Synonymous Codon Substitution Matrices*

Adrian Schneider, Gaston H. Gonnet, and Gina M. Cannarozzi

Computational Biology Research Group,
Institute for Computational Science, ETH Zürich,
Universitätstrasse 6, 8092 Zürich, Switzerland
cgina@inf.ethz.ch

Abstract. Observing differences between DNA or protein sequences and estimating the true amount of substitutions from them is a prominent problem in molecular evolution as many analyses are based on distance measures between biological sequences. Since the relationship between the observed and the actual amount of mutations is very complex, more than four decades of research have been spent to improve molecular distance measures. In this article we present a method called SynPAM which can be used to estimate the amount of synonymous change between sequences of coding DNA. The method is novel in that it is based on an empirical model of codon evolution and that it uses a maximum-likelihood formalism to measure synonymous change in terms of codon substitutions, while reducing the need for assumptions about DNA evolution to an absolute minimum. We compared the SynPAM method with two established methods for measuring synonymous sequence divergence. Our results suggest that this new method not only shows less variance, but is also able to capture weaker phylogenetic signals than the other methods.

1 Introduction

Measures of synonymous substitution are fundamental for many analyses in molecular evolution such as in the calculation of dN/dS to provide information about the degree of selection operating on homologous sequences, in the construction of phylogenetic trees, and for molecular dating. Measuring synonymous change between coding DNA sequences requires a model of evolution at the codon or nucleotide level. Because DNA evolution has properties such as unequal transition (purine-purine or pyrimidine-pyrimidine mutation) and transversion (purine to pyrimidine and vice versa) rates and unequal nucleotide and codon frequencies, modeling of DNA evolution is not trivial. The first methods to estimate the amount of synonymous mutations were introduced more than two decades ago [1, 2] and have since then continuously been improved [3, 4, 5, 6]. Usually, a 4×4 or a 64×64 Markovian model is employed with varying numbers of rate constants and/or nucleotide or codon frequencies. These models are used to estimate the real number of substitutions from the observed number of

* This work was supported by the intramural research program of the National Institutes of Health, National Library of Medicine.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 630–637, 2006.

substitutions. The models differ in the number of variables needed to describe transitions between nucleotides, the equilibrium frequencies and the definition of 'synonymous site' used.

Two measures of synonymous sequence divergence are the dS metric that estimates the ratio of synonymous substitutions per synonymous site and the NED or TREx metric, based on mutations at the third position of two-fold redundant codons at conserved amino acid sites. The most commonly used implementation of dS is found in the *PAML* software package [7] and is based on the maximum-likelihood method by Yang and co-workers [3, 8]. NED was developed by Benner and co-workers [5, 6] to model the decay of the two-fold redundant amino acids at conserved sites with a first-order rate equation. This method has the advantage that it is very simple and can easily be applied in any molecular analysis. The use of only two-fold codons makes it independent of all the complications introduced by transition/transversion rate differences or mutations at the first base positions. However, the drawback is that there are often not enough of this specific type of mutation to yield statistically significant results.

The SynPAM method presented here employs a 64×64 Markov model with transition rates measured from empirical data to estimate the amount of synonymous change between two sequences of coding DNA using maximum-likelihood. SynPAM is different from the other methods in two aspects: Instead of a parameterized model of codon evolution, an empirically determined matrix [9] is used to assign likelihood scores to all the synonymous mutations in an alignment. Maximum-likelihood scoring matrices in the tradition of Dayhoff [10] are then employed to estimate the synonymous distance of the two sequences. Additionally, the SynPAM method directly measures synonymous change in the form of codon substitutions. This makes it unnecessary to identify the sites that are synonymous (which always introduces additional uncertainty).

2 Methods

Synonymous Substitution Matrix. The synonymous substitution probabilities are derived from an empirical codon matrix M where every entry $M_{a,b}$ is the probability that codon b mutates to codon a [9]. Starting with a codon substitution matrix for a particular evolutionary distance, the probabilities of all nonsynonymous substitutions are set to 0. The remaining probabilities are then rescaled such that all possible codon substitutions between codons coding for a given amino acid sum to 1:

$$M^*_{a,b} = \begin{cases} \frac{M_{a,b}}{\sum_{x \in \mathrm{syn}(b)} M_{x,b}} & \text{if } a \in \mathrm{syn}(b) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

with syn(x) being the set of all codons that are synonymous to x.

From the the empirical codon substitution matrix, a wide range of matrices approximating different evolutionary distances are extrapolated through matrix exponentiation [11], thereby providing substitution matrices representing different evolutionary distances that can all be converted to synonymous substitution

matrices. It is important to state that the exponentiation has to be executed on the full mutation matrix and only then the transformation described above and the computation of the scores should be performed. The reason is that the reduced substitution matrix no longer describes the full Markovian process of codon evolution. E.g. some codons have a larger chance to undergo nonsynonymous mutations than others, but this information would be lost, if only the synonymous matrices were exponentiated. In addition, by including all elements of the substitution matrix until the time of reduction, all alternative pathways, back mutations and multiple hits are taken into consideration.

