doi:10.1016/S0022-2836(02)00239-5 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 319, 729–743

JMB

Detecting Compensatory Covariation Signals in Protein Evolution Using Reconstructed Ancestral Sequences

K. Fukami-Kobayashi[1], D. R. Schreiber[2] and S. A. Benner[2,3]*

[1]*Center for Information Biology and DNA Data Bank of Japan National Institute of Genetics Mishima 411-8540, Japan*

[2]*Departments of Chemistry and Anatomy and Cell Biology and the NASA Astrobiology Institute, University of Florida Gainesville, FL, USA*

[3]*Foundation for Scientific Inquiry, P.O. Box 13174 Gainesvilla, FL 32604, USA*

**Corresponding author*

When protein sequences divergently evolve under functional constraints, some individual amino acid replacements that reverse the charge (e.g. Lys to Asp) may be compensated by a replacement at a second position that reverses the charge in the opposite direction (e.g. Glu to Arg). When these side-chains are near in space (proximal), such double replacements might be driven by natural selection, if either is selectively disadvantageous, but both together restore fully the ability of the protein to contribute to fitness (are together "neutral"). Accordingly, many have sought to identify pairs of positions in a protein sequence that suffer compensatory replacements, often as a way to identify positions near in space in the folded structure. A "charge compensatory signal" might manifest itself in two ways. First, proximal charge compensatory replacements may occur more frequently than predicted from the product of the probabilities of individual positions suffering charge reversing replacements independently. Conversely, charge compensatory pairs of changes may be observed to occur more frequently in proximal pairs of sites than in the average pair. Normally, charge compensatory covariation is detected by comparing the sequences of extant proteins at the "leaves" of phylogenetic trees. We show here that the charge compensatory signal is more evident when it is sought by examining individual branches in the tree between reconstructed ancestral sequences at nodes in the tree. Here, we find that the signal is especially strong when the positions pairs are in a single secondary structural unit (e.g. α helix or β strand) that brings the side-chains suffering charge compensatory covariation near in space, and may be useful in secondary structure prediction. Also, "node–node" and "node–leaf" compensatory covariation may be useful to identify the better of two equally parsimonious trees, in a way that is independent of the mathematical formalism used to construct the tree itself. Further, compensatory covariation may provide a signal that indicates whether an episode of sequence evolution contains more or less divergence in functional behavior. Compensatory covariation analysis on reconstructed evolutionary trees may become a valuable tool to analyze genome sequences, and use these analyses to extract biomedically useful information from proteome databases.

Keywords: functional genomics; molecular evolution; structural biology; compensatory covariation; mutation; natural selection

Introduction

The evolution of protein sequences is nearly always described using one of several stochastic models for the accumulation of amino acid replacements.[1] These are captured in algorithms known by widely recognized names (e.g. the Needleman–Wunsch,[2] Smith–Waterman,[3] and

Abbreviations used: ASA, accessible surface area; PAM, point accepted mutations; PDB, Protein Data Bank.

E-mail address of the corresponding author: benner@chem.ufl.edu

Felsenstein maximum likelihood[4] tools). These tools have become more sophisticated in recent years as mathematicians have "inched towards reality"[4] in their mathematical modelling well supported by a rich background in statistics, theorems, and proof.

Nevertheless, patterns of replacement predicted by these mathematical models remain quite different from the patterns that are actually observed in proteins diverging under functional constraints.[1,5] The reason for these differences is well understood. Briefly, simple stochastic models treat proteins as if they were linear strings of letters. In reality, proteins have three-dimensional structures that support behaviors that are important for them to contribute to the fitness of the host ("function"). These behaviours are not a linear sum of the behaviors of their parts. Amino acid replacement is therefore constrained in a way unanticipated for a linear string of letters.

The differences between mathematically convenient models and the reality of organic chemistry need not be paralyzing, however. First-order stochastic treatments of protein sequences can provide "null" hypotheses, statements about how proteins would behave if they were formless, functionless strings of letters. The difference between how proteins actually divergently evolve and how first-order treatments model their divergent evolution, therefore contains a signal about fold and function.[1]

Analyses of this signal have been remarkably productive. They provide practical tools for predicting the folded conformation of proteins from sequences,[1,6] and many useful approaches for extracting functional information from genomic sequence data.[7–9] Accordingly, one goal of contemporary computational biology is to extend the concepts and tools needed to extract signals concerning structure and function from features of divergent evolution that do not meet the expectations of simple stochastic models.

Perhaps the most serious approximation made by first-order stochastic models is their treatment of individual positions in a protein sequence as independently evolving entities. Virtually all of these tools analyze sequence divergence one position at a time under a model where position i in a sequence suffers independent replacement of position j. Even models that recognize that different sites may have different mutability (gamma models) treat sites as being independently evolving.[10] This is, of course, extremely convenient for any statistical model for protein sequence divergence as it enables the probability of a sequence alignment overall to be calculated as the product of probabilities calculated for individual positions in the alignment.

