Enzyme Evolution Explained
(Sort Of)

ANTONY M. DEAN

BPTI, 240 Gortner Laboratories, 1479 Gortner Ave,
St. Paul, MN 55108

G. BRIAN GOLDING

Department of Biology, McMaster University,
Hamilton, ON L8S 4K1, Canada

Sites in proteins evolve at markedly different rates; some are highly conserved, others change rapidly. We have developed a maximum likelihood method to identify regions of a protein that evolve rapidly or slowly relative to the remaining structure. We also show that solvent accessibility and distance from the catalytic site are major determinants of evolutionary rate in eubacterial isocitrate dehydrogenases. These two variables account for most of the rate heterogeneity not ascribable to stochastic effects.

1 Introduction

A great deal of thought and effort has been spent investigating heterogeneity in the rates of molecular evolution, partly because the phenomenon has important implications for reconstructing the historical relationships among various taxa, and partly because it provides an all-important window through which we can glimpse the action of underlying evolutionary mechanisms.

Rates of molecular evolution are ultimately determined by the rates of mutation. These vary substantially across taxa. Drake[1,2] reports that experimentally determined rates of spontaneous nucleotide substitution (per site per generation) range from a low of 2 x 10^{-11} in *Neurospora crassa*, through 5.4 x 10^{-10} for *Escherichia coli*, to a high of 1.5 x 10^{-3} in *Qβ*. The latter, an RNA bacteriophage, lacks the antimutagenic processes of proofreading and mismatch repair common to most double stranded DNA genomes.

Observed rates of molecular evolution are greatly modified from the underlying rates of mutation, as a quick comparison of the units with which these processes are measured will testify - rates of evolution are presented as substitutions per site per *year*; rates of mutation are presented as substitutions per site per *generation*. Precisely why the rate of molecular evolution should appear so regular on an annual basis, rather than on a per generation basis, remains a mystery. Kimura[3] speculated that if the deleterious effects of new mutations were Γ distributed and that if generation time is inversely proportional to the square root of the variance effective population size, then the rate of molecular evolution per site per year should be approximately constant given a constant rate of mutation. Ochmann and Wilson[4]

approached the problem empirically, arguing forcibly that data from silent substitutions provided evidence for a universal annual molecular clock. Others[5,6] denied the very existence of a clock. While it is now evident that viruses evolve at rates orders of magnitude faster than cellular organisms[7], it is equally evident that closely related species display broadly similar rates of evolution[8]. This is not to deny, however, that rates of evolution can vary as witnessed by the slow down in lineages leading to humans[9].

Heterogeneity in rates of molecular evolution is even more apparent when different genes, or different parts of genes, are compared. Following the precedence set by Kimura[10] differences in rates of evolution are now commonly interpreted in terms of constraints placed on neutral evolution. Mutations in regions of little functional importance are deemed less likely to be deleterious than those in regions of critical functional importance. Given equal neutral mutation rates, the former regions evolve faster than the latter, which are said to be more constrained. For example, pseudogenes, their expression silenced by mutation, evolve slightly faster than introns, 3'-flanking regions of genes and the four-fold degenerate sites of coding regions. These, in turn, evolve rather faster than 5'-flanking regions and two-fold degenerate sites which, in their turn, evolve much faster than nondegererate sites[7]. The latter are considered the most constrained because they determine the amino acid sequences of proteins whose functions are directly subject to natural selection.

Invoking constraint risks abuse. The frequent replacement of one hydrophobic amino acid by another in the lipid binding domains of mammalian apolipoproteins has been taken as evidence of a lack of structural constraints[11] which, in turn, have been offered in explanation of the high rate of amino acid replacement in these genes. The concept of constraint carries force only when supported by evidence garnered independently of evolutionary argument so that tautology can be avoided.

Constraints are not the only mechanism offered in explanation of rate heterogeneity. Fisher[12] long ago argued that mutations with smaller phenotypic effects are more likely to be selectively advantageous. A modern interpretation of his argument would suggest that the rapid evolution seen in four-fold degenerate sites within coding regions might be related to selection acting on subtle differences in expression or translational accuracy, and that the slow evolution seen at nondegenerate sites is merely a consequence of mutations with larger phenotypic effects being less likely to be advantageous.

These hypotheses are, however, without the quantity of experimental support that is necessary for their validation. It our purpose here to examine some of the methods used to identify rate heterogeneity, to determine how much of this heterogeneity can be explained by different hypotheses and to use the tertiary structure of the molecules to examine what this may mean for the underlying evolutionary mechanisms.

