Untangling the Effects of Codon Mutation and Amino Acid Exchangeability

L.Y. Yampolsky and A. Stoltzfus

Pacific Symposium on Biocomputing 10:433-444(2005)

UNTANGLING THE EFFECTS OF CODON MUTATION AND AMINO ACID EXCHANGEABILITY

L. Y. YAMPOLSKY

Department of Biological Sciences,
East Tennessee State University, Johnson City, TN 37614-1710

A. STOLTZFUS

Center for Advanced Research in Biotechnology,
9600 Gudelsky Drive, Rockville, MD 20850

Determining the relative contributions of mutation and selection to evolutionary change is a matter of great practical and theoretical significance. In this paper, we examine relative contributions of codon mutation rates and amino acid exchangeability on the frequencies of each type of amino acid difference in alignments of distantly related proteins, alignments of closely related proteins, and among human SNPs, using a model that incorporates prior estimates of mutation and exchangeability parameters. For the operational exchangeability of amino acids in proteins, we use EX, a measure of protein-level effects from a recent statistical meta-analysis of nearly 10,000 experimental amino acid exchanges. EX is both free of mutational effects and more powerful than commonly used "biochemical distance" measures (*1*). For distant protein relationships, mutational effects (genetic code, transition/transversion bias) and operational exchangeability (EX) account for roughly equal portions of variance in off-diagonal values, the complete model accounting for $R^2 = 0.35$ of the variance. For human/chimpanzee alignments representing closely related proteins relationships, mutational effects (including CpG bias) account for 0.52 of the variance; adding EX to the model increases this to 0.67. For natural variation in human proteins, the variance explained by mutational effects alone, and by mutational effects and operational exchangeability are, respectively, 0.66 and 0.70 for SNPs in HGVBase, and 0.56 and 0.60 for disease-causing missense variants in HGMD. Thus, exchangeability has a stronger relative effect for distant protein evolution than for the cases of closely related proteins or of population variation. A more detailed model for the hominid data suggests that 1) there is a threshold in EX below which substitutions are highly unlikely to be accepted, corresponding to roughly 30 % relative protein activity; 2) selection against missense mutants is a slightly convex function of protein activity, not changing much as long as protein activity is low; and 3) the probability of disease-causing effects decreases nearly linearly with EX.

1. Introduction

The evolution of molecular sequences, including the change of one amino acid for another, appears to reflect a two-step process of mutation and fixation, in which first, a mutation introduces a new allele into a population, and second, the allele rises to fixation by some combination of drift and selection. In the case of population variation, such as missense SNPs, the fixation process has not gone to completion. In either case, the observation of an inter-specific difference, or an intra-specific variant, reflects these two primary factors, mutation and natural selection. For missense changes, or missense variants, the mutational factor is the rate of mutation from one DNA codon to another, while the primary selective factor is the operational exchangeability of the amino acids in their protein context. Other potentially relevant fitness consequences of such

a change are differences in the metabolic costs of amino acids, or in translation efficiencies of codons, but here such effects are assumed to be secondary to the effect of an exchange on the operation of a protein.

The goal of this work is to make a preliminary estimate of the relative contributions of mutational and selective factors to observed amino acid changes in protein evolution, as well as to observed amino acid variation in the human population (i.e., missense SNPs). These effects can be confounded easily. Therefore, *to be effective, any effort to untangle the relative contributions of mutation and selection must rely on some predictive model that incorporates prior knowledge of parameter values.* Mutation parameters are available from the comparative analysis of pseudogene divergence (*2*).

Although measures of amino acid similarity or distance have existed for a long time (*3*), the concept of a measure of the exchangeability of amino acids in a protein context is problematic. In the past, analyses that call for such a measure (*4-6*) have relied on so-called "biochemical distances" (*7, 8*). However, these are not pure measures of amino acid exchangeability, but attempts to fit observed propensities of evolution using a small number of biochemical parameters, as is clear from the original work of Grantham (*7*). With respect to the present problem of untangling mutational and selective effects, one cannot use an exchangeability measure that is based on observed propensities of evolution, because this would confound precisely the effects that must be treated separately. Therefore, the analysis here uses EX, a measure of relative effect on protein activity (derived from a statistical meta-analysis of nearly 10,000 experimental amino acid exchanges) that is free of mutational effects and which, in tests of power, out-performs "biochemical distance" measures (*1*).

