J Mol Evol (2005) 61:659–665
DOI: 10.1007/s00239-005-0060-0

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2005

A Collection of Amino Acid Replacement Matrices Derived from Clusters of Orthologs

Rolf Olsen,[1] William F. Loomis[2]

[1] Department of Physics, University of California at San Diego, La Jolla, CA 92093, USA
[2] Division of Biology, University of California at San Diego, 2317 Bonner Hall, La Jolla, CA 92093, USA

Received: 16 March 2005 / Accepted: 4 June 2005 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. Sequence divergence among orthologous proteins was characterized with 34 amino acid replacement matrices, sequence context analysis, and a phylogenetic tree. The model was trained on very large datasets of aligned protein sequences drawn from 15 organisms including protists, plants, *Dictyostelium*, fungi, and animals. Comparative tests with models currently used in phylogeny, i.e., with JTT + Γ ± F and WAG + Γ ± F, made on a test dataset of 380 multiple alignments containing protein sequences from all five of the major taxonomic groups mentioned, indicate that our model should be preferred over the JTT + Γ ± F and WAG + Γ ± F models on datasets similar to the test dataset. The strong performance of our model of orthologous protein sequence divergence can be attributed to its ability to better approximate amino acid equilibrium frequencies to compositions found in alignment columns.

Key words: Protein sequence divergence — Phylogeny

Introduction

In some of the first comparisons of the amino acid sequences of orthologs, Zuckerkandl and Pauling (1965) expected that equivalent positions would be filled by amino acids with similar chemical properties more frequently than amino acids with differences in size, charge, or hydropathy. They pointed out that most mutations resulting in amino acids that alter the local protein context would be eliminated by selection if they were incompatible with function. Their expectations were largely confirmed in the larger database used by Dayhoff et al. (1972, 1978) to construct their PAM matrix. This matrix as well as the BLOSUM series (Henikoff and Henikoff 1993) and the JTT matrix (Jones et al. 1992) use amino acid frequencies observed along the length of many protein sequences. However, as many people have observed, a large fraction of all columns in multiple alignments of protein sequences is dominated by a few amino acids (Taylor 1986; Livingstone and Barton 1993; Brown et al. 1993; Miyamoto and Fitch 1995; Sjolander et al. 1996; Halpern and Bruno 1998). The compositions of these columns are not well represented by average frequencies. Recent efforts to generate improved matrices continue to use database-averaged frequencies (Müller and Vingron 2000; Whelan and Goldman 2001; Lin et al. 2001; Dimmic et al. 2002; Veerassamy et al. 2003; Xu et al. 2004). Adjustments of these matrices to specific proteins focused on length rather than position (Cao et al. 1994; Goldman and Whelan 2002; Yu et al. 2003; Yu and Altschul 2004).

We have developed a model of amino acid replacements that contains a collection of 34 replacement matrices. These were derived from the frequencies observed in alignment columns of a very large number of likely orthologs from organisms separated by hundreds of millions of years. Each

Correspondence to: William F. Loomis; *email:* wloomis@ucsd.edu

position can be characterized by several different matrices with different weights. Twenty of the matrices account for highly conserved positions of each of the 20 amino acids. There is also a generic matrix similar to the PAM for substitutions between any pair of amino acids. Thirteen other matrices are used for modeling replacements between subgroups of amino acids sharing physical similarities. The model also adjusts weights on the matrices in protein regions to capture domain-specific differences in selection pressure. A phylogenetic tree is generated by the model and is expanded as new organisms are iteratively added to the study. When the mature model was compared to models currently used in phylogeny (JTT + Γ ± F, WAG + Γ ± F), it was found to be better.

The model can also be used for classifying homologous relationships between sets of protein sequences by making tests of their historical lineage relationships. Recognizing clusters of orthologous proteins by sequence comparison can assist in predicting related functions and provide useful information on phylogeny.