Even casual inspection of a multiple sequence alignment, however, shows that positions in a protein sequence do not suffer replacement independently. Replacements in positions adjacent in a sequence are strongly correlated.[11] The correlation almost certainly reflects functional constraints on replacement set within the context of a folded structure. Therefore, it has proven to be useful, in particular, for predicting the three-dimensional structure of protein folds.[1]

Position pairs distant in the sequence but near space in the three-dimensional fold might also be expected to suffer replacement in a correlated fashion.[12,13] Many classes of these can be envisioned. For example, a "big-for-small" replacement at position i might be compensated by a coincident "small-for-big" replacement at position j, to conserve overall size in the packed core of the fold, and therefore conserve a functional behavior related to packing (the stability of the fold). Alternatively, a "positive-for-negative" charge replacement at position i might be compensated by a "negative-for-positive" charge replacement at position j to conserve overall charge, and therefore conserve a functional behavior related to net charge.

Behind this expectation stands a model based on the neutral theory of evolution.[14] Under this model, amino acid replacements that dramatically alter the physical property of the side-chain (e.g. its size or charge) will disrupt the performance of a protein already optimized to contribute to the fitness of the host organism. The fitness value of the protein can be restored only by a second change that compensates for the first alteration in physical properties. In the language of neutral theory, we would say that the first replacement was selectively disadvantageous, the second was positively selected (in the context of the first), and both together lead to a result that is neutral.

Analyses of compensatory replacement have found practical application. For example, a pair of compensatory changes in the protein kinase family underlay the successful prediction of the antiparallel sheet in the first kinase domain,[15] contradicting a prediction of a parallel sheet based on motif analysis.[16] This example represents the first example where compensatory covariation analysis was a key to a *bona fide* prediction of protein structure. More recently, Cohen and his co-workers presented an elegant example of compensatory replacement in phosphoglycerate kinase,[17] supporting the suggestion that correlated change in the evolutionary history of two protein families might be taken as evidence that the two proteins operate together. More generally, several laboratories have suggested that compensatory covariation might be used to detect incorrect folds in a structure prediction environment.[18–20]

Unfortunately, comparison of any two sequences diverging at n sites generates $n(n-1)/2$ pairs of candidate sites holding coincident replacements. These might be compensatory; they might be coincidental. Only a fraction of these will be truly compensatory, meaning that either replacement alone would have been rejected by natural selection without the other. Most replacements are presumably not compensated, either because they are neutral (implying that no other changes are needed to prevent their having a negative impact

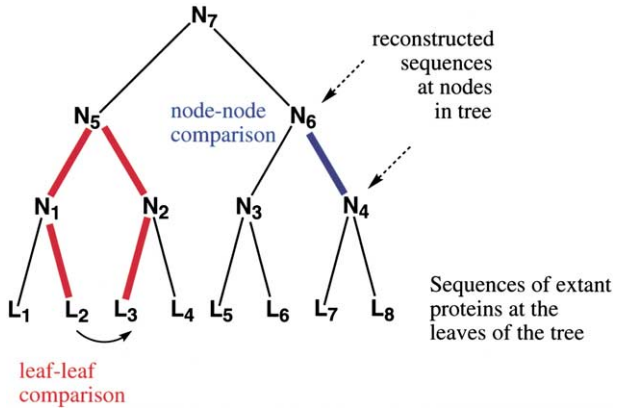


Figure 1. A leaf–leaf comparison (red) traverses more evolutionary distance than a node–node comparison (blue).

on functional behavior), or because they have a positive impact on fitness (implying that no other changes are desired to neutralize their impact on functional behavior).

Thus, while such compensation is easily recognized when analyzed in the context of a known crystal structure (the sites suffering compensatory replacements presumably are near in space), it is difficult to identify the pairs of sites that might suffer compensatory changes without a crystal structure. As a consequence, the compensatory covariation signal, represented by the number of pairs of replacement that are causally interrelated (where either alone would be rejected by natural selection) relative to the $n(n-1)/2$ pairs that arise whenever two protein sequences differing at n sites are compared, is weak, at least as it has been generally calculated.[21–23] Some have suggested that the compensatory replacement signal may never be broadly useful for this reason.[24] Others have suggested that the signal might have value, especially if it can be strengthened.

A variety of laboratories has recently explored approaches to strengthen the compensatory replacement signal.[18,19] For example, the signal for charge compensation appears to be stronger than the signal for size compensation, suggesting that it might be useful to analyze different types of coincident replacement separately. Further, compensatory covariation signals appear to be strongly dependent on the evolutionary distance between the two sequences, measured in PAM units (the number of point accepted mutations per 100 amino acids).[25] For example, charge compensatory changes were found to be more prominent at PAM 25 than at PAM 100.[19]

These observations are not surprising given the model. Charge reversal changes might be expected to be the change most in need of compensation. Likewise, at longer PAM distances, pairwise compensation might be obscured beneath compensation arising from replacements at multiple sites.

This work suggested a general strategy to strengthen the signal for compensatory replacement. At the core of this strategy is the recognition that compensatory replacements can be calculated in two ways. In the first, two extant protein sequences—sequences that are found in organisms living today—are examined. As extant sequences are at the leaves of an evolutionary tree, we call these "leaf–leaf" comparisons (Figure 1). In the past, virtually all compensatory substitution has been sought *via* leaf–leaf comparisons.

An alternative approach is possible. Given a set of homologous protein sequences, a multiple sequence alignment, and an evolutionary tree interrelating them, the sequences of ancestral proteins represented by nodes in the tree can be approximated using well-known heuristics. Given reconstructed sequences of ancestral proteins at nodes in an evolutionary tree, compensatory covariation can be sought using "node–leaf" and "node–node" comparisons.