Analogous to the definition of 1 PAM as the amount of evolution in which 1 percent of the amino acids are expected to undergo mutations, 1 SynPAM is defined as the amount of evolution in which 1 percent of the synonymous positions are subject to a codon substitution.

The substitution matrix $M(t)$ (and derived from it also $M^*(t)$) represents t times the distance of 1 SynPAM and is computed as $M(1)^t$, where $M(1)$ is the full substitution matrix corresponding to 1 SynPAM (which was derived from the original M in an iterative process).

Likelihood Scores. In almost the same way that log-likelihood scores are computed for amino acid scoring matrices, they can be calculated for the synonymous substitution matrices. In addition to the codon frequencies, π, also the synonymous frequencies, π^*, are needed, which is the codon distribution for given amino acids. This means that the sum of all π^*_x which encode the same amino acid is 1. The scores are computed from the ratio of the probabilities of two synonymous codons having mutated from a common ancestor during time t compared to being paired by random chance (which corresponds to $t = \infty$):

$$S_{i,j}(t) = 10 \log_{10} \frac{\pi_j \cdot M^*_{i,j}(t)}{\pi^*_i \cdot \pi_j} = 10 \log_{10} \frac{M^*_{i,j}(t)}{\pi^*_i} \tag{2}$$

Synonymous codon scores are only defined for synonymous substitutions. Because amino acid altering mutations should have no influence on the synonymous scores, all of these mutations are assigned a score of 0.

Given a codon-wise alignment of two coding DNA sequences and a SynPAM matrix for a given distance t, the synonymous score of the entire alignment is obtained by adding up the scores, $S_{i,j}$, for all synonymous codon positions in the alignment. The alignment is scored with a range of SynPAM matrices, the highest-scoring matrix can be determined and the distance of this matrix is chosen as the SynPAM distance of the alignment. This corresponds to a maximum-likelihood estimation of the synonymous distance.

Implementations of NED and dS. The dS values were computed using the *codeml* program from the *PAML* software package [7]. Version 3.14 of the program was used with nine free parameters used to account for codon frequencies (F3x4). Using other models of codon frequencies did not appreciably change the results. Since only pairwise sequence comparisons were done, the parameters were set as follows: runmode=−2 (pairwise comparisons) and the d_N/d_S ratio

was not allowed to vary between sites (NSsites = 0). The other parameters have no influence when pairwise comparison is chosen.

The NED method was implemented in *Darwin* [12]. Given a codon-wise alignment, the fraction $f2$ of conserved codons for all amino acids encoded by two codons is computed. Then the fraction of conserved codons is modeled as an exponential decay to equilibrium via the equation: $NED = -ln\frac{f_2-b}{1-b}$ where b is the codon bias [6] and was originally taken from the Codon Usage Database [13]. Here, the equilibrium fraction of conserved codons was calculated from the codon frequencies from the codon substitution matrix and was found to be .5136.

Databases. The complete genome databases are from *ENSEMBL* [14] (*H. sapiens*, *P. troglodytes*, *C. familiaris*, *B. taurus*, *G. gallus*, *X. tropicalis*, *B. rerio*, *T. rubripes*, *T. nigroviridis*, *C. intestinalis*, *D. melanogaster*, *A. gambiae*, *A. mellifera*, *C. elegans* and *C. briggsae*).

3 Results and Discussion

Distance estimation methods can be assessed either using simulations or real data. Simulations have the advantage that many parameters can be chosen and a detailed analysis of the method can be performed. However, the use of simulations is not appropriate in this case because distance measures are based on a model of evolution while at the same time such a model is needed to create simulated data. Therefore simulations can not be a fair way of comparing distance measures.

The drawback with real data is that the true distance between sequences cannot be known. But at least two criteria can be used when comparing different distance measures: the variance of the estimates and the range in which they yield valid and reliable results.

An unbiased estimation of the variance can be found using each method to measure the distances between many pairs of genes with the same divergence time and then calculating the variance of these different estimates. Sets of gene pairs that fulfill this criteria of equal divergence times are the sets of orthologs between two species. Here, we used the metazoan set from the OMA orthologs project [15], a large-scale cross-comparison of complete genomes with the goal of systematically identifying groups of orthologous proteins.

3.1 Variance of the Estimates

The variation of the different time estimates between the different methods was compared. Because the time metrics of the different methods are on different scales, the variances can not be compared directly. Instead, the ratio of the standard deviation to the average value can be used, which is called the coefficient of variance (CV).