2 Phylogeny

Isocitrate dehydrogenases (IDH) are ubiquitous in nature, catalyzing the oxidation of isocitrate to α-ketoglutarate and CO_2 with concomitant reduction of either NAD or NADP. Many organisms lacking Krebs' cycle for energy production nevertheless retain its IDH to produce the α-ketoglutarate so essential for glutamate biosynthesis, this pathway being the principle means by which nitrogen, in the form of ammonia, becomes incorporated into organic matter.

IDHs belong to an ancient and diverse family of enzymes which include isopropylmalate dehydrogenases, tartrate dehydrogenases, homoisocitrate dehydrogenases and several unidentified orfs[13]. Eubacterial IDHs are approximately 400 amino acid residues long, share at least 40% amino acid sequence identity, with only a few gaps here and there in the alignment generated using ClustalV[14]. Phylogenetic trees constructed using neighbor joining and maximum parsimony have similar topologies. Most eubacterial IDHs belong to a single monophyletic clade (Figure 1), the only exceptions being those of *Mycobacteria*, which more closely resemble the eukaryotic cytosolic IDHs (not shown), and that of *Rickettsia prowazekii* which more closely resembles the eukaryotic mitochondrial IDHs.
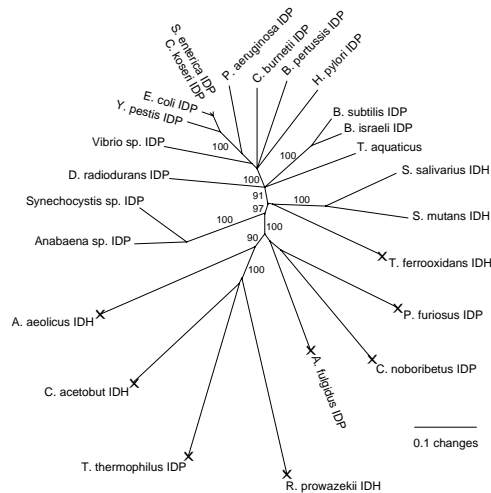

Figure 1—Neighbor joining phylogeny of the eubacterial IDH clade. Bootstrap values are the percentages from 1000 replicates. Lineages with long (> 0.2) branch lengths (5) were removed prior to estimating the number of replacements per site. Parsimony produces a similar tree overall, although the precise branch order of some of the longest lineages differs.

The numbers of amino acid replacements per site in IDH were determined using PAUP. Long branches (> 0.2) were removed from the analysis to avoid the risk of inferring that only single replacements had occurred at a substantial number of sites where multiple events had occurred. Results obtained with neighbor joining and

maximum parsimony trees were very similar (r > 0.99) and further analyses were confined to the distribution of replacements obtained with neighbor joining.

The X-ray structures of *Escherichia coli* IDH are extraordinarily similar to those of isopropylmalate dehydrogenase, a distantly related member within the protein superfamily[15]. This implies a key assumption in our analysis, that secondary structure has not evolved within the eubacterial IDH clade, is robustly justified.

3 Finding the Hot Spots and the Cold Spots

The expected distribution of replacements across sites would be Poisson were all sites to evolve at the same rate. The ratio of the observed variance (s^2_{obs}) to the observed mean ($\overline{y}$), widely known as the coefficient of dispersion, provides the basis for a convenient test[16] for rate heterogeneity since $\chi^2_{n-1} = (n - 1) s^2_{obs}/ \overline{y}$. For virtually all proteins this is significant - indeed a lack of significance would indicate a highly unusual protein.

Detecting significant rate heterogenaities and understanding their cause are two very different matters. We are developing a means to help locate regions in proteins that evolve at unusually fast and slow rates. The method is based on an empirical evaluation of the likelihood of observing replacement rates in small regions throughout the molecule. The likelihoods are calculated using the method of Felsenstein[17], for a given sequence alignment and a given phylogeny. The likelihood of state i at node l in a phylogeny is calculated as

$$L^l_i = (\Sigma_j P^{t_r}_{ij} \cdot L^r_j)(\Sigma_k P^{t_s}_{ik} \cdot L^s_k)$$

where r and s are the descendants of node l and t_r, t_s are their respective branch lengths. The likelihoods throughout the entire topology are then calculated by traversing the tree starting from the tips and at the root, multiplying by the prior probabilities of each state (taken here as their equilibrium frequencies). The transition probabilities of the amino acids are modeled using a PAM matrix although any other transition matrix could be used as easily. An overall rate of replacement is determined that maximizes this likelihood.