We first analyze amino acid differences in variation and evolution using a formal *ad hoc* statistical model in which mutational biases and amino acid exchangeability act as factors. This approach allows us to assess the relative contributions of mutational effects and fitness effects in accounting for patterns of amino acid sequence differences in data sets representing close and distant inter-specific divergence, and in data sets representing intra-specific variation in humans (i.e., missense SNPs), including disease-associated variants in HGMD, as well as the general population sample in HGVBase. Then, we analyze a more specific set of models that includes the CpG mutational effect, applied to the data on human intra-specific variation, as well as to data from human/chimp divergence.

2. Methods

2.1. ***Sources and treatment of data on divergence and variation.*** The BLOSUM series of matrices used to represent distant protein divergence were taken from the supplementary material to (9), which provides five digits of precision in log-of-odds ratios (s_{ij}). Close divergences are represented by the set of human-chimpanzee alignments of over 7000

coding regions from Clark, et al. (10), using mouse as the out-group to polarize substitutions, and including only those codons in which all three sequences are known, polarization is non-ambiguous, and the gene alignment includes the start codon. The resulting set of data comprises 821,180 codons, representing roughly 5% of hominid genome. These dataset will be referred to as HomoPan dataset. Human polymorphism data came from two databases, the "Proven" subset of SNPs from HGVBase (11) and the disease-associated missense variants from HGMD (12). Both HGVBase and HGMD frequencies were divided by frequencies of source amino acid in human coding regions inferred from a codon usage database (13). Combined samples sizes are 1628 and 15373 for HGVBase and HGMD, respectively.

2.2. ***Prior estimates of parameters of the prediction model.*** Estimates of mutation parameters for hominids are taken from (2). To represent the frequencies of CpG sites in hominid genes, including CpG sites that straddle adjacent codons, we computed the frequency distribution of pentamers consisting of a codon with the 5' and 3' flanking nucleotides, from the complete set of coding sequences in the human RefSeq standard (14), omitting any entries with non-canonical start or stop codons, or with nucleotide ambiguities. The exchangeability of amino acids is parameterized in terms of the EX measure of Yampolsky and Stoltzfus (1). Since this measure is not well known, we describe it briefly. EX is based on a statistical meta-analysis of published data on the effects of 9671 amino acid changes in experimental studies carried out on 12 different proteins. Data on mutant protein activity from a subset of the studies provides the basis for a model of the frequency distribution of effects on protein activity; this model is then used to assign scores on a common scale for all of the exchanges. Taken literally, an EX_{ij} value of 0.42 means that, on average, a variant protein with a residue j replacing the wild-type residue i has 42 % of the activity of the original protein. The mean value of EX is 0.28. EX out-performs Grantham's distances and Miyata's distances in an unbiased test of the ability to predict effects of experimental exchanges, and in a test that incorporates a measure of amino acid distance into the mutation-acceptance model of (15) implemented in the PAML package (5).

2.3. ***Statistical analysis.*** The mutational effects that are considered are the effect of the genetic code in imposing a minimum number of mutational steps ("minimum mutational distance") of 1, 2, or 3 (16), which we refer to here as "singlet", "doublet" and "triplet" exchanges; the effect of a transition/transversion bias; and for hominid data, the effect of a CpG context. Transition/transversion and CpG biases are considered only within singlet exchanges. "Transition" factor is assigned to a level of 1 for any singlet exchange that can occur by a transition, and a level of 0 otherwise. The "CpG" factor was a continuous factor represented by the combined frequency of all CpG containing codons and cGNNn and nNNCg pentamers

(codons with flanking neighbors) among the codons of a given source amino acid that can mutate in a single step to each codon of the destination amino acid. For pairs of amino acids that cannot mutate into each other by a mutation at a CpG site such frequency is 0.