As compensatory signals are stronger when the sequences being compared are separated by shorter distances[19] and as the distance between two nodes in a tree must be shorter than the distance between the leaves on the tree (Figure 1), node–node and node–leaf compensatory signals must be stronger than the leaf–leaf signal that contains them. Phrased differently, the number of replacements between average nodes is smaller than the number between average leaves. This means that the $n(n-1)/2$ total number of pairs is smaller, implying that the truly compensatory pairs, those driven by a fitness constraint, will be less obscured by the background of uncompensated events, when they are identified on individual branches of a tree.

Further, although the reconstructed ancestral sequences are probabilistic, and cannot be proven to be correct, even crude heuristics for reconstructing ancestral sequences localize specific changes to specific regions of a tree better than if no reconstructions are done at all. Thus, even poor reconstruction heuristics permit us to focus on briefer episodes of time during which two compensatory changes might have occurred than is possible by leaf–leaf comparisons.

Last, node–leaf and node–node comparisons model the actual evolutionary events that might contain the compensatory signal better than leaf–leaf comparisons. The statement that "position i and position j should suffer replacement in a compensatory way" is equivalent to the statement that if either position i or position j suffer replacement individually, the host organism is less "fit" (in a Darwinian sense). This, in turn, implies that if position i suffers a replacement then position j is under "positive selective pressure" to suffer a replacement. Conversely, this means that a replacement at position j will be fixed in a population faster than expected for a neutrally drifting position. This means that true compensatory changes will normally occur on the same

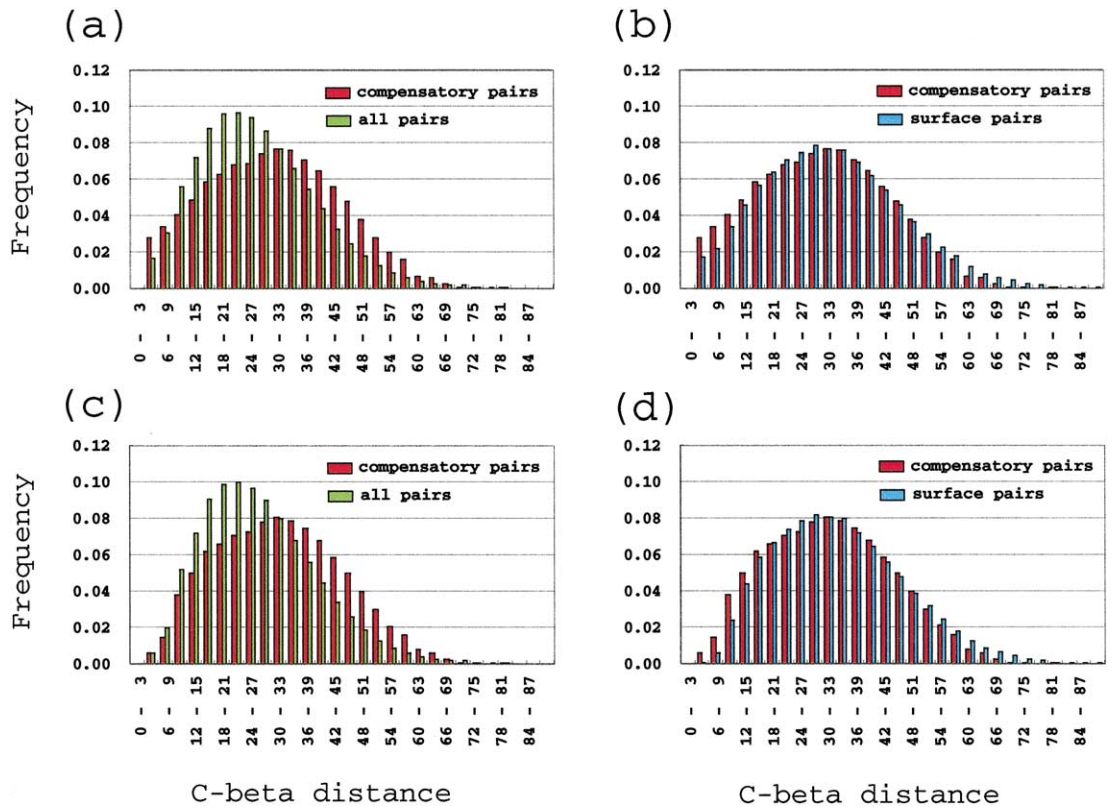


Figure 2. Attempted detection of charge compensatory covariation signal using leaf–leaf comparisons. (a) Distribution of distances in 71 protein families between position pairs displaying charge compensatory substitution (a positive-for-negative substitution at position *i*, and a negative-for-positive substitution at position *j*) (red bars), using as a reference (green bars) the distribution of all pairwise distances in the proteins. The *x*-axis is in angstroms; the *y*-axis is frequency, with each distribution normalized to unity. Note the absence of considerably taller red bars at short distances. Total observations in sample: 13,460. (b) As in (a), but using as a reference curve the distances between all pairs of surface residues (blue bars, >50% exposure, calculated by normalizing its accessible surface area (ASA) by the standard ASA of the residue. ASA was computed with the DSSP program.[46] Total observations in sample: 13,460. (c) As in (a), but considering only non-contiguous pairs, those where *i* and *j* are separated by more than four positions, with reference to a distribution of distances between all pairs of residues in the reference crystal structures (green bars). Total observations in sample: 12,811. (d) As in (c), but with reference to the distribution of distance between surface pairs (blue bars). Total observations in sample: 12,811.

branch of the tree, even on a very short branch of the tree.