To search for rate heterogeneity, we considered the corresponding likelihood for just those amino acids encompassed within a sphere of some specified diameter. Spheres were chosen with diameters ranging from 2 to 20 Å, centered in turn, upon each amino acid in the protein. The maximum likelihood for amino acids within this sphere and the likelihood that would be obtained using the overall optimal amino acid replacement rate can be compared. Spheres with unusually rapid/slow rates of substitution will have unusually large ratios of these two likelihoods. The likelihood ratio for any one sphere will be distributed asymptotically as a χ^2. But due to the

large number of such spheres, likelihood ratio tests would give false significance levels.

Therefore, an empirical measure, making use of permutations is employed. This test has the advantage of simplicity. The permutation shuffles only the rank order of sites within the alignment. It retains the relative frequencies of different amino acids, it retains the composition of individual taxa and it retains the same tree topology. Each residue in the three dimensional structure of the protein has a random set of neighbors after permutation. The identical test is then applied to the permuted order, searching for any regions with unusually high or low rates of amino acid replacement. This is repeated a thousand or more times. The extent to which the regions in the three dimensional structure are unusual in their rates of substitution can be compared to the permuted proteins to give an empirical measure of their significance.

The results are relative to the group of taxa selected and are also relative to the remainder of the molecule. That is, if the entire molecule is evolving at an unusually rapid rate this would not be detected by this method as it compares one region of the molecule to another region. Tests for absolute rates should be done using a method that is not comparative. In addition, these results depend on the species selected. It is possible that a single taxon may have an unusually rapid/slow rate of substitution in one region. When this taxon is excluded the rate variation may disappear. To test if the variation is localized to individual taxa, jackknife or similar subsampling of the taxa is recommended.
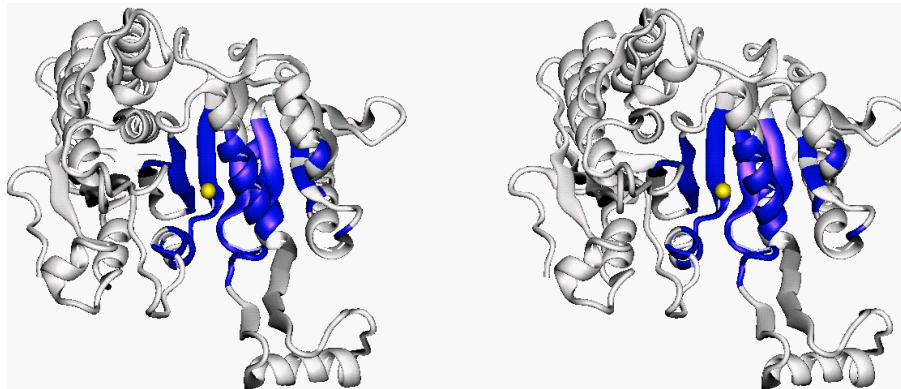

Figure 2—Crosseyed stereo view of a monomer of IDH showing the slowly evolving region (black) near the catalytic Mg^{2+} (sphere). Black residues on the right hand side of the monomer are part of the second active site in the dimer. This slowly evolving region, contiguous in three dimensional space, is

not identified by genic analyses because, as can be seen by following the trace, it is discontinuous in the one dimensional space of the linear sequence.

When analyzed by this methodology, IDH sequences reveal many regions that demonstrate significant heterogeneity. The most abnormal region is a 10Å diameter sphere centered at amino acid residue 153 (Figure 2). This region evolves at a rate that is 2.8 times slower than the overall average of the molecule. The 59 residue region encompasses much of the active site and includes residues ranging from 113 to 338. This simple approach will surely prove useful in identifying functionally important regions in three-dimensional space that are discontinuous in the one dimensional sequence of a gene.

4 Explaining the Hot Spots and the Cold Spots

If the above method provides a means to identify conserved and rapidly evolving regions, it does nothing to explain why such heterogeneity in rates occurs across sites. For example, it takes no account of the influence of protein structure on rates of replacement. Much information pertinent to this problem is available from X-ray crystallographic structures and can be retrieved with suitable hardware and software.