Materials and Methods

Proteome Database

For initial model construction, the complete proteomes of six organisms (*Arabidopsis thaliana, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, Fugu rubripes*, and *Homo sapiens*) were downloaded from major archival and genome sequencing institutions (see supplementary material). In subsequent expansions of the dataset, proteomes of nine more organisms were included (*Plasmodium falciparum, Leishmania major, Oryza sativa, Zea mays, Dictyostelium discoideum, Neurospora crassa, Caenorhabditis elegans, Anopheles gambiae*, and *Ciona intestinalis* (see supplementary material).

Clusters of Orthologs

The "COG" method, introduced by Tatusov et al. (1997), was applied to the six-proteome dataset to find clusters of likely orthologous sequences. Easy-to-align columns were identified by comparing four distinct multiple alignments done on each cluster using T-Coffee (Notredame et al. 2000) and ClustalW with gap opening costs of 6, 10, and 18 (Thompson et al. 1994). Alignment blocks with at least five gap-free columns that were found in all four multiple alignments or received a T-Coffee column score of at least 5 were considered easy-to-align. Amino acids falling outside an easy-to-align column were not considered further. Hard-to-align regions hold little functional or phylogenetic information and can lead to misassignments (Altschul 1998; Fitch 2000). Each cluster of sequences was represented by concatenating blocks of easy-to-align columns. Clusters in which at least half of the amino acids were found in the easy-to-align columns were accepted into the dataset. Clusters with fewer easy-to-align columns were reanalyzed after removing members. Those that passed the 50% threshold and still had at least three members were accepted into the dataset.

Parameters and a Phylogenetic Tree

Equilibrium frequencies of our matrices were established by self-consistent counting. All other model parameters were set using maximum (log)likelihood on the dataset with a phylogenetic tree. The tree topology is well established for the initial six organisms (two vertebrates, an insect, two yeasts, and a plant). The branch length ratios of the tree were computed simultaneously with the model parameters using maximum (log)likelihood on the entire database of multiple alignments. Proportional trees with varying lengths were used for individual clusters. The Pruning algorithm was used to compute the probabilities of an alignment column given a matrix (Felsenstein 1981). The branching pattern of the 15 organisms was essentially identical to that determined by Bapteste et al. (2002).

Iterative Expansion

The clusters were expanded by adding nine proteomes, one at a time, with redefinition of parameters. As each organism was added, a tentative expanded tree was established on a subset of the new proteome assigned to clusters. In the expansion steps cluster additions were based on comparisons of different test topologies and branch length variation rather than the COG methodology used in the initial set. Previously accepted members of clusters were also serially removed from the clusters and subjected to the same tests to establish likely orthology. The seven proteome dataset went through three rounds of cluster database orthology testing and tree and model parameter retraining. The tree and model parameters were retrained when each of the next three proteomes was added. Retraining was not continued thereafter because there was little change or gain in model performance after addition of the tenth proteome.

Regional Sequence Context

The modular structure of many proteins results in significant differences in the rate of amino acid replacement from region to region (Goldman et al. 1998; Tourasse and Li 2000). Hidden Markov models have been used to assign rate categories in multiple alignments of DNA sequences (Yang 1995; Felsenstein and Churchill 1996). Our regional sequence context adjustment is similar to a hidden Markov model. It adjusts the weights (prior probabilities) on the use of the individual matrices to the density of conserved positions. Its five states are weight vectors that are one-parameter variants of the average matrix use probabilities (see supplementary material). The sequence context adjustment differs from a standard HMM in that the state transition probabilities are dependent on how long a state has been occupied. For each state, this dependency is such that the length of occupation of the state (of optimal paths) follows a state-specific, shifted Γ distribution.

