Neutral evolution on mammalian protein surfaces

Gavin C. Conant

Because of their low effective population sizes, natural selection is expected to have reduced effectiveness in organisms such as mammals. By comparing the amino acid substitution rates between mammalian protein surfaces and interiors, it was found that almost a third of the proteins surveyed failed to reject the null hypothesis of neutral substitutions among surface residues. Proteins with such partly neutral evolution nonetheless have no fewer protein interactions than do other proteins. I suggest that natural selection can function to preserve protein interactions without requiring strict conservation of the individual residue contacts that impart those interactions.

Molecular evolution, population size and protein structure

The birth of the field of molecular evolution was attended by a certain parental tension between the then-current state of evolutionary theory and the new ability to study the genetic structures of populations. This neutralist–selectionist debate continues to stimulate discussion as to whether the majority of the fixed sequence differences between species have arisen through genetic drift [1–3].

One avenue for exploring this question is to compare situations where the propensity for neutral variation varies. One such comparison is between the residues of a protein exposed to the solvent versus those buried in the interior of the folded structure. It is well known that interior residues are subject to stronger constraints against substitution than are exposed residues [4–6]. This pattern is maintained over wide taxonomic ranges, including mammals [7], yeasts [8] and bacteria [5], despite the variation in effective population sizes (Ne) among these organisms that might be thought to produce differing selective regimes.

Such differences in the actions of natural selection are suggested by the nearly neutral theory of molecular evolution [9] that relates such selection to Ne. This theory points out that, because selection is less effective in pruning disadvantageous variation from small populations, one should observe an association between Ne and measures of genetic diversity [10]. The implied increased propensity for neutral substitutions is potentially important for a number of reasons: not least because neutral changes can profoundly affect the trajectories of biological evolution [1,11].

Mammals and yeasts show similar ratios of surface to interior constraint

In yeasts, we have recently compared the ratio of nonsynonymous to synonymous substitutions (Ks/Ka or ω) between surface and interior residues (ωs and ωi, respectively), finding not only that ωs > ωi (i.e. interior residues are more constrained) but, more surprisingly, that the ratio of ωs/ωi is quite similar across a range of proteins [12]. Here I extend this analysis to eight mammals (human, chimpanzee, the rhesus macaque, mouse, rat, cow, horse and dog). The approach differs from previous work [4,7,8] in that selective constraints are estimated for individual genes using orthologous genes from several species, rather than by performing correlation analyses between exposure and pairwise sequence divergence. Although this technique results in smaller samples, it allows us to understand the variation between genes in how structure affects selection.

It was hypothesized that the weakening of selection due to mammals’ small Ne would reduce the difference in constraint between the surface and interior. Instead, Ne appeared to manifest itself as an increased propensity toward neutral evolution at the surfaces of mammalian proteins. (Here ‘neutral evolution’ is used as shorthand for the fixation of nearly neutral alleles by genetic drift.)

Using 194 human proteins with known crystal structures [13], the corresponding genes were aligned with their orthologs in the seven other genomes and the relative surface exposure x of each residue was calculated (see the supplementary data online). The resulting alignments were then analyzed using two models of codon evolution.

The first model employed was that of Muse and Gaut/Goldman and Yang (MG/GY) [14,15] where the ratio of nonsynonymous to synonymous substitutions (i.e. selective constraint) is given by the parameter ωe. This model was extended by allowing residues to experience a continuous mix of selective constraints depending on their proportional surface exposure x (see the supplementary data online). Thus, the likelihood L of a particular site s is given by:

\[ L(s|x) = x \cdot L(s|x) + (1-x) \cdot L(s|x) \]  

A likelihood ratio test (LRT, see the supplementary data online) [16] was then used to explore whether there was a significant difference in the selective constraint on surface and interior residues. Of the alignments considered, 151 showed significant improvement under the MG/GY struct model (LRTMG/GY struct; Table 1) with a false-discovery rate (FDR) threshold of 0.05 [17].