Graphical analysis of IDH was conducted on an SGI Indigo II running Quanta software (Molecular Simulations Inc.). Secondary structure assignments were determined from the Phi and Psi torsion angles along the main chain of the polypeptide. Solvent accessibilities of amino acids in the homodimer (with both substrate and coenzyme bound) were determined from the static structure as the proportion of van der Waals surface area their side chains make contact with a 1.4Å diameter spherical probe. Because the positions of hydrogen atoms are not resolved in IDH, H-bonding patterns were inferred from the positions of non-hydrogen atoms using Quanta. Quanta was also used to identify amino acids occupying unusual environments (e.g. a charged residue buried in the hydrophobic core). Active site residues are defined as those with at least one atom within 6.0Å of the bound substrates, Mg^{2+} isocitrate and NADP. Statistical analyses were implemented using JMP (SAS Institute Inc.). Bonferroni corrections for Type I errors were not implemented because so few models have levels of significance near $\alpha = 0.05$.

Table 1 presents the results of regression analyses and analyses of variance for various models. Cα torsion angles, the angles at which amino acid side chains extend from the main chain of a peptide, have no significant impact on the number of replacements per site. Nor do Phi and Psi angles, which are the principle determinants of secondary structure, and nor do H-bonding patterns on the main chain and side chains. Secondary structure assignment alone is barely significant (and were a Bonferroni correction implemented it may not be so) and in any case

explains only 2% of the observed site to site variation. By any criterion, secondary structure has little impact on replacement rates.

Table 1. Coefficients of Dispersion (r^2) of Models

Model	Proportion Observed Variance Explained (r^2)	df	Probability
Cα Torsion Angle	0.041	1	0.4018
Phi Angle	0.004	1	0.1051
Psi Angle	0.000	1	0.9864
H-bonding	0.007	1	0.0918
Secondary Structure (helix, sheet, turn, coil)	0.022	3	0.0333
Mean B-Factor (Å^2)	0.058	1	<0.0001
20 Amino Acids	0.132	19	<0.0001
Pro, Gly, Other	0.073	4	<0.0001
Hydrophyllic, Charged, "Pro, Gly, Other"	0.056	2	<0.0001
Side Chain Exposed (Å^2)	0.259	1	<0.0001
Fraction Side Chain Exposed (FSCE)	0.242	1	<0.0001
Unusual Environment	0.002	1	0.4366
Active/Nonactive Site	0.096	1	<0.0001
Cold Spot	0.099	1	<0.0001
Distance to Catalytic Mg^{2+} (Distance)	0.334	1	<0.0001
FSCE + Distance	0.391	2	
FSCE + "Pro, Gly, Other" + Distance	0.446	4	

The B-factor (or Debye-Waller factor) is a measure of atomic disorder; the higher the B-factor, the less well localized is the atom within the crystal[18]. For example, atoms in the hydrophobic cores tend to have lower B-factors than those in flexible surface loops. An amino acid can hardly be considered constrained if it adopts many conformations. Consequently, B-factors might be taken as an indication of functional constraint. Table 1 shows that, although B-factors contribute significantly to rate heterogeneity among sites, they explain no more that 6% of the site to site variation.

Better results were obtained using the 20 amino acids, which account for 13% of the site to site variation (Table 1). Yet 13% does not seem quite so impressive when 19 degrees of freedom are used. Gly, Pro and Trp are relatively conserved, whereas Lys, Glu and Gln tend to occupy sites that rapidly evolve. Reclassifying amino acids as hydrophobic (Ala, Val, Leu, Ile, Phe, Met, Cys, Trp, Tyr), hydrophyllic (Asn, Gln, Ser, Thr) and charged (Arg, His, Lys, Asp, Glu) and retaining Gly and Pro because of their unusual structural properties (Gly can adopt unusual main chain conformations important in turns and loops; the side chain of Pro is attached to the main chain carbon which greatly restricts conformations) reduces the degrees of freedom to 4 while retaining an ability to explain 7% of the observed

variation. Most of this 7% (5.5%) can be explained using just three classes and 2 degrees of freedom: Gly, Pro, and everything else. In fact, of all the classifications we have investigated only Gly and Pro retain significance when combined with other effects. That only 5.5% of the site to site variability can be explained may seem small, but it should be borne in mind that only 14.5% of residues are Gly or Pro. We conclude that Gly and Pro routinely make a significant, though small, contribution of site to site heterogeneity in the rate of evolution.

