Abstract

Models of molecular evolution are important for species conservation because they help us to determine whether two populations belong to the same species or different species. Thus, we can determine if they need to be preserved as one or as two distinct species. The Tamura-Nei model is the most complex reversible model for molecular evolution for which transition probabilities can be explicitly calculated. Usually numerical optimization is used to estimare parameters for this model but instead we will use expectation-maximization (EM) to optimize parameters. They key for expectation-maximization is that it guarantees to find a local maximum and then can be added into other more complex models that already use EM. With EM we can then integrate the Tamura-Nei model into a larger context of models of conservation biology.

1 What is a model of DNA evolution, and how is it used in phylogenetics?

Comparing genome sequences of different organisms from different or from the same species it can be seen that these change over time. These changes or mutations happen through their evolutionary history and can be produced by different causes. Sometimes mutations can produce a fixed polymorphism, which is the occupation of more than one allele at the same gene locus, and be transmitted to their descendants. Phylogeny is the branch of biology that studies the evolutionary development of a species or a taxonomic group. Phylogenetics is focused on the history of a species through sequencing data. It shows the relationship, differences and similarities, among their evolutionary history. Phylogenetic studies are based on the comparison of genomes from different species allows to estimate the historic relationship between them and their distance in time. Estimating the genetic distance between two homologous sequences is measuring the number of differences accumulated between them since they diverge from a common ancestor. It is non trivial to estimate this distance since multiple substitutions can happened, thus the phylogenetic analysis relays on choosing the appropriate substitution model, also called model of DNA evolution. Each model of DNA evolution sets a ratio of substitution per unit of time as well as the frequency of all four DNA bases.

2 Describe the Tamura-Nei Model

The Tamura-Nei model of DNA evolution contemplates the probability of having, at any site, two kind of events: transitions and transversions. If a site has a purine, it assumes that there is a constant probability of α_R per unit time of replacing the base with a random purine, first case of event of type I (transition). Similarly, if a site has a pyrimidine, the model assumes there is a constant probability of α_Y per unit of time of the base being replaced by a random pyrimidine, second case of event type I (transversion). The Tamura-Nei model also assumes that every site has a constant probability of β per unit of time of the base being replaced by a base drawn from all four bases (overall pool), type II event. The overall pool is assumed to contain the four different bases at some frequencies π_A, π_C, π_G, π_T and these will particularly be the equilibrium frequencies.

3 What is the expectation-maximization algorithm? When is it useful?

The expectation-maximization (EM) algorithm is "an iterative method that attempts to find the maximum likelihood estimator of a parameter θ of a parametric probability distribution" (Gupta and Chen, 224). It is an extended used tool for maximum likelihood estimation and "enables parameter estimation in probabilistic models with incomplete data" (Chuong and Batzoglou, 897), where the model relies on unobserved latent variables (hidden data). Two of the more common applications of EM are "estimating Gaussian mixture models, and estimating hidden Markov models" (Gupta and Chen). Finding a maximum likelihood solution commonly involves calculations taking into account all the unknown parameters and latent variables and solving the resulting equations. That is, usually not possible in statistical models. Thus, the EM uses an iterative process to solve two sets (known as steps) of equations numerically. The EM algorithm is guaranteed to find a local maximum.

4 Describe the hidden log-likelihood for the Tamura-Nei model

The likelihood function is the probability, given a set of parameters (model), for a given outcome (observed data). The likelihood value, usually L, is "used in phylogenetic inference is the probability of observing the data under a given phylogenetic tree and a specified model of evolution" (Kosiol, Bofkin and Whelan, 52). That can be translated as "the product of the probabilities of the differences at each site" (Felsenstein, 200). The hidden log-likelihood is the sum of all probabilities of all possible events, except we do not know all the data. The number of every type of transitions occurred are unknown to us.

5 Describe the full log-likelihood for the Tamura-Nei model

The full log-likelihood is the log of "the sum over all possible assignments of states to positions, weighted by the probability of that assignment given a simple stochastic model" (Felsenstein, 225). The full log-likelihood is the natural logarithm of the sum of all probabilities of all possible events weighted by the observed data. In the Tamura-Nei model, given the following probabilities of events: If the branch starts with a purine:

No events	$\exp\left(-(\alpha_R + \beta)t\right)$
Some events type I, no type II	$\exp\left(-\beta t\right)(1 - \exp\left(-\alpha_R t\right))$
Some type II	$1 - \exp\left(-\beta t\right)$

If the branch starts with a pyrimidine:

