Codon-based Models of Evolution and Applications in Mammalian Phylogeny

Author(s):
Schneider, Adrian

Publication Date:
2009

Permanent Link:

https://doi.org/10.3929/ethz-a-005842984 →

Diss. ETH No. 18355

Codon-based Models of Evolution and Applications in Mammalian Phylogeny

A dissertation submitted to the
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
ADRIAN SCHNEIDER
Dipl. Informatik-Ing. ETH
born 14 April 1979
citizen of Wetzikon and Volketswil, Switzerland

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Dr. Gina M. Cannarozzi, co-examiner
Prof. Dr. Steven A. Benner, co-examiner
Prof. Dr. Axel Janke, co-examiner

2009

Abstract

Many applications and analyses in computational biology require a model to describe the evolutionary process at the level of DNA or protein sequences. Typically a Markov model is employed which defines the probabilities that a nucleotide or amino acid is replaced by another over time. Besides the reconstruction of phylogenetic trees, models of molecular evolution are also needed for reconstructing the evolutionary history of a particular gene, for searching homologous sequences and their classification as paralogs or orthologs, for determining if a gene or part of it was subject to selection and when this occurred, or for dating evolutionary events (such as speciations or duplications).

This thesis is concerned with modeling sequence evolution and its use in phylogenetic reconstruction. In the first part, an empirical model of evolution for codons (nucleotide triplets) is presented. This model combines the DNA level information of nucleotide substitutions with the protein level information of amino acid substitutions. The range of applicability is established by comparing the codon model to an amino acid model for doing alignments of coding sequences and for phylogenetic reconstructions. Furthermore, the codon model is used as the basis for a method called ”SynPAM” to estimate evolutionary distances based only on synonymous substitutions (substitutions that conserve the encoded amino acid). These substitutions are much less dependent on functional selection pressure and thus provide an evolutionary signal which is more clock-like than that of the nonsynonymous substitutions. Comparisons of SynPAM estimates to other distance measures based on synonymous substitutions show that SynPAM estimates have less variance, contain more phylogenetic signal and increase linearly with time over a longer range.

In the second part of my thesis, evolutionary models are applied to problems in mammalian phylogeny. A reanalysis of a large set of genomic data from human, dog and mouse using several different methods results in strong support for a human-dog clade for most, but not all of the methods. A large-scale analysis on thousands of mammalian quartets shows that it is not uncommon that phylogenetic methods can return strong

statistical support for an incorrect topology. This phenomenon is investigated by means of simulations as well as theoretical analysis, to determine the conditions under which a method can be mislead and thus provide strong support for an incorrect tree. Finally, a method is presented which quantifies the fact that support for a branch should decrease, when increasingly distant outgroups are used. Simulations show that this method allows the identification of the correct branchings even when most phylogenetic methods fail.

Zusammenfassung

Viele Bioinformatikanwendungen benötigen ein Modell, das den evolutionären Prozess auf der Ebene von Protein- oder DNA-Sequenzen beschreibt. Typischerweise wird dazu ein Markov-Modell verwendet, welches die Wahrscheinlichkeiten definiert, dass bestimmte Nukleotide oder Aminosäuren im Laufe der Zeit durch andere ersetzt werden. Ausser zum Rekonstruieren von phylogenetischen Bäumen (evolutionären Stammbäumen), werden Modelle für die molekulare Evolution auch verwendet, um die evolutionäre Geschichte eines Gens zu bestimmen, um homologe Sequenzen zu finden und für deren Klassifizierung als Orthologe oder Paraloge, um zu bestimmen, ob ein Gen oder ein Teil davon Selektion ausgesetzt war und wann das geschah, oder zum Datieren von evolutionären Ereignissen (wie Artenbildungen oder Genduplikationen).

Diese Doktorarbeit beschäftigt sich mit der Modellierung von Sequenzevolution und deren Verwendung für phylogentische Rekonstruktionen. Im ersten Teil wird ein empirisches Modell für die Evolution von Codons (Nukleotidtripletts) vorgestellt. Dieses Modell kombiniert die Information aus den Nukleotidsubstitutionen auf DNA-Ebene mit derjenigen aus Aminosäurensubstitutionen auf Proteinebene. Das Codonmodell wird mit einem Aminosäurenmodell verglichen, um den Anwendbarkeitsbereich zum Alignieren von Sequenzen sowie zu phylogenetischen Rekonstruktionen zu bestimmen. Ausserdem dient das Codonmodell als Basis für eine Methode namens ”SynPAM”, um evolutionäre Distanzen nur auf Grund von synonymen Substitutionen (solche, die die Aminosäure nicht ändern) zu schätzen. Diese Art von Substitutionen ist weniger von funktionalem Selektionsdruck beinflusst und beinhaltet daher ein evolutionäres Signal, welches regelmässiger ”tickt” als das von nicht-synonymen Substitutionen. Vergleiche von SynPAM-Schätzungen mit anderen auf synonymen Substitutionen basierenden Distanzmassen zeigen, dass SynPAM-Schätzungen weniger Varianz aufweisen, ein stärkeres phylogenetisches Signal enthalten und über einen längeren Zeitraum linear ansteigen.

Im zweiten Teil meiner Arbeit werden evolutionäre Modelle auf Probleme im Säugetierstammbaum angewandt. Eine Neuanalyse von einer grossen Menge Genom-

daten vom Menschen, dem Hund und der Maus durch mehrere verschiedene Methoden wird präsentiert, wobei die meisten, aber nicht alle Methoden klar eine nahe Mensch-Hund Verwandtschaft unterstützen. Eine ausgedehnte Untersuchung von Tausenden von Säugetierquarteten zeigt, dass es nicht selten ist, dass phylogenetische Methoden zu starken statistischen Ergebnissen für inkorrekte Bäume kommen. Dieses Phänomen wird durch Simulationen und theoretische Analysen untersucht, um festzustellen, unter welchen Bedingungen eine Methode fehlgeleitet werden kann und damit einen inkorrekten Baum stark unterstützt. Schliesslich wird eine Methode vorgestellt, welche die Tatsache quantifiziert, dass die Unterstützung für einen Baum abnimmt, wenn weiter entfernte Aussenseiter verwendet werden. Simulationen zeigen, dass diese Methode oft den korrekten Baum identifizieren kann, auch wenn die meisten phylogenetischen Methoden scheitern.