Comparative Tests

We used a test set of all (380) alignments that contained at least one sequence from each of the five major taxonomic groups represented in the dataset, i.e., protists, plants, *Dictyostelium*, fungi, and animals. These were extracted from a dataset of 5038 multiple alignments of proteins from 15 organisms. The parameters of our model were reestimated on the remaining 4658 multiple alignments to avoid any bias by the test set. Performance of our model and various forms of the JTT and WAG models on the test set was compared by resampling BIC model selection (Schwartz 1978; Efron 1979). In addition to testing the JTT and WAG models with fully adjustable trees, we tested them with proportional and fixed trees. The fixed tree was established on concatenated aligned

Hydrophobic

IV	6644	LM	2095	FI	445	FM	229
IL	4591	FL	1750	MV	341	ILV	212
LV	2739	IM	648	FV	329		

Alanine and Threonine with Valine, Leucine, Isoleucine

AV	1240	AL	488	IT	385
TV	951	LT	405	AI	309

Small

AS	2846
AG	1377
GS	881

Alanine and Glycine with Threonine and Asparagine

AT	865
GN	523
AN	255
GT	185

Small/Medium Hydrophilic

ST	2447	DS	579
DN	1107	NT	484
NS	1013	DT	243

Acidic and Medium/Large Hydrophilic

KR	3956	EK	1143	ER	348	HK	218
DE	3744	KQ	1003	HQ	333	EH	108
EQ	1299	QR	497	HR	218		

Large Hydrophilic with Small Hydrophilic

KS	659	KT	501	QS	381	DK	338
KN	591	EN	416	ET	360	QT	294
ES	562	HN	382	NQ	347		

Large Aromatic

FY	2555	FW	313	FH	147
HY	473	WY	286	HW	29

Proline and Cysteine with Alanine, Serine, Valine

AP	676	PS	488	CS	387
AC	633	CV	470	PV	180

Other Top 60 Ranked Mixed Alignment Columns

AE	674	EP	396	AQ	367	KL	340
AK	581	LY	394	DG	363	KP	318

Fig. 1. Number of multiple alignment columns by composition. Each alignment column counted had six amino acids, one each from *A. thaliana, S. cerevisiae, S. pombe, D. melanogaster, F. rubripes*, and *H. sapiens*. Column counts were placed into bins according to composition. If an amino acid appeared two or more times in a column, it was counted into a bin indexed by that amino acid.

sequences from EF1α, HSP70, DNAJ, ribosomal S14 and L10, a vacuolar proton pump ATPase, and TCTP, which have clear orthologs in all 15 organisms. Proportional trees were constant ratio variants of the fixed tree. The PAML package was used to perform all work with the JTT + Γ and WAG + Γ models (Yang 1997). Eight categories were used for the discrete Γ distribution controlling rate variation (ncatG = 8). The resampled versions of our 380 multiple alignment test dataset were generated by a standard program, seqboot (Felsenstein 1993). Only seven resampled datasets could be analyzed because the four slowest competing models would each require 5 years of 1-GHz processor time for 100 resamplings.

Results

Counting and Choice of Matrices

Our collection of matrices is structured to fit amino acid replacements commonly found in multiple alignment columns. The composition of 329,062 columns in the 976 multiple alignments of the proteins from the six widely separated organisms (see Materials and Methods) was determined. About a third of the columns were filled by a single amino acid. Another 58,250 columns were dominated by a single amino acid but had one row replaced, while another 84,511 were biased to a single amino acid but contained others. Most of the remaining columns (77,339 out of 85,146) were dominated by a mixture of two or three amino acids. These were counted into bins indexed according to composition; 80% of the counts were concentrated in 60 of the 1309 potential bins (Fig. 1).