Since the overwhelming majority of hominid data are singlet differences, we performed two types of analyses. First, we included the genetic code effect, treating lack of observations of doublet and triplet exchanges as zero frequencies. Second, we considered only singlet exchanges, considering only transition/transversion and CpG effects among such exchanges.

Thus, the statistical analysis includes two groups of factors, mutational effects and exchangeability (EX), and four sets of response variables, representing frequencies of amino acid differences in distantly related proteins (BLOSUM30 through 100), in closely related proteins (human-chimpanzee alignment data), among human missense SNPs (HGVBase data) and among disease-associated human missense SNPs (HGMD data).

GLM models including these factors and their interactions were evaluated using JMP statistical package (*17*). For each test, the first model includes only the genetic code effect, then we add transition/transversion bias and (for human data) CpG bias, and finally, EX and its interactions. R^2 values associated with each model are the measure of relative contribution of each factor to the variance in the response variable.

For the case of hominid data (human-chimp divergence, and human missense variation), we also consider a more sophisticated model that takes into account the relevant target size for each individual mutational path, considering enhanced rates of transition and transversion at CpG sites. The target size is simply the relative frequency of codons that participate in a particular mutational path from one amino acid to another, e.g., for the Val-to-Leu change there is some subset of GTN Val codons that are preceded by a C, and thus are subject to an enhanced rate of mutation from GTN to CTN, specifically the transversion rate at CpG sites.

For each of three types of human data a simple model predicting substitution frequencies has been constructed. For human-chimpanzee substitutions we assume that differences are proportional to rates of change, which are in turn described by an origin-fixation process with a rate equal to the rate of mutational origin multiplied by the probability of fixation (18). Then the occurrence of some type of difference is proportional to

$$P_{fix}\,\mu\,(T_{11} + T_{12}t + T_{21}c_v + T_{22}c_t), \qquad (1)$$

where P_{fix} is the unscaled mean probability of a given type of mutant being fixed; μ is the mutation rate for non-CpG transversions; t is transition/transversion bias; c_t and c_v are the biases in transitions and transversions, respectively, at CpG sites; and T_{11} ,T_{12}, T_{21} and T_{22}, respectively, are the target sizes representing the sums over the frequencies of codons that, when subjected to each kind of mutation (non-CpG transversions, non-CpG transitions, CpG transversions and CpG transitions) produce the amino acid

exchange of interest. The values of mutational biases used were: t =2.4; c_t = 23.0, and c_v = 7.0 (*2*). A logistic function was used to describe relationship between P_{fix} and EX:

$$P_{fix} = k/(1+ exp(-a(EX-b)). \quad (2)$$

This constrains the function to be between 0 and 1, and to increase, but allows it to take nearly any shape. The meaning of parameters a and b is steepness of the curve and location of the inflection point, respectively. The meaning of k is simply the number of generations since the common ancestor (twice that number for both lineages combined). This analysis has been done for substitution frequencies in human and chimpanzee lineages separately and for both lineages combined.

An essentially identical model has been used to fit HGMD data, only instead of the probability of being accepted, the HGMD model has the probability of having severe effects, assumed to be related to EX through a logistic function, this time a decreasing one:

$$P_s = k - k/(1+ exp(-a(EX-b)). \quad (3)$$

The parameter k here is a scaling factor reflecting how well the human populations have been screened for deleterious variants. A slightly different model was utilized for HGVBase data. First, we assume that majority of known human SNPs are recessive deleterious mutations segregating at mutation-selection balance, i.e., their frequencies are at $\sqrt{(\mu/s)}$, where s is selection against the mutant variant (18). Second, we assume that the probability of finding a variant is proportional to its frequency. This representation of ascertainment bias is reasonable when most SNPs do not have any clinically important effects and are discovered in population genetics or genomic screens. This seems to be the case for SNPs in HGVBase (*11*). Then, the probability of observing a change is proportional to

$$k\sqrt{\mu/s}\,(T_{11} + T_{12}\sqrt{t} + T_{2n}\sqrt{c_v} + T_{22}\sqrt{c_t}) \quad (4)$$