No events	$\exp\left(-(\alpha_Y + \beta)t\right)$
Some events type I, no type II	$\exp\left(-\beta t\right)(1 - \exp\left(-\alpha_Y t\right))$
Some type II	$1 - \exp\left(-\beta t\right)$

The full log-likelihood can then be calculated as the sum of all probabilities for the sixteen different transitions (subscripts i and j ranging over all four bases A,C,G,T) weighted by the number of observed transitions, where the probability for every transition is given by:

$$
\begin{aligned}
Prob(j|i,t) &= \exp{(-(\alpha_i+\beta)t)}\delta_{ij} && (1)\\
&+ \exp{(-\beta t)}(1-\exp{(-\alpha_j t)})(\frac{\pi_j\varepsilon_{ij}}{\sum_k \varepsilon_{jk}\pi_k}) && (2)\\
&+ (1-\exp{(-\beta t)})\pi_j && (3)
\end{aligned}
$$

Where δ_{ij} is the standard Kronecker delta function, that is one if $i == j$ and 0 otherwise; and ε_{ij} is the Watson-Kronecker function, which is one if either i and j are both purines or both pyrimidines and zero otherwise. Also αi is either α_R or α_Y depending on whether i is a purine or a pyrimidine, respectively.
For the ease of computation it is set,

$$
\begin{aligned}
\exp{(-\alpha_R t)} &= r && (4)\\
\exp{(-\alpha_Y t)} &= p && (5)\\
\exp{(-\beta t)} &= q && (6)
\end{aligned}
$$

Then the full log-likelihood for the Tamura-Nei model,

$$
\begin{aligned}
&S_1[\log{(r)}+\log{(q)}] + S_2[\log{(p)}+\log{(q)}] + S_3[\log{(q)}+\log{(1-r)}] && (7)\\
&+S_4[\log{(q)}+\log{(1-p)}] + S_5[\log{(1-q)}] + S_6\log{(\pi_A)} + S_7\log{(\pi_G)} && (8)\\
&+S_8\log{(\pi_C)} + S_9\log{(\pi_T)} - S_{10}\log{(\pi_R)} - S_{11}\log{(\pi_Y)} && (9)
\end{aligned}
$$

Where

$$
\begin{aligned}
\pi_R &= \pi_A + \pi_G && (10)\\
\pi_Y &= \pi_C + \pi_T && (11)
\end{aligned}
$$

And the constants S_1, S_2, ... , S_{11} are

$$S_1 = X_{AA} + X_{GG} \tag{12}$$
$$S_2 = X_{CC} + X_{TT} \tag{13}$$
$$S_3 = Y_{A-} + Y_{G-} \tag{14}$$
$$S_4 = Y_{C-} + Y_{T-} \tag{15}$$
$$S_5 = Z_{--} \tag{16}$$
$$S_6 = X_{AA} + Y_{A-} + Y_{-A} + Z_{A-} + Z_{-A} \tag{17}$$
$$S_7 = X_{GG} + Y_{G-} + Y_{-G} + Z_{G-} + Z_{-G} \tag{18}$$
$$S_8 = X_{CC} + Y_{C-} + Y_{-C} + Z_{C-} + Z_{-C} \tag{19}$$
$$S_9 = X_{TT} + Y_{T-} + Y_{-T} + Z_{T-} + Z_{-T} \tag{20}$$
$$S_{10} = Y_{-A} + Y_{-G} \tag{21}$$
$$S_{11} = Y_{-C} + Y_{-T} \tag{22}$$

Where X_{ii} is the number of no events (no substitution) on a site, Y_{ij} is the number of events of type I going from i to j, and Z_{ij} is the number of events going from i to j of type II. When a subscript (i or j) is replaced by a dash, we contemplate all possibilities for that subscript (i.e. Y_{A-} contemplates: Y_{AA}, Y_{AG} and Z_{A-} contemplates: Z_{AA}, Z_{AG}, Z_{AC}, and Z_{AT}.

6 Describe the maximum likelihood estimation of parameters using the full log-likelihood

Maximum likelihood is a "long established method for statistical inference" (Kosiol, Bofkin ans Whelan, 52). It is used in phylogenetics to "find the optimal set of parameters contained within the model that best describes the observed data" (Kosiol, Bofkin and Whelan, 52).

