Models of protein and rRNA evolution

Fredrik Ronquist and Peter Beerli

Sep 2007, Sep 2009

In this lecture, we will be covering models for the evolution of two important gene expression products, proteins and ribosomal RNA.

1 Protein models

DNA sequences (genes) that code for proteins are first transcribed into messenger RNA (mRNA) which is translated with the help of transfer RNA (tRNA) into polypeptides (sequences of amino acids). The polypeptides are then folded to form proteins; often there are several polypeptide units in a single protein. The basic flow of information is affected by several additional processes, including so-called splicing of messenger RNA in higher organisms to remove introns (non-coding DNA segments) from exons (protein-coding segments). There is also some post-translational processing of the polypeptides before they are assembled into functioning proteins. Much of this complexity is ignored here even though it can, in principle, be accommodated in stochastic models of protein evolution.

Polypeptide chains are constructed from twenty amino-acid building blocks, which differ considerably in size, hydrophobicity, and chemical properties (Table 1). These features are all significant in the evolution of protein-coding sequences. For instance, transmembrane proteins usually have regions of hydrophobic amino acids where they are embedded in the cell membrane and regions of hydrophilic amino acids where they interact with the aqueous solution inside or outside the membrane. Most cytosol proteins (proteins that float around in the interior of the cell) have interior regions that consist mainly of hydrophobic amino acids and an exterior dominated by more hydrophilic amino acids.

The translation of RNA into polypeptides is governed by the *genetic code*. Each tRNA molecule

Table 1: The amino acids (sorted alphabetically on their three-letter code)

Trivial name	Three-letter code	One-letter code	Volume [1]	Hydrophobicity[2]	Properties
Alanine	Ala	A	88.6	1.8	hydrophobic
Arginine	Arg	R	173.4	-4.5	basic
Asparagine	Asn	N	114.1	-3.5	hydrophilic
Aspartic acid	Asp	D	111.1	-3.5	acidic
Cysteine	Cys	C	108.5	2.5	hydrophilic
Glutamine	Gln	Q	143.8	-3.5	hydrophilic
Glutamic acid	Glu	E	138.4	-3.5	acidic
Glycine	Gly	G	60.1	-0.4	hydrophilic
Histidine	His	H	153.2	-3.2	basic
Isoleucine	Ile	I	166.7	4.5	hydrophobic
Leucine	Leu	L	166.7	3.8	hydrophobic
Lysine	Lys	K	168.6	-3.9	basic
Methionine	Met	M	162.9	1.9	hydrophobic
Phenylalanine	Phe	F	189.9	2.8	hydrophobic
Proline	Pro	P	112.7	-1.6	hydrophobic
Serine	Ser	S	89.0	-0.8	hydrophilic
Threonine	Thr	T	116.1	-0.7	hydrophilic
Tryptophan	Trp	W	227.8	-0.9	hydrophobic
Tyrosine	Tyr	Y	193.6	-1.3	hydrophilic
Valine	Val	V	140.0	4.2	hydrophobic

has three nucleotides coding for a particular amino acid. These three tRNA nucleotides are called a *codon*. There are $4^3 = 64$ different codons and one different kind of tRNA molecule for each. The codon corresponds to three nucleotides in the sense strand of the DNA, except that the tRNA codon uses the nucleotide uracil (U) instead of the DNA nucleotide Thymine (T). Most life uses the same genetic code, called the *universal genetic code* (Table 2), but there are slightly modified codes used by some life forms.

There are several interesting properties of the genetic code suggesting that it has not been assembled haphazardly. First, most of the redundancy in the code is in the third codon position. The matching of mRNA to tRNA codon nucleotides is less robust at the third position, and translational errors are therefore more likely to occur at that position. Apparently, selection has compensated for

Table 2: The universal genetic code

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	stop	UGA	stop
UUG	Leu	UCG	Ser	UAG	stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

this imperfection by making the third codon position largely redundant, masking most of these translational errors and thereby increasing the fidelity of polypeptide production. Second, if the third codon position is not completely redundant, there is almost always partial redundancy, such that any pyrimidine will translate to one amino acid and any purine to a second amino acid. This structure masks the most common DNA mutations, transcriptional errors, and translational mismatches, namely those involving substitution of one purine with another purine or a pyrimidine with another pyrimidine. Finally, the code generally groups the amino acids such that most single DNA substitutions will, at most, result in the substitution of an amino acid with a chemically similar one in the polypeptide the DNA codes for.