No matter how one measures it, by surface area exposed or fraction of side chain surface area exposed, solvent accessibility makes a major contribution to variation among sites. This well known observation[19,20] is not unexpected and has been commonly interpreted as meaning that side chains buried in tightly packed hydrophobic cores are more constrained than those exposed to solvent and which, being free to move, suffer fewer structural constraints. Approximately 25% of the site to site variability can be explained by solvent accessibility alone (Table 1).

By contrast, amino acids occupying unusual environments in the *E. coli* IDH structure (e.g. solvent exposed Trp) do not seem to evolve at unusual rates (Table 1). This may be because in most other proteins these sites are occupied by the appropriate residues. One way to investigate this possibility is to model related and ancestral IDHs, energy minimize their structures and perform dynamics simulations using a force field such as CHARMM[21] to ensure that no other conformations are likely, and then determine which residues occupy unusual environments. The considerable effort involved would be wasted if it did not markedly improve predictive power, or if other simpler methods better account for the data.

Another obvious classification is whether a residue lies inside or outside an active site. In Figure 3 IDH residues are color coded according to the number of replacements at each site. It can clearly be seen that regions surrounding the active sites (marked by the yellow catalytic Mg^{2+}) evolve very slowly (dark blue) , whereas regions outside evolve quite rapidly (pink). This dichotomous classification with 42 active site residues accounts for only 9.6% of the site to site variation. Most of the 59 residues in the cold spot identified by our maximum likelihood method are common to the active site as defined from the X-ray crystallographic structure with bound substrates. Not surprisingly, a dichotomous classification with 59 cold spot residues yields similar results (Table 1).

Inspecting a cross section through the hydrophobic core of the IDH dimer (Figure 3) reveals that a dichotomous classification does not capture the fact that rates of replacement increase with increasing distance from the catalytic Mg^{2+}. A simple regression of replacements against distance to the nearest catalytic Mg^{2+} accounts for 33.3% of the site to site variation. This is by far the greatest proportion of variability explained by a simple regression model (Table 1).

Combining distance and a measure of solvent accessibility results improves the fit considerably (39% of the variation is explained, r = 0.62) and adding the "Pro,

Gly, Other" classification improves matters further, to 44.6% (r = 0.67), at a total cost of only 4 degrees of freedom. When all independent criteria are combined 52% (r = 0.72) of the site to site variability in the number of replacements is explained, but at a cost of 31 degrees of freedom. All these models consist of fixed main effects. No higher order interaction has yet been found to improve the fit significantly.
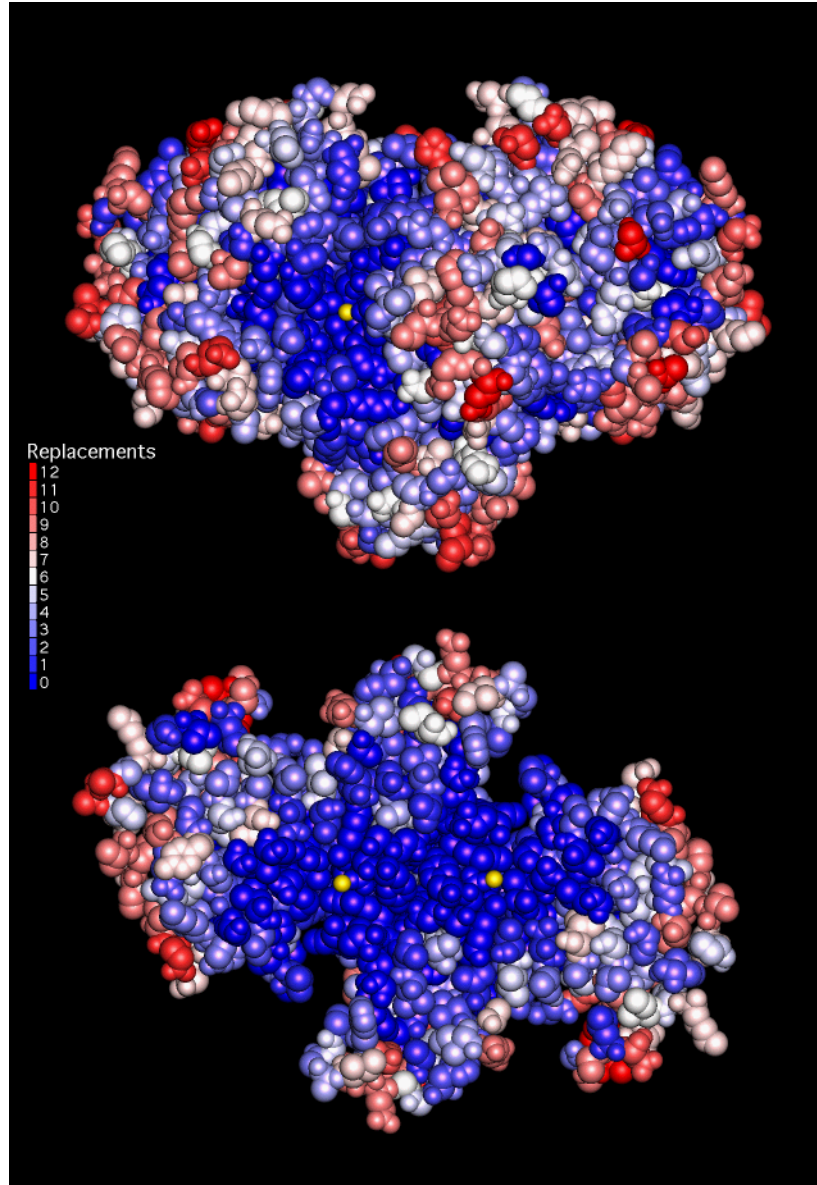
Figure 3—The surface of the IDH dimer (top) and a cross section through its active sites (below) illustrating the distribution of replacements in relation to the catalytic Mg^{2+}s (yellow). It needs no statistical test to confirm that the active sites are highly conserved (dark blue), whereas the remaining

