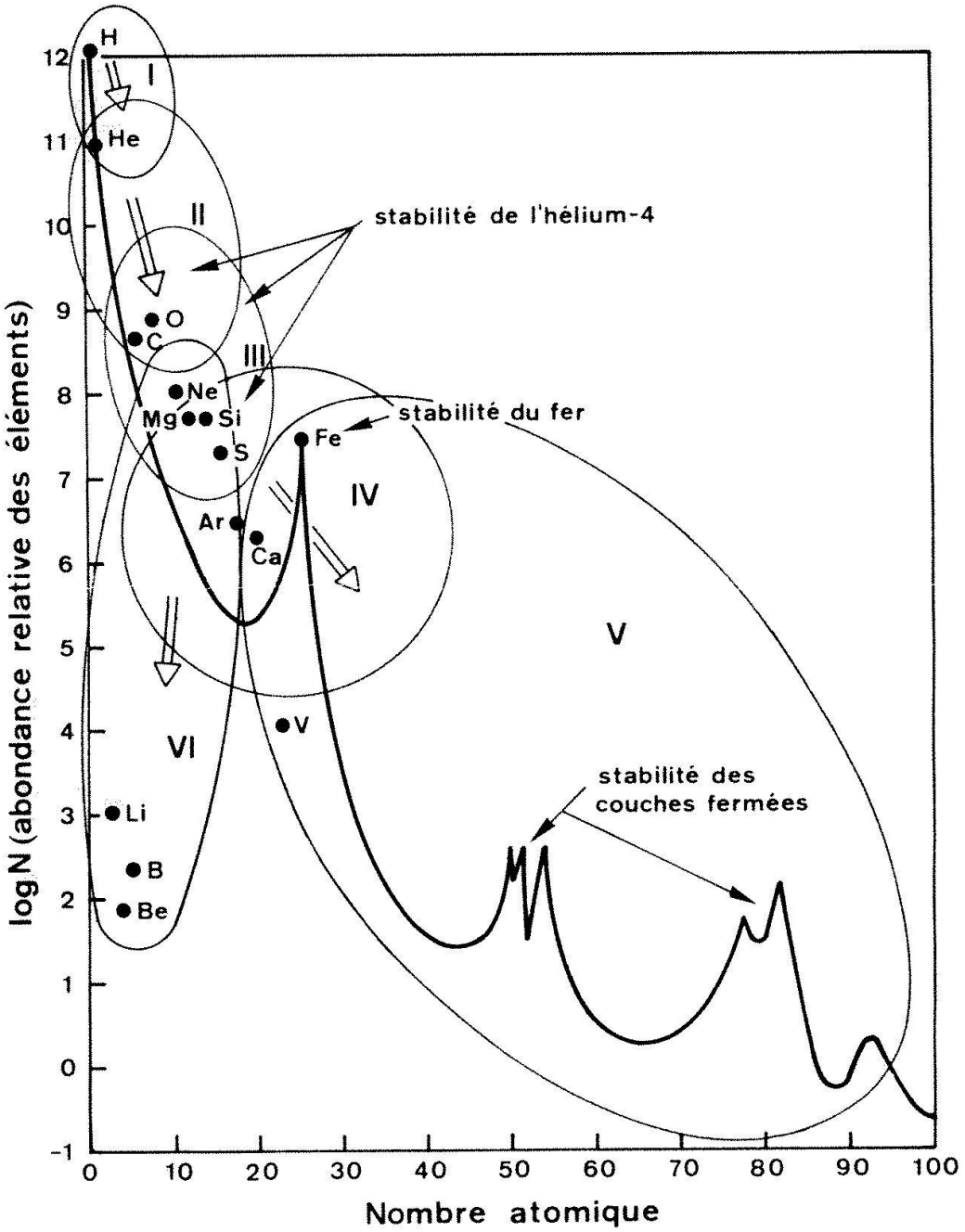



*Fig. 3.* Abondance des éléments dans l'univers. L'ensemble des points forme une courbe relativement régulière, avec quelques exceptions. Les chiffres romains indiquent les phases stellaires associées à l'évolution des éléments responsable des abondances observées:

I Explosion initiale et étoiles de la «série principale»
II Géantes rouges
III Phase stellaire plus tardive
IV Supernovae
V Supernovae
VI Evolution des éléments dans l'espace interstellaire

immédiate de l'étoile, sous forme de contraction accélérée. Il y a ensuite le fait que au-delà d'une certaine température (quelques milliards de degrés) des noyaux commencent à se photo-désintégrer, c'est-à-dire à se casser sous l'effet des photons du rayonnement thermique. Ces désintégrations absorbent de l'énergie qu'elles enlèvent à l'étoile.

A l'appel pressant de ces effets variés, l'étoile se contracte de plus en plus vite. Bientôt, il s'agit d'une avalanche, d'une implosion. Le cœur de l'étoile atteint des densités énormes, des dizaines de milliers de tonnes par centimètre cube.

Par un mécanisme encore mal compris, la partie la plus centrale de l'étoile continue à se contracter jusqu'à devenir une étoile à neutrons. La gravité, ici, est équilibrée par la force nucléaire devenue répulsive tandis que la partie extérieure est rejetée vers l'espace par des phénomènes qui font intervenir à la fois la force électromagnétique et la force faible. C'est une supernova. Les noyaux engendrés par l'étoile se dispersent au loin sous forme d'un rémanent de supernova qui s'étend sur des distances de dizaines d'années-lumière.

## Evolution chimique: phase spatiale

Maintenant cette matière chaude se refroidit en se déployant dans l'espace. A mesure que les vitesses thermiques décroissent, les forces électromagnétiques deviennent en mesure de

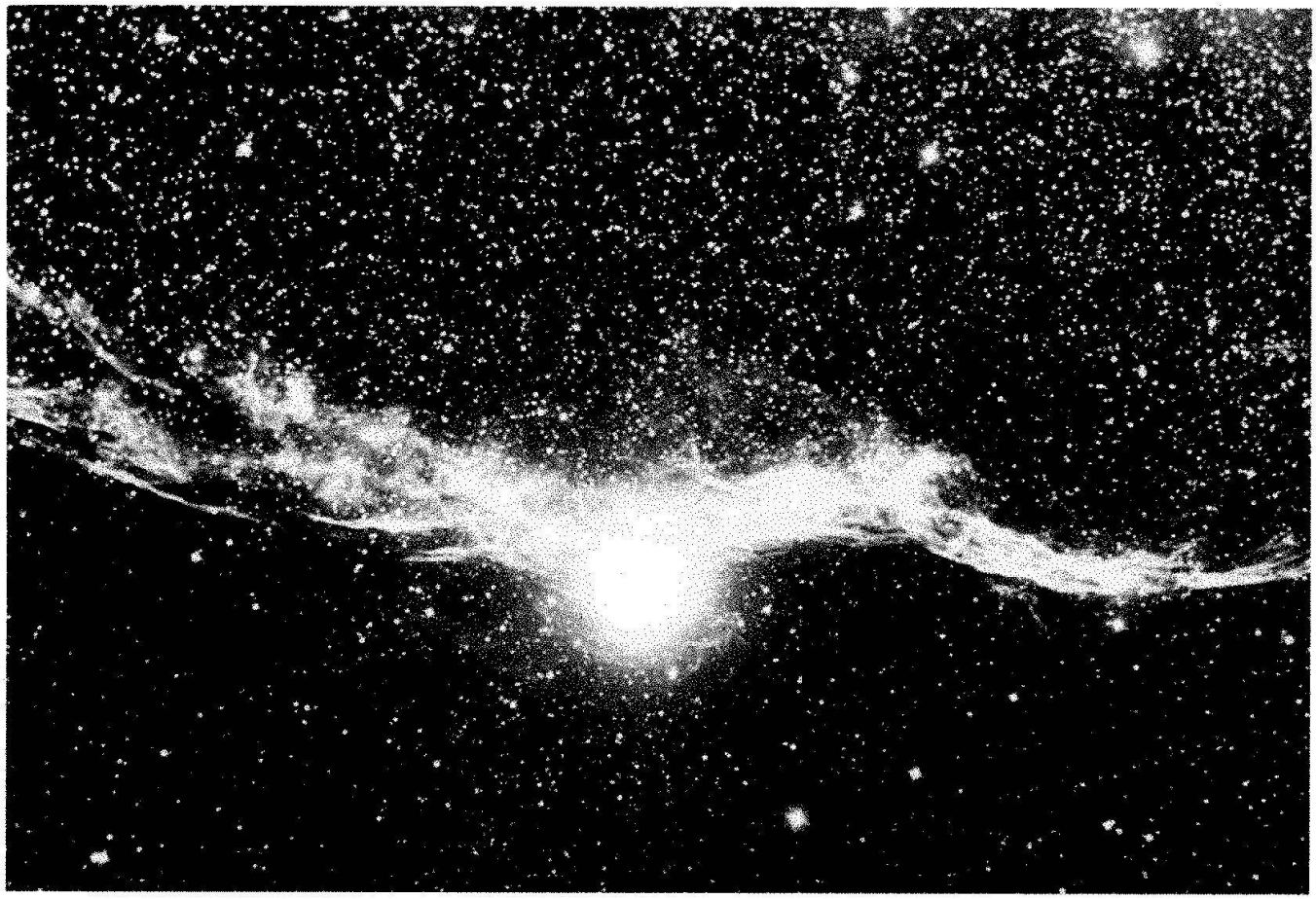


*Fig. 4.* La Dentelle du Cygne. Détail des filaments issus d'une explosion stellaire. Ici, les noyaux s'habillent d'électrons et forment des atomes et des molécules. Des poussières s'y constituent qui donneront plus tard naissance aux planètes. C'est un des hauts lieux de l'évolution chimique.

fixer les électrons en orbite autour des noyaux. D'abord se remplissent les orbites les plus internes – les orbites K – puis, successivement, tous les autres. Quand les atomes atteignent la neutralité, ils commencent à se combiner en molécules. L'hydrogène se joint au carbone, à l'azote et à l'oxygène pour donner le méthane, l'ammoniac et l'eau (appelés souvent les «glaces»). L'oxygène se combine au fer, au silicium au magnésium etc. pour donner les premiers solides: des grains de poussières interstellaires qui incorporent des trillions d'atomes. Dans les grands froids de l'espace, les glaces se déposent sur ces grains comme des océans figés sur des planètes en miniature.

Ici, c'est le début de l'architecture électromagnétique. Après la gravité qui a construit les étoiles, après la force nucléaire qui a construit les noyaux, l'organisation de la matière est maintenant prise en charge par la force électromagnétique. C'est elle qui sera responsable de tous les solides, ainsi que de tous les organismes vivants.

Nous sommes maintenant amenés au problème de l'origine des planètes et des cortèges planétaires tels que notre système solaire. Tout se joue entre deux forces réelles: la gravité et l'électromagnétisme, auxquelles s'ajoute une troisième «fausse» force: la force «centrifuge» (qui est en fait reliée à l'inertie des corps). Les nuages interstellaires contiennent de larges quantités de poussières engendrées dans les rémanents de supernovae des étoiles révolues. Dans certaines régions du ciel, on voit très bien les lueurs bleues que provoque la réflexion de la lumière stellaire sur ces poussières. Quand l'effondrement d'un nuage amène la formation d'une étoile, ces poussières sont entraînées dans l'avalanche. Certaines d'entre elles sont vaporisées par la chaleur stellaire, mais d'autres sont épargnées grâce, en particulier, au fait que les nuages tournent sur eux-mêmes. A cause de cette rotation il se forme alors dans le plan équatorial de l'étoile une sorte de disque, semblable aux anneaux de Saturne, où se retrouvent de vastes quantités

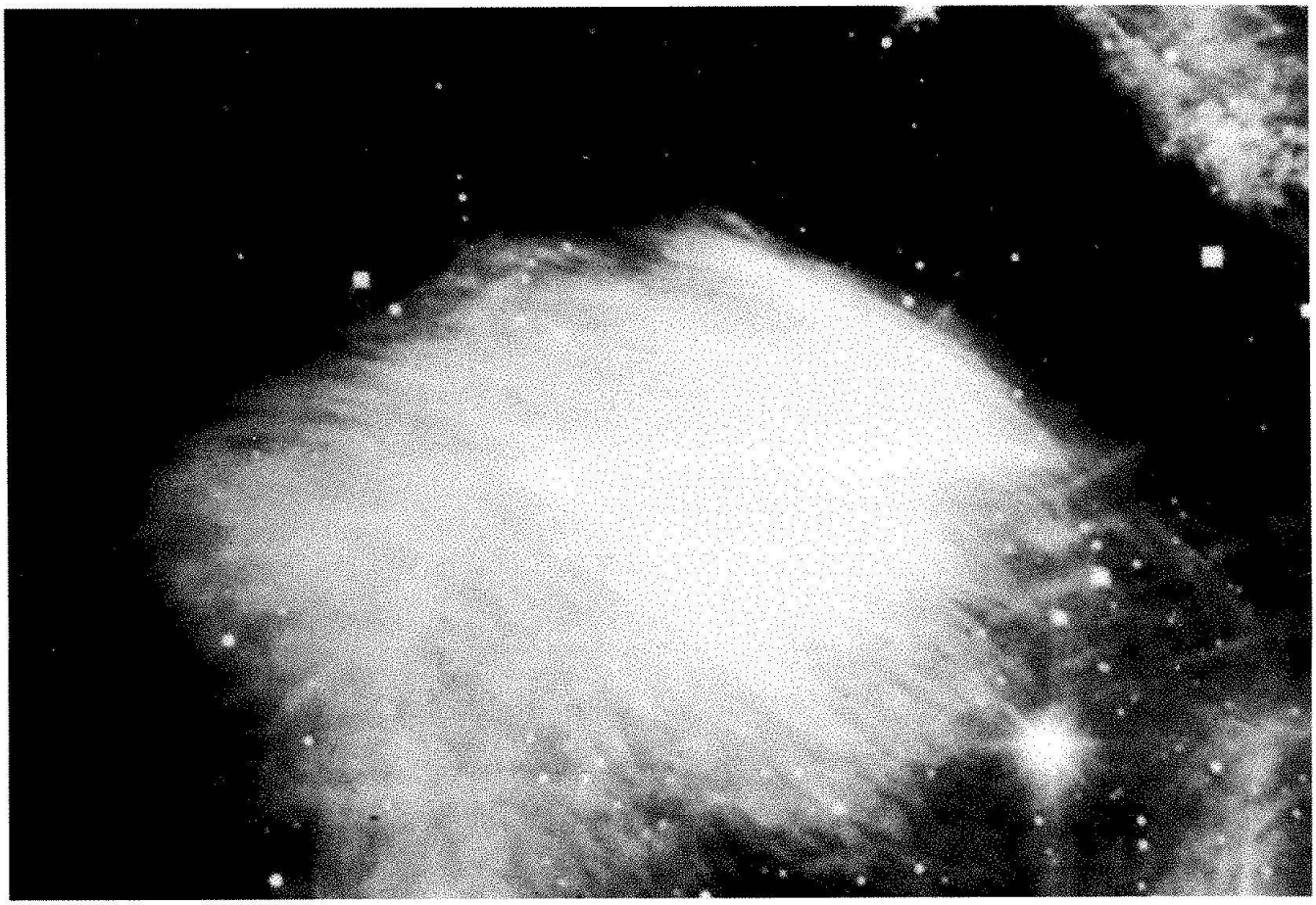


*Fig. 5.* Cette photo montre l'étoile Mérope dans les Pléiades nimbée de luminosités filamenteuses. Les étoiles des Pléiades sont nées toutes ensemble à partir d'un nuage interstellaire il y a environ cent millions d'années. Les filaments sont composés de myriades de «poussières interstellaires». Ils rappellent les cirrus de notre atmosphère, formés de minicristaux de glaces. Ces poussières, éclairées ici par la lumière stellaire, s'assemblent pour former des planètes autour des embryons d'étoiles.

de poussières interstellaires, ainsi maintenues à distance prudente du foyer central.
Là commence un processus d'agglomération, dont les mécanismes sont mal connus, par lequel ces grains se juxtaposent et se transforment en corps de plus en plus massifs: astéroïdes et planètes. Ici, la gravité sert de «filet». Elle permet aux proto-planètes de capturer les bolides interstellaires qui s'approchent et ainsi d'accroître leur masse.
Les planètes de type terrestre sont des structures physiques où la gravité est équilibrée par la force électromagnétique (rigidité des structures cristallines). Cet assemblage permet la construction de grandes plateformes solides où l'eau liquide peut s'accumuler: c'est le lieu de la soupe océanique primitive. Les molécules d'eau de cet océan, comme d'ailleurs les molécules de méthane et d'ammoniac, nous sont parvenues par le biais des poussières interstellaires dont elles formaient la couche extérieure.

Les propriétés des corps du système solaire sont dominées par deux facteurs importants: leur masse et leur distance au soleil. Et il y a tout lieu de penser que la masse des planètes est elle-même largement influencée par la distance au soleil.

## La chaleur des planètes

La distance au Soleil détermine la quantité de chaleur qui parvient (rayonnement ou conduction) au lieu de formation d'une planète, et donc les conditions de température dans lesquelles elle se formera. Ces conditions de température, à leur tour, détermineront quels composés physicochimiques pourront s'y condenser. La grande densité de Mercure nous montre que seuls les composés les plus réfractaires s'y trouvent. Par opposition, l'eau qui apparaît à la distance de la Terre, devient très importante au niveau des

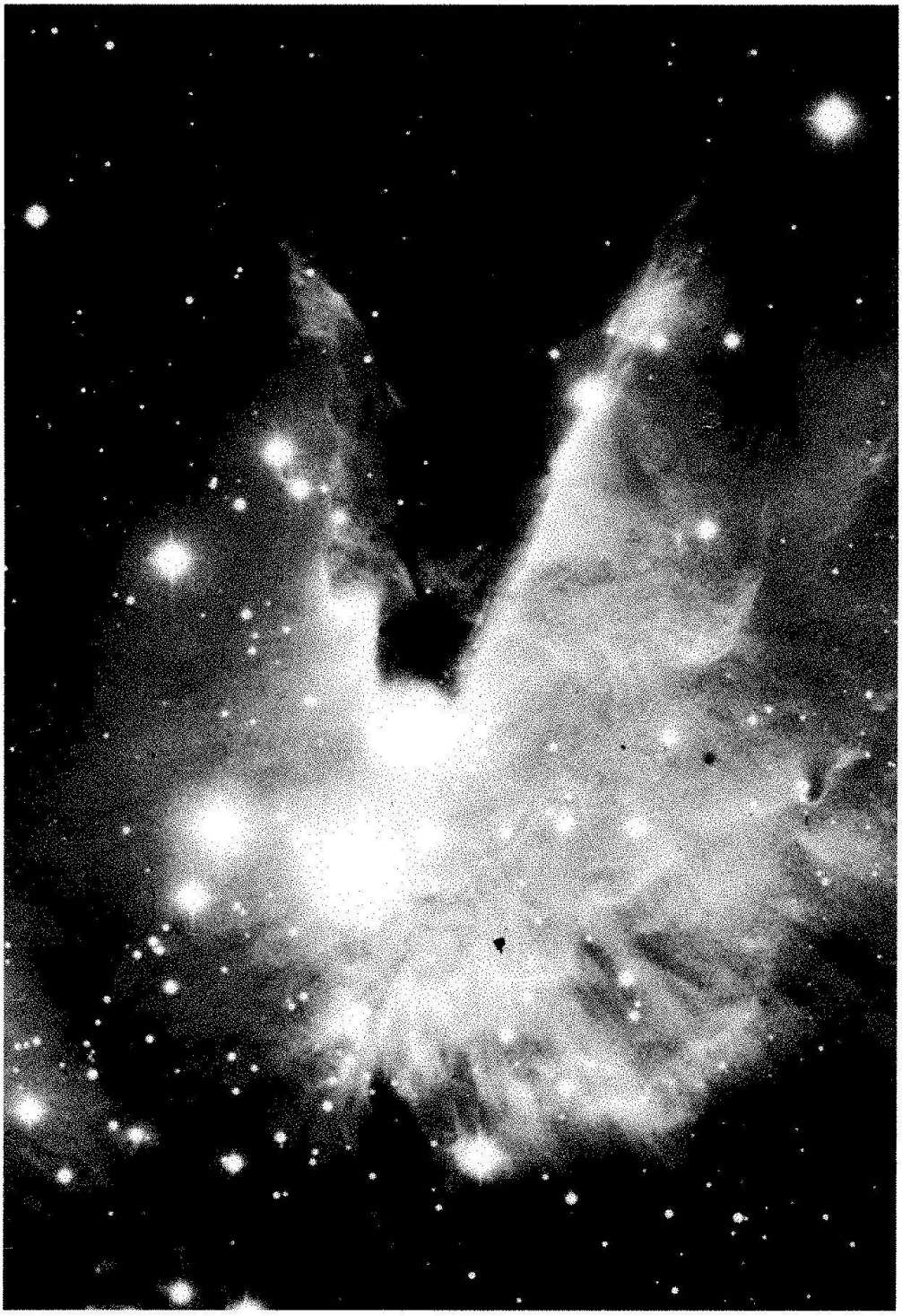


*Fig. 6.* La nébuleuse de la Quille, dans la Licorne. Au-dessus de l'énorme masse opaque, panachée de lumière, on identifie une vingtaine d'étoiles nées il y a moins d'un million d'années, c'est-à-dire après la naissance des premiers hommes sur la Terre.

satellites de Jupiter et de Saturne. De surcroît, les planètes géantes possèdent suffisamment de masse pour retenir dans leur atmosphère la composante normale d'hydrogène et d'hélium de la nébuleuse primitive.
L'accumulation de la masse d'une planète produit de la chaleur (énergie gravitationnelle transformée en énergie cinétique). Plus la masse est grande, plus la température initiale est élevée. L'évacuation de cette chaleur, dans les ères qui suivent, est le moteur de l'évolution de la planète et l'élément dominant de sa constitution physique. La Lune, Mercure, de petites masses, ont évacué leurs chaleurs initiales en quelques centaines de millions d'années. Par la suite, pétrifiées pour l'éternite, elles se sont contentées d'enregistrer l'arrivée des météorites à leur surface. C'est ce qui explique leur relief criblé de cratères météoritiques.
Mars est un cas intermédiaire: quelques volcans, quelques canyons révèlent une activité

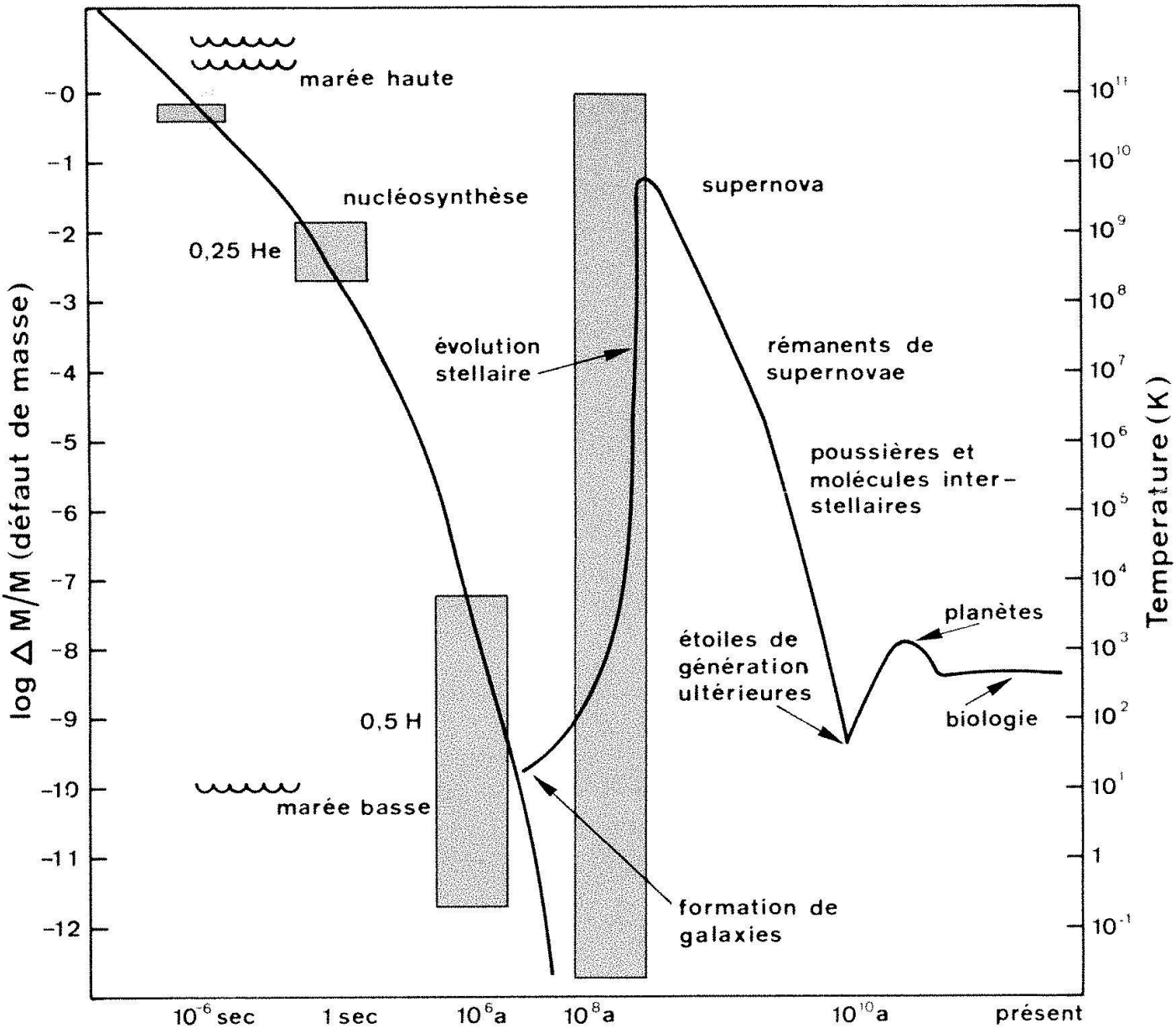



*Fig. 7.* Evolution de l'univers: Semblable à la Fig. 1. En ordonnée à droite est ajoutée une échelle de température, et, en bas, une échelle de temps. La courbe continue décrit d'abord la chute de la température universelle (Big Bang) qui permet la formation des nucléons, des noyaux (H et He) (nucléosynthèse primordiale) et des premiers atomes d'hydrogène et d'hélium. On peut considérer cette première phase comme un échec sur le plan de la complexité croissante: aucun noyau lourd n'a été formée. Grâce à la force de gravité la croissance de la complexité reprend dans les galaxies et les étoiles. La gravité sert d'ascenceur. Par le théorème du viriel elle permet aux étoiles de remonter l'échelle des énergies (thermiques) et de retrouver successivement les domaines électromagnétiques et nucléaires - avec nouvelle nucléosynthèse et formation des éléments lourds. Puis, après l'explosion de l'étoile, la température redescend (dans les remanents de supernovae) et la force électromagnétique engendre les poussières et les molécules interstellaires.
Autour d'étoiles de générations ultérieures, ces poussières vont s'agglutiner pour engendrer des planètes. A nouveau la température va remonter. Dans les océans primitifs reprendra avec vigueur la poursuite de la complexité cosmique.

interne passée, vraisemblablement largement éteinte aujourd'hui. La Terre est le prototype de la planète vivante. Elle est loin d'avoir évacué toute sa chaleur «initiale». Volcans en éruption, séismes, mouvements orogéniques et dérive des continents en font foi. Il est intéressant de noter, dans ce cadre, que les jeux de la sélection naturelle dépendent en partie des variations des conditions physiques dans lesquelles les races animales évoluent. Les changements de climat, amenés par exemple par le mouvement des plaques terrestres, sont de puissants moteurs de l'évolution ...

On a mis récemment en évidence la présence dans certaines météorites de traces «fossilisées» de matières radioactives, de courte durée (moins d'un million d'années). Cette découverte nous rappelle que les étoiles naissent en groupe et qu'à sa naissance notre Soleil était entouré de jeunes étoiles plus ou moins massives. Les plus massives (géantes bleues) vivent toute leur existence en quelques millions d'années. A leur mort, elles ont dispersé sur leur consœurs encore en formation une partie de leur cuisson interne. De là proviennent ces fossiles.
Après avoir revu rapidement les phénomènes physiques qui président à la formation des éléments chimiques et des cortèges planétaires (voir aussi figure 7) je voudrais terminer sur deux points d'une portée plus générale vis-à-vis du problème de «l'origine des choses». Le premier s'adresse à l'évolution des constantes de la physique, le second au rôle de l'expansion universelle.

## Evolution des «constantes» de la physique

On a vu dans les pages qui précèdent le rôle des quatre forces de la nature dans l'élaboration des choses. Chacune de ces forces est caractérisée par une *constante de couplage* et par une portée, qui décrivent l'intensité de son interaction dans l'univers. Pour déterminer une échelle relative, on va donner la valeur unité à la constante de la plus puissante force: la nucléaire. Sa portée est d'environ $10^{-13}$ cm. L'électromagnétique vient en second avec une valeur environ cent fois plus faible (1/137), elle a une portée infinie ($\alpha\ r^{-2}$). Puis la force faible, avec $10^{-5}$, sa portée est de $10^{-15}$ cm. Et finalement, la gravité, avec la valeur extrêmement basse de $10^{-40}$, mais aussi de portée infinie. On peut montrer assez facilement (cela a été fait par Carr and Rees, en particulier) que ces valeurs relatives suffisent, à elles seules, à fixer la structure des atomes, des organismes vivants, des planètes, des étoiles, des galaxies et même de l'univers tout entier. La réponse à la question: Pourquoi les constantes de couplage ont-elles ces valeurs? nous amènerait donc très loin dans la compréhension de l'origine des choses.
La physique moderne nous offre déjà les éléments d'une réponse partielle. Cela s'appelle le programme de la «grande unification». Selon ce programme, toutes les «constantes» se seraient différenciées à partir d'une seule constante universelle caractérisant une force de portée infinie qui aurait régné sans rivale au tout début de l'univers.
D'importants progrès ont été réalisés vis-à-vis de l'unification de trois des quatre forces: la nucléaire, l'électromagnétique, et la faible. Les essais d'intégrer aussi la gravité semblent jusqu'ici n'avoir donné aucun résultat. Affaire à suivre ...
Le schéma le plus populaire aujourd'hui est le suivant. Jusqu'au temps $10^{-35}$ secondes, les forces nucléaire, électromagnétique et faible sont réunies en une seule force, de portée infinie, avec une constante de couplage dont la valeur est d'environ 2%. Cette force s'exerce sur une seule particule, laquelle peut exister dans de multiples états différents: électrons, neutrinos et quarks variés.
La température à cette époque était de $10^{24}$ eV ($10^{28}$ K). Chaque particule possédait assez d'énergie cinétique pour monter un piano du sol jusqu'au dixième étage ... Il se passe alors une quantité de réactions importantes qui pourraient expliquer l'absence d'anti-matière dans notre univers. Avec le refroidissement, les constantes de couplage commencent à diverger. L'électromagnétique et la faible varient lentement tandis que la nucléaire croît jusqu'à la valeur unité. Les portées de ces deux dernières se raccourcissent.
La théorie permet de comprendre le mécanisme de ces variations. Elles font intervenir un ensemble de particules nouvelles dont le rôle est de transporter les forces – au même titre que les photons transportent la force électromagnétique. Le tout est intégré dans une théorie de jauge à transformations locales et à symétries spontanément brisées. Il n'est pas dans le cadre de cet article d'expliquer en détail le sens de ces mots. La chose à retenir c'est que la physique moderne nous ouvre la possibilité de comprendre les mécanismes par lesquels les constantes ont pris leur valeur présente et que ces mécanismes sont liés à la baisse de température et de densité dans l'univers en expansion.

## Expansion et complexité de la matière

Dans l'organisation du monde, l'expansion joue encore un rôle fondamental à un tout

autre niveau. Considérons l'événement par lequel un proton et un électron se combinent pour donner un atome d'hydrogène. C'est un exemple typique d'organisation de la matière.

A ce moment un photon est émis. Ce photon, qui s'en va au loin, transporte à la fois l'excès d'énergie dont la perte va permettre au système de se lier, et l'excès d'entropie dont la perte va permettre au système de rencontrer les exigences de la deuxième loi de la thermodynamique. Si ce photon est réabsorbé par un autre atome d'hydrogène, tout se défait et aucune organisation nette n'a été acquise par l'univers. Il est donc indispensable que les photons émis soient «neutralisés», c'est-à-dire mis dans l'impossibilité d'agir à nouveau. Dans un univers stable, sans expansion, cela serait impossible: le photon finirait toujours par interagir à nouveau ... C'est l'expansion qui fait que la très grande majorité des photons émis n'interagira plus jamais (du moins jusqu'au prochain chapitre de contraction universelle, s'il y en a un ...) puisqu'ils se propagent dans un univers de plus en plus vide, où ils ont de moins en moins de chance d'être absorbés et puisque, à cause de l'expansion, ils perdent progressivement leur énergie, ce qui les rend doublement incapables de revenir sur la scène ...

Chaque étape d'organisation naturelle: nucléons en noyaux, noyaux en atomes, atomes en molécules, molécules en organismes, se passe dans des conditions analogues et exige la neutralisation de photons. Ce phénomène d'expansion qui semble n'intéresser que des entités très éloignées de nous paraît donc jouer un rôle fondamental dans l'organisation des choses, tant par son effet sur les constantes de couplage que par son rôle décisif sur l'organisation du monde à tous les niveaux.

## Littérature

Reeves, H. 1981. Patience dans l'azur. L'évolution cosmique. Editions du Seuil, 27, rue Jacob, Paris VI, France.

*Adresse de l'auteur:*

Prof. Dr. H. Reeves
Centre d'Etudes Nucléaires
de Saclay
F-91190 Gif-sur-Yvette (France)

# Ursprung und Evolution des Lebens auf molekularer Ebene

Manfred Eigen

## Biologische Komplexität

Das auffälligste Merkmal biologischer Organisation ist ihre Komplexität. Das wird besonders deutlich, wenn wir in das molekulare Detail eindringen. Das physikalische Problem der Lebensentstehung kann auf die Frage reduziert werden: Gibt es einen gesetzmässigen Mechanismus für die reproduzierbare Erzeugung von Komplexität? Eine Antwort auf die Frage, wie man sich die Entstehung biologischer Komplexität als gesetzmässigen Prozess vorstellen kann, ist bereits vor ca. 120 Jahren von Charles Darwin gegeben worden. Aus heutiger Sicht sind Darwins Thesen etwa folgendermassen zu formulieren:

- Komplexe Systeme entstehen evolutiv,
- Evolution basiert auf natürlicher Selektion.
- Natürliche Selektion ist eine gesetzmässige Konsequenz der Selbstreproduktion.

Die dritte These ist neo-darwinistischen Ursprungs. Sie geht aus den quantitativen Ansätzen der Populationsgenetik hervor, wie sie in der ersten Hälfte dieses Jahrhunderts vor allem von Haldane, Fischer und Wright ausgearbeitet wurden. Die molekularbiologische Revolution der fünfziger Jahre weckte die Euphorie, dass sich die Gesetze der Genetik auf die einfache Zauberformel

DNA → RNA → Protein → alles weitere

zurückführen liessen.
Dieses Dogma der Molekularbiologie postuliert, dass jedes Detail einer komplexen Struktur informationsgesteuert entsteht, wobei die Information un-umkehrbar von der genotypischen Legislative zur phänotypischen Exekutive der somatischen Seinsebene des Organismus fliesst. Heute, in den achtziger Jahren – nachdem wir reversible Transkriptasen, Restriktionsendonukleasen, Extrons und Introns, kurzum Teile des natürlichen Instrumentariums zur Verarbeitung genotypischer Information besser kennengelernt haben – sind wir mit unseren Aussagen etwas vorsichtiger:

- Alle Lebewesen müssen ihre genetische Information reproduzieren.
- Nur Nukleinsäuremoleküle sind sequenzgetreu reproduktionsfähig.
- Reproduktion ist nicht nur die Grundlage der Informationserhaltung, sondern auch der selektiven Informationsbewältigung und Optimierung.

Dreissig Jahre molekularbiologischer Forschung haben uns gezeigt, wie wir heute fragen müssen. Am Anfang steht die genaue experimentelle Beobachtung des grundlegenden Prozesses, nämlich die Reproduktion der genetischen Information. Daraus folgt die Abstraktion eines biologischen Grundprinzips sowie die experimentelle Verifizierung seiner logischen Konsequenzen. Schliesslich suchen wir in den biologischen Strukturen nach «fossilen» Spuren, die uns bestätigen, dass der historische Prozess der Lebenswerdung sich «im Prinzip» nach eben jenen Grundsätzen vollzogen hat. Im einzelnen sollen in enger Rückkopplung zwischen Theorie und Experiment folgende Fragen behandelt werden:

- Lässt sich zeigen, dass molekulare Selbstreproduktion Selektion und Evolution gesetzmässig bedingen?
- Ist Selbstorganisation auf der Grundlage von Selbstreproduktion und Selektion ein zwangsläufiger Prozess, dessen Voraussetzungen und Konsequenzen sich in natürlichen Systemen nachweisen lassen?

- Gibt es «fossile» Zeugnisse für den molekularen Evolutionsprozess?

## Experimente zur molekularen Evolution

Wir befassen uns zunächst ausführlicher mit dem Reproduktionsmechanismus eines Virus, dessen genetische Information in einem einsträngigen RNA-Molekül niedergelegt ist. Das Virus benötigt für seine Aufgaben im wesentlichen vier Funktionen, die durch Proteineinheiten repräsentiert sind: Ein Kapsid als Verpackungsmaterial zum Schutz gegen hydrolytischen Abbau, ein Penetrationsenzym zum Einschleusen seiner genetischen Information, einen Faktor zur Auflösung der Wirtszelle sowie einen Umfunktionierungsmechanismus, der die gesamte komplexe Maschinerie der Wirtszelle der Befehlsgewalt des Virus unterordnet. In dem von uns untersuchten Fall wird dieser von einem Proteinmolekül wahrgenommen, das sich mit drei ribosomalen Wirtsproteinen assoziiert und damit ein Enzym ergibt, das das Virusgenom exklusiv erkennt und sehr schnell reproduziert. Der gesamte Stoffwechsel- und Übersetzungsapparat wird damit dem ausschliesslichen Zweck untergeordnet, neue Viruspartikel zu produzieren. Allein in diesem Faktor ist das Prinzip der Virusinfektion begründet.
Der Reproduktionsmechanismus eines RNA-Bakterien-Virus ist in Abbildung 1 schematisch dargestellt.
Das aus vier Untereinheiten bestehende Enzym läuft vom 3′- zum 5′-Ende der Matrize. Der neugebildete Strang, die Replica, hat eine zur Matrize komplementäre innere Faltungsstruktur, die verhindert, dass ein Doppelstrang gebildet wird. Sol Spiegelman, der als erster das spezifische Reproduktionsenzym dieses Virus - $Q_\beta$ genannt - isolierte und mit seiner Hilfe in vitro infektiös wirksame Virus-RNA synthetisieren konnte, hat gezeigt, dass durch Tempern die Reproduktionsfähigkeit des Virus verloren geht. Die Reproduktion erfolgt nicht kontinuierlich. Das Enzym pausiert an sogenannten «pause sites». Vermutlich muss es warten, bis ein weiterer Teilbereich der Matrize aufgeschmolzen ist, um dann in relativ schnellem Durchlauf diesen Bereich zu kopieren. Spiegelman verdanken wir auch die Entdeckung und Isolierung einer nicht-infektiösen, etwa 220 Nukleotide langen RNA-Komponente. Dieses «Midivariante» genannte RNA-Molekül besitzt einen «Ausweis», so dass es vom Enzym ebenso wie die echte $Q_\beta$-RNA erkannt und dann - allerdings sehr viel schneller als diese - reproduziert wird. «Midivariante» ist also ein Schmarotzer, der selber keine Infektion bewirkt, da er das spezifische Reproduktionsenzym nicht aufbauen kann.
Wir haben in unserem Laboratorium den Mechanismus der RNA-Replikation mit Hilfe der $Q_\beta$-Replikase quantitativ studiert. Die Untersuchungen wurden von Manfred Sumper und Rüdiger Luce begonnen und später von Christof Biebricher und Rüdiger Luce fortgeführt. Unsere Kenntnisse über den Mechanismus der Replikation resultieren aus experimentellen Untersuchungen der Replikationsgeschwindigkeit als Funktion der Substrat-, Enzym- und RNA-Matrizen-

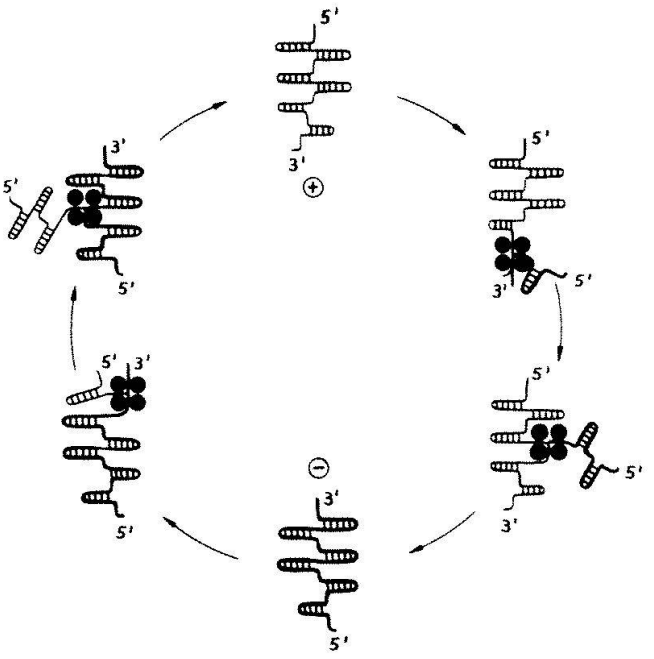



*Abb. 1.* Die einzelsträngige RNA des Bakterienvirus $Q_\beta$ reproduziert sich mit Hilfe eines Enzyms, $Q_\beta$-Replikase genannt, das aus vier Untereinheiten (schwarze Punkte) besteht. Das Enzym erkennt die Matrize spezifisch und läuft bei der Synthese vom 3′- zum 5′-Ende des Matrizenstranges. Die gebildete Replica (−) ist zur Matrize (+) komplementär. Aufgrund einer Symmetrie zwischen 3′- und 5′-Ende haben Plus- und Minus-Strang gleichartige 3′-Enden, die beide von der Replikase spezifisch erkannt werden. Der Minus-Strang wirkt daher ebenfalls als Matrize für die Bildung eines Plus-Stranges. Die innere Faltungsstruktur beider Stränge verhindert die Ausbildung einer Plus-Minus-Doppelhelix.

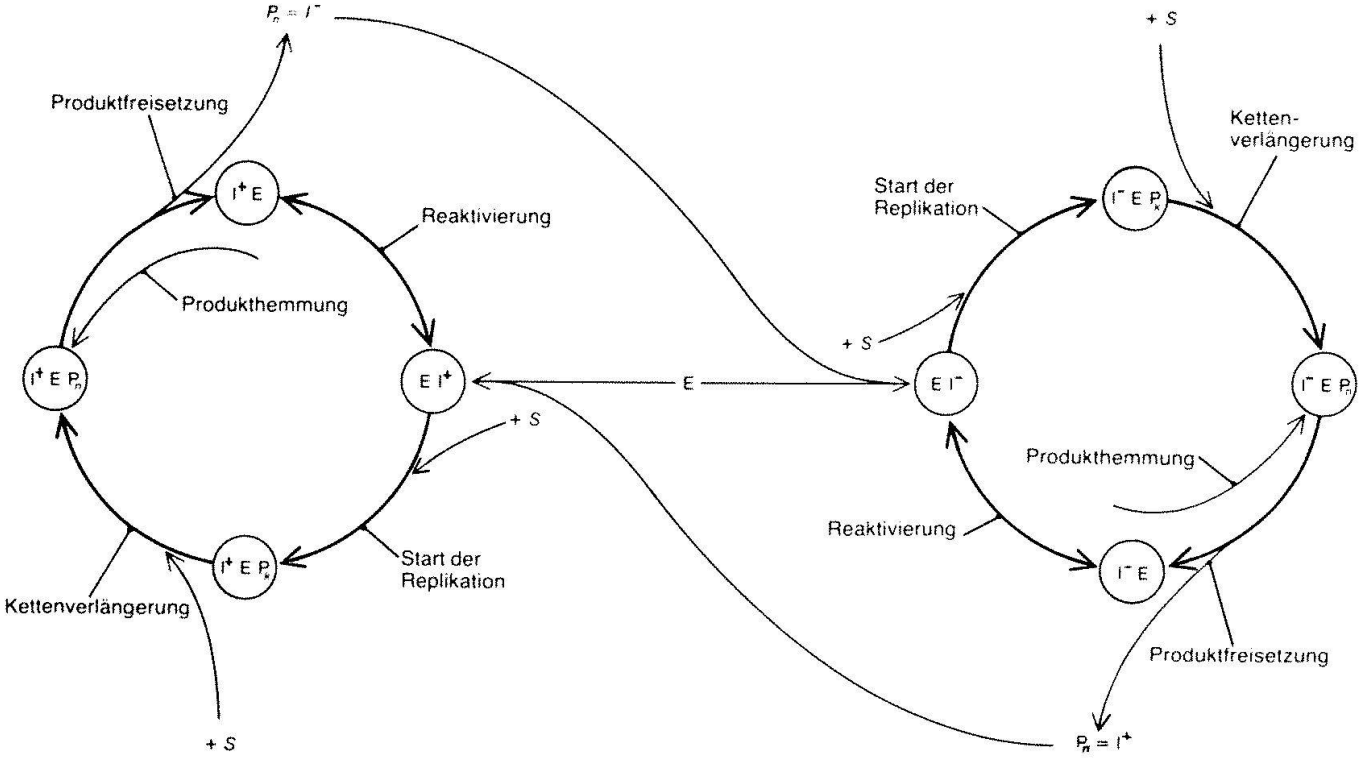
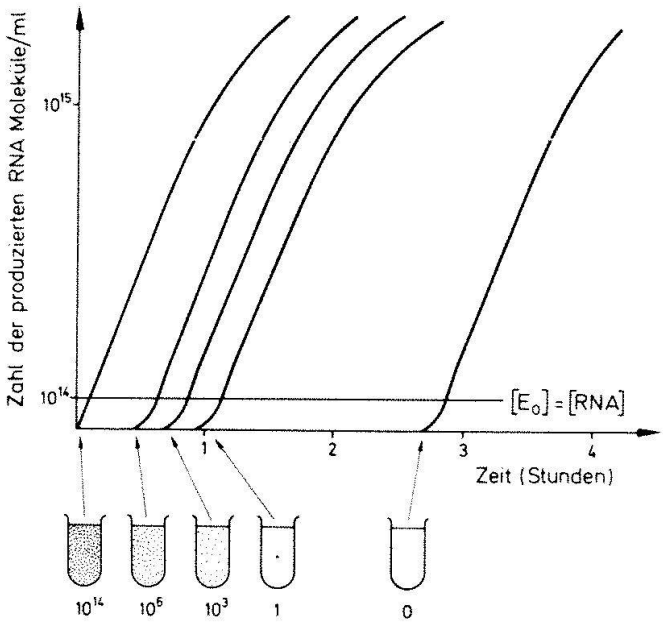



*Abb. 2.* Charakteristisch für den Mechanismus der RNA-Replikation sind die miteinander gekoppelten Synthesezyklen für den Plus- und Minus-Strang. Der katalytisch wirksame Komplex besteht aus dem Enzym, der Replikase, und einem RNA-Matrizen-Strang. Vier Phasen lassen sich in jedem Zyklus unterscheiden: Start der Replikation durch Anlagerung von mindestens zwei Substratmolekülen (Nukleosidtriphosphaten). Elongation des Replica-Stranges durch sukzessiven Einbau von Nukleotiden, Freisetzung der fertigen Replica, Dissoziation des am 5′-Ende der Matrize gebundenen Enzyms und Reassoziation am 3′-Ende eines Matrizen-Stranges. I repräsentiert die Matrix (Information), E das Enzym, P das Reaktionsprodukt, das nach Fertigstellung ($P_n$) und Freisetzung als Matrize (I) wirkt. S ist das Substrat, also jeweils eins der vier Nukleosidtriphosphate von A, U, G und C.

*Abb. 3.* Inkubiert man Syntheselösungen (Puffer, Ionen, $Q_\beta$-Replikase, Nukleosidtriphosphate und RNA-Matrizen) mit einer äquivalenten Menge von Enzymmolekülen und Matrizen-Strängen, so erfolgt die Vermehrung der RNA-Moleküle linear mit der Zeit, denn die Zahl der katalytisch wirksamen Komplexe ist gleich der konstanten Zahl der Enzymmoleküle. (Das Abbiegen der Reaktionskurven bei hohen RNA-Konzentrationen resultiert aus einer Hemmung des Enzyms, hervorgerufen durch Bindung von überschüssiger RNA im aktiven Zentrum). Verringert man die Matrizenkonzentration seriell um jeweils den gleichen Faktor, so verschieben sich die Wachstumskurven um konstante Abschnitte auf der Zeitachse. Diese logarithmische Abhängigkeit der Induktionszeiten von der Verdünnung zeigen ein exponentielles Wachstumsgesetz, das gültig ist, solange das Enzym im Überschuss von RNA ist. Hier nimmt die Zahl der katalytisch aktiven Komplexe aufgrund der RNA-Vermehrung ständig zu (Autokatalyse). Auch wenn kein Matrizen-Strang vorhanden ist, entsteht nach langer Induktionszeit «de novo» RNA, die als Matrize wirkt und schnell vermehrt wird. Diese «de novo»-Synthese von RNA ist eine besondere Eigenschaft der $Q_\beta$-Replikase.



Konzentrationen, der analytischen Behandlung eines Replikationsmodells (siehe Abbildung 2) und der Computersimulation dieses Modells mit realistischen Parametern, die aus den experimentellen Daten gewonnen wurden.
Ein typisches Experiment ist in Abbildung 3 skizziert. Gemessen wird die Menge neu synthetisierter RNA - indiziert durch ein $P^{32}$-markiertes Nukleotid - als Funktion der Zeit. Die Konzentrationen der Substrate (das sind die vier Nukleosidtriphosphate von A, U, G und C) sowie des Enzyms sind konstant. Die Anfangskonzentration der RNA hingegen wird seriell variiert, indem man jeweils um einen konstanten Faktor verdünnt. Die Messkurven, nämlich der Anstieg der RNA-Konzentration mit der Zeit, lassen sich in drei Phasen unterteilen:

1. Eine Induktionsperiode, die mit steigender Verdünnung logarithmisch zunimmt. Verdünnung um einen konstanten Faktor bedeutet jeweils eine konstante Verschiebung auf der Zeitachse.

2. Ein linearer Anstieg der RNA-Konzentration mit der Zeit, der einsetzt, wenn die RNA-Matrizen-Konzentration ungefähr gleich der Enzymkonzentration ist.

3. Ein Plateau, das erst erreicht wird, wenn die RNA-Konzentration sehr gross gegenüber der Enzymkonzentration ist.

Diesem Reaktionsverhalten liegt folgender kinetischer Sachverhalt zugrunde: Die Reaktion, deren Produkt neue RNA-Matrizen sind, wird katalysiert durch einen Komplex aus Enzym und Matrize. Die Affinität zwischen beiden katalytischen Partnern ist so hoch, dass bei der gewählten Enzymkonzentration (ca. $10^{-7}$ Mol/Liter) zunächst jedes RNA-Molekül ein Enzymmolekül bindet. Die Zahl der katalytisch aktiven Komplexe nimmt exponentiell zu, und zwar bis die RNA-Konzentration gleich der Enzymkonzentration ist. Alle Enzymmoleküle sind dann mit Matrizen gesättigt. Die lineare Phase deutet an, dass die Zahl der pro Zeiteinheit neu synthetisierten RNA-Moleküle nunmehr konstant bleibt, eben weil die Zahl katalytisch aktiver RNA-Enzymkomplexe sich nicht mehr ändert. Die neugebildeten RNA-Moleküle binden aber nicht nur als Matrize an das Enzymmolekül, sondern auch - wenngleich mit geringerer Affinität - am Syntheseort des Produkts. In der linearen Phase setzt daher eine Hemmung durch überschüssige RNA-Moleküle ein, die schliesslich die Reaktion vollständig zum Stillstand bringt. Durch Abwandlung der Versuchsbedingungen, wie Veränderung der Substrat-Konzentration, der Enzym-Konzentration oder des Verhältnisses der Anfangskonzentration von Plus- und Minus-Strang der Matrize, konnte eine quantitative Zuordnung der kinetischen Parameter erzielt und das in Abbildung 2 gezeigte Reaktionsschema verifiziert werden. Die wesentlichen Ergebnisse dieser Untersuchung sind:

- Die Replikation wird von einem Komplex katalysiert, der aus einem Enzym- und einem RNA-Molekül besteht.
- Die Wachstumsrate ist demzufolge der kleineren von beiden Bruttokonzentrationen (Enzym bzw. RNA) proportional. Das bedeutet für kleine RNA-Konzentrationen exponentielles, für grosse RNA-Konzentrationen lineares Wachstum.
- Bei Konkurrenz zwischen verschiedenen Mutanten wächst auch im linearen Bereich die vorteilhafte Mutante exponentiell an, bis diese das Enzym vollkommen sättigt.
- Selektionsvorteil in der linearen Phase basiert allein auf der Kinetik der Matrizen-Enzymbindung. In der exponentiellen Phase dagegen werden die Erzeugungsraten der verschiedenen Mutanten bewertet.
- Plus- und Minus-Strang tragen in der Exponentialphase mit dem geometrischen, in der linearen Phase mit dem harmonischen Mittel ihrer Geschwindigkeitsparameter zur Wachstumsrate bei. Das Ensemble von Plus- und Minus-Strang wächst konform in einem von ihren jeweiligen Geschwindigkeitsparametern abhängigen Verhältnis hoch.
- Die Replikationszeit hängt von der Länge der zu synthetisierenden RNA-Kette ab und ist von der mittleren (reziproken) Substratkonzentration (A, U, G, C) abhängig. Diese Abhängigkeit ist schwächer als linear, da die sukzessiv eingebauten Substrate bereits teilweise am Enzym gebunden vorliegen.

Wir kommen nun zur eigentlichen Fragestellung: Was sind die gesetzmässigen Konsequenzen der Replikation? Lassen sich mit Hilfe eines replikativen Systems evolutiv neue Eigenschaften erzeugen? Sind die Merkmale des replikativen Systems hinreichend, oder bedarf es noch anderer essentieller Eigenschaften?

Sol Spiegelman hatte auch hier schon eine wichtige Anregung gegeben. Durch serielle Übertragungen von RNA-Matrizen in Nährmedien hatte er am Ende der Versuchsreihe Varianten des $Q_{\beta}$-Genoms selektiert, die zwar nicht mehr infektiös waren, sich dafür aber durch eine höhere Replikationsrate (bezogen auf das einzelne Nukleotid) auszeichneten. Sie entkamen auf diese Weise dem Selektionsdruck der Ausdünnung durch schnellere Replikation. Das Wesentliche dabei war, dass die neuen Varianten nicht nur spezifisch schneller wuchsen, sondern dass sie anstelle der ursprünglich 4500 nur noch 500 Nukleotide besassen und somit auch die Replikationsrunde sehr viel schneller beenden konnten. Eine solche Evolution, bei der Information verlorengeht, mag eher als Degeneration bezeichnet werden. Doch zeigen die Untersuchungen, dass dieses Replikationssystem sehr anpassungsfähig ist – eine wesentliche Voraussetzung für evolutives Verhalten.

Die Versuche gewannen an Aktualität, als Manfred Sumper im Jahre 1974 eine überraschende Beobachtung machte. In Verdünnungsreihen, die soweit getrieben wurden, dass in der Probe nur noch mit sehr geringer Wahrscheinlichkeit überhaupt eine Matrize vorhanden sein konnte, entstand dennoch reproduzierbar und homogen ein Molekül, das etwa die Grösse und die Struktur der von Spiegelman isolierten «Midivariante» hatte. In Abbildung 3 ist dieses Phänomen angedeutet. Zum Unterschied von den matrizengesteuerten Replikationen ist diese Synthese mit einer unverhältnismässig langen und von den Versuchsbedingungen in kritischer Weise abhängigen Induktionsperiode verknüpft. Sumper war sofort überzeugt, dass er eine vom $Q_{\beta}$-Enzym «erfundene» und «de novo» synthetisierte Variante in den Händen hatte. (Seine Fachkollegen dagegen plädierten fast sämtlich auf eine durch das Enzym eingeschleppte Verunreinigung.) Konnte Sumper bereits durch eigene Versuche die «Verunreinigungs»-Hypothese ausschliessen, so liegt heute, vor allem durch Christof Biebrichers und Rüdiger Luces Experimente, der Beweis für die «de novo»-Synthese vor. Aus den nunmehr bekannten kinetischen Daten folgt, dass die Induktionsperioden bei der matrizengesteuerten und der «de novo»-Synthese vollkommen verschiedenartigen Gesetzen gehorchen. So wird beispielsweise bei der matrizengesteuerten Synthese lediglich ein Enzymmolekül benötigt, und die Substrate werden sukzessive einzeln zugeführt. Für die «de novo»-Synthese ist dagegen ein Komplex aus mehreren Enzymmolekülen erforderlich, und der geschwindigkeitsbestimmende Schritt besteht im Aufbau eines Keimes von mindestens drei Substratmolekülen. Das wesentliche Beweisstück aber wird geliefert durch den Nachweis, dass in der Frühphase der Synthese stets unterschiedliche «de novo»-Varianten auftreten, die unter Selektionsdruck in ihrer Länge zunehmen und bei Variation der Versuchsbedingungen zu verschiedenartigen Endprodukten führen. Letzteres war auch schon von Manfred Sumper gezeigt worden. Er erhielt verschiedene «Mini-Varianten», die unter Bedingungen – z. B. in Gegenwart von Reaktionshemmern – aufwuchsen, unter denen der Wildtyp gar nicht mehr existenzfähig war.

Das entscheidende Experiment von Biebricher und Luce ist in Abbildung 4 dargestellt: Eine mit hochgereinigten Enzymen und Substraten angesetzte Syntheselösung wird durch Temperaturerhöhung inkubiert, und zwar für eine Zeit, die ausreicht, etwa vorhandene Matrizen hochzuverstärken, die aber zu kurz ist, fertige «de novo»-Produkte zu ermöglichen. Anschliessend wird die Lösung kompartimentiert. In den einzelnen Kompartimenten wird sodann für eine Zeit inkubiert, die für eine «de novo»-Synthese ausreichend ist. Die erhaltenen Produkte werden nach der Fingerprint-Methode analysiert und miteinander verglichen. Im Falle der Gültigkeit der «Verunreinigungs»-Hypothese sollten aufgrund der Verstärkung in der Anfangsphase alle Kompartimente das gleiche Produkt enthalten. Im Falle der «de novo»-Hypothese sollten sich dagegen die Produkte unterscheiden, da die Synthese an verschiedenen Enzymmolekülen mit verschiedenen «de novo»-Produkten beginnt. Selektion, das heisst bevorzugte Reproduk-

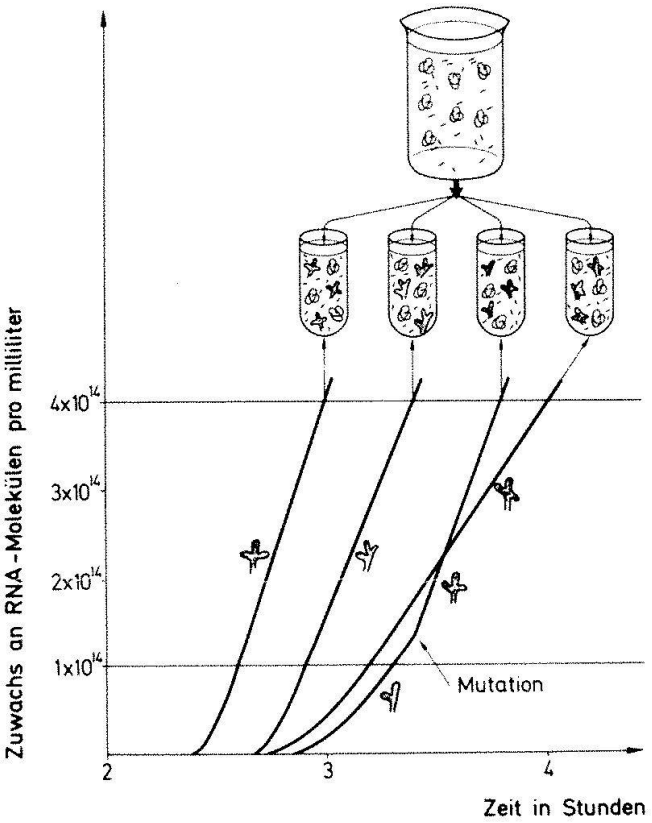



*Abb. 4.* Eine Lösung aus Nukleosidtriphosphaten wird in Gegenwart von $Q_\beta$-Replikase gerade lange genug inkubiert, dass jede als Verunreinigung des Enzyms vorhandene Matrize vielfach vermehrt würde. Der Inkubationsprozess wird aber unterbrochen, bevor auch nur eine Matrize «de novo» entstehen konnte. Anschliessend kompartimentiert man diese Lösung und inkubiert wiederum, nunmehr lange genug, dass «de novo»-Produkte entstehen und sich vermehren können. Man analysiert die in den verschiedenen Kompartimenten gebildete RNA nach der Fingerprintmethode. Es ergeben sich unterschiedliche Reaktionsprodukte. Manchmal gibt sich in der Wachstumskurve noch eine Mutation zu erkennen. Während für die matrizengesteuerte Synthese die Induktionszeiten eindeutig festgelegt sind (als Folge einer Überlagerung vieler Einzelprozesse), ist hier eine Streuung der Induktionszeiten zu beobachten. Das weist darauf hin, dass der auslösende Schritt ein molekularer Einzelprozess ist, der anschliessend sehr schnell «verstärkt» wird.

tion eines «de novo»-Stranges konnte noch nicht stattfinden, da innerhalb der kurzen Inkubationzeit keines der «de novo»-Produkte fertiggestellt war.

Das Experiment ergab viele unterschiedliche Produkte. Erst wenn man diese wieder vereinigte und die Lösung weiter inkubierte, wuchs schliesslich eine Variante homogen und reproduzierbar auf. Die frühesten Produkte, die abgefangen werden konnten, hatten Längen von etwa 70 Nukleotiden. Im Verlauf der Evolution erschienen längere Ketten, z. B. bei hohen Salzkonzentrationen die «Midivariante», die etwa 220 Nukleotide einschliesst.

Das fundamentale Ergebnis dieser Experimente liegt jedoch nicht allein in der Aufklärung der Besonderheit des $Q_\beta$-Systems. Man hat nunmehr ein flexibles Replikationssystem zur Verfügung, mit dem man eine Reihe höchst interessanter Produkte aufbauen kann. Hier zeigt sich vor allem, dass Selektion und Evolution gesetzmässige Konsequenzen der Selbstreplikation sind und sich als solche quantitativ studieren lassen. So konnte die Frage der schnellen Optimierung unter extremen Versuchsbedingungen bis ins Detail beantwortet werden. Die Resultate der beschriebenen Evolutionsexperimente können in vier Hauptaussagen zusammengefasst werden:

- «De novo»-Synthese von RNA durch $Q_\beta$-Replikase erfolgt nach einem grundlegend anderen Mechanismus als matrizengesteuerte RNA-Synthese. Der aktive Reaktionskomplex enthält mindestens zwei Enzymmoleküle und bedarf eines Keimes von drei bis vier Substratmolekülen.
- Keimbildung ist der geschwindigkeitsbestimmende Schritt, während Elongation und Reproduktion sehr schnell erfolgen. Die Einmaligkeit des molekularen Prozesses der Reaktionsauslösung ist in der Streuung der Induktionszeiten für die «de novo»-Synthese zu erkennen.
- «De novo»-Synthese erzeugt ein breites Spektrum von Mutanten unterschiedlicher Länge, das für verschiedenste Umweltbedingungen leicht adaptierbare Sequenzen enthält.
- Initiation von Selbstreproduktion ist offensichtlich hinreichend, evolutive Optimierung in Gang zu setzen.

Dieser Befund macht es möglich, einen Evolutionsreaktor zu entwickeln, in dem sich in relativ kurzer Zeit optimal reproduzierende RNA-Sequenzen herstellen lassen. Daraus kann auch ein Prinzip zur Evolution von RNA-Strukturen mit optimalen Übersetzungsprodukten entwickelt werden. Versuche in dieser Richtung sind im Gange.

Selbstreplikation und Mutagenität in einem offenen System (weitab vom Gleichgewicht) sind also hinreichend für selektives und evolutives Verhalten. Auch in relativ einfachen Replikationssystemen lassen sich (gemessen

am Wildtyp) optimale Eigenschaften in vitro innerhalb kurzer Generationsfolgen reproduzierbar erzeugen. Derartige Auswirkungen müssen die Konsequenz eines physikalischen Prinzips sein. Kann ein solches Prinzip quantitativ formuliert werden?