This study examined the feasibility of detecting node–node and node–leaf compensatory covariation signals by examining reconstructed evolutionary events within families of proteins.

Results

A total of 71 families of proteins were examined in this study. For each family, a multiple sequence alignment and a phylogenetic tree were constructed using Clustal-W.[26] Reconstructed ancestral sequences at nodes throughout the tree were generated using the Fitch method as described in Methods.[27]

We then sought charge reversal replacements in these families. For each family, positions were identified where an amino acid replacement caused a charge reversal between two ancestral sequences (representing a replacement event that occurred on a branch between two nodes) or between a node sequence and a leaf sequence. Fractional changes were included. Each was associated with a particular branch of the tree.

From these, pairs of replacements displaying charge compensatory behavior were collected, as described in Methods. A pair of replacements was defined as being charge compensatory if they were coincident (both occurring on the same branch of the tree), if they each individually would reverse a charge, and taken together, if no change in the overall charge of the protein resulted from the two.

A three-dimensional crystal structure for a member of the protein family was then extracted from the PDB and an estimate was made for the strength of the signal. In making this estimate, we assumed that only proximal charge compensatory charges, near enough in space in the folded structure that the two side-chains could interact coulombically, could be functionally correlated. We therefore tabulated the distances between the sites in pairs that suffered charge compensatory replacements.