Inspection of the bins immediately suggested the use of matrices with only a subset of the 20 amino acids since the composition of many individual

IV

I	V	L
0.458	0.442	0.100

Medium/large hydrophobic

I	V	L	M	F	W
0.230	0.033	0.554	0.086	0.091	0.006

Small/ medium hydrophobic

I	V	L	M	F	A	T
0.217	0.296	0.238	0.070	0.058	0.067	0.055

Small

A	G	S	T	V	N
0.370	0.245	0.205	0.078	0.073	0.028

Small/medium hydrophilic

S	T	N	D
0.24	0.301	0.209	0.248

Acidic

D	E	Q	N
0.447	0.484	0.033	0.033

Basic

K	R	Q	N	H
0.444	0.370	0.082	0.066	0.038

Medium/large hydrophilic

E	Q	K	R	H
0.34	0.117	0.252	0.242	0.041

Hydrophilic

S	T	N	D	E	Q	K	R	H	A	G
0.101	0.057	0.083	0.139	0.193	0.080	0.133	0.060	0.026	0.064	0.063

Large aromatic

F	Y	W	H	L	M
0.354	0.404	0.135	0.046	0.053	0.007

Large

F	Y	W	H	L	M	Q	K	R
0.07	0.089	0.02	0.04	0.21	0.081	0.05	0.19	0.21

Proline and cysteine with small/medium

P	C	A	S	V	T	G	E	K
0.064	0.149	0.265	0.155	0.107	0.114	0.062	0.054	0.031

Fig. 2. Matrix equilibrium frequencies. The equilibrium frequencies are taken from the results of retraining done on the 15-organism dataset. The equilibrium frequencies of the 20 highly conserved and two 20 × 20 matrices in the collection are not shown. The equilibrium frequencies of the amino acid associated with each highly conserved matrix are all higher than 0.93.

positions was restricted to amino acids with shared physical properties, particularly hydropathy and size. Together with 20 "highly conserved" matrices, one for each amino acid, we included matrices dedicated to frequently observed pairwise exchanges. We included "IV," "acidic," and "basic" matrices for positions dominated by isoleucine and valine, aspartic and glutamic acids, lysine and arginine. The collection also contains matrices for exchange of "small," "small/medium hydrophobic," "medium/large hydrophobic," "small/medium hydrophilic," "medium/large hydrophilic," and "large aromatic" amino acids. The amino acids in the IV matrix overlap with those in the hydrophobic matrices (I,V,L,F,M,W and I,V,L,A,T,F,M); however, the equilibrium frequencies of these three matrices are very different from each other (Fig. 2). The remaining matrices in our collection cover more varied replacement patterns. One is a 11 × 11 "hydrophilic" matrix which can model replacements between amino acids that are not modeled by the more specific hydrophilic matrices. Another is a 9 × 9 "large" matrix. There are two 20 × 20 matrices, "hydrophobic biased" and "generic," so that any amino acid replacement can be modeled with nonzero probability. The generic matrix was initialized to be a 0.1 PAM version of the JTT matrix (Jones et al. 1992), although its parameters were repeatedly retrained and ultimately took on a noticeable bias for hydrophilic replacements.

We tested collections with different matrices and kept only those that improved the performance of the model. The 34 matrices in the final collection were given mnemonic names that suggest, but do not define, their sets of interchangeable amino acids. Although not hydrophilic themselves, alanine and glycine were included in the hydrophilic matrix because of the relatively high frequency with which they were seen to exchange with hydrophilic amino acids. Tyrosine was excluded from this matrix because it was seldom observed to exchange with other hydrophilic amino acids. The equilibrium frequencies in individual matrices are shown in Fig. 2. Many of the frequencies are much higher than typical frequencies seen in standard 20 × 20 matrices. In some matrices, such as IV, Acidic, and Basic, the frequencies of two amino acids account for a high proportion, effectively limiting the alphabet size.

Comparison Between Models

The performance of several forms of the JTT + Γ ± F and WAG + Γ ± F models (Jones et al. 1992; Yang 1994 ; Cao et al. 1994; Whelan and Goldman 2001) were compared to our model by resampling BIC model selection (Schwartz 1978; Efron 1979). Likelihood ratio tests could not be used because the models were not nested. In the comparative BIC tests, models that used amino acid frequencies specific to each multiple alignment (+ F) were found to perform less well than models with fixed frequencies. The strongest performing models used proportional trees rather than freely adjustable or fixed trees (see Materials and Methods). Models that used the WAG matrix always outperformed the corresponding JTT model. The strongest performing WAG + Γ model has two adjustable parameters for each multiple alignment, Γ and absolute branch length, while our model only adjusts absolute branch length.