## Selektion und Evolution als naturgesetzliche Vorgänge

In einer Reihe von früheren Veröffentlichungen ist gezeigt worden, dass das Selektionsprinzip aus den Voraussetzungen eines selbstreplikativen Systems heraus als Extremalprinzip ableitbar ist. Es besagt, dass bei inhärenter linearer Autokatalyse die relativen Populationsvariablen Werte annehmen, die einer optimalen Reproduktionseffizienz des Gesamtsystems entsprechen. Die relativen Konzentrationsverhältnisse der stationären Population sind nach kurzer Induktionszeit unabhängig von den zeitlichen Veränderungen des Gesamtsystems. Die Population besteht aus einem singulären Wildtyp (oder mehreren gleichwertigen, d.h. «entarteten» Varianten) und einem Mutantenspektrum. Der Wildtyp erscheint innerhalb seiner Mutantenverteilung relativ zu jeder Einzelmutante am häufigsten, macht aber in einer gut adaptierten Population nur einen kleinen Bruchteil der Gesamtmenge aus. Der Quotient der Populationsvariablen ($x_i$) von Einzelmutanten (i) und Wildtyp ist durch die jeweiligen Geschwindigkeitsparameter für Mutation $W_{im}(m \rightarrow i)$ und Reproduktion $W_{mm}(m \rightarrow m)$ bzw. $W_{ii}(i \rightarrow i)$ bestimmt:

$$\frac{x_i}{x_m} = \frac{W_{im}}{W_{mm} - W_{ii}} .$$

Von Bedeutung sind ferner die Variablen:

$\bar{q}$ = mittlere Kopierungsgenauigkeit eines Nukleotids

bzw. $1 - \bar{q}$ = mittlere Fehlerrate pro Nukleotid

$\nu_i$ = Zahl der Nukleotide in der Sequenz i

$Q_i \approx \bar{q}^{\nu_i}$ = Bruchteil korrekter Replikationen

$\sigma_m$ = Superiorität des Wildtyps gegenüber seinem Mutantenspektrum (entspricht im allgemeinen dem Verhältnis von Replikationsrate des Wildtyps und mittlerer Replikationsrate der Mutantenverteilung.)

Die Konsequenzen dieses für Darwinsche Systeme gültigen Extremalprinzips sind:

- Selektion einer vom Wildtyp beherrschten Mutantenverteilung. Diese ist nur solange stabil, wie die Bedingungen $\sigma_m > 1$ und $Q_m > \sigma_m^{-1}$ erfüllt sind.
- Evolution durch Selektion neu auftretender Mutanten, die aufgrund eines Selektionsvorteils die Stabilitätsbedingung $\sigma_m > 1$ verletzen und daher eine Destabilisierung des bis dahin dominanten Wildtyps bewirken.
- Begrenzung des Informationsgehaltes aufgrund der Bedingung $Q_m > \sigma_m^{-1}$. Der Grenzwert errechnet sich zu $\nu_{max} = \frac{\ln \sigma_m}{1 - \bar{q}_m}$.

  Er entspricht etwa der reziproken mittleren Fehlerrate $(1 - \bar{q}_m)$, sofern $\sigma_m$ genügend gross gegen 1 ist. Wenn der Informationsgehalt $\nu_m$ des Wildtyps nahe an den Grenzwert $\nu_{max}$ herankommt, wird $Q_m$ etwa gleich $\sigma_m^{-1}$. Dann ist der Anteil des Wildtyps an der Gesamtpopulation sehr klein:

$$\frac{x_m}{\sum_k x_k} = \frac{Q_m - \sigma_m}{1 - \sigma_m^{-1}} .$$

Beide Prozesse: Selektion als Stabilisierung einer bestimmten Verteilung sowie Evolution als sukzessive Etablierung neuer Populationen resultieren aus «innerem Zwang». Sie sind das unausweichliche Resultat selbstreproduktiven Verhaltens.

Dass ein solcher Optimierungsprozess der Evolution tatsächlich auf «hohe Berge» führt und nicht auf «niedrigen Hügeln» stehen bleibt, liegt an der Topologie des vieldimensionalen Mutantenraumes. Betrachten wir als Beispiel eine binäre Sequenz mit $\nu$ Positionen. Wir können jeder Position der Sequenz eine Koordinate mit zwei Punkten zuweisen und erhalten dann einen $\nu$-dimensionalen Phasenraum, in dem jeder der $2^\nu$ Punkte eine Mutante darstellt. Der Evolutionsprozess kann dann als eine Route in

diesem Raum beschrieben werden, die durch einen ständig ansteigenden Selektionswert charakterisiert ist. Die Topologie eines solchen vieldimensionalen Raumes ist unserer Anschauung nicht leicht zugänglich; die «Gebirge» sind hier äusserst bizarr. Denn obwohl es $2^{\nu}$ Punkte gibt, ist die grösste Entfernung nur $\nu$. Es gibt Sattelpunkte verschiedener Ordnung, bei denen ein Fortschreiten in k-Richtungen bergan und in $\nu$–k Richtungen bergab führt. Aufgrund dieses Sachverhaltes genügen relativ kleine Mutationssprünge, um das System immer wieder eine ansteigende Route finden zu lassen. Es gibt eine optimale Zahl $\nu$, bei der die Anzahl der Routen bereits gross genug und Mehrfachmutationen so wahrscheinlich sind, dass ein optimaler «Gipfel» erreicht werden kann.

Fassen wir zusammen: Selektion, Evolution und Anpassung an ein Optimum sind Prozesse, die nach physikalischen Gesetzen ablaufen und die sich quantitativ formulieren lassen. Hier ist zu beachten, dass eine Theorie keineswegs den realiter in der Natur ablaufenden Prozess beschreibt, für den ja die Ausgangssituation, die komplexen Randbedingungen sowie die mannigfach überlagerten Störeinflüsse weitgehend unbekannt sind. Die Theorie sagt uns lediglich, was aus bestimmten Voraussetzungen bei Einhaltung bekannter Randbedingungen folgt. Sie erklärt die in der Natur zu beobachtenden, reproduzierbaren Regelmässigkeiten in einer «Wenn - dann»-Beschreibung. Das gilt auch für die hier dargestellte Theorie. Sie hilft uns zunächst, die oben geschilderten Experimente zu verstehen und auszuwerten. Unter den klar definierten Anfangs- und Randbedingungen im Laboratorium lässt sich die Theorie quantitativ bestätigen. Für die in der Natur ablaufenden Ereignisse zeigt sie lediglich Tendenzen, Minimalforderungen, Begrenzungen und eventuelle Auswirkungen auf. Hier muss wiederum experimentell überprüft werden, ob die Schlussfolgerungen, die sich aus der Theorie ergeben, auch für die natürlichen Prozesse relevant sind.

Als Folgerungen sind vor allem zu nennen:

Die Informationsmenge, die in einer molekularen Population zur Selektion gelangen kann, ist von der mittleren Fehlerrate und dem mittleren Selektionsvorteil des Wildtyps abhängig. Beim Überschreiten der kritischen Fehlerschwelle akkumulieren sich die Fehler derart, dass die Information der Wildtypsequenz restlos verloren geht.

Die Adaptationsfähigkeit des Wildtyps ist in der Nähe der Fehlerschwelle am grössten. Bei der für eine stabile Verteilung erreichbaren Informationsmenge zeigt das Mutantenspektrum grösstmögliche Variabilität. Ein solches System ist gegenüber Änderungen der Umweltbedingungen äusserst flexibel. Der Wildtyp ist als individuelle Sequenz absolut dominierend, ist jedoch im Vergleich zur Gesamtheit des Mutantenspektrums nur in geringer Menge vorhanden.

## Molekulare Zeugnisse der natürlichen Evolution

Die Aussagen der Evolutionstheorie lassen sich an natürlichen Systemen testen. Charles Weissmann und Mitarbeiter haben für $Q_{\beta}$-Viren die folgenden Ergebnisse erhalten:

Der Wildtyp hat eine definierte Sequenz, was aber nicht heisst, dass die überwiegende Zahl der Viren exakt die gleiche RNA-Sequenz besitzt. Es bedeutet lediglich, dass sämtliche Sequenzen bei Überlagerung eine eindeutige Nukleotidabfolge, nämlich die des Wildtyps ergeben.

Klonierung einzelner Viren bzw. einzelner Virus-RNA-Moleküle mit anschliessender schneller Vermehrung führt zu Populationen mit unterschiedlichen Sequenzen. Die Sequenzen weichen im allgemeinen in einer oder in zwei Positionen vom Wildtyp ab, der selbst kaum in irgendeinem Klon auftritt (siehe Abbildung 5). Aus der Tatsache, dass der Wildtyp nur einen (vernachlässigbar) kleinen Anteil des Mutantenspektrums ausmacht, lässt sich folgern, dass der Schwellenwert des Informationsgehalts mit der tatsächlichen Informationsmenge des Wildtyps ($\nu_m$) nahezu identisch ist.

Gezielt erzeugte extra-cistronische Einfehlermutanten (das sind nicht-lethale Mutationen ausserhalb der zur Übersetzung gelangenden Sequenzen) revertieren zum Wildtyp. Sie regenerieren ihre Fehler mit einer Wahrscheinlichkeit von ca. $3 \times 10^{-4}$. Die quantitative Auswertung erlaubt die Bestimmung von Fehlerrate und Wachstumsvorteil des Wildtyps und damit auch eine Festlegung der kritischen Fehlerschwelle für die optima-

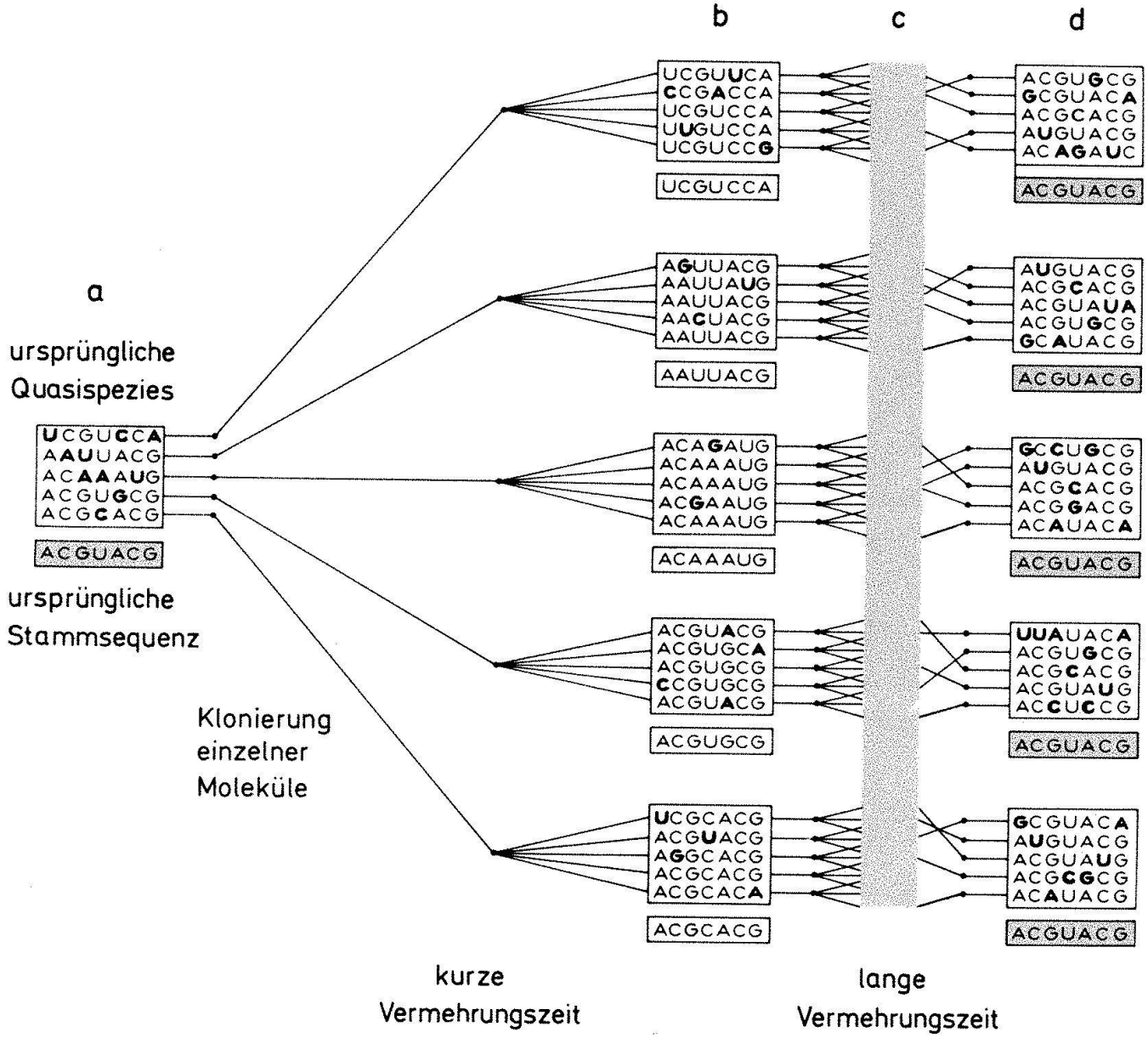



*Abb. 5.* In diesem von Charles Weissmann und Mitarbeitern ausgeführten Experiment wurden einzelne $Q_{\beta}$-RNA Moleküle (bzw. Viren) aus einer Wildtypverteilung (a) in Coli-Bakterien kloniert. Die nach schneller Vermehrung der jeweils eingefangenen RNA-Moleküle erhaltenen Klone (b) wurden nach der Fingerprintmethode (zweidimensionale Elektrophorese teilweise gespaltener RNA-Sequenzen) analysiert und miteinander verglichen. Es zeigten sich durchweg Unterschiede in ein oder zwei Positionen der Sequenzen. Nach langer Vermehrungszeit (c) bildet sich in allen Klonen schliesslich wieder die Wildtypverteilung (d) aus, das heisst die Mittelwerte der Sequenzen werden sämtlich wieder identisch.

le Reproduktion der Informationsmenge. Dieser Wert stimmt innerhalb der Messfehler mit der vorhandenen Informationsmenge (4500 Nukleotide) überein. Die Tatsache, dass eine Klonierung von Einzelmutanten überhaupt möglich ist, liegt daran, dass in der Mutantenverteilung des Wildtyps die Mutanten dominieren, deren Wachstumsrate der des Wildtyps sehr ähnlich ist. Diese werden bei der zum Klonieren von Einzelmolekülen notwendigen seriellen Verdünnung bevorzugt «herausgefischt». Da sie sich fast so schnell vermehren wie der Wildtyp, bauen sie zunächst ein Mutantenspektrum auf, dessen mittlere Sequenz mit der der klonierten Einzelmutante identisch ist. In diesem Mutantenspektrum muss irgendwann auch der Wildtyp erscheinen. Er kann sich aber nur langsam durchsetzen, und zwar mit einer Rate, die der (kleinen) Differenz der Wachstumsgeschwindigkeit von Wildtyp und klonierter Mutante entspricht. Am Ende dominiert natürlich in jedem Klon wieder der Wildtyp (siehe Abbildung 5).

Weiterhin kann der Schluss gezogen werden, dass alle einsträngigen RNA-Viren ähnlichen Einschränkungen hinsichtlich des Informationsgehaltes unterworfen sind. Denn in der Natur gibt es keine (einsträngigen) RNA-Viren, deren replikative Einheit mehr Information als in der Grössenordnung von $10^4$ Nukleotiden enthält. Alle grösseren Viren besitzen doppelsträngige Nukleinsäuren oder setzen sich aus mehreren replikativen Einheiten zusammen. Für diese gelten wiederum analoge Beziehungen zwischen Fehlerschwelle und Reproduzierbarkeit der Informationsmenge. DNA-Polymerasen müssen im allgemeinen wesentlich genauer als RNA-Replikasen arbeiten. Dieses geschieht vermittels zusätzlicher Möglichkeiten zur Fehlererkennung und Korrektur.

Was bedeuten diese Ergebnisse für die frühe Evolution?

Die ersten replikativen Einheiten müssen einen erheblich geringeren Informationsgehalt besessen haben als die mit einer optimalen Reproduktionsmaschinerie arbeitenden RNA-Viren. Die Reproduktionsgenauigkeit hängt in Abwesenheit von (wohl-adaptierten) Enzymen allein von der Stabilität der Basenpaare ab. Das GC-Paar hat unter diesen Umständen gegenüber dem AU-Paar einen ca. zehnfachen Selektionsvorteil. Modellexperimente zeigen, dass für GC-reiche Polynukleotide eine Fehlerrate von $10^{-2}$ kaum unterboten werden kann. Die ersten «Gene» müssen demnach Polynukleotide einer Kettenlänge $\lesssim 100$ gewesen sein.

Molekulare Evolution erfordert inhärente Selbstreproduktivität. RNA scheint diese Forderung am besten zu erfüllen. In der zeitlichen Abfolge entstand RNA aufgrund ihres komplexeren Aufbaus sicherlich später als Proteine oder proteinähnliche Substanzen. Proteine können an gewisse Funktionen zufällig adaptiert sein, doch folgte eine solche Adaptation lediglich strukturellen und nicht funktionellen Kriterien. Anpassung an optimale Funktion erfordert dagegen einen inhärenten Reproduktionsmechanismus. Der einzige logisch begründbare Weg, die ungeheure funktionelle Kapazität der Proteine evolutiv zu erschliessen, liegt in der Verknüpfung beider Stoffklassen, also in der Übersetzung der in den selbstreproduktiven RNA-Strukturen gespeicherten Information.

Daraus ergibt sich sofort die Frage: Konnte RNA ohne enzymatische Unterstützung, das heisst ohne Replikasen überhaupt entstehen? Experimente, die von Leslie Orgel und Mitarbeitern ausgeführt wurden, bejahen diese Frage. So wurde gefunden, dass $Zn^{2+}$-Ionen – die heute in allen Replikasen als Ko-faktoren enthalten sind – ausgezeichnete Katalysatoren für die 3′-5′-Verknüpfung von Nukleotiden darstellen und eine matrizengesteuerte Synthese von Polymeren ermöglichen. Gezeigt wurde dies zunächst für poly-C als Matrizenstrang. Bietet man diesem aktivierte G- und A-Nukleotide in gleichem Konzentrationsverhältnis an, so wird G je nach Reaktionsbedingungen 30- bis 200fach bevorzugt.

Damit wird nahegelegt, dass in einem chemisch genügend reichhaltigen Milieu GC-reiche RNA-Stränge einer Kettenlänge von ca. 100 Nukleotiden spontan entstehen, sich stabil reproduzieren und evolutiv adaptieren können.

Lassen sich für diese ersten «Gene» heute noch Zeugnisse finden?

Der Informationsgehalt solcher «Gene» reicht nur für relativ kleine, sicherlich noch nicht optimal angepasste Proteine aus. Das bedeutet aber, dass diese als Informationsträger inzwischen längst überholt und somit verdrängt worden sind. Der Verdrängungsprozess lief mit der Ausbildung der Translationsmaschinerie einher. Im Translationsapparat sind RNA-Strukturen nicht nur als Informationsquellen, sondern auch als Funktionsträger und Zielstrukturen wirksam. Es besteht viel eher die Möglichkeit, dass sie in dieser Funktion im Übersetzungsapparat, z.B. als Adaptoren in Form der Transfernukleinsäuren (t-RNA) oder als ribosomale Nukleinsäuren (r-RNA) bis auf den heutigen Tag überlebt haben. Da die funktionellen RNA-Moleküle keine genetische Information zu speichern hatten, unterlagen sie nach erfolgter struktureller Anpassung kaum noch einem Evolutionszwang. Rekrutiert wurden die funktionellen Nukleinsäuren zunächst aus dem gleichen Reservoir wie ihre informationstragenden Schwestermoleküle, die m-RNAs. Wir erwarten daher, dass uns die t-RNA zum Beispiel als Zeuge der frühen Evolution des Translationsapparates Auskunft über Aufbau und Struktur der ersten «Gene» zu liefern vermag. Dieses Molekül entspricht mit einer Kettenlänge von ca. 76

Nukleotiden ideal den von der Theorie geforderten, an gegenwärtigen Strukturen experimentell bestätigten und für präbiotische Bedingungen relevanten Kriterien.

Von der t-RNA sind heute sehr viele Sequenzen bekannt, und zwar sowohl bei gegebenem Anticodon für eine Vielzahl phylogenetischer Stufen, als auch bei gegebener Spezies für eine grosse Zahl von unterschiedlichen Anticodons. Beide Fälle sind für eine komparative Analyse gleichermassen interessant: Die phylogenetische Analyse offenbart, ob t-RNA noch Information aus präbiotischer Zeit bewahrt hat oder ob diese im Verlauf der Phylogenie weitgehend verloren gegangen ist. Der Vergleich verschiedener t-RNA-Moleküle innerhalb einer Spezies kann dann zu einer weitgehenden Rekonstitution der Urstrukturen führen und über die frühe Evolution des Translationsapparates Aussagen machen.

Der in Abbildung 6 gezeigte phylogenetische Stammbaum für t-RNA$^{met\,in}$ zeigt, dass t-RNA zu den konservativsten Strukturen gehört, die wir kennen. So unterscheiden sich in dem gezeigten Beispiel die Fruchtfliege (Drosophila) und der Seestern lediglich in einem einzigen Nukleotidpaar und beide vom Menschen nur in 4 Paaren. Organismen, die sich vor Milliarden von Jahren voneinander separiert haben, wie etwa die Eubakterien, Chloroplasten und Archaebakterien, erscheinen auf der Ebene der t-RNAs noch als «nahe Verwandte», nämlich mit nur geringfügig abgeänderten Sequenzen.

Damit wird eine Rekonstitution der Urstruktur in den Bereich des Möglichen gerückt. Vergleicht man die Sequenzen verschiedener t-RNAs, zum Beispiel für Coli-Bakterien, für Hefezellen, oder für Archaebakterien, so zeigen alle einen sehr hohen GC-Gehalt, der bei einer Überlagerung der Sequenzen noch zunimmt. Andererseits ist aus den mitochondrialen Sequenzen, die alle sehr AU-reich sind, zu ersehen, dass G und C im Verlaufe der Evolution – wohl aufgrund eines grossen Angebots des Stoffwechselprodukts A in den Mitochondrien – ersetzbar und nicht aus Gründen struktureller Stabilität erforderlich ist. Darüber hinaus deuten die rekonstituierten Vorläufersequenzen eine periodische Triplettstruktur an (siehe Abbildung 7), die auf einen Ur-Code GNC hinweist, in dem N jeweils eins der vier Nukleotide A, U, G, C symbolisiert. t-RNA war offenbar nicht nur der Ur-Adaptor, sondern fungierte auch als genetischer Informationsträger, eine Eigenschaft, die im Verlaufe der Evolution verloren gegangen ist. Die vergleichende Sequenzanalyse der t-RNAs hat – zusammengefasst – zu folgenden Aussagen geführt:

- t-RNA ist ein «alter» Adaptor, der sich im Verlaufe der Phylogenese relativ geringfügig verändert hat.
- Verschiedene t-RNA Moleküle einer Spezies erscheinen als Mutanten einer Stammsequenz.

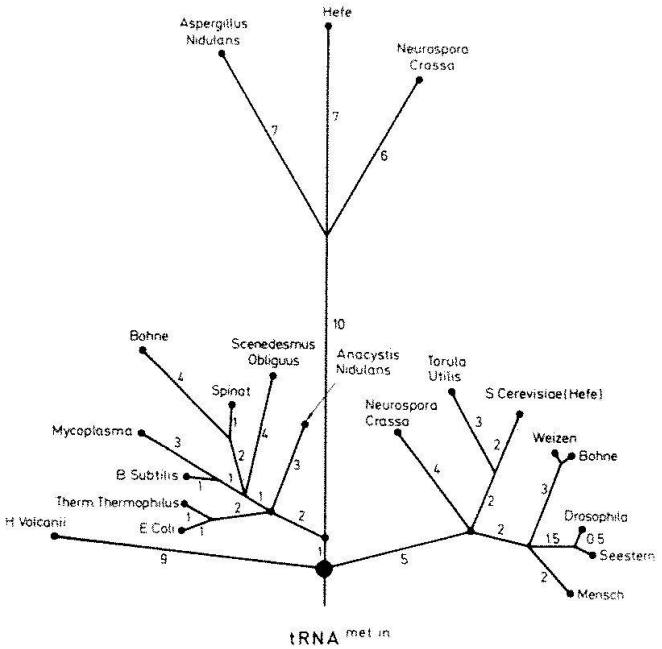



*Abb. 6.* Der Stammbaum einer Transfer-RNA, hier derjenigen, die beim Start der Translation eine Rolle spielt, lässt auch nach Milliarden von Jahren nur wenige Veränderungen in der Nukleotidsequenz (Zahlen an den Astabschnitten) erkennen. Bei allen bislang untersuchten Säugetieren ist die Sequenz nahezu dieselbe. Die Quotienten geben das Verhältnis von Guanin plus Cytosin zu Adenin plus Uracil an. In der Nähe der frühesten Verzweigungen ist es am grössten und an den Enden der langen Äste am kleinsten. Mit einem Wert von etwa 1:2 in den Mitochondrien (den «Kraftwerken» der Zelle) hat es sich gegenüber einem Wert von 2:1 nahe den frühesten Verzweigungen praktisch umgekehrt. Der Stammbaum lässt vier Gruppen deutlich hervortreten: Archaebakterien (nur ein Vetreter: H. Volcanii), Eubakterien und Blaualgen, die sich kaum von den Chloroplasten unterscheiden, Eukaryonten sowie Mitochondrien. Der grosse Abstand der Mitochondrien ist durch eine starke Substitution von G und C durch A und U gekennzeichnet. In der Purin-Pyrimidinabfolge weisen sich die Mitochondrien als nähere Verwandte der Eubakterien aus, während ihr Abstand zu den Halobakterien und den Eukaryonten relativ gross bleibt.

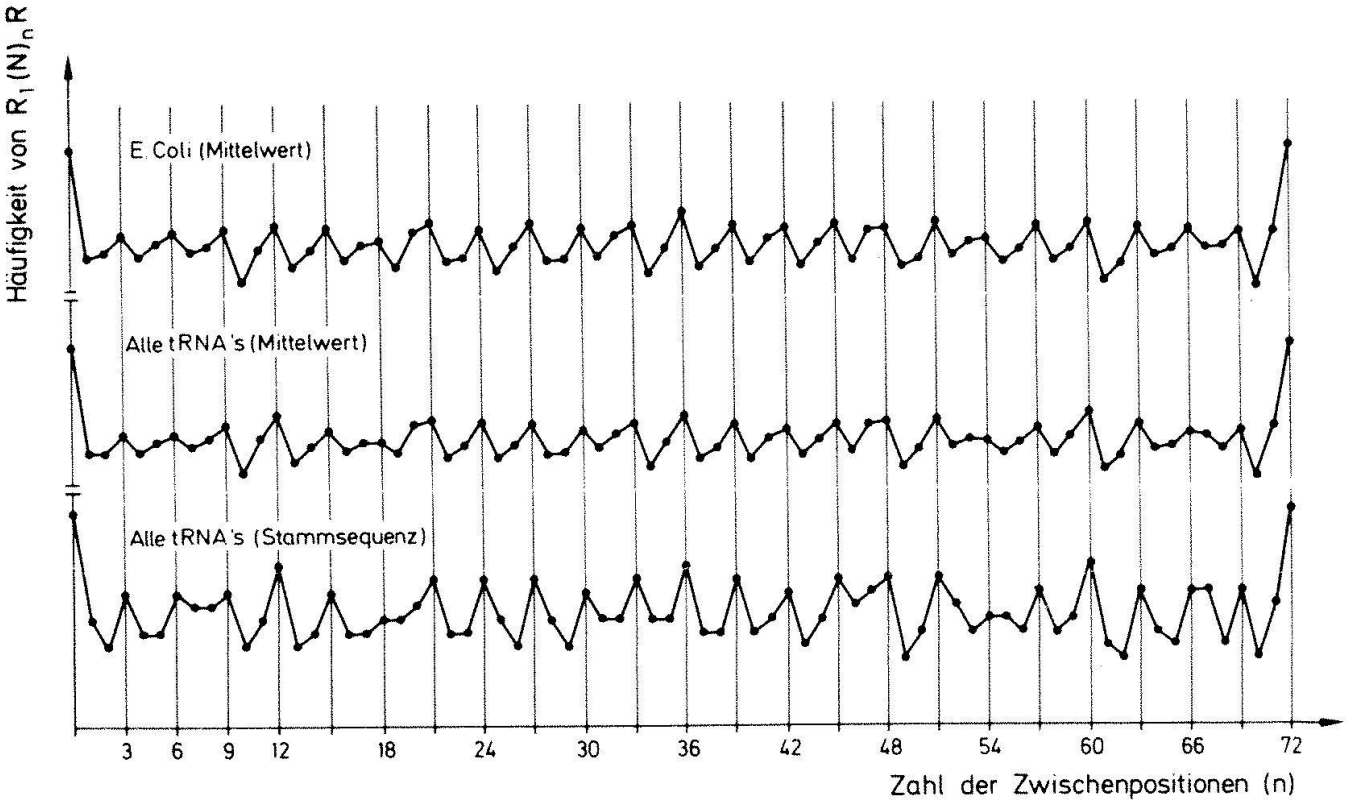



*Abb. 7.* Korrelationsanalyse der Repetition von Purin in t-RNA. Man unterteilt eine t-RNA Sequenz in Tripletts, beginnend am 5'-Ende und im Raster mit dem Anticodon, und zählt, wie oft ein Purin (R) in erster Position des Tripletts nach n Zwischenpositionen von einem Purin (R) gefolgt wird. Die deutlich sichtbare Dreierperiode weist auf eine Triplettstruktur der Form RNY hin. Die Kurven repräsentieren die Mittelwerte für die Sequenzen von E.coli bzw. von allen bisher untersuchten tRNAs im Vergleich zu der Sequenz, die sich aus einer Überlagerung aller tRNAs ergibt. Die Tatsache, dass die Korrelation in der «Überlagerungssequenz» deutlicher ist, zeigt an, dass es sich um eine «Erinnerung» an die Frühzeit der Evolution handeln könnte.

- Die ursprüngliche Stammsequenz ist zum grössten Teil rekonstruierbar.
- Sie zeichnet sich durch hohen GC-Gehalt sowie durch ein RNY-Code-Muster aus.
- Alle Befunde sind mit einem Ur-Code GNC für die in der Natur mit grösster Häufigkeit auftretenden natürlichen Aminosäuren: gly, ala, asp und val kompatibel.

## Synopsis

Wir finden eine Kongruenz von Theorie, Modellexperiment und historischem Zeugnis. Doch sollten desungeachtet die folgenden Punkte berücksichtigt werden:
Die Evolutionstheorie – wie jede andere physikalische Theorie – beschreibt lediglich ein «Wenn-dann-Verhalten». Unter der Voraussetzung, dass die Theorie frei von Fehlern ist, zeigt sie, was aus gegebenen Anfangsbedingungen zwangsläufig folgen muss.

Der Wert der Theorie ist allein an der Möglichkeit ihrer experimentellen Überprüfung zu messen. Modellexperimente liefern quantitative «Eichwerte», mit deren Hilfe sich Wahrscheinlichkeiten für die Entstehung der molekularen Vorstufen des Lebens abschätzen lassen.

Weder Theorie noch Modellexperiment sagen etwas über den tatsächlichen historischen Ablauf der Evolution aus. Dazu bedarf es spezieller historischer Zeugnisse.

Kongruenz von Theorie, Modellexperiment und historischem Zeugnis ermöglicht uns, das Prinzip «Leben» als eine Regelmässigkeit der Natur zu begreifen.

Auf der Grundlage der erkannten Prinzipien wird der Evolutionsprozess im Laboratorium nachvollziehbar.

## Literatur


Spiegelman, S., et al. Proc. Nat. Acad. Sci. USA, 50 (1963) 905; 54 (1965) 579, 919; 60 (1968) 866; 63 (1969) 805.

Mills, D.R., Kramer, F.R., Spiegelman, S., Science 180 (1973) 916.

Kramer, F.R., Mills, D.R., Proc. Nat. Acad. Sci. USA 75 (1978) 5334.

Sumper, M., Luce, R. Proc. Nat. Acad. Sci. USA, 72 (1975) 162.

Biebricher, Ch.K., Eigen, M., Luce, R., J. Mol. Biol. 148 (1981) 369, 391.

Biebricher, Ch.K., Eigen, M., Gardiner, W.C. Jr., to be published.

Mills, D.R., Peterson, R.I., Spiegelman, S., Proc. Nat. Acad. Sci. USA 581 (1967) 217.

Eigen, M., Naturwiss. 58 (1971) 465.

Eigen, M., Schuster, P., Naturwiss. 64 (1977) 541. 65 (1978) 7, 341.

Domingo, E., Flavell, R.A., Weissmann, Ch., Gene 1 (1976) 3.

Batschelet, E., Domingo, E., Weissmann, Ch. Gene 1 (1976) 27.

Weissmann, Ch., Feix, G., Slor, H., Cold Spring Harbor Symp. Quant. Biol. 33 (1968) 83.

Lohrmann, R., Bridson, P.K., Orgel, L.E., Science 208 (1980) 1464; J. Mol. Evol. 17 (1981) 303.

Bridson, P.K., Orgel, L.E., J. Mol. Biol. 144 (1980) 567.

Eigen, M., Winkler-Oswatitsch, R., Naturwiss. 68 (1981) 217, 282.



*Anschrift des Verfassers:*

Prof. Dr. Manfred Eigen
Max-Planck Institut für Biophysikalische Chemie
Abteilung für Biochemische Kinetik
D-3400 Göttingen (BRD)

# Origin of the Brain

David Hubel

When I was first invited to talk to you on some subject related to the Origin of the Brain I was appalled, because I am neither an expert in evolution, nor in neuroanatomy, nor in neurodevelopment. Although I do work to some extent in anatomy and in development, the work is confined to a small part of the mammalian brain: for much of the rest I am almost as ignorant as the astronomers in this audience.

But lest those who invited me appear too far mistaken in their choice, let me say that when it comes to the ultimate origin of the brain - the question of exactly what it came from, what forces molded it, and, in detail, how it got the way it is, I think all of us are in a profound state of ignorance. All I or any other neurobiologist can do is to make a few very general statements, which will be almost truisms; one can also point out what some of the major puzzles are.

In considering the origins of something like our own brain, one has three main ways of going about the problem. One can approach it head-on and examine the fossil record, to ask what the brains of our ancestors looked like. When it comes to the nervous system all fossils can do is indicate the bony housing of the brain or spinal cord. While we all sense the far reaching nature of the deductions that paleontologists can make from a single bicuspid tooth about events that occurred over millions of years of history, most people will agree that the insides of skulls are rather limited in the kinds of things they can tell us about the development of an incredibly complex structure like the brain. It is as though some future archeologists were to try to trace the development of computers having only some fragments of the outer box containing them. Of course some valuable hints could be inferred just from the outer box, such as the spectacular decline in size of a computer, and hence the decreasing size of the elements of which it was composed. Perhaps the increasing occurrence of computers in the bedrooms of family dwellings will indicate, among other things, the spectacular decrease in their price.

The second tool for tracing the origin of the brain is comparison with brains of simpler contemporary species of animals. This assumes that there are plenty of animals around today that at least to some extent resemble our ancestors, an assumption that has all of evolutionary theory to back it up. It is an obvious tool, and a powerful one. Third, and finally, there is the well-known fact that the development of an animal to some extent recapitulates evolution - in the case of neurobiology, that the development of the brain in the embryo of an animal can give useful hints about the way in which that brain evolved.

Two hundred years ago one would have to add a fourth source of information, the Bible. The whole question of the origin of the brain is ignored in Genesis Chapter 1, which sadly for my status at this meeting surely does deal with the origin of almost everything else. All botany is taken care of in one or two verses, similarly for zoology; two verses are devoted to man, and curiously almost as many to women.

In order even to begin a discussion of brain origins I feel I should say a few words about nervous systems. It is relatively easy to draw a schematic diagram of a very general brain or nervous system (figure 1). This is to give the astronomers in the audience a rough idea of what kind of thing the brain is. Strictly speaking, the nervous system includes brain and any other aggregates of nerve cells, such as spinal cord and other clumps of nerve cells that are scattered here and there in the body. I will often use the word brain when I really mean 'nervous system'. In the very final analysis the brain can be regarded as

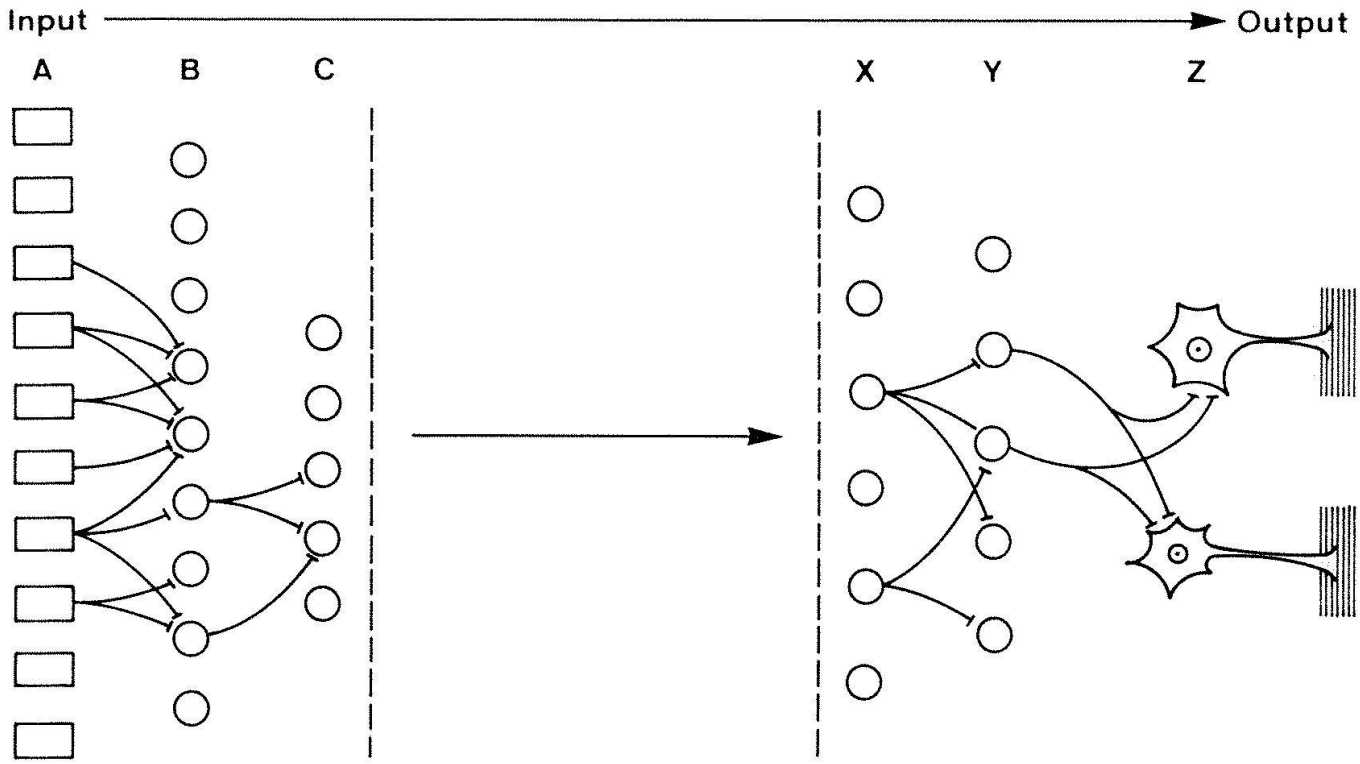



*Fig. 1.* Overall organization of the brain is indicated in a rough caricature that suggests the flow of information from the input of sensory signals by receptor cells (A) to the eventual output by motor neurons (Z) terminating on muscle cells. The outputs of receptors and neurons usually branch to send diverging signals to the next stage. Most neurons receive converging inputs, both excitatory and inhibitory, from earlier stages. Something is known about the significance of the connections near the input end of the brain (B, C) and near the output end (X, Y). Far less is known about the working of regions in between, which make up most of the brain.

forming a link between the outside world and an animal's muscle cells or gland cells. On the left of the diagram we have cells specialized to take in information from the outside world – information in the form of light (vision), mechanical energy (touch, hearing), and chemicals (smell, taste) – and on the right end, cells that can react in some way so as to allow us to manipulate the outside world. The brain, between these two, can be incredibly complex and many storied, or it can consist, as it does in the simplest animals, of one set of direct connections between input and output. In the most complicated of brains it is both at once. When we see a face, recognize it and smile or grimace, the circuit (path) involved in conveying the messages from the eyes to face muscles is so complex that beyond a certain stage we lose all trace of it. Between the rods and cones in the eye and the muscles of the face must be circuits that allow us to compare the face with others we remember and to recall previous experiences with that particular face, and circuits that have to do with emotion, delight or anger, and finally circuits that make just the appropriate sets of muscles contract together to produce the smile or scowl. No one has any clear idea how many sets of nerve cells are interposed between input and output. A wild guess would be 50–100, but almost certainly there is no single number, but many parallel pathways of varying lengths.

At the other extreme we have the simplest reflexes, very short paths with one or two connections. An example is the knee jerk, in which tapping the knee extensor tendon produces contraction of the thigh muscle. Here only two connections are involved, one in the spinal cord and one at the muscle itself.

The simplest animals have only rather direct paths between input and output. An example would be a worm which responds to poking it in the side by bending away, for example by contracting muscles on the opposite side. This takes a not entirely trivial set of connections going from the skin on the right side to the muscles that run fore-and-aft in the worm on the left, and of course a

similar set of connections linking skin on left with muscle on right.

Another primitive example would involve pressure on the inside of the intestine wall, produced by food, leading to constriction of circular muscle so as to propel the food along.

For circuits even as simple as these, animals need to have evolved a special bizarre kind of cell specialized for conveying information – the nerve cell, with its more or less round cell body and long cylindrical or tapering branches that convey electrical signals. Prior to the invention of nervous systems, long-range signalling in the organism had to be done chemically by an endocrine system. In the endocrine system a cell when stimulated releases a chemical that is discharged into the circulating fluid, into the blood stream, for example; when molecules of the chemical reach a sensitive target cell, they interact with some complex molecule on the surface of that cell's membrane, in a specific way, reminiscent of a key fitting a lock, and the union in some complicated way leads to the cell's responding appropriately, by changing shape, growing, releasing some other chemical, or some other change. A certain amount of specificity can be obtained in such a system by making use of dozens or hundreds of different chemicals and the same number of kinds of target receptor molecules. Such a system, in which a chemical is released by one set of cells, diffuses out through the entire animal and influences some matching set of target cells, is simply not refined enough to look after problems such as swimming or locomotion or feeding in a large multicellular animal. It lacks speed and specificity.

The nervous system was quite possibly derived from this endocrine system. We think so because nerve cells and endocrine cells are similar in many respects. At the point where two nerve cells communicate, called the synapse, an electrical signal arriving at the terminal of the first cell causes the release of a chemical, which diffuses out and in less than 1 msec arrives at the closely apposed membrane of the second cell. The result is to produce contraction if the second cell is a muscle, or the release of a second chemical, if the cell is a gland cell, or the production of a second 'postsynaptic' electrical nerve signal if the second cell is a nerve cell. What the endocrine system and the nervous systems have in common is the release from one cell of a chemical that interacts specifically with molecules of a second chemical on the recipient cell, the 'receptor molecules'. The first diffusing molecule, called a hormone in endocrinology and a neurotransmitter in the case of the nervous system, is usually small; several aminoacids are transmitters. The recipient molecule is large and complex, generally a protein. The two systems also have in common an elaborate chemical machinery for release of the first substance, and an elaborate machinery in which the change brought about by the combination or the union of the first substance with the receptor molecule is finally translated into a response – a nerve impulse, contraction, or whatever. Some of the same compounds, such as adrenalin, may occur in the two systems, as the endocrine in the one and the transmitter substance in the second. Many of the details of the release and translation processes are identical in the endocrine and nervous systems and involve substances like cyclic AMP and calcium ions, and multiple phosphorylation steps. The similarity of the two systems is a powerful reason for concluding that one was derived from the other. The advantages of the nervous system are, of course, speed and the possibility of extreme complexity and specificity.

I now want to make a completely different kind of statement about the brain, a statement that would be obvious and a truism to any zoologist in the audience, but which for a non-zoologist is probably worth enunciating. The brain of any animal is not so much a general device for computation as it is a set of devices to solve or take care of a large number of very special problems, often very mundane ones that may give us no very great aethetic satisfaction to contemplate. Take the example of vomiting. We eat a poison, say a liter of schnaps. Now just imagine the intricacy of what happens and the intricate specific circuits that must be called into play in a particular part of our brain, a subdivision called the medulla oblongata. The stomach wall is irritated; signals pass along one set of nerves to the brain; as a result the stomach contracts, the

diaphragm and abdominal muscles contract, the pharynx opens, the epiglotis closes (so bile and alcohol don't enter our lungs and give us pneumonia), respiration stops, we sweat. We are likely to feel some emotion, probably not elation. All very elaborate, and not something we learn from our parents or at school. The circuit is there at birth and gets there because of genetic instructions. We have hundreds of such circuits: coughing, sneezing, laughing, crying, blinking, defecating, urinating, copulating, sleeping and moving around while we sleep, walking, running, pulling away from something painful, turning head and eyes when we see or hear something new or interesting, following objects with our gaze, grasping objects placed in the hand, holding and keeping our bodies erect, phonating, yelling, singing, speaking – the list goes on and on, and is of course different for different animals – we don't warble like birds though some of us may yodel, we don't preen our fur or swim like leeches. The point is that we can never understand the specifics of how brains are connected without a knowledge of the organism's behavioral repertoires. Our various little rituals are very particular and mundane, but we have to think about them if we are to understand why our brain is hooked up the way it is. Mathematicians and physicists love generality and all-inclusive ideas, and I think that that is why they sometimes don't get very far in thinking about the brain, an organ that evolved to start with as something to take care of a host of small and special, but for survival very important problems.

Of course, I hardly need say that most of the more complicated problems arise in the part of the body where the sensory organs are placed, where food is taken in, and for most animals except man, the part that leads during locomotion, namely the front end or head. That is the part that requires the most control machinery and that is why the front part of the nervous system becomes larger than any other part, and is consequently dignified by a special term: Brain. Now I have already consumed 25 of my 50 minutes and have barely introduced the subject. From here on I have to be selective and will just touch on a few very diverse questions relating to how our brain got there and why it is designed as it is.

One thing we might wish to consider is the degree to which brains are similar or dissimilar in different animal species. In one respect the nervous system is similar to most other systems in zoology: the brains of invertebrates all have much in common, and are wildly different from brains of vertebrates, all of which likewise have very much, even more, in common. Invertebrate brains tend to be distributed in the form of aggregates of relatively small numbers of cells, called ganglia. A worm or leech has dozens of ganglia, each containing hundreds of cells (not millions), one for each body segment and a larger one for the head. The numbers of cells involved may be small by vertebrate standards but the number of branches or processes of each cell, and their intricacy of interconnections, may be formidable. Invertebrate brains are not at all simple. Invertebrate ganglia generally have one property that our brains utterly lack, something we may term 'cell identifiability'. In a particular lobster abdominal ganglion we can speak of cell No.321 in two different animals, and know we are talking about the same cell, with almost identical connections and functions; for example, a cell whose firing leads to extending the terminal part of a walking leg of the lobster. In a vertebrate one can almost never speak of a certain particular cell, the way one can of a tooth or a limb or a bone, any more than one can speak of a particular hair on someone's head. For doing research on brains identifiability can be very useful.

Sydney Brenner some years ago began a project whose aim was to take a relatively simple invertebrate, a worm less than 1 mm long and having only the order of 1000 nerve cells, all identifiable (in the technical sense mentioned above) and work out, using the electron microscope, a wiring diagram of all the connections. This is feasible, but it took many years just to get 10% of the way through this humble animal. My point is that 'simplicity' is relative: even the simplest of invertebrates can be incredibly complex. An animal as far evolved as a fly is not simple by any standards.

When most of us think of a brain we of course think of our own, or ours plus our cat's, and not that of a tapeworm or a fly. In the case of vertebrates – such as fish, birds,

reptiles, amphibia, mammals, the thing I have always found most unexpected is the similarity of design of the nervous system. Most of the basic parts, or substructures or subdivisions, such as cerebellum, spinal cord, medulla oblongata, occur in all vertebrates; and nothing like any of these occurs in any invertebrate. Vertebrate brains differ in details and emphasis, but all have the same basic plan. The origin of brains like ours is thus tied up very much with the origin of vertebrates from some invertebrate ancestor – a problem on which I can speak with absolutely no knowledge or authority, except to give my vague impression that evolutionists really don't know very much about the origins of vertebrates. It does seem likely that our invertebrate ancestor was much more like a worm than like a clam, fly or lobster. It was what the author of Genesis would have called a 'creeping thing'.

The subject, the origin of the brain, is closely related to the problem of how any given brain originates – the problem of the development of the nervous system. This is a branch of embryology, a vast subject in which much is known, and much still unknown. Here I would like to discuss two aspects of neurodevelopment that particularly interest me.

The first is the distinction between the development of the brain before birth (here I am thinking of mammals) and the development after birth. Prior to birth most or all development must obviously be dictated, ultimately, by genetic information. After birth, the development of connections in the brain obviously has the possibility of being influenced both by genetic and by environmental process, and this is what constitutes learning, in the very broadest sense. The brain of higher organisms, to develop in a normal way to maturity, must be in a normal environment postnatally during a period of flexibility usually termed the 'critical period'. One good example of a critical period occurs in the visual system. If an adult of 65 years develops cataracts, he may become blind for some years, but on removal of the cataracts he will almost invariably see perfectly well provided glasses are fitted to replace the lens of the eye. In contrast, if a newborn baby has cataracts that are removed only say at age 5, vision will at first be totally absent. It develops only slowly and imperfectly. The cause of this blindness is not in the eye, but in the visual part of the cerebral cortex, to which the eye connects. Whether the disuse leads to a withering of connections that were there at birth, or prevents a flowering of connections whose specific details depend on experience, is the question of nature vs. nurture, at present a vexed subject. Most modern neurophysiologists would probably agree that the amount of precoded information in the brain has been seriously underestimated by psychologists of the past century. On the other hand we all agree that the proportion of information contained in the mature brain of a human that got there by postnatal learning as opposed to prewiring is far higher than in any other animal. This may be somehow related to the greatly increased size of the cerebral cortex of man, a subject that I will come back to. It is probably related to the long period in infancy during which a human is completely dependent on its parent. 'Critical periods' tend to be very long in humans, up to 5–10 years in some systems, as opposed to a few months in a kitten and one year or less in a monkey.

No one can look at the hydraulic engineering accomplishments of a family of beavers without being greatly impressed. I doubt that anyone would seriously try to maintain that their abilities came about through attending lectures at 10 o'clock every morning. Just how the civil engineer accomplishes the same thing – and much more – by attending lectures, is the subject of memory and learning. Today we know very little about the mechanisms of learning. Some parts of the brain, such as the temporal parietal and frontal lobes of the cortex, doubtless play a more important role than others (the spinal cord; retinas, occipital lobes). Learning almost certainly involves changes in efficacy of transfer of information across synapses. It involves a fast transient component and a slower consolidation component. But beyond these things very little is known, especially about the changes that occur at the synapse, or how they occur.

The second subject that interests most neurobiologists concerns the origin of specific neural connections. There are countless examples in the brain in which two gigantic structures, each with millions of nerve cells,

are connected by a cable, one-way or two-way, that interconnects the cells in an incredibly specific manner so that any tiny region of one becomes closely associated with some tiny region of the other. A vivid example is the optic nerve, a bundle of just over 1 million nerve fibers that link the retina with the brain, specifically with two grape-size nests of cells deep in the brain called the lateral geniculate nuclei, and the optic radiations, which connect the lateral geniculate nuclei to cerebral cortex. The optic nerves in mammals are one-way, eye-to-brain; the radiations are two-way with some fibers carrying messages from geniculates to cortex and others carrying messages back. Each cable contains in the order of a million nerve fibers.

When the brain develops each of the structures (retina, geniculate cortex) develops by itself up to some point; then fibers grow out of the cells, bundle together and proceed to their targets, a distance of many centimeters. They do so with all the precision that exists in the fully developed brain. How each fiber knows exactly where to go is one of the great unsolved problems of neurodevelopment. Several very different kinds of theories exist, each with fairly strong experimental support, but none of them, given the present state of the subject, seems fully compelling. Everyone at least agrees that the brain wires itself up, and does so with an almost unbelievable precision. The question is a fascinating one, even if at the moment the field is rather a mess.

A curious feature of many of these fiber bundles concerns the frequency with which they are crossed – that is, they originate on one side of the brain and end up on the other. For example, as almost everyone surely knows, the left side of our cerebral cortex governs movement of the right face, arm and leg; things happening to the left of where we are looking produce a reaction in both of our eyes, but in our right lateral geniculate bodies and occipital lobes. Sounds coming from one side and arriving at our two ears at slightly different times and at different intensities, by an elaborate process of interaction have their main effects on the opposite side of the brain. The rule is not invariable: the left side of the cerebellum governs movement of the left side of the body. Nevertheless it turns out, reasonably, that the left side of the cerebellum is closely connected to the right side of the cerebral cortex, by a massive cable that crosses over from the left side to the right. But why all this crossing? No one would have imagined a priori that the left brain governs the right body, to the extent that it does. It is astonishing, and seems quite unnecessary, irrational, uncalled for, even silly. It does exist, however, and the truth is that no one has any good idea why. It is easy to predict the existences of certain crossed paths in any nervous system – consider the examples given earlier of a poked worm arcing away from the stimulus. Something like that rudimentary reflex has been advanced as explaining all the subsequently evolved crossing fibers in the nervous system. Most people find the idea possible, but not very compelling.

Now finally I want to turn to the cerebral cortex, a structure that exists except in rudimentary form only in mammals, and which in going from the lowly marsupial to man expands in a more spectacular way than any other part of the brain. The cortex happens to be the part of the brain I work on, and virtually the only part that I can talk about from first-hand experience.

Francis Crick, and for all I know others before him, has made the general statement that aggregations of cells in the nervous system can be divided into two main categories, plates and globs. There are many examples of plate-like structures, but the cortex of the cerebrum happens to be the most imposing. Spread out, the cerebral cortex occupies an area of about ½ square meter; its thickness is 2 mm. It is packed with cells, some 100,000 under each $mm^2$ of surface, and the cells are aggregated into six or so layers, which differ in the kinds of neurons that each contains, their form, packing densities, and their connections.

The ½ $m^2$ sheet is itself subdivided probably into a hundred or so regions, whose functions and connections differ profoundly from one to the next. These include about 8 or 10 separate visual areas and about as many auditory, motor, sensory (touch) and speech areas, and certainly a host of others whose functions are known vaguely or not at all. Within any one of these areas the machinery seems to be relatively constant: one millimeter is about the same as the next, and carries

out the same tasks. The best known of these areas have topographic maps – a term I have touched on earlier. In our visual system one part of the visual world (say, up and to the left of where we are looking) connects to and activates one part of the visual cortex, another part of the visual world activates another. Information reaching any one small part of the cortex is disseminated across the cortex by neuronal connections over a distance of a few mm at most. Thus in the case of vision, signals related to what is happening in one part of the visual environment arrive at the visual cortex, undergo some kind of analysis or transformation, and leave, with no interaction with what takes place in remote parts of the visual field. I speak here just of the primary visual cortex, a region about 2000 $mm^2$ in extent, at present the best understood region of the cortex. Here we know quite well what happens to the information between its arrival and departure. To discuss it would take all of a second hour, or a separate paper.

That, at any rate, is the kind of thing the cortex is, and does. It represents things (visual, auditory, ideational, and who knows what else?) in two-dimensions, and it makes local operations on information – to use a rather unhappy jargon.

Once invented, in the course of evolution, the cerebral cortex obviously caught fire and took off, reaching its peak in primates and, in particular, in man. For a structure that is so similar throughout its extent, it serves an astonishingly large variety of functions; as an all-purpose device it outdoes any computer and certainly anything else in the known universe.

There are great differences in the levels of our understanding of different areas of the cortex. The primary visual cortex is the only area where there exists a known difference between the information entering and leaving. For about a dozen other areas something is known descriptively about how cells behave, and beyond that our knowledge varies from vague to zero. For example, certain areas in the left hemisphere of man are well known to be involved in language, including speech, comprehension, and reading. But exactly what goes on in these regions is not understood at all. For many other areas, the functions are not known even in any global sense.

Perhaps one should ask in concluding what the future limits may be in the evolution of the brain, or if you wish, of the cortex. How far can the development go? How big, how much more complex, can it get? In the first place, to me it is far from clear what, if any, the relationship is between brain size and intelligence. Among animals of the cat family (lions, tigers, cats and so on) generally the bigger the animal, the bigger its brain. A lion's brain looks just like a magnified cat brain. It hardly seems likely that intelligence is related in any way to animal size. A lion may be smarter or more agile or better in some other way than a cat, but I know of no evidence for this. The same thing applies to members of the primate family. Perhaps bigger animals have bigger brains simply because there is room; it's not clear to me what good it does them. Of course this law breaks down completely when we compare brains of different families of animals. In general, primates tend to have bigger brains than brains of carnivores, and here there is an obvious correlation with intelligence. Whales have enormous brains, about the size of the human brain, and whales are undoubtedly very intelligent.

One anatomist friend of mine has suggested that an important limitation to the size of our brain may simply be the size of the female birth canal. Certainly in any normal delivery the head is what offers the most difficulty. Human head and brain grow considerably after birth. Thus the fact that humans, compared to other mammals, are so immature and helpless at birth may be partly explained by the problem of delivering a head of ever increasing size.

In the past few months I have had (after resisting it for years) to learn how to work a computer, and have spent more time than I like to admit writing programs. I have noticed that as one works on a program, adding to it, trying to debug it and refine it, there comes a time when the whole thing gets so cumbersome and unwieldly that it has to be torn down and built up again, often with a quite different basic design. One can't help wondering if similar problems do not arise in brain evolution. How many structures of a seemingly bizarre form owe their

presence to some blind alley in evolution that has had to be abandoned? There are some suggestions that this happens. The nerve bundle that supplies our face runs in a path in the brain, before leaving it, that displays a most extraordinary loop whose presence is easily understood in terms of development. A rather large structure in our brain concerned with movement called the caudate nucleus (it degenerates in Parkinson's Disease) is rather glob-like and compact in lower animals. In man, because of the way the brain enlarges late in development, backwards, downwards and forwards, this poor structure becomes deformed almost to the shape of a pretzel. Such aberrations are gross ones, and hence easy to detect; we have no idea at all whether the detailed circuits involving information processing get into similar contortions.

There is one curious kind of apparent constraint in brain size and behavioral repertoire. The spider monkey has an extremely dexterous prehensile tail, whose ventral surface lacks hair and possesses ridges closely resembling fingerprints. The first time Torsten Wiesel and I experimented with a spider monkey, we caught it and I held the legs, Torsten held the arms and he began to inject the anesthetic. Out of nowhere came the tail, seizing the syringe out of Torsten's grasp and waving it on high, wrapped around it like a boa. The spider monkey also has feet that look like hands, with an apposable thumb that is far more dexterous than our big toe. But it pays for these assets. When you look at the hand, you see no thumb at all, just a mound-like prominence. Either there is only enough sensory and motor cortex for some limited number of agile parts - hands, feet, tail, trunk or beak - or else the limit is in something like the available attention span or the ability to coordinate these cortical areas.

Motor and sensory cortical areas of man are probably relatively old, in an evolutionary sense, and well entrenched. By comparison the speech or language areas are much more recently developed. Of these speech areas I find particularly interesting the parts (or part) concerned with reading. Reading, writing, understanding speech and producing it are functions that are at least to some extent dissociable, or so we are told by neurologists who study deficits after localized brain damage. The ability to read must surely be quite recent - probably even today not nearly every human learns to do it, though those who do learn may spend a lot of time at it. For those who read, what part of the brain, if any, becomes devoted to the task, and what does that region do in people who don't read at all? The question is a little bit related to the problem of whether someone like Mozart, who possesses some extraordinary abilities, pays some price by having some glaring defects.

To the entire question of where the brain came from or is going, and a host of more or less related questions, there are no very clear answers, as the past hour has probably convinced you. At most I can hope that the range of questions is greater than you might have supposed. As we come to understand the brain better, we will doubtless come to ask, and fail to answer, an even greater variety of questions about its origin.

*Address of the author:*

Prof. Dr. David Hubel
Department of Neurobiology
Harvard Medical School
Boston, MA 02115 (USA)

# Ursprung, Grenzen und Zukunft der Naturwissenschaft

Gunther S. Stent

Bevor ich mich meinem Thema von Ursprung, Grenzen und Zukunft der Naturwissenschaft zuwende, möchte ich kurz erklären worüber ich nicht sprechen werde. Ich beabsichtige nicht, die wirtschaftlichen Grenzen der Naturwissenschaft zu erwägen, die dadurch entstehen, dass die Forschung immer teurer wird. Noch habe ich vor, die rein physikalischen Grenzen zu besprechen, die unserer Kenntnis des Weltalls oder der Materie prinzipielle Schranken setzen, wie die Lichtgeschwindigkeit oder die praktisch maximal mögliche Beschleunigung von Elementarteilchen. Auch werde ich nicht die sozialen und politischen Grenzen erwähnen, die durch das nicht vollkommen unbegründete Aufblühen von Anti-Szientismus in technologisch hoch entwickelten Ländern heute offensichtlich sind. Statt dieser praktischen Grenzen möchte ich drei kognitive Grenzen der Naturwissenschaft erörtern, die in diesem Jahrhundert zu Tage gefördert wurden: eine semantische Grenze, eine strukturelle Grenze, und eine subjektive Grenze. Da das Bestehen dieser Grenzen dem Fortschritt der Forschung Schranken setzt, ist es wahrscheinlich, dass die Naturwissenschaft der Zukunft anders sein wird als die Naturwissenschaft der Vergangenheit.

## Naturwissenschaft und Wahrheit

Zu Anfang meiner Ausführungen möchte ich die für mein Thema grundlegenden Begriffe «Naturwissenschaft» und «Wahrheit» erläutern. Eine gründliche Erläuterung dieser Begriffe wäre jedoch schon allein ein stundenfüllendes Programm, und daher werde ich im Rahmen dieses Symposiums über den Ursprung der Dinge ganz einfach behaupten, dass die Naturwissenschaft eine Tätigkeit ist, die das Ziel hat, den Zusammenhang der Dinge darzustellen. Mit dieser Tätigkeit versuchen wir, die Welt zu verstehen und Macht über ihre Dinge auszuüben. Selbstverständlich gibt es auch andere Tätigkeiten, die eine der Naturwissenschaft ganz ähnliche Rolle spielen, wie zum Beispiel die Religion oder die Magie. Diese werden zwar gewöhnlich als von der Naturwissenschaft grundsätzlich verschieden angesehen. Aber das Errichten einer scharfen Demarkationslinie zwischen Naturwissenschaft und Nicht-Naturwissenschaft ist ebenfalls ein philosophisch schwieriges Problem, mit dem ich hier nicht ringen möchte. Ich werde also davon absehen zu klären, in welchem Sinne die Astronomie sehr wohl und die Astrologie mitnichten eine Naturwissenschaft ist. Unerlässlich ist jedoch der Hinweis, dass die Naturwissenschaft eine semantische Tätigkeit ist, insofern, als es ihr Zweck ist, Darstellungen mit einem sinnvollen Inhalt mitzuteilen.

Was bedeutet es zu sagen, dass ein dargestellter Zusammenhang der Dinge, also eine naturwissenschaftliche Theorie, wahr ist? Hier verkürze ich abermals die eigentlich notwendige stundenlange Diskussion und behaupte einfach, dass eine Darstellung für mich wahr ist, sofern sie mit meinem Weltbild im Einklang ist und meine Zustimmung gebietet. Diese Auslegung des Wahrheitsbegriffes ist offensichtlich keine objektive, sondern eine subjektive. Sie führt zum Begriff der objektiven Wahrheit nur dadurch, dass ich davon überzeugt bin, dass die für mich wahre Darstellung auch die Zustimmung jeder anderen, für dieses Urteil qualifizierten Person gebieten würde. Das Ideal der absolut objektiven Wahrheit wird auf diesem Weg nur erreicht, wenn auch Gott der Darstellung zustimmt.

Wie kann eine naturwissenschaftliche Theorie Zustimmung gebieten? Vor allem, indem sie eine einleuchtende Antwort auf eine «warum?» Frage über den Zusammenhang

der Dinge liefert, und weiter, indem durch diese Theorie vorausgesagte Ereignisse tatsächlich eintreffen und sie bestätigen. Auch kann die Antwort auf eine theoretische «warum?» Frage manchmal zu einer Antwort auf eine praktische «wie?» Frage über die Dinge führen. Die uns dadurch verfügbare Macht kann als eine weitere Bestätigung der Theorie gelten. Jedoch ist auch die Frage der Verifizierung naturwissenschaftlicher Theorien unter Philosophen sehr umstritten. So wird beispielsweise oft argumentiert, dass die Wahrheit einer Theorie überhaupt nicht durch eine endliche Anzahl von Beobachtungen zu beweisen ist. Dessen ungeachtet muss ein praktizierender Wissenschaftler immerhin daran glauben, dass er sich dem Beweis der Wahrheit einer naturwissenschaftlichen Theorie wenigstens annähern kann. Denn wie könnte er sonst sein Leben damit zubringen, den Zusammenhang der Dinge zu verstehen, wenn er nicht überzeugt wäre, dass seine Anstrengungen beweisbare Wahrheiten hervorbringen?