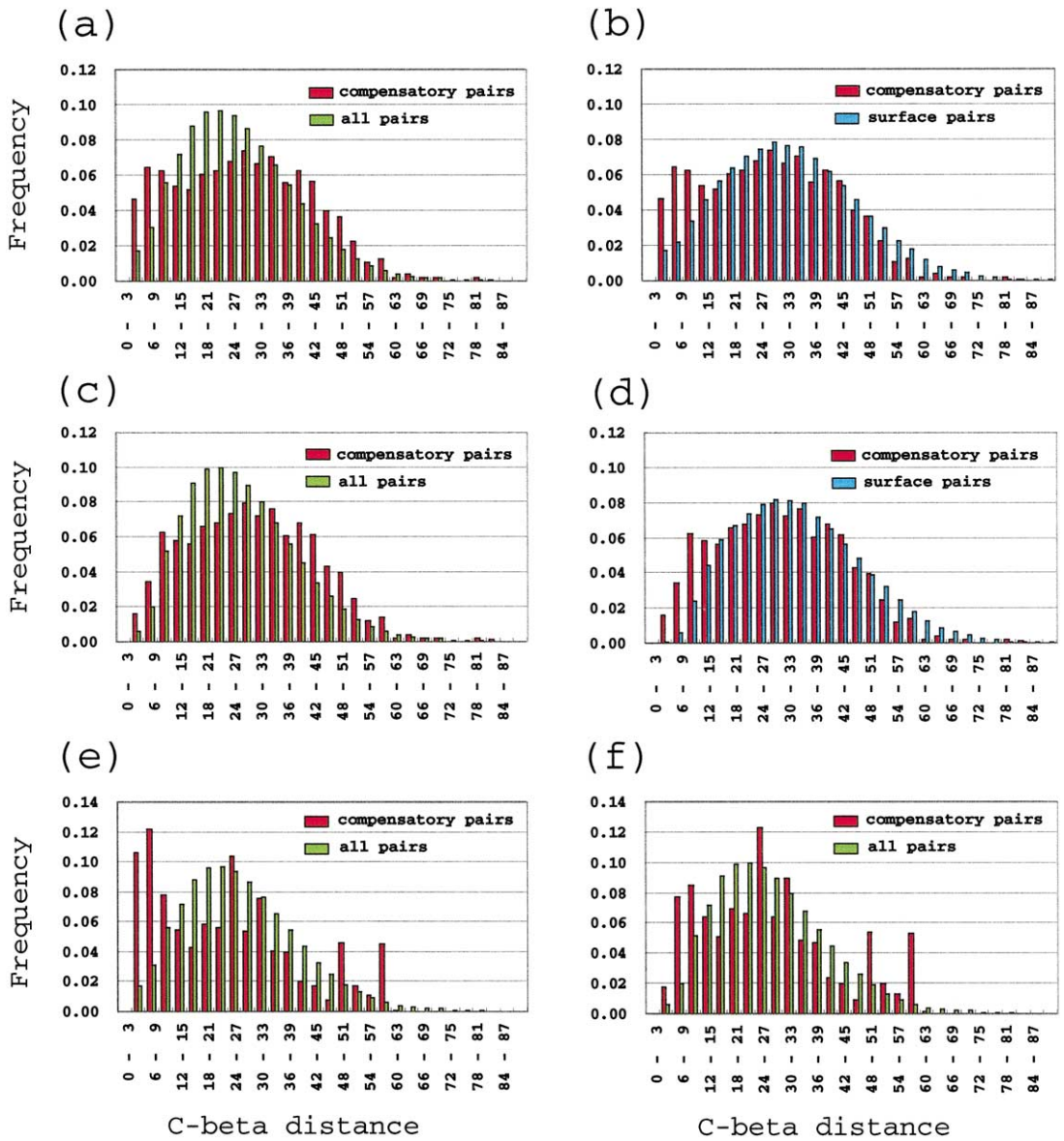


Figure 3. Detecting charge compensatory covariation signal using explicitly reconstructed ancestral sequences. (a) Distribution of distances in 71 protein families between position pairs displaying charge compensatory substitution (a positive-for-negative substitution at position *i*, and a negative-for-positive substitution at position *j*) (red bars), using as a reference (green bars) the distribution of all pairwise distances in the proteins. The *x*-axis is in angstroms; the *y* axis is frequency, with each distribution normalized to unity. Note the greater height of the red bars at both short distances and at long distances. Total observations in sample: 803.3 (note fractional number reflecting fractional character assignments in ancestral states; precision is less than implied by the decimal). (b) As in (a), but using as a reference curve the distances between all pairs of surface residues (blue bars, >50% exposure, calculated by normalizing its ASA by the standard ASA of the residue. ASA was computed with the DSSP program.[46] Note the greater height of the red bars at short distances only. This is the compensatory covariation signal. Total observations in sample: 803.3. (c) As in (a), but considering only non-contiguous pairs, those where *i* and *j* are separated by more than four positions, with reference to a distribution of distances between all pairs of residues in the reference crystal structures (green bars). Total observations in sample: 745. (d) As in (c), but with reference to the distribution of distance between surface pairs (blue bars). Total observations in sample: 745.5. (e) Distance distribution of charge compensatory pairs where only one such pair occurs on a specific branch of the evolutionary tree between reconstructed ancestral nodes. Total observations in sample: 57.2. Note the small sample size. (f) Same as (e), but where the position pair is separated by more than four positions in the linear sequence. Total observations in sample: 48.3. Note the small sample size.

Histograms were then constructed to show the distance distribution of pairs suffering compensatory replacements. As a standard, a distance distribution for all pairs of sites was calculated for the proteins. If the number of charge compensatory pairs was located disproportionately more in proximal positions than the average pair (where the square root of the number of pairs was a crude measure of significance), a signal was considered significant. As previous studies predicted, leaf–leaf comparisons found an only barely perceptible charge compensatory signal (Figure 2).

That is, pairs of positions suffering charge compensatory replacement in two extant sequences were not much more likely to be near in space than the average pair. The analogous analysis for each evolutionary branch in the tree, however, identified a signal that was clearly perceptible (Figure 3), even by eye. This signal was then analyzed.

Charge compensation and the surface of the folded protein

Surprisingly, the initial analysis (Figure 3(a)) showed that position pairs suffering node–node or node–leaf compensatory charge replacement were more likely to lie both proximally and distally in the fold, when compared with the average distance between all position pairs in the proteins. That is, pairs of positions near in the fold displayed charge compensatory covariation more frequently than the average pair, as expected should charge compensatory replacement be functionally correlated. But pairs of positions distant in the fold also displayed charge compensatory covariation more than the average pair.

Distal charge compensatory replacement suggested two explanations. First, overall net charge might be a selected trait in a protein. It is conceivable, for example, that a constant isoelectric point is desired by natural selection. If so, reversing a charge at one position would require an adaptive replacement leading to the opposite reversal somewhere (anywhere), even at a second position distant in the fold from the first.

Alternatively, charge changes are more likely to be tolerated by natural selection, and therefore are more likely to be observed, if they occur on the surface of the protein. The mean distance between a pair of surface residues is greater than the mean distance between the average pair of residues. Therefore, charge changes are more likely by chance to occur in pairs more distant than the average pair if they occur predominantly in positions on the surface, whether the pair is under direct selection or not (i.e. if the replacement is neutral).

These considerations suggested that the distance between the average position pair might not be the most informative reference distribution for these studies. Instead, an alternative reference distribution was calculated (blue bars, Figure 3(b)) for the distance between pairs of surface residues in the model proteins. This new distribution fit nicely the distal portion of the distribution of distances between position pairs displaying charge compensatory replacement. The result implies that the apparently high occurrence of charge compensatory replacement in distal pairs of residues reflects simply the greater likelihood that charge reversal replacements occur on the surface of the protein. This surface pair reference distribution does not, however, fit the proximal portion of the distribution. The probability that a pair of positions displaying charge compensatory replacement is near in the folded structure is considerably higher than expected (Figure 3(b)). This represents the "signal" in the charge compensatory pattern of replacement. When a side-chain changes its charge, that charge is more likely to be accompanied by a compensatory change in the charge of another side-chain near in space to the first.

Charge compensation in both contiguous and non-contiguous position pairs

We then explored several features of this signal. First, we asked whether the signal arises only in positions that were also nearby in the polypeptide sequence (contiguous pairs), or whether it was also observed in residue pairs >4 positions distant (those of i, $i+q$ relationship, where $q>4$) in the sequence (non-contiguous pairs). A strong signal was also observed for non-contiguous pairs (Figure 3(c) and (d)) as well as for contiguous pairs. This implies that a charge compensatory replacement signal arises when two residues are near in space as a consequence of the tertiary fold, as well as if they are near in space because they are near in the sequence.