Our model was preferred over the strongest performing WAG + Γ model in 7 of 7 resampling BIC tests. Further, the margin of preference was very large. The average difference in BIC values between our model and the best WAG + Γ model was over 21 times their standard deviation. The distribution of BIC differences is most probably Gaussian due to the large size of the test dataset. Differences found in the seven tests fit a Gaussian well (χ^2 was better than what is expected of data drawn from a real Gaussian). Although we could not extend these trials much further (see Materials and Methods), the 21 standard deviations and the compact distribution mean that our model would be preferred over the best WAG + Γ model on any resampled dataset with a probability very close ($\sim 10^{-90}$) to 1.

Discussion

The constraints that selection imposes on amino acid sequences differ from position to position along the length of proteins. Some positions are forced to be almost invariant, while others can vary among members of sets of amino acids with similar physical properties and still maintain function. Other positions can vary among a broad variety of amino acids. Phylogenetic models that rely on a matrix of amino acid replacements based on observed frequencies that are averaged over the length of proteins cannot capture position-specific frequencies nor accurately estimate the degree of divergence among orthologs. We have generated a collection of 34 matrices that more closely model the inherent variation from position to position. While selection pressures on each position differ, the use of weighted matrices from the collection better approximates sequence divergence.

The number of possible matrix collections that can be considered is almost unlimited. However, inspection of position-specific amino acid frequencies in a very large set of alignment columns immediately suggests a set of pertinent matrices that should be included in any collection. The large number of invariant positions in easy-to-align regions requires that each of the 20 amino acids has a dedicated matrix in the collection. Replacements of amino acids with similar hydropathy and size, such as isoleucine and valine, aspartate and glutamate, and lysine and arginine, are frequently observed at equivalent positions and can be captured by matrices that focus on these hydrophobic, acidic, and basic amino acids to the exclusion of most others. A generic 20 × 20 matrix has to be included so that any amino acid replacement can be modeled with nonzero probability.

The choice of the remaining 10 matrices in our collection was directed by the observed position-specific replacement frequencies of amino acids with shared properties. We found that small amino acids were often found at equivalent positions and generated a matrix for these exchanges. Likewise, we observed a high frequency of replacements among aromatic amino acids and so included a matrix dedicated to these amino acids. We explored a considerable number of collections of matrices and kept only those that improved the performance of the model. The final collection was based on amino acid replacements observed in 5038 multiple alignments generated from the proteomes of 15 diverse organisms and so covers a large portion of eukaryotic evolution. While we cannot be certain that we have the optimal set of matrices, the sparing use of the 20 × 20 matrices indicates that most common replacements are accounted for. Moreover, addition of the last three matrices, those for "large," "broadly hydrophobic," and "proline and cysteine with small medium" amino acids, only marginally improved the performance of the final model over a preceding model that used 31 matrices. When the model with the final set of 34 matrices was compared to models presently used in phylogenetic analyzes of orthologs (JTT + Γ ± F and WAG + Γ ± F), it outperformed them.

Our model has the advantage of being able to handle large databases efficiently. During the final steps of the iterative addition of proteins predicted from fully sequenced genomes, there were over 10 million amino acids in multiply aligned positions. Müller et al. (2002) said that a training dataset of 6 million amino acids was unfeasibly large for maximum likelihood estimation. As each proteome was added, individual proteins were initially assigned to established clusters on the basis of Blast results. However, once the tentative tree position was established for a new organism, all members of the clusters were reanalyzed for historical evidence of orthology and the clusters readjusted. This technique has advantages over mutual best Blast hit approaches to establishing likely orthologs that can be confounded by gene loss.