**Intuitive Begriffe**

Nachdem ich Ihnen den Naturwissenschaftler als einen Darsteller des Zusammenhangs der Dinge vorgestellt habe, möchte ich mich nun den kognitiven Grundlagen der Naturwissenschaft zuwenden, der Frage also, wie der menschliche Verstand überhaupt zu einer Darstellung der Welt gelangen kann. Die Empiristen des späten siebzehnten und frühen achtzehnten Jahrhunderts behaupteten, dass der Verstand bei Geburt eine leere Tafel ist, auf der sich mehrende Erfahrungen allmählich ein Bild der Welt skizzieren. Dieses Bild ist geordnet oder strukturiert, weil wir Regelmässigkeiten in unseren Erfahrungen mit Hilfe des Prinzips der logischen Induktion erkennen und Ereignisse, die wiederholt zusammen auftreten, kausal verbinden. Jedoch wusste schon David Hume, einer der hervorragendsten Vertreter des Empirismus, dass diese Auffassung an einem logischen Fehler Schiffbruch erleidet. Denn die Gültigkeit der Induktion, die ja die Grundlage unserer kausalen Verbindung beobachteter Ereignisse sein soll, kann weder logisch noch durch Erfahrung bewiesen werden. Das Vertrauen auf die Induktion und der Glaube an den kausalen Zusammenhang zwischen Ereignissen entsteht intuitiv und kann nicht aus der Erfahrung gefolgert werden.

Wenige Jahre nach Hume zeigte Immanuel Kant, dass die empiristische Lehre auf einer unzureichenden Einsicht in das Wesen der Vernunft beruht. Kant wies darauf hin, dass Sinneswahrnehmungen überhaupt nur dann Erfahrung werden, das heisst, sinnvoll werden, wenn sie mittels apriorischer intuitiver Begriffe wie Zeit, Raum und Objekt gedeutet werden. Andere intuitive Begriffe wie der kausale Zusammenhang von Ereignissen erlauben uns den Aufbau aus der Erfahrung von einem Bild der Welt der Dinge. Und mit diesem Weltbild ist auch der Begriff der Wahrheit verbunden, mein intuitiver Glaube also, dass die Dinge tatsächlich so sind wie ich sie mir eingebildet habe.

Aber wie ist es möglich, dass, wenn wir Begriffe a priori zu unseren Sinneswahrnehmungen bringen, jene Begriffe so ausgezeichnet auf die Dinge der Welt passen? Angesichts der Unzahl unpassender Begriffe, die wir uns vor jeglicher Erfahrung ausdenken könnten, scheint es schlicht ein Wunder zu sein, dass unsere intuitiven Begriffe ausgerechnet die Passenden sind. Konrad Lorenz erkannte vor rund 40 Jahren, dass dieser scheinbar wundersame Umstand leicht von der Evolutionslehre erklärt werden kann. Unser Gehirn ist ja das Ergebnis der natürlichen Selektion unserer fernen Urahnen und kann daher selbstverständlich auch vererbtes, das heisst angeborenes, Wissen über die Welt besitzen, vor jeder persönlichen Erfahrung. Oder, wie Lorenz schrieb, «das Passen des Apriorischen auf die reale Welt ist ebensowenig aus der ‹Erfahrung› entstanden wie das Passen der Fischflosse auf die Eigenschaften des Wassers».

Obwohl die apriorischen Begriffe Kants daher als Teil unserer biologischen Ausstattung anzusehen sind, sind sie doch nicht in dem Sinn «angeboren», dass sie bereits bei der Geburt im Gehirn gegenwärtig sind. Sie entstehen erst, und das ist der grosse Beitrag Jean Piagets, im Laufe einer kognitiven Entwicklung währen der Kindheit. Piaget fand, dass diese Entwicklung eine Folge von klar erkennbaren Stufen durchläuft und durch Wechselwirkungen des Kindes mit seiner Umwelt gesteuert wird. So schreibt das Kleinkind den Dingen in seiner Umgebung

zuerst noch keine konstante Grösse und keine Identität zu. Der Begriff eines Objekts, das Identität und charakteristische Eigenschaften hat, tritt erst in einer späteren Stufe auf. Aus solchen konkreten Begriffen entwikkeln sich später die abstrakten sprachlichen, logischen und mathematischen Denkweisen. Zum Beispiel fand Piaget, dass das Kind erst den Gedanken der Invarianz entwickeln muss, ehe es Wörter gebrauchen kann, die sich auf bestimmte Objekte beziehen oder ehe es Zugang zum Begriff der Zahl haben kann. Die abstrakten Kantschen Begriffe von Raum und Zeit erscheinen erst in einer noch späteren Stufe in ihrer reifen Form.

Für meine Ausführungen liegt die Bedeutung von Piagets Entdeckungen in der Einsicht, dass unsere intuitiven Begriffe während der Kindheit jeder normalen Person als Resultat einer genetisch bestimmten Dialektik zwischen dem sich entwickelnden Gehirn und der Kindeswelt entstehen. Daher stellen diese Begriffe eine biologische Tatsache dar, und nicht kontingente oder zufällige Produkte sozialer oder philosophischer Konventionen. Diese Begriffe sind also immanente Eigenschaften menschlicher Vernunft und sie zu erwerben bedeutet, als geistig gesunde Person aufzuwachsen. Aus dieser Einsicht folgt, dass der Homo sapiens schon bei seinem Auftreten vor einigen hunderttausend Jahren das zum Naturwissenschaftler notwendige geistige Rüstzeug besass. Aber erst vor etwa zehntausend Jahren gelang die Zucht von Haustieren und Erntepflanzen, wurden Metallurgie, Töpferei und Ziegelei erfunden, die ersten menschlichen Leistungen also, die wir als mit der Naturwissenschaft verwandt anerkennen können.

Die eigentliche Naturwissenschaft begann erst vor ungefähr zweieinhalbtausend Jahren, als die Griechen auf den Gedanken kamen, dass die Natur von einer begrenzten Anzahl von Naturgesetzen regiert werde, die vom Menschen entdeckt werden können und aus denen Antworten auf «warum?» Fragen abzuleiten seien. Damit betrachteten die Griechen die Ereignisse der Natur als unabhängig von menschlichen Gefühlen und stellten den Menschen als Beobachter ausserhalb der Natur, obwohl sie nicht leugneten, dass auch er den Naturgesetzen unterworfen ist. Mit der Vorstellung, dass objektive, gesetzmässige Wahrheiten über den Zusammenhang der Dinge existieren, gründeten die Griechen also das Projekt der Naturwissenschaft als Suche nach diesen Wahrheiten. Der raketenartige Aufstieg der modernen Naturwissenschaft nahm dann vor vier Jahrhunderten seinen Anfang mit Galileos Entdeckung, dass Naturgesetze mathematisch ausdrückbar sind. Galileo fand, dass es möglich ist, mathematische Modelle, sozusagen quantitative Bilder der Welt, zu entwikkeln, die von genau messbaren Eigenschaften natürlicher Phänomene Rechenschaft geben können.

Ich möchte nun meine bisherigen Ausführungen kurz zusammenfassen: Die Naturwissenschaft bemüht sich Wahrheiten über den Zusammenhang der Dinge darzustellen. Zu diesem Projekt bringt der Naturwissenschaftler seine biologisch begründeten, intuitiven Begriffe mit, mit deren Hilfe er aus Sinneswahrnehmungen Erfahrung gewinnt und aus der Erfahrung sein Weltbild aufbaut. Weiter beruht die Naturwissenschaft auf dem Postulat objektiv wahrer Naturgesetze, die vom menschlichen Verstand, der sie erkennt, unabhängig sein sollten. Dieser Zweck und diese kognitiven Grundlagen bestimmen die Gestalt der Naturwissenschaft, sind aber zugleich auch der tiefste Grund für ihre Grenzen. Im folgenden werde ich versuchen zu erklären, auf welche Weise diese Grundlagen die Begrenztheit der Naturwissenschaft verantworten.

## Kognitive Ungereimtheiten

Am Ende des 19ten Jahrhunderts hatte das von den Griechen begonnene naturwissenschaftliche Projekt unerhört Frucht getragen. So zeigte sich, dass die Natur dem Verstand mit seinen intuitiven Begriffen weitgehend zugänglich ist, und dass man durch die so erreichten Darstellungen enorme Macht über die Dinge ausüben kann. Der hervorragende Dienst, den die griechische Weltanschauung der Technik geleistet hatte, bestätigte in eindrucksvoller Weise ihre Gültigkeit. Gerade zu diesem Zeitpunkt aber brachten weitere Fortschritte in der Erkenntnis der Natur kognitive Ungereimtheiten ans Licht. Es war Niels Bohr, der erkannte, dass diese Ungereimtheiten daraus entstehen,

dass wir auch für naturwissenschaftliche Darstellungen auf die Begriffe der Alltagssprache angewiesen sind, der Alltagssprache, die wir für die Orientierung in unserer Umwelt und die Organisation unserer Gemeinschaften entwickelt haben. Die Modelle, die die Naturwissenschaft als Erklärung der Welt anbietet, sind also sprachliche Bilder, die mit den aus der Alltagssprache entlehnten Ausdrücken konstruiert sind. Solange nur Phänomene von Dimensionen derselben Grössenordnung wie die unserer Alltagserfahrung dargestellt wurden, waren diese Bilder durchaus zufriedenstellend. Dies änderte sich jedoch, als sich die Physik in subatomare und kosmische Dimensionen hineinwagte. In Bereichen von Raum und Zeit, die milliardenfach kleiner und grösser sind als diejenigen unserer unmittelbaren Erfahrung, treten beim Versuch sich mit unseren Sprachmitteln zu orientieren Schwierigkeiten auf. Es zeigte sich nämlich, dass sprachliche Darstellungen von Phänomenen dieses nur mittelbaren Erfahrungsbereichs versteckte Widersprüche enthalten. Um diese zu vermeiden, müssen gewisse Voraussetzungen hinter sprachlichen Begriffen, die zu diesen Widersprüchen führen, abgeändert werden. Derartige Abänderungen haben jedoch zur Folge, dass der Sinn dieser Begriffe mit der Intuition nicht mehr in vollem Einklang steht.

Die ersten schwerwiegenden Abänderungen von Begriffen der Alltagssprache wurden zu Anfang unseres Jahrhunderts von Albert Einstein zur Entwicklung der Relativitätstheorie vorgenommen, nachdem er erkannt hatte, dass die experimentell gesicherte Konstanz der Lichtgeschwindigkeit nicht mit dem intuitiven Begriff der Zeit vereinbar ist. Die sich aus dieser Tagsache ergebenden Widersprüche entstehen aus der im intuitiven Zeitbegriff verborgenen Voraussetzung, dass der Ablauf der Zeit absolut ist. Um diese Widersprüche zu vermeiden, entfernte Einstein jene Voraussetzung aus dem intuitiven Zeitbegriff und kam somit zum Schluss, dass der Zeitpunkt eines Ereignisses nicht absolut bestimmt ist, sondern vom Bewegungszustand des Beobachters abhängt. Es gibt daher nicht nur eine Zeit, sondern für jeden Beobachter die Seine. Einstein löste auch die grundlegende Unabhängigkeit der intuitiven Begriffe von Raum und Zeit auf, deren Entwicklung im Verstand des Kindes Piaget ja ein vollkommen natürlicher Vorgang darstellte.

Zwei Jahrzehnte später führte die Entwicklung der Quantenmechanik zu einer weiteren Erosion der intuitiven Begriffe. Heisenberg's Unbestimmtheitsprinzip zeigte, dass die unvermeidliche Wechselwirkung zwischen Beobachter und Beobachtetem eine instrumentale Grenze der Objektivität setzt mit der Phänomene im dimensionalen Bereich der Atome dargestellt werden können. Auch führte die Quantenmechanik zu dem Schluss, dass wir nicht nur den Ort und den Impuls eines Elektrons nicht mit unendlicher Schärfe messen können, sondern auch, dass das Elektron überhaupt an keinem bestimmten Ort ist und keinen bestimmten Impuls hat. Daher entspricht das Elektron nicht mehr vollkommen dem intuitiven Begriff eines Objektes, das zu einer gegebenen Zeit nur an einem Ort sein und sich nur in einer Weise bewegen sollte. Der Quantenmechanik gröbste Verletzung der Intuition ist wahrscheinlich ihre Behauptung, das die Dynamik des Elektrons nicht den kausalen Verbindungen unterworfen ist, durch die die Ereignisse der Alltagswelt zusammenhängen. Ereignisse der Welt der Atome hängen nur durch probabilistische, indeterminierte Gesetzmässigkeiten zusammen. Einstein weigerte sich, dieser Behauptung stattzugeben, denn er wollte nicht daran glauben, dass Gott mit der Welt Würfel spielt.

In den zwei Jahrzehnten seit Bohr's Tod machte die Kernphysik weitere Siebenmeilenschritte im Marsch der begrifflichen Entfremdung der Naturwissenschaft. Während den Elementarteilchen der Quantenmechanik - Elektronen, Protonen und Neutronen - bereits Eigenschaften zugeschrieben wurden, die mit den intuitiven Begriffen von Raum, Zeit, Objekt, und Kausalität nicht mehr im vollen Einklang waren, hatte doch der technische Sinn der Wörter, wie Masse, Ladung und Spin, mit denen diese Eigenschaften beschrieben wurden, noch einiges gemeinsam mit dem Alltagssinn. Jedoch mit der Entwicklung der Theorie der «Quark» Teilchen, oder der sogenannten «Chromodynamik», trat ein Sprachgebrauch in die Physik ein, der, obwohl er von Wörtern der (englischen) Alltagssprache, wie «up», «down», «strange» und «charm», Gebrauch macht, diese Wörter nicht mehr metaphorisch be-

nützt. Hier haben «up», «down», «strange» und «charm» nichts mit «oben», «unten», «seltsam» und «Zauber» zu tun und können daher keinerlei bildliche Vorstellung erwekken. Offensichtlich sind die so benannten Eigenschaften der Quarks rein formale, semantisch sinnlose Symbole, die sich nicht auf die Welt beziehen.

Nun können wir uns aber fragen, in wie fern die zeitgenössische Kernphysik eigentlich noch Naturwissenschaft ist. In welcher Hinsicht können die Darstellungen der Chromodynamik der Quark-Teilchen überhaupt als «wahr» bezeichnet werden? Wie können sie meine Zustimmung gebieten, wenn sie mir nicht erlauben, mir ein Bild zu machen, das ich mit meinem Weltbild in Einklang bringen kann? Ist der Zweck dieser Theorien noch die Darstellung gesetzmässiger Zusammenhänge zwischen Ereignissen einer realen Welt der Dinge oder sollen sie nur Ergebnisse von hochgezüchteten Experimenten voraussagen? Ist es möglich, dass die Theorien der Chromodynamik unsere Macht über die Dinge erweitern und Antworten auf praktische «wie?» Fragen liefern, selbst wenn sie nicht zu greifbaren Bildern und sinnvollen Antworten auf theoretische «warum?» Fragen führen? Offenbar treffen wir hier auf eine der wichtigsten Kernfragen für die Zukunft der Naturwissenschaft.

Um diesen Abschnitt meiner Ausführungen kurz zusammenzufassen weise ich nochmals darauf hin, dass die Naturwissenschaft, in den zweieinhalbtausend Jahren seit ihrem Stapellauf in Griechenland, einen triumphalen Aufstieg machte. Sie zeigte, dass die Natur dem Verstand mit seinen intuitiven Begriffen weitgehend zugänglich ist, und sie lieferte uns umfassende Macht über die Dinge. Dann aber, anfangs unseres Jahrhunderts, brache der Fortschritt in der Physik plötzlich kognitive Ungereimtheiten ans Licht, die es nötig machten, unsere intuitiven Begriffe abzuändern. Die von der Relativitätstheorie und der Quantenmechanik geforderten Abänderungen näherten die Naturwissenschaft einer Grenze, die dann von der Chromodynamik überschritten wurde. Diese Grenze ist eine semantische Grenze, jenseits derer die Welt nicht mehr mit unseren intuitiven Begriffen in Bildern verstanden werden kann. Es ist fraglich, ob Naturwissenschaft jenseits dieser Grenze überhaupt noch die Darstellung gesetzmässiger Zusammenhänge zwischen realen Ereignissen ist oder lediglich ein Formalismus, der im Stande ist, Ergebnisse von Experimenten vorauszusagen. Sollte es sich dennoch herausstellen, dass die Chromodynamik tatsächlich praktische Folgen hat, so wie uns die Relativitätstheorie und die Quantenmechanik das Atom als eine praktische Energiequelle erschlossen, dann wäre das in meinen Augen ein echtes Wunder, das man nicht so einfach mit Argumenten über die natürliche Selektion des Gehirns vom Homo sapiens vom Tisch fegen könnte.

## Indeterminismus der zweiten Stufe

Bisher habe ich also versucht, Ihnen zu zeigen, dass sich die Naturwissenschaft einer semantischen Grenze nähert. Nun möchte ich eine zweite Grenze der Naturwissenschaft erörtern, die ebenfalls erst in unserem Jahrhundert in Sicht kam. Diese Grenze wurde in den fünfziger Jahren vom Mathematiker Benoit Mandelbrot erkannt, als er vergeblich versuchte, die Schwankungen des Baumwollpreises statistisch zu erfassen. Die Schwierigkeiten, denen er im Laufe dieser Arbeit begegnete, veranlassten Mandelbrot, ein erkenntnistheoretisches Argument zu entwickeln, das sein Scheitern erklären sollte. Dieses Argument ist nicht nur auf die Nationalökonomie anwendbar, sondern hat allgemeine Gültigkeit für die Suche nach Antworten auf «warum?» Fragen. Zur Entwicklung von Mandelbrots Argument bemerken wir zuerst einmal, dass die Naturwissenschaft ein statistisches Unterfangen ist. Der Naturwissenschaftler sucht immer nach einem gemeinsamen Nenner, oder einer Struktur in der Gesamtheit von Ereignissen, an denen er interessiert ist. Sobald er glaubt, eine solche Struktur erkannt zu haben, vermutet er, dass die Ereignisse zusammenhängen, und versucht ein Gesetz zu finden, das die Ursache dieses Zusammenhangs erklärt. Um ursächlichen Zusammenhängen auf die Spur zu kommen, ist der Naturwissenschaftler also auf eine Mehrzahl von ähnlichen oder verwandten Ereignissen angewiesen. Ein einzigartiges Ereignis, oder wenigstens der Aspekt von dem aus gesehen das Ereignis einzigartig ist, ist einer solchen Untersuchung nicht zu-

gänglich und kann daher nicht der Gegenstand einer naturwissenschaftlichen Theorie sein. Einzigartige Ereignisse sind zufällig, an ihnen ist nichts zu erklären, und der Beobachter nimmt sie als Rauschen wahr. Da aber jedes wirkliche Ereignis etwas einzigartiges an sich hat, ist in jedem wirklichen Phänomen Rauschen zugegen, und es ist die Aufgabe jeglicher naturwissenschaftlichen Untersuchung, die sinnvolle Struktur eines Phänomens vor diesem Hintergrundsrauschen zu erkennen. Je kleiner der Anteil von Einzigartigkeit an den Ereignissen, aus denen ein Phänomen besteht, desto rauschärmer ist es und desto verlässlicher kann seine Struktur erkannt werden. Fast alle Phänomene, die bis vor ungefähr hundert Jahren erklärt wurden, sind verhältnismässig rauschfrei. Derartige Phänomene können durch deterministische Theorien erklärt werden, die sich dadurch auszeichnen, dass ein gewisser Ausgangszustand nur zu einem einzigen Endzustand führen kann. Erst gegen Ende des neunzehnten Jahrhunderts wurden statistische Methoden zur Analyse vormals unergründlicher, rauschreicher Phänomene herangezogen. Aus der statistischen Behandlung ergaben sich indeterministische Theorien, zum Beispiel die kinetische Gastheorie und die statistische Thermodynamik. Eine indeterministische Theorie erlaubt, dass ein gewisser Ausgangszustand zu mehreren alternativen Endzuständen führen kann, und bestimmt lediglich die Wahrscheinlichkeit, mit der die alternativen Endzustände eintreten werden.

Mandelbrot behauptet nun, dass viele der rauschreichen Phänome, die sich noch immer einer erfolgreichen theoretischen Analyse entziehen, nicht nur für deterministische Theorien, sondern auch für die gewöhnlichen indeterministischen Theorien der sogenannten «ersten Stufe», unzugänglich sind. Diese Phänomene kennzeichnen vielmehr einen «Indeterminismus der zweiten Stufe». Das Kriterium, das laut Mandelbrot indeterministische Phänomene der ersten Stufe von denen der zweiten Stufe unterscheidet, ist der statistische Charakter ihrer spontanen Aktivität, das heisst des sie begleitenden Rauschens. Die Energie einzelner Gasmoleküle zum Beispiel, unterliegt grossen Schwankungen, während die durchschnittliche kinetische Energie einer steigenden Anzahl von Molekülen rasch auf einen Grenzwert konvergiert. Es gibt jedoch viele Phänomene, die eine spontane Aktivität haben, die ganz anders verteilt ist und für die es bisher nicht möglich war, erfolgreiche Theorien zu entwickeln. In diesen Fällen konvergiert der Durchschnitt einer Reihe von Messwerten nicht oder nur sehr langsam auf einen Grenzwert. Für derartige Phänomene, so behauptet Mandelbrot, kann sich der Beobachter nur sehr schwer davon überzeugen, ob eine wahrgenommene Struktur real oder nur ein Stück seiner Einbildung ist.

Statistische Verteilungen dieser Art werden «Pareto» Verteilungen genannt nach dem italienischen Ökonom der Jahrhundertwende, der sie erstmals in der Verteilung von Einkommen fand. Wie sich herausstellte, ist die spontane Aktivität vieler geographischer, meteorologischer und astronomischer Phänomene durch Pareto Verteilungen gekennzeichnet. Die Anzahl vergleichbarer Ereignisse, die dem Forscher verfügbar sind, ist daher meist zu klein um ursächliche Zusammenhänge erkennen zu können. Daher ist das Wahrnehmen von Strukturen in diesen Phänomenen keine Garantie dafür, dass sie nicht durch Zufall entstanden sind, so wie zum Beispiel eine Wolke, in der wir ein uns bekanntes Gesicht wahrnehmen, ein Produkt des Zufalls ist.

Der Indeterminismus der zweiten Stufe erklärt das offensichtliche Manko von erfolgreichen Theorien in den Sozialwissenschaften, wie Ökonomie und Soziologie, die in dieser Hinsicht weit hinter den Erfolgen der Naturwissenschaft zurückbleiben. Im Gegensatz zur Naturwissenschaft sind in den Sozialwissenschaften die interessanten Phänomene, die es quantitativ zu erfassen gilt, überwiegend im Nebel der Pareto Verteilungen verhüllt. Daher ist ein baldiges Aufblühen der Sozialwissenschaften leider nicht zu erwarten, obwohl wir doch gerade ihre Antworten auf «warum?» und «wie?» Fragen so dringend nötig hätten. An der Mehrzahl ihrer Theorien wird notwendigerweise der jede praktische Möglichkeit der Verifizierung ausschliessende Indeterminismus der zweiten Stufe haften bleiben.

Lassen Sie mich zusammenfassen: Viele bisher ungeklärte Phänomene, vor allem in den Sozialwissenschaften, sind dadurch gekennzeichnet, dass ihre ursächlich zusammenhän-

genden Aspekte unter einer grossen Zahl zufälliger, das heisst rauschartiger Aspekte verborgen sind. Die Anzahl vergleichbarer Ereignisse, die für Untersuchungen solcher Phänomene verfügbar sind, ist dann viel zu gering, um eine zweifelsfreie Unterscheidung zufälliger und ursächlicher Aspekte zu erlauben. Dies hat zur Folge, dass wissenschaftliche Theorien solcher Phänomene nur in einem sehr beschränkten Ausmass fähig sind, Zustimmung zu gebieten, das heisst, Anspruch auf Wahrheit zu erheben. Wir stossen hier also auf eine weitere Grenze der Naturwissenschaft, eine Grenze die ich eine strukturelle Grenze nennen möchte, da sie aus der Struktur gewisser komplexer Forschungsgegenstände erwächst.

**Hermeneutik**

Nach meiner Erörterung einer semantischen und einer strukturellen Grenze der Naturwissenschaft möchte ich schliesslich eine dritte kognitive Grenze zur Diskussion stellen. Diese Grenze ist den Geisteswissenschaften schon lange unter dem Namen «Hermeneutik» bekannt. Der Name wurde ursprünglich von Theologen auf die Deutung der Heiligen Schrift angewendet und ist hergeleitet von Hermes, dem Götterboten. Hermes, in seiner Eigenschaft als Informationskanal, der Götter und Sterbliche verbindet, interpretiert oder macht explizit den Sinn, der in den göttlichen Botschaften verborgen oder lediglich implizit enthalten ist. Der wahrscheinlich wichtigste Beitrag der Hermeneutik ist ihre Einsicht, dass verborgener Sinn eine Schwierigkeit für die Interpretation von Texten darstellt. Denn es ist notwendig den Kontext zu verstehen, in dem der ganze Text eingebettet ist, ehe die Möglichkeit besteht, verborgenen Sinn in irgend einem seiner Teile aufzudecken. Hier stehen wir vor einem logischen Dilemma, dem hermeneutischen Kreis. Einerseits machen die Wörter und Sätze, aus denen der Text aufgebaut ist, keinen Sinn, bevor man den Sinn des ganzen Textes kennt. Andererseits kann man aber zu dem Sinn des ganzen Textes nur durch das Verstehen seiner Teile kommen. Um diesen Kreis zu sprengen, ruft die Hermeneutik den Begriff des Vorverständnisses zur Hilfe. Vorverständnis ist die Summe der Erfahrungen und Einsichten, die der Deuter von Anfang an zur Interpretation des Textes mitbringt und die es ihm erlauben, den Sinn des Ganzen intuitiv zu erfassen.
Insofern als die Hermeneutik eine Wissenschaft darstellt, können wir fragen, ob ihre Interpretationen objektive Gültigkeit beanspruchen können. Eine objektiv gültige Interpretation müsste den im Text verborgenen «wahren» Sinn explizit gemacht haben. Aber da das vom Interpreten dem Text entgegengebrachte Vorverständnis von seinem höchst eigenen historischen, sozialen, und psychologischen Hintergrund abhängt, ist seine Interpretation notwendigerweise subjektiv. Daher kann eine Interpretation nicht objektiv «wahr» sein, und sie kann Zustimmung nur in einem Personenkreis gebieten, dessen Mitglieder das gleiche Vorverständnis mitbringen. Hinsichtlich der Unmöglichkeit von allgemein, ewig gültigen Wahrheiten in der Interpretation von Texten, unterscheidet sich die Hermeneutik offensichtlich von der von den Griechen konzipierten Naturwissenschaft und ihrem Ideal der objektiven Wahrheiten in der Interpretation von der Natur. Aber gerade diese ideale Auffassung der Naturwissenschaft wird von manchen zeitgenössischen Philosophen zurückgewiesen. Sie behaupten, dass auch der Naturwissenschaftler sein subjektives Vorverständnis zur Interpretation der Welt der Dinge mitbringe und dass daher auch die Naturwissenschaft keine objektiven Wahrheiten zu Tage fördern könne. Dennoch, selbst wenn wir dieser Ansicht stattgeben, bleibt doch die Tatsache bestehen, dass manche naturwissenschaftlichen Darstellungen von weniger Vorverständnis abhängen, und daher relativ höhere objektive Gültigkeit haben, als andere. So können wir den Grad der objektiven Gültigkeit einer wissenschaftlichen Darstellung abschätzen, wenn wir das Ausmass feststellen, in dem Vorverständnis eine Rolle in ihrer Entwicklung gespielt hat. Eine solche Abschätzung zeigt, warum der Glaube an die Möglichkeit von objektiv gültigen Darstellungen wenigstens in den «harten» Naturwissenschaftszweigen, wie der Physik, mehr berechtigt ist, als in den «weichen» Human- und Sozialwissenschaften, wie der Psychologie, der Ökonomie, und der Soziologie. Eine der Hauptursachen für diesen Unterschied im Gültigkeitsanspruch, das

heisst in der Fähigkeit, Zustimmung zu gebieten, ist, dass die Phänomene, die die «weichen» Wissenschaften zu erklären versuchen, von höherer Komplexität sind als die Phänomene, mit denen sich die «harten» Wissenschaften befassen.

Als zwei extreme Beispiele – das eine sehr hart und das andere sehr weich – können wir die Mechanik und die Psychoanalyse vergleichen. Die Theorien der Mechanik überzeugen uns von ihrer objektiven Wahrheit, weil die für die Mechanik bedeutungsvollen Phänomene, wie Stahlkugeln die einen Abhang herunterrollen, von geringer Komplexität sind. Es ist ohne viel Vorverständnis möglich, dieses Phänomen in seine wesentlichen Bestandteile – Stahlkugel und Abhang – zu zerlegen, die von den kausalen Verbindungen der Theorie bestimmt werden. Zur Verifizierung der Theorie können dann kritische Beobachtungen und Experimente – mit verschiedenartigen Abhängen und Kugeln – angeführt werden. Im Gegensatz zur Mechanik fehlt den Theorien der Psychoanalyse die Überzeugungskraft objektiver Wahrheiten, weil die Phänomene der Psyche ausserordentlich komplex sind. Ohne weitgehendes Vorverständnis kann der Psychoanalytiker überhaupt keine Strukturen erkennen, geschweige denn das Phänomen des Analysanden in seine wesentlichen, kausal zusammenhängenden Bestandteile zerlegen. Von kritischen Beobachtungen oder Experimenten kann hier keine Rede sein, weil es fast immer möglich ist, das Nichteintreffen einer von psychoanalytischen Theorien abgeleiteten Voraussage wegzuerklären, indem man sein Vorverständnis des Phänomens einfach abändert. Psychoanalytische Theorien sind daher kaum zu verifizieren, weshalb manche Naturwissenschaftler der Psychoanalyse schlichtweg den Anspruch verweigern – meiner Ansicht nach unberechtigterweise – überhaupt ein Wissenschaftszweig zu sein.

Die Neurobiologie, über die David Hubel hier gerade berichtet hat und die auch mein eigener Forschungsbereich ist, überspannt einen weiten Bereich auf dieser «hart-weich» Skala der Naturwissenschaft. An ihrem harten Ende ist die Neurobiologie durch elektrophysiologische, anatomische, und biochemische Untersuchungen von Nervenzellen vertreten. Obwohl die mit Nervenzellen verbundenen Phänomene schon komplexer sind als rollende Stahlkugeln, können die Beobachtungen immer noch mit Theorien erklärt werden, die überzeugenden Beweisen unterliegen. Aber an ihrem weichen und für die Mehrzahl unserer Kollegen reizvolleren Ende ist die Neurobiologie mit systemanalytischen Untersuchungen der Struktur und Funktion grosser und sehr komplizierter Nervenzellnetze befasst. Und die mit diesen Netzen verbundenen Phänomene sind annähernd so komplex wie die Psyche selbst, ja sie schliessen die Psyche sogar ein. Daher nimmt die Neurobiologie an ihrem weichen Ende den Charakter der Hermeneutik an: zur Analyse eines komplexen Nervenzellennetzes muss der Forscher erhebliches Vorverständnis der Ganzheit des Systems mitbringen, ehe er versuchen kann, die Funktion eines Teils des Systems zu deuten. Und so wird auch zukünftigen Versuchen, die komplexe Arbeitsweise des Gehirns zu erklären, der Anschein objektiver Wahrheit weitgehend fehlen.

Zusammenfassend können wir über die Hermeneutik sagen, dass die von ihr erkannte, unabkömmliche Rolle von subjektivem Vorverständnis nicht nur der Interpretation von Texten, sondern auch der Interpretation der Natur eine Grenze setzt, die ich eine subjektive Grenze nennen möchte. Der hermeneutische Begriff des Vorverständnisses erlaubt auch eine einleuchtende Erklärung für den augenscheinlichen Unterschied zwischen «harten» Wissenschaften, die sich mit Phänomenen geringer Komplexität befassen und zu Theorien grosser Überzeugungskraft führen, und den «weichen» Wissenschaften, die Phänomene grosser Komplexität zum Gegenstand haben und zu Theorien führen, die kaum fähig sind, allgemeine Zustimmung zu gebieten.

## Coda

Ich schliesse meine Ausführungen von Ursprung, Grenze und Zukunft der Naturwissenschaft, indem ich nochmals daran erinnere, dass die Naturwissenschaft eine semantische Tätigkeit ist, die das Ziel hat, Wahrheiten über gesetzmässige Zusammenhänge zwischen Dingen der Welt darzustellen. Die Naturwissenschaft bemüht sich daher, Antworten auf theoretische «warum» Fragen zu

finden, die ihrerseits zu Antworten auf praktische «wie?» Fragen führen können. Die durch Antworten auf «wie?» Fragen erweiterte Macht über die Dinge stellt eine von jeder Logik unabhängige, ontologische Bestätigung der Wahrheit auf «warum?» Fragen dar. Dieser, von mir hier viel zu kurz skizzierte Frage und Antwort Vorgang, beruht auf intuitiven Begriffen, die wir im Laufe einer biologisch gegebenen, kognitiven Entwicklung in unserer Kindheit erwerben. Mit dieser begrifflichen Ausrüstung wurde das von den Griechen konzipierte Projekt der Naturwissenschaft mit enormem Erfolg vorangetrieben, und Ende des neunzehnten Jahrhunderts war es klar, dass die Natur diesen Begriffen nicht nur zugänglich ist, sondern dass der Mensch durch die auf diesem Weg erhaltenen Einsichten auch umfassende Macht über sie ausüben kann. Als sich jedoch in unserem zwanzigsten Jahrhundert die Naturwissenschaft anschickte, die Geheimnisse der Natur bis in die Tiefe der Nacht zu verfolgen, stiess sie auf kognitive Grenzen. Drei Grenzen habe ich hier besprochen. Die erste ist eine semantische Grenze, auf die wir in Bereichen stossen, die weit ausserhalb unserer unmittelbaren Erfahrung liegen. Dort versagt unsere Fähigkeit, die Welt in sprachlichen Bildern darzustellen, die wir mit unserem intuitiven Weltbild in Einklang bringen können. In diesen entfernten Bereichen wird uns daher die Erkenntnis dessen, was wir Wahrheit nennen, versagt bleiben.

Die zweite kognitive Grenze ist eine strukturelle, die auf dem undurchsichtigen statistischen Charakter vieler, bisher noch ungeklärter Phänomene beruht. Bei diesen Phänomenen können wir uns nur schwer überzeugen, dass eine wahrgenommene Struktur überhaupt ein Bestandteil der Wirklichkeit und nicht lediglich ein Produkt unserer Einbildung ist. Die dritte kognitive Grenze ist eine subjektive, und sie entsteht dadurch, dass wir bei der Darstellung eines komplexen Phänomens weitgehend auf subjektives Vorverständnis angewiesen sind.

Wir können daher erkennen, dass sich die Naturwissenschaft, unabhängig von wirtschaftlichen, physikalischen, sozialen oder politischen Grenzen, rein kognitiven Grenzen nähert. Einerseits hat die Naturwissenschaft unsere intuitiven Begriffe, mit denen wir die Welt erfassen, abgeändert und einen semantisch immer sinnloser werdenden Sprachgebrauch eingeführt. Anderseits sind Zweifel aufgetreten, ob die Strukturen überhaupt existieren, die wir in den bisher noch nicht erfolgreich behandelten, komplexen Phänomenen wahrzunehmen glauben. Sehr viel hat die Naturwissenschaft schon geschafft, und sehr viel müsste sie noch schaffen. Aber was für einen Sinn wird sie haben?

Ich danke Jochen Braun für anregende Diskussionen und Hilfe bei der Redaktion des deutschen Textes.

## Literatur

In den folgenden Veröffentlichungen des Autors sind die hier erörterten Themen eingehender besprochen:

Stent, G.S. 1978: Paradoxes of Progress, San Francisco. W.H. Freeman & Co., 23 pp.

— 1979: Science and Morality as Paradoxical Aspects of Reason. In: Knowing and Valuing: The Search for Common Roots, H.T. Engelhardt and D. Callahan, eds. Hastings-on-Hudson, New York. The Hastings Center. pp 79–101.

— 1979: Does God Play Dice? The Sciences. pp 18–23.

— 1979: Naturwissenschaft und Ethik als paradoxe Schöpfungen der Vernunft. Naturwissenschaften 66. S. 354–357.

— 1981: Cerebral Hermeneutics. Journal of Social and Biological Structures 4. pp 107–124.

*Anschrift des Verfassers:*

Prof. Dr. Gunther S. Stent
University of California
Department of Molecular Biology
Wendell M. Stanley Hall
Berkeley, CA 94720 (USA)

# Round Table

Werner Arber, Manfred Eigen, David Hubel, Hubert Reeves,
Günther S. Stent, Victor F. Weisskopf

*Arber:* In the course of the sessions I felt that there were some problems of cognitive limits. I cannot express this concern in the language of the specialists, but one simple question which I would like to ask now in order to start the discussion is the following: It is our habit to consider space as 3-dimensional, measured in metres or centimetres in a linear scale, and we are accustomed to measure time in minutes or hours, also in a linear scale. Particularly, thinking of the first talk on the universe, but also of all others, I just want to ask: is it relevant to consider other scales than the linear, both for space and for time? exponential for example?

*Weisskopf:* I believe that this is a very interesting question which one cannot answer suddenly in any definite way but I would like to mention one point in this connection. Evidently many of you, including myself, have asked: what about the «Big Bang» and the time zero? What came before? Is there not something wrong with our way of measuring the time? For example, the real unit could be a logarithm of the time which then goes to minus infinity at zero. This is not only a mathematical trick in order to avoid zero, because what is the essence of time? The essence of time, for example, is frequency. Let us take the frequency of the light that fills the universe, as Haydn has described so well. At times when the universe was much denser, the frequency was much higher. So if you use that frequency as your time element, then the time element at the beginning was very much shorter, and so time goes to infinity. Indeed, if you do this quantitatively you find the logarithm of time as the right measurement. One might even in that sense, avoid the question of «before» by just changing the time unit.

*Eigen:* But what about the direction of time?

*Weisskopf:* Well, it is very interesting in the spectacle of the universe, because there is this idea of the – as our colleague Sorkin has expressed it – harmonica world. For example, if the mass density in the universe is big enough, the universe expands, then comes to a stand, comes back again and starts over again. The direction of time depends upon what period you live in. But it does seem that this idea runs into serious difficulties. First into experimental difficulties: The density which we observe seems to be too low to pull the universe back again. The second difficulty is a statistical one: if there was really a harmonica, the entropy should be somewhat larger after each expansion and then, after infinite time, at the end, there would be all in disorder. So this is a very questionable idea. In some ways therefore, the direction of time is seemingly inscribed in the fundamental laws which, of course, we do not know yet.

*Arber:* Would you say that if everything were in equilibrium the direction of time would disappear?

*Weisskopf:* Yes. But of course the infinite expansion of the universe prevents that, it isn't in equilibrium.

*Arber:* If we compare time with 3-dimensional space or, if you like, a 2-dimensional space: we know that – and you mentioned it in your presentation – if we have a sphere, you walk on that sphere and you never come to an end and you don't see the beginning. Does something like that exist for time?

*Weisskopf:* I must admit that I cannot answer this question. H. Reeves, maybe you could?

*Reeves:* No. But I was going to ask about something you just said. This increase of

entropy would be represented ultimately by photons. And in the case that the universe contracts back, then the photons will re-equilibrate with all the particles, and then it seems to me that this increase of entropy is levelled and comes to zero. I have worried very much about what would be carried from one chapter of this harmonica to the other one and I have no answer to this. Do you have one?

*Weisskopf:* No, no. But I do believe it is an important question.

*Stent:* I ask about this logarithmic time. If that were so, then the word Big Bang would not be an appropriate metaphor. There would not have been a Big Bang, everything was just «going on».

*Weisskopf:* Absolutely, yes. It is the singularity which is anyway fathomless.

*Stent:* There is no singularity then.

*Weisskopf:* That singularity is in minus infinity. This means, whenever you are on a finite time measured in logarithm you will have a finite density which increases and going backwards.

*Stent:* The conclusion will be, as I take it, if we consider time logarithmically, things which we think happened tremendously fast during the Big Bang, actually occurred no faster than anything which is happening today.

*Weisskopf:* It depends what you mean by fast. If you define fast relative to the average frequency of the radiation, then it is not fast.

*Reeves:* I like very much this exponential scale because it reminds me of the absoluteness of the velocity of light and of the zero of temperature. In the linear scale you would say «Why is light going at that speed and why is there a zero of temperature? In the logarithmic scale you understand this in terms of the effort you have to make to reach the velocity of light or to reach zero Kelvin. In the same way, the effort you have to make to understand the universe becomes larger and larger, the more you approach its origin. This is why, when we make a step from $10^{-35}$ to $10^{-31}$ seconds, it takes chapters of physics to understand what is going on, even for such a small period of time.

*Eigen:* But I think that such an extrapolation, c.e. to say that time has a logarithmic scale is worth nothing, because there is evidence for many phenomena in physics that they started by bifurcations, through an instability. A similar phenomenon might as well be behind the Big Bang.

*Weisskopf:* Yes, I fully agree. I think this would be a good moment to remind everybody including myself that whatever we say about things that happened in these first fractions of seconds, they are purely hypothetical. They are on a much less safe basis than what we say in any other field of physics, including high-energy physics and astrophysics.

*Eigen:* We are facing similar problems in the evolution of life. There is a certain continuity as to prepare the conditions for something to appear, but in between we have many discontinuous processes. In other words, the sudden appearance of a certain mutant might completely change the scene of evolution.

*Reeves:* However I would qualify the uncertainty: when we say that after one second or one minute there is a chapter of primordial nucleosynthesis, we are on much better ground than if we say that at $10^{-35}$ seconds we have the reaction which explains why we live in a universe which is made of matter and not matter and antimatter. Although the second one is much more speculative, it is more interesting than the first one which is, I would say, almost believable.

*Weisskopf:* I would like to make a remark concerning Gunther Stent's lecture which I found extremely fascinating, stimulating and, how shall I say, irritating. In particular, I would like to take issue with his first point. Namely, the first of these three limits where he says that, for example, quantum chromodynamics is already outside of the legitimate limits of science.

*Stent:* I didn't use the word «legitimate».

*Weisskopf:* No. You know, always in order to be short, one has to be rough. I fully agree with the fact that science begins with the natural concepts that personal evolution has brought us, as Piaget has shown. But after all, science already for a long time added still further new concepts to these concepts. For example, a child certainly does not know anything about electrical charge. And the electrical charge plays, as you know, an extremely important role just in that type of physics which G. Stent certainly agrees to be reasonable physics. Now, in many ways, those categories that G. Stent has criticized, for example, those different quark types are very much in the nature of charges. Indeed they were sometimes called hypercharges. So, in the whole development of science, always new concepts were added to those which were naturally in us from childhood. The concept of atoms, by the way, at the end of the 19th century, was criticized as non-scientific in the same way as G. Stent now has criticized chromodynamics as being non-scientific. Now we can see the atoms. And let me now make a third remark. I do not think that the reality, the truth of a scientific recognition, depends on the practical applications. As long as one can make provable predictions, as long as, for example, chromodynamics says if you make this and this observation you will find that consequence, or you find in the stars these and these phenomena because of the mechanism of chromodynamics, that, I think, is as good as any practical application.

*Stent:* I am afraid that apparently I did not make myself sufficiently clear. First, of course, I do not wish to suggest that chromodynamics is illegitimate or that their activity is not admitted. But the point I am trying to make is one of pictures and I am very much inspired by Bohr's argument about the necessity of picture-making in scientific theories. Now, as to your first point about the electric charge, I certainly do not wish to claim that one is not allowed to do science that cannot be understood by a child, because naturally most of the things children cannot understand. But the new concepts to which you referred and which have been indroduced as science developed, nevertheless have some nexus continuous with infantile ideas. You mention electric charge. Sure, a baby has no experience with electricity. When, finally, the child learns about electricity in elementary school, the teacher tries to explain in metaphorical terms, what electric charge is, always connecting it with something. I think that they told me it was some kind of a fluid or something like that - I can't remember - but nevertheless I did finally get some understanding of electricity in terms of metaphorical pictures: repulsion, things were being pushed away and they showed me amber and all that stuff. All I say is that - I am only repeating Bohr - scientific theories are in the end pictures which are built from everyday language, although they are altered and modified. What to me is novel, at least alleged in chromodynamics, is that no such attempt is made any more. Frankly, the words are without any metaphorical content, they are purely formal operational symbols.

The second point I want to make is about atoms: you cite the 19th century. I think you probably were referring to Mach and the criticism was not semantic or linguistic, it was positivistic. Mach said that atoms are nonsense, not because conceptually it was nonsense that they should be little balls; he criticized that they had never been seen and that no-one at the time was making any kind of empirical experiment. He was some kind of an early member of the Wienerkreis, a person who said that if you have no empirical proof or experiments or something like that, then it is nonsense. So, it was a different criticism that was made of the atoms, not a conceptual or a linguistic one, but an empirical one.

*Weisskopf:* The «charm» and «strangeness» will be seen too, very soon.

*Stent:* However, the novelty is, that at least at the time that the terms were passed there was no obvious connection between the terms and properties.

As to the last point: I do not wish to say that practical applications are a necessary condition for proof. On the contrary, I was claiming that they can be an additional methodological proof, and when some theory leads to

practical results, then you have a good feeling that your theory was not all that bad. It is the same level of abstractum. Dirac equations are here to stay, probably chromodynamics also. As an example, you could invent a way in which cars would run on water or something like that. That would then to me be a true miracle which could not be explained by the Darwinian hocus pocus. That is the point I tried to make.

*Weisskopf:* Quantum mechanics is also hocus pocus and has tremendous practical applications.

*Reeves:* There is no major difference between Dirac equations and quantum chromodynamics. It is the same level and Dirac equations here are good to state, and probably chromodynamics also.

*Arber:* Let's now shift to talk on the origin of life. From several of the questions received and also from private discussions it seemed to me that some people had difficulty to see where life really comes in. Not everybody is ready to admit that relatively short, replicating RNA molecules already represent life. It is my feeling, that an answer to these questions was hidden in M. Eigen's last slide.

*Eigen:* In this context, the question was raised: what is the criterium for an optimum template. An answer is found in the experiments of Spiegelman with the bacterial virus $Q_\beta$. The genome of phage $Q_\beta$ is an RNA molecule which can be isolated from the virus particle. Spiegelman isolated the enzyme $Q_\beta$-replicase and he put the enzyme and the RNA template together and fed the mixture with energy-rich nucleotides. The system then started to make new copies of the genome, but after several generations all the new copies were as infectious as the original viral RNA. But then he went on and put the system under selective pressure for fast replication. Then the copies go shorter and shorter and the original, 4500 nucleotides long template got shortened down to about 500 nucleotides. In addition, the shorter templates were able to replicate faster; the speed of replication per nucleotide was increased by a factor of 3 to 4. It was estimated that this very efficient replication reaches the upper limits possible within the limits of physics and chemistry. Hence the criterium of optimal replication was fulfilled by these short templates, but they were not infectious any more. They had lost all the information to enable them to penetrate a bacterial cell. All they could was to replicate quickly.
In the experiments I reported, we did not start with the $Q_\beta$ genome as a template, we rather started only with the replication enzyme and energy-rich nucleotides. And the enzyme started to line up the nucleotides and to connect them to new templates. Again, the rates which we found approach the upper limits we could think of, and furthermore the process is reproducible, regardless of the environment. That is our criterium of an optimum. In this case, it is still trivial that we get reproducible results, because the estimated number of possible alternative products under the particular conditions is about $10^{12}$, and having $10^{14}$ enzyme molecules around, we could always hope to materialise the best template. Now, of course, for the true living beings, which comprise much more information, one cannot scan through all possibilities. Here we have the constraint that larger sequences must evolve from shorter ones until they reach their optimal length. Thus, in order to define a criterion for the optimum, we have to consider that historical route becomes part of the boudary condition.

*Arber:* Let's assume you have these primary elements. You start to build up, at what time does the new principle «life» come in?

*Eigen:* We could also ask: what is the lowest molecular weight one could associate with life? There is always the question: what is life? and I don't like the question very much because such a definition doesn't tell us very much. There is no disagreement to call bacterial cells alive, and we all agree that we are living beings, but what do we have in common with the bacteria? - Just the chemistry: To give a definition of the term «life» doesn't tell us very much about the living beings. Now there is the question how far can we go down? Do we want to call a virus a living being? If you are willing to do so, you should admit that the smallest viruses now known are plant viruses with only a few hundred

nucleotides in their genome. Their size is of the same order of magnitude as the molecules we produced de novo in our experiments.
One could use an operational definition. Namely, that the system, in order to be alive, has to be able to reproduce itself and to adapt itself by mutation to environmental changes. Furthermore it requires a metabolism. The metabolism is even a necessary prerequisite because of the fact that all these processes can only occur far away from equilibrium. If self-reproduction for instance occurred at equilibrium, one would not expect an effect of selection because of microscopic reversibility which is effective at equilibrium. So, self-reproduction as a prerequisite of selection works only far from equilibrium. You might then say: a molecular system which fulfills these conditions, which starts to reproduce itself, mutates and can thereby adapt to any condition, might be called a living system. But that might not yet satisfy molecular biologist who request the system also to make proteins, in other words, to translate its genetic information in order to gain an unlimited functional capacity. In any case, if we talk about a living system, we have to specify what type of system we mean.

*Arber:* That reminds me of the problem of cognitive limits. In some way, we are unable to define life properly.

*Stent:* Oh, I think not, but we might perhaps better devote the discussion to the origin of a bacterium such as E.Coli And we might try to trace a reasonable history starting from the atoms and ending up with E.Coli. We would of course like to know what is in between.

*Eigen:* I agree that there is a big jump to an E.Coli. It is quite clear to me, however, that an evolutionary process which takes that direction can only do so after self-replication was established. It is thus clear that the nucleic acid had to start this kind of process although the proteins, being chemically simpler, might have been around long before. But they had no way to optimize their functions.

*Arber:* Let us go back to the universe: In fact, can one expect, if the universe is infinite, that there is an infinite number of planets on which life has developed? If so, what is the chance that there is also on some of these planets actually intelligence developing?

*Weisskopf:* Probably we should first ask the astronomers. This question entirely depends on the probability to find planets with conditions and histories like our own.

*Reeves:* Yes, this you can divide into two questions. What is the probability that stars have planets and then, what is the probability that these planets are habitable.
The probability that stars have planets is, I would say, very large. In fact, when we look at the stars, we find that the stars which are alone are quite rare. At least two-thirds of the stars live in couples or triples or more complex systems. If you look at a star which is a double star and you then improve the visibility, you often find that it is triple or quadruple and so on. So there is a very large probability that planets in existence like ours are very, very common and perhaps one star out of two or three has planets. Since we have a hundred billion ($10^{11}$) stars in our galaxy and since in the observable universe there is in the order of a billion ($10^{9}$) galaxies, you see that this makes a lot of planets.
Then you ask the next question: what is the probability that some of these stars have planets with life? One limitation to the development of life is given by the orbit of the planet. If our earth, for instance, had a very elliptical orbit, if it was going far away from the sun, and close by again, we would have large variations in temperature which would probably be very harmful to development of life. If our system belonged not to one star but to two and was going in an eight around two stars like you can have with double stars, you would probably have a similar type of difficulty. Nevertheless, I am ready to assume that probably a good fraction of stars, certainly hundreds of billions, have habitable planets getting some ultra-violet light although not too much, and where the temperature is quite constant. The astronomers cannot go past this statement. Everybody wants to know how many of these habitable planets have developed life. We have experts

here and I am glad to pass the question to them.

*Eigen:* Well, the question again is difficult to answer, because of the word «life». The earlier organizations about which I have talked are almost so deterministic that whenever appropriate conditions are created, those organizations will show up; Thus, whenever there is a planet whith earthlike conditions such as reducing atmosphere, it will start to synthetize all these chemicals, which then start to polymerize. If it were not so, we wouldn't have a chance to find it in the laboratory. I would therefore propose that those primitive states of life certainly must exist if habitable planets are around. Now comes the difficult question, on which no experimental results are available, namely whether a cellular organism such as E. Coli also could have formed just due as a consequence of environmental conditions. And what about the evolution of higher life? We know that life came to a standstill almost at the level of unicellular organisms and we guess also why: because bacteria already have a genome of a few million nucleotides. In order to keep it stable, it had to reproduce it with an accuracy of 1 in $10^7$, or so which means that a particular mutation which could have brought about progress, would have been very here. If the mutation rate is 1 in 100, it happens every day, as we have shown by experiments. If the mutation is 1 in one million, it takes much longer time. The rate of evolution decreases. The way out of this situation was a mechanism to exchange information between two organisms, i.e. recombinative processes or sex. That immediately allowed to spread advantages through the whole population. One can estimate that it took about 3 billion ($3 \cdot 10^9$) years from the existence of unicellular life up to evolution of higher organisms. Mankind is not older than a million or a few million years.
The process of evolution from unicellular to higher life is subject to fluctuation. Suppose that fluctuation in time is only 1 percent. One percent in a billion years is 10 million years. It appears thus unlikely that the appearance of higher forms of life on neighbouring planets is synchronous. It is thus doubtful if we will ever have an overlapping time to communicate with intelligent organisms of other planets and find them in a state in which they are able and willing to communicate with us.
And who knows, whether we are still willing to communicate with anybody outside after another hundreds of thousands of years. But that is a completely different question.

*Arber:* I have just one little restriction to what you said, to which I otherwise fully subscribe. We do know that viruses, which we usually do not consider as organisms with sex, are also able to transport genetic material from one cell to another, and various molecular mechanisms are known to promote this exchange. That is one point, and the second: I could imagine that bacteria which usually divide into two every half hour, loose the capacity to separate completely, which would probably lead to a very primitive multicellular organism. This might allow for a compartmental evolution by mutation, perhaps helped by viral infections, bringing segments of genetic material from foreign sources.

*Stent:* In his new book, Francis Crick has developed an argument which is relevant to what was just said, and whose devilish, fiendish conclusion is that it is entirely possible that life, as we know it, has no natural origin. The argument, that he makes, goes in short as follows: If all calculations come out the way that was suggested here, in a large number of stars, many of which have their planets, sometimes with the conditions for life, it is a necessary consequence that life arises, and one may perhaps assume even intelligent life. There is then also a likelihood that – maybe two billion ($2 \cdot 10^9$) years before us, this waiting period which M. Eigen has decribed did not take place on some planets. Some two billion years ago people like us, or little more advanced than us, could then have existed on a planet. They might have had a space committee and high technological means. They had perhaps also molecular biologists. They might have known that there was a planet here, Earth, in the solar system and so the said committee could have designed a bacterium as being a perfect organism for the conditions of this planet. As to themselves, their life could be based on

silicon and selenium, and have arisen under conditions entirely different from ours. But their acquired knowledge might have enabled them to construct a terrestrial bacterium and to send it in a rocket to the Earth. It then could have infected our ocean, and the rest is history. So, at first, this seems like science-fiction, but it shows that if you develop all the general arguments about necessity, if you really believe that there is something necessary about the origin of life and that the probability of its arising is high, then the credibility is also seriously diminished that our own life has actually a natural origin rather than being created somewhere else according to a particular design. This is the argument developed by F. Crick.

*Weisskopf:* If it is really true that the universe is infinite, one can say that life must be there. Intelligence must be somewhere because even if the probability is extremely small, multiplied with infinity it becomes 1. Now the problem is only this: the communication radius. We can only speak sensibly about that part of the universe with which we can communicate. This radius, of course, expands as time goes on. Therefore, the question should really not be: «Is there life in the universe?», because I think that somehow logically the answer must be «yes», but: «Is it within our communication radius».

*Eigen:* Well, I have no idea what the probability is to find life within that radius. That is a question you should answer. The fact is that we haven't yet received any message.

*Weisskopf:* The probability that two civilisations are in exactly the same state or in approximately the same state is rather improbable. Either they are ahead of us and they are not interested in us, or they are behind us, then they cannot communicate with us.

*Arber:* Another problem is that if they are really far away and even if they are some time ahead of us, their communication may not have reached us yet.

*Reeves:* There is also the problem of how long a civilisation is developing technology and whether it will survive its technology?

*Weisskopf:* May I just add one point: It is not only the ability of technical development. It is by no means sure that any civilisation develops technically. They might be interested in completely different things – in writing poetry for example.

*Hubel:* The original question was what are the chances that intelligent life could develop. However, some other species than ours are enormously successful, insects for example, although they don't have anything that we would call intelligence. It isn't entirely clear that intelligence is a very great advantage and I don't see any way of knowing how much of a chance occurrence the development of intelligence is. I don't see any way of assigning any kind of number to that. Of course, if you multiply anything by infinity you get something, so that doubtless, somewhere, there will be people with two legs, two arms and perhaps ten fingers. It could just as easily have been that life would not have any intelligence even as yet and still have many very successful animal species.

*Arber:* On the other hand, we have a tendency to believe that our intelligence is of the highest level that could develop. Perhaps this is a very wrong idea.

*Hubel:* Certainly, one cannot think about the origins of the brain without wondering how much further this can evolve and, of course, there isn't any answer to that. Our cortex, our thalamus is so much bigger than that of the highest ape. It is rather hard to think that it won't go on getting bigger. But its even hard to know how to think about it, would we be allowed to.

*Reeves:* May I bring in a pessimistic view? If intelligence, I should probably say technology, should bring selfdestruction, then it may not be the best thing that happens to a civilisation.

*Arber:* We could then also ask, how far selfdestruction can go? We can probably destroy life. Can we also destroy the planet as such?

*Reeves:* No. Not really, not at the moment. Perhaps later.

*Arber:* Then, according to M. Eigen's view, things can start again.

*Reeves:* Yes, with a delay of a few hundred million years.

*Weisskopf:* It's a good hope.

*Arber:* Maybe, we should close with this hope.

# Cosmogony of Celestial Bodies and the Formation of the Chemical Elements

Symposium of the Swiss Society for Astrophysics and Astronomy, Davos (Switzerland), September 24–25, 1981

# General Introduction

This symposium took place in Davos, together with the 161st annual meeting of the Swiss Academy of Sciences. The Central Committee of the latter had proposed to place all discussions under the very broad theme of the 'Origin of Things and Beings' (vom Ursprung der Dinge) and for us astronomers, this reflects itself on the origin of the different bodies which populate the cosmic universe.

The six lectures delivered at this symposium deal successively with the origin of planets, of the solar system, stars, galaxies and finally of the universe itself. This ordering is thus correlated with increasing distance scales, starting with close objects like planets and meteorites, about which we have a precise knowledge, even if incomplete. Reversing this order should have meant first laying the stage for the universe at large, of which we can draw today an improved picture, but still controversial in many respects however. Another ordering of the cosmic objects reviewed here could have been chronological, but this ordering appears to be questionable for reasons connected to the conditions prevailing at the formation of celestial bodies or of systems of such bodies.

In the first lecture, by I.P. Williams, the most recent data on the planets and meteorites of the solar system are reviewed, with the purpose of reaching a better understanding of planetary formation and of describing the 'best' current model.

Now it appears that many chemical elements in planetary bodies and in the sun contain isotopic anomalies, often uncorrelated. Going through recent investigation in the field of isotopic research, P. Bochsler points, in the second lecture, to some implications of new observations about several of these elements, on the early history of the solar system.

In the third lecture, P. Bouvier summarizes the successive stages of stellar evolution as we know them today, from the protostars forming of diffuse stellar clouds, to the stars in which thermonuclear reactions ignited inside the central regions provide the main source for the energy radiated away. The great majority of the chemical elements building the matter of our world appears to have originated inside stars, while some of the lighter elements presumably formed in interstellar spallation reactions or during the first quarter of an hour of the early universe, according to the standard Big Bang picture.

It has been customary to think of galaxies as being born from gravitational instabilities in an initially more or less uniform universe, although the growth of such perturbations in time appears to be too low in the case, suggested by the observed baryon density, of an open universe.

At present the recent measurements of the electron neutrino rest mass opens new perspectives in cosmology and several authors have pointed out that massive neutrino condensations may trigger the formation of baryonic matter condensation in the universe at large, probably on the scale of clusters of galaxies. This work is being reviewed in the fourth lecture by F. Occhionero who gives also some new results on the linear growth of baryon condensations, from decoupling onwards, when self-gravitation has overcome the gravitational coupling to preexisting neutrino condensations.

In the fifth lecture, R. Buser alludes to recent work on the spectral evolution of galaxies, using models which allow the calculations of galaxy magnitudes, k-corrections, evolutionary corrections and colours as functions of red shift, to be carried out in a variety of photometric systems. These results are used to interpret the observed colours of faint galaxies, thus providing a test for the deceleration parameter and for constraints on the

ages of galaxies and the history of star formation inside galaxies.

The Hubble discovery of the recession of the galaxies (1929) is a far-reaching observational fact, although not a proof for an expanding universe starting from a singularity (Big Bang); there was still no definite agreement, two decades ago, on whether the expanding universe had a beginning or whether it remained eternally in a steady state.

The present observational evidence, revised by G. Tammann in the sixth and last talk of this symposium, strongly indicates that the universe is indeed evolving and, in addition to the Hubble recession, the colour distribution of galaxies and their counts at different wavelengths require cosmological evolution. Moreover, the universal helium content of celestial bodies and the discovery (in 1965) of the cosmic background radiation can only be explained if we assume that the universe started in a 'Big Bang' some 18 billion years ago.

Such were the topics treated in these six lectures, outlining in a fairly coherent way the state of knowledge reached today about the origin of the physical world.

Pierre Bouvier
Chairman

# The Origin of the Planets

Iwan P. Williams

## Introduction

Pondering on the origin of the earth and planets has been one of the hobbies of the human race since early times. In earliest times actual data was very scarce and so the restriction on the types of theories advanced were very few, and most of those arose because of philosophical or religious reasons rather than conflict with hard data. The first major constraint came with the Copernican realisation that the Sun was the central object in the system. With the discovery of the telescope and the development of dynamics following the work of Newton and Kepler, the general features of the system became known and only minor changes in these have been recorded in the last decade. For information the current values are given in table 1. From the study of the mean densities of the planets, it becomes apparent that major compositional differences exist between certain groups of planets. Based on this, and the dynamical data, an alarming number of theories have been proposed, and these are described in reviews such as ter Haar and Cameron (1963), Williams and Cremin (1968). In the intervening period there have been a number of meetings devoted to the subject, and as a consequence, books of the proceedings have been published, for example Reeves (1972), Gehrels (1978), Dermott (1979).

One of the major difficulties facing any prospective cosmogonist is the diversity of topics which combine to define the whole problem. Ideally, one needs to understand dynamics (orbits and general motion) hydrodyamics (equilibrium of gaseous condensations) plasma physics (solar wind and magnetic effects) radiative transfer theory (temperature of the solar environment) solid state physics (behaviour of solids) chemistry (production of compounds) atomic physics (isotopic properties) geology (behaviour of solids under pressure) and many others. It is impossible to become an expert in all of these, and specialisation is inevitable, leading to communicational problems. The cosmochemist wants a single simple dynamical model of the situation so that he can apply his chemistry while the dynamicist wants a single chemical scenario so that he can perhaps build a computer model.

The task I have set myself, and inevitably I shall fail to achieve this, is to gather together the material from these diverse fields, to present it in a way which is comprehensible and to stress its cosmogonic importance. This I do by giving separate sections to each important fact, the ordering of the sections being of no significance.

*Table 1.*

| Body | Mass (Earth = 1) | Inclination of orbit to ecliptic | Mean distance from Sun (A.U.) | Eccentricity |
|---|---|---|---|---|
| Mercury | 0.055 | 7° 0′ | 0.39 | 0.2056 |
| Venus | 0.815 | 3° 24′ | 0.72 | 0.0068 |
| Earth | 1.000 | - | 1.00 | 0.0167 |
| Mars | 0.108 | 1° 51′ | 1.52 | 0.0934 |
| Jupiter | 317.8 | 1° 18′ | 5.20 | 0.0485 |
| Saturn | 95.15 | 2° 29′ | 9.55 | 0.0557 |
| Uranus | 14.54 | 0° 46′ | 19.2 | 0.0472 |
| Neptune | 17.23 | 1° 46′ | 30.1 | 0.0086 |
| Pluto | 0.003 | 17° 10′ | 39.5 | 0.0250 |

In the last section I present a possible scenario for the process of planetary formation. Most parts of the scenario have already been proposed in isolation in the other theories. What I have done is to take the best parts of a number of theories and joined them together into one, occasionally reversing the temporal order of events.