Like most text languages the genetic code has a start signal, which is the same as the codon used for Methionine (AUG). There are also three stop codons (UAA, UAG and UGA), which are used to signal the end of the polypeptide chain.

Because of the redundancy of the genetic code, some DNA substitutions will not result in changes at the amino acid level (synonymous substitutions) whereas others will (nonsynonymous substitu-

tions). Similarly, some amino acid changes require several nucleotide substitutions whereas other require only one. There are basically two kinds of stochastic models for protein evolution: those that deal with only the amino acid changes and disregard the changes at the DNA level, and those that take changes both at the DNA level and at the amino acid level into account. The former are often referred to as *amino acid models* and the latter as *codon models*.

1.1 Amino acid models

Amino acid models only take the changes between the amino acids into account. We can formulate such models by simply applying the four by four nucleotide substitution models described in the last lecture to a larger state space. For instance, if we assume that all stationary state frequencies and rates are identical, we would obtain the analogue of the Jukes Cantor model, sometimes referred to as the *Poisson model* in the context of amino acids. It appears to have been formulated originally by Neyman (1971). The instantaneous rate matrix (unscaled) for this model is a twenty by twenty matrix of the following form

$$Q = \{q_{ij}\} = \begin{Bmatrix} - & 1 & 1 & \cdots & 1 \\ 1 & - & 1 & \cdots & 1 \\ 1 & 1 & - & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & - \end{Bmatrix}$$

Note that we omit the rates along the diagonal (which are -19 in this example); this is quite helpful for more complicated rate matrices because the expressions along the diagonal can be quite involved and they are easily calculated anyway since we know that each row of the rate matrix sums to 0, that is, $q_{ii} = -\sum_{j \neq i} q_{ij}$. Like the Jukes Cantor model, the Poisson model does not have any free parameters (if scaled to a mean rate of 1.0 at stationarity). This means that it does not take the different properties of the amino acids into account, resulting in a poor fit to most real data sets.

A simple extension is to allow the stationary state frequencies of the amino acids to be different. Let π_A be the stationary state frequency of alanine (A), π_R the stationary state frequency of arginine (R), etc. Then the instantaneous rate matrix (unscaled) of the *equalin* model, an extension of the Felsenstein 1981 four by four model of nucleotide (DNA) evolution, is:

$$Q = \{q_{ij}\} = \begin{Bmatrix} - & \pi_R & \pi_N & \cdots & \pi_V \\ \pi_A & - & \pi_N & \cdots & \pi_V \\ \pi_A & \pi_R & - & \cdots & \pi_V \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_A & \pi_R & \pi_N & \cdots & - \end{Bmatrix}$$

We can write this in more compact form by saying that the instantaneous rate for the change from amino acid i to the amino acid j is:

$$q_{ij} = \begin{cases} \pi_j & : \quad i \neq j \\ -\sum_{j \neq i} \pi_j & : \quad i = j \end{cases}$$

or, if we omit the negative rate:

$$q_{ij} = \pi_j \quad : \quad i \neq j$$

The number of free parameters in the equalin model is 19 $(20 - 1)$.

The most general time-reversible model may be referred to as the *GTR model* for amino acids, and is completely analogous to the GTR model for four by four DNA data:

$$q_{ij} = \pi_j r_{ij} \quad : \quad i \neq j$$

The number of free parameters in this model is 19 $(20 - 1)$ stationary state frequencies and 189 $(((20 * 20 - 20)/2) - 1)$ relative rate parameters, in total 208 free parameters. Depending on the amount of data we have at hand and the statistical approach we are using, the large number of parameters in this model may cause problems even though it usually provides the best fit to the data of the models discussed thus far.