It is clear that since DNA mutations are a stochastic process, different sequence pairs are likely to have different amounts of synonymous mutations. Also, selection pressure on synonymous mutations can not be ruled out [16, 17]. Therefore, although we are looking at orthologous sequences between the same two

Table 1. Ratio of the standard deviation to mean for date estimates between human and animals using different dating methods

	# values	SynPAM	NED	dS
Canis familiaris	13235	0.29	0.32	0.38
Mus musculus	13316	0.23	0.28	0.30
Gallus gallus	8515	0.26	0.32	0.62
Xenopus tropicalis	8288	0.26	0.30	1.61
Brachydanio rerio	5845	0.26	0.33	1.07
Ciona intestinalis	1696	0.59	0.32	0.56
Drosophila melanogaster	1643	1.95	0.38	0.44
Caenorhabditis elegans	1038	1.60	0.33	0.48

species, it is in the nature of these sequence pairs to already have a certain amount of variation in the true sequence divergence. But when this sequence divergence is measured, an additional random component is added because the models of evolution and the ways to estimate the divergence can never precisely reflect the real pattern of molecular evolution.

Since in this comparison the same set of sequence pairs was used for all the tested methods, the inherent variance due to these stochastic (and possibly selection) effects are exactly the same for all methods. The only difference is how much additional noise is introduced by the model of evolution and the measurement technique used.

Table 1 displays these values for distance estimates of human against some other metazoans. For each pair, the coding DNA was aligned using local dynamic programming [18, 19] with the codon substitution matrices. Only alignments with at least 100 synonymous positions were considered in order to gain some significance in the results. For $f2$ values lower or equal to the codon bias b, no NED estimate can be computed. Alignments with this problem were discarded for all methods, therefore the comparison is done on the same set of alignments for all three methods. In addition, for each method, the 10% of alignments with the highest distance estimate were discarded to exclude obvious outliers. The number of sequence pairs used is also shown in Table 1 and decreases with increasing distance because there are less orthologs between distant species. Also, as the distance increases, the synonymous mutations approach their equilibrium values and produce more invalid results.

Table 1 shows that the SynPAM method consistently causes the least amount of variation for species up to the human-fish divergence. Since SynPAM is the estimator with the least amount of variance, it is the preferred method to estimate distances within this time range.

It is remarkable that the CV for dS clearly decreases after the human-frog distance and that it becomes very small for the longer distances. In order to explain this unexpected behavior, we plotted the histograms for all the pairwise values between any two species. A selection of those are shown in Figure 1. There, the explanation of this artifact can be seen: Gene pairs, where the synonymous substitutions are close to saturation, are assigned a seemingly arbitrary dS value around 65. Since this high amount of substitutions per site is impossible to

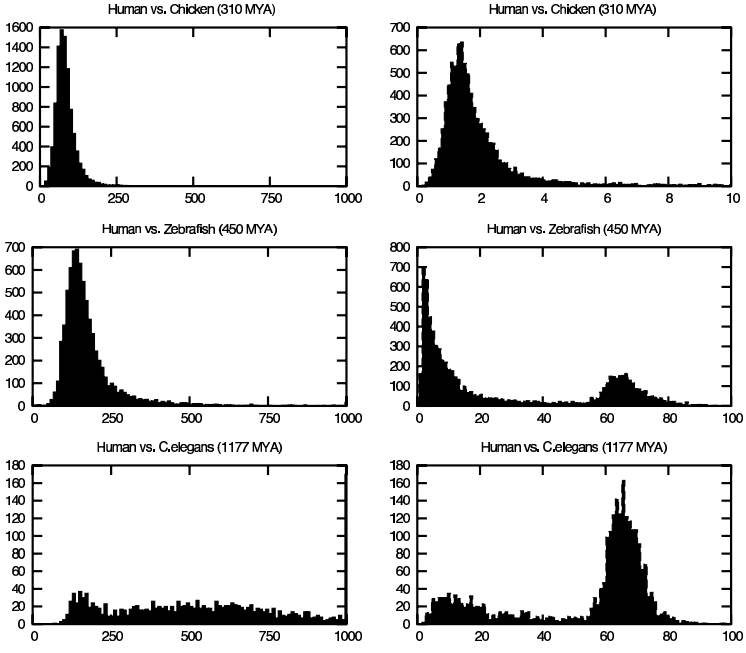

Fig. 1. Histograms for SynPAM (left) and dS (right) of values for the alignments of orthologs between human and other metazoans. There are 9130 alignments of human with with chicken, 7647 with zebrafish and 2512 with *C. elegans*.

estimate with confidence, this is clearly a limitation in the calculation of dS. Already at the primate-fish distance, a significant number of dS values are in this nonsense range, while the SynPAM distribution still forms a sharp curve.