The constraint ratio ωs/ωi is similar across these 151 proteins (Figure 1b) and does not differ significantly from
that seen in yeasts (Wilcoxon rank test, \(P = 0.42\), Figure 1a,b). This observation reinforces our previous speculation that the limited variation in \(\omega_{s}/\omega_{a}\) results from substitutions in the protein interior being more likely to hinder folding [18]. Note however that both \(\omega_{s}\) and \(\omega_{a}\) are significantly larger in mammals (Wilcoxon rank test, \(P < 0.002\)). Moreover, contrary to the situation in yeast, there is a significant association between \(\omega_{a}\) and \(\omega_{s}/\omega_{a}\) (Spearman’s \(r = 0.54\), \(P < 10^{-10}\)).

Neutral evolution at mammalian protein surfaces
Fourteen of these 151 alignments had estimates of \(\omega_{s} > 1.0\). Under the MG/GY model, \(\omega > 1.0\) indicates that the selection operating on nonsynonymous substitutions is no greater than that (potentially operating on synonymous substitutions, whereas \(\omega > 1.0\) is indicative of directional/diversifying selection. To test the extent to which (nearby) neutral evolution might be occurring in these genes, the MG\(_{struct}\) model was introduced, requiring \(\omega_{s} = 1.0\). For 47/151 (31%) of proteins (hereafter Neu), the null hypothesis of neutral evolution among the surface residues could not be rejected (\(P > 0.05\); LRT\(_{MG-struct-neu}\); Table 1). Notably, only 1/61 (2%) of yeast proteins fail to reject neutrality (\(P < 0.05\), LRT, Figure 1b), suggesting that yeasts, with their larger \(N_{e}\), more efficiently remove mildly deleterious surface substitutions. Of the remaining 104 mammalian alignments (hereafter Sel), three have \(\omega_{a}\) greater than 1.0 (FDR-corrected \(P\) value \(\leq 0.061\), an observation suggesting directional selection. The human genes in question are: \(OSM\) encoding oncostatin M, \(CD8A\) encoding the immune system protein CD8a, and \(CLEC1B\) encoding a member of the protein family which contains C-type lectin domains. Oncostatin M is a cytokine involved in controlling cell differentiation [19], whereas CD8a contributes to the binding between the MHC and T-cell receptor [20]. The \(CLEC1B\) gene product has a less well-defined function in the immune system [21].

One might suspect that Neu consists of cases where the power to detect differences in \(\omega_{s}\) is low. To examine this issue, sequence data were first simulated under models where \(\omega_{s} \neq 1.0\) and the power of the LRT to reject \(\omega_{s} = 1.0\) using these data was measured (see supplementary material online). As can be seen in Figure 1d, power is low for values of \(\omega_{s}\) close to 1.0 but higher in other parameter ranges. To study the degree of statistical support for the value of \(\omega_{s}\) for each gene in Neu, I simulated data under the ML parameter estimates and determined whether each simulated dataset was able to reject two null hypotheses: \(\omega_{s} = 1.0\) and \(\omega_{a} = 0.5\). The second hypothesis is included as a test of whether a sequence alignment has an overall strong signal of the value of \(\omega_{a}\) even if that alignment cannot reject \(\omega_{s} = 1.0\). Thus, in cases where \(\omega_{a}\) is near unity, the power to reject \(\omega_{s} = 1.0\) can be low, but \(\omega_{a}\) might still be clearly distinguishable from 0.5. The choice of \(\omega_{a} = 0.5\) is of course somewhat arbitrary but represents a natural and moderate level of selection. I next defined a measure of support \(q\) for the value of \(\omega_{a}\), given by the proportion of simulated datasets that rejected either \(\omega_{a} = 0.5\) or \(\omega_{a} = 1.0\) at \(P = 0.05\) for a given gene in Neu. Thus, \(q\) can be thought of as how often an alignment like the one used would be able to distinguish between \(\omega_{a} = 0.5\) and \(\omega_{a} = 1.0\) with statistical significance. The average value of \(q\) across the 47 genes is 0.54 and all but three genes have \(q > 0.25\). Thus, while it is very difficult to determine the number of true-positive neutral proteins, one can at least say with some confidence that the test would have the power to distinguish between neutrality and \(\omega_{a} = 0.5\) with relatively high statistical confidence (\(P = 0.05\)) for approximately half of the 47 genes considered.