A practical alternative to using the full GTR model is to fix the rates to those determined from some large empirical data set. This approach accounts for the differences in substitution rates among pairs of amino acids without necessitating the estimation of these rates in a particular analysis. However, the success of the method obviously depends heavily on the accuracy of the fixed rate matrix and its applicability to the data at hand. And we still need to estimate the rates using some large data set.

Many different rate matrices have been derived for different types of proteins using more or less sophisticated approaches. Commonly, one simply scores the number of observed amino acid differences between pairs of closely related sequences. Closely related sequences are used so that the

possibility of multiple amino acid changes at a particular site can be ignored. More sophisticated estimation of rate matrices involves statistical inference on a phylogeny.

1.1.1 Computational complexity

Even though fixed rate matrices are helpful in that they reduce the number of free model parameters to estimate, they do not change the basic computational complexity of calculating parsimony scores or likelihoods under amino acid models, which is determined by the number of states. The computational complexity for inference under a discrete model with k states is $O(k^2)$ as we have shown previously for Sankoff parsimony and will show later for likelihood calculations. Amino acid models with 20 states thus require roughly $(20 * 20)/(4 * 4) = 25$ times the computational effort of a standard four by four model of DNA substitution. If the rate matrix is not fixed but estimated during inference, then we need to recalculate the eigenvalues and eigenvectors, which is an operation of time complexity $O(k^3)$. Thus, inference under the equalin and GTR models is roughly $20^3/4^3 = 125$ times slower than the comparable four by four models of nucleotide substitution.

1.2 Codon models

Codon models accommodate both changes at the DNA level and changes at the amino acid level in protein-coding sequences. The basic unit in these models is the codon. In theory, this requires a state space of 64 states, one for each possible codon. However, there are typically three stop codons (Table 2) that can be disregarded, so the state space can be shrunk down to 61 states.

A GTR model for codons would be extremely parameter-rich and has, to our knowledge, not been implemented yet. Instead, various simplifications have been introduced in order to model codon evolution realistically with a limited number of parameters.

A standard assumption in the codon models that have been explored so far is that the instantaneous rate of change between two codons is zero if the change involves more than one nucleotide change. This does not mean that the change is impossible, only that it has to occur through one or more intermediate steps, each involving the replacement of a single nucleotide.

The remaining non-zero non-diagonal entries in the instantaneous rate matrix now fall into two different types: those that represent a synonymous change (the two codons produce the same amino acid) and those that involve nonsynonymous changes (the two codons produce different

amino acids). In the simplest possible case, we can assume that synonymous nucleotide changes occur at rate μ and that nonsynonymous nucleotide changes occur at rate $\omega\mu$. Assume that $d(i,j)$ is the number of nucleotide changes required to go from codon i to j and that a_i is the amino acid specified by codon i. Then we can formulate the rate of going from codon i to codon j as

$$q_{ij} = \begin{cases} 0 & : \quad d(i,j) > 1 \\ \omega\mu & : \quad d(i,j) = 1, a_i \neq a_j \\ \mu & : \quad d(i,j) = 1, a(i) = a(j) \end{cases}$$

if we omit the diagonal (negative) rates as usual.

There are several different ways of accommodating unequal state frequencies in codon models. An approach that is parsimonious with the number of parameters is to assume that there are overall stationary state frequencies for the four nucleotides (A, C, G, and T), and that the stationary frequencies of the codons is simply determined by multiplying these frequencies together. Thus, the codon `AUG`, corresponding to the DNA triplet `ATG`, would have the stationary state frequency $\pi_A\pi_T\pi_G$. This solution requires only four (three free) parameters.

An alternative model is to allow a stationary state frequency for each of the 61 codons. Although this introduces a lot more parameters, there is ample evidence of so-called codon usage bias in many types of organisms. To avoid estimating all 60 free parameters of this model, it is common to determine the stationary codon frequencies by simply counting the frequencies of the codons in the data matrix. This tends to be fairly accurate but it does not take the relatedness of the sequences into account, and can therefore lead to biased estimates.