For distances larger than primate-fish (i.e. approximately 450 million years), all methods tend to suffer from saturation effects and the single gene estimates become very unreliable. In the SynPAM histograms, many values of 1000 can be found, which corresponds to the highest matrix used and means that the synonymous mutations have reached an equilibrium state.

4 Conclusions

A new method called SynPAM for the estimation of the amount of synonymous change between two codon sequences has been introduced which is unique in its use of empirical codon transition probabilities combined with a maximum likelihood formalism in the tradition of the Dayhoff scoring matrices. This method has the property that it is works well without estimating mutational parameters. Due to the Markovian model of evolution, multiple hits, alternative pathways and back mutations are not only included in the model, but are also based on empirical observations. Since the distance is expressed in the form of expected number of synonymous codon substitutions, no individual nucleotide sites have

to be assigned as 'synonymous' or not. The direct way of obtaining the synonymous distance makes it faster than other methods while still employing the strength of maximum likelihood estimation.

As the codon mutation matrix was built from vertebrate DNA, its usable range is for the vertebrates and may be extendable to other metazoans. Tests with plants, yeasts and bacteria have so far not been successful, although mutation matrices more specific to specific species sets are under consideration and could improve the performance for these subgroups. The creation of all new matrices for only pairs of species would allow for the incorporation of species-specific substitution patterns without the need to adjust parameters to the actual sequences.

A comparison of SynPAM with dS and NED using large sets of orthologous genes revealed that the SynPAM estimates have lower variances than those from the other methods. Also, their behavior with increasing time is favorable compared to other methods since it captures also weak signals where the synonymous substitutions are close to saturation.

Because synonymous substiutions are under less functional selection pressure than amino acid mutations, they are expected to happen more regularly and are therefore often used for molecular dating, in which clock-like distance measures and known divergence times are used to extrapolate actual time estimates for unknown divergences. Since the SynPAM method presented here for measuring synonymous divergence shows advantages over previous such distance metrics, we propose the use of SynPAM also for dating purposes in the realm of vertebrates and are currently investigating this topic in more detail.

References

1. Miyata, T., Yasunaga, T.: Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16** (1980) 23 – 36
2. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., Dodgson, J.: The evolution of genes: the chicken preproinsulin gene. Cell **20**(2) (1980) 555–566
3. Goldman, N., Yang, Z.: A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**(5) (1994) 725–736
4. Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M.K.: Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155** (2000) 432–449
5. Benner, S.A.: Interpretive proteomics– finding biological meaning in genome and proteome databases. Advances in Enzyme Regulation **43** (2003) 271–359
6. Caraco, M.D.: Neutral Evolutionary Distance: A New Dating Tool and its Applications. PhD thesis, ETH Zürich, Zürich, Switzerland (2002)
7. Yang, Z.: Paml: A program package for phylogenetic analysis by maximum likelihood. CABIOS **13** (1997) 555 – 556
8. Yang, Z., Nielsen, R.: Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**(1) (2000) 32 – 43
9. Schneider, A., Cannarozzi, G.M., Gonnet, G.H.: Empirical codon substitution matrix. BMC Bioinformatics **6**(134) (2005)

10. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model for evolutionary change in proteins. In Dayhoff, M.O., ed.: Atlas of Protein Sequence and Structure. Volume 5. National Biomedical Research Foundation (1978) 345–352
11. Cox, D., Miller, H.: The Theory of Stochastic Processes. Chapman and Hall, London (1965)
12. Gonnet, G.H., Hallett, M.T., Korostensky, C., Bernardin, L.: Darwin v. 2.0: An interpreted computer language for the biosciences. Bioinformatics **16**(2) (2000) 101–103
13. Nakamura, Y., Gojobori, T., Ikemura, T.: Codon usage tabulated from the international DNA sequence database. Nucleic Acids Res. **28** (2000) 292
14. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Birney, E.: Ensembl 2005. Nucleic Acids Res. **33**(suppl-1) (2005) D447–D453
15. Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., Gonnet, G.: OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. In McLysath, A., Huson, D.H., eds.: RECOMB 2005 Workshop on Comparative Genomics. Volume LNBI 3678 of Lecture Notes in Bioinformatics., Springer-Verlag (2005) 61 – 72
16. Bielawski, J.P., Dunn, K.A., Yang, Z.: Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. Genetics **156** (2000) 1299–1308
17. Dunn, K.A., Bielawski, J.P., Yang, Z.: Substitution rates in drosophila nuclear genes: Implications for translational selection. Genetics **157** (2001) 295–305
18. Waterman, M.S., Smith, T.F., Beyer, W.A.: Some biological sequence metrics. Advances in Mathematics **20** (1976) 367–387
19. Gotoh, O.: An improved algorithm for matching biological sequences. J. Mol. Biol. **162** (1982) 705–708