It is reasonable to assume that selection against mutant variants with relative activity equal to that of the wild type is 0, so the function describing the relationship between s and EX must contain (0,1) point. Thus, logistic function cannot be used. Instead, a power function has been used:

$$s = b\,(1 - EX^a). \quad (5)$$

The quantity in Equation 1 was fitted to the observed frequencies of substitutions in human and chimpanzee lineages and to HGMD frequencies, substituting Equations 2 and 3 for P_{fix} and P_s, respectively. Quantity (4), substituting (5) for s, was fitted to observed frequencies in HGVBase. For HGMD and HGVBase data, target size was recalculated for the entire genome, assuming that sampled 821180 codons in human/chimpanzee alignment represent 5% of the entire length of coding regions and 25% of Clark at al alignments. All fitting was done in the non-linear fit platform in JMP (*17*) using the least-squares model. Profile likelihood confidence intervals were calculated

iteratively when possible; when there was no convergence, approximate standard errors were calculated by the derivative cross-product inverse matrix method (*17*).

3. Results

Contributions of mutational biases, amino acid exchangeability and their interactions to the explained portion (R^2) of variance among amino acid substitution rates are shown in Figure 1. Adding EX to the model more than doubles R^2 for distant protein substitutions represented by BLOSUM log-of-odds values (fig. 1A), raising it from around 0.15 to about 0.35. Note that the variance of off-diagonal values explained by this model increases monotonically from BLOSUM30 to BLOSUM100 (an issue addressed further below). EX alone explains much smaller portion of the variance of substitution frequencies: 0.036, 0.002 and 0.076 for HGVBase, HGMD and HomoPan frequencies, respectively and 0.12-0.20 for BLOSUM. It is worth mentioning that EX is a lot more successful in predicting the ratio between HGMD and HGVBase frequencies ((*1*)), explaining nearly 50% of the variance. This is because the contributions of mutational biases (presumably identical in both datasets) cancel out, leaving the exchangeability effect intact. Table 1 provides ANCOVA results for BLOSUM62 (the matrix most commonly used for protein alignments).