A second model of sequence evolution was also employed, the similarity groups (SG) model, to see if the patterns of the amino acid substitutions support neutrality. This model categorizes amino acid substitutions according to charge [12,22], allowing one substitution rate

---

Table 1. Likelihood ratio tests (LRT) performed

<table>
<thead>
<tr>
<th>LRT</th>
<th>Null model parameters</th>
<th>Alternative model</th>
<th>n</th>
<th>df (^{a})</th>
<th>No. of significant results (^{b})</th>
<th>Groups differ (^{c})</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRT(_{MG})-struct</td>
<td>MG (\omega_{s}) (\omega_{a})</td>
<td>MG(<em>{struct}) (\omega</em>{s}), (\omega_{a})</td>
<td>194</td>
<td>1</td>
<td>151*</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>MG(<em>{struct}) (\omega</em>{s}=1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} \neq 1.0), (\omega_{a})</td>
<td>151</td>
<td>1</td>
<td>104**</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>SG (\omega_{s} \neq 1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} \neq 1.0), (\omega_{a})</td>
<td>151</td>
<td>2</td>
<td>150*</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>SG (\omega_{s} = 1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} = 1.0), (\omega_{a})</td>
<td>47</td>
<td>1</td>
<td>18**</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>SG (\omega_{s} = 1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} = 1.0), (\omega_{a})</td>
<td>47</td>
<td>1</td>
<td>18**</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>SG (\omega_{s} \neq 1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} \neq 1.0), (\omega_{a})</td>
<td>47</td>
<td>1</td>
<td>18**</td>
<td></td>
</tr>
<tr>
<td>LRT(_{MG})-struct-neu</td>
<td>SG (\omega_{s} = 1.0), (\omega_{a})</td>
<td>MG(<em>{struct/neu}) (\omega</em>{s} = 1.0), (\omega_{a})</td>
<td>47</td>
<td>1</td>
<td>18**</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Model parameters describing the rate of nonsynonymous substitutions and their constraints. For the MG and SG models, no structural information was included: each model has only a single set of nonsynonymous parameters (\(\omega_{s}\) and \(\omega_{a}\), respectively). For the remaining models, the surface residues and interior residues were allowed to take on distinct nonsynonymous rates (\(\omega_{s}\) and \(\omega_{a}\) for the MG-derived models and \(\omega_{s}\), \(\omega_{a}\), \(\omega_{s}\), and \(\omega_{a}\) for the SG-derived models).

\(^b\)The degrees of freedom (df) used for calculating the significance of the LRT under a chi-square distribution. For a given test, df is determined from the number of excess parameters in the more complex model. Thus, for the comparison of the MG and MG\(_{struct}\) models, df=1 because the MG\(_{struct}\) model has one extra parameter (\(\omega_{s}\) and \(\omega_{a}\) as opposed to simply \(\omega_{a}\)).

\(^c\)Hypothesis test that the Neu and Sel sets have equal proportions of alignments that reject the null hypothesis for a given LRT (chi-square test).

\(^d\)Alignments that do not reject the null hypothesis of MG\(_{struct}\) for LRT\(_{MG-struct-neu}\).

\(^e\)Alignments that reject the null hypothesis of MG\(_{struct}\) for LRT\(_{MG-struct-neu}\).
between residues with the same charge (I_w) and a second (I_b) between those of differing charge. Under neutrality, substitutions altering the charge should be no less probable than charge-preserving substitutions. The full model (SGstruct) employed has independent values of I_w and I_b for the surface and interior residues (I_w, I_w, and I_b, I_b). The signature of neutral evolution is thus I_w = I_b.

The SG model distinguishes surface and interior residues: all but one of the 151 alignments where LRTMG-struct is significantly large also show a significant improvement in fit with the SGstruct model (LRTSG-struct; Table 1): the remaining case shows marginal significance (FDR-corrected P = 0.057). The SGstruct model was therefore compared to the neutral SGstruct/neu model (I_w = I_b). In the Neu set, 66% of the genes do not reject SGstruct/neu (LRTSG-struct/neu; Table 1): significantly fewer of the genes in Sel (38%) fail to do so (P = 0.001, chi-square test).

The I_w and I_b parameters in the SG model give the ratio of nonsynonymous charge-preserving substitutions to synonymous substitutions, and of nonsynonymous charge-altering substitutions to synonymous substitutions, respectively. In other words, they are akin to ω in the MG/GY model, and one can test for neutrality by requiring I_w or I_b = 1.0. For both parameters, significantly more cases are found in Sel where neutrality cannot be rejected than in Neu (LRTSG-I_w-neu and LRTSG-I_b-neu; Table 1). Of the 16 genes in Neu that reject the SGstruct/neu model, 15 genes fail to reject the null hypothesis of (partial) neutrality for I_w and four genes fail to reject neutrality for either I_w or I_b.