## The Present Solar System

### 1. The Angular Momentum Distribution

It has long been recognised that there is a large discrepancy between the amount of specific angular momentum residing in the sun, and that in the planets, roughly 99% of the angular momentum resides in 0.15% of the mass. Indeed this single feature has been the central theme of a number of theories. It is however, not a fact which should be considered in isolation. With measurement of the interstellar medium becoming more common, it is clear that the planets have a roughly similar amount of specific angular momentum to that of dark clouds in the interstellar medium (in the region of $10^{20}$ $cm^2/s$). The problem is not therefore one of the distribution of angular momentum but rather of why the sun is rotating so slowly. The first question is obviously whether the sun is rotating differently from other stars. This was investigated by McNally (1965). His plot of angular momentum against mass for various stellar classes is shown as Fig. 1. This indicates that stars of spectral type A or earlier were fast rotators, while late type stars (including the sun) are slow rotators. It is thus not just a question of why the sun rotates slowly, but rather of why do all late type stars rotate slowly. In the sixties, the stock answer was 'because they have planetary systems'. However, there is another explanation. The discontinuity at type A/F occurs at just the spectral type where deep convective zones are developing in the stars. Such convective zones can drive a stellar wind of the same type as the well observed solar wind. In the case of the sun, the solar magnetic field interacts with the wind in such a way that the wind co-rotates with the sun out to about 20 solar radii. Because of this, the wind becomes an effective transporter of angular momentum from the sun to interstellar space. (A full description of this phenomenon and other aspects of the solar wind is given by Hundhausen 1972). According to this view, stars with a convective zone i.e. late type stars, will become slow rotators. Additional support for this view comes from the observation (Kraft 1967) that G-type stars in the two young clusters, the Pleiades and the Hyades, are rotating somewhat faster than the sun. The sun being older, has lost more angular momentum via the wind and rotates slower. Indeed the deduced rate of decrease in rotation is in agreement with calculations using the solar wind and it seems that the slow rotation of the sun may be irrelevant to the planetary formation process and is rather a natural consequence of normal stellar evolution.

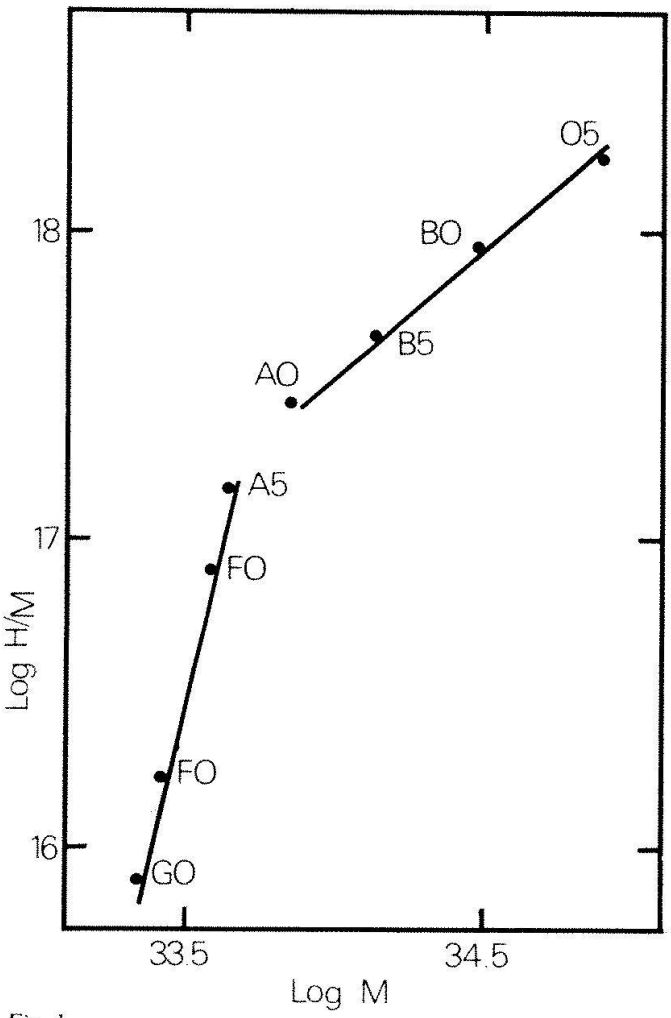


*Fig. 1.*

### 2. The Inturnal Structure of Jupiter and Saturn

The discovery that Jupiter and Saturn both emit more radiation than they receive from

the sun led to the formulation of evolutionary models where these planets contract under gravity from some large initial state, the excess energy now observed coming from the very slow current rate of contraction. Early models (e.g. Donnison and Williams 1974) showed that rough agreement between this theory and observations existed. With the advent of space probes, our knowledge of these planets has increased considerably, and from moment of inertia calculations, we now have a knowledge of the central density as well as the mean density of these planets. Static models (i.e. those ignoring contraction) found it difficult to reconcile this data with a homogeneous solar type composition for the planets, and (Stevenson 1978) a large rocky core was found to be necessary. The evolutionary modelists also found it easier to match theory and observation if a rocky core was adopted. The topic was recently reviewed by Grossman et al. (1980) and in Table 2, the current estimates for the mass of the rocky core is given for both Jupiter and Saturn.

*Table 2.*

| Author | Jupiter (Earth Masses) | Saturn (Earth Masses) |
|---|---|---|
| Pololak (1977, 1978) | 16–18 | 21–25 |
| Slattery (1977) | 14–16 | 15–17 |
| Hubbard & McFarlane (1980) | ~15 | ~15 |
| Grossman (1980) | 19–20 | 19–20 |

There are two important and distinct points arising from this development:

(a) The need for a rocky core of about 20 $M_\oplus$ inside both Jupiter and Saturn implies that neither of them has a cosmic abundance as this requires them to have a mass of about 6000 $M_\oplus$. At some stage they must therefore have either not accumulated all the available hydrogen and helium or they must have lost these gases. A corollary is that the mass of any pre-planetary nebula has to be at least 15,000 $M_\oplus$, or close to $10^{32}$ g.

(b) Jupiter and Saturn were of larger radius in the past. Of course the evolutionary models are insensitive to the earlier stages of contraction and so one cannot deduce what the initial radius was. However, any large radius is consistent with the observations.

## 3. The Composition of Terrestrial Planets

There are of course no direct measurements of the internal composition of the terrestrial planets. However, their mean densities, given in Table 3, show considerable variation. Some of this difference arises because of the different levels of compression present in the different planets, but even when account of this is taken, major differences exist.

*Table 3.*

| Planet | Mean Density ($gm/cm^3$) | Uncompressed density |
|---|---|---|
| Mercury | 5.4 | 5.4 |
| Venus | 5.25 | 4–4.5 |
| Earth | 5.52 | 4–4.5 |
| Mars | 3.94 | 3.7–3.8 |
| Moon | 3.34 | 3.3 |

The only abundant element which has a significantly different molecular weight from other abundant elements is iron. Iron can also appear as a mineral with other elements with vastly different densities, ranging from metallic iron with a density of 7.6 $gm/cm^3$ through Troilite (FeS) with a density of 4.7 $gm/cm^3$ to Fayalite ($Fe_2SiO_4$) with a density of about 3 $gm/cm_3$. It therefore seems obvious that the differences in uncompressed densities between the different planets is explainable in terms of the amount and form of iron present. Both of these depend directly on the environment (pressure and temperature) at the time of the condensation of the terrestrial material.

There are two distinct lines of thought. One argues that initially all the grains present in the solar neighbourhood were vaporized and that as the temperature there decreases, so materials condense out at the appropriate temperature. Thus, Mercury, being the hottest acquires mostly metallic iron; Mars, the coolest, is formed from Troilite and Iron-Silicates, while the Earth and Venus form from an intermediate mixture. This scenario obviously requires a source of heat of sufficient strength to vapourize all the grains. When Hayashi (1961) proposed a high luminosity initial pre-main sequence phase for stellar evolution, it appeared that the source

of heat had been identified. However, it has now become apparent that the early evolutionary stages will not be in hydrostatic equilibrium so that Hayashi's model is invalid. Larson (1972) shows that the Sun was never more luminous than about 10 times its present luminosity, too little to generate the high temperature in the planetary neighbourhood. Other possible sources of heat may be collisions, or friction and the contraction of the gas component in the preplanetary phase. It is very difficult to quantify these suggestions and so I will simply leave them as possibilities for now.

An alternative view for the whole process suggests that initially the ambient material was cool and that all the material capable of condensing at about 200 K had condensed. This produces a basic material of composition similar to Cl Carbonaceous chondrites (see later), that is a material commonly found in bodies which are thought to be primitive. Heating (to a lesser extent than in the first view) during the process of accumulation into planets in the presence of carbon, acting as a reducing agent, modifies the basic composition, turning iron oxides into metallic iron and outgassing CO and $CO_2$, this process again being more pronounced near the Sun.

It is possible to match the planetary composition in either model. The point to remember is that either an initial hot phase is called for, or a subsequent heating in the presence of reducing environment.

## 4. Meteorites

Meteorites produce a sample of interplanetary material which can be studied in the laboratory. It is important to realize that it does not give a random sample of the solar system. The preponderance of meteorites originates from a few asteroids and comets with near Earth-grazing orbits. Nevertheless, since both asteroids and comets produce a very inactive environment, they can yield information regarding parts of the system at an early epoch. Meteorites can be subdivided into four main classes.

a) Carbonaceons chondrites

These meteorites are distinguished by having hard mineral aggregates about 1 mm long embedded in a matrix of earthy material. The mineral deposits come in two forms, chondrules (which tend to be spherical) and irregular forms. The matrix consist of a mixture of minerals which condense at low temperatures. It is usual to assume that the chondrules are condensed dropletts from a liquid state while the irregular forms are condensates from a vapour state. To obtain a liquid state, it is usually necessary to have a pressure higher than prevalent in the interplanetary medium.

b) Ordinary chondrites

These were called ordinary as they are the most common on Earth, presumably because a number of similarly composed asteroids are on earth crossing orbits. They are composed entirely of minerals like olivine and troilite which condense at fairly high temperatures.

c) Achondrites

Chemically these meteorites are very similar to igneous rock and most have been broken up by some violent event in the past.

d) Iron

As their name suggest, these are essentially composed of a nickel-iron alloy. In fact, a large number consist of two discrete metallic alloys arranged in a characteristic geometry, called the Widmanstätten structure. This structure develops during the slow cooling of the material, cooling from 800 K to 500 K occurring at a rate of only a few degrees per million years. Clearly, for such a slow cooling rate to occur, the meteorite must have been enclosed, or be part of, a much larger body.

Thus a number of points clearly emerge, namely that iron and minerals became segregated, that in some, the liquid state implies pressure, the iron meteorites implies slow cooling, the irregular minerals imply condensation from a vapour state. All this suggests the existence of a number of large parent bodies, heated during formation, and subsequently cooling very slowly, collisions leading to the existence of the present small pieces.

## 5. The Allende Meteorite

In addition to the general chemical features in meteorites discussed above, the study of isotopic ratios and their comparison with terrestrial and solar ratios have revealed a

number of anomalies. The Allende chondrite surpassed all others in the extraordinary mineralogy and isotopic anomalies associated with it. An anomaly involving oxygen was pointed out by Clayton (1973) but Allende is best known for the anomaly involving $Mg^{26}$ and $Al^{26}$ (Wasserburg et al. 1977). The half life for radioactive decay is short ($7 \times 10^5$ y) while the only known source is subsequent to a supernova explosion. Accordingly, there is a requirement that a supernova explosion occurred in the vicinity (ie. a few hundred parsecs at the most) of the solar system close in time to its origin. Indeed, Cameron and Truran (1977) have suggested that this supernova was the trigger for the process of star formation which lead to the existence of the Sun. Such a scenario is also discussed by Schramm (1978). I am sure Bochsler (1981) will discuss these points further and so I will leave this topic now.

6. The Mass and Density of Pluto

The discovery of the satellite to Pluto has enabled an accurate determination of its mass to be made. This turned out to be considerably smaller than all previous estimates with a value of about $1.5 \times 10^{25}$ g and leads to a density of about 0.5 g $cm^{-3}$. The most likely composition is thus methane ice, rather than water ice. The very low mass means that for all practical purposes we can regard the main part of the solar nebula (or whatever one wishes to call the solar envelope) as terminating with Neptune's orbit, that is about 30 A.U.

7. Dynamics and Computer Simulations

The development of computing hardware and computing technology have made it possible for models of pre-planetary situations to be developed and their evolution investigated. We are still a long way from being capable of producing a simulation which takes account of all the phenomena encountered in the real solar system but some progress has been made. For example, Greenberg et al. (1978) have shown that growth can occur within a distribution of matter consisting of an interacting family rotating about the Sun. In this simulation, orbital dynamics were ignored and the particles given a random velocity in addition to rotational velocity (ie a kinetic theory approach within a rotating box). Williams and Donnison (1973) were interested in the settling of a three dimensional distribution to a plane while orbital dynamics played an important part in Wetherill's (1978) simulation. In general terms these simulations succeeded in reaching their objectives (of necessity almost, otherwise they would not have been published), and it is becoming clear that growth into larger bodies is possible within a dust cloud. It has also become clear that growth is much faster if nucleii are assumed to exist for growth to occur around. One such simulation by Dole (1970) showed that the final product of a number of experiments produced a family of star systems, out of which it was impossible to pick out our own.
In the foregoing, I have discussed a number of new results and constraints. These are in addition to the standard list of constraints discussed in earlier reviews such as Williams and Cremin (1968). If he so desires, the reader can judge for himself how many of the theories described there are consistent with this new data. I shall now give a brief outline of a scenario which I think is consistent with most of the data. It contains no fundamentally new ideas – they have all turned up, though not all together, in previous theories.

## A Scenario for Planetary Formation

Star formation can be observed to be ongoing in complexes like the Orion. It is clearly therefore an event which occurs and I will not concern myself with the details of the triggering process, remembering the earlier discussion regarding the solar angular momentum. The formation of a nebula surrounding the protosun has also been extensively discussed in the literature (eg Cameron 1962) and I will not discuss it further here, but am rather more interested in the evolution within the nebula. However, it is just as well to establish the general characteristics of the nebula. By the arguments in 2) above, it must have a mass of the order of $10^{32}$ g and by 6) extend out to 30 AU. Since radially pressure and rotation balance gravity, while perpendicular in the

plane, only pressure balances gravity, it is relatively easy to obtain an estimate for the height of the disk as something of the order of $1/40$ of the radius. Thus the average density in the disk is of the order of $6 \times 10^{-12}$ g/cm$^3$. It should be noted that these are just values for information to give the general picture and should not be taken as quantitative estimates.

There exists a critical mass, known as the Jeans Mass (see Williams 1974) such that any mass larger than this critical mass, for a given mean density $\rho$ and temperature T, will fragment into elements with the critical mass. This is given by the expression

$$M_J = \left(\frac{5RT}{\mu G}\right)^{3/2} \left(\frac{4\pi p}{3}\right)^{-1/2} \qquad (1)$$

$\mu$ being the mean molecular weight, $R$ the gas constant and G the gravitational constant.

Substitution of numerical values into (1) shows that $M_J \sim 10^{32}$ g and so the nebula has no tendency to fragment due to the Jeans instability.

Even if this tendency had been present, fragmentation need not have occurred for the Sun has also a tidal disruptive force on any condensation. This is expressed by the Roche limit (or the distance of the inner Legrangian point from the Sun). Accordingly, a condensation with density $\rho$ (Williams 1975) can only exist external to a distance L given by

$$L = \left(\frac{9 M_\odot}{4\pi\rho}\right)^{1/3} \qquad (2)$$

For the given value of $\rho$, this gives 40 AU and so again no condensations could be expected.

In this situation, any non-volatiles that had been vapourized in the formation process will condense out, and together with any that had not been vapourized, will form grains within the nebula. Indeed, there is no reason to assume any condensation sequence drastically different from one of those described by Wood (1979) or Anders and Owen (1977). By the mechanism of gravitational segregation, these grains will settle to the mid plane of the disk as for example in Williams and Handbury (1974) (a poor nebula model but the general principle is clear there). Within this disk, some agglomeration into larger bodies may occur. At this point I diverge from the popular picture, partly as a result of the voyager pictures of the rings of Saturn. I suspect that the family of large rocks would tend to set up all kinds of resonance and that this together with the tendency for orbits to circularize, results in a very long time scale for accumulation.

I assume that the Allende meteorite gives a clue as to the next event, namely the occurrence of a supernova explosion in the solar neighbourhood and which was responsible for injecting $Al^{26}$ into the solar system. By the snowplough effect, it also pushed into the system intervening parts of interstellar space, polluted by a lifetime of other supernova and stellar ejections, thus accounting for the isotopic anomalies. It is generally assumed that the shock wave following a supernova explosion can trigger star formation (e.g. Cameron (1978), Elmegreen and Lada (1978)). It does this through the shock wave compressing the gas so that a gravitationally stable unit is formed.

It is somewhat difficult to obtain the degree of compression to be expected, but we can obtain a rough estimate. Using standard conservation equations across a shock with $x = \rho_2/\rho_1 \equiv$ compression ratio and u as the ratio of shock to sound speed, we obtain:

$$u^2 = x\left(\frac{\gamma}{\gamma-1}\right)\left(\frac{x^\gamma - 1}{x-1}\right), \qquad \gamma = 1,$$

$$u^2 = x, \qquad \gamma = 1 \mu^2 = x, \qquad \gamma = \%$$

For a typical supernova, $u \sim 5 \times 10^3$, so that with $\gamma = 5/3$

$$x \sim \frac{2}{5} u^{6/5}, \quad \text{or} \quad x \sim 10^4.$$

with such a compression, we see from (1) that condensations with a mass of $10^{32} \times (10^4)^{-1/2} \equiv 10^{30}$ g would be stable, while from (2), these could exist beyond a distance of $40 \times (10^4)^{-1/3} = 1.8$ AU.

Thus there would be no change in the terrestrial planet region, but external to this, what might be termed giant gaseous protoplanets would be formed. The evolution of such

planets has been discussed in the literature. For example, McCrea and Williams (1965) showed that the settling grains would form a core while Donnison and Williams (1974) showed that such objects could contract to become Jupiter and Saturn. Handbury and Williams (1975) even suggested that the segregation of ammonia and methane grains liberated enough energy to drive away hydrogen and helium and so form the outer planets.

Of course, these giant gaseous protoplanets may also be involved in collisions. This will result in the rapid heating of them followed by a cooling under pressure which may have relevance to meteoritic evolution. Indeed, the existence of a liquid phase is more than likely. Another consequence of collisions is that protoplanets could cross into the terrestrial planet region. As soon as they do this, they become totally unstable and will be disrupted. However, any nonvolatile agglomerations within them will survive, to be injected, on initial eccentric orbits, into the non-volatile disk in the terrestrial planet region where they immediately serve as a nucleus for accretion, terminating in the terrestrial planets. In addition, of course, the asteroid belt region will have been polluted with minor agglomerations that had not reached the centre of a protoplanet when it disrupted.

It may also be possible to account for the major satellites of the system in terms of young cores lost during a glancing collision or interaction between two protoplanets.

## Conclusions

In the foregoing I have attempted to describe our current state of knowledge regarding the planetary system in a fairly objective way. The use I have made of these facts in the preceding section is very subjective. Other authors reconcile these facts with either a solar nebula with accretion from a disk occurring throughout or with a protoplanetary picture in which protoplanets form the preplanetary stage of all the planets. What I have produced is a hybrid qualitative model, which to my mind is a logical deduction from the facts.

I would like to thank E. Anders, G. Arrhenius, W.K. Hartmann and G. Wetherill for discussions and correspondence which have directly or indirectly had influence on the formulation of my scenario.

## Summary

During the last decade our knowledge concerning the individual members of the solar system has considerably increased, and a review of this recent data is given, and its implication for the process of planetary formation discussed. A scenario for the process, taking account of all these developments is also given.

## References

Anders, E., and Owen, T. 1977, Science, 198, 453.
Bochsler, P., 1981, This volume.
Cameron, A.G.W., 1962, Icarus 1, 13.
Cameron, A.G.W., 1978, in Origin of the Solar System Ed. Dermott, J. Wiley.
Cameron, A.G.W., and Truran, J.W., 1977, Icarus, 30, 447.
Clayton, D.D., 1977, Earth, Plan. Sci. Lett. 36, 381.
Dermott, S.F., 1978, The Origin of the Solar System J. Wiley Pub. Co.
Dole, S.H., 1970, Icarus, 13, 494.
Donnison, J.R., and Williams, I.P., 1974, Astroph. Sp. Sci, 29, 387.
Elmegreen, B.G., and Lada, C.J., 1977, Ap. J. 214, 725.
Gehrels, T., 1978, Protostars and Planets, Univ. of Arizona Press.
Greenberg, R., Hartmann, W.K., Chapman, C.R., and Wacker, J.F., 1978, in Protostars and Planets Ed. Gehrels, Univ. of Arizona Press.
Grossman, A.S., Pollack, J.B., Reynolds, R.T., and Summers, A.L., 1980, Icarus, 42, 358.
Handbury, M.J., and Williams, I.P., 1975, Astroph. Sp. Sci, 38, 29.
Hayashi, C., 1961, Pub. Astron. Soc. Japan, 13, 450.
Hubbard, W.B., and McFarlane, J.T., 1980, J. Geoph. Res., 85, 225.
Hundhausen, A.J., 1972, Solarwind and Coronal Expansion, Springer-Verlag.
Kraft, P.R., 1967, AP. J., 150, 551.
Larson, R.B., 1972, Mon. Not. R. astr. Soc, 157, 121.
McCrea, W.H., and Williams, I.P., 1965, Proc. Roy. Soc, A287, 143.
McNally, D., 1965, The Observatory, 85, 166.
Podolak, M., 1977, Icarus, 30, 155.
Podolak, M., 1978, Icarus, 33, 342.
Slattery, W., 1977, Icarus, 32, 58.
Reeves, H., 1972, The Origin of the Solar System, C.N.R.S. Paris.

Schramm, P.N., 1978, in Protostars and Planets, Ed. Gehrels, Univ. of Arizona Press.
Stevenson, D.J., 1978, in Origin of the Solar System, Ed. Derriott, J. Wiley. Pub. Co.
ter Haar, D., and Cameron, A.G.W., 1963, in Origin of the Solar System Ed. Jastrow and Cameron, Academic Press.
Wasserburg, G.J., Lee, T., Papanastassiou, P.A., 1977, Geoph. Res. Lett, 4, 299.
Wetherill, G., 1978, in Protostars and Planets Ed. Gehrels, Univ. of Arizona Press.
Williams, I.P., 1974, Origin of the Planets A. Hilger Pub. Co.
Williams, I.P., and Cremin, A.W., 1968. Qt. Jl. R. astr. Soc., 9, 40.
Williams, I.P., and Donnison, J.R., 1973, Mon. Not. R. astr. Soc., 165, 295.
Williams, I.P., and Handbury, M.J., 1974, Astroph. Sp. Sci, 30, 215.
Wood, J.A., 1979, The Solar System, Prentice-Hall.

*Address of the author:*

Dr. Iwan P. Williams
Queen Mary College (London University)
Dept of Applied Mathematics
Mile End Road
London E14NS (England)

# Isotopic Research Related to the Origin of the Solar System

Peter Bochsler

## Introduction

Until a decade ago it was generally assumed that the solar system has condensed from a hot, homogeneous nebula. Elemental heterogeneities as found between different classes of meteorites were interpreted as consequences of local elemental fractionation within the nebula following the onset of condensation. Some isotopic heterogeneities in meteorites, such as $^{129}$Xe-anomalies originating from the decay of $^{129}$I ($T_{1/2} = 15.7 \cdot 10^6$ y), could be explained to be variations in time elapsed between the last nucleosynthetic event and the solidification of the different fragments. All investigations for other isotopic heterogeneities, due to incomplete mixing of the solar nebula, had led to the conclusion, that the solar system material was indeed well mixed at least to the 1‰ level.

In the beginning of the last decade this view of the history of the early solar system was gradually changed: Black (1972) argued that the isotopic pattern of neon as observed in stepwise heating experiments on the Orgueil-meteorite could be due to an admixture of an 'extrasolar component' rich in $^{22}$Ne, in the following called 'Neon-E'. Clayton et al. (1973), on the other hand, found that oxygen in meteorites and lunar samples could not simply be explained by fractionation of an originally homogeneous reservoir but rather by fractionation and variable admixture of a component rich in $^{16}$O.

In the following Lee et al. (1976) investigated several inclusions in the Allende-meteorite which were considered to be early condensates from the solar nebula and found anomalies of $^{26}$Mg correlated with the aluminium content of the samples. They could convincingly interpret this anomaly to be due to 'fossil' $^{26}$Al ($T_{1/2} = 0.72 \cdot 10^6$ y) which had been incorporated into their samples before decay. From this Lee et al. (1977) could conclude that the inclusions of Allende must have been formed only a few million years after a nucleosynthetic event which produced $^{26}$Al.

Since 1977 isotopic anomalies have been found in many other chemical elements. More recent reviews have been given by Clayton (1978) and by Wasserburg et al. (1980).

The following short review will be limited to the cases of four elements: Hydrogen, nitrogen, neon, and magnesium. In this context we discuss some consequences of recent discoveries for our understanding of the early history of the solar system.

## Deuterium

The solar wind is thought to reflect quite closely isotopic ratios in the outer convective zone of the sun and hence, except for the light elements Li and Be, the composition of the sun at the time of ignition of nuclear burning. Also deuterium does not survive the temperatures at the bottom of the convective zone and is converted to $^3$He.

From the isotopic ratio $^3$He/$^4$He in the solar wind as determined with the Apollo foil experiments (Geiss et al., 1972); Geiss and Reeves (1972) could set a firm upper limit to the deuterium content of the primeval sun. The best estimate for the D/H-ratio in the early sun is now $2.0 \cdot 10^{-5}$ (Geiss and Bochsler, 1978). As can be seen from figure 1, the solar value agrees well with the one for interstellar atomic hydrogen (York and Rogerson, 1976 and Vidal-Madjar et al., 1977) and possibly for Jupiter which is controversal (Combes et al., 1978; Trauger et al., 1977). On the other hand, the solar value is well below the terrestrial and the meteoritical values. Special fractions of the Chainpur

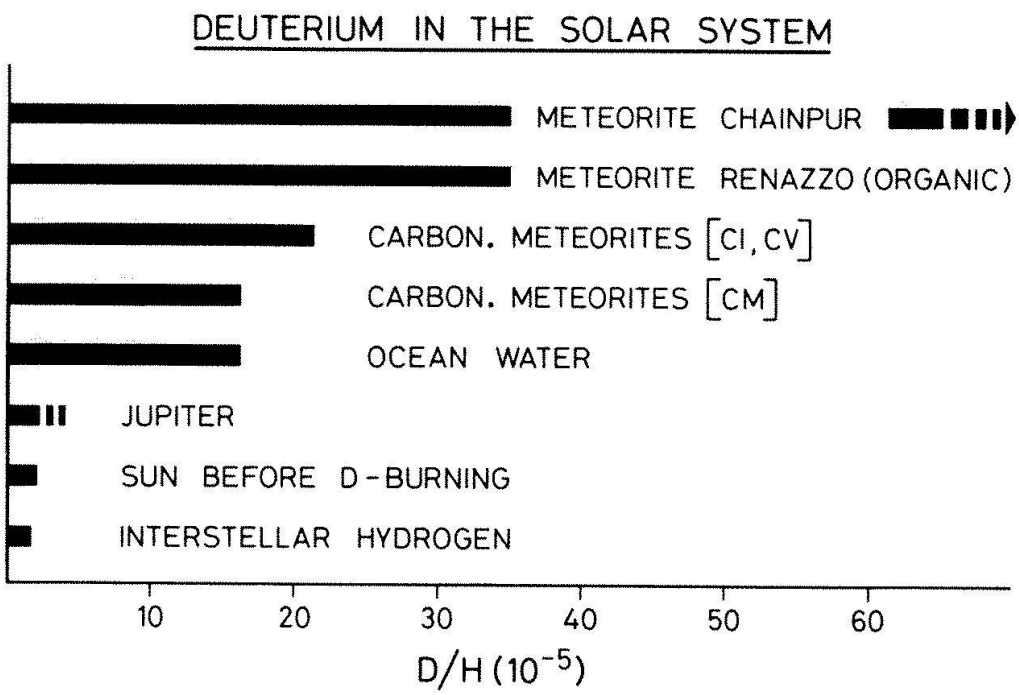



*Fig. 1.* Compilation of deuterium abundances in the solar system. (See text for references.)

and Renazzo meteorites still show much stronger deuterium enrichments relative to the primordial solar composition (Robert et al., 1979; Kolodny et al., 1980).

The information on the primeval D/H-ratio in the sun is of special importance for modeling the 'big-bang'. It poses limits to the present baryon density $\rho_B$ and hence the deceleration parameter $q_0$. $\rho_B$ derived from the D/H-value in the early sun and the interstellar gas is compatible with densities derived by other methods, leading to the conclusion that our universe is open.

Geiss and Reeves (1981) have pointed out that the strong enrichment of deuterium relative to normal solar hydrogen as observed in the Chainpur and Renazzo meteorites cannot be due to equilibrium reactions nor to kinetic effects since temperatures below 200 K would have been required to produce such large isotopic shifts. At such low temperatures the equilibrium time would be of the order of $10^{30}$ years. Geiss and Reeves (1981) therefore suggest that the enrichment has taken place in a similar way as postulated for the strong deuterium enrichments observed in molecular clouds, i.e. via ion-molecule reactions which would proceed at a sufficiently fast rate. It is not clear whether the presence of strong D-enrichments in the organic fractions of carbonaceous meteorites and also in terrestrial hydrogen excludes the possibility that all solar system material had been heated to temperatures above 2000 K, disrupting existing molecules, previous to formation of the planetary bodies. This conclusion will arise later from different observations; however, as indicated in the introduction, the formation of the solar system most probably was completed within a few million years while the lifetime of molecules in clouds might be of the order of $10^7$ years, which would leave some more time for D-enrichment. At present it looks more likely that the organic matter in Chainpur and Renazzo has survived the process of formation of the solar system and is a witness of the existence of the presolar molecular cloud.

## Nitrogen

In the previous section we have shown that deuterium anomalies in the solar system are entirely due to chemical and not to nuclear effects.

In a recent paper (Geiss and Bochsler, 1982) we have investigated evidence for the presence of an anomaly in the isotopic composition of nitrogen in different planetary bodies and the sun. Our motivation for this study was the ongoing discussion on the question whether the isotopic composition of nitrogen at the solar surface has changed by 30% during the history of the sun or not: Kerridge

(1975) discovered a strong anticorrelation of the $^{15}N/^{14}N$-ratio in surface implanted nitrogen in lunar soil with the cosmic-ray produced $^{21}Ne$ in the soil. It was already known before, that most of this surface implanted nitrogen (as an element) must be due to the solar wind. Clearly the observed anticorrelation of $^{15}N$ with $^{21}Ne$ could not be the consequence of spallation produced $^{15}N$ in lunar soil since then one would rather expect a correlation. Kerridge (1975) explained this correlation as evidence for a secular increase of the $^{15}N/^{14}N$-ratio in the solar wind by approximately 30% during the last $4 \cdot 10^9$ years. Becker and Clayton (1975) suggested another explanation for the apparent secular change in the isotopic composition of surface implanted nitrogen: They attributed the secular change to a varying admixture of a second component with a low $^{15}N/^{14}N$-ratio which was emanating from the lunar interior. However no trace of such a component in lunar rocks was found and this hypothesis was abandoned. Despite the difficulty to explain a secular increase of the $^{15}N/^{14}N$-ratio in the outer convective zone of the sun by nuclear reactions at the solar surface this remained the most favoured hypothesis.

In our paper (Geiss and Bochsler, 1982) we have reassessed the evidence against nuclear production of $^{15}N$ and destruction of $^{14}N$ at the solar surface and provided several further arguments against it. We conclude that the apparent secular trend in the isotopic composition of nitrogen trapped in lunar soils must be due to an admixture of a component of light nitrogen ($\delta^{15}N \lesssim -400‰$)*) from a source with decreasing yield during the age of the regolith. Evidence for such a component is found in some meteorites: If one assumes the hypothesis of a secular increase at the solar surface to be true, one is led to the conclusion that solar nitrogen, and hence the bulk of the solar system nitrogen consists of a light isotopic mixture. According to our explanation however, bulk solar system nitrogen is a heavy isotopic mixture. This idea is supported by the fact that the class of meteorites which is least depleted in volatiles, the carbonaceous chondrites of class 1, contain heavy nitrogen ($\delta^{15}N \cong +40‰$). Meteorites depleted in volatiles tend towards lighter nitrogen. It appears that the postulated anomalous light component is ubiquitous in the solar system and could possibly also be used to explain the difference between the nitrogen composition of the sun and the terrestrial atmosphere. Principally it is possible, that the anomalous light component could have been depleted in $^{14}N$ by kinetic effects from an originally heavier component. In our paper we favour the idea that this light component stems from a different nucleosynthetic process and was incorporated into refractory phases.

*) We here use the 'delta-notation':

$$\delta^{15}N_{sample} = \frac{(^{15}N/^{14}N)_{sample} - (^{15}N/^{14}N)_{air}}{(^{15}N/^{14}N)_{air}} * 1000‰$$

## Neon

When Black and Pepin (1969) investigated neon in carbonaceous meteorites by the stepwise heating technique they found that always one component released at 1000 °C was enriched in $^{22}Ne$ relative to $^{20}Ne$ and other temperature steps by factors up to 3. Black (1972) postulated that an extrasolar component – in the following called 'Neon-E' – must be present in carbonaceous meteorites. This hypothetical component should be rich in $^{22}Ne$ and it should be concentrated in a special mineral phase which releases neon at about 1000 °C. Eberhardt (1974) succeeded to isolate a Ne-E rich carrier phase from the Orgueil carbonaceous meteorite. Since then, the upper limits for $^{20}Ne$- and $^{21}Ne$-contents of Ne-E have been steadily lowered by further improvement of the experimental techniques. This is illustrated in a logarithmic three-isotope plot (figure 2) taken from Jungck and Eberhardt (1979). From this diagram it is evident that a component consisting essentially of pure (>99%) $^{22}Ne$ must be present in Orgueil. Eberhardt et al. (1979a,b) have shown that there are at least two carriers for Ne-E. This is shown in fig. 3 taken from Eberhardt et al. (1979b): The first carrier has a density of less than 2.3 g/cm$^3$. It releases Ne-E at temperatures below 900 °C and is poor in target elements for cosmogenic (= cosmic ray produced) neon such as Mg, Al, and Si, as can

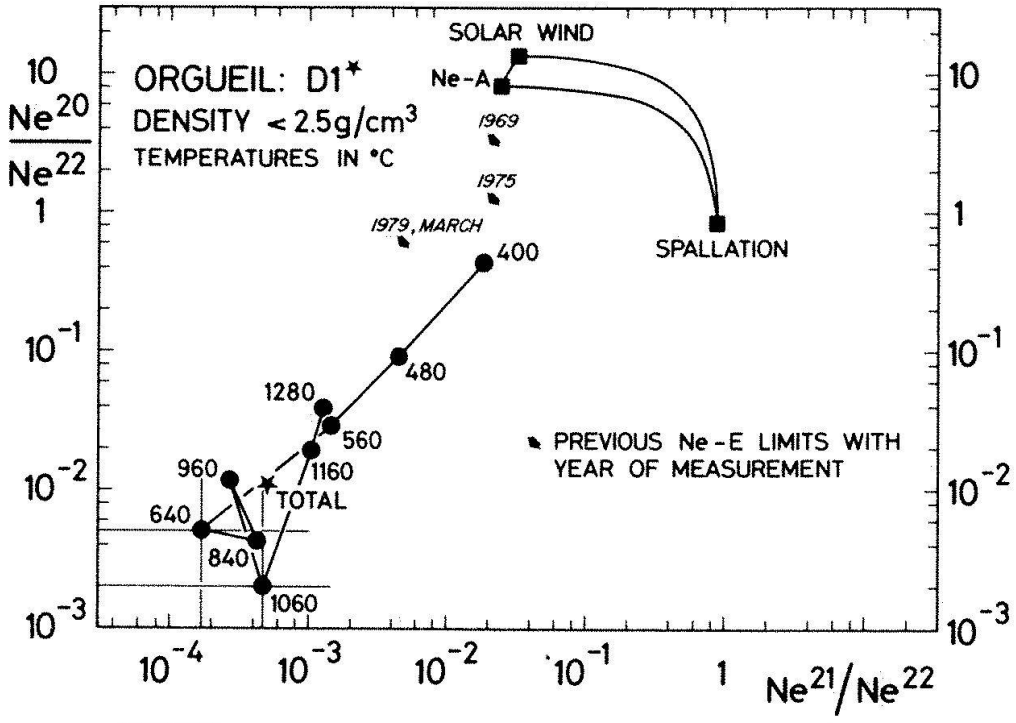



*Fig. 2.*
Neon three isotope diagram with the results of stepwise heating experiment on the Ne-E rich phase D1* separated from the C1 chondrite Orgueil.
(From Jungck and Eberhardt 1979)

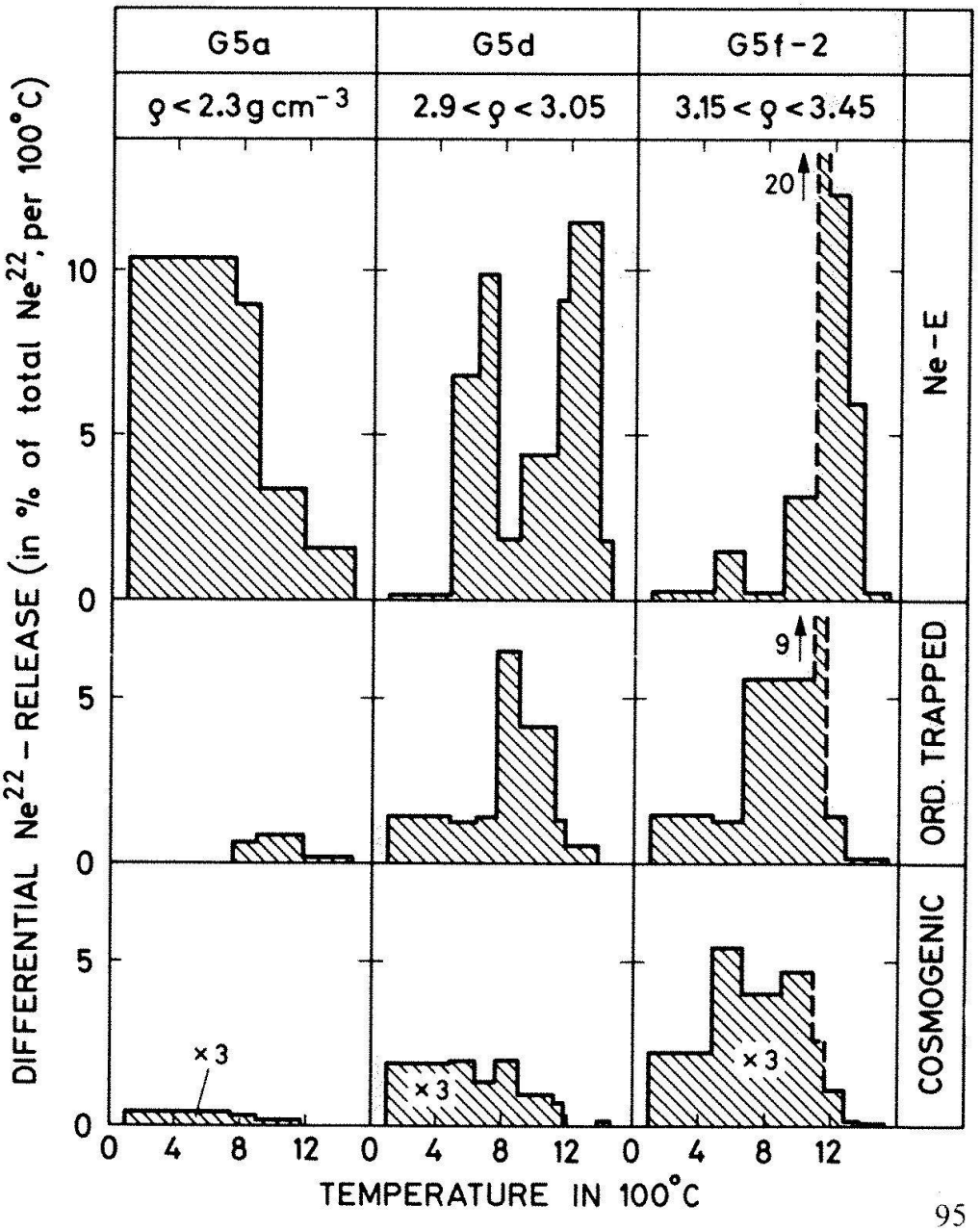



*Fig. 3.*
Differential gas release observed for three density separates from Orgueil. Shown are the absolute amounts of gas released per °C of temperature increase, and the areas in the histograms correspond to the gas concentrations.
(From Eberhardt et al. 1979a)

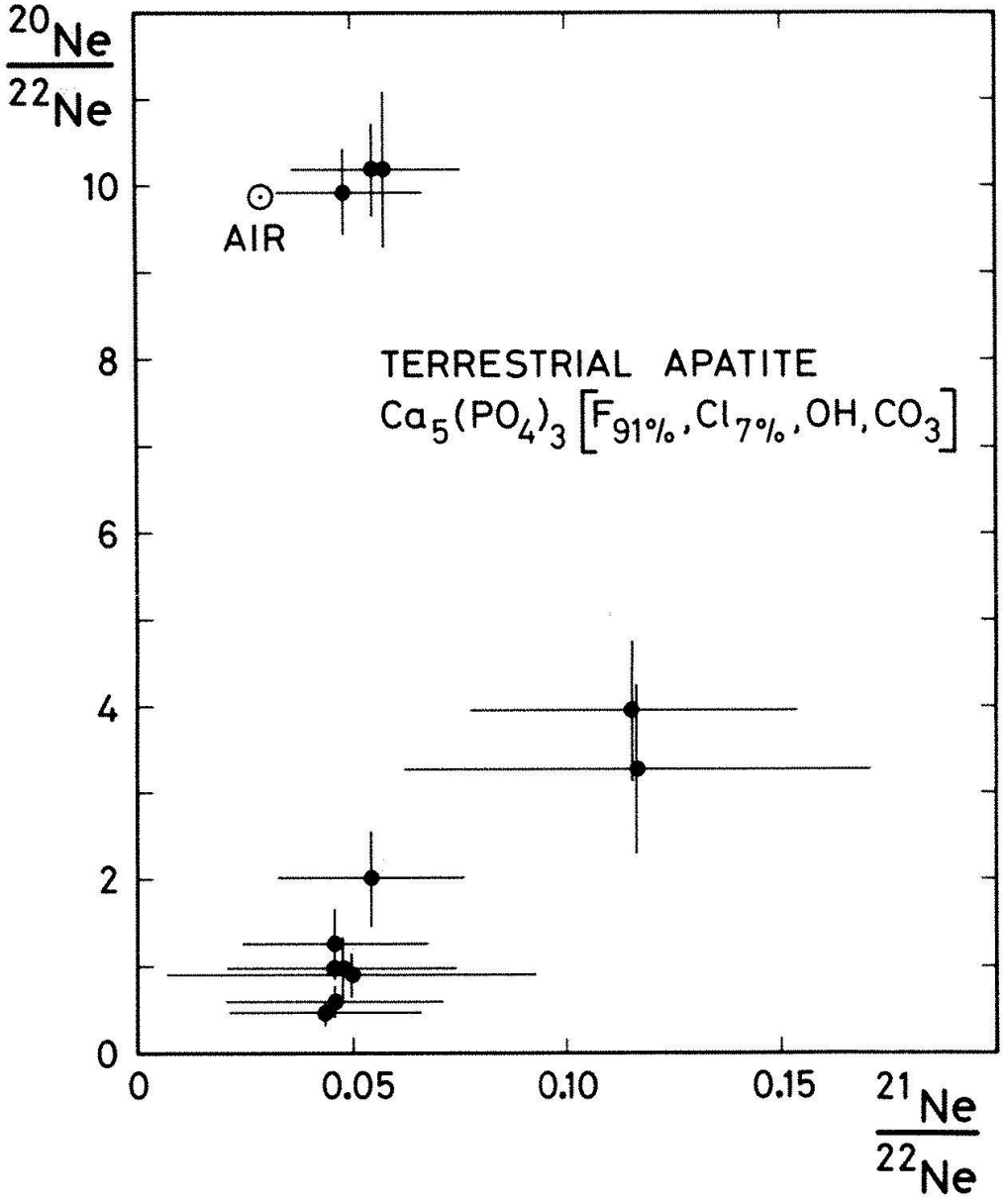



*Fig. 4.*
Neon three isotope diagram with results of stepwise heating experiment on a terrestrial fluorapatite.

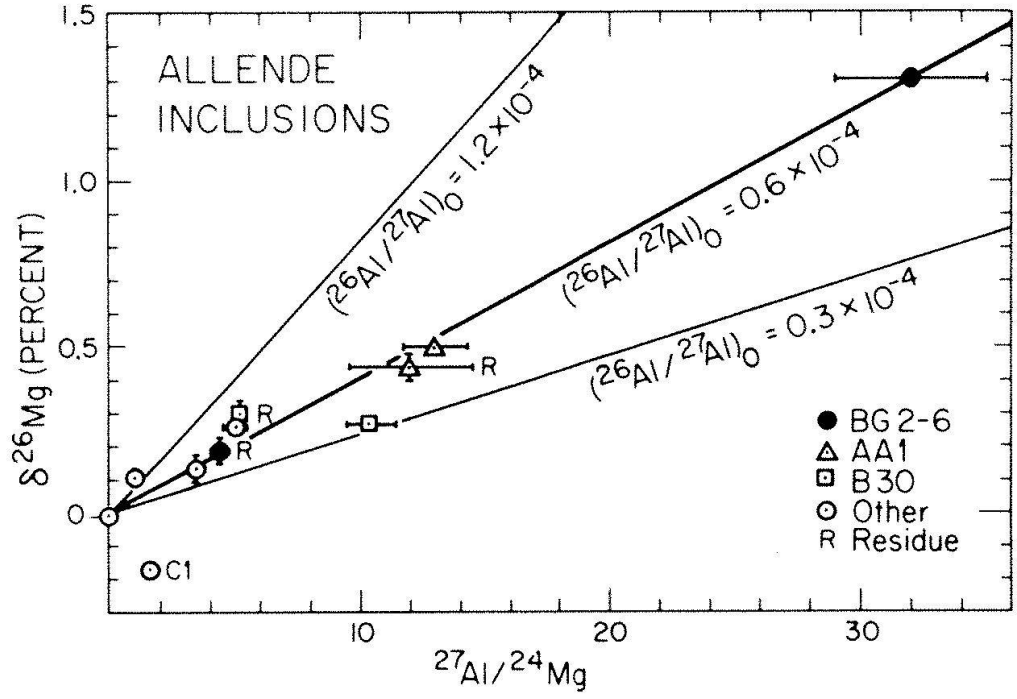



*Fig. 5.*
Al-Mg evolution diagram for Allende samples. The correlation line for BG2–6 yields $(^{26}Al/^{27}Al)_0 = 6 \times 10^{-5}$ and contrasts sharply with the correlation line for B30 which has essentially zero slope and much higher initial $^{26}Mg/^{24}Mg$.
(From Lee et al. 1976)

be derived from the low concentration of spallogenic $^{21}Ne$. The second carrier has a density between 3.15 and 3.45 g/cm$^3$. It releases Ne-E at temperatures above 900 °C and must contain some target elements for cosmogenic Ne. More recently Jungck et al. (1981) have provided strong evidence that the heavier carrier is, or must be closely attached to apatite. The fraction G5f-2 also contains some olivine which could provide the target elements for cosmogenic neon as visible in figure 3 (Eberhardt et al., 1981).

## Origin of Neon-E

Clearly, $^{22}Ne$ in such a degree of purity as obtained by Jungck and Eberhardt (1979) cannot be produced by fractionation of solar neon: If we assume Rayleigh distillation with a fractionation factor of $(1 - \sqrt{m_{20}/m_{22}})$ this would require a depletion of neon by a factor $10^{-65}$ in order to obtain the necessary concentration of $^{22}Ne$. The incorporation of Ne-E into the samples is still not understood, particularly the fact that apatite might be one of the carriers is puzzling: As Eberhardt et al. (1981) point out, $^{22}Ne$ can be produced by reactions of $\alpha$-particles with $^{19}F$. Unfortunately the cross section for this reaction is not known. By the stepwise heating technique we have investigated neon in an almost pure fluorapatite which contained approximately 3.4 weight percent F. The sample investigated contained a sizeable amount of Th and U which produced a large concentration of $^4He$. The neon data plotted in a conventional three-isotope plot (Fig. 5) indicate the presence of three components: The first component, being essentially atmospheric neon with $^{20}Ne/^{22}Ne = 10$ is released at temperatures below 600 °C. The second component is enriched in $^{22}Ne$ and contains also some $^{21}Ne$. The third is enriched in $^{21}Ne$ and $^{22}Ne$. The first component can be interpreted as atmospheric neon in small inclusions or dissolved in the crystal lattice, the second component is probably the result of the reactions of $\alpha$-particles from the U/Th-decay with fluorapatite. The third might be due to irradiation of a second unidentified mineral with a lower F/O-ratio than in fluorapatite present in the sample. From our measurement we can derive a production ratio for $^{21}Ne/^{22}Ne$ in fluorapatite of 0.030. Related to the abundances of $^{18}O$ and $^{19}F$ this would give a cross section ratio (integrated over the spectrum of $\alpha$-particles from U/Th-decay) of 1.5 close to the value of 1, assumed by Eberhardt et al. (1981).

Since the $^{19}F(\alpha,n)^{22}Na$ reaction has a higher coulomb-barrier and a higher threshold energy than $^{18}O(\alpha,n)^{21}Ne$, assuming a steeper energy spectrum of $\alpha$-particles would even increase the $^{21}Ne/^{22}Ne$-production ratio. Thus we are probably on safe ground if we conclude that Ne-E in apatite was not produced by irradiation of this mineral with energetic $\alpha$-particles. There are two more arguments against this: As Eberhardt et al. (1981) point out, no evidence for excess $^4He$ is found in their samples. In addition, in any astrophysical environment which generates energetic $\alpha$-particles one would also expect a certain amount of energetic protons. This would then produce spallogenic argon from calcium. In G5f Eberhardt et al. (1981) find only $1.2 \cdot 10^{-8}$ cm$^3$ spallogenic $^{38}Ar$. If G5f is essentially apatite, this amount could be produced within $10^6$ years by irradiation with galactic cosmic rays (Bochsler et al., 1969). This has to be compared with the exposure age of $10^7$ y determined by Jeffery and Anders (1970). Obviously there is no room for additional irradiation. Therefore Eberhardt et al. (1981) conclude, that Ne-E was most likely incorporated in interstellar grains in the form of $^{22}Na$. Since $^{22}Na$ has a half-life of only 2.6 years this incorporation must have taken place very soon after production of $^{22}Na$.

## Extinct Al-26

Schramm et al. (1970) have investigated several feldspar samples of meteorites, in which extinct $^{129}I$ ($T_{1/2} = 15.7 \cdot 10^6$ y) had been found before, for extinct $^{26}Al$ ($T_{1/2} = 0.72 \cdot 10^6$ y). No trace of this isotope was found and the authors concluded that $^{26}Al$ could not have been important as heat source in parent bodies of meteorites at the time of solidification of feldspar. After the discovery of anomalous oxygen in carbonaceous chondrites by Clayton et al. (1973) and after the finding of minerals considered to be 'high temperature condensates' in the Allende meteorite it

appeared to be worthwhile to investigate these minerals for extinct $^{26}Al$, despite the previous experience with other meteorites. Subsequently Lee et al. (1976) discovered a strong correlation of excessive $^{26}Mg$ with the Al/Mg ratio in several inclusions. It was obvious at once that not all inclusions fell on the same isochrone as can be seen in figure 5 which is taken from Lee et al. (1976). In the following, Lee et al. (1978) could show that in one single chondrule from Allende, coexisting mineral phases provided an isochrone with an initial $^{26}Al/^{27}Al$-ratio of $(5.1 \pm 0.6) \cdot 10^{-5}$. Lee et al. (1978) could therefore exclude the possibility that the isochrone was the result of a mixture of two phases with different amounts of inherited $^{26}Mg$ excess and came to the conclusion that $^{26}Al$ must have decayed in situ. It should be mentioned however, that not all early condensates of Allende contain fossil $^{26}Al$. Lee et al. (1979) discovered a hibonite inclusion ($CaAl_{12}O_{19}$) with Ca-anomalies but no trace of extinct $^{26}Al$. The authors point out that the question of the relation of $^{26}Al$ with other nuclear anomalies is still open.

Nevertheless two conclusions can be drawn: $^{26}Al$ can again be considered as an important potential heat source for planetary bodies. Secondly, the formation of the solar system must probably have started within a few million years after the last nucleosynthetic event.

## Conclusions

The observations by Lee et al. (1976, 1978), that the formation of the solar system took place only within a few million years after the last nucleosynthetic event, strongly supports present ideas on the close relation of nucleosynthesis and star formation. It is now generally accepted, that stars form within clusters. An initial trigger for the onset of star formation in a dense molecular cloud might be the passage through a galactic spiral density wave. Massive stars can contribute in two ways to the sequence of the following events: First, by formation of H II-regions and possibly by supernova explosions they can fragment the molecular cloud in which they have been formed (Silk, 1979) and hence propagate star formation. Secondly: Supernovae explosions inject freshly synthesized material into the protostellar gas. Such a last injection might have also contained $^{26}Al$, although $^{26}Al$ could as well have been produced in a red giant (Nørgaard, 1980). Another piece of evidence for the close relation of star formation and nucleosynthesis has been given by Reeves and Johns (1976) who showed that nucleosynthesis occurs in bursts which can be correlated to the passage of galactic spiral density waves through the galactic medium.

The survival of some volatile anomalies such as Ne-E requires that the anomaly-bearing grains have not been heated to temperatures above 1200 K. Thus we should expect that these grains, after their formation, could not have been exposed to strong heating of nearby massive stars or near supernovae explosions. We expect that the life expectancy of a dust grain in the environment of star formation can only be rather short so that only a very limited number of nucleosynthetic sources for Ne-E can be involved. It could therefore well be that in our view the importance of nucleosynthesis of refractory anomalies such as $^{26}Al$, in the era shortly before formation of the solar system is exaggerated due to the preferred survival of anomalies in refractory elements.

Theory of star formation shows that the presence of dust grains in collapsing molecular clouds is important, since grains are the most efficient cooling agent at temperatures of a few hundred K, thus keeping the collapse going on. The presence of volatile anomalies shows that grains indeed have been around during star formation and could even survive the initial stages of star birth.

The presence of $^{26}Al$ is of enormous importance for heat generation within planetesimal bodies. As Lee et al. (1977) have pointed out, as soon as bodies of a few 100 km are formed, the heat generated by the decay of $^{26}Al$ can bring temperatures in the interior of such a body to several thousand degrees. This might cause disruption of smaller bodies or segregation of different phases in larger bodies where the interior gravitational field is sufficiently strong to keep viscous melts moving.

Unsegregated large bodies can only form if they incorporate little or no $^{26}Al$, i.e. if they

form practically without Al (and other refractory elements) or late ($10^7$ to $10^8$ y after synthesis of $^{26}Al$).
Many questions remain unsettled, one of the most intriguing problems is the lack of correlation between many anomalies. It appears that e.g. the neon anomaly is decoupled from all the other anomalies except maybe xenon and krypton. On the other hand there is a clear clustering of anomalies in refractory elements in some inclusions of the Allende meteorite. As pointed out earlier, the distribution of anomalies is not uniform in these inclusions. It appears that the early solar system was chemically and isotopically heterogeneous and contained many distinct reservoirs which themselves result from incomplete mixing of different nucleosynthetic products. Clearly, the field of nuclear and chemical anomalies in the solar system will continue to evolve rapidly in the next years, some of the questions raised will be solved, many new questions will be open.


## Acknowledgements

I gratefully acknowledge many suggestions and stimulating discussions with Drs. J. Geiss and P. Eberhardt. Thanks are due to Dr. A. Stettler for his participation in the acquirement and interpretation of the apatite data. There have also been several instructive discussions on the latest developments in the field of neon-E with M.H.A. Jungck and F.O. Meier. I thank L. Reichert, who carefully typed this manuscript. This work is supported by the Swiss National Science Foundation.



## Abstract

A decade ago it was generally accepted that the solar system has condensed from a chemically homogeneous cloud and that the cloud material has been heated to temperatures above 2000 K shortly before the formation of the bodies of the solar system. Recent investigations in the field of isotopic research have demonstrated that this picture no longer holds: Many chemical elements in planetary bodies contain isotopic anomalies, showing that matter has not been completely homogenized previous to formation of the solar system. In the case of some elements we show some implications of the recent discoveries on our understanding of the early history of the solar system.


## References


Becker, R.H., and R.N. Clayton, Nitrogen abundances and isotopic compositions in lunar samples. Proc. Lunar Sci. Conf. 6th (1975) 2131–2149.

Black, D.C., On the origins of trapped helium, neon and argon isotopic variations in meteorites-II. Carbonaceous meteorites. Geochim. Cosmochim. Acta *36* (1972) 377–394.

Black, D.C., and R.O. Pepin, Trapped neon in meteorites-II. Earth Planet. Sci. Lett. 6 (1969) 395–405.

Bochsler, P., P. Eberhardt, J. Geiss, and N. Grögler, Rare gas measurements in separate mineral phases of the Otis and Elenovka chondrites. In Meteorite Research, D. Reidel Publishing Company, Dordrecht, Holland (1969) 857–874.

Clayton, R.N., Isotopic anomalies in the early solar system. Ann. Rev. Nucl. Part. Sci. 28 (1978) 501–522.

Clayton, R.N., L. Grossman, T.K. Mayeda, A component of primitive nuclear composition in carbonaceous meteorites. Science 182 (1973) 485–488.

Combes, M., T. Encrenaz, and T. Owen, On the abundance of deuterium in Jupiters atmosphere. Astrophys. J. 221 (1978) 378–381.

Eberhardt, P., A neon-E rich phase in the Orgueil carbonaceous chondrite. Earth Planet. Sci. Lett. 24 (1974) 182–187.

Eberhardt, P., M.H.A. Jungck, F.O. Meier, and F. Niederer, Neon-E. New limits for isotopic composition. Two host phases? Lunar and Planet. Sci. 10 (1979a) 341–343.

Eberhardt, P., M.H.A. Jungck, F.O. Meier, and F. Niederer, Presolar grains in Orgueil: Evidence from Neon-E. Astrophys. J. 234 (1979b) L169–L171.

Eberhardt, P., M.H.A. Jungck, F.O. Meier, and F.R. Niederer, A Neon-E rich phase in Orgueil: Results obtained on density separates. Geochim. Cosmochim. Acta 45 (1981) 1515–1528.

Geiss, J., and H. Reeves, Cosmic and solar system abundances of Deuterium and Helium-3. Astron. Astrophys. 18 (1972) 126–132.

Geiss, J. and P. Bochsler, On the abundances of rare ions in the solar wind. Proc. 4th Solar Wind Conf., Burghausen (1978).

Geiss, J. and P. Bochsler, Nitrogen isotopes in the solar system. Geochim. Cosmochim. Acta 46 (1982) 529–548.

Geiss, J. and H. Reeves, Deuterium in the solar system, Astron. Astrophys. 93 (1981) 189–199.

Geiss, J., F. Bühler, H. Cerutti, P. Eberhardt, and Ch. Filleux, Solar wind composition experiment, Apollo 16 Prelim. Sci. Rep. NASA SP-315 (1972).

Jeffery, P.M. and E. Anders, Primordial noble gases in separated meteoritic minerals. Geochim. Cosmochim. Acta 34 (1970) 1175–1198.

Jungck, M.H.A. and P. Eberhardt, Neon-E in Orgueil density separates, Meteoritics 14 (1979) 439–441.

Jungck, M.H.A., F.O. Meier, and P. Eberhardt, Apatite in Orgueil carrier phase for Neon-E? Meteoritics 16 (1981) 336–337.
Kerridge, J.F., Solar nitrogen: Evidence for a secular increase in the ratio of nitrogen 15 to nitrogen 14. Science 188 (1975) 162.
Kolodny, Y., J.F. Kerridge, and I.R. Kaplan, Deuterium in carbonaceous chondrites. Earth Planet. Sci. Lett. 46 (1980) 149–158.
Lee, T., D.A. Papanastassiou, and G.J. Wasserburg, Demonstration of $^{26}$Mg excess in Allende and evidence for $^{26}$Al. Geophys. Res. Lett. 3 (1976) 41–44.
Lee, T., D.A. Papanastassiou, and G.J. Wasserburg, Aluminium-26 in the early solar system: Fossil or Fuel? Astrophys. J. 211 (1977) L107–L111.
Lee, T., W.A. Russell, and G.J. Wasserburg, Calcium isotopic anomalies and the lack of aluminum-26 in an unusual Allende inclusion. Astrophys. J. 228 (1979) L93–L98.
Norgaard, H., $^{26}$Al from red giants. Astrophys. J. 236 (1980) 895–898.
Reeves, H., and O. Johns, The long live radioisotopes as monitors of stellar, galactic and cosmological phenomena. Astrophys. J. 206 (1976) 958.
Robert, F., L. Merlivat, and M. Javoy, Deuterium concentration in the early solar system: Hydrogen and oxygen isotope study. Nature 282 (1979) 785–789.
Schramm, D.N., F. Tera, and G.J. Wasserburg, The isotopic abundance of $^{26}$Mg and limits on $^{26}$Al in the early solar system. Earth Planet. Sci. Lett. 10 (1970) 44–59.
Silk, J., Molecular clouds and star formation. In: 10th Advanced Course of the Swiss Society of Astronomy and Astrophysics (1980) A. Maeder and L. Martinet, editors.
Trauger, J.T., F.L. Roesler, and M.E. Mickelson, The D/H ratios on Jupiter, Saturn und Uranus based on new HD and $H_2$ data. Bull. Amer. Astron. Soc. 9 (1977) 516.
Vidal-Madjar, A., C. Laurent, R.M. Bonnet, D.G. York, The ratio of deuterium to hydrogen in interstellar space III. The lines of sight to Zeta Puppis and Gamma Cassiopeiae, Astrophys. J. 211 (1977) 91.
Wasserburg, G.J., D.A. Papanastassiou, and T. Lee, Isotopic heterogeneities in the solar system. In: Early Solar System Processes and the Present Solar System (D. Lal, editor) North Holland Publishing Company, Amsterdam (1980).
York, D.G., and J.B. Rogerson, The abundance of deuterium relative to hydrogen in interstellar space. Astrophys. J. 203 (1976) 378.

*Address of the author:*

Dr. Peter Bochsler
Physikalisches Institut der
Universität Bern
Sidlerstrasse 5
CH-3012 Bern (Schweiz)

# Formation and Nucleosynthetic Evolution of the Stars

Pierre Bouvier

## 1. Diffuse and Condensed Matter in the Universe

On the cosmical scale, the observable universe reveals itself as a collection of galaxies marking out the ways of space-time. Each galaxy is made up of matter, either condensed into stars or scattered in extensive patches of gas and dust. Several phenomena, namely stellar winds, mass loss by stars, explosive events, bear witness of some conversion of condensed into diffuse matter; on the other hand, how do stars come into existence?

Before answering this delicate question, let us recall here that innumerable stellar models have been built since the middle of the present century, using evermore powerful computing means and increasingly refined input physics, allowing us to follow the history of a star of given mass and chemical composition, starting from initial conditions such as those expected to prevail when nuclear burning is first switched on at the star's centre.

The results given by these model computations, which build the main body of the theory of stellar evolution, have made it possible to ascribe a fairly definite age to any star with known observable characteristics and it soon became clear that the younger bright stars, as we observe them in the galaxies, are most often associated with interstellar clouds of diffuse matter from which they presumably originate. What we find in those clouds in gas, mostly hydrogen often in molecular form, and dust grains.

Data on the masses of clouds come from the observation of the 21 cm hydrogen line, of interstellar absorption lines (due to ionized calcium in particular) and of obscuration caused by dust grains. The mean mass density inside large interstellar clouds should be on the order of $10^{-23}$ g cm$^{-3}$ (Allen 1973), but our attention will be drawn especially towards some dark dust clouds showing little if any 21 cm emission, so that hydrogen seems to be essentially in molecular form in these fairly dense clouds, which are likely to form stars. The largest values of the density within such clouds are close to $10^{-19}$ g cm$^{-3}$.

## 2. Heating and Cooling Processes

An isolated interstellar cloud is submitted to the effects of two adverse forces: self gravitation which tends to produce a collapse of the cloud towards some central region and pressure forces conversely leading to a dispersion of the cloud in space. Although turbulent motions, magnetic fields and rotation of the cloud might play a significant role in certain cases, we shall generally consider here a pressure of thermal nature only. In order to estimate the temperature of an interstellar cloud, we have to call on some sort of balance between the heating and cooling processes going on inside the cloud.

At densities of $10^{-22}$ g cm$^{-3}$ at least, we should retain primarily (Larson 1974) the following heating and cooling processes:

1) Heating effects by low energy galactic cosmic rays, ultraviolet radiation, soft X-rays; when, during some contraction, the optical depth of the cloud becomes much larger than unity, and the density larger than $10^{-19}$ g cm$^{-3}$, these mechanisms lose their importance to the advantage of compressional heating.

2) Among the cooling processes, the most important appears to be collisional excitation with free electrons and protons, of ions (often CII), atoms (OI), molecules ($H_2$, CO) followed by infra-red emission of radiation but here again, in very dense and opaque clouds, another process shall prevail, namely cooling due to inelastic collisions of mole-

cules with dust grains, followed by infra-red radiation emission from the grains.
By equating the heating and cooling rates for the gas and dust, one manages to obtain a density-temperature law illustrated in the (log $\rho$, T) plane for a given mass and given chemical composition of the cloud (Fig. 1). This curve is in fact practically unaffected by a change in the mass of the cloud and, as regards chemical composition, it is sensitive in the lower values of the density range to the assumed abundances of the coolants CII and OI. It is worth noticing here that in the density interval between $10^{-21}$ and $10^{-12}$ g $cm^{-3}$, the temperature varies only slightly, remaining around 5 to 20 K, a result in good agreement with radio observations of the comparative strengths of the 21 cm and Ly$\alpha$ interstellar line absorptions in dark clouds.

