

Basic Models of Nucleotide Evolution

15.6.2008

Background and Motivation. The first protein sequences were determined in the 1950s and the first DNA sequences in 1960s. The availability of two homologous (derived from common ancestor) sequences immediately posed the problems of how to count the true number of events in their common history, since only the totally effect of many events could be seen. If an "A" had been turned into a "C" this could be done in many ways, for instance "A" → "C" or "A" → "G" → "C" using 1 or 2 event respectively.

This lead to the need for probabilistic models that could evaluate the probability of different histories. The first such model was presented by Jukes and Cantor in 1969 for proteins. The analogous model was used for DNA a year later by Jerzy Neyman. Almost all models were continuous time Markov Chains, ie the probability of the next transition only depends on the present state

$$P(X_{t_1} = x_1, \dots, X_{t_{i-1}} = x_{i-1}, X_{t_i} = x_i) = P(X_{t_1}) \prod_{j=2}^i P(X_{t_j} | X_{t_{j-1}})$$

Again most models used have following properties

- have state space the set of nucleotides {A,C,G,T} (or proteins) and extended to a complete sequence by assuming independence between different positions.
- *time homogenous* (the same process at any time point). Thus there will be a probability that a nucleotide *i* evolves into nucleotide *j* or a time period *t*, $P_{ij}(t)$, that does not depend on actual start and end of the period, but only its duration.
- Is defined via a rate matrix Q (4 * 4 entries) that describes the probabilistic change over a very short time interval.

The transition probability matrix, $P(t)$, has positive entries and row sums of 1. This gives 3 parameters to vary for each row, thus 12 parameters in total. The rate matrix can be defined via the limits:

$$\lim_{\epsilon \rightarrow 0} \frac{P_{i,j}(\epsilon)}{\epsilon} = q_{ij} \quad \text{and} \quad \lim_{\epsilon \rightarrow 0} \frac{P_{i,i}(\epsilon) - 1}{\epsilon} = -q_{ii}$$

but the definition of most models proceed by defining a rate matrix. Thus the rate matrix, $R = \{r_{ij}\}$, is given, not the other way around $P(t)$ and then Q derived. The rate matrix has negative diagonal elements and row sums 0 (thus positive off-diagonal entries). This again gives 3 parameters to vary for each row, thus 12 in total. $P(t)$ can be retrieved from Q by the matrix Taylor series:

$$P(t) = \exp(tQ) = \sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \dots$$

A series of elementary intuitive properties can be observed from this representation: $P(0) = I$ (identity), $P(\epsilon) \approx I + \epsilon Q$. If any state can be reached from any state and transitions are not deterministic, $P(\infty) = \Pi$, where Π has rows that corresponds to the equilibrium frequencies of the process. All these models have the property/problem, that in $P(t)$ rates and time will always appear together as a product $r_{ij}t$, and thus cannot be separated. A large rate and short time versus large rate and long time, cannot be distinguished without extra information. Such information could be knowledge of the time back to the common ancestor or experimental measurement of mutation rate.

As stated above, only the evolution of one nucleotide was considered. In reality we will have sequences of length L . The model is extended from a single nucleotide to a sequence by assuming that evolution of different nucleotides is independent. Thus

$$P(s1, s2) = \prod_{i=1}^L P(s1_i, s2_i)$$

We will focus on the situation where two sequences are observed and one is assumed to evolve into the other. The probability of this observation, nucleotide *i* in sequence *s1* evolves into sequence *i* in sequence *s2* is

$$P(s1_i, s2_i) = P(s1_i)P(s1_i \rightarrow s2_i)$$

To be able to put the model into this form, the process will have to be time reversible. If the probabilities will be the same if you reverse time direction. This can be shown to be equivalent to $\pi_i P_{ij} = \pi_j P_{ji}$ or to $\pi_i Q_{ij} = \pi_j Q_{ji}$. Time reversible rate matrices have 9 parameters to vary – they are symmetric around the diagonal (i=j) except for a weighting with the equilibrium frequencies (π_i).