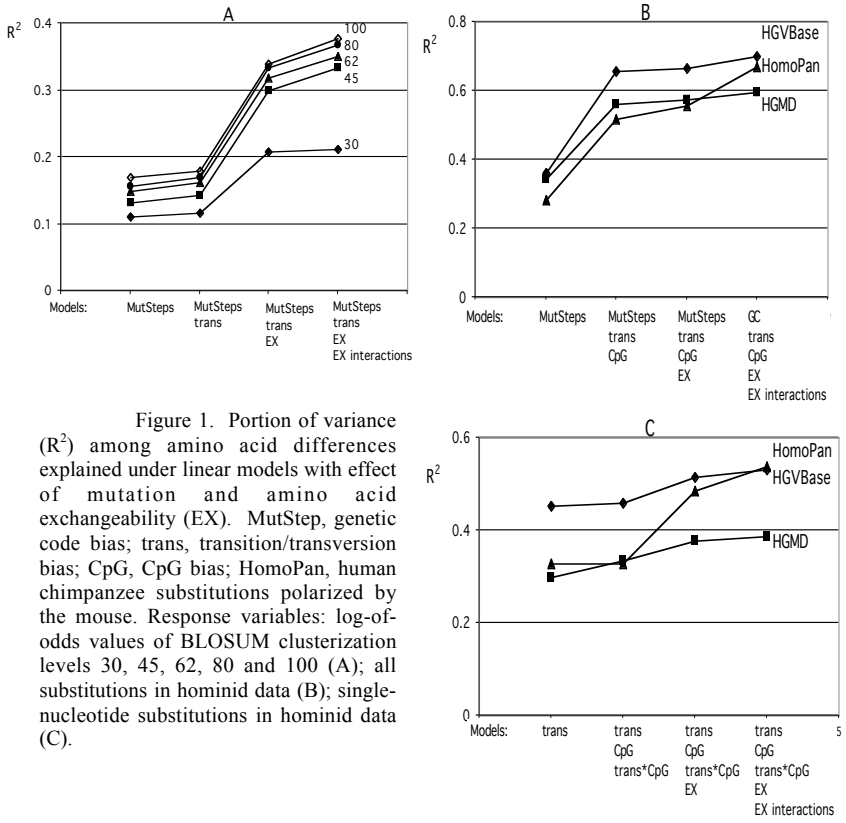

Figure 1. Portion of variance (R^2) among amino acid differences explained under linear models with effect of mutation and amino acid exchangeability (EX). MutStep, genetic code bias; trans, transition/transversion bias; CpG, CpG bias; HomoPan, human chimpanzee substitutions polarized by the mouse. Response variables: log-of-odds values of BLOSUM clusterization levels 30, 45, 62, 80 and 100 (A); all substitutions in hominid data (B); single-nucleotide substitutions in hominid data (C).

Table 1. ANCOVA of the effects of mutation biases and EX on log-of-odds score of BLOSUM62, with and without interactions. Abbreviations as on Fig. 1A. P values < 0.015 in bold.

Response: BLOSUM62	Without interactions			With interactions		
Source	df	F	P	df	F	P
MutSteps	2	19.1	**1.0E-08**	2	1.4	0.24
trans	1	4.2	0.04	1	2.1	0.15
EX	1	90.0	**3.0E-19**	1	57.6	**3.0E-13**
EX*MutSteps				2	8.1	**0.0004**
EX*trans				1	0.7	0.39
Error	369			366		

Transition/transversion bias contributes strongly to the power of the model when applied to close (hominid) divergence data and human variation data (fig. 1B), unlike the case for distant divergence data. Most of the increases in R^2 values are due to the transition/transversion factor, while the CpG factor adds very little (data not reported). Adding EX and its interactions with mutational biases to the model improves the ability of the model to predict SNP frequencies surprisingly little, although the effect of EX is highly significant (Table 2). There is, however, a large increase in predicting power of the model attributable to EX when the response variable is frequency of human/chimpanzee substitutions, particularly when single-nucleotide substitutions alone are considered.

Although incorporating the interactions between mutational biases and EX to the models adds little to the explained variance, some of these interactions are significant and of interest (Tables 1 and 2). In particular, it is striking that well known genetic code component of BLOSUM scores statistically is made up entirely of the interaction with amino acid exchangeability. The nature of this interaction is the presence of a strong EX effect among single nucleotide substitutions, a weaker effect among double-nucleotide substitutions and lack of such effect among triple-nucleotide substitutions. Interaction between EX and transition/transversion bias is ubiquitous and highly significant in hominid data (Table 2). The nature of this interaction is that there is a covariance between observed frequency of substitutions and EX for substitutions that can occur through a transition, but is absent or much weaker in the group of substitutions that can only occur through a transversion.

Results of fitting non-linear models to hominid data are shown on Figure 2. The probability of fixation as a function of EX appears to have a critical range of mutant protein activity in which the drop of P_{fix} occurs quickly (Fig. 2A). This inflection point is located around EX = 0.3. Fig. 2A shows the fit to combined substitution frequencies in both human and chimpanzee lineages; parameters of the model fitted to these lineages separately are not significantly different from each other or from the combined date fit. The curve

for the chimpanzee is slightly less steep and inflection point is shifted slightly towards higher EX values, possibly indicating stronger stabilizing selection in chimpanzee lineage than in human one. The third parameter of the fitted model, K, has the meaning of the number of generations since the common ancestor. The best fit for this parameter is, assuming baseline mutation rate of $5*10^{-9}$, 250,000 for humans and 300,000 for chimpanzee. Assuming $5*10^{6}$ my since the common ancestor, this corresponds to generation times of 20 years for humans and 17 years for chimpanzees, a remarkably meaningful estimate.

Table 2. ANCOVAs of the effects of mutation biases and EX (continuous variable) on frequencies of single-nucleotide amino acid substitutions in hominid data, with and without interactions. Abbreviations as on Fig. 1C.