### 3. Gravitational Instability, Fragmentation

If self-gravitation is to dominate over the pressure forces (gravitational instability), the gravitational potential energy of the cloud must, in the context of the virial theorem, exceed twice its kinetic energy.
This condition for collapse, called the Jeans criterion, is expressed in the case of a spherical homogeneous cloud of radius R, mass M, temperature T, by the inequality

$$\frac{3}{5} G \frac{M}{R} > 3 \frac{\mathscr{R} T}{\mu} \qquad (1)$$

where G, $\mathscr{R}$ are respectively the gravitation and the gas constant, $\mu$ being the mean molecular mass number of the cloud's medium.
Therefore, in a cloud of given M, T, gravitational instability will occur when

$$R < R_c \equiv 0.2 \frac{GM}{\mathscr{R} T} \mu \qquad (2)$$

To the critical radius $R_c$ corresponds, for a cloud of a given mass density $\rho$ and temperature T, a critical minimum mass $M_J$, allowing the Jeans criterion to be written also:

$$M > M_J \equiv 5.45\, \rho^{-1/2} \left( \frac{\mathscr{R} T}{G \mu} \right)^{3/2} \qquad (3)$$

Adopting a more realistic approach than that of a mere homogeneous cloud would only slightly modify the numerical factors in (2) and (3).
Turning back to the equilibrium curve of section 2, obtained as a result of heat balance inside the cloud, we may then express $M_J$ as a function of $\rho$ only; thus for a rather typical large interstellar cloud where $\rho \sim 10^{-22}$ g $cm^{-3}$, the minimum unstable $M_J$ given by (3) is of one thousand solar masses at least. However, the normal range of stellar masses extends from about 0.03 $M_\odot$ to an upper limit which, according to the latest massive star evolutionary models (Maeder 1980) should be taken around 150 to 200 $M_\odot$. Under external conditions of strong compression due either to galactic density waves or to shock waves from a supernova event, $M_J$ could perhaps fall into this upper mass range, but in general, a large interstellar cloud is not likely to collapse and condense into a single normal star. Consequently, fragmentation of the initial cloud is expected, presumably in several steps, the number of these steps being sensitive to the inner motions of the cloud (rotation, turbulence) and also to the combination of atoms into molecules (Reddish 1978).
Replacing in (2) $R_c$ by $R_c = (3\ M/(4\pi\rho_c))^{1/3}$, the Jeans criterion for the gravitational instability of a cloud of given M, T now reads

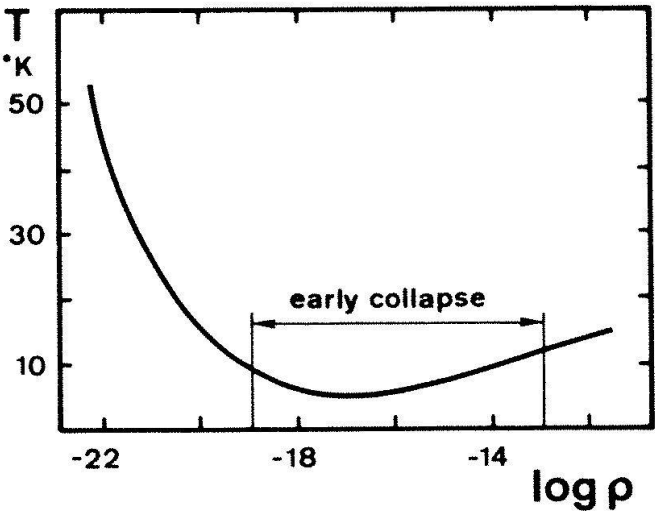



*Fig. 1.* Equilibrium temperature-density curve resulting from heat balance inside in interstellar cloud ($\rho$ in $gcm^{-3}$). The density range for which $-19 < \log\rho < -13$ corresponds to the isothermal collapse of a protostar (see text).

$$\rho > \rho_c = \frac{375}{4\pi M^2}\left(\frac{\mathscr{R}T}{G\mu}\right)^3 \qquad (4)$$

If $\rho \gg \rho_c$, the cloud is unstable against fragmentation and tends to break into smaller clouds in which, owing to their smaller masses but nearly same temperature T, the density $\rho$ will get closer to the corresponding critical $\rho_c$. At the end of the fragmentation process, $\rho = \rho_c$ inside the fragments and the initial large cloud has now become an aggregate of protostars.

The concept of fragmentation agrees well with the observational situation suggesting that most of the stars if not all of them, and especially the more massive ones, are born inside cloud (protostar) aggregates later becoming stellar associations or clusters. In spiral galaxies, the associations and dark interstellar clouds are both found within the spiral arms.

Finally, let us mention that model computations of cloud collapse yield a mass distribution of the ultimate fragments which agrees well with the observed mass spectrum of the youngest stars: the small mass stars are much more numerous than the massive ones and the distribution function f(m) behaves like $m^{-2}$.

## 4. Evolution of a Protostar: Isothermal Phase

The Jeans mass $M_J$ given by (3) with the ($\rho$, T) dependence of Figure 1 is a sharply decreasing function of $\rho$, going down over the entire range of stellar masses, approximately from $10^2$ to $10^{-2}$ solar masses, as $\rho$ increases from $10^{-21}$ to $10^{-12}$ g cm$^{-3}$. We thus expect protostars to form by fragmentation at minimum density $\rho = \rho_c$ in the former density range and since the corresponding temperature remains, as mentioned at the end of section 2, close to 10 K right up to $\rho \sim 10^{-12}$ g cm$^{-3}$, the early contraction phase of a protostar will develop in a quasi-isothermal condition.

Numerical computations of the collapse of a spherical protostar are based on the set of Eulerian equations of hydrodynamics together with a mass distribution in spherical shells around the centre. Boundary and initial conditions are not easy to specify; Larson (1974) chose a constant boundary of the protostar (corresponding to an adequate value of the external pressure) and started the computation with a cloud of uniform density $\rho \approx \rho_c$; this avoids having to specify in detail the initial conditions of the protostar, connected to the state reached at the end of the fragmentation phase.

Results obtained with different assumptions and by different authors appear qualitatively similar. Starting with constant $\rho$ and T values (Larson), such as $\rho = 10^{-19}$ g cm$^{-3}$ T $\approx$ 10 K, the pressure will be initially constant throughout the protostar and for lack of a pressure gradient, we find ourselves in the case of an initially freely falling configuration.

A pressure difference shall develop immediately after with the surroundings, at the cloud's surface, and the contraction of the outer layers decelerates, whence the occurrence of a density and of a pressure gradient. Meanwhile the deep interior keeps falling freely at a time-scale $t_{ff}$ proportional to the inverse square root of the initial density. This is also the time needed for the boundary separating the still homogeneous core from

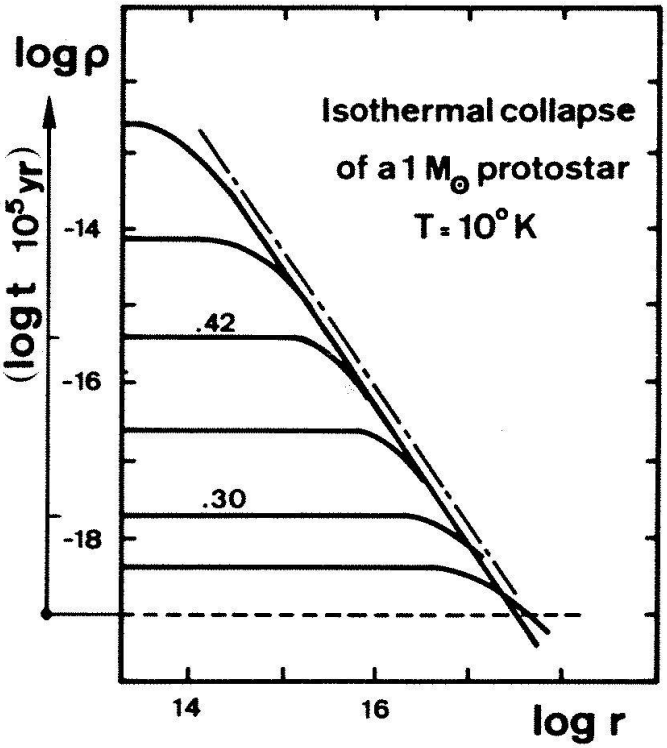



*Fig. 2.* Density distribution inside a 1 $M_\odot$ protostar at different times during isothermal collapse, starting at t = 0 when log$\rho$ = − 19 and ending at t = 300,000 years when log$\rho$ = − 13.
The limiting envelope curve follows a $\rho \propto r^{-2}$ law.

the inhomogeneous envelope, to travel inwards with the speed of sound (rarefaction wave). The evolution depicted in figure 2 concerns the isothermal collapse of a spherical non rotating protostar of one solar mass, and initial density $10^{-19}$ g cm$^{-3}$; This collapse is highly non homologous, exhibiting an increasingly strong central density peak, while the density profile $\rho(r)$ in terms of the distance r to the centre gradually unfolds according to the law $\rho \propto r^{-2}$.

As we saw in section 2, the cooling processes involved lead to infra-red emission from the grains and this radiation escapes freely from the protostar, which is therefore an infra-red source during isothermal collapse. This picture is in good agreement with observational evidence: infra-red sources have indeed been found in direction of many dark clouds, indicating that some of these clouds are the sites of current star formation. Recently, Snell (1981) has studied the atomic and molecular properties in nine interstellar dark clouds, obtaining a partial mapping of these clouds in infra-red.

The kinetic temperature in the dense regions of the clouds is between 10 and 20 K and their central density on the order of $10^{-19}$ g cm$^{-3}$. The relevant clouds have masses of order 10 to 140 $M_\odot$ (which has no bearing on the curves of figures 1 and 2) and exhibit density gradients in which the density, strongly peaked at the centre, diminishes outwardly according roughly to a $r^{-2}$ power law.

The end of the isothermal phase is attained when the central density reaches $10^{-13}$ g cm$^{-3}$; the medium then completely absorbs the radiation and a central core appears, heating up rapidly under further collapse.

## 5. Further Evolution towards the Main Sequence Stage

The protostar now consists of a small very dense core, containing 1% of the total mass, surrounded by an extended rarefied envelope still falling on the core, hence a sharp inward directed velocity gradient quickly steepening into a shock front at the core's surface. The opaque core increases in mass but shrinks in radius by losing energy from its outer layers, while the collapse of its central part is more adiabatic than isothermal, a large fraction of the energy released there being spent in dissociating and further ionizing the former molecular hydrogen. At the end of this short (a few centuries) and rather violent non-isothermal evolutionary step, during which the temperature rises only slowly, the central collapse gets balanced by the pressure forces, and this leads finally to the formation of a second core, termed the stellar core by Larson and having a mass less than $10^{-2}$ $M_\odot$, a density close to $2 \times 10^{-2}$ g cm$^{-3}$ and a temperature around 20000 K.

The evolution of the protostar is now characterized by the infall of the outer layers of the protostar on the stellar core, the latter continuing to grow in mass at the expense of the envelope and eventually acquiring the bulk of the protostellar mass.

In the case of fairly low mass protostars ($M < 3$ $M_\odot$), the envelope accumulates on the stellar core while the heating by slow contraction of the core becomes increasingly important as compared to the energy generated by the shock front. After about one million years for a 1 $M_\odot$ protostar, nearly all the envelope has fallen on the core and this 1 $M_\odot$ object has become a visible star (point No 6 on track in figure 3).

In case of a protostar of mass $M > 3$ $M_\odot$, the stellar core evolves more rapidly and the central temperature will soon become high enough for hydrogen nuclear burning to ignite before the infall of the envelope matter has declined to zero. Furthermore, in very massive stars, H-burning produces a strong temperature increase of the core, thus a large radiation pressure at its surface, causing a partial reversal of the infall of the remaining envelope and the ejection of the outer layers, finally leaving a massive main-sequence star. According to these model computations, massive protostars transform directly into main-sequence stars, characterized by core hydrogen burning (Appenzeller 1980).

However, as we just saw before, such is not the case for the lower mass protostars; consider again the 1 $M_\odot$ protostar which, 1.3 million years after the beginning of its collapse, became a visible star. It is now subject to a slow contraction, undergoing a quasi-equilibrium stage during which about half of the energy released is used up in heating the inner region, while the other half is radiated

away. The opacity is quite large, so that the energy is transported inside the star mainly by convection at first; later on, however, the medium gets hotter and less opaque, allowing radiative energy transfer to set up within most of the star.

Accordingly, the point representing the 1 $M_\odot$ star in the HR diagram follows a quasi-vertical path downward (the so-called Hayashi track, HT) and after a luminosity minimum, ascends slowly to the left, with surface temperature and luminosity both increasing (figure 3, points 6′–7). Near the centre, the first nuclear reactions between H and light elements D, Li, Be will rapidly destroy most of the latter and when the central temperature goes beyond $8 \times 10^6$ degrees, steady conversion of hydrogen into helium sets in; the 1 $M_\odot$ star has now reached, 50 million years after it came first visible, the main sequence stage where it will spend most of its life, in a state very close to hydrostatic and thermal equilibrium (point 8).

The larger the masses, the shorter the evolutionary timescales; thus for a 3 $M_\odot$ star the time for H burning to settle is around 2 million years, which is comparable to the accretion time for the envelope to fall on the stellar core. A 5 $M_\odot$ protostar will become, only after 500 thousand years, visible as a main-sequence star but H-burning had started already 100 thousand years earlier in the core.

In all the former considerations, we had neglected rotation effects which might indeed alter significantly the stellar evolution. Many investigations are presently on the way, with more elaborate computing codes. Let us just mention a recent paper by Regar and Shaviv (1981) who studied collapse and star formation processes in rotating turbulent interstellar gas clouds, without assuming a mechanism for the transport of angular momentum. It seems likely that the collapse with turbulent viscosity leads to the formation of a central protostar surrounded by a disk when the rotation is low enough and the turbulence efficient, but in the present treatment of the problem, it cannot yet be concluded that one reaches the state where a central star is formed, together with a planetary system. When the rotation is fast or turbulent transport less efficient, the central opaque object is very flat and this raises the possibility of multiple star formation.

Anyway, the picture of protostar evolution outlined above still remains somewhat controversial; new very recent infra-red mappings of the large Orion molecular cloud complex reveal that the compact infra-red

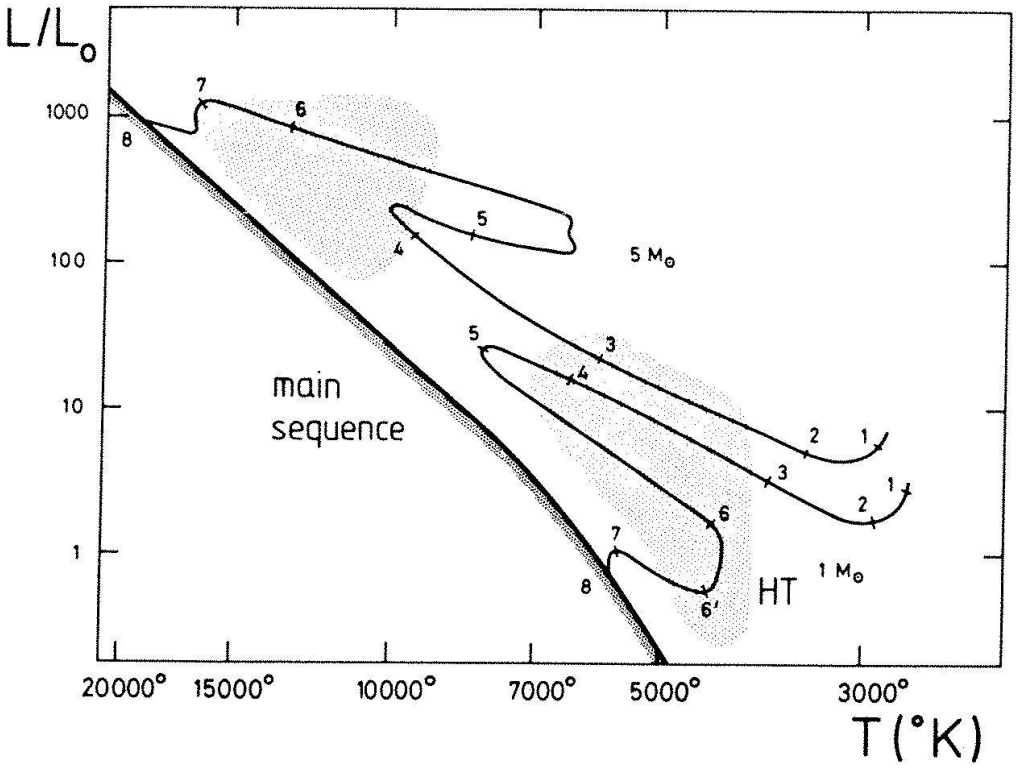



*Fig. 3.* Evolution of a 1 $M_\odot$ and a 5 $M_\odot$ protostar towards the main sequence. Numbers along tracks refer to specific evolutionary stages; it is a point No 6 that the protostar becomes a visible star. The upper shaded area denotes the Herbig Ae stars and the lower one the T Tauri stars.

objects thought at first to be protostars might well consist of very young and massive stars, dramatically shedding matter in the form of a dense stellar wind (Wynn-Williams 1981).

## 6. The Main Sequence as the First Nuclear Burning Stage

On account of the low potential barrier of its nucleus (1.4 MeV) and further of its cosmically preponderant abundance, hydrogen appears as the first major nuclear fuel met in stellar interior conditions. Consequently, a star spends most of its life in the narrow band of the colour-luminosity diagram termed the main-sequence, starting from its lower envelope and moving slowly upwards. If we follow the main-sequence from right to left, we pass from the lower to the higher mass stars (figure 4). The time spent in the main-sequence stage is shorter for the massive stars, since the thermonuclear reactions are enhanced in a hotter environment; if a one solar mass star stays about 10 billion years in the main-sequence, the lifetime drops to 10 million years for a 15 $M_\odot$ star.
The H-burning reactions occurring in stars, converting H into He are of two kinds: the proton-proton chains and the CNO cycles (first investigated at the end of the thirties by C.F. von Weiszäcker and H.A. Bethe). The p-p chains dominate in cool stars; of these three chains, the first can operate from pure hydrogen, while the two others work with $^4$He as a catalyst. On the other hand, the CNO cycles require the presence of catalytic isotopes of carbon, nitrogen, oxygen and are more efficient in hotter stars. On the approach to an equilibrium state, these C, N, O isotopes have their initial abundances modified, and when the equilibrium temperature is attained (for instance 25 million degrees for a central density of some 100 g cm$^{-3}$), the most abundant of these isotopes is $^{14}$N.
Anyway, the main outcome of this main-sequence stage is the nucleosynthesis of helium from hydrogen within the stellar core. The energies liberated by these thermonuclear reactions in the central regions of the main-sequence stars have enabled us, fixing a suitable chemical composition of the stellar matter, to account for the observed luminosities of these main-sequence stars in a most satisfactory way.

## 7. The Red Giant Stars as Second Nuclear Burning stage

When hydrogen is exhausted in the stellar core, a gravitational contraction of the core follows, together with a wide expansion of the outer envelope: in the colour-luminosity diagram, the representative point of the star moves far away to the right of the main-sequence and the star has now become a red giant (figure 4), with high luminosity, large radius and a deep convective envelope. The central temperature has now risen beyond 120 million degrees, favouring the ignition of helium-burning through triple $\alpha$-process, where three $^4$He nuclei react together simultaneously, to form a carbon nucleus $^{12}$C. When enough $^{12}$C is present, further $\alpha$-capture by $^{12}$C can synthesize oxygen $^{16}$O, while $^{14}$N resulting from the CNO cycles which still operate in a H-burning shell source outside the He core may also react with $^4$He to give some $^{18}$O.
At the end of such reactions, about 2% of $^{18}$O has been formed (starting with a standard chemical composition for population I stars), which is enough to induce further reactions with $\alpha$ particles, releasing free neutrons, important for the later building of elements heavier than those of the iron group.
The core He-burning phase (together with a shell H-burning) lasts five or six times less than the core H-burning on the main-sequence; the end products are now carbon, oxygen, and some neon, magnesium. The reaction rate of the C($\alpha\gamma$)O reaction contains uncertain parameters, so that we may always adjust the C and O abundances resulting from the He-burning to the observed values; on the other hand, the long lifetime of $^{16}$O against $\alpha$-capture makes it impossible, through O($\alpha$, $\gamma$)Ne, to obtain as much $^{20}$Ne as $^{16}$O as the observations seem to require.

## 8. Carbon and Oxygen Burning Stages

Stellar evolution followed hitherto has proceeded through two successive hydrostat-

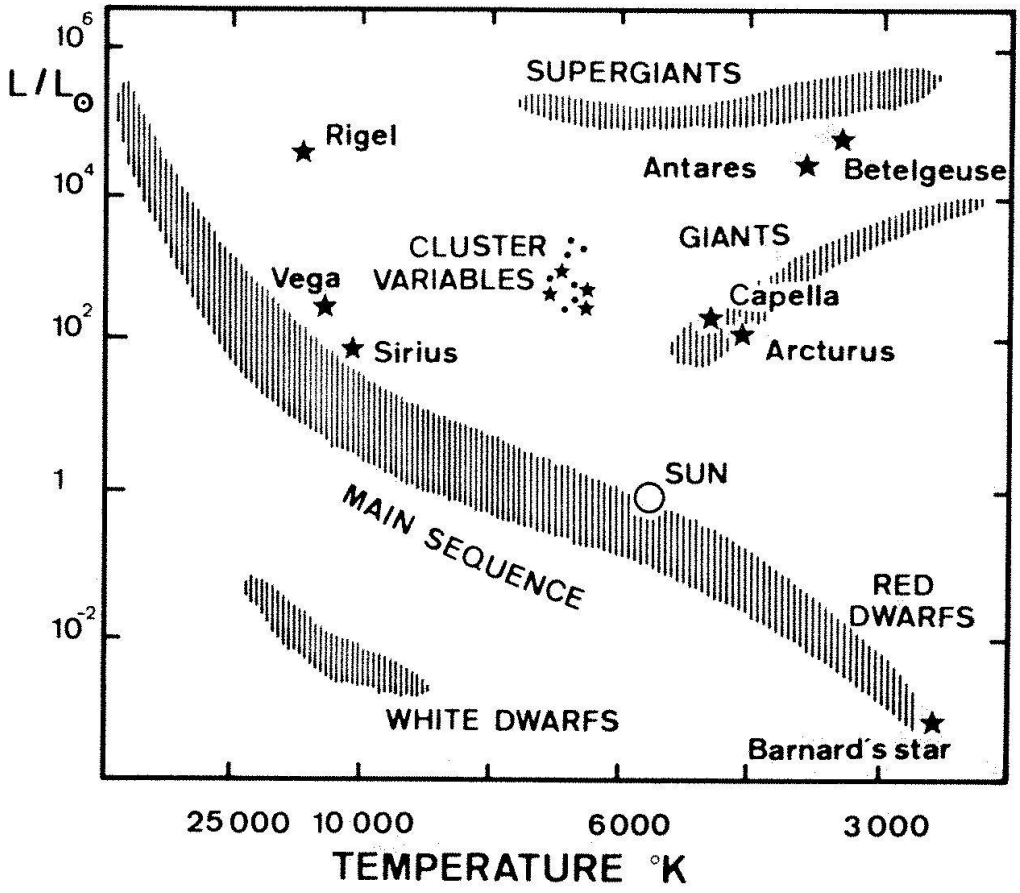



*Fig. 4.* Qualitative Hertzsprung-Russell Diagram.

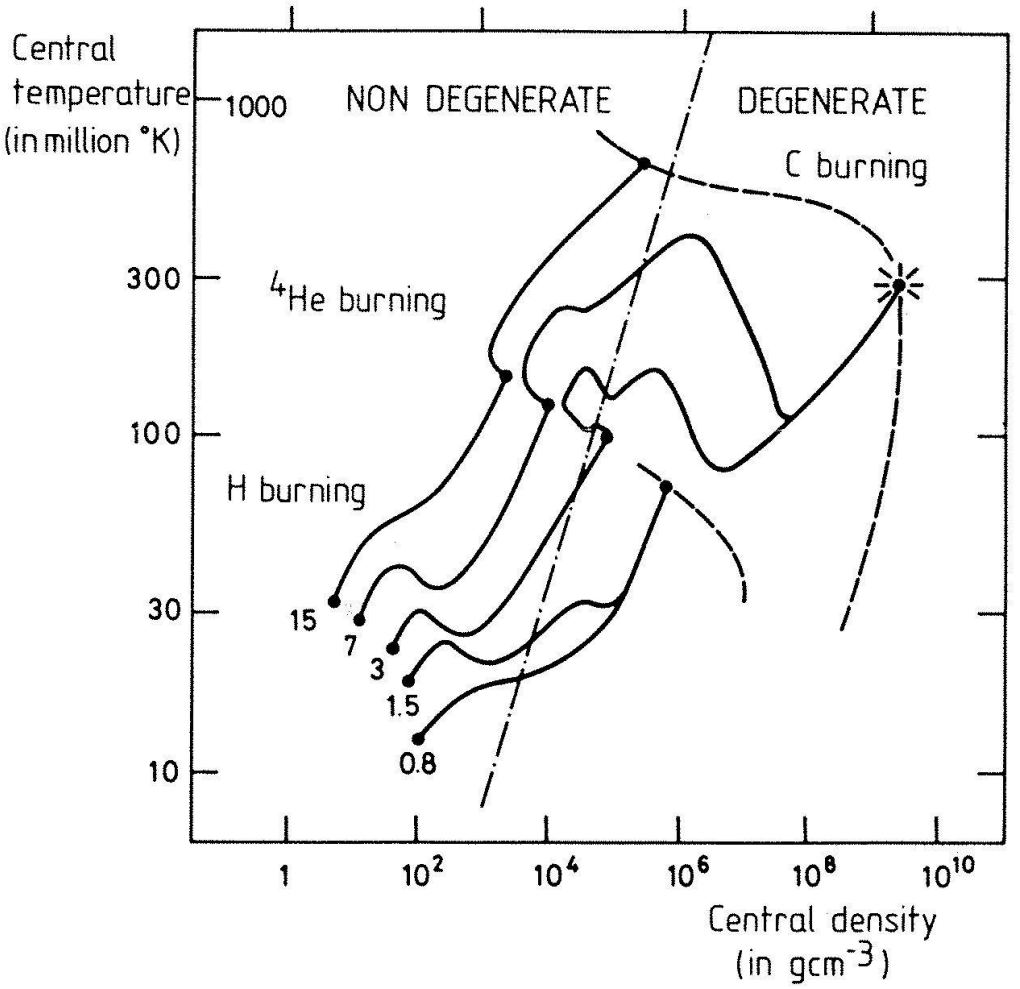



*Fig. 5.* Central conditions inside stars of masses ranging from 0.8 to 15 $M_\odot$ evolving from main sequence to carbon ignition. The dash-dotted line separates roughly the degenerate and non-degenerate regions of the (logT, log$\rho$) plane.

ically stable nuclear burning stages, separated by a gravitational contraction of the star's core. Similarly, further hydrostatic nuclear burning stages, first examined by G.R. Burbidge, W.A. Fowler, F. Hoyle more than 20 years ago, can occur inside very hot and massive stars, located in the supergiant region of the colour-luminosity diagram (figure 4). When the temperature T rises beyond half a billion degrees K under the contraction following the helium-burning phase, the carbon nuclei will start to react between themselves and due to the high excitation of the compound nucleus $^{24}Mg$ associated to the reaction $^{12}C + ^{12}C$, we are faced with several outgoing channels, the two most probable being $^{23}Na + p$ and $^{20}Ne + \alpha$. At the typical temperatures of 600 to 700 million degrees required for this carbon-burning stage, the liberated p (protons) and $\alpha$-particles are rapidly recaptured, eventually yielding free n (neutrons) through subsequent reactions like $^{12}C(p,\gamma)^{13}N(,\beta^+\nu)^{13}C(\alpha,n)^{16}O$. Such free neutrons are also produced by the channel $^{12}C + ^{12}C \rightarrow ^{23}Mg + n$, which is endothermic and much less probable than the two former ones.

Anyway, at the end of the preceding reactions, $^{24}Mg$ being the most stable of all the nuclei involved, becomes the most abundant product element of carbon burning.

If ever T attained a billion degrees, the oxygen-burning may set in and, similarly to the case of carbon, the $^{16}O + ^{16}O$ reaction has two prominent outgoing channels, yielding respectively free p's and free $\alpha$'s while a third, less important one, produces free n's.

Here, as a result of the relevant O-burning reactions, isotope $^{28}Si$, the nucleus of which is the most stable among all those involved, becomes the most abundant product element of oxygen-burning.

## 9. C-Detonation

In the course of stellar evolution, the central conditions exhibit increasing densities and temperatures; therefore electron degeneracy sets in, and fairly soon for low mass stars (figure 5). At 0.6 billion degrees, neutrino emission due to universal weak interaction processes becomes significant and further, around a billion degrees, half the energy radiated by the star is in form of neutrinos. The subsequent cooling of the C–O degenerate core tends to render the latter still more degenerate, thereby postponing the ignition of carbon. The situation is now similar to the one met in the degenerate helium core of a less massive star at the end of its core hydrogen-burning phase: the ignition of helium leads to the so-called He flash which mixes somewhat certain inner layers without destabilizing the whole star which goes on evolving to the final white dwarf stage, after undergoing significant mass loss.

But in the very hot and dense environment ($T \sim 0.6 \times 10^9$ K, $\rho \sim 10^9$ g cm$^{-3}$) reached in the degenerate centre of a supergiant star, the ignition of carbon entails an explosive C-detonation blowing up the star entirely.

Such an evolution should be envisaged for stars of moderate mass, say between 3 and 8 solar masses $M_\odot$ although the complete absence of a dense remnant (like a pulsar) after the explosion probably discards this scenario as an explanation of a supernova event.

Below $M = 3\ M_\odot$ nuclear burning stops, as we just saw, at the He-flash and no further nucleosynthesis will occur. Further, in very small mass stars ($M < 0.1\ M_\odot$), degeneracy sets in before the firing of any nuclear reaction. On the other hand, C-ignition inside a massive star (more than $8\ M_\odot$) can take place in nondegenerate matter, eventually followed later by O-burning. The massive stars (around $22\ M_\odot$) develop an inner composite structure consisting of an iron nucleus, surrounding by concentric layers successively enriched mainly in Si, Ne, $^{16}O$, $^{12}C$, $^4He$, H (onion-skin structure) (figure 6). Nuclear burning goes on at the inner interfaces between contiguous layers.

## 10. Photodisintegration. Silicon Melting

Between 1 and 2 billion degrees, the photodesintegration of nuclei by the radiation of the fantastically hot photon bath will set in (comparable, on another scale, to photoionization of atoms occurring around 10 thousand degrees), thus excluding the possibility of new hydrostatic nuclear burning phases of the fusion type seen before.

The less stable nuclei (like those of odd

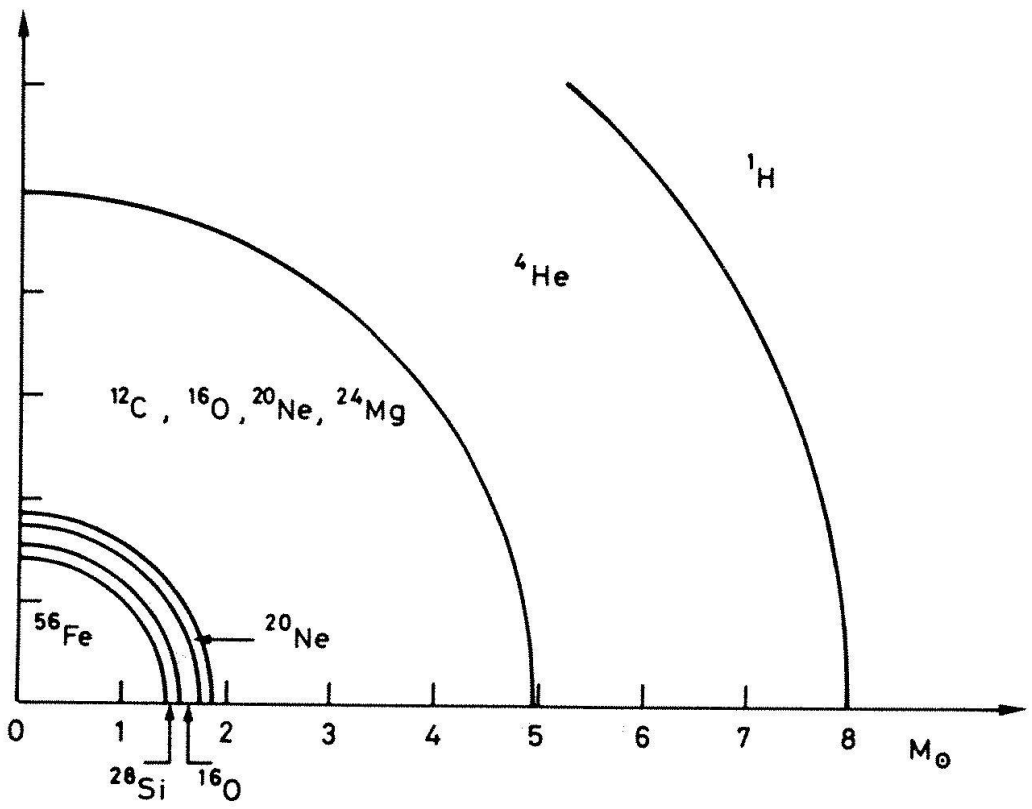



*Fig. 6.* Internal structure of a massive star (around 22 $M_\odot$) having presumably reached the pre-supernova stage. The composition of the successive concentric layers reflects the successive nuclear burning stages.

atomic mass number) shall be destroyed in favour of more stable ones, so that photodisintegration tends to redistribute the ejected loosely bound nucleons $(\alpha, p, n)$ into nuclei where they become more tightly bound.

In the range of the relevant intermediate nuclear masses, $^{28}Si$ is the most stable nucleus against photodisintegration and when it starts to be photodisintegrated, at temperatures drawing close to 3 billion degrees, many reactions of type $(\alpha,\gamma)$, $(p,\gamma)$, $(n,\gamma)$ and their inverses are at work.

The nucleons slowly photoejected by $^{28}Si$ are principally recaptured by $^{28}Si$ nuclei (which are the most abundant): reactions like $^{28}Si(\alpha,\gamma)\,^{32}S$ will quickly be partly counteracted by the inverse $^{32}S(\gamma,\alpha)^{28}Si$, but the $\alpha$-particles ejected by the overwhelmingly numerous $^{28}Si$ nuclei will also lead to the gradual building of heavier nuclei, through $^{32}S(\alpha,\gamma)$ $^{36}Ar$ etc in a slow progression towards stabler nuclei, up to those of the iron group which are known to have the maximum binding energies per nucleon. The rates of these photodisintegration and capture reactions being much larger than the rate of variation in elemental abundances, we therefore have here a kind of quasi-equilibrium state of the medium, during which $^{28}Si$ slowly melts away.

The characteristic time for the photodisintegration redistribution of the nuclei from silicon to the iron group is governed by the photodisintegration rate of $^{28}Si$, which itself depends sharply on the temperature T. Remember also that the most tightly bound nuclei in the iron group are not those with equal numbers Z of protons and N of neutrons, but these having a neutron excess by 2 or 4.

When T is around 3 billion degrees, this redistribution time is on the order of one day, which allows $\beta$ decays to occur so that one reaches the iron group with isotope $^{56}_{26}Fe$ $(Z < N)$, but for T between 4 or 5 billion degrees, no such decays have time to appear and we finally attain isotope $^{56}_{28}Ni$ $(Z = N)$, or $^{54}Fe + 2p$ for still larger T's.

As $^{28}Si$ disappears, we reach a statistical equilibrium of the medium, in which the abundance of each nuclear species finally depends on the density and temperature of the medium, together with the ratio of the total number of protons to the total number of the neutrons (bound and free, in a given volume).

Independently of the photodisintegration process, the chemical elements of mass number $A \gtrsim 65$, located beyond the iron-group, are less stable on account of their high inner

Coulomb repulsion; they cannot be built by fusion but only by neutron capture as we shall see later (section 14).

## 11. Abundances of the Elements in the Universe

The well known Standard Abundance Distribution (SAD) curve gives, in terms of the atomic mass number A (figure 7) the relative abundances of the chemical elements as they have been determined in the carbonaceous chondrite meteorites, in the solar and many stellar atmospheres; furthermore a large number of stars and of galaxies follow the SAD curve, as well as the interstellar matter. H and He are by far the most abundant observed elements and when A increases, the abundances decrease sharply before rising again to the iron group ($50 \lesssim A \lesssim 65$). We shall come back later to the light underabundant elements Li, Be, B (section 15) and to the heavy elements (section 14).

We notice further the presence of several peaks in the range of heavy elements, corre-

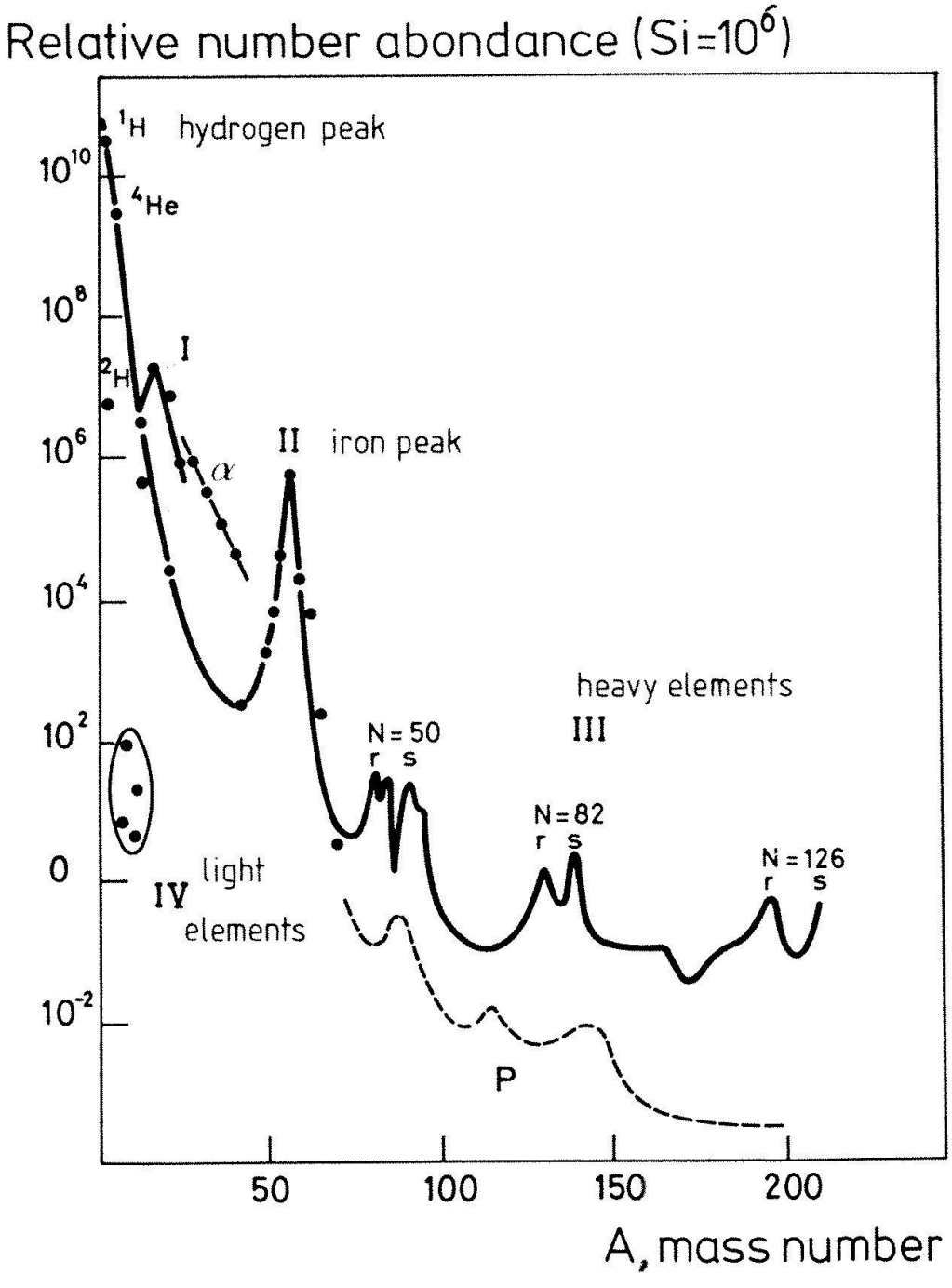



*Fig. 7.* Standard Abundance Distribution of the Elements (see text).

sponding to the so-called magic nuclei ($N = 50, 82, 126$) for which the n-capture cross-section shows a significant minimum.
The labels s, r, p in Fig. 7 refer to heavy elements formed respectively by the s-, r- or p-processes (section 14).
In order to compare the products of nucleosynthesis which took place deep inside the stars with the data collected in the SAD curve and referring to observations made on the solar system, the solar and stellar surfaces, the interstellar matter, we must consider the possible ejection of the inner nucleosynthesized stellar layers into the outer galactic medium. Let us therefore turn our attention to explosive objects.

## 12. Explosive Events

Among the known stellar explosive objects, the supernovae are undoubtedly the most conspicuous and powerful, exhibiting a very sudden increase in luminosity followed by a steep and more or less regular decrease of their light curves. The mechanism that triggers a supernova explosion must be able to release a tremendous amount of energy (around $10^{51}$ ergs) and also to allow the formation of a remnant star.
Recalling the evolution of massive stars ($M \gtrsim 8\ M_\odot$) mentioned at the end of section 9, we saw that these stars acquire an onionskin structure made up of concentric shells, the outermost one containing essentially hydrogen, while the internal layers appear to be successively enrichted in He, C, O, Ne, Si as a result of the different hydrostatic nuclear burning stages which they underwent (figure 6). In the central Fe-Ni core, the temperature now tends to exceed 5 billion degrees, while contraction favoured by the cooling effect of a strong neutrino emission accelerates the star's final evolution. However, at such high temperatures photodisintegration of the iron group nuclei becomes possible, leading to

$$^{56}\mathrm{Fe} + \gamma \leftrightarrows 13\,^{4}\mathrm{He} + 4\mathrm{n} - \mathrm{Q}$$

and similar reactions with $^{54}$Fe, $^{56}$Ni. Since these nuclei are particularly stable, such reactions are highly endoenergetic (large $Q > 0$) therefore capable of triggering a violent implosion of the stellar core, immediately followed by the explosion of the surrounding matter still undergoing nucleosynthetic processes. The central core, which will very quickly be neutronized, finally transforms either into a neutron star or pulsar if its mass is below about $2\ M_\odot$, or otherwise eventually into a black hole. The mechanism just outlined might explain the supernova II phenomenon, observed in galaxies with population I stars.
Much less important with respect to the amount of energy released ($10^{43}$ ergs), are the novae outbursts, which may last quite as long as a supernova event (a few months) but are much more frequent and often recurrent. Without entering here into details (see Audouze and Vauclair 1980), we notice that the nova phenomenon affects only the outer layers of a prenova star, which is probably a white dwarf enriched in C, N, O isotopes. When such a white dwarf finds itself belonging to a binary system, the other companion being a star evolved to the red giant stage, matter from this companion, essentially H and He of its outer layers, is transferred to the white dwarf and heats its surface to 100 million degrees or more, thus triggering the so-called hot CNO cycle and causing a nova outburst, likely to be the site for the formation of some rare odd-A isotopes such as $^{13}$C, $^{15}$N, $^{17}$O.

## 13. Explosive Nucleosynthesis

The thermonuclear reactions causing the explosive events described in section 12 arise on a much shorter time scale (a few seconds or hours) and at much higher temperatures and/or densities than during hydrostatic stable nuclear burning. The problem, first tackled by W.D. Arnett and J.W. Truran about 12 years ago, is very complicated in general, since we have to follow the abundance variation of many nuclear species through a fairly large network of reactions; moreover the rapidity of the reactions in the present environment renders all the more important their dependence on density and temperature with time, actually involving a hydrodynamic treatment of the explosive process. In attempting to describe such a treatment, we should use the equations of

state and of energy transport within the stellar medium.
Relevant calculations have been performed for some novae outbursts, but not yet for supernovae explosions, where one usually relies on simplifying assumptions. One first assumes that $\rho$ and T evolve according to a theoretical profile for which the characteristic time is the free fall time scale $t_{ff} = 446\ \rho_0^{-1/2}$ as expressed (in cgs units) in the collapse of a uniform gas sphere of initial density $\rho_0$.
We then write

$$\rho(t) = \rho_0 \exp\left(-\frac{t}{t_{ff}}\right),$$

$$T(t) = T_0 \exp\left(-\frac{t}{3\,t_{ff}}\right)$$

by assuming further the medium (submitted to a strong shock wave) to undergo an adiabatic transformation.
Now, the initial values $\rho_0$, $T_0$ are connected by equating the characteristic life-time $t_c(\rho_0, T_0)$ of the relevant nuclear fuel as obtained in usual hydrostatic conditions, to the free-fall time scale $t_{ff}(\rho_0)$; the point is then to compare the possible $(\rho_0, T_0)$ values to the density and temperature attained, during the evolution of a massive star ($22\ M_\odot$) in the different layers (figure 6) just before the supernova explosion.
Thus, for the couple of initial values $(\rho_0, T_0) = (10^5, 2)$ in $gcm^{-3}$ and billion K respectively, explosive ignition of carbon could take place; the $\rho_0$ value agrees indeed with that met in the C-rich layer of the $22\ M_\odot$ presupernova star while $T_0$ is a little below the T-value prevailing in that layer.
The same can be said about explosive oxygen burning at $(\rho_0, T_0) = (2 \times 10^5, 2.6)$ and explosive silicon burning at $(\rho_0, T_0) = (2 \times 10^7, 4.7)$.
The final output of explosive nucleosynthesis is very sensitive not only to the $(\rho_0, T_0)$-values, but also to the initial chemical composition currently parametrized by the neutron enrichment factor $\eta = (N - Z)/(N + Z)$.
The resulting scenario for a supernova event is then the following: the strong shock, induced by the core's implosion due to iron photodisintegration, propagates rapidly outwards through the successive Si-, O-, C-, He-rich layers of the massive star, heating them to bring their temperature (main parameter) at the $T_0$ value, thereby switching on the corresponding explosive nuclear burning in a very short time.
D.N. Schramm and W.D. Arnett showed that when such exploding stars have expelled all the layers surrounding their dense inner iron core, one does obtain the relative abundances of the elements (at least those with $\alpha$-nuclei) observed in the solar system. In reality, the presolar interstellar medium has been enrichted by stars which had different masses, but on the average, the nucleosynthesis output appears to be satisfactorily represented by the contribution of a $22\ M_\odot$ star (Schramm 1977).

## 14. The Formation of the Heavy Elements

As recalled at the end of section 10, the heavy nuclei found beyond those of the iron group ($A > 65$) can no longer be formed by thermonuclear fusion reactions between charged particles; the very high potential barrier of these nuclei becomes insuperable and the only mode of formation for such heavy elements is through neutron capture.
In the (N, Z) plane, where Z is the number of protons and $N = A - Z$ the number of neutrons per nucleus, the known stable heavy elements are located along a curve corresponding to the so-called 'valley of stability' in the (M, Z, A) space, where M is the mass of the (Z, A) nucleus (figure 8).
Adding more and more neutrons to a given nucleus will in general remove it from the curve, in a zone of instability, from which it tends to fall back into the valley of stability through a Z-increase, namely a radioactive decay.
In contrast with fusion reactions, the cross section $\sigma(v)$ for capture of neutrons having velocity v, decreases with energy and in the astrophysical environment met here, the average product $\langle \sigma v \rangle$ giving the rate of neutron absorption, is roughly independent of energy; consequently the life-time $t_n$ of a heavy nucleus against neutron capture is inversely proportional to the neutron density.
The building rate of a heavy nucleus by n-capture depends on the ratios of $t_n$ with the

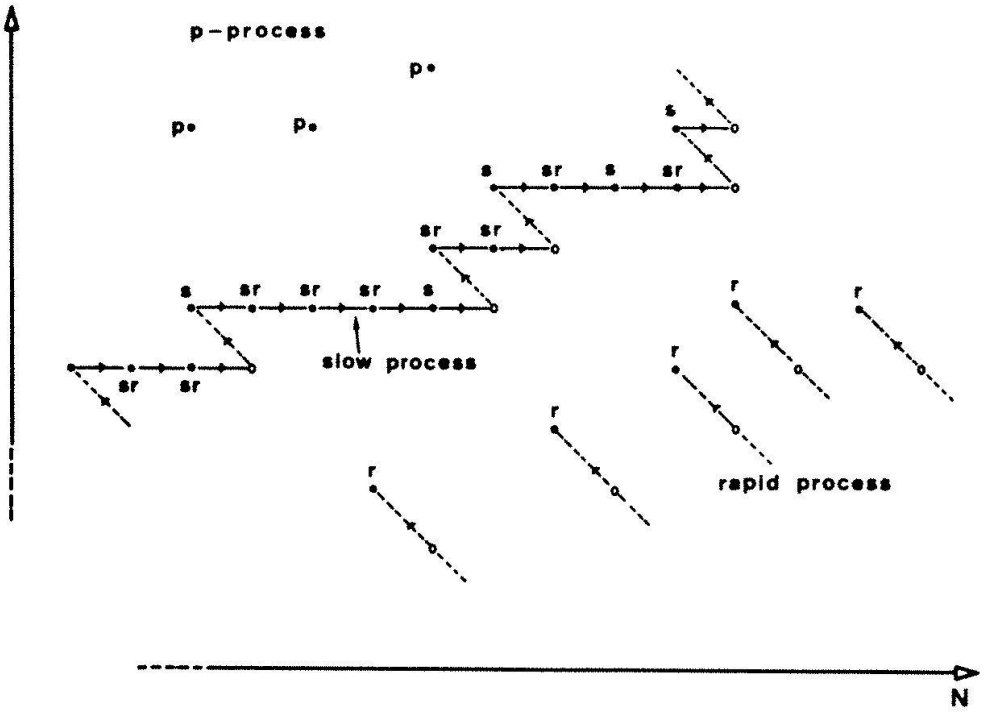



*Fig. 8.* Schematic illustration of the heavy element building processes in the (Z,N) plane. Full circles denote stable elements; those labelled s,r,p, correspond to the respective s,r,p-processes with produced them and for those labelled sr, formation can occur either by s- or by r-process. Open circles denote $\beta$-unstable elements.

average period $t_\beta$ of $\beta$-decay; this ratio is usually much larger or much smaller than unity. In the first case, we are faced with the s-process, where the free neutron flux is sufficiently weak to allow $\beta$-decays to occur, but in the second case the n-flux is very intense, no $\beta$-decays have time to appear and we get the r-process.

The s-process is assumed to take place during the red giant stage of stellar evolution, in particular when helium flashes induce an incomplete mixing of the H and He zones; $^{12}C$ resulting from the He-burning may become more abundant than H in the mixed matter, where the CNO cycle reactions then reduce to $^{12}C(p\gamma)\ ^{13}N(,\beta^+\nu)^{13}C$ and $^{13}C$ can react with He, according to $^{13}C(\alpha,n)^{16}O$ which is an important source of free neutrons. In the He-burning region, the former presence of $^{14}N$ can also, by $\alpha$ capture and subsequent steps, lead to the release of free neutrons.

The iron-group nuclei are usually considered to be the seed nuclei for the s-process, allowing the gradual building of heavier nuclei from $^{56}Fe$ to $^{206}Pb$; the main property emerging here is the continuous decrease, with increasing A, of the product $\sigma N$ of the cross-section for n capture by nuclei of mass A, times their concentration N.

Observationally, the working of the s-process in stars is illustrated by the presence of Tc (A = 99), detected in the atmosphere of certain red giant stars; this element has indeed been formed in the star, since its radioactive lifetime of $10^5$ yr is much shorter than the age of the star.

Moreover the variable star FG Sagittae, to which much attention has been paid lately, revealed a large annual increase in the abundances of s-process elements like Ba, Y, Zr; a similar increase of rare earth s-process elements has been noticed in CI Cygni.

In contrast to the fairly smooth dependence of $\sigma N$ on A in the s-process, yielding a curve in good agreement with the solar values pertaining to s-process abundances, the r-process leads to quite a disorderly picture for $\sigma N$ in terms of A.

The r-path in the Z,N plane brings us far to the right of the valley of stability, viz. to very high n-rich and unstable nuclei (figure 8). As more neutrons are added to the (Z,A) nucleus, a point (Z,A′) is reached, where A′ is distinctly larger than A, beyond which the nuclear binding energy is too weak to allow further n-capture; at this so-called waiting point, the nucleus is subject to no reaction until a $\beta$-decay changes Z into Z + 1, so that n-capture may go on for the (Z + 1,A′) nucleus.

It is the poor knowledge about the binding energy of very n-rich nuclei not observed in nature, that makes the r-process not yet fully

understood, but nevertheless, this process appears quite necessary to account for the formation of elements heavier than $^{206}Pb$, like Th, U, Pu.

Finally to the left of the valley of stability, we find p-process elements disclosing a proton excess, therefore impossible to be formed by n-capture. Several mechanisms have been proposed to explain their formation; all of them are connected to the s- and r-processes and we shall not go into detail here (see Audouze 1980). Recently, a nucleosynthetic process involving rapid p-capture has been identified in a study on explosive H-burning (Wallace and Woosley 1981); this so-called rp-process has significant implications for element synthesis and energy production in H-rich material at high temperatures; substantial enrichment occurs in elements heavier than oxygen.

## 15. A Glance at the Light Elements. Spallation Reactions

The light elements D, Li, Be, B are destroyed in stars by p-capture occurring when the temperature exceeds already a million degrees K; they are therefore not produced by stellar nucleosynthesis. On the other hand, He is formed in stars during the main sequence stage (section 6) but not in sufficient amount to account for the observed He in the universe. We have to resort to other nucleosynthetic processes to explain the presence of these light nuclear species, and since the aim of the present lecture was to concentrate on nucleosynthesis in course of stellar evolution, we shall remain satisfied by giving here only a very brief outline about these other processes, of which we retain essentially two: spallation reactions and primordial cosmological nucleosynthesis, described in section 16.

In a spallation reaction, a heavy or intermediate mass target nucleus is hit by a very high energy ($E > 10MeV$) incident particle (p or $\alpha$) and subsequently broken into several fragments. These reactions are very endoenergetic; f. ex., the following one, studied in laboratory,

$$p + {}^{12}C \rightarrow {}^{6}Li + 3n + 3p + p' + Q$$

requires a threshold energy $E_T = -Q = 30$ MeV. Note that p′ is the incident projectile p with a different energy.

A spallation reaction is usually described by a 2-step procedure: first, when crossing the target nucleus, the incoming p collides with a few nucleons which in turn generate more collisions (cascade) and some of the nucleons have enough energy to leave the target nucleus with the incident $p \rightarrow p'$. After this rapid step ($\lesssim 10^{-21}$sec), the resulting excited nucleus will dissipate its energy more slowly ($10^{-16}$sec) by evaporating a few more nucleons.

This scheme (Serber model) accounts for the behaviour of the spallation cross-section $\sigma(E)$ which, according to experimental measurements, increases with energy E up to a maximum before decreasing to an asymptotic value, the more sharply the lower the threshold energy $E_T$.

It has been suggested by Reeves and others about 10 years ago, that the formation of the light elements Li, Be, B is due to the interaction between the Galactic Cosmic Rays (GCR) and the interstellar medium.

These highly energetic GCR, presumably originating in explosive objects such as supernovae, anyway coming from outside the Solar System, interact strongly with the terrestrial atmosphere before reaching the ground. This GCR flux $\phi$ appears to be isotropic, time-independent and follows, for the higher energies ($E \geq 2GeV$/nucleon) a power law spectrum $\phi \propto E^{-2.7}$ while at lower energies ($< 1GeV$/nucleon) $\phi$ is strongly perturbed by the Sun.

As rather conspicuous signatures of the interaction between interstellar medium and GCR, let us mention the following peculiarities in the GCR chemical composition: the ratio of odd Z to even Z nuclei is larger than in the standard abundances (cf. figure 7) and the ratio of the light elements to the medium mass elements is very much ($10^4$) larger than its value in the universal cosmic abundances. This means that along their path in space, the primary rapid (C, N, O, ...) GCR particles hit interstellar H and He atoms, resulting in spallation reactions creating odd nuclei and light elements Li (in general more $^{6}Li$ and $^{7}Li$), $^{9}Be$, $^{10}B$, $^{11}B$.

On the other hand, it appears quite impossible to explain the observed deuterium

abundance ($D/H \sim 10^{-5}$) in the solar system and neighbourhood by spallation reactions due to cosmic rays.

## 16. Big Bang Nucleosynthesis

Two basic observational facts, namely the recession of the galaxies and the 2.8 K blackbody background radiation, both point to a very dense and hot initial state called Big Bang, from which the universe has started its present evolution. G. Tammann will come back on the evidence for the Big Bang in his forthcoming lecture.

The basic assumptions on which rests any Big Bang model are the validity of the same physical laws (not depending on gravity) everywhere in the universe and the existence of this very dense and hot initial state of the universe when all particles were in equilibrium. Other more restrictive assumptions have to be made, such as homogeneity, isotropy of the universe and a positive baryon number, in order to define the so-called Standard Model which has served as a framework for the nucleosynthesis calculations initiated by R. Wagoner, W. Fowler and F. Hoyle a decade ago.

Without speculating here about the very first instants of expanding universe, let us recall that 10 seconds after the beginning of the expansion, the universe is thought to have reached the end of the so-called lepton era, when energetic photons ($>1$ MeV) maintained their equilibrium around 10 billion degrees with light particles (leptons) like electrons, muons, neutrinos, in the presence of some baryons left over from previous hadron annihilation, namely protons and neutrons, in about equal amounts through exchange of leptons. When the temperature sinks below $10^{10}$ K as the universe continues to expand, the neutrons, more massive, become less numerous than the protons, soon stabilizing at a ratio of about 1 n for 5 p's. Meanwhile, the electron-positron pairs have annihilated, leaving a small remainder of electrons.

The primordial cosmological nucleosynthesis is now ready to take place: the first reaction to retain, namely $n + p \rightarrow D + \gamma$, is still practically reversible above 5 billion degrees, so that no deuterium $D = {}^2H$ is yet really produced. Below that temperature, however, when the age of the universe is close to 3 minutes, there will be enough D obtained for subsequent reactions (already met in the p-p chains of H-burning in stars) to produce $^4$He, with a little $^3$He, $^7$Li. The calculations were carried out assuming, in this very early evolutionary stage, a flat Friedmann model universe submitted to an adiabatic expansion: equilibrium is achieved for any reaction when the reaction rate largely outweighs the expansion rate. After 1000 seconds, the temperature has dropped below 300 million degrees and the density is then too low for any other reaction to proceed; consequently, in a quarter of an hour, D and $^4$He have reached their equilibrium values and no heavier element can be built.

The $^4$He abundance in mass, quite sensitive to the values of the n/p ratio and of the expansion rate, is close to 25% in the present conditions of the Standard Model for the Big Bang: this is in good agreement with the overall abundance of helium observed in the universe. Thus, the only nuclear species produced in the primordial cosmological nucleosynthesis are D, $^3$He, $^4$He, $^7$Li and, apart from $^7$Li also synthesized in stars, these are precisely those isotopes that neither stellar nucleosynthesis nor interstellar spallation reactions could account for. Finally we should mention as an important result of the former considerations, that deuterium in the Standard Big Bang is produced in sufficient amount to explain the present observed abundance value $D/H \sim 10^{-5}$, only if today's density of the universe does not exceed $5 \times 10^{-31}$ gcm$^{-3}$ and this points definitely to an open universe (insofar as the neutrinos reveal no significant rest mass).

## 17. Conclusion

At the end of this condensed survey about star formation and nucleosynthesis, it is worth emphasizing first the still approximate character of the description obtained today and outlined above. The picture of protostar evolution, mentioned in section 5, might soon improve on account of the infra-red measurements presently being carried out on interstellar molecular cloud complexes.

With respect to nucleosynthesis, many uncertainties stand in our way to follow accurately the advanced stages of nuclear burning in stars, and if the chemical element abundances resulting from explosive nucleosynthesis do represent an improvement over the ones deduced from mere hydrostatic burning, in the sense that they appear to reproduce rather successfully the Solar System abundances, we should keep in mind that what we know of the Solar System may not reflect a truly representative sample of the galactic medium.

Moreover, the exact way to describe an exploding star remains badly known and, as regards the cosmological nucleosynthesis, we must remember that its interpretation rests on the Standard Big Bang model, itself subject to many simplifying assumptions and therefore possibly rather far from the real physical universe.

In spite of all these shortcomings, this description outlined here as resulting from the investigations of the last two decades, does not lack a certain majesty; it leads us to face a universe starting in the hot Big Bang initial state with pre-existing radiation, matter and antimatter, expanding then during more than 10 billion years, undergoing various inner modifications, and giving birth to stars and stellar systems.

In consequence of this evolution, it turns out that all the different chemical species composing every bit of matter existing today (including our own bodies!) have originated, in a remote past, inside stars or eventually during the first minutes after the lepton era, when our universe was still in its glowing earliest infancy.

## References


Allen, C.W.: 1973, Astrophysical Quantities (3rd ed., The Athlone Press, London).

Appenzeller, I.: 1980, in 'Star Formation' 10th Saas-Fee Course (Ed. Geneva Observatory).

Audouze, J., and Vauclair, S.: 1980, 'An Introduction to Nuclear Astrophysics' (Reidel Pub. Co.).

Larson, R.B.: 1974, 'Fundamentals of Cosmic Physics', V. 1 (A.W. Cameron, ed.).

Maeder, A.: 1980, Astron. Astrophys. 92, 101.

Reddish, V.C.: 1978, 'Stellar Formation' (Pergamon Press, Oxford).

Regar, O., and Shaviv, G.: 1981, Astrophys. J. 243, 934.

Schramm, D.N.: 1977, in 'Advanced Stages in Stellar Evolution', 7th Saas-Fee course (Ed. Geneva Observatory).

Snell, R.L.: 1981, Astrophys. J. Suppl. Series 45, 21.

Wallace, R.K. and Woosley, S.E.: 1981, Astrophys. J. Suppl. Series 45, 388.

Wynn-Williams, G.: 1981, Scientific Amer. 245, No. 2.



*Address of the author:*

Prof. Dr. Pierre Bouvier
Observatoire de Genève
Chemin des Maillettes
CH-1290 Sauverny (Switzerland)

# The Growth of Structure in the Universe

F. Occhionero, N. Vittorio, M. Boccadoro, S. De Luca

## 1. Visible and Invisible Matter

a) Cosmological Deuterium

In cosmology we use as a reference value for the present cosmic density the so-called critical density,

$$\rho_{crit} = 3\,H_0^2/(8\,\pi\,G) = 2\times 10^{-29} h^2\,g\,cm^{-3},$$

$$H_0 = 100\,h\ (km/s/Mpc), \qquad \frac{1}{2} \gtrsim h \gtrsim 1\,; \quad (1)$$

determinations of $H_0$ are done by several authors, Sandage and Tammann (1976), de Vaucouleurs and Bollinger (1979), Aaronson et al. (1980). It is also convenient to use the ratios

$$\Omega_0 = \rho_0/\rho_{crit}, \qquad \Omega_{B_0} = \rho_{B_0}/\rho_{crit}, \ldots, \qquad (2)$$

for the total density and its partial components. (The subscript "0" refers here and below to the present epoch.) For $\Omega_0 \leqslant 1$, the Universe is spatially open or flat and energetically hyperbolic or parabolic - i.e. it will expand forever - while for $\Omega_0 > 1$ the Universe is spatially closed and energetically elliptic - i.e. it is bound to collapse. This beautiful interaction between energetics and geometry is born from General Relativity, as we will see.

Primordial nucleosynthesis, occurring at the end of the first three minutes (e.g. Weinberg 1972 and 1977) gives us a powerful theoretical tool to evaluate $\rho_0$ or - better - its baryonic component, $\rho_{B_0}$. At temperatures $T > 10^{10}$ K ($t < 1$ sec) weak interactions and $\beta$-decay keep the neutron to proton number density ratio to its equilibrium value,

$$\exp\{-\Delta m\,c^2/k\,T\},$$

where $\Delta m$ is the mass difference between neutron and proton; meanwhile deuterium is formed and destroyed:

$$n + p \rightleftarrows D + \gamma\,.$$

As the temperature drops the weak interaction rate falls below the expansion rate ($\sim 1/t$): at $T = 10^{10}$ K the two rates are equal and the neutron to proton ratio freezes out; this ratio decreases then slightly further due to neutron decay. At $10^9$ K deuterium is no longer destroyed, the bottleneck is broken and $He^4$ is formed; due to the absence of any stable nucleus at mass 5, all the nucleons end up in $He^4$. The abundance of the latter, Y, is therefore basically twice the abundance of neutrons. Therefore the cosmological abundance of $He^4$ depends essentially on the rate of the cosmological expansion during the nucleosynthesis, which is directly related to the energy density of the photons and the relativistic particles ($e^{\pm}$ and $\nu$) at $10^9$ K. The abundance of deuterium depends instead on the competition between formation and destruction in two body reactions: it is therefore sensitive to the nucleon density at nucleosynthesis, which is related to the present nucleon (or baryon) density. The above argument is made precise by detailed study of the time evolution of the abundance of the various nuclei by the numerical integration of the appropriate differential equations (e.g. Schramm and Wagoner 1977, Steigman 1979): in particular the amount of deuterium that survives decreases steeply as the present baryon abundance, $\rho_{B_0}$ or $\Omega_{B_0} h^2$, increases.