Examples where I_w or I_b are significantly greater than 1.0 are indicators of directional selection. Of these 151 genes, only two of the three genes identified previously as indicative of directional selection were consistent with such selection under this test: OSM (I_b > 1.0; FDR-corrected P = 0.007) and CD8A (I_w > 1.0; FDR-corrected P = 0.001).

Neutral substitutions and the protein interaction network
Although surface residues are less implicated in protein folding, they still mediate protein interactions, making this apparent neutrality somewhat surprising. This is especially the case as coevolution between interacting residues has been documented (reviewed in Ref. [23]). Of course, if the Neu set had been drawn from proteins with few interactions, that might explain the low surface constraint.

However, there is little difference between the two sets in the proportion of genes with protein interactions: 47% for Neu versus 42% for Sel (see the supplementary material online). The average number of interactions per gene is also similar (4.7 and 4.6, respectively; see the supplementary data online). Neither of these figures differs significantly between the two sets (P = 0.61 and P = 0.48, chi-square test and Wilcoxon rank test, respectively). There is also no difference in the proportion of proteins in the two sets annotated with the GO term ‘protein binding’ [24] (P=0.5, Fisher’s exact test). Of course, it remains possible that a few interacting surface residues are under selective constraint but that the model used here is insufficiently sensitive to detect them.

Implications of surface neutrality for molecular evolution
Of the proteins surveyed, at least 21% show neutral evolution among surface residues (the 31 gene products that fail to reject either MGstruct/neu or SGstruct/neu). Note however, that relatively few residues are completely exposed: on average only 7% of a protein has solvent exposure of 75% or more. Although these results might appear at variance with a previous analysis that rejected neutrality [7], the
difference is likely to be due to the fact that the former study aggregated data across genes.

My simulations suggest that low statistical power alone does not account for all of the observed neutrality. This contention is supported by the fact that all 151 alignments have sufficient evolutionary signal to distinguish between surface and interior residues ($o_{3}$ and $o_{5}$ are statistically distinct). On a related note, Berglund et al. [25] recently suggested that some inferred instances of directional selection might be better explained by mutational biases. There are three reasons for thinking it unlikely that similar forces are at work here. First, the approach used here identifies residue exposure independently of and prior to constraint, rather than detecting positive selection directly from sequences. Second, no association was found between base composition and neutrality (see the supplementary material online). Finally, biased mutations cannot explain the neutrality seen in the SGstruct/neu model as it is based on amino acid substitutions. This model likewise avoids issues with non-neutrality of synonymous substitutions. It is also to be noted that, to avoid potential difficulties matching orthologous genes to protein structures, protein structures were only accepted if their associated sequences exactly matched the human genome, another source of the smaller sample used here.

Two possibilities suggest themselves when considering the implications of these finding for the evolution of protein interaction networks. First, one might imagine that such neutral evolution would give rise to a highly pliable protein interaction network that would follow Maynard-Smith's conception of neutral paths through the space of genotypes (neutral networks) [26–28]. These paths allow very distinct genotypes to evolve from each other through a series of individually neutral mutations [1].

The second interpretation of apparently neutral substitution rates at protein surfaces is that it might reflect the alternating fixation of mildly deleterious substitutions, weakening protein interactions, and mildly advantageous compensating substitutions that strengthen them. Note that this hypothesis is consistent with the observation of coevolution between interacting residues. The details of such a model remain to be worked out, because obviously if a deleterious substitution has a fitness effect small enough to allow fixation through drift, the advantage of a compensating back-mutation is likely to be similarly small. Nonetheless, such weakly advantageous substitutions are known [29]. Note that, given the models employed, the relatively few cases of directional selection detected above do not rule out this effect.

It seems likely that it will eventually become possible to distinguish between these two explanations of the pattern of selection at protein surfaces using population genetic data. However, an initial analysis of this type was unsuccessful because the relatively few variable sites per protein gave insufficient power to explore evolution among the rare surface residues (see the supplementary material online). In the meantime, it is intriguing that both explanations imply a weak relationship between the conservation of protein sequence and of protein interactions. We might thus expect to find a constant turnover of substitutions among interacting proteins that is not accompanied by any necessary change in the function of the network.