Enhancing the charge compensation signal

These results showed that node–node and node–leaf analyses generated a more perceptible charge compensatory covariation signal than leaf–leaf analyses. Nevertheless, the signal remained small.

We considered several explanations for the small size of the signal. First, we considered a case where four sites suffer charge reversal replacements in a single evolutionary episode. Site a suffers a ($+$ to $-$) replacement compensated by a ($-$ to $+$) replacement at site b. Site c suffers a ($+$ to $-$) replacement compensated by a ($-$ to $+$) replacement at site d. Sites a and b are proximal; Sites c and d are proximal. Compensatory changes at sites b and d are required to maintain a functional protein given changes at sites a and c, respectively. Two pairs that do not represent adaptively significant compensation (a,d and b,c) arise in addition to the two pairs that do (a,b and c,d). This is simply another way of saying that changes that need compensation are more noticeable when the total number of changes is small.

Therefore, we asked whether the signal was stronger if the only branches examined were those holding exactly one pair of charge compensatory replacements (Figure 3(e) and (f), for all pairs and non-contiguous pairs, respectively). The likelihood that two positions undergoing charge compensatory replacement are near in space was indeed more perceptible after we excluded all of the events occurring on branches where more than two positions suffered charge reversal. The number of cases was small, however (57 for all pairs, and 48 for non-contiguous pairs, respectively), and the plots accordingly displayed substantial

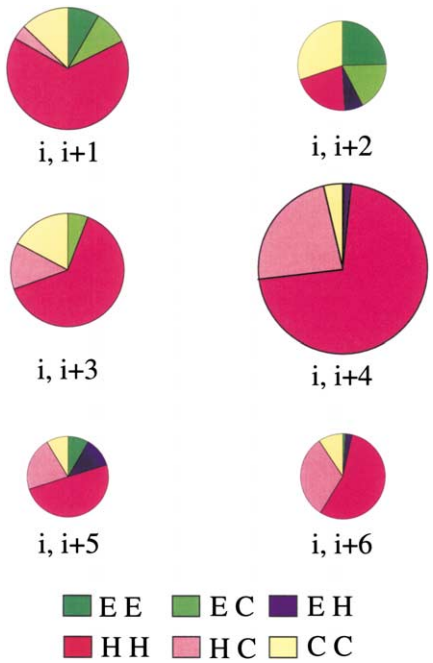


Figure 4. Predicting secondary structure using contiguous pairs of compensatory changes. The relative sizes of the "pies" indicates the relative numbers of examples of each pair. Where position pairs are separated by 1–5, or 6 positions, the likelihood that residue pairs are both found in helices (red), both found in strands (dark green), one in a helix and the other in a coil (pink), one in a strand and the other in a coil (light green), one in a helix and the other in a strand (violet), and both found in coils. Note an observation that if two charge compensatory substitutions are observed with a $i, i+4$ relationship, one or both are 98% likely to lie in a helix. The total number of observations; $i, i+1 = 12.82$; $i, i+2 = 6.47$; $i, i+3 = 11.38$; $i, i+4 = 23.37$; $i, i+5 = 4.79$; $i, i+6 = 6.48$. Note fractional values and the small size of these samples.

variances. As the number of sequences in the database grows, trees will become more articulated, individual compensatory events will be more likely to be isolated from others on a single branch of the tree, and the signal should strengthen.

Charge compensation in specific secondary structural elements

We then examined more closely the contiguous position pairs (i, $i+4$ or nearer) displaying charge compensatory covariation. The most striking feature of the charge compensatory signal within contiguous pairs was its dependence on the nature of the secondary structural element that held those positions (Figure 4). In 98% of the cases where the positions showing compensatory replacement had an $i, i+4$ relationship, one or both of the residues was found in an α helix. In only 1.6% of these cases was one of the residues found in a β strand, and in none of the cases were both found in a strand. This was significantly larger than the 47% of the position pairs with an $i, i+4$ relationship having one or both residues in a helix found in the dataset as a whole (Table 1). Conversely, in 49% of the cases where the position pair showing compensatory substitution had an $i, i+2$ relationship, one or both of the residues was found in a β strand.

Two alternative explanations can be proposed to account for the secondary structure bias. First, surface helices present residues to solvent (water) once every 3.6 turns. Surface residues are expected to be less constrained by function from diverging (that is, single replacements are more likely to be neutral), and are more likely than the average residue to suffer charge reversal substitutions. Therefore, the abundance of compensatory changes occuring with $i, i+3$ or $i, i+4$ relationship in the sequence might simply reflect unconstrained (i.e. neutral) charge reversal at surface positions.

Alternatively, loss of a coulombic interaction between residues i and $i+3$ or $i+4$ might lead to a protein less able to contribute to the fitness of the host. This view implies that once a charge reversal substitution occurs at position i, position $i+4$ is under sufficient positive selection to acquire a compensating charge reversal substitution.