surfaces evolve rapidly (pink). The cross section reveals that residues buried deep in the hydrophobic cores of the domains evolve more rapidly the further they are from the catalytic Mg^{2+}.

5 The Expected Correlation Coefficient of a Poisson Process

How well do these mixed models explain the observed rate variation? The adequacy of a model's fit to data is usually judged through the use of a correlation coefficient. A correlation coefficient near ±1 indicates a good agreement between the model's predictions and the observed data. By this common interpretation the correlation coefficient of our three parameter model with r = 0.67 (for distance, solvent accessibility and the "Pro, Gly, Other" classification) is unimpressive, indicating a model with only limited predictive power. We should construct a better model.

Or should we?

Many phenomena are inherently stochastic, their data inevitably scattered around an expectation by chance alone. The accumulation of amino acid replacements at a site in a protein is one such example. If, with a priori knowledge, we expect 5 replacements at a site, we should not be altogether surprised if 4 or 7 replacements are actually observed during the course of evolution. This scattering ineluctably lowers correlation coefficients, and a value of 1 can no longer be achieved even by a perfect model. We should judge the adequacy of our model not on a correlational scale of 0 to 1, but on a correlational scale of 0 to something less than 1.

The problem is to determine the correlational range over which a model is to be judged, specifically the upper limit of that range - the "something less than 1." At first glance it might seem absurd to suggest that this, the maximum possible correlation coefficient for a perfect yet unknowable model of molecular evolution, can be estimated without ever specifying the biological basis of the model. Yet that is precisely what we shall now do.

We begin by partitioning the observed variance in replacements across sites (s^2_{obs}) into two portions, the first ascribable to the expectation of a perfect model of molecular evolution (s^2_{exp}) and the second ascribable to residual errors (s^2_{err}):

$$s^2_{obs} \;=\; 1/n \sum (y_i - \bar{y})^2$$
$$=\; 1/n \sum (Ey_i - \bar{y})^2 \;+\; 1/n \sum (y_i - Ey_i)^2$$
$$s^2_{obs} \;=\; s^2_{exp} \;+\; s^2_{err}$$

Consider again the expected number of replacements at a single site in a protein, and consider too the distribution around that expectation. It is Poisson. This is true regardless of changes in rates of mutation, the effects of selection, of drift, and of lineage effects. This is true regardless of whether amino acids at different sites

interact. The only requirement is that replacements at one site do not influence the rates of replacement at other sites. Let the number of replacements at a site, y_i, be drawn from a Poisson distribution, $Ey_i = \lambda_i$ and $\sigma_{yi}^2 = \lambda_i$. The variance across all sites is given by

$$
\begin{aligned}
\sigma^2 &= \sigma_{exp}^2 + \sigma_{err}^2 \\
&= 1/n \, \Sigma(\lambda_i - \bar{\lambda})^2 + 1/n \, \Sigma(y_i - \lambda_i)^2 \\
&= 1/n \, \Sigma(\lambda_i - \bar{\lambda})^2 + 1/n \, \Sigma\sigma_{y_i}^2 \\
&= 1/n \, \Sigma(\lambda_i - \bar{\lambda})^2 + \bar{\lambda}
\end{aligned}
$$