A compromise approach is to use stationary state frequencies for the four nucleotides but allow these frequencies to be different for the different codon positions. Thus, there would be one stationary frequency for nucleotide `A` at codon position 1, $\pi_A^{(1)}$, another frequency at position 2, $\pi_A^{(2)}$, and a third frequency at position 3, $\pi_A^{(3)}$. The stationary state frequency of a codon such as `AUG` would now be determined as $\pi_A^{(1)}\pi_T^{(2)}\pi_G^{(3)}$.

The rate parameter ω is interesting; it is the ratio of the nonsynonymous to the synonymous substitution rates (observe that this is a rate ratio just like κ in our formulation of the HKY and K2P models of DNA evolution). If $\omega < 1$, then nonsynonymous changes are more rare than synonymous changes. This is evidence of natural selection discriminating against change of amino acid, often called negative or constraining selection. If $\omega = 1$ then the two rates are identical and evolution is selectively neutral. Finally, if $\omega > 1$ then there is positive selection favoring amino acid changes. Although relatively rare in nature, positive selection often indicates that something interesting is going on. For instance, a virus protein that is targeted by the host immune response

might be expected to be subject to positive selection. Thus, methods that estimate ω could be used to find such proteins. Another similar example is that animal breeding should result in positive selection on the genes responsible for the traits being selected for; this should enable identification of such genes with methods that estimate ω.

The basic type of codon model described here can easily be extended. Assume for instance that we allow each codon i to have its stationary frequency π_i. Furthermore, assume that a_i is the amino acid specified by codon i and that n_i and n_j are the two nucleotides that differ between two codons i and j separated by only one nucleotide difference. We can now use any four by four model of DNA substitution and any twenty by twenty model of amino acid substitution to give us a codon model with the (unscaled) structure:

$$q = \begin{cases} 0 & : \quad d(i,j) > 1 \\ \pi_j \omega r_{a_i a_j} r_{n_i n_j} & : \quad d(i,j) = 1, a_i \neq a_j \\ \pi_j r_{n_i n_j} & : \quad d(i,j) = 1, a(i) = a(j) \end{cases}$$

It is often assumed that the parameter ω varies across sequences (from one part of a sequence to another) and across lineages. There has been some progress in accommodating this variation, particularly the variation across sequences. There are two general approaches that can be used for the latter: Hidden Markov Models and mixture models. Hidden Markov Models simply allow variation in a value like ω across sites while favoring solutions that assign adjacent sites similar ω values. The mixture model assumes that the sites are drawn independently from a mixture of a fixed number of categories of ω values. For instance, we can assume that there is a probability p_+ that the site belongs to a positively selected category with $\omega_+ > 0$, a probability p_0 that the site is neutral with $\omega_0 = 1$, and a probability p_- that the site is under negative selection with $\omega_- < 1$. All of these parameters can then be estimated from the data at hand. Note that the mixture model does not favor solutions with adjacent sites having similar ω values.

2 Ribosomal RNA models

Some genes do not code for proteins but are translated to RNA. The secondary structure of these RNA molecules involve two basic structural elements: loops and stems. In the loop regions, the RNA is single-stranded, while it pairs with itself using standard Watson-Crick nucleotide matching in the stem regions (Figure 1).