Table 1. Frequencies of the average contiguous position pairs participating in a helix *versus* strand

	Relationship between the pair in the protein sequence					
Pair	$i, i+1$	$i, i+2$	$i, i+3$	$i, i+4$	$i, i+5$	$i, i+6$
EE	0.185	0.146	0.119	0.099	0.087	0.079
EC	0.077	0.148	0.192	0.219	0.234	0.239
Strand	0.262	0.294	0.311	0.318	0.321	0.318
HE	0.003	0.011	0.022	0.033	0.044	0.055
HH	0.307	0.274	0.243	0.223	0.208	0.194
HC	0.072	0.132	0.184	0.214	0.233	0.249
Helix	0.382	0.417	0.449	0.470	0.485	0.498
CC	0.356	0.289	0.241	0.212	0.195	0.185
Total	1.00	1.00	1.00	1.00	1.00	1.00
Total sample	12.8	6.5	11.4	23.4	4.8	6.5

EE, both positions involved in a charge compensatory event found in strands; EC, one in a strand and the other in a coil; Strand: HE, one in a helix and the other in a strand; HH, both in helices; HC, one in a helix and the other in a coil; Helix: CC, both in coils. Calculated from the test set of reference crystal structures using DSSP.[46] Note the small number of observations in each sample, and the fractional number of changes arising from parsimony reconstructions.

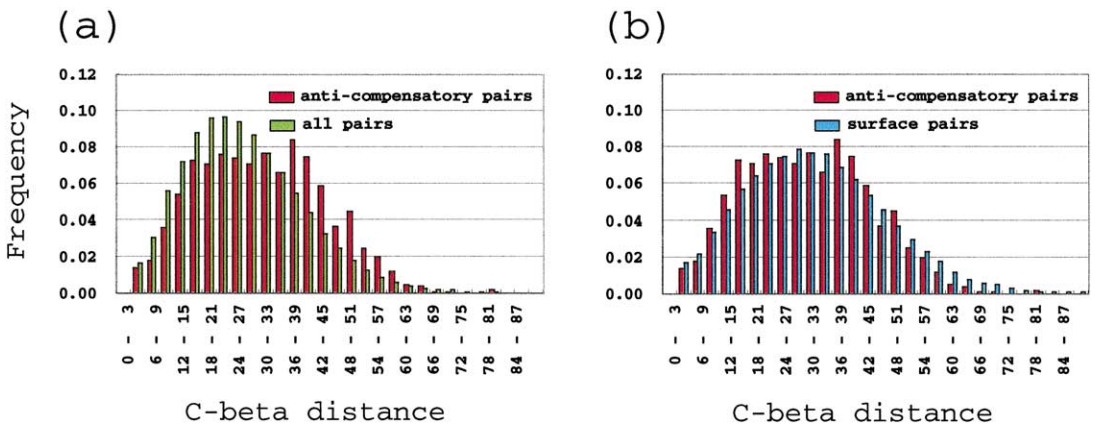


Figure 5. Distribution of distances between charge anti-compensatory pairs (red bars) (a) relative to all pairwise distances (green bars) and (b) relative to pairwise distances between all pairs of surface residues (blue bars, >50% exposure). Notable is the absence of the increased probability of proximal anti-compensatory pairs (compare with Figure 3). The total number of observations in both distributions is 793.3.

To explore these alternatives, we sought examples of anti-compensatory charge reversals, where a charge reversal (+ to −, for example) was accompanied by another charge reversal substitution in the same direction (+ to −) (Figure 5). The histogram showing the distance distribution in pairs suffering compensatory covariation is shown in Figure 5. Here, the distribution of distances between position pairs carrying charge anti-compensatory substitution was not noticeably different from the distribution of distances between all surface position pairs (Figure 3(b)). Notable is the absence of an increased probability of proximal anti-compensatory pairs (compare with Figure 3).

As a predictive tool, this enhanced charge compensatory signal may prove to be most valuable in secondary structure prediction. The accuracy of a prediction made on the relative positions of a compensatory pair is extremely high. The "coverage" is low, however. Only a few dozen examples were observed; only 6.8% of the helices contained within the 71 test families have one or more charge compensatory position pair. This number will, of course, grow as the size of the protein families grows with the increasing size of the genomic database (Table 2).

Charge compensation in buried residues

The high dielectric constant of water is known to weaken coulombic interactions between charged species. We therefore asked whether a stronger signal might be found in position pairs where one or more of the side-chains was buried. Figure 6(a) shows the distribution of surface accessibility calculated for all charged amino acids (DEKR, or Asp, Glu, Lys, and Arg) (blue), those that participate in a position pair suffering charge compensatory substitution (red), and those that participate in a position pair suffering charge anti-compensatory substitution (green). Not surprisingly, most buried charged residues do not suffer charge reversal within the test set. Compensatory changes near in space (Figure 6(b)) were slightly more likely to be found in positions that were more buried, a modest signal that suggests that compensation is more necessary in partly buried sites shielded from the high dielectric constant presented by the solvent water. This is the case for those residue pairs in protein kinase[15] and phosphoglycerate kinase[17] where charge compensatory substitution has had predictive value.

Discussion

Explicit reconstruction of ancestral proteins has been shown to provide insight into the structure and function of protein families, both when done *in silico*,[27–29] and when recombinant DNA technology is used to resurrect ancestral proteins from extinct organisms so they can be studied in the laboratory.[30–33]

The work reported here applies ancestral reconstructions towards a new goal. We suggest several conclusions. First, node–node and node–leaf comparisons between reconstructed ancestral sequences provide a stronger compensatory covariation signal than the leaf–leaf comparisons that have been used previously in the search for a compensatory covariation signal.