Jukes and Cantors' model was the simplest conceivable where all rates were the same: $r_{ij} = \alpha$ if i not equal j and $r_{ii} = -3\alpha$. The transition probabilities can then be calculated to $P_{ij}(t) = 3/4(1 - e^{-3\alpha t})$ and $P_{ii}(t) = 3/4(1 - e^{-3\alpha t})$. Not surprisingly given the symmetry of model, the equilibrium frequency of the model is $1/4$ for each of the 4 nucleotides.

Around 1980 the first mitochondrial sequences were published and was immediately apparent that there was a much higher frequency of *transitions* (the four events: C \leftrightarrow T and A \leftrightarrow G), than *transversions* (the remaining 8 events). This conflicts the total symmetry postulated by the JC69 model. A model accommodating this was proposed by Kimura (K80 model), by having two rates, α for transition events and β for transversion events. The transition probabilities can be calculated as $P_{ij}(t) = 3/4(1 - e^{-3\alpha t})$. Again the equilibrium frequencies are $1/4$ for each nucleotide. Kimura (1981) also introduced a 3-parameter model (extra parameter γ) (K3P), where the transversion events were split into two types.

Even in one sequence it is clear that the frequency for the four nucleotides are highly uneven – often a high frequency of Cs and Gs are observed. This led Felsenstein (1981) to introduce a simple model (F81) that has a desired equilibrium frequency, π_j . Very simply the rates, r_{ij} , is proportional to the popularity of a given nucleotide: $C\pi_j$. This model also has transition/transversion bias. To see this, assume C and T has very high equilibrium frequencies say .49 each. Then most events will be toggling between these two nucleotides. This model has 4 parameters.

In 1995 Hasegawa, Kishino and Yano combined features of K80 and F91 by the model (5 parameters):

$$r_{ij} = \begin{cases} \alpha\pi_j & \text{if } i \text{ and } j \text{ differ by a transition} \\ \beta\pi_j & \text{if } i \text{ and } j \text{ differ by a transversion} \end{cases}$$

Yang (1993) investigated the class of reversible models (REV). Time reversibility is a computational convenience that biology might not obey. For time irreversible models, the position of the root will influence the probability of the data. This complicates computations as assuming that one sequence is the ancestor of the other will give the wrong result. However, it can also be used as an advantage, as now different nucleotides at the root will give different probabilities and the left leg and the right leg might have different lengths that can be distinguished.

In the analysis of real data, the model and parameters will not be known and both model and parameters will have to be estimated. If the model is given, one way of doing this is to choose the parameters that maximizes the probability of the data (maximum likelihood).

Plan:

- Read Yang (2007) chapter 1. Read Python programming in PYTHON
- Make exponentiation program. Exponentiate JC69, K80, K3P, F81, HKY85. Plot $P_{ij}(t)$ for these models. Write down closed expressions for $P_{ij}(t)$, where possible. Check that closed expression matches with exponentiation result.
- Simulate pairs of pairs of sequences of length L nucleotides (L=10, 100, 1000). Plot the probability of the data under model and parameters it was simulated under and also for other models/parameters.
- Find maximum likelihood estimates
- Finish program and report

Comment. i. This project is designed for two high school students (Gregory Craven and Ken Mawhinney) in 6 weeks. ii. Skills obtained through this project: a. increased programming proficiency. B. understanding of basic probabilistic models of DNA evolution. iii. Introduction to an important field within modern biosciences: comparative genomics

Discussion times. These times are only meant as guidance to make sure that students and supervisors meet at regular times.

References.

- J. Felsenstein. Evolutionary trees from DNA sequences. J. Mol. Evol., 17:368--376, 1981.
 J. Felsenstein and G. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol., 13:93--104, 1996.
 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725-736.M.
 Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol., 22: 160--174, 1985.
 J. Jensen and A.-M. Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob., 32:499--517, 2000
 Jukes, TH and Cantor, CR. 1969. Evolution of protein molecules. Pp. 21-123 in H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.
 Motoo Kimura Estimation of Evolutionary Distances between Homologous Nucleotide Sequences Proceedings of the National Academy of Sciences of the United States of America, Vol. 78, No. 1, [Part 2: Biological Sciences] (Jan., 1981), pp. 454-458
 SV Muse and BS Gaut (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome Molecular Biology and Evolution, Vol 11, 715-724
 J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S. Gupta and J. Yackel, editors, Statistical Decision Theory and Related Topics, pages 1--27. Academic Press, New York, 1971.