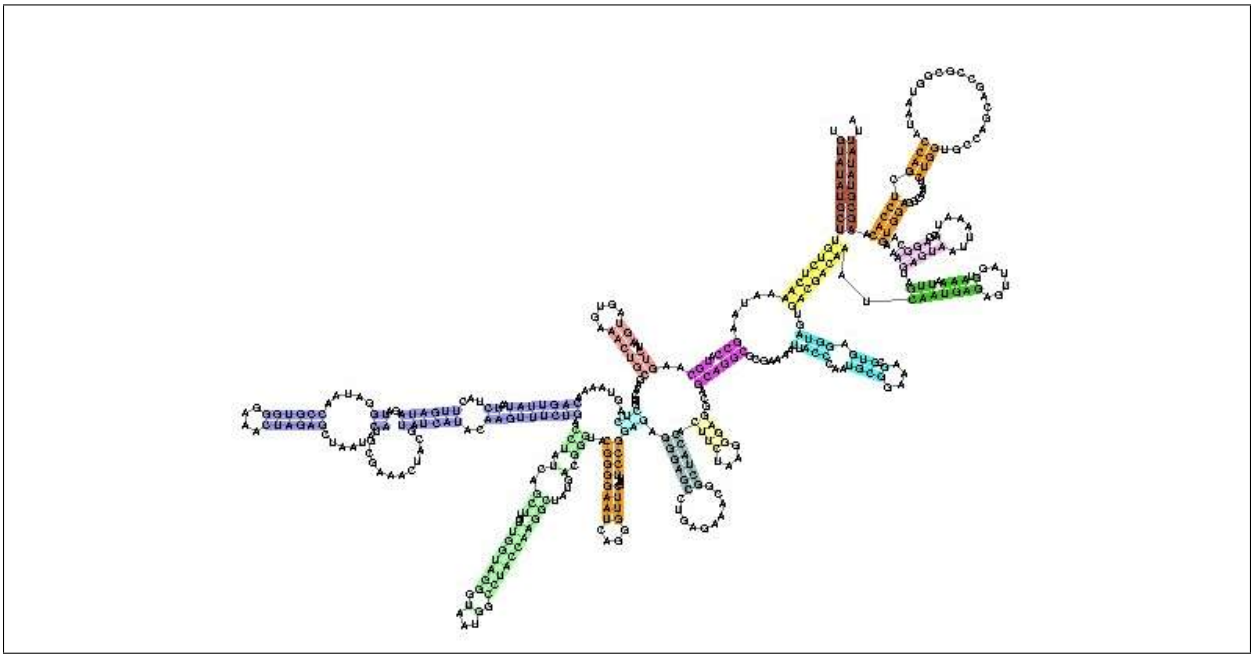

Figure 1: Secondary structure of ribosomal RNA. Note the stem regions where the RNA pairs with itself.

While a standard four by four model can be used for the nucleotide sites in loop regions, it is obvious that changes in stem-region sites will be strongly correlated with changes in their matching sites. If this correlation is not accounted for, then estimates of phylogeny and other model parameters will be excessively precise.

Models of stem regions essentially take two different forms. The first form is similar to the codon models described above. The state space is all of the 16 possible state pairs (or doublets) (AA, AC, AG, AT, ..., TT), and we allow changes of only one nucleotide at a time. The nucleotide substitutions are modeled using a standard four by four model of DNA evolution, while the stationary state frequencies are allowed to be different for all the sixteen doublets. If we use similar notation introduced above for the codon models, the rate of changing from one doublet i to another doublet j is now given by:

$$q_{ij} = \begin{cases} 0 & : \quad d(i,j) > 1 \\ \pi_j r_{n_i n_j} & : \quad d(i,j) = 1 \end{cases}$$

There has been some work (Jow et al. 2002; Savill et al., 2001) suggesting that such a model is relatively poor for two reasons. First, there are only six doublets that occur with any significant frequency (AU, GU, GC, UA, UG and CG). Second, the rate of double substitutions is high because of a tendency for compensatory mutations to occur once a single nucleotide has been replaced. Thus, it is not realistic to model evolution as occurring in single nucleotide substitution steps. Thus, a six

by six GTR model generally tends to outperform the single-step sixteen by sixteen model (Savill et al. 2001). The GTR model would have the rates

$$q_{ij} = \pi_j r_{ij}$$

like a normal GTR model. If one wanted to model the rare doublets in such a model, they can be introduced by adding a seventh state (Jow et al. 2002).

3 Study questions

1. What is the difference between an amino acid and a codon model?
2. How can an amino acid rate matrix be estimated?
3. Describe the significance of the ω parameter in codon models.
4. Why are special models needed for stem regions of ribosomal RNA?
5. Why do you think that double substitutions are so common in stem regions?
6. Is there any reason to suspect that double substitution are common in codons as well?