Wenig beachtete Problemkreise in der phylogenetischen Analyse: Invariante Positionen und die Wahl von Substiutionsmodellen

Autor(en): **Haase, Martin / Misof, Bernhard**

Objekttyp: **Article**

Zeitschrift: **Contributions to Natural History : Scientific Papers from the Natural History Museum Bern**

Band (Jahr): **- (2003)**

Heft 2

PDF erstellt am: **27.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-786949

Wenig beachtete Problemkreise in der phylogenetischen Analyse: Invariante Positionen und die Wahl von Substitutionsmodellen.

Martin Haase und Bernhard Misof

Kurzfassung

Contrib. Nat. Hist. 2: 35–40.

Die Basenzusammensetzung eines Datensatzes hat oft entscheidende Auswirkungen auf die phylogenetische Rekonstruktion. Invariante Positionen können dabei einen grossen Einfluss haben. Das ist der Fall, wenn sie Heterogenität der Basenzusammensetzung der phylogenetisch informativen Positionen überdecken, also Homogenität vortäuschen. Die Basenzusammensetzung eines Datensatzes muss daher vor einer phylogenetischen Analyse ausschliesslich an den variablen Positionen untersucht werden. Die Wahl des einer Maximum-Likelihood-Analyse zugrundeliegenden Substitutionsmodells ist ebenfalls kritisch für die Stammbaumrekonstruktion und sollte nicht willkürlich erfolgen, da es die Topologie entscheidend mitbestimmt. Likelihood-Ratio-Tests ermöglichen die Bestimmung des auf einen Datensatz am besten passenden Modells.

Einleitung

Die Erfindung und Automatisierung der Polymerasekettenreaktion (PCR) hat auch die phylogenetische Forschung nachhaltig beeinflusst: molekulare Daten, insbesondere Sequenzdaten, sind zu einer wichtigen Grundlage in der Erforschung der Verwandtschaftsverhältnisse der Organismen und der Evolution von Merkmalen (im weitesten Sinne) geworden. Der Umfang dieser neuen Datenquelle zusammen mit dem rasanten Fortschritt, den die Computertechnologie vorlegt, hat auch zu einem enormen Aufschwung in der Entwicklung von teilweise sehr komplexen Analysemethoden geführt. Jede Methode, jedes Modell, das einer Methode zugrunde liegt, hat freilich seine Voraussetzungen und Limiten. Diese zu kennen und zu erkennen ist essentiell, die Möglichkeiten dazu sind aber oftmals nicht trivial. Die Gründe dafür liegen u.a. in der „Natur der Sache", also in unserer oft ungenügenden Kenntnis von Eigenschaften der DNS, ihrer Evolution und den Einschränkungen, der letztere unterliegt. Ander-

erseits sind Testverfahren, die die Voraussetzungen von phylogenetischen Analysemethoden hinterfragen, zumeist ähnlich komplex und rechentechnisch aufwändig wie die Methoden selber. In ihrer Entwicklung hinken derartige Tests den Rekonstruktionsmethoden naturgemäss meist hinterher. Zudem sind auch sie nicht frei von Voraussetzungen und Limiten, die ihrerseits exploriert werden müssen.

In der vorliegenden Arbeit wollen wir zwei Probleme und Fragen diskutieren, die erst in jüngerer Zeit erkannt worden sind bzw. für die erst seit kurzem Lösungsmöglichkeiten, auch in Form von Computerprogrammen, vorliegen. 1) Wie beeinflussen invariante Positionen der DNS, also Positionen, die aus funktionellen Gründen keine Veränderung erfahren können, meine Analyse? 2) Wie finde ich das meinen Daten zugrunde liegende Substitutionsmodell für eine Maximum-Likelihood-Analyse? Beiden Problemen sollte man sich grundsätzlich vor jeder phylogenetischen Analyse stellen, die auf molekularen Daten beruht, da sie entscheidend die Wahl der Rekonstruktionsmethoden beeinflussen.

Invariante Positionen

Primär scheint das Auftauchen von invarianten Positionen in molekularen Datensätzen kein grosses Problem zu sein. Offensichtlich haben invariante Merkmale keinen Einfluss auf Stammbaumrekonstruktionen und können geflissentlich ignoriert werden. Dies ist auch als Standard-Option in fast allen frei zugängigen Programmpaketen enthalten.

Konzentrieren wir uns auf die simple Parsimonie-Rekonstruktion, so stimmt es in der Tat, dass invariante Positionen keinen Einfluss auf das Rekonstruktionsverfahren besitzen. Wird aber versucht, mittels Resampling-Methoden (z.B. Bootstrapping) die Robustheit einer Rekonstruktion abzuschätzen, so stossen wir auf ein ernsthaftes Problem. Uninformative, also invariante und autapomorphe Merkmale terminaler Taxa beeinflussen zwar nicht das Resultat der Parsimonie-Rekonstruktion, also die Topologie des Baumes, aber erniedrigen die Trefferwahrscheinlichkeit für jede Position im Resampling-Verfahren. Diese ist abhängig von der Gesamtzahl der Merkmale und nicht von der Gesamtzahl der informativen Merkmale. Konsequenterweise ist ein Resampling-Verfahren also nur mit den parsimonie-informativen Merkmalen durchzuführen, uninformative Merkmale sind aus den Daten zuvor zu entfernen. Ein leicht zu behebendes Problem, das aber gerne übersehen wird (siehe z.B. Zharkikh & Li 1992).