The observed amount of deuterium (York and Rogerson 1976, Vidal-Madjar et al. 1977) is large, $\chi_D \cong 2\times 10^{-5}$. If it is of cosmological origin - which is not obvious since it might have been created and destroyed elsewhere; see the discussion by Greenstein (1980) - the implication that

follows (Gott et al. 1974) is that the present baryon density is low, of the order of some units in $10^{-31}$ g/cm$^3$: hence

$$\Omega_{B_0} h^2 \cong 10^{-2}. \quad (3)$$

This takes into account all the baryonic matter that has been processed in the Big-Bang, irrespective of whether such matter at present is or not in a luminous form.
The standard Big-Bang scenario sketched above has been recently further exploited (Yang et al. 1979; see also Shvartsman 1969, Dolgov and Zel'dovich 1981) to set an upper limit on the number of lepton species, $N_L$. In this case, $He^4$ abundance must be used: in fact adding the associated neutrino flavors to the cosmic medium increases the energy density of relativistic particles and decreases the age of the Universe,

$$t \propto \varepsilon^{-1/2}.$$

Hence more neutrons are present and more $He^4$ is formed: from a limit $Y \gtrsim 0.25$, it is concluded that $N_L \gtrsim 3$; thus there should not be many more leptons beyond the known electron, the muon and the newly discovered tau. We will use this estimate later on (see, however, Stecker (1980) for a different point of view).

b) Luminous Matter

Cosmology has also straightforward observational arguments to evaluate the mean cosmic density. Zwicky in the 30's observed that in order to bind the Virgo Cluster it is necessary to have 500 times more mass than it is apparently there; this started an important line of research, that of the missing mass, wherein cosmologists try to discover whether or not in galaxies and their associations there is more mass than is directly responsible for the observed electromagnetic emission. The issue has become more compelling, already at the level of individual galaxies, after the remark by Ostriker and Peebles (1973) that the thin disks of spiral galaxies would not be stable against bar instability unless they were embedded in massive halos. Observational support of this theoretical speculation has been strong particularly from 21-cm observations showing constant rotational velocities of HI-clouds at large distances from galactic centers. For this and other reasons, we have now little doubts about the existence of the missing mass and we prefer to consider it hidden or dark and we rather speak of missing light.
A convenient tool for the investigation of this problem are mass-to-light ratios (in solar units, M/L); a recent review of this subject is given by Faber and Gallagher 1979. What we see there is an escalation of M/L from the small to the large systems: thus M/L is of the order of unity (or slightly larger) in the solar neighborhood, of the order of 10 for spiral galaxies, around 20 for ellipticals and SO's, of the order of 100 for binaries and small groups and finally of several hundreds for cluster of galaxies (on the latter issue see also Bahcall 1977 and Hoffman et al. 1980).
In particular it happens that M/L is large whenever evaluated by dynamical methods: thus for spiral galaxies, for instance, while M/L has acceptable values for the inner regions out to 20 kpc, when we move beyond 50 kpc the flatness of the rotation curves pushes M/L above 100. Clearly dark matter dominates there; incidentally its spatial distribution can be easily inferred: on the assumption of centrifugal equilibrium, $v^2 = G M(r)/r$, the constancy of the rotational velocity implies that $M(r) \propto r$ and hence that $\rho \propto r^{-2}$, which is reminiscent of isothermal spheres. Likewise, estimates of the mass of our own Milky Way from tidal effects on globular clusters or from globular cluster radial velocities place the mass above $10^{12}$ $M_\odot$ and M/L around 70. A value consistent with the latter can also be found from the dynamics of the Local Group, which, as we know, is dominated by M31 and the Milky Way: if the velocity of approach of the two galaxies arises from their mutual gravitational interaction, the total mass must be of the order of some units in $10^{12}$ $M_\odot$ and M/L consistently ranges up to 60.
Dynamical methods are also used to determine the mass of great clusters: most commonly the virial theorem; it yields values of M/L of several hundreds and thus much larger than the M/L's of the constituent galaxies. On the contrary the X-ray emission from the cluster cores (Lea et al. 1973, Cava-

liere and Fusco-Femiano 1976, Malina et al. 1978) accounts only for a small fraction, $\sim 10\%$, of the virial mass.

We may now try to correlate mass-to-light ratios to masses and hence to representative densities: the standard technique multiplies M/L by an average luminosity function (Kirschner et al. 1979) and obtains a mass density. When we use the (M/L)'s of the solar neighborhood we obtain density limits in qualitative agreement with (3) above; the same is true when we estimate the mean density from the mass of hot gas in clusters of galaxies. On the contrary, the virial theorem, mass-to-light ratios of great clusters and the analysis of the correlation function send $\Omega_0$ to values of the order of unity (Davis et al. 1978).

Thus observational cosmology suggests the view that most of the cosmic matter is in some hidden form, of which all we know is that it gravitates and that it is very likely dissipationless (Gunn 1978). If so, the fact that unseen matter is needed more at the larger scales, may be related to the fact that on the scales of galaxies ordinary baryonic matter did have the time to cool and sink in the potential wells; on the scales of cluster of galaxies instead, cooling times are longer that the age of the Universe and the separation between visible and invisible matter has not yet occurred.

## 2. Gravitational Instability in the Universe

### a) The Jeans and the Silk Masses

Our Universe is homogeneous on the large scales ($> 100$ Mpc), but shows a considerable amount of clumpiness and structure at the small ones ($\lesssim 10$ Mpc). Among the main tasks of modern cosmology is the explanation of that degree of structure; we think it is a problem of following theoretically the evolution of this structure as it grows by self-gravitation from a slight perturbation in an initially uniform and expanding medium. This view meets serious difficulties on the mass scales of galaxies, as we will see, even if we postulate very "ad hoc" initial conditions. Thus, the situation is far from satisfactory; there are hopes however that we are close to a major breakthrough. In the sequel we will review the basic facts following Weinberg (1972).

We define our vocabulary starting from the elementary theory of Jeans instability: in a uniform (hence infinite) self-gravitating medium the evolution of a small, linearizable perturbation of all the quantities describing the fluid is studied via the equation of continuity, Euler's and Poisson's equation. We condense all this in a simple, second-order, partial differential equation for the density enhancement $(\delta\rho/\rho)$,

$$\left(\frac{\partial^2}{\partial t^2} - c_s^2 \nabla^2\right) \frac{\delta\rho}{\rho} = 4\pi G \rho \frac{\delta\rho}{\rho}, \tag{1}$$

where $c_s$ is the sound speed. Clearly we seek a solution of the form

$$\frac{\delta\rho}{\rho} \propto \exp\{i(\mathbf{k}\cdot\mathbf{x} - \omega t)\}, \tag{2}$$

and we find the elementary dispersion relation

$$\omega^2 = k^2 c_s^2 - 4\pi G \rho. \tag{3}$$

The latter tells us that on the small wavelength side of the perturbation spectrum, we have genuine sound waves, $\omega^2 > 0$, while on the large wavelength side we have an instability, $\omega^2 < 0$, with two exponentially growing and decaying modes; in the limit $k \to 0$, the e-folding time is given by

$$\frac{1}{\tau} = \sqrt{4\pi G \rho}. \tag{4}$$

The separation between the two regimes occurs at a Jeans wavenumber

$$k_J = (4\pi G \rho / c_s^2)^{1/2}; \tag{5}$$

the corresponding wavelength is

$$\lambda_J = 2\pi / k_J. \tag{6}$$

In the sequel in order to apply the concept of gravitational instability to expanding cosmological models, where a length would not be an invariant quantity, we prefer to introduce a Jeans mass

$$M_J = \frac{4\pi}{3} \rho \left(\frac{\lambda_J}{2}\right)^3. \tag{7}$$

Furthermore it is convenient to modify and generalize (7) in two ways. Firstly, we want to make sure that $\rho$ in (7) contains only the proper mass density of the constituent particles without any contribution from the internal energy; thus we write explicitly in place of $\rho$ the product $n \times m_H$ of the number density times the mass of the individual particles, protons, say. In this way we can compare the behavior of a given rest mass under different conditions in the history of the Universe, regardless of the associated thermal energy. Secondly, we observe that (6) gives us a measure of the strength of the gravitational field and, in order to generalize (7) to the case of strong fields, we recall that in general relativity not only the rest-mass but any form of energy and pressure feel the gravitational field: we thus replace $\rho$ in (5) by $(\varepsilon + p)/c^2$. The more general expression for the Jeans mass,

$$M_J = \frac{\pi}{6} n m_H \left[ \frac{\pi c^2 c_s^2}{G(\varepsilon + p)} \right]^{3/2}, \tag{8}$$

is of interest in radiation dominated cases.
We can now apply (8) to various regimes of interest: thus, for instance, before hydrogen formation, $T > 4000$ K, the cosmic medium is a mixture of black-body radiation, $\varepsilon = a T^4$, $p = \varepsilon/3$, and non-relativistic protons, $\varepsilon = n \times m_H \times c^2$, $p = 0$. Then (8) yields

$$M_J \cong \eta^2 \left( 1 + \eta \frac{kT}{m_H c^2} \right)^{-3} M_\odot, \tag{9}$$

where

$$\eta = \frac{4 a T^3}{3 n k}, \tag{10}$$

is the specific entropy or photon-to-baryon ratio. The latter is a large number, $10^8 \div 10^{10}$ (see also the recent estimates due to Olive et al. 1981, from nucleosynthesis and mass-to-light ratios). Then (9) has a high temperature limit

$$M_J \cong (m_H c^2/kT)^3/\eta \, M_\odot, \tag{11}$$

and a low temperature limit where it levels off at the very high value

$$M_J \cong \eta^2 M_\odot. \tag{12}$$

As hydrogen forms, $z_{dec} \cong 10^3 \gtrsim z_{eq}$, the picture changes substantially because radiation disappears from the budget

$$\varepsilon = n m_H c^2 + \frac{3}{2} n k T, \qquad p = n k T, \qquad c_s^2 = \frac{5}{3} \frac{kT}{m_H}, \tag{13}$$

then (8) yields

$$M_J \cong \frac{1}{2} \left( \frac{5 k T}{G} \right)^{3/2} n^{-1/2} m_H^{-2}, \tag{14}$$

which starts as low as

$$M_J \cong 10 \, \eta^{1/2} M_\odot \cong 10^5 \div 10^6 M_\odot$$

at $z_{dec}$ and drops thereafter as $R^{-1.5}$. The above behavior of the Jeans mass is sketched together with the Jeans mass of the massive neutrinos in figure 4.
Comparing the Jeans mass with the typical mass of galaxies, $M_G \cong 10^{11} M_\odot$, we see that there are three regimes of interest. In the first phase, $M_G > M_J$, any oscillation involving a mass of the order $M_G$ will grow due to self-gravity; in reality this growth would occur in a radiation dominated epoch which must be studied with a general-relativistic treatment (see sect. 2b). After this, there is a second phase where $M_G < M_J$ in which any perturbation on the scale of $M_G$ oscillates at constant amplitude; whether or not a relativistic treatment is needed, depends on the value of $\eta$ in the sense that, as before, a large amount of radiation cannot be described properly in Newtonian terms. Finally, there is a third and last phase $z < 10^3$, where $M_G > M_J$ and any matter perturbation on the scale $M_G$ grows by self-gravitation in a matter dominated background.
The above discussion may not seem to relate very strongly to objects like galaxies, since it does not single out a mass of the order of $M_G$. Silk (1968), Peebles and Yu (1970) and Weinberg (1971) show instead that a mass around $M_G$ comes very naturally into play when dissipation mechanisms are taken into account.
Dissipation arises as a consequence of the imperfect coupling between the photon and the baryon component of the cosmic

medium; this is the case when the photon mean free path for collisions with the electrons

$$l_\gamma = (n_e \sigma_T)^{-1}, \tag{15}$$

(where $\sigma_T = 2/3 \times 10^{-24}$ cm$^2$ is the Thomson cross-section) becomes long enough for the photon to random-walk out of the perturbation. Obviously this occurs increasingly as ionization decreases.
In first approximation the physics of this phenomenon may be studied by treating photons and baryons as a non-perfect fluid endowed with shear and bulk viscosity and heat conductivity (they are all proportional to $l_\gamma$). It is found that a sound wave of mass M will be damped at the end of ionization by a factor

$$\exp\{-(M_D/M)^{2/3}\}. \tag{16}$$

Thus, the attenuation will be negligible if $M \gg M_D$ and will be substantial in the opposite case: recent estimates (Jones 1976) give for the Silk mass the expression

$$M_D \cong 10^{13} (\Omega_{B_0} h^2)^{-5/4} M_\odot. \tag{17}$$

A fluctuations of the mass of a galaxy will suffer substantial damping and will unlikely survive to the matter dominated era.

b) The Influence of the Cosmic Expansion

The expansion of the Universe is conveniently described by the time evolution of the familiar scale factor R(t) of the Friedmann-Robertson-Walker (FRW) line element. In the simplest case R(t) obeys the Einstein field equation (see later)

$$\dot{R}/R = \sqrt{8\pi G \rho/3}. \tag{1}$$

Thus, the expansion time scale is uncomfortably close to the time scale for gravitational collapse given in (a.4). Since the formation of galaxies occurs in the expanding Universe, we are forced to generalize our theory of gravitational instability to the case of an expanding medium. This has been done by Lifshitz in 1946 in the full glory of general relativity; a much simpler yet very indicative Newtonian approach to this problem was given by Bonnor 1957; a recent review is given by Field 1975. On the other hand, when we consider sound waves

$$k \gg k_J,$$

frequencies are very large and the expansion of the Universe may be neglected.
Let us review briefly the description of the unperturbed model. The assumptions that are commonly accepted in cosmology are that the correct theory of gravity is general relativity where the gravitational field is the metric tensor.

$$ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta, \qquad \alpha, \beta = 0, 1, 2, 3, \tag{2}$$

and the dynamical equations of motion are Einstein's field equation,

$$G_{\alpha\beta} - \Lambda\, g_{\alpha\beta} = \frac{8\pi G}{c^4} T_{\alpha\beta}. \tag{3}$$

$$T^\beta_{\alpha;\beta} = 0. \tag{4}$$

We have introduced for complete generality the cosmological term, $\Lambda$; $G_{\alpha\beta}$ is the Einstein tensor and $T_{\alpha\beta}$ the energy-momentum tensor for which we use the perfect fluid expression,

$$T^{\alpha\beta} = (\varepsilon + p)\, u^\alpha\, u^\beta - p\, g^{\alpha\beta},$$
$$u^\alpha = dx^\alpha/ds. \tag{5}$$

Once we specialize the metric tensor (2) to the FRW form,

$$ds^2 = c^2\, dt^2 - \frac{R^2(t)}{(1 + k r^2/4)^2}(dx^2 + dy^2 + dz^2), \tag{6}$$

appropriate for dealing, in comoving coordinates ($u^0 = 1, u^k = 0$) with a uniform medium expanding isotropically, from (3) we obtain

$$\frac{\ddot{R}}{R} = -\frac{4\pi G}{3c^2}(\varepsilon + 3p) + \frac{1}{3}\Lambda c^2, \tag{7}$$

$$\left(\frac{\dot{R}}{R}\right)^2 + \frac{k c^2}{R^2} = \frac{8\pi G}{3c^2}\varepsilon + \frac{1}{3}\Lambda c^2. \tag{8}$$

Equation (7) shows the deceleration of the cosmic expansion under self-gravity (to

which pressure gives a contribution) and the accelerating role played by a positive $\Lambda$; (8), which generalizes (1), is a first integral of (7). Under the assumption (5), (4) yields an energy conservation equation,

$$\frac{d}{dt}(\varepsilon R^3)+p\frac{d}{dt}R^3=0. \tag{9}$$

Together with an equation of state, which we usually assume of the form

$$p=\gamma\varepsilon, \qquad \gamma=0,\frac{1}{3},1, \tag{10}$$

to deal at least schematically with dust, radiation and a maximum stiffness fluid ($c_s=c$), we have now stated all the rules, equations (8), (9) and (10), which a cosmological model must obey.

For every particle component we have a continuity equation

$$(n u^\alpha)_{;\alpha}=0, \tag{11}$$

which yields the conservation of total number in comoving volume:

$$n R^3=\text{const}. \tag{12}$$

We now perturb the equilibrium solution given above by introducing small and linearizable changes in the thermodynamic quantities, $\delta n, \delta\varepsilon, \delta p, \delta u^\alpha$ and in the gravitational field, $\delta g_{\alpha\beta}$; for the latter we limit ourselves without loss of generality to the gauge $\delta g_{0\alpha}=0$.

Lifshitz' solution contains also radiative and rotational modes which decay away with the expansion of the Universe and which we will not consider. The compressional normal modes are found to obey a relatively simple equation in the long wavelength limit appropriate to study the gravitational instability (a straightforward derivation is given in Harrison 1967).

By perturbing the (00)-component of the field equations (3), we find

$$\frac{1}{R^2}\frac{d}{dt}R^2\frac{d}{dt}\delta g = \frac{8\pi G}{c^2}\left[1+3\left(\frac{c_s}{c}\right)^2\right]\delta\varepsilon, \tag{13}$$

where

$$\delta g=-(\delta g_{11}+\delta g_{22}+\delta g_{33}), \tag{14}$$

and the assumption of adiabaticity, $\delta p=c_s^2\delta\varepsilon$, has been used. By perturbing the (0)-component of (4) and the equation of continuity (11), we find for $\lambda\to\infty$, respectively

$$\frac{d}{dt}\frac{\delta\varepsilon}{\varepsilon+p}=\frac{1}{2}\frac{d}{dt}\delta g, \tag{15}$$

$$\frac{d}{dt}\frac{\delta n}{n}=\frac{1}{2}\frac{d}{dt}\delta g. \tag{16}$$

The comparison between the last two gives the result

$$\frac{\delta n}{n}=\frac{\delta\varepsilon}{\varepsilon+p}, \tag{17}$$

which is again a statement of the adiabaticity of the perturbation. If we replace in (13) $\delta g$ and $\delta\varepsilon$ by (16) and (17), we end up with a single differential equation

$$\frac{1}{R^2}\frac{d}{dt}R^2\frac{d}{dt}\frac{\delta n}{n} - \frac{4\pi G}{c^2}(\varepsilon+p)\left(1+3\frac{c_s^2}{c^2}\right)\frac{\delta n}{n}. \tag{18}$$

This is the sought generalization of the limiting case of (a.1) for $\lambda\to\infty$; it takes into account the expansion and the strong gravitational fields (with the regeneration of the pressure) and is valid for any value of the curvature and of the cosmological constant.

The most important novelty of (18) is that it replaces the exponential growth of the Jeans instability with a much more modest law, tipically a power law. To see this, let us assume $k=0$ for simplicity and let us look for a solution in the form

$$\frac{\delta n}{n}\propto t^\alpha. \tag{19}$$

For the three values of $\gamma$ in (10) we find the results of table 1, where $a_\pm$ give the growing and the damping mode, respectively.

The most important result we read in table 1 is that in an Einstein-de Sitter Universe

$$\Omega_0=1, \qquad \frac{1}{1+z}\propto t^{2/3}.$$

*Table 1*

| | R | $\varepsilon$ | $\frac{c_s}{c}$ | $a_+$ | $a_-$ |
|---|---|---|---|---|---|
| $\gamma=0$ | $t^{2/3}$ | $\frac{c^2}{6\pi G t^2}$ | 0 | $\frac{2}{3}$ | $-1$ |
| $\gamma=\frac{1}{3}$ | $t^{1/2}$ | $\frac{3c^2}{32\pi G t^2}$ | $\frac{1}{\sqrt{3}}$ | 1 | $-1$ |
| $\gamma=1$ | $t^{1/3}$ | $\frac{c^2}{24\pi G t^2}$ | 1 | $\frac{4}{3}$ | $-1$ |

the growing mode amplifies in the linear regime according to the law

$$\frac{\delta n}{n} \propto t^{2/3} \propto \frac{1}{1+z}. \tag{20}$$

In particular, the amplification available between decoupling $1+z_{dec}=10^3$, and the present is just by a factor $10^3$; thus a perturbation which enters non-linearity at the present ($\delta n/n=1$ at $1+z=1$) at decoupling had an amplitude $\delta n/n=10^{-3}$. The latter amplitude should be observed in the form of a small scale ($\gtrsim 1^0$) distortion in the microwave background: under the assumption of perfect adiabaticity, in fact

$$\left(\frac{\delta T}{T}\right)_{dec} = \frac{1}{3}\left(\frac{\delta n}{n}\right)_{dec}. \tag{21}$$

On the contrary (see Partridge 1980) the experimental limit, which is also the present limit of sensitivity of our detectors, is already down to

$$\frac{\delta T}{T} \gtrsim 10^{-4}, \tag{22}$$

and seems to indicate the existence of a conflict between theory and observations. There are ways out: the most obvious is that there might have been a reheating (early star formation) and consequent reionization of the cosmic medium. In that case further Thomson scattering would have blurred any imprint left in the microwave background by condensing galaxies or larger objects. Alternatively we must take into account the possibility that the coupling between matter and radiation fluctuations is minimum (as in isothermal fluctuations) and not maximum as in (21) (Davis and Boynton 1980): in that case the theoretical predictions for purely isothermal fluctuations are certainly below the present detection threshold, but an order of magnitude improvement of the current instrumental sensitivity will critically test the gravitational instability picture for galaxy formation.

Equation (18) is valid in all generality: let us consider the case of dust ($p=0$, $c_s=0$); it reduces to

$$\delta=\frac{\delta\rho}{\rho},$$

$$\left[\frac{1}{R^2}\frac{d}{dt}R^2\frac{d}{dt}-4\pi G\rho\right]\delta=0, \tag{23}$$

and in this form analytic solutions are known for any value of $\Omega_0$ (provided $\Lambda=0$). In particular $\Omega_0<1$ ($k=-1$) is of interest according to the considerations of section 1 for a low-density baryon cosmology. We plot in figure 1 the solution for the growing modes given in Weinberg 1972. Inspection of the figure gives us as a rule of the thumb the notion that in the $\Omega_0<1$ case the growing mode grows as in (20) only for $z\gtrsim\Omega_0^{-1}$, but levels off thereafter. Thus for $\Omega_0=0.1$ the total amplification available is only 100 (and not 1000 as in the Einstein-de Sitter case): temperature fluctuations in the microwave background of amplitude around $10^{-2}$ could be expected and the lack of their detection is disturbing (Boynton 1978).

A solution of this problem is given by Doroshkevich et al. (1974) (see also Gott 1979): they consider the case of perturbations obeying the Zel'dovich (1970) condition which is the assumption that a) all the perturbations are purely adiabatic and that b) they have the same amplitude $A=10^{-4}$ on all mass scales when they enter the horizon (see also Press 1980). These authors find that on scales smaller than the Silk mass ($\sim 10^{14}\ M_\odot$) all the fluctuations are damped as expected by photon viscosity, while on scales just larger than the Silk mass the fluctuations not only survive, but also undergo a two-order of magnitude amplification due to "velocity overshoot".

Press and Vishniac (1980) have however

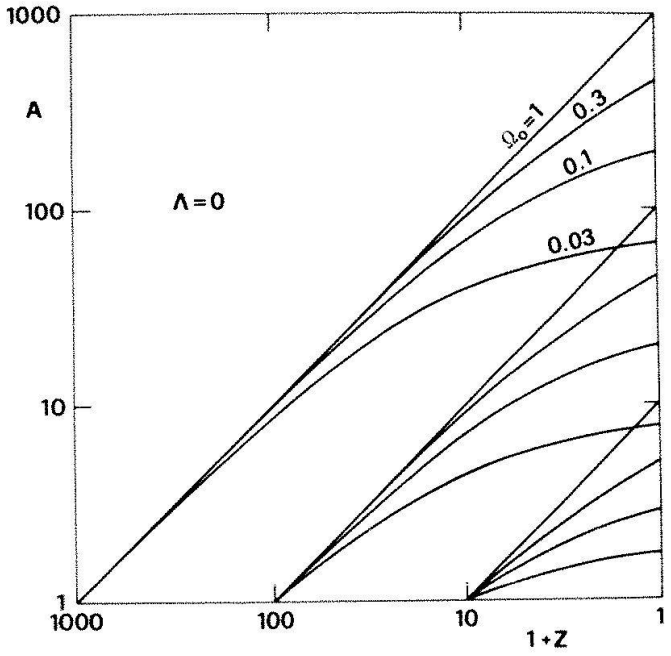



*Fig. 1.* Plot of the amplification of the growing modes in cosmological dust models with $\Lambda = 0$ and $\Omega_0 \leqslant 1$ vs. redshift. In the Einstein-de Sitter ($\Omega_0 = 1$) case the straight diagonal lines apply: thus if $\delta\rho/\rho = 1$ at the present the initial $\delta\rho/\rho$ was $10^{-3}$, $10^{-2}$ and $10^{-1}$ at $1 + z = 10^3$, $10^2$ and 10, respectively. For open models, $\Omega_0 < 1$, however, the amplification is reduced considerably with disturbing implications on the microwave background: however, a low-density model may be closed ($k = +1$) when $h \cong 1$ and $\Lambda > 0$ because

$$k \propto (\Omega_0 - 1) + \Lambda c^2/(3 H_0^2).$$

In that case the growing modes amplify better than in the $\Omega_0 = 1$ case, as shown in figure 2.

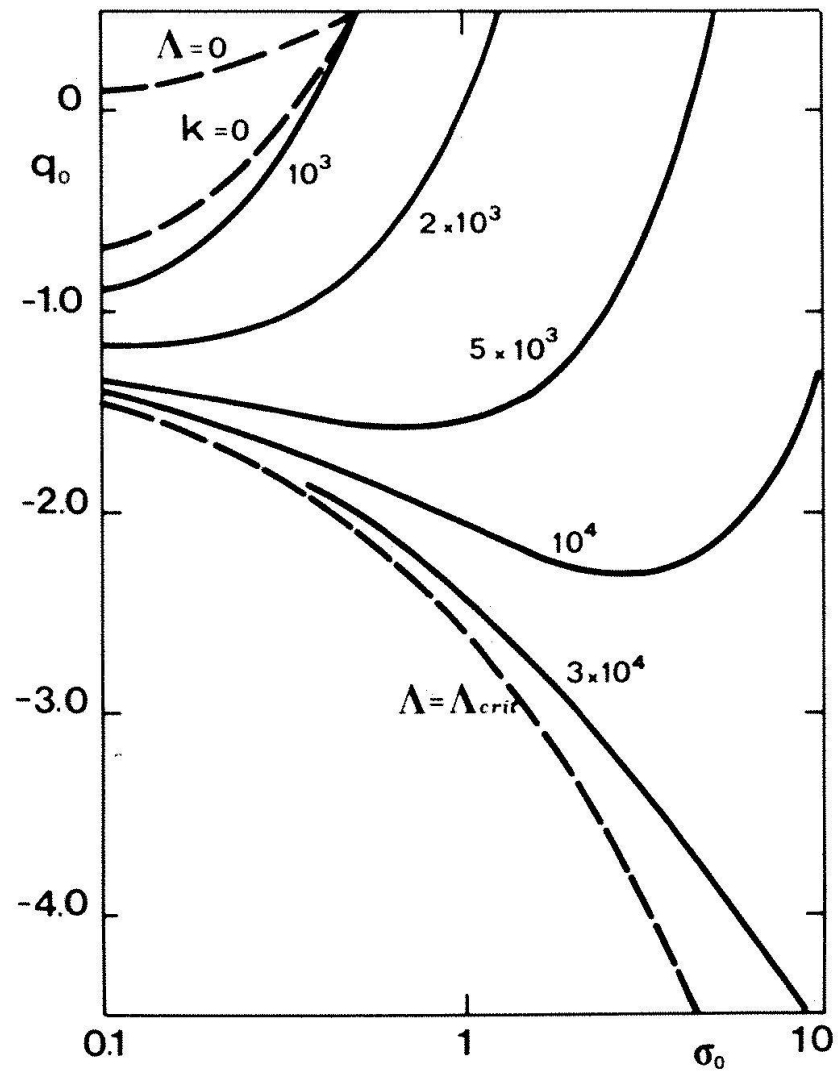



*Fig. 2.* Level curves of the amplification of the growing modes in a pure baryon Universe assuming that the linear growth starts at $z_{dec} = 1000$. Broken lines define the constraints $k = 0$, $\Lambda = 0$ and $\Lambda = \Lambda_{crit}$. The labels on each curve define the total amplification available on that curve; thus the curve labelled $10^3$ goes through the Einstein-de Sitter model $\sigma_0 = q_0 = 0.5$. This plot shows that in the region of the plane where $k = +1$ and $\Lambda > \Lambda_{crit}$ the linear amplification may exceed considerably that of an Einstein-de Sitter model. These results must be compared with those given in figure 10 below.

shown that the effect is in reality a consequence of an inaccurate treatment of decoupling, i.e. of the assumption that it is instantaneous. An accurate treatment of hydrogen formation shows no sign of overshooting in agreement with previous results of Peebles and Yu (1970).

A possibility of removing the mentioned difficulties with the microwave background has been described by Occhionero et al. (1980); it is found investigating the dust solution of (23) after allowing, for complete generality, also for a non-vanishing $\Lambda$ term. In that case, dust models form a biparametric set: the first parameter is the density parameter $\Omega_0$ we met in (1a.2) and which we now replace by

$$\sigma_0 = \Omega_0/2 . \tag{24}$$

in order to agree with the established conventions (e.g. McVittie 1965), while the second parameter is the deceleration parameter

$$q_0 = -(R\ddot{R}/\dot{R}^2)_0 . \tag{25}$$

In terms of these

$$\begin{aligned} \frac{k c^2}{(H_0 R_0)^2} &= 3\sigma_0 - q_0 - 1 , \\ \frac{\Lambda c^2}{3 H_0^2} &= \sigma_0 - q_0 ; \end{aligned} \tag{26}$$

clearly for $\Lambda = 0$ the $(\sigma_0, q_0)$ parameter plane degenerates into a straight line.

In figure 2 we present some of the results: broken lines define the curves $k = 0$, $\Lambda = 0$ and $\Lambda = \Lambda_{crit}$. (It is known from the general theory of the FRW models that when $k = +1$ we must have $\Lambda > \Lambda_{crit}$ otherwise the cosmologic expansion reverts into a collapse.) Solid lines define the loci of points where the amplification of the density contrast between $z_{dec} = 10^3$ and the present has a well defined value, which is the label on each curve. Thus we see that the interesting region on the plane satisfies both $k = +1$ and $\Lambda > \Lambda_{crit}$: indeed these growing modes amplify better than a factor $10^3$ (which is the curve going through the Einstein-de Sitter model $\sigma_0 = q_0 = 0.5$).

In the framework of a low-density Universe, these consideration apply only if $H_0 = 100$ km/s/Mpc, $H_0^{-1} = 10 \times 10^9$ y and we must resort to the cosmological term for the age problem (see also figure 3). If so, the order of magnitude improvement of the growing mode amplification results partly from curvature and partly from the increased time span available for growth; indeed in the limiting case $\Lambda = \Lambda_{crit}$, the cosmic expansion is suppressed and power laws are again replaced by exponentials.

Let us now return to the era immediately before decoupling when baryons and photons are coupled by Thomson scattering and energetically equivalent: incidentally we may define the equivalence redshift between matter and radiation at

$$1 + z_{eq} = 4 \times 10^4 \Omega_0 h^2 , \tag{27}$$

and say that for $z > z_{eq}$ radiation dominates while matter ($p = 0$) dominates afterwards, $z < z_{eq}$. The perturbations of the cosmic medium we have discussed above are adiabatic in the sense that radiation and matter are perturbed together and the ratio of photon to baryon number is kept constant. We have another fundamental mode of perturbation, however; the isothermal one, where only baryons are perturbed, but the background radiation is left unperturbed. In this case since the number of baryons per photon is changed we have an entropy perturbation. Mészàros (1974) addresses the question of whether given a completely uniform distribution of particles and radiation, can a perturbation of the particle distribution only grow. Under the assumption of flatness of space-time ($k = 0$) and, more importantly, of no interaction between the particles and the relativistic substratum, beside of course gravitation, the answer is that no growth is possible until the Universe is radiation dominated, but growth becomes possible thereafter.

## 3. High Density Universes

### a) A Neutrino Dominated Universe

The experimental measure of the electron neutrino rest mass by Lubimov et al. (1980),

$$14\ \mathrm{eV} < m_{\nu e} c^2 < 46\ \mathrm{eV}, \qquad (1)$$

makes it quite plausible that the sought hidden mass is in fact in the form of massive neutrinos, as it was suggested with considerable foresight by many authors. The conventional view holds instead that the unseen matter is ordinary baryonic matter of low luminosity such as dust, subluminous stars, black holes, rocks, etc.

Gershtein and Zel'dovich (1966) and Cowsik and McClelland (1972) compared the present known cosmic density in baryons with the theoretical cosmic density in neutrinos (see later) and derived an upper limit for the mass of the latter. Later on, Cowsik and McClelland (1973) assumed that massive neutrinos might dominate the gravitational dynamics of large clusters of galaxies and did build on this basis a simple model for the Coma cluster. Szalay and Marx (1976) called attention to the fact that density fluctuations in a primordial neutrino gas may initiate the formation of clusters of galaxies. An early review of neutrino cosmology is given by Bludman (1976) while Markov (1964) calls attention to degenerate massive neutrino superstars.

We will now examine the cosmological impact of the neutrino rest-mass as it has been studied by many authors (Zel'dovich et al. 1980, Bisnovatyi-Kogan et al. 1980, Schramm and Steigman 1980 and 1981, Bond et al. 1980, Klinkhamer and Norman 1981, Sato and Takahara 1981) who have come essentially to similar conclusions; it seems possible that we may solve at the same time the hidden mass problem of section 1 and the gravitational instability problem of section 2. In particular we will hold the view that the condensations of galactic or larger scale started out as massive neutrino condensations at $z \cong 10^4$; only after photon-baryon decoupling, $z \cong 10^3$, were the baryons capable of falling into the neutrino gravitational wells. The possibility that massive neutrinos are distributed like the galaxies is made plausible by the remark that the Universe does not possess a significant smooth component (Yahil et al. 1978).

The neutrinos we have around today in our Universe originated in the Big-Bang (we assume left-handed neutrinos of the Majorana type); at temperature in excess of 1 Mev ($\gg m_\nu$) all neutrinos of the three types ($e, \mu, \tau$) were in thermal equilibrium and an extremely relativistic (ER) Fermi distribution was established for each flavor (i)

$$\frac{dn_{\nu i}}{d^3 q} = (g_i/h^3)\,[\exp(q\,c/k\,T) + 1]^{-1}. \qquad (2)$$

We assume a vanishing chemical potential (Weinberg 1972) and $g_i = 2$ (as it is the case for Majorana neutrinos, while it would be $g_i = 4$ for Dirac neutrinos). As the temperature drops below 1 Mev the weak interaction rate falls below the expansion rate and thermal equilibrium is lost; however since both T and momentum fall like $1/R$, the distribution function (2) remains formally unchanged down to the non-relativistic (NR) region and the present. Clearly T has not the physical meaning of a temperature. By integration of (2) over momentum space we can relate the number density of neutrinos to the number density of photons:

$$n_{\nu i} = \frac{3}{4}\,\frac{1}{2}\,g_i\,n_\gamma = \frac{3}{4}\,n_\gamma, \qquad (3)$$

where

$$n_{\gamma 0} \cong 400\,(T_{\gamma 0}/2.7\ \mathrm{K})^3. \qquad (4)$$

As the temperature drops below the electron mass, electron-positron pairs annihilate and generate photons: it is known (Weinberg 1972) that

$$\frac{n_\gamma\,(T < 0.5\ \mathrm{MeV})}{n_\gamma\,(T > 0.5\ \mathrm{MeV})} = \frac{11}{4}; \qquad (5)$$

hence the photon temperature jumps by a factor $(11/4)^{1/3}$ due to the electron-positron annihilation and remains higher by the same factor during the whole history of the Universe.

Thus the total number density of neutrinos now is given by

$$n_{\nu 0} = \sum_i n_{\nu i 0} = 3 \times \frac{3}{4} \times \frac{4}{11}\,n_{\gamma 0} \cong 300\,(T_{\gamma 0}/2.7\ \mathrm{K})^3. \qquad (6)$$

Nowadays the neutrinos are non-relativistic; the associated mass density is the sum of

their proper masses. Assigning each neutrino flavor the same average mass

$$m_\nu = m_{30}\,(m_\nu c^2/30\ \mathrm{eV}),$$
$$30\ \mathrm{eV} \cong 5\times 10^{-32}\ \mathrm{g}, \qquad (7)$$

we end up with a present density in neutrinos which is very large

$$\rho_{\nu_0} = 2\times 10^{-29}\, m_{30}\, \mathrm{g\,cm^{-3}}. \qquad (8)$$

When we compare this with the critical density (1a.1) we have:

$$\Omega_{\nu_0} = m_{30}\, h^{-2} \cong 1. \qquad (9)$$

Thus there is a valid candidate for hidden matter of the density required in large systems; on the other hand, we must also hold that baryonic matter is scarce: we define here a new parameter

$$\varepsilon = \rho_{B_0}/\rho_{0\,\mathrm{tot}}, \qquad (10)$$

which is small.
The cosmological model we want to explore now is one of high density

$$\Omega_{0\,\mathrm{tot}} = \Omega_{\nu_0} + \Omega_{B_0} \cong \Omega_{\nu_0} \cong 1,$$

where ordinary baryonic matter represents only a minor contamination (clearly very important for us!).
One question we must face immediately is whether neutrino rest-masses would affect the standard Big-Bang predictions: the answer given by Shapiro et al. (1980) and Dolgov and Zel'dovich (1981) is negative. Indeed although both left-handed and right-handed neutrinos could be present, right handed neutrinos would not be in equilibrium at 1 Mev, but would have decoupled much earlier ($kT \gg 100$ Mev) according to the Weinberg-Salam-Glashow theory of weak interactions.
Another issue that must be taken up if (9) is valid is whether massive neutrinos affect the theoretical estimates of the age of the Universe in a way that is still consistent with the ages derived from nucleocosmochronology and stellar evolution (see, e.g., Symbalisty et al. 1980). Nucleocosmochronology gives a lower limit of the order $10\times 10^9$ y, which does not pose us any particular problem. However, the ages inferred from globular cluster stars (Iben 1974) are very large (more than 12 billion years) for standard helium abundances and may even exceed 20 billion years as the helium abundance is decreased: we must recall that the Hubble time

$$H_0^{-1} = 10\times 10^9\, h^{-1}\, y,$$

is the upper limit to the age of a $\Lambda = 0$ FRW model valid when $\Omega_0 \to 0$; a high density model has a short age, $< 2/3 \times H_0^{-1}$.
The suggestion by Zel'dovich and Sunyaev (1980; see also Luminet and Schneider 1981) is to revitalize the cosmological term because a suitably chosen positive $\Lambda$-term, can make the age of the Universe arbitrarily long.
In figure 3 we give some numerical results: from the equation of motion (2b.8) we first evaluate numerically the age

$$t_0 = \int_0^{t_0} dt = \int_0^{R_0} dR/\dot{R},$$

as a function of the pair $(\sigma_0, q_0)$ and then we plot on the $(\sigma_0, q_0)$-plane isochrones, loci of the points of the same age. We are interested in exemplificative ages of 12, 14, 16 and 18 billion years; the corresponding curves are conveniently parametrized by the dimensionless number $H_0 t_0$ which assumes the two sets of values a) 0.6, 0.7, 0.8 and 0.9 for $H_0 = 50$ km/s/Mpc and b) 1.2, 1.4, 1.6 and 1.8 for $H_0 = 100$ km/s/Mpc.
According to (9) $\Omega_0$ ranges between 4 and 1 and $\sigma_0$ ranges between 2 and 0.5: in order to find a value for the $\Lambda$-term all we have to do is to find a value for $q_0$, (2b.26), by the intersections of the vertical lines $\sigma_0 = 2$ (labelled 50 to remind us the Hubble constant) and $\sigma_0 = 0.5$ (labelled 100) with the ages curves. These intersections occur at negative values for $q_0$, $\cong -2$. For a not unlikely intermediate value of $H_0$ (see, e.g., Van der Bergh 1981) and from (9), $\Omega_0 \cong 1$ for $m_{30} \cong 0.5$: to get an age $= 13$ billion years ($\cong H_0^{-1}$) we should look in figure 3 at the intersections between the vertical line $\sigma_0 = 0.5$ and the curve $H_0 t_0 = 1$ which has not been drawn to avoid further crowding of figure 3. Again the intersection yields a negative $q_0$, $\cong -1.5$. As before this implies that the Universe expansion is accelerating which formally calls for a positive $\Lambda$, (2b.7).

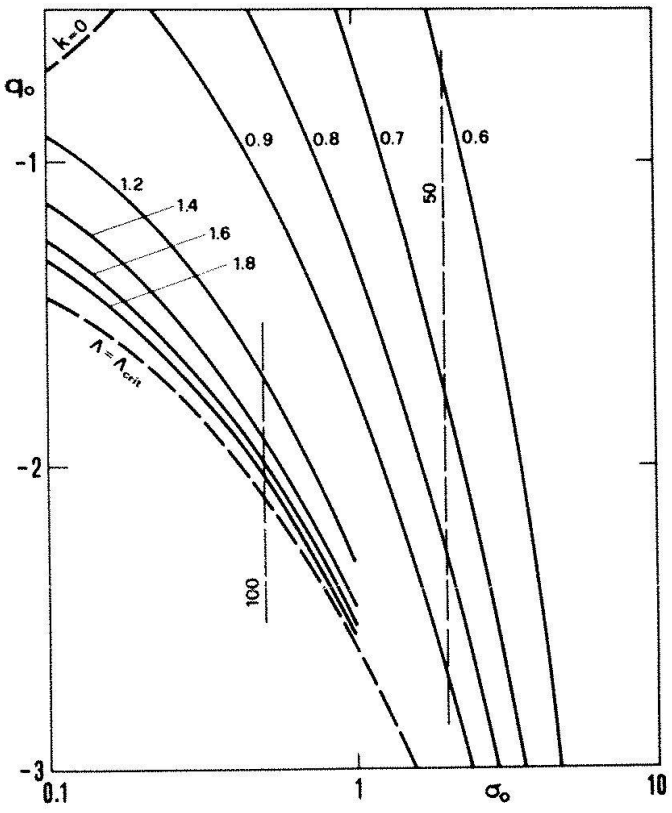



*Fig. 3.* Plot on the $(\sigma_0, q_0)$-plane of isochrones, loci of points where the cosmological models have the same age. We consider ages of 12, 14, 16 and 18 billion years. We label the curves on the graph by the values of $H_0 t_0$: for $H_0 = 50$ km/s/Mpc we have a first set of values $H_0 t_0 = 0.6$, 0.7, 0.8 and 0.9 and of corresponding curves; for $H_0 = 100$ km/s/Mpc we have a second set of values $H_0 t_0 = 1.2$, 1.4, 1.6 and 1.8 and of corresponding curves. Assuming $m_{30} = 1$, according to (3a.9) we have $\sigma_0 = 0.5$ for $h = 1$ or $\sigma_0 = 2$ for $h = 0.5$. The intersections of the vertical lines $\sigma_0 = 0.5$ (labelled 100) and $\sigma_0 = 2$ (labelled 50) with the corresponding age curves yield the values for $q_0$: in all cases negative value of $q_0$ are found (of the order $-1$ or $-2$). An acceleration of the cosmic expansion is implied, which means formally $\Lambda > 0$.

In the history of cosmology the $\Lambda$-term has seen many ups and downs (for a review see Petrosian 1974, and, more recently, Gunn and Tinsley 1975 and Tinsley 1977); from figure 3 we see that we are now far from the $\Lambda = \Lambda_{crit}$ curve which seemed interesting some years ago due to an apparent accumulation of quasar redshifts around 2 (the issue has now disappeared; Tytler 1981). Our difficulties with $\Lambda$ stem from our failure to understand its physics, aside from an attempt by Zel'dovich (1968) to relate $\Lambda$ to the quantum fluctuations of vacuum.

b) The Infall of Baryons onto Massive Neutrino Condensations

The growth of neutrino density fluctuations is of fundamental importance for the formation of the structure we observe in the Universe. To understand this we must remark that the growth of baryon fluctuations is inhibited by photon viscosity until $z_{dec} = 10^3$, while neutrinos decouple from equilibrium at $T = 1$ Mev and are collisionless ever since. This collisionless feature deserves some attention: the approaches to gravitational instability by Jeans, Lifshitz and Bonnor were all based on the hydrodynamic description of matter, that is on the assumption that the mean free path between particle collisions is small in comparison with any characteristic length of the problem. When we deal with neutrinos the opposite is true and we must resort instead to the distribution function and the evolution of its perturbations: we will quote here the results of the pioneering work of Gilbert (1966).

Under the Newtonian approximation, one describes the uniform cosmic dust by a collisionless Boltzmann equation, superposes the cosmic expansion and introduces a small perturbation in the distribution function.

The Fourier transform of the density contrast obeys a Volterra integral equation, which Gilbert (1966) studies numerically. He finds that the large wavelength density contrasts grow monotonically under self-gravitation and that the small wavelength density contrasts do not oscillate like sound waves, but first decrease due to Landau damping and eventually grow too. The separation between large and small wavelengths is given by a Jeans wavelength

$$\lambda_J = [\pi < v^2 > / 3 G \rho]^{1/2} \qquad (1)$$

which is built with a characteristic mean square particle velocity rather than with the speed of sound as (2a.5). The reason why small wavelength modes increase again after an initial Landau damping is that in (1) $v^2$ behaves as $R^{-2}$ while $\rho$ behaves as $R^{-3}$; the Jeans length thus increases as $R^{1/2}$ while the wavelength of any perturbation increases as R and eventually overtakes $\lambda_J$.

Stewart (1972) also adopts a kinetic theory rather than a hydrodynamical approach with the aim of studying the evolution of condensations of collisionless and massless neu-

trinos in a homogeneous isotropic FRW Universe. After generalizing the formalism to deal correctly with strong gravitational fields, he confirms the qualitative picture that emerges from the Newtonian analysis and in particular the Landau damping (see also Lynden-Bell 1967) of the short wavelength modes.

We must therefore compute the Jeans mass of the massive neutrinos: let us first set (somewhat conventionally) at

$$1+z_{NR} \cong \frac{m_\nu c^2}{3\,(4/11)^{1/3}\,k\,T_{\gamma 0}} \cong 6\times 10^4\, m_{30}\,, \qquad (2)$$

the redshift at which the neutrinos become non-relativistic in their adiabatic cooling. For $z \gg z_{NR}$ the neutrinos are considered to be extremely relativistic (ER), while they are considered to be fully non-relativistic (NR) for $z \ll z_{NR}$; strictly speaking in the middle neither approximation is true, but either is satisfactory for the purpose of obtaining order of magnitude estimates. The Jeans mass is a qualitative concept any way.

From (2.7) we have

$$M_J = \frac{\pi}{6}\, n\, m_\nu\, \lambda_J^3\,, \qquad (3)$$

where $\lambda_J$ is given in (1) and n is related to the distribution function by the generalization of (a.6) above:

$$n(z) \cong 300 \left(\frac{T_{\gamma 0}}{2.7\ K}\right)^3 (1+z)^3\,. \qquad (4)$$

For the ER regime, $v=c$, the Jeans mass coincides with the horizon mass and we find

$$M_{JER} \cong \frac{1}{16}\, m_p \left(\frac{m_p}{m_\nu}\right)^2 \left(\frac{m_\nu c^2}{k\,T_\gamma}\right)^3, \qquad (5)$$

where

$$m_p = \left(\frac{\hbar c}{G}\right)^{1/2} \cong 2\times 10^{-5}\,g\,,$$

is the Planck mass; this expression appears as a generalization of that found by Bisnovatyi-Kogan et al. (1980), which is interesting because it is constructed in terms of fundamental constants. Alternatively we may write for present purposes

$$M_{JER} \cong 2\times 10^{29} \frac{m_{30}}{(1+z)^3}\, M_\odot\,; \qquad (6)$$

this is the growing straight solid line in figure 4 which should extend only to $z = 6\times 10^4\, m_{30}$ but is used also to slightly lower redshifts.

For the NR regime, we find

$$\langle v^2 \rangle^{1/2} \cong 6\, \frac{1+z}{m_{30}}\ \text{km/s}\,, \qquad (7)$$

and in place of (5)

$$M_{JNR} \cong 20\, m_p \left(\frac{m_p}{m_\nu}\right)^2 \left(\frac{k\,T_\gamma}{m_\nu c^2}\right)^{3/2}; \qquad (8)$$

alternatively we write this as

$$M_{JNR} \cong 8\times 10^8\, \frac{(1+z)^{3/2}}{m_{30}^{7/2}}\, M_\odot\,, \qquad (9)$$

which we plot as the decreasing straight solid line in figure 4. The two lines meet at a redshift

$$z_{max} \cong 3.5\times 10^4\, m_{30}\,, \qquad (10)$$

where the Jeans mass attains its maximum

$$M_{J\,max} \cong 5.5\times 10^{15}\, m_{30}^{-2}\,. \qquad (11)$$

Strictly speaking this is a slight overestimate of the true value, which would result from a better approximation for the intermediate zone. As many authors have remarked it is especially gratifying to cosmology that (11) gives the order of magnitude of the mass of a cluster of a galaxies; in the scenario we are presenting here, however, the mass in baryons must be smaller than (11) by a factor $\varepsilon$, which may mean even two orders of magnitude.

Condensations on mass scales larger than $M_{J\nu}$ can grow for $z > z_{NR}$, but not if they involve only neutrinos for an argument used in section 2a. Indeed the Universe is radiation dominated from the epoch when the neutrinos become non-relativistic, $z = z_{NR}$, until

$$1+z_{EQ} \cong 4\times 10^4\, m_{30}\,, \qquad (12)$$

(compare with (2b.27) which applies to baryonic matter-radiation equivalence), whereafter neutrinos take over and the Universe becomes dust dominated. During this phase, $z_{NR} > z > z_{EQ}$, a density perturbation involv-

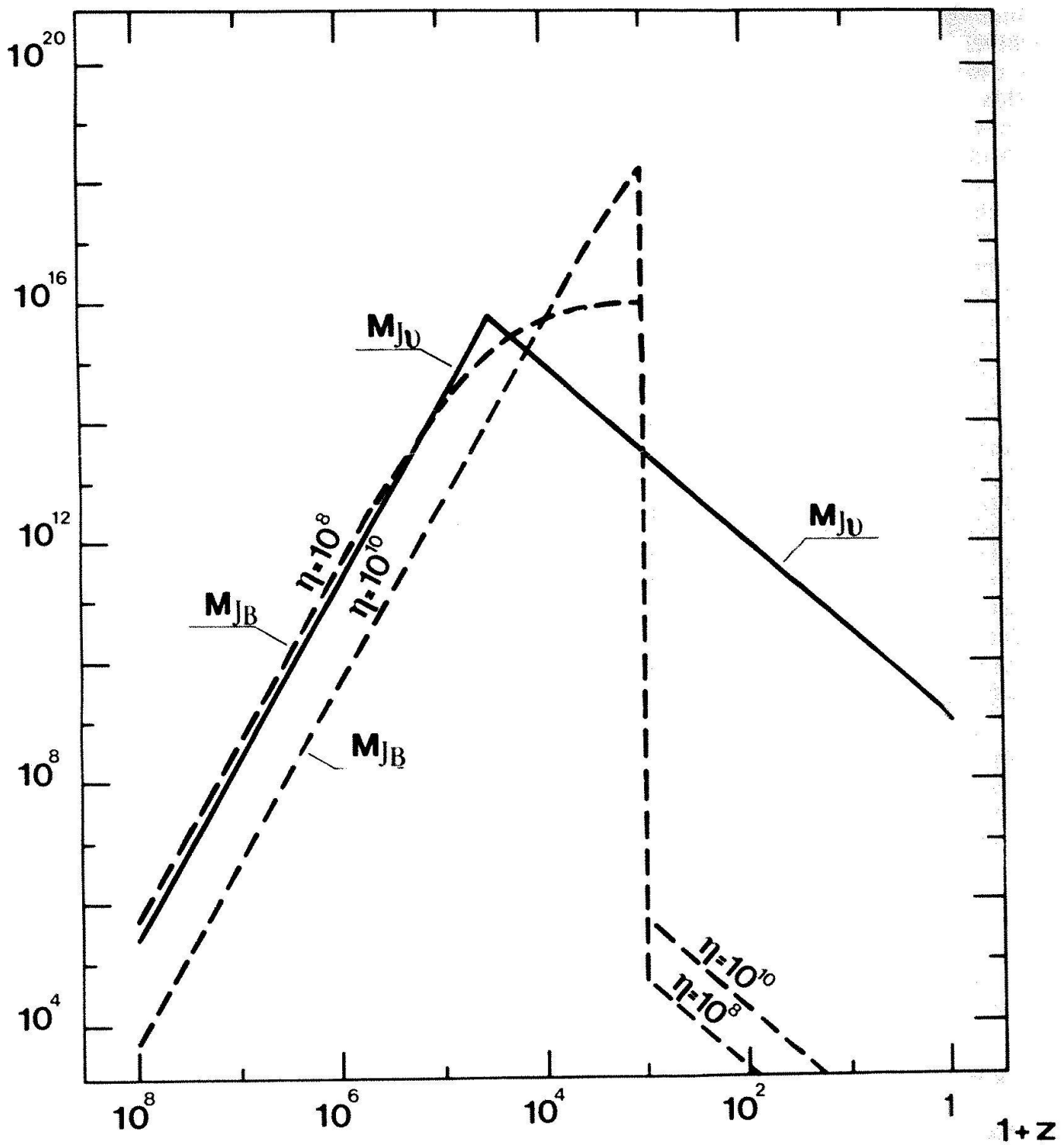



*Fig. 4.* Plot of the Jeans mass in solar masses in a massive neutrino Universe vs. redshift for $m_{30} = 1$. In the extreme relativistic regime (ER) the Jeans mass grows with time as shown by the straight solid line to the left, i.e. according to (3b.6), this expression being valid only for $z > z_{NR} = 6 \times 10^4$; in the non-relativistic (NR) region the Jeans mass decreases as shown by the straight solid line to the right, i.e. according to (3b.9). The intersection of the two straight lines gives an indication - perhaps slightly in excess - of the maximum of the Jeans mass: this occurs at

$$z_{max} \cong 3.5 \times 10^4,$$

which is $\gtrsim z_{NR}$ and close to the value $z_{EQ}$ when the Universe starts being dominated by neutrinos rather than by radiation and the neutrino perturbations can start to grow. The maximum Jeans mass is of the order

$$M_{J\,max} \cong 5 \times 10^{15} \frac{1}{m_{30}^2} M_\odot .$$

Also shown by broken lines are the Jeans masses for a canonical baryon Universe evaluated with the formulae given in the section 2 for two values of the photon-to-baryon ratio, $\eta = 10^{10}$ and $10^8$. The vertical drop at $z = 1000$ is a consequence of the fall of the sound speed due to hydrogen formation, assumed to occur instantaneously.

ing only the dust component cannot grow, but continues to expand with the substratum (Mészàros 1974).
For the Silk mass in baryons in a baryon-neutrino Universe, generalizing (2a.17), we have (Bond et al. 1980)

$$M_D \cong 10^{13}\,\Omega_{B_0}^{-1/2}\,\Omega_{\nu_0}^{-3/4}\,h^{-5/2}\,M_\odot\,; \qquad (13)$$

again in a qualitative sense, the latter reminds the correct range of masses.
In order to study the gravitational coupling between the baryon and the neutrino components we will limit ourselves here for semplicity to mass scales $\gtrsim M_{J\nu max}$ (which is the most interesting range due to the likely decrease of the fluctuation spectrum with mass) and to epoch $z < z_{max} \cong z_{EQ}$; by itself the baryonic content of such a perturbation ($\sim \varepsilon/(1-\varepsilon)\,M_{J\nu max}$) would undergo damped oscillations. The formalism of section 2b yields directly the differential equations for

$$\delta_\nu = \left(\frac{\delta\rho}{\rho}\right)_\nu, \qquad \delta_B = \left(\frac{\delta\rho}{\rho}\right)_B. \qquad (13)$$

The physics we want to impose to our model is that $\delta_\nu$ can grow in the interval $z_{EQ} > z > z_{dec} = 10^3$, where instead $\delta_B = 0$ due to photon viscosity:

$$\delta_\nu \propto t^{\alpha_\pm},$$
$$\alpha_\pm = \frac{1}{6}\left[-1 \pm \sqrt{25-24\varepsilon}\right]. \qquad (14)$$

For $z < z_{dec}$, we have (Doroshkevich et al. 1980, Bond et al. 1980, Wasserman 1981)

$$\frac{1}{R^2}\frac{d}{dt}R^2\frac{d}{dt}\delta_B - 4\pi G\rho\varepsilon\delta_B = 4\pi G\rho(1-\varepsilon)\delta_\nu, \qquad (15)$$

$$\frac{1}{R^2}\frac{d}{dt}R^2\frac{d}{dt}\delta_\nu - 4\pi G\rho(1-\varepsilon)\delta_\nu = 4\pi G\rho\varepsilon\delta_B. \qquad (16)$$

These are valid in all generality for any pair of $\sigma_0$ and $q_0$. The power solution (14) is obtained from (16) setting $\delta_B = 0$ and using the Einstein-de Sitter approximation valid for high redshift.

The interesting remark that has been made is that, due to the inhomogeneous nature of (15), baryon perturbations can grow pulled into the neutrino gravitational wells even if $\delta_B = \dot{\delta}_B = 0$ at $z_{dec}$ (Doroshkevich et al. 1980, Bond et al. 1980, Wasserman 1981). Therefore even if all the baryon perturbations have been damped by photon viscosity, we only need postulate an initial spectrum of neutrino perturbations earlier on at $z_{NR}$ or $z_{max} \cong z_{EQ}$. An idea of the solution of (15) and (16) can be obtained in the limit $\varepsilon \to 0$ and for large redshifts (Einstein-de Sitter approximation): indicating with a subscript "1" a reference epoch when $\delta_B = \dot{\delta}_B = 0$ around $z_{dec}$, one has

$$\delta_\nu = \Delta_\nu\,(t/t_1)^{2/3},$$
$$\delta_B = \Delta_\nu\,[(t/t_1)^{2/3} + 2\,(t_1/t)^{1/3} - 3], \qquad (17)$$

where for simplicity only the growing mode of $\delta_\nu$ has been considered.
From (17) we draw from the conclusion that $\delta_B$ and $\delta_\nu$ lock together and attain the same values at the present time; it also appears that the baryon density amplification defined as

$$A = \delta_B\,(1+z=1)/\delta_B\,(1+z_1), \qquad (18)$$

is formally infinite. We are rather interested in using in the denominator of (18) the redshift of the later epoch when baryons last interact with the radiation background: let this be

$$1+z = (1+z_1) - \Delta z, \qquad \Delta z \ll (1+z_1).$$

Then from (17) we find

$$\delta_B\,(1+z) = \frac{3}{4}\,\Delta_\nu\left(\frac{\Delta z}{1+z_1}\right)^2, \qquad (19)$$

which tells us that for $1+z_1 = 1000$, $\Delta z = 100$, $\delta_B\,(900) \cong 10^{-2}\,\Delta_\nu$. Thus an amplification of physical interest

$$A = \delta_B\,(1+z=1)/\delta_B\,(1+z) \cong \delta_\nu\,(1)/(10^{-2}\,\Delta_\nu), \qquad (20)$$

would result larger than the amplification of the neutrino modes by a couple of orders of magnitudes.

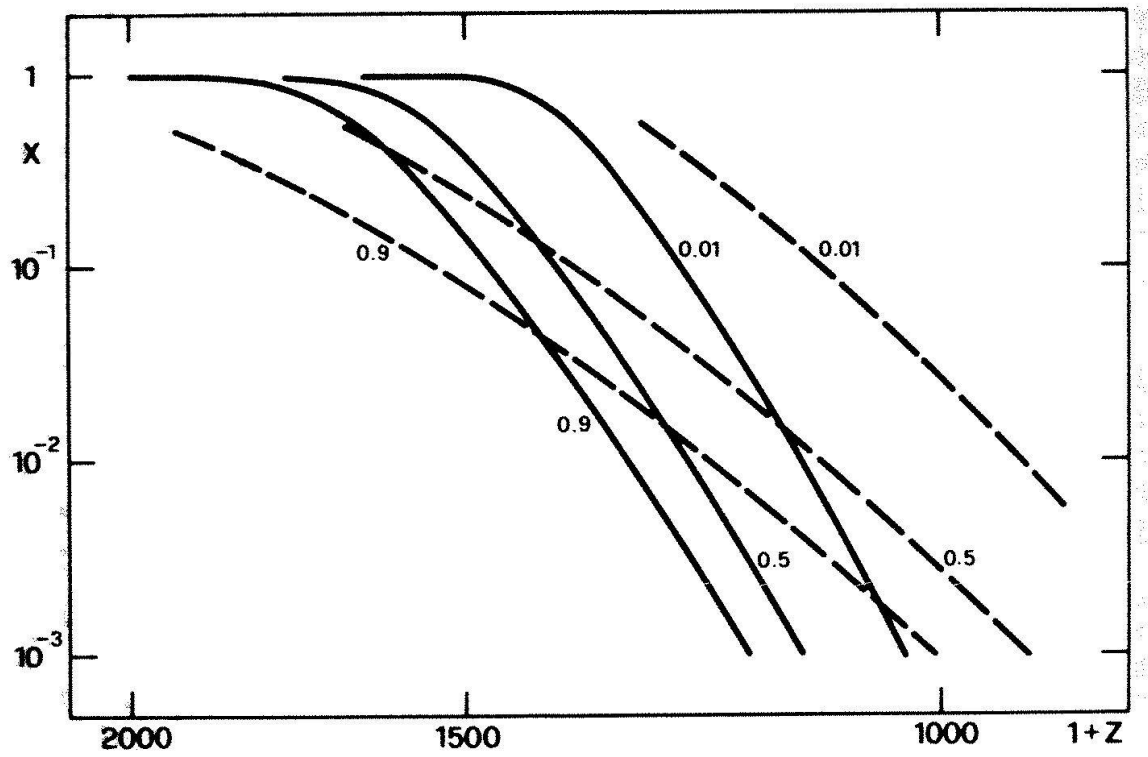



*Fig. 5.* Plot of the ionization fraction

$$x = \frac{n_p}{n_p + n_H},$$

vs. redshift for various values of $\varepsilon$, defined in (3a.10). Full lines are evaluated via the Saha equation (e.g. Peebles 1971). As expected, the smaller $\varepsilon$, the later the ionization drop occurs: for $\varepsilon = 0.01$, x drops very rapidly only for $z \lesssim z_{dec} = 1000$. This is the basis for the approximation used in the text to let the baryons fall freely in the interval $z_{dec} + \Delta z > z > z_{dec}$, $\Delta z = 100$. For comparison, for the same values of $\varepsilon$ also the approximate formula by Sunyaev and Zel'dovich (1970 and 1980) is shown by broken lines: in this case ionization lasts much longer than indicated by the Saha equation.

In fact, decoupling between baryons and photons is not an instantaneous process but lasts a certain time. The corresponding thickness in redshift is of the order of several hundreds as we see in figure 5 where we plot the degree of ionization vs. redshift as a function of $\varepsilon$ evaluated either via the Saha equation (see, e.g., Peebles 1971) or via a more elaborate treatment due to Sunyaev and Zel'dovich (1970 and 1980).

Although during this phase the coupling between baryons and photons is described correctly only by the method of Peebles and Yu (1970; for more recent work see Silk and Wilson 1980), an order of magnitude information can be extracted from (15) and (16) as well. For this purpose we assume that baryons start falling freely, i.e. obeying (15) and (16), at $z_1 = z_{dec} + \Delta z$, $\Delta z = 100$ with $\delta_B(z_1) = \dot{\delta}_B(z_1) = 0$. We then place conventionally at $z_{dec} = 10^3$ the end of any interaction between baryons and photons having in mind to evaluate an upper limit to the perturbations on the microwave background via

$$\left(\frac{\delta T}{T}\right)_{dec,max} = \frac{1}{3}\,\delta_B(z_{dec}). \tag{21}$$

We have integrated numerically (15) and (16) on the $(\sigma_0 - q_0)$ plane and for various values of $\varepsilon$ with the purpose of understanding what sort of modifications have been generated with respect to the results of figure 2 by the two fluid nature of the background model.