Source	Without interactions			With interactions		
Response: HGMD	df	F	P	df	F	P
trans	1	46.1	**2.7E-10**	1	23.1	**4.0E-06**
CpG	1	2.2	0.14	1	2.0	0.16
trans*CpG	1	8.0	**0.0053**	1	0.023	0.88
EX	1	19.9	**1.7E-05**	1	3.6	0.058
EX*trans				1	7.6	**0.007**
EX*CpG				1	0.79	0.37
EX*trans*CpG				1	0.44	0.50
Error	145			142		
Response: HGVBase	Df	F	P	df	F	P
trans	1	74.6	**9.5E-15**	1	0.44	0.51
CpG	1	1.9	0.17	1	0.0034	0.95
trans*CpG	1	0.4	0.52	1	2.3	0.13
EX	1	18.5	**3.1E-05**	1	3.3	0.072
EX*trans				1	4.7	0.032
EX*CpG				1	0.2	0.64
EX*trans*CpG				1	3.2	0.076
Error	145			142		
Response: HomoPan	df	F	P	df	F	P
trans	1	66.1	**1.7E-13**	1	6.3	**0.014**
CpG	1	0.16	0.69	1	0.02	0.89
trans*CpG	1	0.004	0.95	1	1.6	0.20
EX	1	44.1	**5.8E-10**	1	8.1	**.0052**
EX*trans				1	30.8	**1.4E-07**
EX*CpG				1	0.03	0.87
EX*trans*CpG				1	1.2	0.28
Error	145			142		

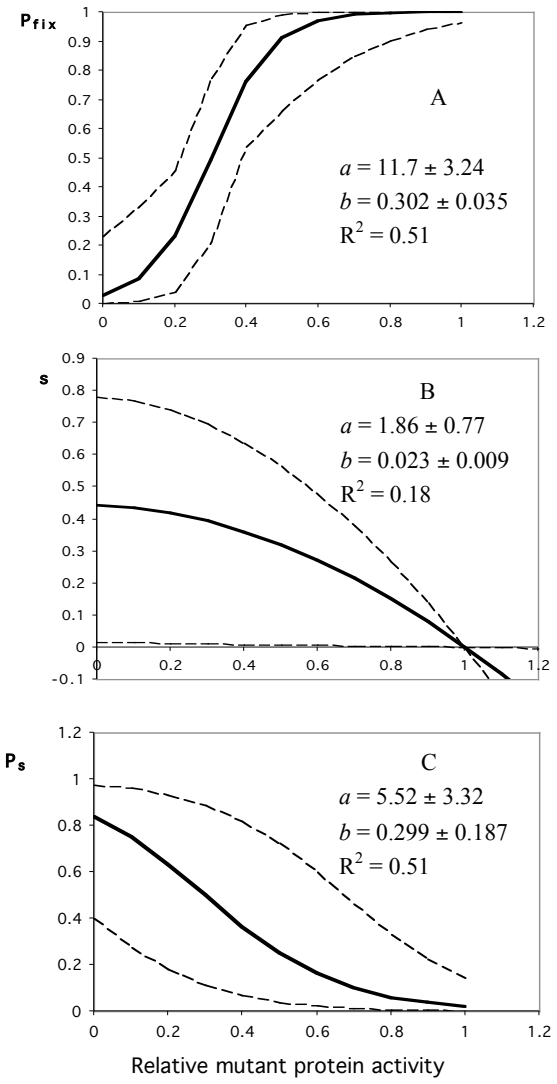

Figure 2. Probability of evolutionary acceptance (P_{fix}), mean selection coefficient (s) and probability of having clinically severe effects (P_s) as functions of relative mutant protein activity. Dashed lines, 95 % confidence intervals.

The shape of the relationship between selection coefficient and EX estimated from HGVBase frequencies under the assumption of mutation selection balance is slightly convex (fig. 2B). 95% confidence intervals are quite

large, so this result needs to be taken with caution. The value of *b* reported on fig. 2B is an arbitrary value calculated for the scaling factor $k = 0.1$. The relationship between *b* and *k* is such it is a possible to estimate *k* from very broad limits on *b*. For $b = 1$ (all loss of function mutations lethal, clearly the upper limit of *b*) $k = 0.15$. For $b = 0.01$ (probably a gross underestimate of selection against loss of function mutations) $k = 0.015$. In other words, the set of "proven" SNPs in HGVBase constitutes somewhere between 1.5 and 15% of all non-synonymous SNPs in existence.