The model depends on a phyogenetic tree built up iteratively from the initial tree of animals, fungi, and a plant where the branching topology is well established. It was rooted on a set of seven archaebacteria whose genomes had been fully sequenced. The positions of the nodes and the branch lengths were determined on the basis of maximum likelihood tests with the whole genome datasets. The measurements of divergence are robust to further additions as well as minor changes in the model parameters. As each organism was added, various positions on the tree were tested and the best performing node and branch length established.

The strong performance of our model of orthologous protein sequence divergence can be attributed to its ability to better approximate amino acid equilibrium frequencies to compositions found in alignment columns. These compositions appear to be the result of position contextual constraints imposed by functional selection which are taken into account in our model. At least on the timescales of more than 1 billion years separating the major groups of eukaryotes represented in our dataset (Feng et al. 1997), the close fit of our model to the data supports the predictions of Zuckerkandl and Pauling (1965).

The datasets, computer programs, and scripts used in this study are freely available upon request to the authors.

Acknowledgments. We thank T. Hwa for guidance and encouragement and R. Konecny of the W.M. Keck Laboratory for Integrated Biology II for helping to keep the processors running. R.O. was supported in part by a grant from the NIH to W.F.L. (GM-062350), grants from the NSF to T.H. (DMR-9971456, DMR-0211308, MCB-0083704), and the Center for Theoretical Biological Physics.

References

Altschul SF (1998) Generalized affine gap costs for protein sequence alignment. Proteins 32:88–96

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, Philippe H (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A 99(3):1414–1419

Brown M, Hughey R, Mian IS, Sjolander K, Underwood R, Haussler D (1993) Using Dirichlet mixture priors to derive Hidden Markov Models for protein families. In: Hunter L, Searls D, Shavlik J (eds) Proceedings of First International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA

Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 39:519–527

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of Protein Sequence and Structure, Vol 5. National Biomedical Research Foundation, Washington, DC, pp 89–99

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, Vol 5, National Biomedical Research Foundation, Washington, DC, Suppl 3, pp 345–358

Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55:65–73

Efron B (1979) Bootstrap methods: another look at the jacknife. Ann Stat 7:1–26

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package), version 3.5c. Distributed by the author. Department of Genetics University of Washington, Seattle

Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13(1):93–104

Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. Proc Natl Acad Sci USA 94:13028–13033

Fitch WM (2000) Homology, a personal view on some of the problems. Trends Genet 16:227–231

Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458

Goldman N, Whelan S (2002) A novel use of equilibrium frequencies in models of sequence evolution. Mol Biol Evol 19:1821–1831

Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917

Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. Proteins 17:49–61

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Lin K, May AC, Taylor WR (2001) Amino acid substitution matrices from an artificial neural network model. J Comput Biol 8:471–481 [abstract]

Livingstone CD, Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 9:745–756

Miyamoto MM, Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503–513

Müller T, Vingron M (2000) Modelling amino acid replacement. J Comp Biol 7:761–776

Müller T, Spang R, Vingron M (2002) Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 19:8–13

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12:327–345

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Taylor WR (1986) The classification of amino acid conservation. J Theoret Biol 119:205–218

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tourasse NJ, Li WH (2000) Selective constraints, amino acid composition and the rate of protein evolution. Mol Biol Evol 17:656–664

Veerassamy S, Smith A, Tillier ER (2003) A transition probability model for amino acid substitutions from blocks. J Comput Biol 10:997–1010

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Xu W, Miranker DP (2004) A metric model of amino acid substitution. Bioinformatics 20:1214–1221

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z (1995) A space–time process model for the evolution of DNA sequences. Genetics 139:993–1005

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556

Yu YK, Altschul SF (2004) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics 21:902–911

Yu YK, Wootton JC, Altschul SF (2003) The compositional adjustment of amino acid substitution matrices. Proc Natl Acad Sci USA 100:15688–15693

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166