In figure 6 we study high density models: full lines give the growth of $\delta_B$ (normalized to unity at the present) vs. redshift, broken lines give the growth of $\delta_\nu$. Broken and full lines merge together. For $\varepsilon$ we consider a very low value $\varepsilon = 0.01$, and an intermediate one, $\varepsilon = 0.5$; it turns out that the dependence on $\varepsilon$ is minor in the range of physical interest (which excludes $\varepsilon \to 1$). For $\varepsilon = 0.01$ the first

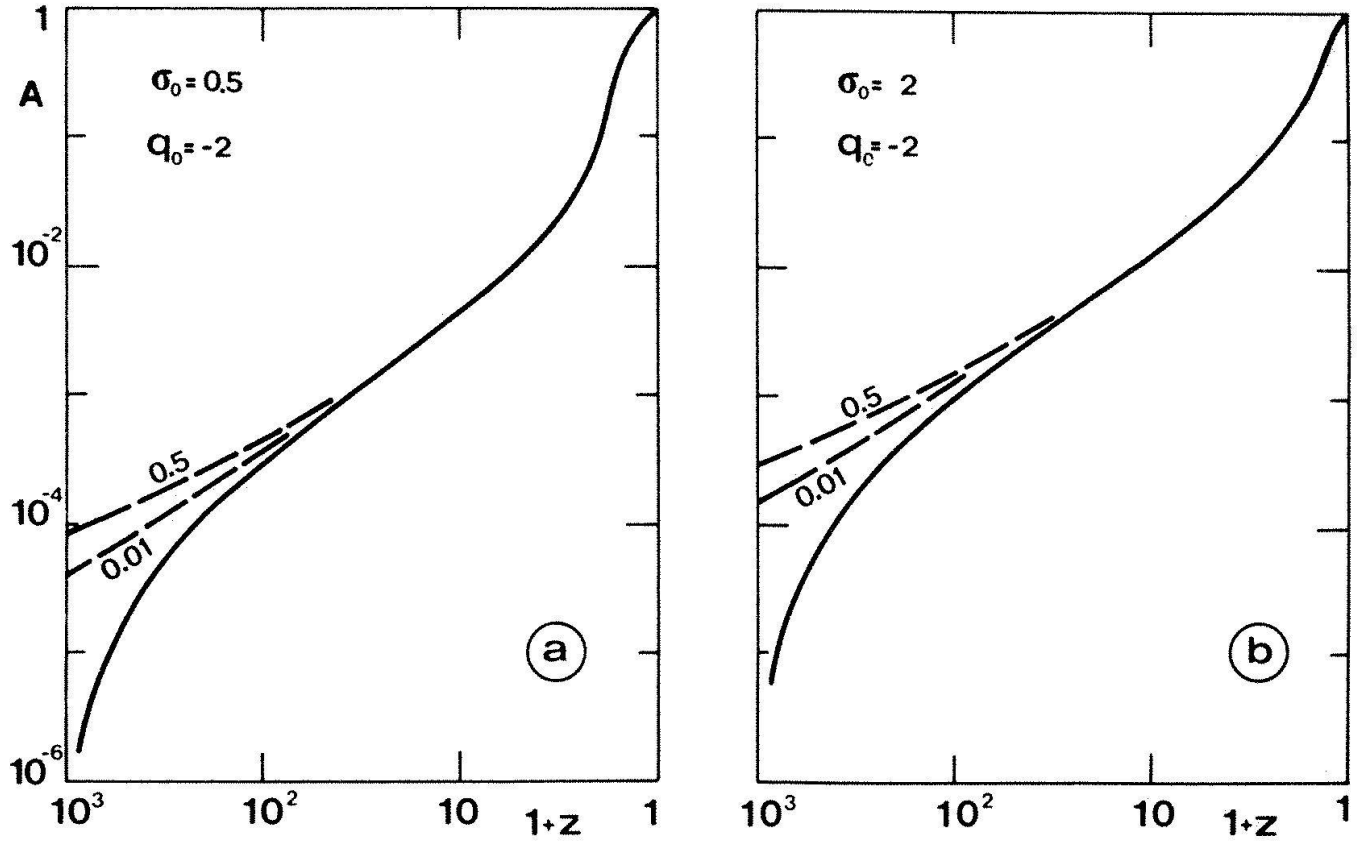



*Fig. 6.* Amplification of the growing modes of baryons and neutrinos in high-density Universe models: $\sigma_0 = 0.5$, $q_0 = -2$ in a); $\sigma_0 = 2$, $q_0 = -2$ in b). Full lines refer to baryons, dashed lines to neutrinos; we show the cases $\varepsilon = 0.01$ and $\varepsilon = 0.5$. When baryon density contrasts are normalized to unity at the present, the baryon curves are undistinguishable from each other, so that we plot only one of them. This means that the total amplification available to baryons is not very sensitive to the actual value of $\varepsilon$, as long as the latter is $\varepsilon \ll 1$, but depends mainly on the nature of model, i.e. the coexistence of two self-gravitating fluids, one of which is able to begin its gravitational condensation very early, at $z \cong z_{EQ} \cong 10^4 > z_{dec} = 1000$. Baryon fluctuations in high density models are shown to amplify by six orders of magnitude. As far as the neutrinos are concerned instead, their amplification depends strongly on $\varepsilon$; for small $\varepsilon$ the baryons do not matter and neutrino fluctuations amplify as in (2b.20). On the contrary for larger $\varepsilon$ neutrino self-gravity is weak and neutrino fluctuations grow little for $z_{EQ} > z > z_{dec}$.

part of the neutrino growth is virtually that of an Einstein-de Sitter model (2b.20); for lower z the neutrino growth rate goes to what is expected from figure 2. For $\varepsilon = 0.5$ the neutrino condensations build up very slowly in the beginning because 50% of the total matter content (all the baryons) is unable to condense before $z_{dec}$; later on, the neutrino and the baryon growth curve are undistinguishable from each other and from the curves valid for $\varepsilon = 0.01$. We have in fact drawn only one curve for the baryons.

The exit from linearity for the baryons and the neutrinos occurs simultaneously; the amplification available to $\delta_B$ is of six orders of magnitude.

In Figure 7a we examine the paradigmatic Einstein-de Sitter case, where $\delta_B$ amplifies by five orders of magnitude between $z_{dec}$ and the present - rather than three as in the pure baryon model - due to the presence of the neutrinos, provided they are a major component of the total density ($\varepsilon < 0.5$). In figure 7b we show that this is basically true even in a low density Universe, though to a lesser extent than in high density models of figure 6.

The dependence of our results from $\varepsilon$ is shown in figure 8 where the exemplificative models of figures 6 and 7 are studied in the range $0 < \varepsilon < 1$: we plot the baryon and the neutrino amplification vs. $\varepsilon$ and we consider the two cases, that the exit from linearity occurs a) at the present, $1+z=1$, or b) at $1+z=5$. For baryons the dependence on $\varepsilon$ is insignificant, for neutrinos it becomes relevant only for $\varepsilon \to 1$, which is not interesting. When we assume that linear growth applies all the way to the present, then the total amplification depends to a certain

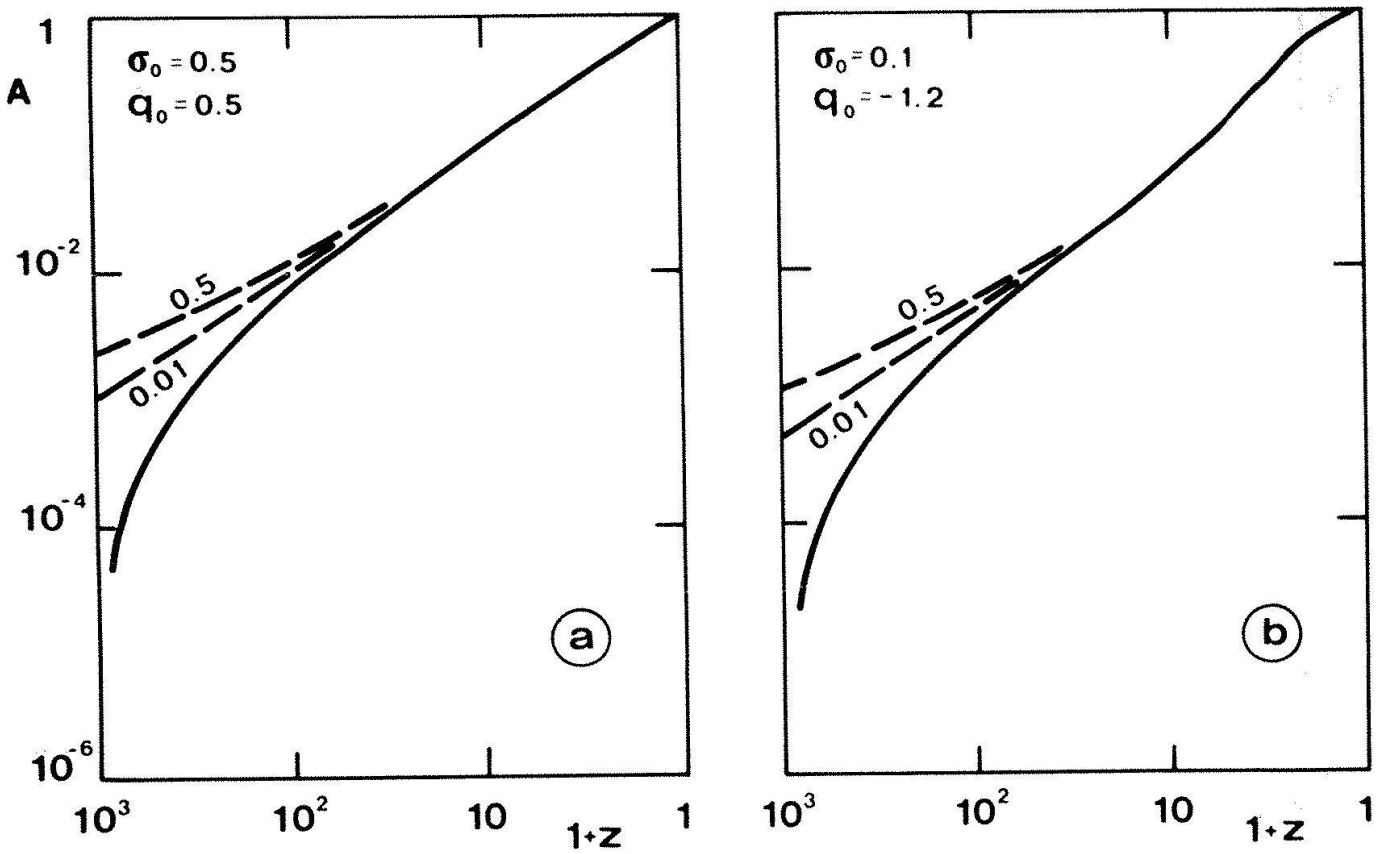



*Fig. 7.* The same as in figure 6 for the Einstein-de Sitter case in a) and for a low density model in b); in the latter case $q_0$, taken from the age curves of figure 3, yields an age of the order of 15 billion years. Basically the same comments of figure 6 apply: due to the early locking of the baryon to the neutrino fluctuation, even when the total density is low the amplification of the baryon fluctuations exceeds four orders of magnitude, thus justifying the lack of detection in the microwave background of the condensation's footprints.

extent on the actual values of $\sigma_0$ and $q_0$; but when non-linearity is attained earlier at $1+z=5$, which is physically more interesting, the amplifications available are largely independent of $\sigma_0$ and $q_0$ (since the linear growth is interrupted during the Einstein-de Sitter phase).

Depending on whether the condition $\delta_\nu = 1$ is reached at $1+z=1$ or at $1+z=5$, we may evaluate $\delta_\nu(z_{EQ})$ by combining the growth resulting from the numerical integration of (15) and (16) with the initial growth given by (14) between $z_{EQ}$ and $z_{dec}$. In figure 9 we plot $\delta_\nu(z_{EQ})$ vs. $\varepsilon$: in the interesting region $\varepsilon \ll 1$, the initial neutrino amplitudes are below $10^{-4}$.

In figure 10 we present our results for the amplifications at the present on the whole $(\sigma_0 - q_0)$ plane for $\varepsilon = 0.01$; the level curves we obtain there should be compared with those of figure 2, where the case of a one-fluid Universe is examined. A couple of orders of magnitude are gained everywhere; when we translate this in upper limits for the microwave background temperature fluctuations at $z_{dec}$ according to (17), which we plot in figure 11, we see that we are many orders of magnitude below the present detectability. Even for $\delta_B(1+z=5)=1$, we read from figure 8 that $(\delta T/T)_{dec} < \delta_B(z_{dec})/3 \cong 10^{-5}$.

As we have seen baryon-neutrino, high-density cosmological models hold great potentialities for the linear growth of baryon density enhancements and make less remote the understanding of the condensed structures we see. Progress has not been equally fast in the realm of non-linear condensations, but some results in agreement with the above considerations are already available.

The study of the (non-linear) evolution of spherically symmetric condensations may be of interest for modelling clusters of galaxies (Occhionero et al. 1981a and b, and references therein). During the formation of such a condensation the matter which accumulates at the center is swept away from the space around the condensation itself, so

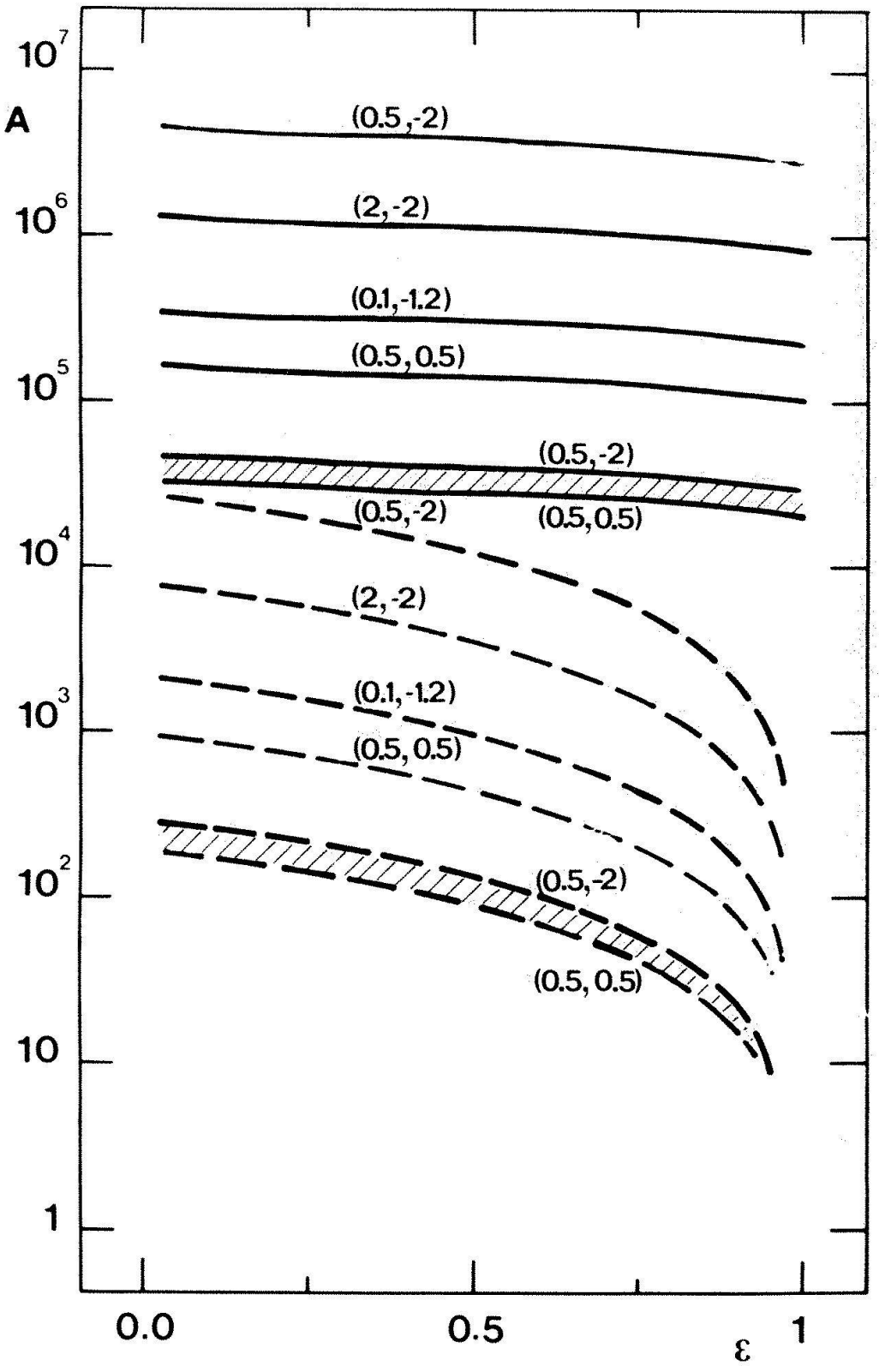



*Fig. 8.* Plot of the baryon and of the neutrino mode amplification vs. $\varepsilon$; full lines refer to baryons, broken lines to neutrinos. In one set of curves, exit from linearity is assumed to occur at the present and the amplifications are defined as

$\delta_B(1)/\delta_B(1000)$,
$\delta_\nu(1)/\delta_\nu(1000)$;

in this case each curve is labelled by the $\sigma_0$ and $q_0$ to which it refers. In the second set of curves, exit from linearity is assumed to occur at $1+z=5$; in this case the amplifications are defined instead as

$\delta_B(5)/\delta_B(1000)$,
$\delta_\nu(5)/\delta_\nu(1000)$,

and their spread is confined within the dashed regions bounded by the indicated pairs of $\sigma_0$ and $q_0$.

that two things occur simultaneously: a density enhancement at the center and a density deficit in a spherical concentric shell. For such an empty shell a theoretical dimension is given by

$$L = 10(\Omega_0 h^2)^{-1/3} m_{15}^{1/3} \text{ Mpc}, \qquad (22)$$

where $m_{15}$ is the mass in units of $10^{15}$ $M_\odot$. This must be confronted with the observations which call for L = 50 Mpc (Kirschner et al. 1981).

A low density solution, $\Omega_0 h^2 = 0.01$, is certainly possible; it requires a very strong initial density contrast because the binding condition (Sunyaev 1971) is

$$\left(\frac{\delta\rho}{\rho}\right)_1 \gtrsim \frac{1}{\Omega_0}\frac{1}{1+z_1}, \qquad (23)$$

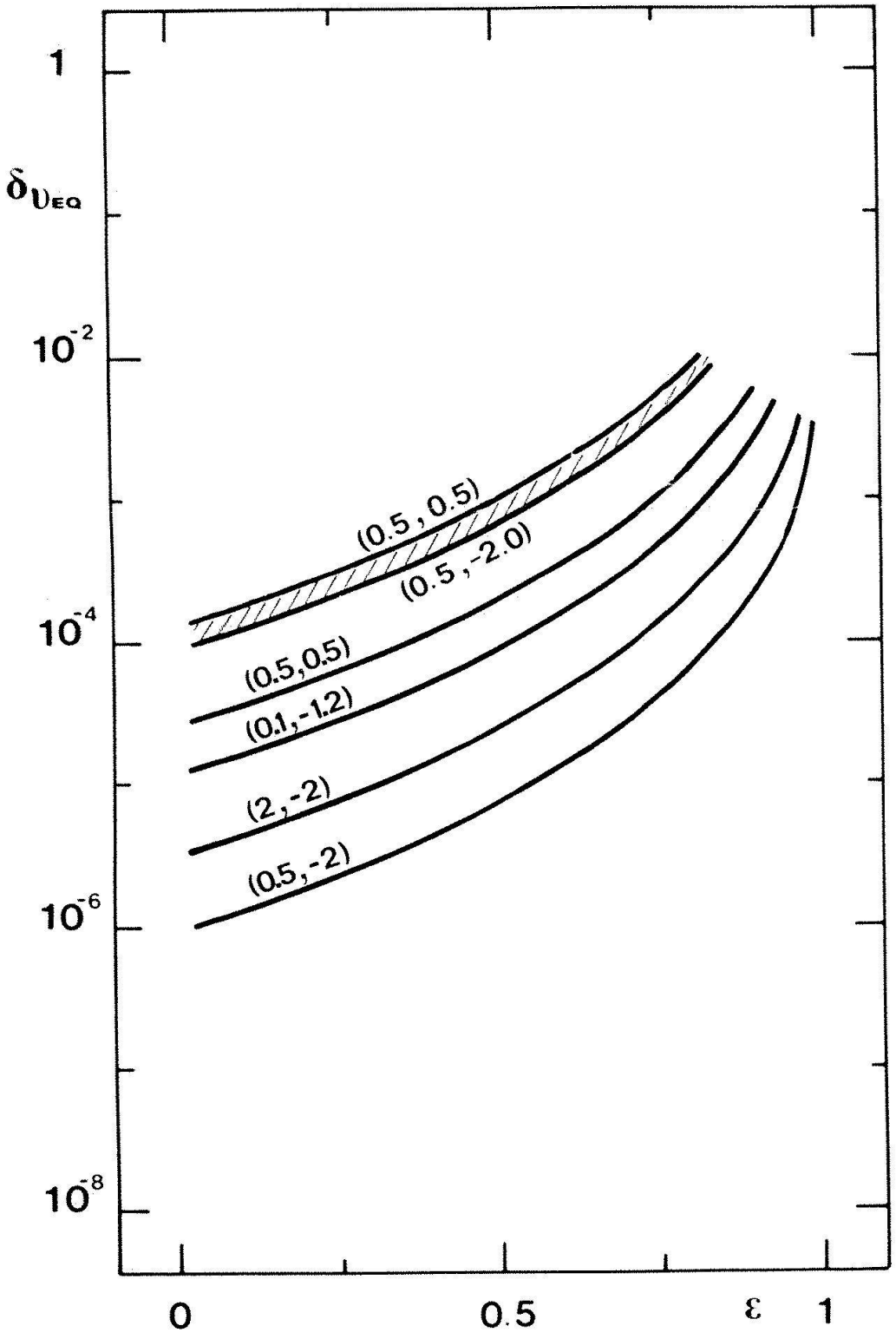



*Fig. 9.* Amplitude of the neutrino fluctuation at $z_{EQ}$ when exit from linearity occurs at the present (the four lower curves) and when exit from linearity occurs at $1+z=5$ (curves within the dashed region). Labels define the pairs $\sigma_0 - q_0$ of the unperturbed cosmological models. For $\varepsilon \ll 1$ these initial amplitudes of the order of $10^{-4}$, as in Zel'dovich's assumption, are enough to guarantee entrance into the non-linear growth well before the present; on the other hand, when $\varepsilon \lesssim 1$, larger initial amplitudes are required because neutrino self-gravity is weaker.

where $z_1$ may be assumed again to be 1000. This implies an excess binding energy per unit mass

$$b = \left(\frac{\delta\rho}{\rho}\right)_1 = \frac{\delta W}{W}, \qquad (24)$$

for which (23) translates into

$$B = b\,(1+z_1) \gtrsim \frac{1}{\Omega_0} = 10^2.$$

In a high density cosmological model, $\Omega_0 h^2 \cong 1$, the dimension of the cavity is right if one assumes $m_{15} = 100$ or $10^{17}\ M_\odot$ in such a condensation. This mass cannot be in baryons and may be in massive neutrinos; when spread over a sphere of 50 Mpc, it amounts to the density of $2 \times 10^{-29}$ g/cm$^3$, in agreement with (a.8). From an energetic point of view the binding condition is much weaker than (23) and may be formulated as

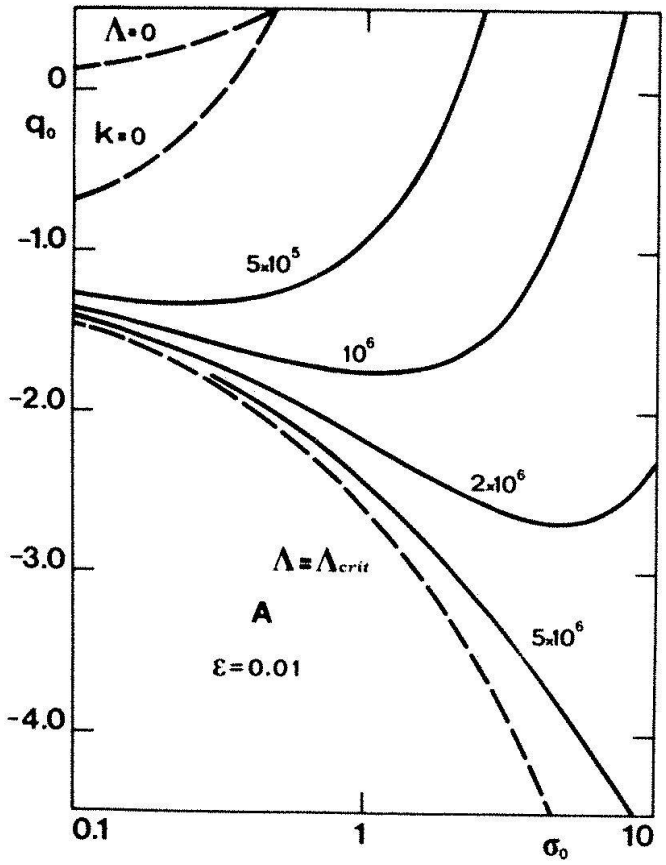



*Fig. 10.* Level curves of the amplification of the baryon growing modes in a baryon-neutrino Universe with $\varepsilon = 0.01$. The results given here must be compared with those given in figure 2: broken lines have the same meaning; the label on each solid line is the amplification evaluated again as

$$A = \delta_B(1)/\delta_B(1000);$$

it is recalled that $\delta_B(1100) = \dot{\delta}_B(1100) = 0$. The qualitative trend of the solid lines here is the same as in figure 2; from a quantitative point of view, however, there is a gain of two orders of magnitude due to the gravitational coupling between the two fluids.

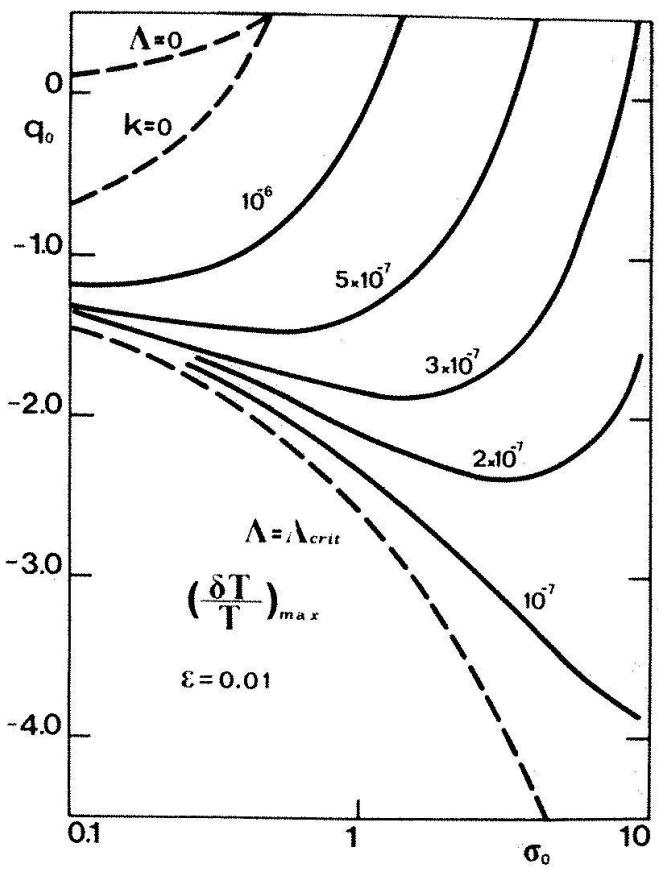



*Fig. 11.* Level curves for an upper limit on the temperature fluctuations in the microwave background evaluated under the adiabatic assumption given in (3b.21). As in figure 10, $\varepsilon = 0.01$. Labels on each curve define the expected $(\delta T/T)_{max}$ for perturbations that enter non-linearity only at the present. If more realistically we assume that this occurs at a redshift of the order of 5 or 10, the linear growth is reduced by not more than two orders of magnitude; the expected temperature fluctuations, increased by the same amount, may remain under the present detectability.

$$B \geqslant B_{crit}$$

$$= \{3[\sigma_0^2(\sigma_0 - q_0)]^{1/3} - (3\sigma_0 - q_0 - 1)\}/(2\sigma_0).$$

For example for $\sigma_0 = 0.5$, $q_0 = -2$, $B_{crit} = 0.065$ and for $\sigma_0 = 2$, $q_0 = -2$, $B_{crit} = 0.140$. We give in figure 12 density profiles for spherical condensations developing from $1+z = 1000$ with $B = B_{crit}$ in high density cosmological models. We underline that the structure of each condensation is fully non-linear by the present. By contrast in the standard open model, $\Omega_0 = 0.01$, the choice $B = 0.1$ implies $(\delta\rho/\rho)_1 = B/(1+z_1) = 10^{-4}$ and a linear growth only by a factor 10. In these conditions the density excess at the center would amount to an inconspicuous $10^{-3}$; of the same order of magnitude would result the depth of the surrounding hole.

## Abstract

Several authors have pointed out that massive neutrino condensations may trigger the formation of baryonic matter condensations in cosmology, probably on the scale of clusters of galaxies. We review their work and we give new results on the linear growth of baryon condensations from decoupling onwards under the influence of self-gravitation and the gravitational coupling to pre-

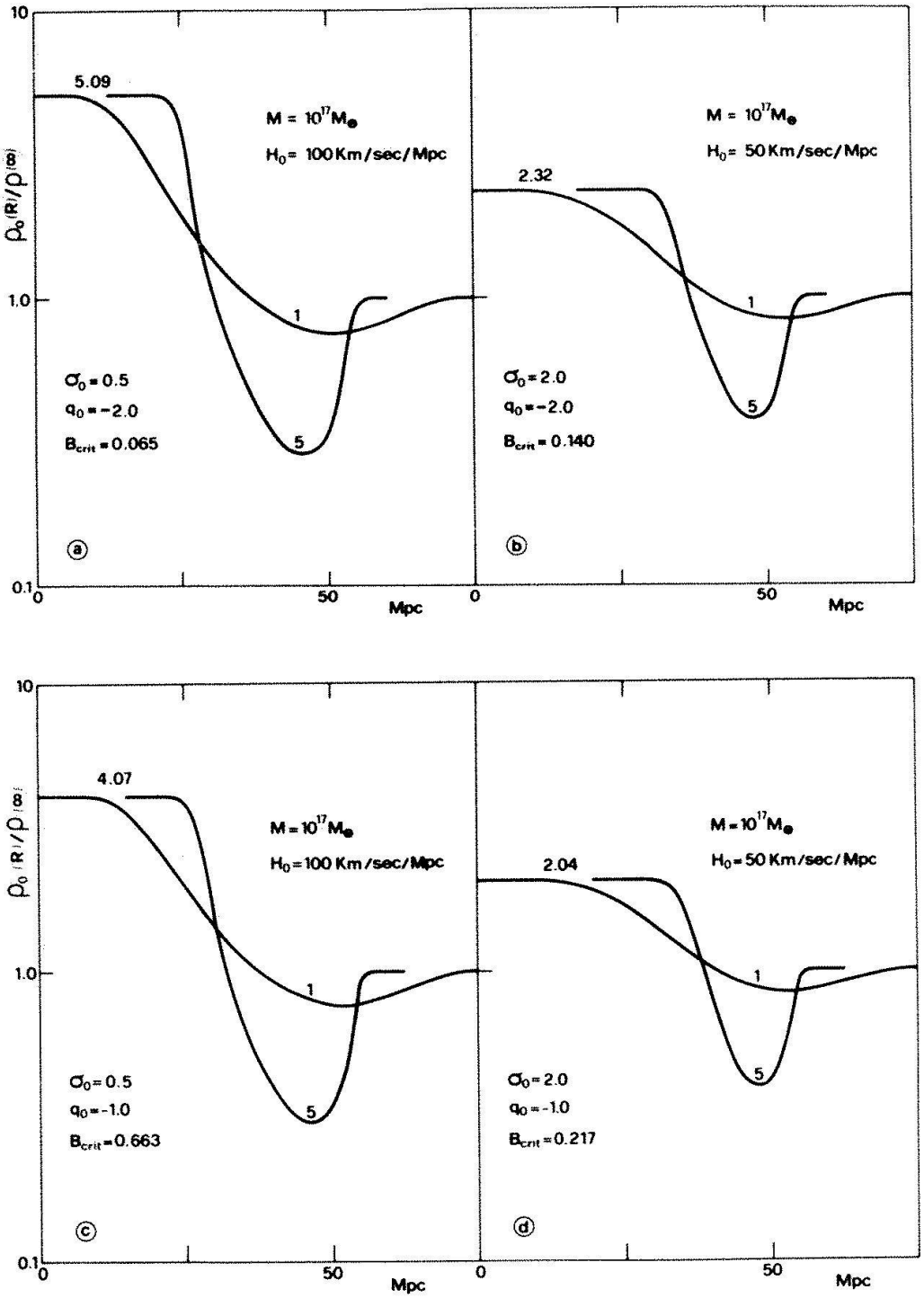



*Fig. 12.* Plot of the present density profiles in units of the asymptotic density vs. radius in spherical condensations developing in high density cosmological models. The mean density is assumed to be $2 \times 10^{-29}$ g/cm$^3$; this translates in $\sigma_0 = 0.5$ for $H_0 = 100$ km/s/Mpc (a and c) and $\sigma_0 = 2.0$ for $H_0 = 50$ km/s/Mpc (b and d). An unusually large total mass of $10^{17}$ $M_\odot$ is involved, mostly in neutrinos with only a small fraction $\varepsilon$ in baryons. Each condensation is marginally bound at the center ($B = B_{crit}$). The other details of the energetics are specified in the references quoted in the text; thus the labels 1 and 5 attached to the curves define an integer n with which we parametrize our models. The evolution is fully nonlinear; cavities are shown to develop around each condensation with dimensions of the order of 50 Mpc.

existing neutrino condensations. We parametrize our work by the ratio of the present density in baryons to the present total density; such a number is likely to be small. We also allow for a positive cosmological constant, which – as it has been suggested – may be needed if the cosmic density in neutrinos is around the closure value.

As it was already known, we find that the fractional baryon density enhancements reach quickly the level of the fractional neutrino density enhancements and remain locked to the latter thereafter. Secondly, in agreement with previous work of ours, we find that at low redshift the linear growth of the condensations (of either component) is stronger in that region of the parameter space where curvature is positive and the cosmological constant exceeds the critical value. If this really applies to our Universe, the latter argument may further justify the lack of detection of small scale fluctuations in the microwave background or, at least, help push down their theoretical upper limit.

Finally we give some results on the formation of non-linear condensations with spherical symmetry; the motivation for this work lies in the observation of large scale voids (linear dimensions of the order of 100 Mpc). A high density Universe is again preferred because in a low density model similar condensations would not have reached non-linear growth.

## References


Aaronson, M., Mould, J., Huchra, J., Sullivan, W.T., Schommer, R.A., and Bothun, G.D.: 1980, Astrophys. J., 239, 12.

Bahcall, N.A.: 1977, Ann. Rev. Astron. Astrophys., 15, 505.

Bisnovatyi-Kogan, G.S., and Novikov, I.D.: 1980, Sov. Astron., 24, 516.

Bisnovatyi-Kogan, G.S., Lukash, V.N., and Novikov, I.D.: 1980, Paper presented at the Fifth European Regional Meeting (I.A.U.), Liege, Belgium.

Bludman, S.A.: 1976, Gen. Rel. Grav., 7, 569.

Bond, J.R., Efstathiou, G., and Silk, J.: 1980, Phys. Rev. Lett., 45, 1980.

Bonnor, W.B.: 1957, Monthly Not. Roy. Astron. Soc., 117, 104.

Boynton, P.E.: 1978, in "The Large Scale Structure of the Universe", eds., M.S. Longair, J. Einasto, D. Reidel Publ. Co., Dordrecht, Holland.

Cavaliere, A., and Fusco-Femiano, R.: 1976, Astron. Astrophys., 49, 137.

Cowsik, R., and McClelland, J.: 1972, Phys. Rev. Lett., 29, 669.

Cowsik, R., and McClelland, J.: 1973, Astrophys. J., 180, 7.

Davis, M., Geller, M.J., and Huchra, J.: 1978, Astrophys. J., 221, 1.

Davis, M., and Boynton, P.: 1980, Astrophys. J., 237, 365.

de Vaucouleurs, G., and Bollinger, G.: 1979, Astrophys.J., 233, 433.

Dolgov, A.B., and Zel'dovich, Ya.B.: 1981, Rev. Mod. Phys., 63, 1.

Doroshkevich, A.G., Sunyaev, R.A., and Zel'dovich, Ya.B.: 1974, in "Confrontation of Cosmological Theories with Observational Data", Longair, M.S., ed., Reidel Publ. Co., Dordrecht, Holland.

Doroshkevich, A.G., Zel'dovich, Ya.B., Sunyaev, R.A., and Khlopov, M.Yu.: 1980a, Sov. Astr. Lett., 6, 252 and 257.

Doroshkevich, A.G., Khlopov, M.Yu., Sunyaev, R.A., Szalay, A.S., and Zel'dovich, Ya.B.: 1980b, "Proceedings of the Xth Texas Symposium", Baltimore, MD.

Faber, S.M., and Gallagher, J.S.: 1979, Ann. Rev. Astron. Astrophys., 17, 135.

Field, G.B.: 1975, in "Stars and Stellar Systems", vol.9, University of Chicago Press, Chicago, IL.

Gershtein, S.S., and Zel'dovich, Ya.B.: 1966, J.E.T.P. Lett., 4, 174.

Gilbert, I.H.: 1966, Astrophys. J., 144, 233.

Gott, J.R., Gunn, J.E., Schramm, D.N., and Tinsley, B.M.: 1974, Astroph. J., 194, 543.

Gott, J.R. III: 1979, in "Physical Cosmology", Balian, R., et al. eds., North-Holland Publ. Co., Dordrecht, Holland.

Greenstein, J.L.: 1980, Physica Scripta, 21, 759.

Gunn, J.E.: 1978, in "Observational Cosmology", Maeder, A., Martinet, L., Tamman, G., eds, Swiss Society of Astronomy and Astrophysics, Geneva.

Gunn, J.E., and Tinsley, B.M.: 1975, Nature, 257, 454.

Harrison, E.R.: 1967, Rev. Mod. Phys., 39, 862.

Hoffman, G.L., Olson, D.W., and Salpeter, E.E.: 1980, Astrophys. J., 242, 861.

Iben, I.: 1974, Ann. Rev. Astron. Astrophys., 12, 215.

Jones, B.J.T.: 1976, Rev. Mod. Phys., 48, 107.

Kirschner, R.P., Oemler, A., Jr., and Schechter, P.L.: 1979, Astron. J., 84, 951.

Kirschner, R.P., Oemler, A., Jr., Schechter, P.L., and Shectman, S.A.: 1981, Astrophys. J., 248, L57.

Klinkhamer, F.R., and Norman, C.A.: 1981, Astrophys. J., 243, L1.

Lea, S.M., Silk, J., Kellogg, E., and Murray, S.: 1973, Astrophys. J., 184, L105.

Lubimov, V.A., Novikov, E.G., Nozik, V.Z., Tretyakov, E.F., and Kosik, V.S.: 1980, Phys. Lett., 94B, 266.

Luminet, J.P., and Schneider, J.: 1981, Astron. Astrophys., 98, 412.

Lynden-Bell, D.: 1967, in "Relativity Theory and Astrophysics", Ehlers, J., ed., Am. Math. Soc., Providence, R.I..

Malina, R.F., Lea, S.M., Lampton, M., and Bowyer, S.: 1978, Astrophys. J., 219, 795.

Markov, M.A.: 1964, Phys. Let., 10, 122.

McVittie, G.C.: 1965, "General Relativity and Cosmology", The University of Illinois Press, Urbana, IL.

Mészàros, P.: 1974, Astron. Astrophys., 37, 225.

Occhionero, F., Vittorio, N., Carnevali, P., and Santangelo, P.: 1980, Astron. Astrophys., 86, 212.

Occhionero, F., Veccia-Scavalli, L., and Vittorio, N.: 1981a and b, Astron. Astrophys., 97, 169 and 99, L12.

Olive, K.A., Schramm, D.N., Steigman, G., Turner, M.S., and Yang, J.: 1981, Astrophys. J., 246, 557.

Ostriker, J.P., and Peebles, P.J.E.: 1973, Astrophys. J., 186, 467.

Partridge, R.B.: 1980, Physica Scripta, 21, 624.

Peebles, P.J.E.: 1981, "Physical Cosmology", Princeton University Press, Princeton, NJ.

Peebles, P.J.E., and Yu, J.T.: 1970, Astrophys. J., 162, 815.

Petrosian, V.: 1974, in "Confrontation of Cosmological Theories with Observational Data", Longair, M.S., ed., Reidel Publ. Co., Dordrecht, Holland.

Press, W.H.: 1980, Physica Scripta, 21, 702.

Press, W.H., and Vishniac, E.T.: 1980, Astrophys. J., 236, 323.

Sandage, A., and Tammann, G.A.: 1976, Astrophys. J., 210, 7.

Sato, H., and Takahara, F.: 1981, Prog. Theor. Phys., 65, 374.

Schramm, D.N., and Wagoner, R.V.: 1977, Ann. Rev. Nucl. Part. Sci., 27, 37.

Schramm, D.N., and Steigman, G.: 1980, First Prize Essay, Grav. Res. Foundation.

Schramm, D.N., and Steigman, G.: 1981, Astrophys. J., 243, 1.

Shapiro, S.L., Teukolsky, S.A., and Wasserman, I.: 1980, Phys. Rev. Lett., 45, 669.
Shvartsman, V.F.: 1969, JETP Lett., 9, 184.
Silk, J.: 1968, Astrophys. J., 151, 459.
Silk, J., and Wilson, M.L.: 1980, Physica Scripta, 21, 708.
Stecker, F.W.: 1980, Phys. Rev. Lett., 44, 1237.
Steigman, G.: 1979, Ann. Rev. Nucl. Part. Sci., 29, 313.
Stewart, J.M.: 1972, Astrophys. J., 176, 323.
Symbalisty, E.M.D., Yang, J., and Schramm, D.N.: 1980, Nature, 288, 143.
Sunyaev, R.A.: 1971, Astron. Astrophys., 12, 190.
Sunyaev, R.A., and Zel'dovich, Ya.B.: 1970, Astrophys. Space Sci., 7, 3.
Sunyaev, R.A., and Zel'dovich, Ya.B.: 1980, Ann. Rev. Astron. Astrophys., 18, 537.
Szalay, A.S., and Marx, G.: 1976, Astron. Astrophys., 49, 437.
Tinsley, B.M.: 1977, Phys. Today, 30, 32.
Tytler, D.: 1981, Nature, 291, 289.
Van der Bergh, S.: 1981, Paper presented at 158th Meeting of AAS, Calgary, Canada.
Vidal-Madjar, A., Laurent, C., Bonnet, R.M., and York, D.G.: 1977, Astrophys. J., 211, 91.
Yahil, A., Sandage, A., and Tammann, G.A.: 1980, Physica Scripta, 21, 635.
Yang, J., Schramm, D.N., Steigman, G., and Rood, R.T.: 1979, Astrophys. J., 227, 697.
York, D.G., and Rogerson, J.B.: 1976, Asrophys. J., 208, 378.
Wasserman, I.: 1981, Astrophys. J., 248, 1.
Weinberg, S.: 1971, Astrophys. J., 168, 175.
Weinberg, S.: 1972, "Gravitation and Cosnology", J. Wiley, New York, NY.
Weinberg, S.: 1977, "The First Three Minutes", Basic Book, Inc., Publ., New York, NY.
Zel'dovich, Ya.B.: 1968, Sov. Phys. Uspekhi, 11, 381.
Zel'dovich, Ya.B.: 1970, Astron. Astrophys., 5, 84.
Zel'dovich, Ya.B., and Sunyaev, R.A.: 1980 Sov. Astr. Lett., 6, 249.

*Address of the authors:*

F. Occhionero
N. Vittorio
M. Boccadoro
S. De Luca
Istituto di Astrofisica Spaziale
C.N.R., via E. Fermi, 21
I-00044 Frascati-Roma (Italy)

# On the Colors of Faint Galaxies

Roland Buser

**Abstract**

The role of color determinations in the context of galaxy evolution and cosmology is briefly described. A rough sketch is given of spectral evolutionary models of galaxies and their use in computing synthetic galaxy colors and magnitudes as functions of redshift. Bruzual and Kron's (1980) application of these theoretical calculations in the analysis of the observations of a complete sample of faint galaxies is discussed as an important and promising step toward the solution of the perennial riddle of the universe and its galaxies, via multicolor photometry.

## Cosmology and Galaxy Evolution

Naturally, the most prominent index of a star's evolution is its luminosity changing with time. If a galaxy is a 'closed' system of stars and gas that was or still is transforming gas into stars, we must expect the galaxy's luminosity to change with time accordingly. As for the stars themselves, the time scale for galactic evolution is so large as to prevent us from observing the evolution of any one galaxy directly. Yet we can hope that the simultaneous observation of many galaxies reveals us different stages in the lives of galaxies, from which we might be able to recover a natural sequence of stages in the evolutionary history of the typical galaxy. – How, then, would we wish to tackle the problem of galaxy evolution?

Consider, first, the real home of the galaxies: the universe. The universe is so large that by looking at its galaxies in its ever deeper realms we effectively see these galaxies at ever more distant past times. Since galaxy luminosities are likely to be limited by general astrophysical constraints, galaxies at larger distances have, on the average, fainter apparent magnitudes; the relation is not a strict one because, for a given apparent magnitude, for each galaxy there is of course a luminosity-distance ambiguity. Still, it can be said that comparing galaxies at successively larger distances holds a clue to the evolution of galaxies, and this necessarily involves observations at faint magnitude levels.

Consider now the real context of galaxy evolution: the evolution of the universe itself. The most fundamental observable dimension of the evolving universe is not the distance, but the redshift. There is no ambiguity in the redshift-distance relation for galaxies. Hence galaxies at successively larger redshifts are also at successively larger distances and thus sample increasingly past cosmic epochs. The study of faint galaxies at different redshifts is therefore a promising approach to the evolution of galaxies.

Consider, finally, the number of galaxies as a function of apparent magnitude and redshift, A(m,z). If we assume space to be homogeneously and isotropically populated with galaxies, then the number of galaxies, A(z), in a redshift shell of radius z and thickness dz is proportional to the volume of this redshift shell, i.e., A(z) samples the (differential) volume element dV/dz, which is predicted different for different values of the deceleration parameter, $q_0$, in Friedmann models of the universe. Hence A(z) is a cosmological test, providing $q_0$. Knowledge of this value of $q_0$ then allows us to determine the (luminosity) evolution as a function of redshift by studying the Hubble diagram (i.e., the (m,z) relation) and the count-magnitude diagram (i.e., the A(m) relation) of the sample galaxies. Adoption of a value for $H_0$, the Hubble expansion parameter, eventually gives the evolution as a function of time via the time-redshift relation, which depends on $q_0$ and scales as $H_0^{-1}$ (Sandage 1961).

This line of thought, then, suggests that in order to get hold of the evolution of galaxies, magnitudes and redshifts be obtained for large numbers of faint galaxies. – But how can redshifts be determined for large numbers of faint galaxies? That's the time for the colors to enter the scene.

## Color and Redshift

The color of a galaxy is obtained from observations of its integrated brightnesses in two or more fixed wavelength bands typically several hundred Angstroms wide. Such a color is a coarse measure of spectral information and is therefore apt to reflect the redshift of its underlying spectral energy distribution in principle. In practice we can compute, by artificially redshifting the known intrinsic spectral energy distribution of a galaxy, the fluxes falling into the fixed photometric bands at successive redshift steps and produce colors as a function of redshift for that particular galaxy. Indeed, galaxy colors as a function of redshift represent the very kind of structure that we are looking for in order to determine redshifts from observed colors.

In the present context, the practical application requires that at least two rather important complications to the above simple procedure be taken into account.

First, different galaxies at zero redshift span a range of colors to begin with, due to their different intrinsic spectral energy distributions. The distribution of colors of bright galaxies may lead us to think of the morphological type of galaxy as the fundamental variable behind the observed variations in the spectral energy distributions. Relations between color and redshift must therefore be evaluated for different morphological galaxy types.

Second, any relations between color and redshift computed by artificially redshifting available bright galaxy spectra (observed at $z \approx 0$) necessarily ignore the time-redshift relation. In order to interpret the observed colors of truly reshifted galaxies, color evolution as a function of redshift (hence time) needs to be taken into account.

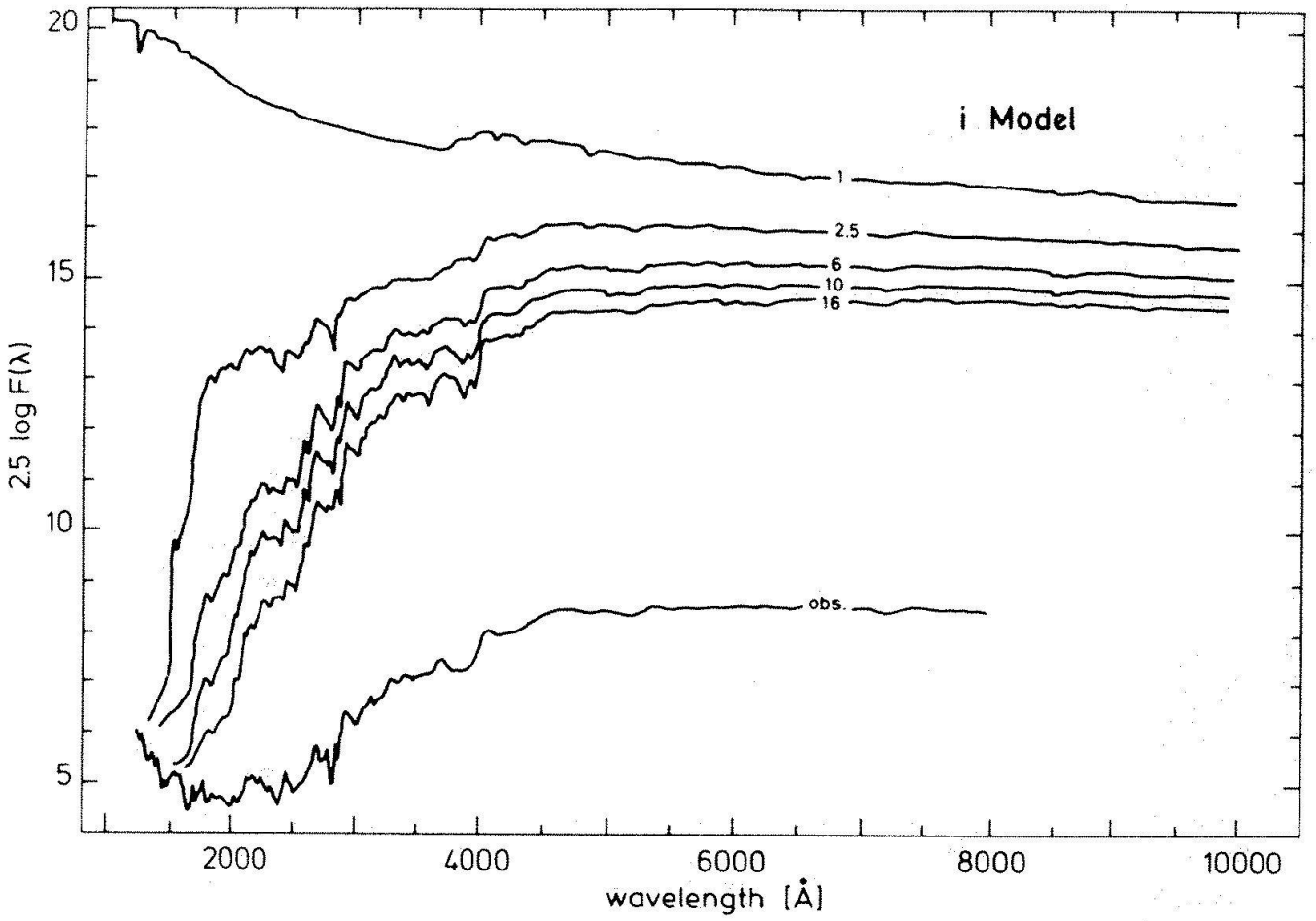
For lack of observed spectral energy distributions of galaxies of different morphological types and redshifts, these requirements imply that evolving spectra of galaxies be modeled.

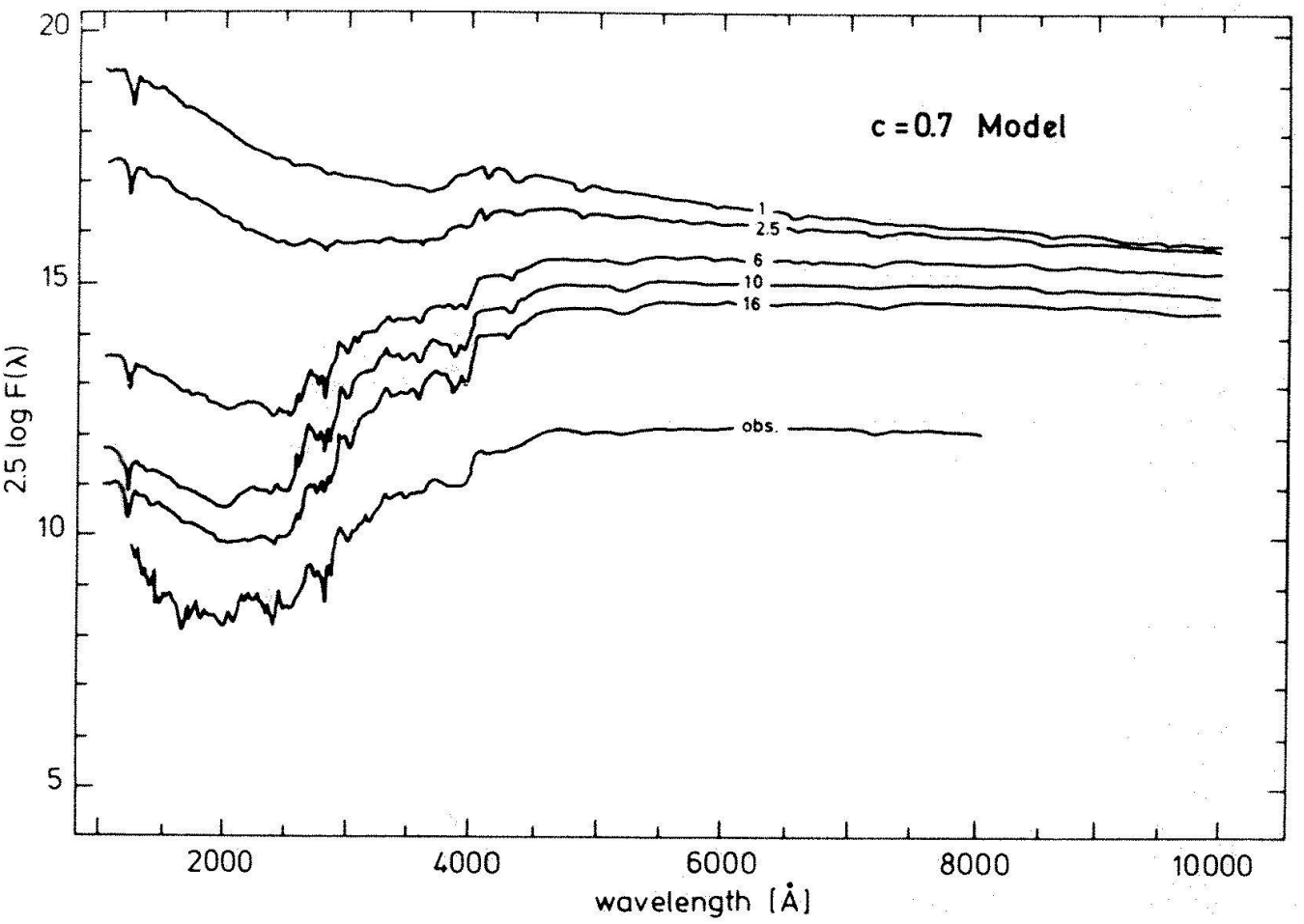
## Spectral Evolution of Galaxies

Modeling the spectral evolution of galaxies assumes that a galaxy is a closed system of gas and stars. At any given time the emitted spectral energy distribution of the galaxy is the sum of the absolute spectral energy distributions of all the stars existing in the galaxy at that time. The number of stars that contribute to the galaxy's luminosity at a given time is the difference between the number of stars that have been born out of the galaxy's gas, and the number of stars that have finished their active lives by that time. Both numbers of stars in this difference are functions of time: the first is given by the star formation rate, and the second results from applying stellar evolution theory to the stars formed. Since the time scale for stellar evolution is a function of the stellar mass, it is assumed that via the star formation rate gas is transformed into stars according to a stellar mass spectrum (i.e., the initial mass function), which itself is assumed to be independent of time and to be of the same general form as the stellar mass function observed in the solar neighborhood (Salpeter 1955).

For each stellar mass and for any given time, a theoretical evolutionary track provides the observable quantities like spectral type or B-V color, and absolute magnitude, which in turn determine the selection of the associated stellar spectral energy distribution from a comprehensive library of spectrophotometric data. Adding all the stellar spectra corresponding to the stars in the galaxy at a given time yields, finally, the spectral energy distri-

*Fig. 1.* Examples of evolving energy distributions for galaxy models with different assumed histories of the star formation rate (see text). The distributions are shown at different ages from 1 to 16 Gyr, as indicated. For comparison, the observed spectral energy distribution of a typical nearby elliptical galaxy (obs.) is displayed on the same flux scale, but with an arbitrary shift of the zero point. Note the pronounced differences between the models in the evolution of their ultraviolet spectra ($\lambda \lesssim 4000$ Å), which will show up as color differences in the visual wavelength range if such galaxies are observed at high redshift ($z \approx 1$).
(Figure adapted from Bruzual 1981.)


i Model
1
2.5
6
10
16
obs.
20
15
10
5
2.5 log F(λ)
2000
4000
6000
8000
10000
wavelength [Å]


c = 0.7 Model
1
2.5
6
10
16
obs.
20
15
10
5
2.5 log F(λ)
2000
4000
6000
8000
10000
wavelength [Å]

bution of the whole galaxy, at that particular time.

A galaxy spectral evolutionary model then consists of a series of absolute spectral energy distributions calculated, according to the above scheme, for different times.

The most important parameter of such a model is the star formation rate as a function of time. For example, observations of elliptical galaxies suggest that these systems contain mainly old stars and little gas. Hence star formation was probably very efficient and exhaustive at early stages of their evolution. On the other hand, observations of spiral and irregular galaxies reveal considerable amounts of gas and significant star formation still going on at their present - i.e., presumably late - evolutionary phases. Dwarf galaxies, still, appear to have started their lives as rather low-density systems, with a low star formation rate that may reach its highest value only at late times in their evolution.

Most currently available models (e.g., Tinsley 1980, Bruzual 1981) attempt to approximate analytically these different histories of star formation in galaxies, as suggested by the observed properties of different morphological types. 'Initial burst (i) models' have a high constant star formation rate during the first billion years or so, after which star formation ceases. For 'continuous burst (c) models' the star formation rate is highest at the beginning and thereafter decreases exponentially, the time scale for the decrease being a free parameter, which allows the gas mass to total mass ratio at the present time to cover the observed range. 'Delayed burst (d) models' start with a low but growing star formation rate which peaks after an adjustable time scale (e.g., 10 Gyr).

Figure 1 illustrates a few important aspects of such theoretical models constructed by Bruzual (1981). Evolving galaxy spectra at ages between 1 and 16 Gyr are shown for an i-model (upper panel) and for a c-model (lower panel). While there is a similarly slow change with time of the slope of the visual and near-infrared spectra of both models, the evolutionary differences between the models are most prominent in the ultraviolet. The absence of star formation after 1 Gyr makes the i-model get rapidly darker at ultraviolet wavelengths and its whole spectrum dominated by late-type giant stars; on the other hand residual formation of massive hot stars throughout the life of the c-model provides the main source of its less rapidly dimming ultraviolet light.

The curves labeled 'obs.' represent the observed spectrum of a typical nearby elliptical galaxy and are displayed to illustrate how the models can be checked on their ability to match real galaxies. In view of the scarcity of presently available observations of ultraviolet galaxy spectra - which are most important in putting constraints on the adopted histories of star formation -, the main purpose of such comparisons is to ascertain that the variety of observed galaxy spectra be covered by a variety of evolving model energy distributions at the same present age.

The variety of model energy distributions as functions of time can then be used to calculate the colors and magnitudes of the model galaxies as functions of redshift, after a cosmological model has been specified (i.e., values for $q_0$ and $H_0$ have been adopted), providing the time-redshift relation and an age of the universe consistent with the adopted epoch for galaxy formation.

Figures 2a and 2b give examples of the resulting relations for the J-F color and the F magnitude defined by Kron's (1980a) photographic broad-band system. The curves in both figures are for the same representative models of galaxies having present ages of 16 Gyr, adopted in a Friedmann universe with $H_0 = 50$ km s$^{-1}$ Mpc$^{-1}$ and $q_0 = 0$. The top curves are for a nonevolving galaxy, i.e., with an i-model spectrum frozen at its present age. The other curves are for evolving galaxy models, labeled according to their particular values for the star formation rate parameter, a lower numerical value for the c-models meaning that gas consumption - or star formation - decreases more slowly with time.

Qualitatively, these results demonstrate quite generally - although Figures 2 are for a particular color system - that evolution has a significant effect on the colors and magnitudes of galaxies, making them bluer and brighter, at all redshifts, than an hypothetically unevolving source of the same present color and absolute magnitude. Galaxies which by the present time have the same intrinsic colors and absolute magnitudes may

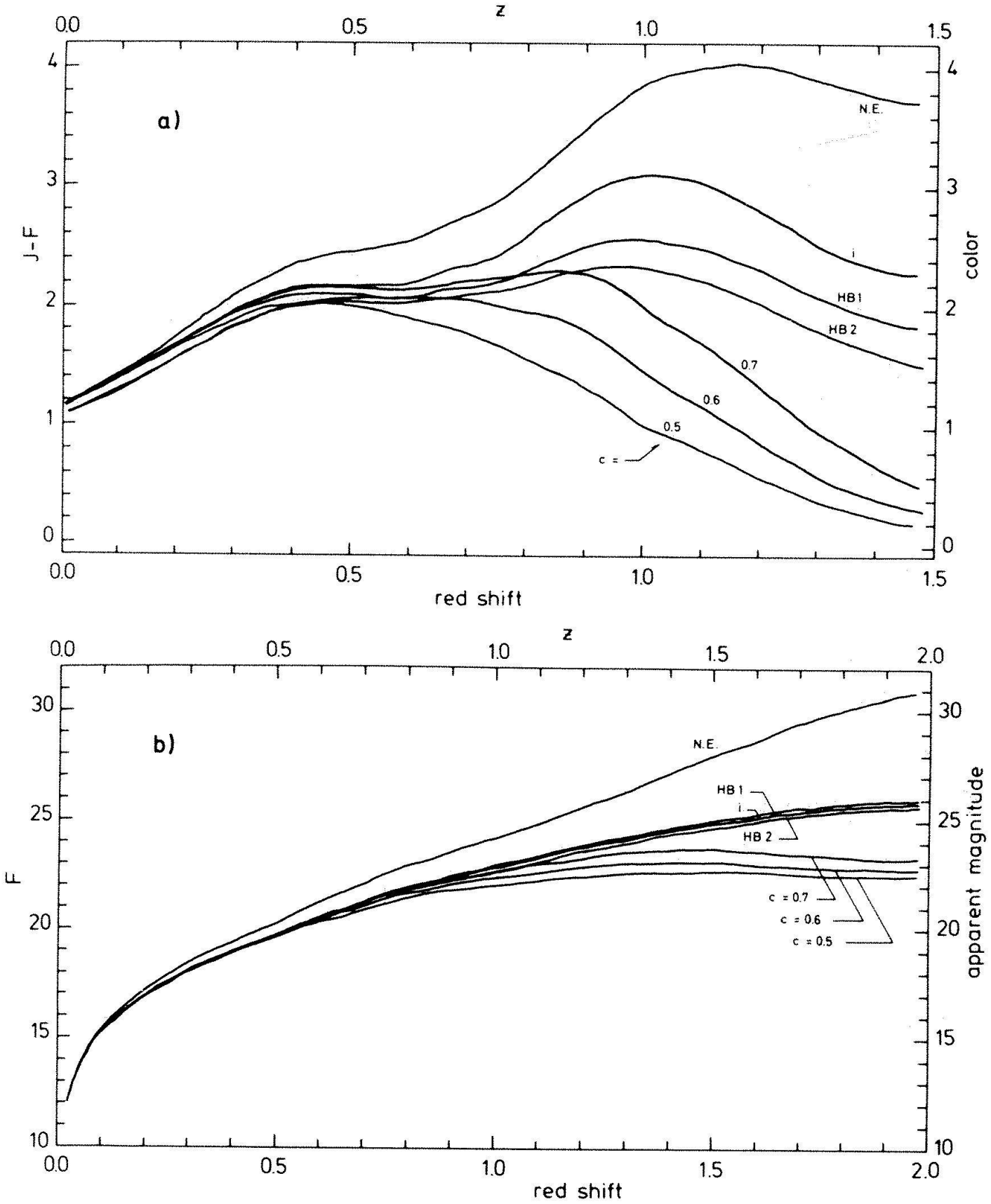



*Fig. 2.* Curves of intrinsic J-F color (a) and apparent F magnitude (b) as functions of redshift, for representative galaxy models. The curves labeled N.E. were computed from the i-model spectrum at the present age ($z = 0$), i.e., from an unevolving source. The differences between the N.E. and the i-model curves therefore illustrate the evolutionary effects, which operate in addition to the k-correction. Curves labeled HB1 and HB2 refer to the i-model with horizontal branch star light added. For the c-models, a lower numerical value indicates that the star formation rate decreases more slowly with time. In Figure 2b, all models were normalized to the same absolute magnitude at their present age.