Acknowledgements

I would like to thank M. Woolfit and four anonymous reviewers for critical comments on the manuscript and Á. Pérez-Bercoff, D. Gonzales Knowles and A. McLyshaght for helpful discussions regarding ortholog inference. This work was supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2009.07.004. Alignments, surface exposures and tree files for the 194 alignments analyzed are provided on the author's website: (http://web.missouri.edu/~conant/data/mammal_struct_evol/alignments_trees_rates.tar.gz).

References

Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals

Gertraud Burger, Yifei Yan, Pasha Javadi and B. Franz Lang

Placozoa – the simplest known free-living animals – have been considered primitive, early diverging metazoans based on mitochondrial genome structure and phylogeny. Here we reanalyze placozoan mitochondrial DNAs, reported to include a highly unorthodox, fragmented and incomplete cox1 gene. We discover overlooked exons and split group I introns that mediate trans-splicing of the discontinuous placozoan cox1. Furthermore, we find that cox1 expression involves U-to-C editing, reconstituting an otherwise invariant, essential histidine involved in copper binding. These atypical features qualify placozoan mitochondrial gene and genome organization as derived rather than primitive. Whether the Placozoa diverged early or late during metazoan evolution remains unresolved by mitochondrial phylogeny.

Are Placozoa primitive or derived animals?

Trichoplax adhaerens, an organism thriving in tropical waters, is the simplest known free-living animal [1]. It grows as a flat disk, 0.5–3.0 mm wide and ~25 μm thick, has only four different somatic cell types, and lacks organs, differentiated tissues, and body symmetry [2]. Traditionally, T. adhaerens has been the only recognized species of the phylum Placozoa, but recent rRNA and ITS genotyping studies distinguish five different ‘highly divergent’ placozoan clades. The simple placozoan body plan has been interpreted by some authors as genuinely primitive (placing Placozoa as the deepest divergence within Metazoa), but by others as reduced-derived (see Refs. [2–5] and references therein).

The phylum’s phylogenetic position inferred from molecular data has been equally controversial (Figure S1). With only marginal statistical support, nuclear rRNA data with broad metazoan taxon sampling place Placozoa either within or close to Cnidaria (e.g. Ref. [6]) or as a sister clade to Porifera (sponges) / Ctenophora (comb jellies) [7]. Multi-protein phylogenies are similarly inconclusive. A report on the Trichoplax nuclear genome [8] presents a phylogeny in which Placozoa emerge before Cnidaria plus Bilateria, but statistical tests of alternative tree topologies leave the position of Trichoplax as unsupported, and taxon sampling is limited (only one demosponge, one placozoan, and two cnidarians; there are no representatives of other sponge lineages or ctenophores). Independent analysis using nuclear plus mitochondrial sequences, but with similarly limited species sampling (for example, outgroup species close to Metazoa, such as Ministeria and Capsaspora, were not included) provides support for monophyly of Placozoa plus non-bilaterian animals [9], but this is probably a long-branch attraction (LBA) artifact. Finally, a phylogenomic analysis with rich taxon sampling and the more reliable CAT model recovers sponge monophyly and certain deep animal relationships, yet fails to position the placozoan representative in the tree with confidence (62% bootstrap support [10]).

By contrast, two phylogenies constructed with concatenated mitochondrial-encoded proteins strongly indicate that the Placozoa diverged later than the Bilateria, a sister group to the Cnidaria and sponges (in other words, Placozoa plus non-bilaterian Metazoa are a sister group to Bilateria) [5,11]. Nevertheless, as in the examples discussed above, Placozoa might be misplaced owing to poor taxon sampling (only three Porifera and four to five Cnidaria), and potential LBA might artifactually draw Placozoa towards the distant, fast-evolving fungal outgroup and/or the extremely fast-evolving Bilateria. Indeed, the same set of mitochondrial-encoded proteins, but with broader taxon sampling, fails to provide significant support for placozoan divergence, and even yields different tree topologies with different evolutionary models [12]. The recent availability of numerous complete mitochondrial genomes...