We estimate the population mean and variance, $\bar{\lambda}$ and σ^2, using the sample mean and variance, $\bar{y}$ and s_{obs}^2. Hence,

$$
s_{obs}^2 = \hat{\sigma}_{exp}^2 + \bar{y}
$$

Rearranging yields

$$
\frac{\hat{\sigma}_{exp}^2}{s_{obs}^2} = 1 - \frac{\bar{y}}{s_{obs}^2}
$$

The left-hand side of this equation is simply an estimate of the proportion of variation that a perfect model of molecular evolution would explain. The square root

$$
\hat{\rho} = \frac{\hat{\sigma}_{exp}}{s_{obs}} = \sqrt{1 - \frac{\bar{y}}{s_{obs}^2}}
$$

is an estimate of the correlation coefficient of the perfect model. The right hand side reveals that this can be calculated using the observed mean and variance alone. Hence, the correlation coefficient of a perfect yet unknowable model of molecular evolution can be estimated in complete ignorance of the biological basis of that model. The only assumption is that, in common with virtually all models of molecular evolution, replacements are Poisson distributed. In analogy to r (the model-dependent correlation coefficient), we call $\hat{\rho}$ the model-independent correlation coefficient (MICC).

The mean and variance of the number of replacements across sites in IDH are $\bar{y}$ = 4.66 and s_{obs}^2 = 8.79 which yield $\hat{\rho}$ = 0.68. Our simple model's correlation coefficient of 0.67 is rather more satisfactory when judged on a scale of 0 to 0.68. This approach also reveals the danger of over parameterizing a model: with 31 degrees of freedom we can account for $52/(1 - 4.66/8.79) = 111\%$ of the variability

that actually needs explaining. A failure to appreciate the appropriate correlational scale on which the outcome of a stochastic process is judged may often result in attempts to interpret random noise.

Acknowledgments

References

1. J. W. Drake, *Proc. Natl. Acad. Sci. USA* **88**, 7160 (1991)
2. J. W. Drake, *Proc. Natl. Acad. Sci. USA* **90**, 4171 (1991)
3. M. Kimura, *Proc. Natl. Acad. Sci. USA* **76**, 3440 (1979)
4. H. Ochmann and A. C. Wilson, *J. Mol. Evol.* **26**, 74 (1987)
5. M. Goodman, *Prog. Biophys. Mol. Biol.* **38**, 105 (1981)
6. J. Czelusniak, M. Goodman, D. Hewitt-Emmett, M. L. Weiss, P. J. Venta, and R. E. Tashian, *Nature* **298**, 297 (1982)
7. W.-H. Li, *Molecular Evolution* (Sinauer Ass., Sunderland, Mass., 1997)
8. C. O'hUigin and W.-H. Li, *J. Mol. Evol.* **35**, 377 (1992)
9. S. Seino, G. I. Bell, and W.-H. Li, *J. Mol. Evol.* **9**, 193 (1992)
10. M. Kimura, *Nature* **267**, 275 (1977)
11. C. C. Luo, W.-H. Li and L. Chan, *J. Lipid. Res.* **30**, 1735 (1989)
12. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930)
13. A. M. Dean and G. B. Golding, *Proc. Natl. Acad. Sci. USA* **94**, 3104 (1997)
14. D. G. Higgins, A. J. Bleasby, and R. Fuchs, *CABIOS* **8**, 189 (1992)
15. K. Imada, M. Sato, N. Tanaka, Y. Katsube, Y. Matsuura, T. Oshima, *J. Mol. Biol.* **222**, 725 (1991)
16. R. A. Fisher, *Statistical Methods for Research Workers, 10th ed.* (Oliver and Boyd, Edinburgh, 1948)
17. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981)
18. T. E. Creighton, *Proteins: structures and molecular properties* (Freeman, New York, 1993)
19. M. Kimura and T. Ohta, *Genet. Suppl.* **73**, 19 (1973)
20. N. Goldman, J. Thorne and D. Jones, *Genetics* **149**, 445 (1998)
21. B. Brooks, *J. Comp. Chem.* **4**, 187 (1983)