The results of fitting the model to HGMD frequencies are shown in Figure 2C. Again, as for HGVBase data, the confidence intervals are quite high, but one can conclude that P_s decreases more or less linearly from over 0.5 for mutations with the lowest protein activities to about 0.2 for the ones with relative activity 0.5 of the wild type or more. Mean P_s (averaged across all amino acid substitutions types) is 0.52 (standard dev. 0.14).

4. Discussion

Mutational biases and amino acid exchangeability have roughly equal effects on frequencies of amino acid substitutions among distantly related proteins, while on a smaller evolutionary scale, mutational biases add a relatively higher portion to the amount of variance explained by the model. The effect of amino acid exchangeability is seldom a pure effect, acting most often through interactions with mutational biases. Specifically, exchangeability matters for singlet exchanges more than for doublet and triplet exchanges, and for transitions more than for transversions. The first of these two interactions is easy to explain. While singlet differences may often reflect a single origin-fixation event, doublets and triplets probably only do so rarely, resulting instead from multiple changes such that the exchangeability of the original source and final destination amino acids is largely irrelevant.

This interaction is observed in all BLOSUM matrices except BLOSUM30 and is illustrated by Figure 3 showing regression coefficients of the effect of EX on off-diagonal BLOSUM values. Regression of the EX effect on singlet pairs of amino acids monotonically and significantly increases when the BLOSUM level is increased from 30 to 100; this increase is less significant for doublet exchanges and is reversed (though insignificantly) for triplet exchanges.

There is, therefore, also an interaction between the strength of the effect of EX on off-diagonal BLOSUM elements and the level of BLOSUM clustering, with the effect being strongest in the BLOSUM matrix that represents all degrees of relationship among aligned sequences (BLOSUM100), and the weakest in the matrix based only on sequences that are 30 % similar or less (BLOSUM30; (*9*)). Presumably, the reason for this is that a difference in closely related proteins (these being more strongly emphasized in BLOSUM100 than in BLOSUM30) is a difference in a nearly identical protein context, thus the pattern of occurrence of differences in closely related proteins is restricted to amino acids that are compatible with the same contexts, whereas for distantly related proteins, the context has been degraded so that it is no longer so well

shared, but instead each protein is exploring a different region of the possibility-space for the family, so that there are essentially more degrees of freedom in the pattern of divergence.

It is much more difficult to explain why exchangeability has a greater effect on singlet differences that arise via transitions than on those that can only arise via transversions. If this result were found only in a single dataset, we would be inclined to treat it as an artifact of smaller sample size of transversions. But this effect is present in all three hominid datasets, including the very large human-chimpanzee alignments dataset. At present, we don't have a biological explanation for this observation.

More detailed models connecting observed frequencies of amino acid substitutions with mutational biases and amino acid exchangeability can yield information about the shape of relationship between mutant protein activity and how human medicine (HGMD data) or stabilizing selection (HGVBase and HomoPan data) perceive the strength of the deleterious effect such mutations. Of the three datasets analyzed, only the HomoPan dataset has enough data to yield both reasonably high R^2 of the fitted model and tangible narrow confidence intervals around the best fit. Fitting an evolutionary model to this dataset results in the conclusion that there is a relatively sharp transition from mutations with low probability of fixation to those with a high probability, and that the critical value of mutant protein activity is approximately 30% of the wild type activity. Note that the fitted logistic function was not forced to be close to 0 when EX is low or close to 1 when EX is high.