PYTHON

A.-M. Pedersen and J. Jensen. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18:763–776, 2001.
Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for the maximum likelihood alignment of DNA sequences. *J Mol Evol* 33:114-12
Yang, Z. (2007) "Computational Molecular Evolution" OUP

Possible extensions. Due to the enormous amount of sequence data it is now possible to test models to a much higher degree than when the classic models were proposed and many refinements have been introduced. To list some major refinements

- Evolution can be context dependent for instance creating CG bias (Jensen and Pedersen, 2000;Lunter and Hein,).
- Models for triplets of nucleotides coding for an amino acid (codon) and operating on a state space of size 61 ($4^3 - 3$ stop codons) (Goldman and Yang, ; Gaut and Muse,).
- Nucleotide evolution might dependent on which structure is it part of (gene, RNA or regulatory signal).
- The process can be time heterogeneous, both in the rate matrix () or the rate of evolution.
- Real DNA sequences will also encounter insertion-deletions that can be modelled. (Bishop and Thompson, Thorne Kishino and Felsenstein, 1991)
- We have only considered pairs of sequences, but in reality many sequences can available and they will be related by a phylogeny (tree) with the observed sequences at the leaves and ancestral sequences at the internal nodes. Felsenstein (1981) also presented a method to calculate the probability of a sequences related by a tree evolving according to a specified process.

To do for Jotun: make drawings of little tree. Drawing of Q and P. Illustrate the 4 nucleotides. A nucleotide pair, a sequence pair. Assumption by assumption. Estimating Distance.

History of methods of Molecular Evolution

1958 Sokal and Michener publishes UGPMA method for making distance trees with a clock.
1964 Parsimony principle defined, but not advocated by Edwards and Cavalli-Sforza.
1962-65 Zuckerkandl and Pauling introduces the notion of a Molecular Clock.
1967 First large molecular phylogenies by Fitch and Margoliash.
1969 Heuristic method used by Dayhoff to make trees and reconstruct ancestral sequences.
1969 Jukes-Cantor proposes simple model for amino acid evolution.
1971: Neyman analyzes three sequence stochastic model with Jukes-Cantor substitution.
1971-73 Fitch, Hartigan & Sankoff independently comes up with same algorithm reconstructing parsimony ancestral sequences.
1973 Sankoff treats alignment and phylogenies as on general problem – phylogenetic alignment.
1979 Cavender and Felsenstein independently comes up with same evolutionary model where parsimony is inconsistent. Later called the "Felsenstein Zone".
1979: Kimura introduces transition/transversion bias in nucleotide model in response to publication of mitochondria sequences.
1981: Felsenstein Maximum Likelihood Model & Program DNAML (i programpakken PHYLIP). Simple nucleotide model with equilibrium bias.
1981 Parsimony tree problem is shown to be NP-Complete.
1985: Felsenstein introduces bootstrapping as confidence interval on phylogenies.
1985: Hasegawa, Kishino and Yano combines transition/transversion bias with unequal equilibrium frequencies.
1986 Bandelt and Dress introduces split decomposition as a generalization of trees.
1985-: Many authors (Sawyer, Hein, Stephens, M.Smith) tries to address the problem of recombinations in phylogenies.
1991 Gillespie's book proposes "lumpy" evolution.
1994 Goldman & Yang + Muse & Gaut introduces codon based models
1997-9 Thorne et al., Sanderson & Huelsenbeck introduces the Almost Clock.
2000 Rambaut (and others) makes methods that can find trees with non-contemporaneous leaves.
2000 Complex Context Dependent Models by Jensen & Pedersen. Dinucleotide and overlapping reading frames.
2001- Comparative genomics underlines the functional importance of molecular evolution.

www-pages

<http://tolweb.org/tree/phylogeny.html>
<http://www.treebase.org/treebase/>
<http://evolution.genetics.washington.edu/phylip.html>
<http://paup.csit.fsu.edu/>
<http://morphbank.ebc.uu.se/mrbayes/>
<http://evolve.zoo.ox.ac.uk/blast/>
<http://abacus.gene.ucl.ac.uk/software/paml.html>
<http://www.ncbi.nih.gov/Entrez/>
<http://www.sanger.ac.uk>