Invariante Positionen halten allerdings auch weniger triviale Überraschungen bereit, die nicht allgemeinen Niederschlag in der speziellen Literatur ge-

funden haben. Wie erinnerlich, setzt sich die DNS aus den vier Grundbausteinen Adenin, Guanin, Cytosin und Thymidin zusammen. Es ist schon lange bekannt, dass die prozentuale Zusammensetzung aus diesen vier Grundbausteinen nicht ausgeglichen, etwa 25% Adenin, 25% Guanin, 25% Cytosin und 25% Thymidin, sein muss, sondern dass sehr wohl Abweichungen von dieser Verteilung in vielen DNS-Abschnitten bzw. spezifischen Genomen zu finden sind (vergleiche Simon & al. 1994) . Besonders die Zusammensetzung der mitochondrialen DNS ist mit ihrem hohen A- und T-Gehalt auffällig. Dieser AT-Gehalt kann in manchen Taxa, ein Beispiel wären etwa die Hymenopteren, bis zu rund 80% betragen (Simon & al. 1994).

Welchen Effekt hat dies auf phylogenetische Analysen? Die Einschränkung der DNS-Varianz auf vier Merkmalszustände hat zur Folge, dass DNS-Sequenzen auch zufällige Übereinstimmungen zeigen können. Je geringer die Anzahl der beobachtbaren Merkmalszustände ist, desto höher ist diese zufällig mögliche Übereinstimmung. Demzufolge werden Sequenzen mit z.B. unabhängig erworbenem hohem AT-Gehalt rein zufällig einen grösseren Ähnlichkeitsgrad aufweisen als Sequenzen mit balancierter Basenkomposition. Ist in verschiedenen evolutiven Linien eine Verschiebung der Basenkomposition konvergent erfolgt, so kann die dadurch erhöhte, zufällige Ähnlichkeit phylogenetisches Signal überdecken (z.B. Lockhart & al. 1994). Es gilt daher, solche möglichen Verschiebungen der Basenkomposition vor, bzw. unabhängig von einer phylogenetischen Analyse zu identifizieren, um mögliche Ursachen für widersprüchliche Signale im Datensatz auszumachen.

Konventionellerweise wird die mögliche Heterogenität der Basenkomposition eines Datensatzes mit einem χ^2-Test getestet. Dieser Test kann allerdings nicht phylogenetische Abhängigkeiten dokumentieren. Er kann nur helfen, ein generelles Problem im Datensatz zu identifizieren. Üblicherweise wird der χ^2-Test über den gesamten Datensatz ausgeführt, also über informative und uninformative Merkmale. Es ist offensichtlich, dass invariante Merkmale, also Merkmale, die aus funktionellen Zwängen keine Variabilität zeigen können, keine Heterogenität der Basenkomposition im Datensatz stützen können. Ganz im Gegenteil, ein χ^2-Test, der invariante Merkmale einschliesst, wird uns meist ein klares Signal für eine homogene Basenkomposition liefern. Häufig überwiegen invariante Merkmale in molekularen Datensätzen. Es ist daher kaum verwunderlich, dass Homogenität der Basenkomposition fast überall dokumentiert wurde. Man bedenke nun allerdings, dass einzig die informativen Merkmale die Topologie eines Stammbaumes bestimmen. Betrachtet man nur informative Merkmale, so kann eine Heterogenität der Basenkomposition durchaus vorkommen, etwa ausgelöst durch nicht zufällige Substitutionsereignisse in einzelnen Artengruppen, und diese Heterogenität kann entscheidenden Einfluss

auf die Rekonstruktion haben. Mit einem simplen χ^2-Test über alle Merkmale wäre das Problem nicht zu identifizieren. Informative und invariante Merkmale müssen in Tests auf Heterogenität der Basenkomposition getrennt behandelt werden (vergleiche Lockhart & al. 1996; Misof & al. 2001).