For the Tamura-Nei model and give the log-likelihood,

$$S_1[\log(r) + \log(q)] + S_2[\log(p) + \log(q)] + S_3[\log(q) + \log(1-r)] \tag{23}$$
$$+S_4[\log(q) + \log(1-p)] + S_5[\log(1-q)] + S_6\log(\pi_A) + S_7\log(\pi_G) \tag{24}$$
$$+S_8\log(\pi_C) + S_9\log(\pi_T) - S_{10}\log(\pi_R) - S_{11}\log(\pi_Y) \tag{25}$$

The parameters, given our notation specified above, for the maximum likelihood are S_1, S_2 ,S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11} for calculating the values for p, q, r, π_A, π_C, π_G, π_T. For determining the maximum estimation for each value variable, it is obtained from the log-likelihood. The log-likelihood is differentiated with respect to each variable, then set equal to zero and solved for the interested variable. Finally, we obtain a value depending on our parameters,

$$r = \frac{S_1}{S_1 + S_3} \tag{26}$$

$$p = \frac{S_2}{S_2 + S_4} \tag{27}$$

$$q = \frac{S_1 + S_2 + S_3 + S_4}{S_1 + S_2 + S_3 + S_4 + S_5} \tag{28}$$

$$\pi_A = \frac{S_6(S_6 + S_7 - S_{10})}{(S_6 + S_7)(S_6 + S_7 + S_8 + S_9 - S_{10} - S_{11})} \tag{29}$$

$$\pi_C = \frac{S_8(S_8 + S_9 - S_{11})}{(S_7 + S_8)(S_6 + S_7 + S_8 + S_9 - S_{10} - S_{11})} \tag{30}$$

$$\pi_G = \frac{S_7(S_{10} - S_6 - S_7)}{(S_6 + S_7)(S_{10} + S_{11} - S_6 - S_7 - S_8 - S_9)} \tag{31}$$

$$\tag{32}$$

Once we have values for π_A, π_C and π_G, since:

$$\pi_A + \pi_C + \pi_G + \pi_T = 1 \tag{33}$$

Then,

$$\pi_T = 1 - \pi_A - \pi_C - \pi_G \tag{34}$$

7 Describe the E-step and M-step for the EM

The E-step (expectation step) for the EM algorithm consists on “guessing a probability distribution over completions of missing data give the current model” (Chuong and Batzoglou, 898). We calculate, numerically, the estimated values for the missing (hidden) data given our observations. In the Tamura-Nei model and given our notation, the estimated value for the parameters is given by the definition of each parameter and weighted by the observed data. If the observed data, number of transitions, is of the form

From/To	A	C	G	T
A	N_{AA}	N_{AC}	N_{AG}	N_{AT}
C	N_{CA}	N_{CC}	N_{CG}	N_{CT}
G	N_{GA}	N_{GC}	N_{GG}	N_{GT}
T	N_{TA}	N_{TC}	N_{TG}	N_{TT}

The estimated values for the parameters are

$$\overline{S_1} = q \cdot r \cdot (N_{AA} \cdot \pi_A + N_{GG} \cdot \pi_G) \tag{35}$$

$$\overline{S_2} = q \cdot p \cdot (N_{CC} \cdot \pi_C + N_{TT} \cdot \pi_T) \tag{36}$$

$$\overline{S_3} = q \cdot (1 - r) \cdot \frac{N_{CC} \cdot \pi_A^2 + N_{GG} \cdot \pi_G^2 + \pi_A \cdot \pi_G \cdot (N_{AG} + N_{GA})}{\pi_R} \tag{37}$$

References

[1] Chuong, B. D. and Batzoglou, S. *What is the expectation maximization algorithm?*. Nature biotechnology volume 26 number 8, 2008. Available at `http://www.nature.com.ezproxy1.lib.asu.edu/nbt/journal/v26/n8/pdf/nbt1406.pdf`

[2] Felsenstein, J. *Inferring phylogenies* Sunderland, MA: Sinauer Associates, Inc., 2008. Print.

[3] Gupta, M. R. and Chen, Y. *Theory and Use of the EM Algorithm* Foundations and Trends in Signal Processing Vol. 4, No. 3 (2010) 223–296 c 2011. doi: 10.1561/2000000034. Available at `http://mayagupta.org/publications/EMbookGuptaChen2010.pdf`

[4] Kosiol, C. , Bofkin, L. and Whelan, S. *Phylogenetics by likelihood: Evolutionary modeling as a tool for understanding the genome* Journal of Biomedical Informatics Volume 39, Issue 1, (2006). Available at `http://ac.els-cdn.com/S1532046405000766/1-s2.0-S1532046405000766-main.pdf_tid=a3`

[5] Liò, P. and Goldman, N. *Models of Molecular Evolution and Phylogeny* Genome Res. 1998, 8: 1233-1244. doi:10.1101/gr.8.12.1233. Available at `https://www.cs.us.es/ fran/students/julian/phylogenetics/phylogenetics.html`