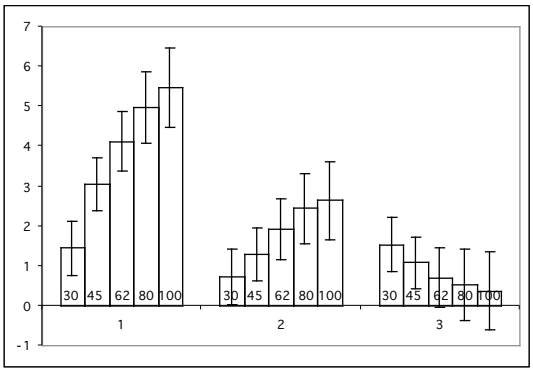

Figure 3. Regression coefficients (slopes) of the regression of BLOSUM s_{ij} values on EX for singlet, doublet, and triplet exchanges. BLOSUM clusterization levels of 30, 45, 62, 80 and 100 % identity cut-off are used.

For the model fitted to HGVBase data, the main assumption is that all substitutions are recessive deleterious alleles at mutation-selection balance. Clearly, since some portion of known SNPs (unknown and probably over-represented) are entirely neutral (or epistatically deleterious) alleles segregating at much higher frequencies, and others almost certainly have some co-dominant deleterious effects and therefore segregate at much lower frequencies, the results

of the fit based on mutation-selection balance is almost certainly biased and it is impossible to estimate whether these two violations of the assumptions tend to compensate for each other. Therefore, it is not surprising that, of the three models, the mutation-selection balance model fitted to HGVBase data has the lowest R^2 values and the broadest confidence intervals.

The model fitted to HGMD data is free from the assumptions about allele frequencies, since each substitution at a given site is reported at HGMD only once. Just as in the case of hominid model, the logistic function shown on fig. 2C was not forced to be close to 1 when EX is small or to be close to 0 when EX is large. Although it is hard to make conclusions about the exact shape of the function relating P_s to EX, there is clearly a strong effect.

5. Acknowledgements

This work was supported by the East Tennessee State University, and by the Center for Advanced Research in Biotechnology (a research institute jointly supported by the National Institute of Standards and Technology and the University of Maryland Biotechnology Institute). The identification of specific commercial software products in this paper is for the purpose of specifying a protocol, and does not imply a recommendation or endorsement by the National Institute of Standards and Technology.

6. References

1. L. Y. Yampolsky, A. Stoltzfus, (in review).
2. I. Ebersberger, D. Metzler, C. Schwarz, S. Paabo, *Am J Hum Genet* **70**, 1490-7 (Jun, 2002).
3. P. H. A. Sneath, *Journal of Theoretical Biology* **12**, 157 (1966).
4. M. Krawczak, D. N. Cooper, *Hum Mutat* **8**, 23-31 (1996).
5. Z. Yang, R. Nielsen, M. Hasegawa, *Mol Biol Evol* **15**, 1600-11. (1998).
6. D. Vitkup, C. Sander, G. M. Church, *Genome Biol* **4**, R72 (2003).
7. R. Grantham, *Science* **185**, 862-864 (1974).
8. T. Miyata, S. Miyazawa, T. Yasunaga, *J Mol Evol* **12**, 219-36 (Mar 15, 1979).
9. S. Henikoff, J. G. Henikoff, *Proc Natl Acad Sci U S A* **89**, 10915-9. (1992).
10. A. G. Clark *et al.*, *Science* **302**, 1960-3 (Dec 12, 2003).
11. D. Fredman *et al.*, *Nucleic Acids Res* **30**, 387-91 (Jan 1, 2002).
12. P. D. Stenson *et al.*, *Hum Mutat* **21**, 577-81 (Jun, 2003).
13. Y. Nakamura, T. Gojobori, T. Ikemura, *Nucleic Acids Res* **28**, 292 (Jan 1, 2000).
14. K. D. Pruitt, D. R. Maglott, *Nucleic Acids Res* **29**, 137-40 (Jan 1, 2001).
15. N. Goldman, Z. Yang, *Mol Biol Evol* **11**, 725-36 (1994).
16. W. M. Fitch, *J. Mol. Biol.* **16**, 9-16 (1966).
17. SAS_Institute, *JMP ® Statistics and Graphis Guide* (SAS Institute, Cary, NC, 2002).
18. J. F. Crow and M. Kimura. An Introduction to Population Genetics Theory. Burgess Publ., Minneapolis, 1970.