(Figures adapted from Bruzual 1981.)

in fact have evolved into this specific condition on rather different routes and through widely dispersed colors and luminosities at earlier epochs.
Quantitatively, such predictions for galaxy colors and magnitudes as functions of redshift can now be exploited in the interpretation of the observed colors and magnitudes of faint galaxies.

## Counts and Colors of Faint Galaxies

As we have seen in the previous sections, cosmology and galaxy evolution are two strongly connected subjects. The present approach to these subjects pursues the questions: How many galaxies are there in the universe in redshift shells of successively larger radius, and what is the distribution of the apparent galaxy brightnesses in each of these shells? Technically, this problem translates into the quest for the number of galaxies as a function of apparent magnitude and redshift, A(m,z). Since this necessarily involves large numbers of faint galaxies, a more practicable step toward A(m,z) consists in studying the number of galaxies as a function of apparent magnitude and color, A(m,c), where the color c serves as a substitute for the redshift, due to the existence of a color-redshift relation, c(z) (Figure 2a). Because different morphological galaxy types trace out different c(z) relations, and because each of these may be multiple-valued (i.e., a given value of c may correspond to different values of z), several colors rather than a single one will eventually be needed to determine A(m,z) from observations.
The interpretation of faint galaxy data rests on our knowledge of the corresponding data for bright galaxies. More to the point: from the local properties of the universe, as reflected by the number density, the luminosity function, and the intrinsic color distribution of galaxies observed at bright magnitudes (i.e., $z \approx 0$), the numbers and the distributions of apparent magnitudes and intrinsic colors of galaxies expected to be observed at faint magnitudes (i.e. at $z > 0$) can now be computed via the assumption of the global properties of the universe, which are specified by the values assigned to the parameters of Friedmann models. This calculation of the past from the present is mediated by the evolutionary models of galaxies. Its results can be expressed in the form A(m,c), which readily allows comparison with observations.
Bruzual and Kron's (1980) interpretation of Kron's (1980a) photometry of a complete sample of faint galaxies is an adequate representation of the current state of the art. In Figures 3a and 3b the observed data are compared with model predictions of the $A(m = J)$ and $A(c = J - F)$ distributions, respectively, calculated for the parameters listed in Table 1. For all these A(m,c)-models in the Table, a variety of evolutionary galaxy models were selected, whose present age colors match the observed colors of bright galaxies (i.e., at $J \approx 15$). In the case of the no-evolution model, the spectral energy distributions of these representative galaxy models were frozen at $z = 0$, i.e., their colors and magnitudes as a function of redshift were computed from unevolving sources.
The main feature of Figure 3 is the fact that the models are capable of reproducing the observed distributions to first order. This can be taken as to indicate that the general model structure is reasonable in that at least its more important basic building blocks have been identified correctly.
The particular choice of values assigned to the parameters in Table 1 serves to illustrate the sensitivity of the A(m,c)-model predictions with respect to the two principal ingredients: the cosmological model and galaxy evolution.

*Table 1.* Parameters of A(m,c)-models.

| | Cosmological Model | | Galaxy Models | |
|---|---|---|---|---|
| A(m,c)-model | $q_0$ | $H_0$ [km s$^{-1}$ Mpc$^{-1}$] | Present Age, $t_0$ [Gyr] | Evolution |
| A | 0 | 60 | 16 | yes |
| B | 0 | 60 | 16 | no |
| C | 0.5 | 50 | 13 | yes |

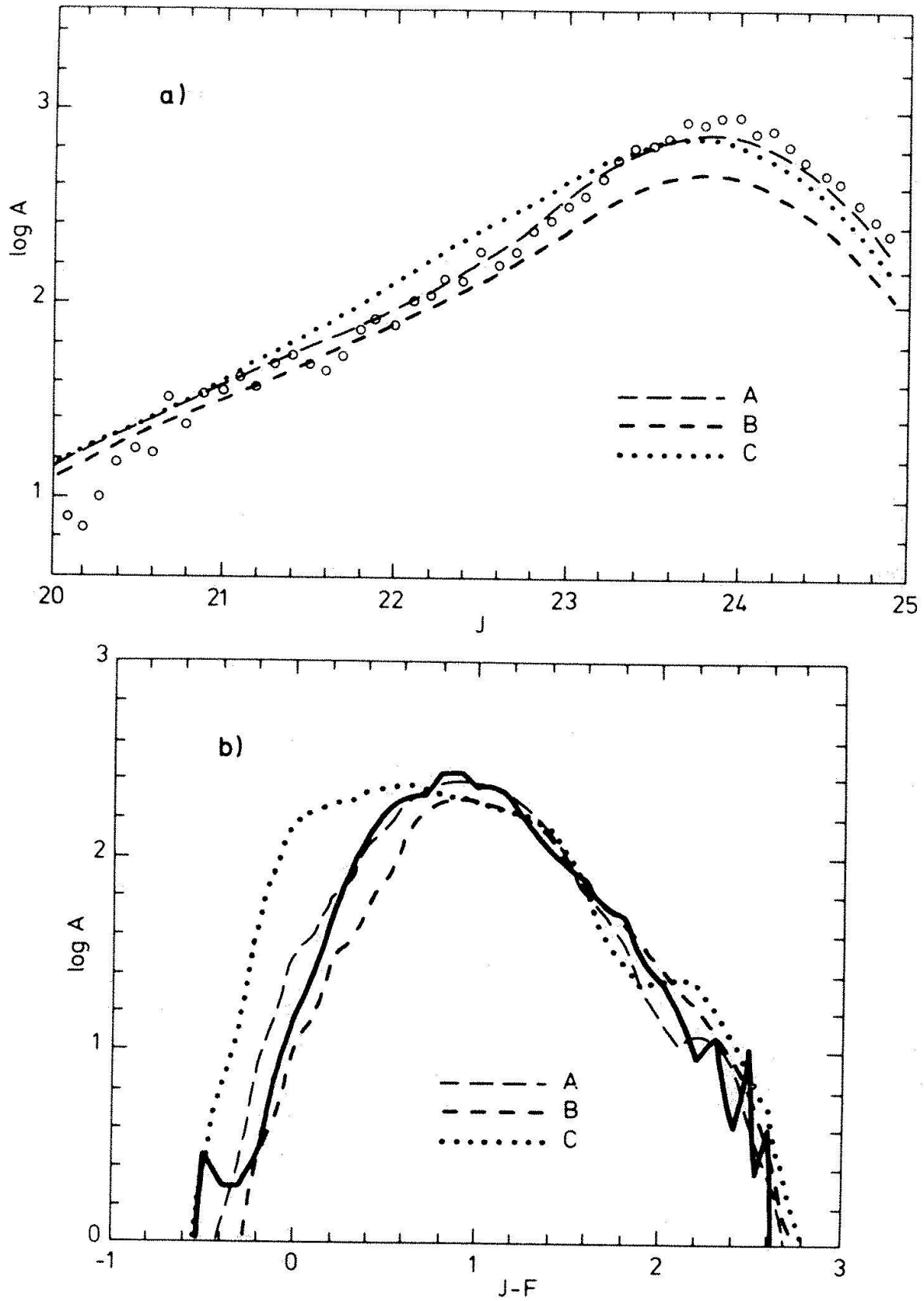



*Fig. 3.* (a) Differential galaxy counts per 0.1 magnitude interval per 1080 square arcmin, as a function of apparent J magnitude. Open circles represent observations by Kron (1980a) in a field centered on Selected Area 57. Curves represent models listed in Table 1, with random errors in J and increasing count incompleteness with increasing J magnitude incorporated.
(b) Differential galaxy counts per 0.1 magnitude interval in J-F color, per 1080 square arcmin, as a function of J-F color, for the magnitude interval $22 \leq (J+F)/2 < 23$. The heavy solid line represents observations by Kron (1980a) in a field centered on Selected Area 57. Other curves represent models listed in Table 1, with random errors in color and magnitude incorporated.

(Figures adapted from Bruzual and Kron 1980.)

The evolutionary effects are demonstrated in Figure 3 by models A and B, whose only difference is that model A has, and model B has not, its galaxy energy distributions evolved. Roughly speaking, the no-evolution model B predicts slightly redder colors (especially on the blue side of the distribution in Figure 3b) and smaller counts (Figure 3a) than observed for $J \gtrsim 22$. The desired improvement, namely to get slightly bluer colors and larger counts at all magnitudes $J \gtrsim 22$, is just about achieved by simply letting the galaxy energy distributions evolve in model A. This behavior is explained by the fact that the evolving galaxy models have higher star formation rates in the past, making them bluer than nonevolving sources at all redshifts (Figure 2a). But then, intrinsically bluer galaxies get dimmer less rapidly with redshift than do redder ones (Figure 2b), so for a given apparent magnitude interval or limit, the bluer galaxies are seen out to larger redshifts (hence distances). The net effect of evolution is then a shift toward bluer average galaxy color and a larger number of galaxies seen at a given (faint) apparent magnitude.

The cosmological effects are demonstrated in Figure 3 by models A and C. While model A ($q_0 = 0$) provides a good fit to the observations in both the number-magnitude and the number-color diagrams, model C ($q_0 = 0.5$) fits both observed distributions poorly, rising too steeply at the brighter magnitudes in Figure 3a, and predicting too many blue galaxies in Figure 3b. These differences in model behavior arise mainly from the different time-redshift and volume-redshift relations involved. At a given redshift, the look-back time is a larger fraction of the adopted present galaxy age, $t_0$, for $q_0 = 0.5$ than it is for $q_0 = 0$. Therefore, at a given redshift, galaxies are being observed in earlier stages of their evolution in the $q_0 = 0.5$ case than they are in the $q_0 = 0$ case, which means that – in the present picture where galaxies are in general assumed to have had higher star formation rates, i.e. have been evolving more rapidly, at earlier times – these galaxies are also being observed bluer and brighter in a $q_0 = 0.5$ universe. Alternatively, for a given apparent magnitude, the redshifts sampled for $q_0 = 0.5$ span a larger range than for $q_0 = 0$, making the evolutionary effects appear more pronouncedly in model C than in model A. However, at sufficiently faint apparent magnitudes (or high enough redshifts) in Figure 3a, the number of galaxies predicted from model C drops below the model A results essentially because the size of the volume element at large z is much smaller for $q_0 = 0.5$ than for $q_0 = 0$.

While the models presented in Figure 3 illustrate well their feasibility as possible self-consistent schemes for the interpretation of faint galaxy observations, they do of course not exhaust all possible combinations of parameters. Consequently, even though model A provides a very good fit to the data, no claims of uniqueness can be made. However, the very fact that a successful fit has been found may be summarized in the following conclusion: The universe at faint magnitudes ($J \approx 24$) appears the way it would be expected to appear from extrapolation of what is known about galaxies at bright magnitudes ($J \approx 15$), if a Friedmann cosmological model with $q_0 \approx 0$ is assumed. The case for a $q_0 > 0.5$ universe would require galaxy evolutionary schemes very different from those considered by Bruzual and Kron and by most other authors.

The next logical step then consists in finding independent constraints for the histories of

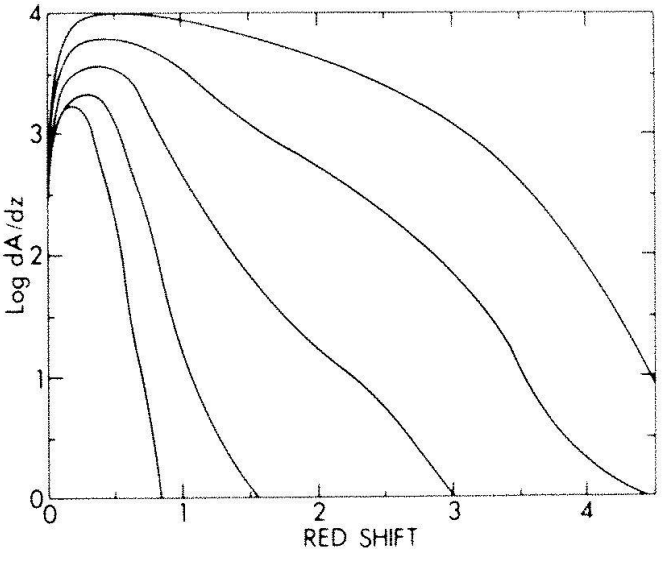



*Fig. 4.* Distribution of galaxy redshifts as a function of apparent magnitude, predicted for model A in Table 1. Log dA/dz gives the number of galaxies per unit redshift interval. The five curves represent, from top to bottom, intervals of one magnitude centered on $(J + F)/2 = 24.5, 23.5, 22.5, 21.5$, and 20.5.
(Figure taken from Bruzual and Kron 1980.)

star formation in the different galaxy types. Figure 4 gives, for model A, the distributions of galaxy redshifts predicted at different apparent magnitude levels. The ordinate is the (logarithm of the) number of galaxies per unit redshift interval, and the five curves from top to bottom represent intervals of one magnitude centered on $(J+F)/2 = 24.5$, 23.5, 22.5, 21.5, and 20.5.

What the Figure tells us is that as we go to fainter magnitudes, galaxies of an increasingly large redshift range, and of those an increasingly large fraction of high-redshift galaxies contribute to the counts at each magnitude. According to model A, then, a substantial number of galaxies in Kron's deep survey have been detected at redshifts $z \gtrsim 1.5$. Since the shapes of the redshift distributions of Figure 4 are sensitive to galaxy evolutionary effects, the reality of the assumed underlying histories of the star formation rate can be tested, in principle, by observing the redshifts of a small random subsample of Kron's faint galaxies and examining the corresponding shape(s) of their distribution(s).

This is how we are led back to the determination of redshifts. In order to complete the task as sketched in the first section, one cannot however do with only a few of them. In fact, measuring two or more colors and constructing two-color diagrams for faint galaxies may prove a more practicable solution.

In a two-color diagram, for each galaxy type the theoretical locus of variable redshift is always a well-defined line - as opposed to a scatter diagram -, and the individual loci for different galaxy types are separate. As a result, for the variety of galaxy types there exists in a two-color diagram a similar locus at each redshift, and these individual loci of constant redshift are separate as well, the extent to which they are spread across a two-color plane essentially depending on the choice of the photometric passbands involved in the color measurements.

As an example, Figure 5 shows the two-color diagrams constructed from Kron's (1980b) photographic J,F-system supplemented by an ultraviolet (U) and a near-infrared (N) passband. Constant-redshift loci were computed from Bruzual's (1981) evolutionary galaxy models described above. Note that

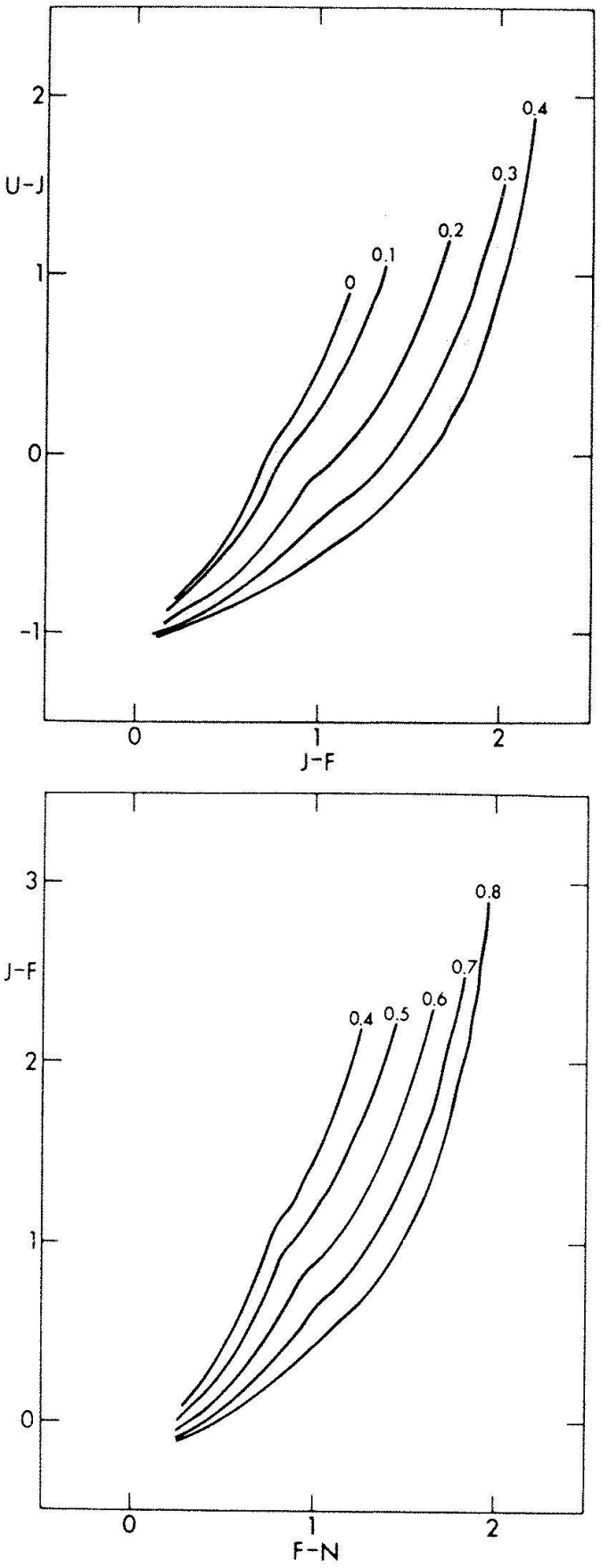



*Fig. 5.* Two-color diagrams for Kron's UJFN photographic broad-band system. Curves are loci of constant redshift, as labeled, computed for evolving galaxy models with present-day colors matching the range observed for bright galaxies.
(Figure taken from Kron 1980b.)

the two diagrams display useful sensitivities to different redshift ranges; this is due mainly to the existence in galaxy spectra of a break in continuum slope around 4000 Å, which is shifted into the redder passbands at higher values of z (see Figure 1).

In principle, then, the general properties of two-color diagrams offer a powerful tool for fast determination of types and redshifts for large numbers of galaxies. It is now largely left to the practice of multicolor photometry to keep this promise of the principle.

## Acknowledgements

This review could not have been written without the inspiring efforts of Richard Kron and Gustavo Bruzual, who have been generous in sharing with me many of their ideas on many occasions. I am also grateful to the Swiss National Science Foundation for financial support.

## References

Bruzual, A.G., 1981, Ph.D. thesis, University of California, Berkeley.
— , Kron, R.G. 1980, Astrophys. J. 241, 25.
Kron, R.G., 1980a, Astrophys. J. Suppl. 43, 305.
— 1980b, in: Two Dimensional Photometry, P. Crane, K. Kjär (eds), ESO Workshop Proceedings, Noordwijkerhout, p.349.
Salpeter, E.E., 1955, Astrophys. J. 121, 161.
Sandage, A.R., 1961, Astrophys. J. 134, 916.
Tinsley, B.M., 1980, Astrophys. J. 241, 41.

*Address of the author:*

PD Dr. Roland Buser
Astronomical Institute
University of Basle
CH-4102 Binningen BL (Switzerland)

# Evidence for the Big Bang

Gustav A. Tammann

## Cosmic Evolution

Hubble's discovery in 1929 of the linear distance-redshift relation of external galaxies was widely accepted as proof for an expanding universe, which must have begun in a singularity (Big Bang). However, the postulation in 1948 of a steady-state universe by H. Bondi, T. Gold, and F. Hoyle suggested that the Big Bang could be avoided: even an expanding universe can in principle be infinitely old and expand forever, provided that mass is continuously created. In this way the steady-state universe fulfills what has been called the 'perfect cosmological principle', viz. it appears the same to any fundamental observer in the universe at *any* time. Only during the last 20 years overwhelming evidence has become available that the universe as a whole is in fact evolving, thus discriminating the steady-state theory and restoring the faith in a Big Bang. This evidence is indeed a vivid illustration of the enormous progress observational cosmology has made during the last decades.

While in the following the emphasis lies on the presently available evidence for an expanding and cooling universe, i.e. for a hot Big Bang, it must be remembered that there is in addition an increasing set of data, which require cosmic evolution with time. An outstanding example for this is the increasing (co-moving) space density of quasars out to redshifts of $z \approx 2$ (Schmidt and Green 1980, 1981) and the cutoff of quasars at higher redshifts (Sandage 1972). This cutoff at $z \approx 4$ is most probably not a selection bias, but it seems now to be a real effect (Osmer 1982). A cutoff is also required by the fact that quasars brighter than $m_B \approx 20^m$ account already for a large fraction of the X-ray background (Setti and Woltjer 1979; Setti 1981).

Another striking result is the narrow colour range of first-ranked (elliptical) cluster galaxies, after they are reduced to a fiducial redshift $z = \text{const}$. Because elliptical galaxies must become redder with time due to the evolution of their individual stars, the near constancy of their colours can only be explained if they were formed during a relatively short epoch (Sandage 1973; cf. Fig. 1). A simple geometrical consideration shows that the number of extragalactic objects brighter than m, N(m), increases for fainter magnitudes as

$$\log N(m) \propto 0.6\, m\,. \qquad (1)$$

(If, for simplicity, radio fluxes are expressed in magnitudes, the relation holds here also for radio sources.) The redshift effect (K-correction) can only flatten the relation in eq. 1. But for several objects steeper relations are observed, which can only mean that these objects were brighter and/or more frequent at larger distances, i.e. at earlier epochs. Such steep count rates are observed for quasars (e.g. Green and Schmidt 1978), X-ray sources (Schwartz 1980), high-flux density radio sources (e.g. Longair 1978), as well as normal galaxies at faint levels (Kron 1980a, 1980b). It should be remarked here that the discussion of the slope of the count rates at large distances is, of course, in addition to the K-correction further complicated by the problem of the exact volume sampled, which depends on the (unknown) space curvature, or – what is essentially the same – on the deceleration parameter $q_0$ (cf. Fig. 2).

In addition to first-ranked cluster galaxies, which clearly have secular colour changes for variable z, also other galaxies are suspected to have surprisingly strong colour evolution. This is quite well established for radio galaxies, which are very blue at redshifts of $0.4 \lesssim z \lesssim 0.7$ (van der Laan, Katgert and de Ruiter 1980; van der Laan 1981). The suggested increasing blueness at large redshifts

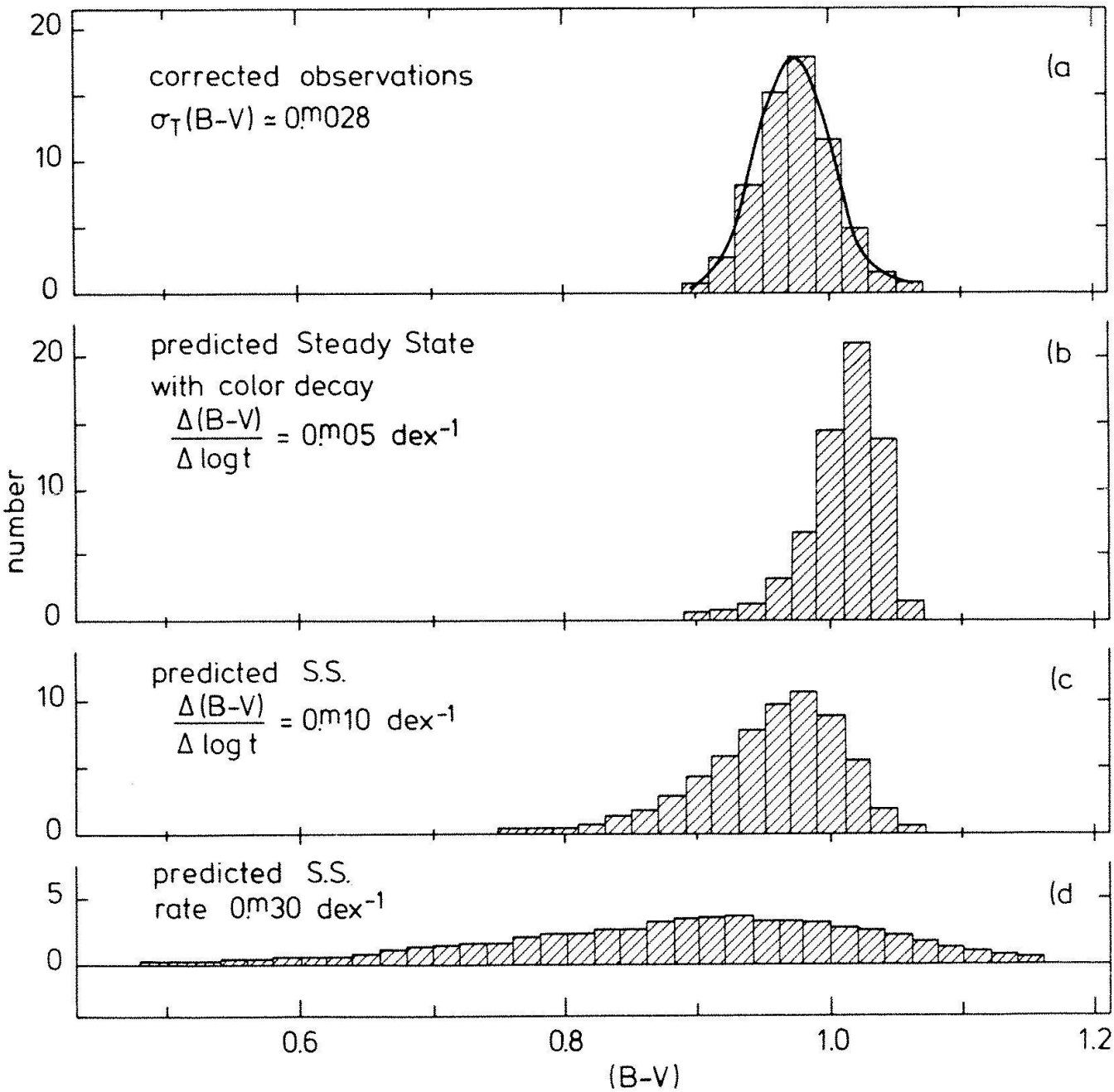



*Fig. 1.* The observed colour distribution of first-ranked cluster galaxies (upper panel). Note the narrow range in (B-V)-colours. The lower panels show the expected colour distribution, if galaxies were formed with constant rate (steady-state situation). Three different cases of colour evolution are assumed. Considering that the assumed colour evolution is improbably small in case b), first-ranked cluster galaxies must have had a preferred formation epoch in the past. (Adapted from Sandage 1973.)

($z > 0.2$–$0.3$) of field galaxies (Turner 1980) and of some cluster galaxies (Butcher, Oemler, and Wells 1980; Spinrad 1980; Oke 1980) may still need further confirmation (cf. also Buser 1981).

These short and incomplete remarks on evolutionary effects in the universe already show that it cannot be questioned anymore, that the universe has not always been the same, and that it is in a constant state of evolution.

However, it is the present goal not only to demonstrate that the universe reveals several aging effects, but rather that it partakes of a general expansion since a singularity in the past. The overwhelming evidence for this expansion rests mainly on four independent arguments, which shall be discussed here in turn: (1) the redshift-magnitude relation of standard candles, (2) the cosmic microwave background radiation (CMB), (3) the nucleo-

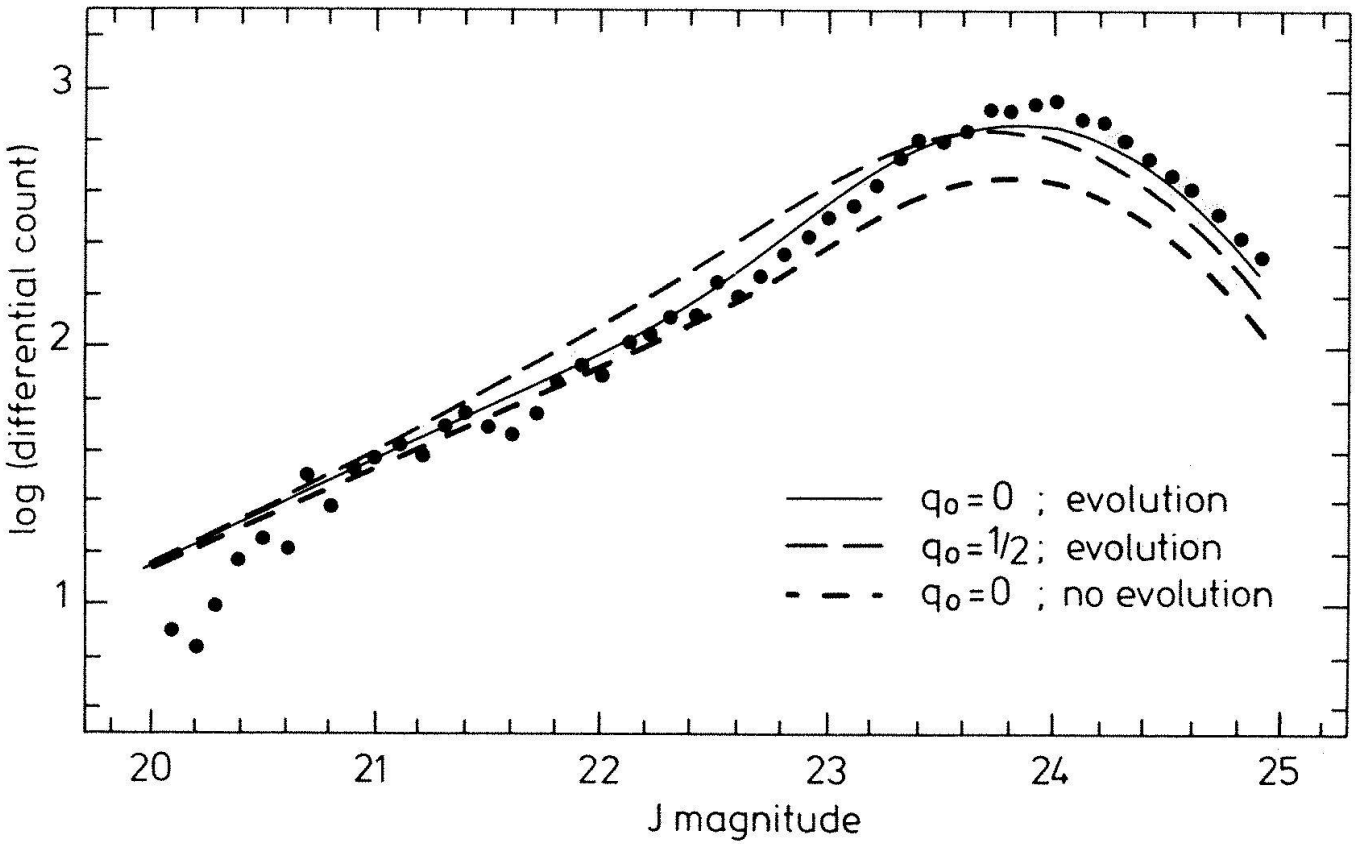



*Fig. 2.* The counts of galaxies (on a logarithmic scale) at very faint (J-) magnitudes. It is clear that the models with no evolution give poor fits to the counts, – whatever (reasonable) value of the deceleration parameter $q_0$ is assumed. (Adapted from Kron, 1980a.)

synthesis of the lightest elements, and (4) the agreement of the time scales of the expansion and of the oldest objects in our Galaxy.

## The Redshift-Magnitude Relation

It is well known that in the Friedmann model (with zero cosmological constant), which is the most simple and therefore exceptionally attractive model of an expanding universe, the apparent bolometric magnitude $m_{bol}$ of an object with absolute bolometric magnitude $M_{bol}$ is a function only of

$$m_{bol} = f(M_{bol}, H_0, z, q_0). \qquad (2)$$

$H_0$ is the present value of the Hubble constant measuring the expansion rate. Generally $H = \dot{R}/R$, where R is the cosmic scale factor. The redshift z is defined by $z = \lambda/\lambda_0$, where $\lambda_0$ is the rest wavelength of a given spectral line. The present value of the deceleration parameter is defined by $q_0 = \ddot{R}_0/R_0 H_0^2$; it measures the total deceleration of the expansion due to matter and energy in the universe. In a matter-dominated universe $q_0$ is related to the mean mass density $\rho_0$ in a simple way:

$$q_0 = \frac{4\pi G \rho_0}{3 H_0^2} \qquad (3)$$

Also the space curvature is determined by $q_0$: if $0 \leq q_0 \leq 1/2$ space is negatively curved (the gravitational binding energy in a representative volume is smaller than its kinetic expansion energy, and the universe will expand forever); if $q_0 = 1/2$ space is flat (the expansion of the universe will stop only in the infinite future); and if $q_0 > 1/2$ space is positively curved (the potential energy exceeds the kinetic energy, and the universe will eventually collaps). (For a rigorous derivation of these correlations see Sandage 1961).

For standard candles with $M = \text{const}$ and small redshifts ($cz \ll 1$), which are observed in heterochromatic magnitudes, – for in-

stance in red magnitudes $m_R$, - equation 2 becomes:

$$m_R = 5 \log cz + const. \quad (4)$$

Here the constant term contains (besides of $M_R$ and $H_0$) the K-correction, which allows for the photometric redshift effect of stretching the energy distribution curve of the emitter and shifting it through the respective filter band, - in the present example through the R-band (cf. Sandage 1975b).

In Figure 3 the log cz - versus - $m_R$ relation ('Hubble diagram') for the brightest galaxy in various clusters of galaxies is shown. The small scatter ($\sigma = 0^m3$) in the diagram proves that these first-ranked cluster galaxies are indeed good standard candles. The most remarkable fact, however, is that these galaxies define a line of slope 5, as predicted by the theory in equation 4. The implied linear relation between recession velocity and distance has two fundamental properties: (1) it requires a singularity in the past (Big Bang), and (2) a linearly expanding universe is the only one which can accomodate an unlimited number of fundamental observers having the same aspect of the universe. With other words: all non-linearly expanding universes would violate the Copernican Principle.

The most distant first-ranked cluster galaxies in Figure 3 with $z \approx 0.5$ (corresponding very roughly to recession velocities of 120,000 km $s^{-1}$) deviate somewhat from the straight line with slope 5. This is expected, because the effect of $q_0$ is neglected in equation 4: looking very far out in space (and back in time) one sees galaxies, which had little time to be decelerated and which lie therefore too high in the Hubble diagram. One has hoped for a long time that from the exact position of very distant galaxies in the Hubble diagram one could just determine the value of $q_0$.

One knows now that this hope is unrealistic, because first-ranked cluster galaxies are not perfect standard candles over cosmic epochs: they do suffer luminosity evolution, as discussed in Section 1, and this effect is accentuated by dynamical interactions within their respective clusters.

While first-ranked cluster galaxies establish the linear expansion out to the largest redshifts, less luminous, but locally more numerous standard candles are important to put limits on any deviations from an ideal Hubble flow. Such additional standard candles are brightest galaxies in small clusters and groups of galaxies (Sandage 1975a), the mean luminosity of the 10 or 5 brightest cluster galaxies (Weedman 1976), the maximum luminosity of supernovae of type I in elliptical galaxies (where they do not suffer absorption in their parent galaxies) (Tammann 1982; Fig. 4), as well as very local groups of galaxies with known individual distances (Tammann, Sandage and Yahil 1980). These various data show that the deviations from a pure Hubble flow are surprisingly small on all scales: the random velocities in the line of sight of field galaxies lie at the detection limit, which means for nearby galaxies $\lesssim 100$ km $s^{-1}$, and the upper limit for any streaming velocity our Galaxy is involved in (probably together with a very large volume of supercluster size) is 600 km $s^{-1}$. The latter value is derived from a small dipole anisotropy of the cosmic background radiation (Smoot 1980; Boughn, Cheng and Wilkinson 1981; cf. Section 3). In view of the large density fluctuations, which in fact are strong enough to have built up gravitationally bound clusters within an expanding universe, it is one of the most surprising facts of cosmology that the (density enhanced) peculiar motions of field galaxies are so small.

Against overwhelming evidence it is still occasionally put forward that the observed extragalactic redshifts are due to other effects than the Doppler effect. To the extent that these alternative interpretations involve 'non-physical' photon propagation, it should be stressed that in addition to the redshift-magnitude relation of standard candles there exists a redshift-angular diameter relation of standard rods. The latter relation for the diameters of first-ranked cluster galaxies (Fig. 5) and for whole clusters of galaxies (Fig. 6) independently support the linear expansion of the universe. The reader is referred to the original literature (Sandage 1961; 1975b) for the correct definition of isophotal diameters; it is these diameters which must be used for galaxies because their metric diameters are ill defined. In the case of galaxy clusters it is in principle possible to measure metric diameters, but

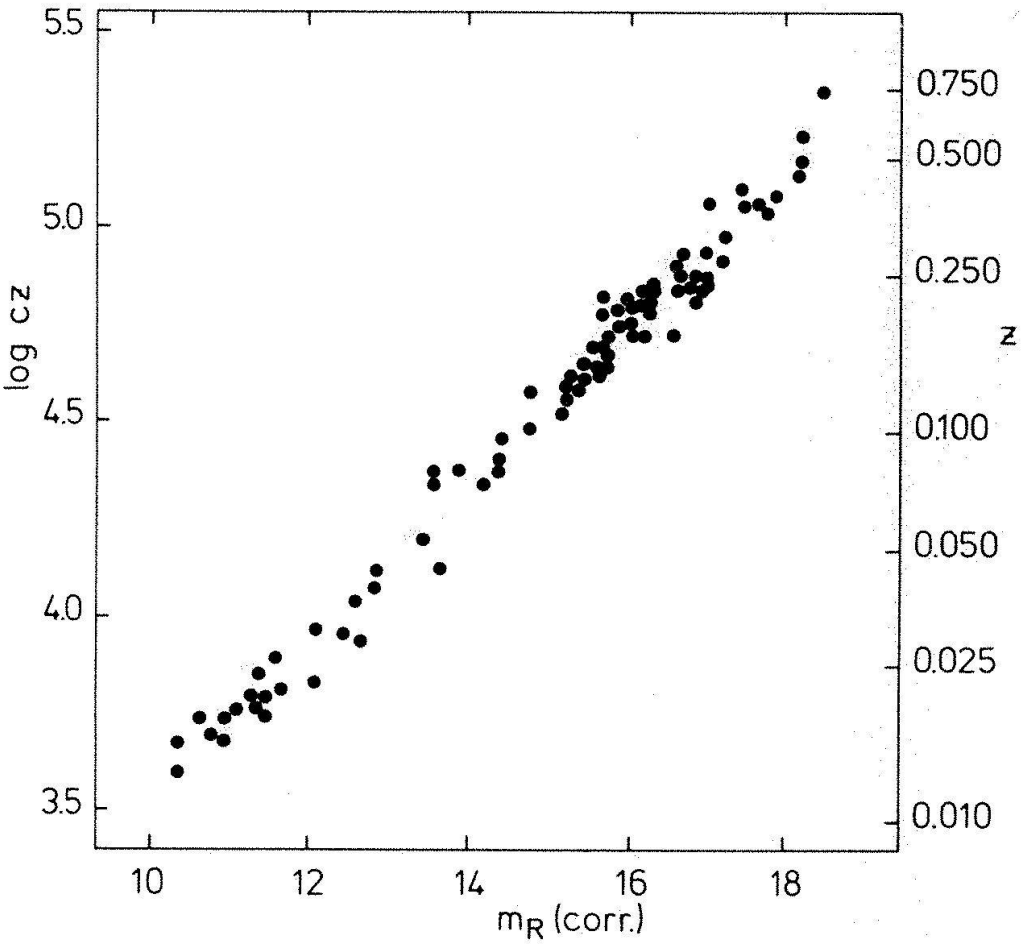



*Fig. 3.* The Hubble diagram of first-ranked cluster galaxies (adapted from Kristian, Sandage and Westphal 1978). The red magnitudes of the abscissa are corrected for galactic absorption, the K-correction, for the richness of the parent cluster and its Bautz-Morgan class. The Bautz-Morgan class is determined by the relative brightness of the first-ranked galaxy in comparison to that of the other cluster galaxies.

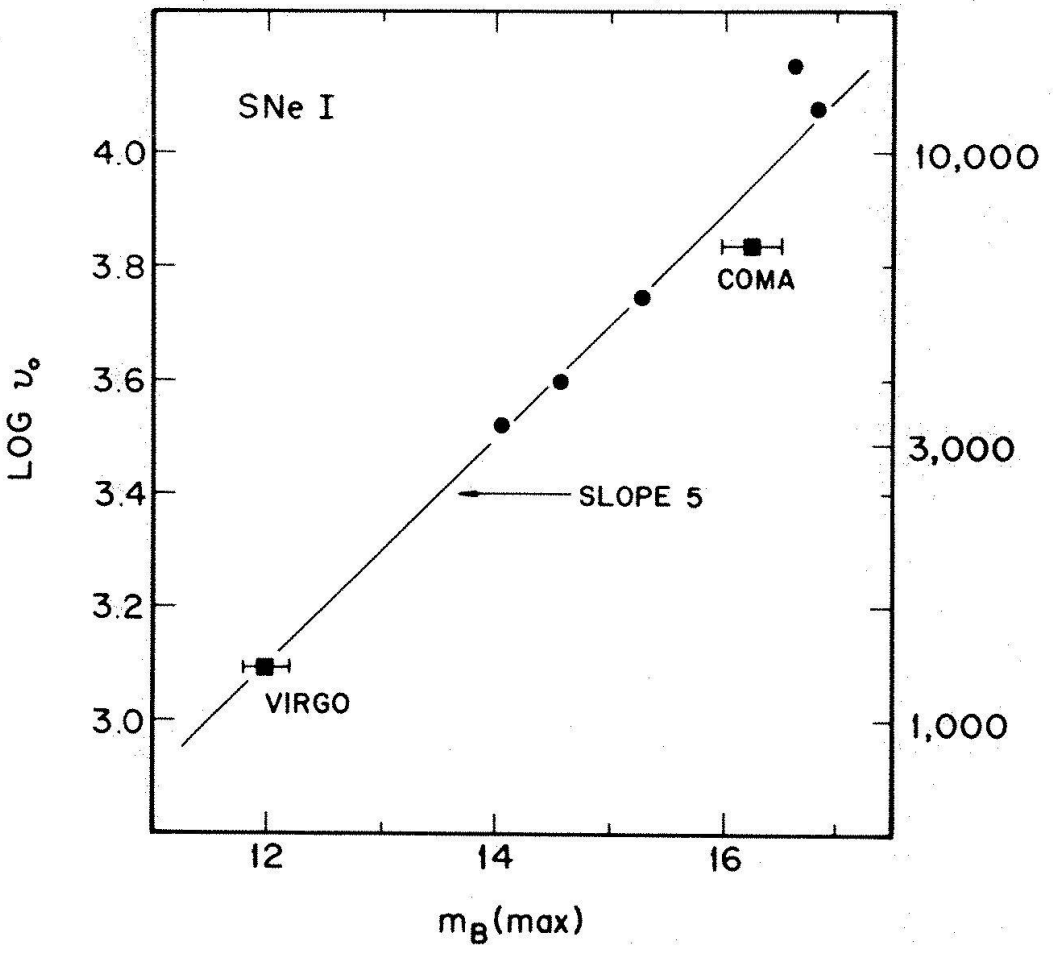



*Fig. 4.* The Hubble diagram of supernovae of type I in absorption – free elliptical galaxies at maximum blue light. The square for the Virgo cluster is the mean of 6 SNe, that for the Coma cluster is the mean of 5 SNe.

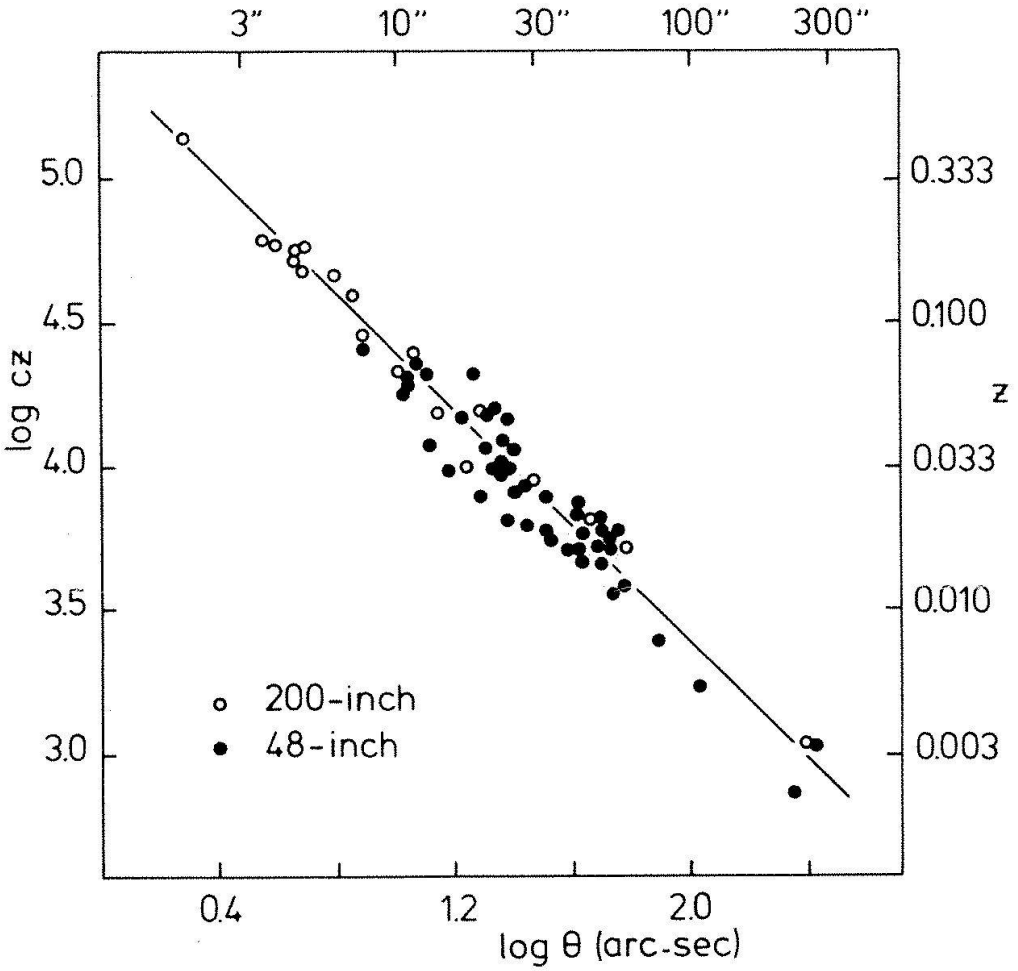



*Fig. 5.* The redshift-diameter relation of first-ranked cluster galaxies (adapted from Sandage 1972a). The abscissa shows the logarithm of the angular isophotal diameter. The different symbols refer to different telescopes.

their unambiguous, distance-independent definition poses also formidable difficulties.

## The Cosmic Microwave Background Radiation (CMB)

The ability of a theory to make predictions, which are subsequently validated by observations, is considered to be the ultimate scientific test of any theory. It is therefore of the highest significance that the CMB radiation, first detected by Penzias and Wilson in 1965, had been predicted already in 1946 and subsequent years by Gamov and his collaborators for an expanding, hot Big Bang universe.

At early epochs the universe was opaque, and particles and photons had the same (high) temperature. About 650,000 years after the Big Bang (at redshifts of $z \approx 1100$ and a temperature of $T \approx 3000$ K) the universe had sufficiently cooled that protons could 're'-combine with free electrons to form neutral hydrogen, and the universe became transparent. From then on photons cooled adiabatically with the expansion of the universe, such that their present temperature is [note that $T_0 \propto R/R_0 = (z+1)$; cf. Sunyaev and Zel'dovich 1980]:

$$T_0 = T_{z=1100}/1101 = 2.7\ \mathrm{K}. \qquad (5)$$

Indeed the best determinations of the blackbody temperature give 2.7–3.0 K (e.g. Wilkinson 1980; Richards 1980; Weiss 1980).

All attempts to explain alternatively the CMB by large numbers of unresolved objects have failed mainly because of three reasons:

(1) The observed spectrum of the CMB is nearly perfectly Planckian (Weiss 1980). No individual known objects, even with redshift effects taken into account, can match such a spectrum;

(2) At microwave wavelengths, i.e. $\lambda = 0.1$–$6$ cm, the temperature fluctuations over angular scales of 2′ to 2° are less than $\Delta T/T = 10^{-3}$ to $10^{-5}$ (Boynton 1978; Partridge 1980). To account for this degree of isotropy it would require an excessive num-

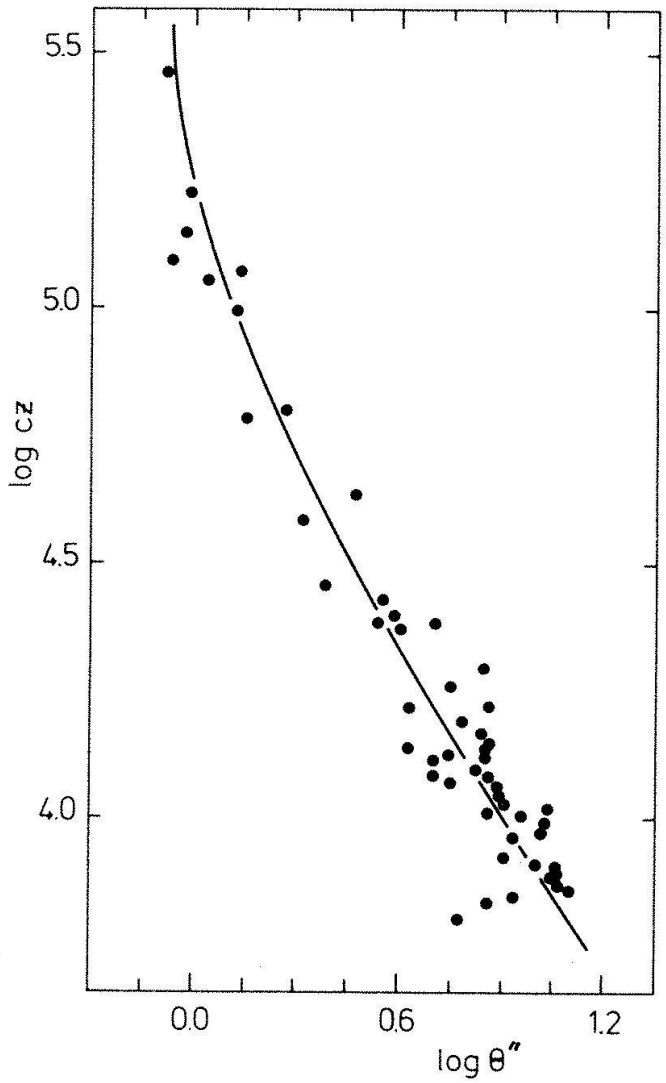



*Fig. 6.* The redshift-diameter relation for galaxy clusters (adapted from Bruzual and Spinrad 1978). Because of the large distances reached, deviations from an intuitively expected straight line become apparent. The full-drawn line is calculated for a linearly expanding universe and an assumed value of $q_0 = 0.5$. The fit, however, does not give a reliable determination of $q_0$ because of conceivable evolutionary diameter effects of clusters, due to their dynamically evolution.

ber of individual sources (cf. Longair 1978); and

(3) The number density of CMB photons is $\sim 400$ cm$^{-3}$, which corresponds to an energy density of $\sim 0.25$ eV cm$^{-3}$. This is at least 95% of the present total mean energy density in form of radiation (Longair 1978). In spite of its low temperature the energy content of the CMB radiation is therefore very high.

While the CMB has widely been accepted as the final proof for a hot Big Bang universe, it should be added that Rees (1980) has argued that a different origin of the CMB cannot totally be excluded: if a hypothetical first generation of stars had formed at a very early epoch ($z > 600$) and generated enough energy, then their radiation could have been thermalized by a pregalactic medium, of which one can possibly postulate that it were opaque at epochs $z > 100$ due to dust and molecules (cf. also Woltjer 1981).

## The Nucleosynthesis of the Lightest Elements

Wherever the $^4$Helium abundance can be measured, i.e. in young and old stars, in the interstellar medium of our Galaxy, in the Magellanic Clouds and other galaxies, and even in metal-poor galaxies, one finds that its fractional abundance (in mass) is *at least* $Y = 0.24 \pm 0.02$ (Peimbert and Torres-Peimbert 1976; Greenstein 1980). This is a highly significant result which can hardly be explained by the assumption that this large amount of $^4$He were processed in stars from hydrogen. It can be shown, that if this were the case the energy released would result in higher galaxian surface brightnesses than actually observed (Burbidge 1958; Hoyle and Tayler 1964).

The only way out of this dilemma is to assume that the $^4$He abundance is primordial, which is also supported by other observational arguments (Greenstein 1980). Upon detailed calculations of the early phases of a standard hot Big Bang model, it turns out that the production of $^4$He with $Y \approx 0.24$ is a nearly unavoidable consequence (cf. Weinberg 1972). The quantitative agreement of the calculated and the observed value of Y gives strong support to the adopted model.

The $^4$He argument is further strengthened by the light isotopes $^2$D, $^3$He, and $^7$Li (David and Reeves 1978; Audouze 1981). In particular it came as a great surprise that deuterium, which is effectively distructed in stars, is among the 12 most frequent isotopes in the interstellar medium (cf. Laurent, Vidal-Madjar and York 1979). Its origin must be primordial, and it turns out that it is indeed easily produced in the Big Bang model, as well as $^3$He and $^7$Li.

The nucleosynthesis of the light elements occurred about 220 sec after the Big Bang at a temperature of $T \approx 9 \cdot 10^8$ K. While the yield of $^4$He is quite insensitive to other

parameters, sufficient amounts of the other light isotopes were produced only, if the mean baryonic mass density was relatively low. The allowed range of the baryonic density at that epoch ($z = 3.6 \cdot 10^8$) can be fixed to within a factor of $\sim 1.5$; if this density is scaled down to the present epoch ($z = 0$) one finds a present baryonic density of $\rho_{b,0} = (1.9\text{–}3.8) \cdot 10^{-31}$ g cm$^{-3}$ (Yang, Schramm, Steigman, and Rood 1979), which gives for the deceleration parameter through equation 3:

$$0.02 < q_0 < 0.04 .$$

This, and the above-mentioned small peculiar motions of field galaxies (Yahil, Sandage, and Tammann 1978) are presently the strongest arguments for an open universe. If the universe should actually be closed, the closing matter is probably not in baryons, and it is not clumped like the visible galaxies. It has been speculated therefore that the closing matter consists of neutrinos with non-zero mass ($m_\nu \approx 30$ eV). In that case it remains to be explained, however, why at most a fraction of all neutrinos cluster like galaxies.

## The Time Scale of the Universe

If the present expansion rate, i.e. the Hubble constant $H_0$, is known as well as the deceleration parameter $q_0$, it is clear that the expansion age of the universe (the 'Friedmann time') can be determined. Obviously this age must be larger than the age of the oldest known objects.

The oldest objects, ever found in our Galaxy, are the globular clusters. Their age is conventionally determined from the absolute magnitude of the stars, which are just leaving the main-sequence (cf. Demarque and McClure 1977). The enormous observational difficulty of the method is, that this turn-off point lies at very faint magnitudes, and that unavoidable errors in the magnitudes and colours (!) have a large effect on the age. Moreover the result depends on the position of the fiducial main-sequence, which is gouverned by the adopted He abundance, and the (obviously variable) He content cannot be directly observed for the cool globular cluster stars.

In view of this difficulty a recent important progress must be mentioned: it is possible to fit the colour-magnitude diagrams of globular clusters by means of their (relatively bright!) horizontal-giant branches. The zero-point calibration of the luminosities is provided by the now understood period-luminosity relation of RR Lyrae variables (Sandage 1982). The resulting colour (temperature)-luminosity diagrams of individual globular clusters can then be fitted to existing model calculations of stellar evolution (Iben and Rood 1970; Demarque and McClure 1977), allowing for the known differences in metal content and almost independently of the He content. The result is for all globular clusters a uniform age of $(17 \pm 2) \cdot 10^9$ years (Sandage 1981). The near equality of the ages of globular clusters, even for those with widely different metal content, is in perfect agreement with the expectation, that the globular cluster system was formed during the relatively short collaps time of our early Galaxy.

The ages derived from evolutionary stellar models depend sensitively on the assumed constancy of the gravitational constant G. It is therefore important that the G-independent ages of the radioactive elements (particularly for the $^{187}Re/^{187}Os$ ratio) yield a time since the beginning of stellar nucleosynthesis of $(13\text{–}22) \cdot 10^9$ years (Luck, Birck and Allegre 1980), – where the relatively wide error range reflects the uncertainty of the necessary assumptions on the stellar birth rate function in our Galaxy.

The minimum age of the universe from two totally different methods is therefore probably $\geq 17 \cdot 10^9$ years, and almost certainly $> 15 \cdot 10^9$ years. How does this compare with the expansion age?

The Hubble constant $H_0$ has been determined to be 50 km s$^{-1}$ Mpc$^{-1}$ (Sandage and Tammann 1976). It has since been pointed out (Tammann, Yahil, and Sandage 1979) that some assumptions going into this determination (viz. mainly the reliability of the morphological luminosity classification of spiral galaxies) are not as rigid as originally assumed. On the other hand, values as high as $H_0 = 100$ have occasionally been proposed. However, these very high values

depend either on distance indicators, whose usefulness cannot be proven independently, or are dominated by selection effects, which are known to stellar astronomers as 'Malmquist bias'. This bias makes that distant objects must be intrinsically brighter (or larger) than their nearby counterparts in order to enter the same magnitude-limited (or angular-diameter-limited) catalogue; it always leads to an underestimate of large distances and hence to an overestimate of $H_0$.

A new way to $H_0$ may resolve the discrepancy. Through Cepheid variables as primary distance indicators one knows the absolute magnitude of the brightest red supergiants in a number of galaxies. This absolute magnitude is well defined, because these stars have a sharp upper luminosity cutoff, and it is a very stable distance indicator, as judged from the widely different calibrating galaxies. Using these red supergiants as distance indicators one can obtain distances to a number of galaxies, including IC 4182 and NGC 4214. The latter two galaxies have produced each a well observed supernova of type I, from which one can derive a mean absolute magnitude at maximum light of $M_B(\max) = -19^m65$ (Sandage and Tammann 1981).

In Figure 4 it was shown already that supernovae of type I are very good standard candles (banning also any appreciable effect of the Malmquist bias). If one assumes $H_0 = 100$, the straight line in Fig. 4 is best represented by

$$M_B(\max) = -18^m22 + 5 \log H_0/100 . \qquad (6)$$

Inserting into this equation the above value of $-19^m65$ gives

$$H_0 = 52 \pm \sim 10 \text{ km s}^{-1} \text{ Mpc}^{-1}.$$

This result for $H_0$ must represent the truly cosmic value, beyond any local peculiar motions, because equation 6 is defined by supernovae out to recession velocities of $\sim 10000$ km s$^{-1}$ (Sandage and Tammann 1981).

In the absence of any deceleration ($q_0 = 0$) the expansion age is $H_0^{-1} = (19 \pm 4) \cdot 10^9$ years. This age is decreased by any amount of deceleration (for a tabulation of the Friedmann time in function of $q_0$ see Sandage 1961a). With $q_0 = 0.05$, which is indicated by the baryonic matter in the universe (cf. Section 4), the Friedmann time becomes $(16 \pm 3) \cdot 10^9$ years. If there should be still enough, yet undiscovered matter to close the universe ($q_0 = 1/2$), the Friedmann time is reduced to $(13 \pm 3) \cdot 10^9$ years.

In any case the near agreement of the minimum age of the universe and its expansion age is stunning, and it is hard to see how this could only be an accident.

In this short presentation only Friedmann cosmologies were considered. Many other fancy and more complicated cosmologies have been put forward. But at present just the simplicity of the Friedmann Big Bang models and the fact that they are in perfect agreement with all available observations, makes them exceptionally attractive.

The author thanks the Swiss National Science Foundation for continued support, which only has made possible some of the work reported here.

## References

Audouze, J. 1981, Vatican Study Week on Cosmology and Fundamental Physics, in press.
Boughn, S.P., Cheng, E.S., and Wilkinson, D.T. 1981, Ap. J. Letters 243, L 113.
Boynton, P.E. 1978, I.A.U. Sypm. 79, 317.
Bruzual, G., Spinrad, H., 1978, Ap. J. 220, 1; 222, 1119.
Burbidge, G., 1958, Publ. Astron. Soc. Pacific 70, 83.
Buser, R. 1981, this symposium.
Butcher, H., Oemler, A., and Wells, D. 1980, I.A.U. Symp. 92, 49.
David, Y., and Reeves, H. 1978, in: Physical Cosmology, ed. R. Balian, J. Audouze, and D.N. Schramm, Amsterdam: North-Holland, p. 443.
Demarque, P., and McClure, R.D. 1977, in: The Evolution of Galaxies and Stellar Populations, ed. B.M. Tinsley and R.B. Larson, New Haven: Yale University Observatory, p. 199.
Green, R.F., and Schmidt, M. 1978, Ap. J. Lett. 220, L 1.
Greenstein, J.L. 1980, Physica Scripta 21, 759.
Hoyle, F., and Tayler, R.J. 1964, Nature 203, 1108.
Kristian, J., Sandage, A., and Westphal, J.A. 1978, Ap. J. 221, 383.
Kron, R.G. 1980a, Physica Scripta 21, 652.
Kron, R.G. 1980b, I.A.U. Symp. 92, 9.
Laan, H. van der, 1981, Vatican Study Week on Cosmology and Fundamental Physics, in press.
Laan, H. van der, Katgert, P., and Ruiter, H.R. de 1980, Physica Scripta 21, 669.
Laurent, C., Vidal-Madjar, A., and York, D.G. 1979, Ap. J. 229, 923.
Longair, M.S. 1978, in: Observational Cosmology, ed. A. Maeder, L. Martinet, and G. Tammann, Geneva: Observatory, p. 125.

Luck, J.-M., Birk, J.-L., and Allegre, C.-J. 1980, Nature 283, 256.
Oke, J.B. 1980, Scientific Research with the Space Telescope, ed. M.S. Longair and J.W. Warner, Washington: NASA, p.309 (=I.A.U. Coll. No.54).
Osmer, P.S. 1982, Ap. J. 253, 28.
Partridge, R.B. 1980, Physica Scripta 21, 624.
Peimbert, M., and Torres-Peimbert, S. 1976, Ap. J. 203, 581.
Rees, M. 1980, in: Physical Cosmology, ed. R. Balian, J. Audouze, and D.N. Schramm, Amsterdam: North-Holland p.615.
Richards, P.L. 1980, Physica Scripta 21, 610.
Sandage, A. 1961, Ap. J. 133, 355.
Sandage, A. 1961a, Ap. J. 162, 841.
Sandage, A. 1972, Quart. J.R.A.S. 13, 282.
Sandage, A. 1972a, Ap. J. 173, 485.
Sandage, A. 1973, Ap. J. 183, 711.
Sandage, A. 1975a, Ap. J. 202, 563.
Sandage, A. 1975b, in: Galaxies and the Universe, ed. A. and M. Sandage, J. Kristian, Chicago: University of Chicago Press, p.761.
Sandage, A. 1982, Ap. J. 252, 553.
Sandage, A., and Tammann, G.A. 1976, Ap. J. 210, 7.
Sandage, A., and Tammann, G.A. 1981, in: Vatican Study Week on Cosmology and Fundamental Physics, in press.
Schmidt, M., and Green, R.F. 1980, I.A.U. Symp. 92, 73.
Schmidt, M., and Green, R.F. 1981, in: Vatican Study Week on Cosmology and Fundamental Physics, in press.
Schwartz, D.A. 1980, Physica Scripta 21, 644.
Setti, G. 1981, in: Vatican Study Week on Cosmology and Fundamental Physics, in press.
Setti, G., and Woltjer, L. 1979, Astron. Astrophys. 76, L1.
Smoot, G.F. 1980, I.A.U. Symp. 92, 489.
Spinrad, H. 1980, I.A.U. Symp. 92, 39.
Sunyaev, R.A., and Zel'dovich, Ya.B. 1980, Ann. Rev. Astron. Astrophys. 18, 537.
Tammann, G.A. 1982, in: Supernovae: A Survey of Current Research, ed. M.J. Rees and R.J. Stoneham, Dordrecht: D. Reidel, p.371.
Tammann, G.A., Sandage, A., Yahil, A. 1980, Physica Scripta 21, 630.
Tammann, G.A., Yahil, A., and Sandage, A. 1979, Ap. J. 234, 775.
Turner, E.L. 1980, I.A.U. Symp. 92, 71.
Weidman, D.W. 1976, Ap. J. 203, 6.
Weiss, R. 1980, Ann. Rev. Astron. Astrophys. 18, 489.
Weinberg, S. 1972, Gravitation and Cosmology, New York: John Wiley and Sons, p.545.
Wilkinson, D.T. 1980, Physica Scripta 21, 606.
Woltjer, L. 1981, Vatican Study Week on Cosmology and Fundamental Physics, in press.
Yahil, A., Sandage, A., and Tammann, G.A. 1978, Physical Cosmology, ed. R. Balian, J. Audouze, and D.N. Schramm, Amsterdam: North-Holland, p.127.
Yang, J., Schramm, D.N., Steigman, G., and Rood, R.T. 1979, Ap. J. 227, 697.

*Address of the author:*

Prof. Dr. Gustav A. Tammann
Astronomical Institute
University of Basle
CH-4102 Binningen (Switzerland)