Wahl des geeigneten Substitutionsmodells in Maximum-Likelihood-Analysen

Die kurze Skizze über invariante Positionen und Heterogenität in der Basenkomposition führt uns unmittelbar zum Problem der Wahl eines geeigneten Substitutionsmodells in Maximum-Likelihood-Analysen. Im Gegensatz zur Parsimonie-Analyse eines molekularen Datensatzes basiert die Maximum-Likelihood-Analyse auf expliziten Substitutionsmodellannahmen. Bei gegebenem Substitutionsmodell wird die Topologie gesucht, die mit grösster kumulativer Wahrscheinlichkeit über alle Merkmale das gefundene Muster der Nukleotidvariationen darstellt. Es ist somit klar, dass die Rekonstruktion eines Stammbaumes vom angenommenen Substitutionsmodell abhängt. Ad hoc-Annahmen zu Substitutionsparametern des Substitutionsmodells waren noch bis vor kurzem üblich. Es war folglich zu erwarten, dass unrealistische Annahmen zu Verzerrungen bzw. Verfälschungen der Ergebnisse führten, eine in der Tat sehr unbefriedigende Situation. Mit der Etablierung des Likelihood-Ratio-Tests (Navidi & al. 1991; Goldman 1993a, 1993b) in der Phylogenetik konnte diesem Problem der Maximum Likelihood Analyse abgeholfen werden. Der Likelihood-Ratio-Test vergleicht bei konstanter Topologie die kumulativen Wahrscheinlichkeitswerte dieser Topologie unter unterschiedlich komplexen Substitutionsmodellen. Im Likelihood-Ratio-Test wird stufenweise ermittelt, ob ein komplexeres Substitutionsmodell einen signifikant besseren Wahrscheinlichkeitswert für die Topologie liefert als ein weniger komplexes. Dieser Test ist mittlerweile gut etabliert, auch wenn teilweise kontroverse Ansichten in der Literatur zu finden sind. Eine sehr übersichtliche Testverfahrensroutine ist im Programm Modeltest (Posada & Crandall 1999) implementiert. Zusammen mit dem Programmpacket PAUP (Swofford 1998) ist die Ermittlung eines relativ besten Substitutionsmodells für einen gegebenen Datensatz möglich.

Aber auch hier gibt es nicht so offensichtliche und vielfach ignorierte Probleme. Substitutionsparameter werden ausgehend vom aktuellen Datensatz unter der Annahme der Gültigkeit für den gesamten Datensatz geschätzt. Sollte sich etwa die Substitutionswahrscheinlichkeit von A nach G innerhalb des Datensatzes ändern, so kann dies durch herkömmliche Modelle nicht beschrieben werden und wird somit zu einer weiteren Ursache für unrealistische Annahmen.

Hier gelangen wir wieder zum Kapitel über invariante Positionen zurück. Eine signifikante Heterogenität der Basenkomposition im Datensatz, ermittelt mit einem χ^2-Test, zeigt uns, dass es im vorliegenden Merkmalssatz zu mindestens einer Veränderung der Substitutionswahrscheinlichkeiten gekommen ist. Die im Maximum-Likelihood-Modell angenommen Substitutionsparameter sind für den Datensatz nicht mehr allgemein gültig. Die Verletzung einer wichtigen Voraussetzung für die Anwendung von Maximum-Likelihood-Methoden liegt vor. In solchen Fällen werden wir auf Methoden der Stammbaumrekonstruktion zurückgreifen, die sich als insensitiv gegenüber Veränderungen der Modellparameter erweisen, etwa das LogDet Verfahren (Lockhart & al. 1994). Die Entwicklung solch robuster Verfahren steckt allerdings leider noch in den Kinderschuhen (siehe auch Galtier & Guoy 1998).

Literatur:

Galtier, N. & Gouy, M. (1998): Inferring pattern and process: Maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. — Mol. Biol. Evol. 15: 871–879.

Goldman, N. (1993a): Statistical tests of models of DNA substitution. — J. Mol. Evol. 36: 182–198.

Goldman, N. (1993b): Simple diagnostic tests of models of DNA substitution. — J. Mol. Evol. 37: 650–661.

Lockhart, P.J., Steel, M.A., Hendy, M.D. & Penny, D. (1994): Recovering evolutionary trees under a realistic model of sequence evolution. — Mol. Biol. Evol. 11: 605–612.

Lockhart, P.J., Larkum, A.W., Steel, M.A., Waddell, P.J. & Penny, D. (1996): Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. — Proc. Nat. Acad. Sci. USA 93: 1930–1934.

Misof, B., Rickert, A.M., Buckley, T.R., G. Fleck, G. & Sauer, K.P. (2001): Phylogenetic signal and its decay in mitochondrial SSU and LSU rRNA gene fragments of Anisoptera. — Mol. Biol. Evol. 18: 27–37.

Navidi, W.C., Churchill, G.A & von Haeseler, A.V. (1991): Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. — Mol. Biol. Evol. 8: 128–143.

Posada, D. & Crandall, K.A. (1998): Modeltest: testing the model of DNA substitution. — Bioinformatics 14: 817–818.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. (1994): Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. — Ann. Entomol. Soc. Am. 87: 651–701.

Swofford, D.L. (1998): PAUP. Phylogenetic analysis using parsimony, Version 4*. — Sinauer, Sunderland.

Zharkikh, A. & Li, W.-H. (1992): Statistical properties of bootrstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. — Mol. Biol. Evol. 9: 1119–1147.

Autorenadressen:

Martin Haase
Institut für Natur-, Landschafts- und Umweltschutz
Universität Basel
St. Johanns-Vorstadt 10
CH-4056 Basel
Schweiz

Aktuelle Adresse:
National Institute of Water and Atmospheric Research
PO Box 11-115 Hamilton
New Zealand

Bernhard Misof
Zoologisches Institut und Museum Alexander Koenig
Abteilung Entomologie
Adenauerallee 160
D-53113 Bonn
Deutschland