These results are consistent with the model outlined in Introduction, which holds that amino acid replacements that dramatically alter the physical property of the side-chain will disrupt the performance of a protein to an extent that requires a compensatory change elsewhere in the protein structure a significant number of times. The fitness value of the protein can be restored only by a second change that compensates for the first alteration in physical properties. In the language of neutral theory, we would say that the first replacement was selectively advantageous, the second was positively selected (in the context of the first),

Table 2. List of 71 protein families used in this analysis

ID	*L*	*N*	Protein name
121P	166	60	H-Ras P21 protein
193L	129	43	Lysozyme
1AAF	55	38	Hiv-1 nucleocapsid protein
1AAK	152	20	Ubiquitin conjugating enzyme
1ARS	396	28	Aspartate aminotransferase
1ATP(E)	350	20	c-AMP-dependent protein kinase
1BET	107	14	Beta-nerve growth factor
1BP2	123	64	Phospholipase A2
1CCR	112	93	Cytochrome *c*
1CPC(B)	172	44	C-Phycocyanin
1CYO	93	17	Cytochrome B5 (oxidized)
1DLH(A)	180	28	Hla-Dr1 (Dra, Drb1 0101) human class II histocompatibility protein (extracellular domain)
1DLH(B)	188	45	Hla-Dr1 (Dra, Drb1 0101) human class II histocompatibility protein (extracellular domain)
1EFT	405	57	Elongation factor Tu (Ef-Tu)
1FRP(A)	335	18	Fructose-1,6-bisphosphatase (D-fructose-1,6-bisphosphate 1-phosphohydrolase)
1FVL	70	29	Flavoridin
1FXI(A)	96	69	Ferredoxin I
1GDD	353	51	Gaunine nucleotide-binding protein G(I)
1GP1(A)	198	16	Glutathione peroxidase
1HAR	216	36	Hiv-1 reverse transcripts (amino-terminal half)
1HCN(A)	92	26	Human chorionic gonadotropin
1HDG(O)	332	109	Holo-D-glyceraldehyde-3-phosphate dehyrogenase
1HGE(A)	328	70	Hemagglutinin
1HLE(A)	345	14	Horse leukocyte elastase inhibitor
1HMT	132	16	Fatty acid binding protein
1HPM	386	128	44K Atpase fragment (N terminal) of 70 kDa heat-shock cognate protein
1HRA	80	42	Retinoic acid receptor
1HRY(A)	76	43	Human sry
1HTB(A)	374	56	Beta-3 alcohol dehydrogenase
1HTM(D)	138	92	Hemagglutinin ectodomain (soluble fragment, Tbha2)
1HUR(A)	180	35	Human Adp-ribosylation factor 1
1HUW	166	31	Human growth hormone
1HVD	319	23	Annexin V
1IRK	306	17	Insulin receptor (tyrosine kinase domain)
1ITG	166	42	Hiv-1 integrase (catalytic domain)
1IVD	388	23	Influenza A subtype N2 neuraminidase (sialidase)
1LDM	329	27	M4 lactate dehydrogenase
1MHC(A)	282	143	Mhc class I antigen H2-M3
1MLS	154	74	Myoglobin
1NDH	272	27	Cytochrome B5 reductase
1NHK(L)	144	26	Nucleoside diphosphate kinase
1NIP(A)	289	35	Nitrogenase iron protein
1OCT(C)	156	33	Oct-1 (Pou domain)
1OSA	148	64	Calmodulin
1PBX(A)	143	56	Hemoglobin
1PDN(C)	128	28	Prd paired domain
1PLQ	258	13	Proliferating cell nuclear antigen
1PVC(2)	271	24	Poliovirus type 3, Sabin strain
1REC	201	21	Recoverin
1SXC(A)	151	51	Superoxide dismutase
1TGX(A)	60	51	Toxin gamma
1TIV	86	24	Hiv-1 transactivator protein
1TPH(1)	247	35	Triosephosphate isomerase
1YTB(A)	180	18	TATA-box binding protein
1ZAA(C)	87	18	Zif268 immediate early gene
2BTF(A)	375	121	Beta-actin-profilin complex
2CPL	165	35	Cyclophilin A
2GDM	153	23	Leghemoglobin
2HMX	133	40	Human immunodeficiency virus type 1 matrix protein
2HPE(A)	99	36	Hiv-2 protease
2REB	352	58	Rec A protein
2TGI	112	18	Transforming growth factor-β 2
3RUB(S)	123	79	Ribulose 1,5-bisphosphate carboxylase/oxygenase (form III)
4ENL	436	35	Enolase
4FGF	146	16	Basic fibroblast growth factor
4GCR	174	31	Gamma-B crystallin
4MT2	62	50	Metallothionein isoform II
4RHV(1)	289	24	Rhinovirus 14
4RHV(3)	236	24	Rhinovirus 14
7RSA	124	48	Ribonuclease A
8CAT(A)	506	34	Catalase

ID, Protein Data Bank identifier, protein subunit in parentheses; *L*, chain length; *N*, number of sequences in multiple sequence alignment.

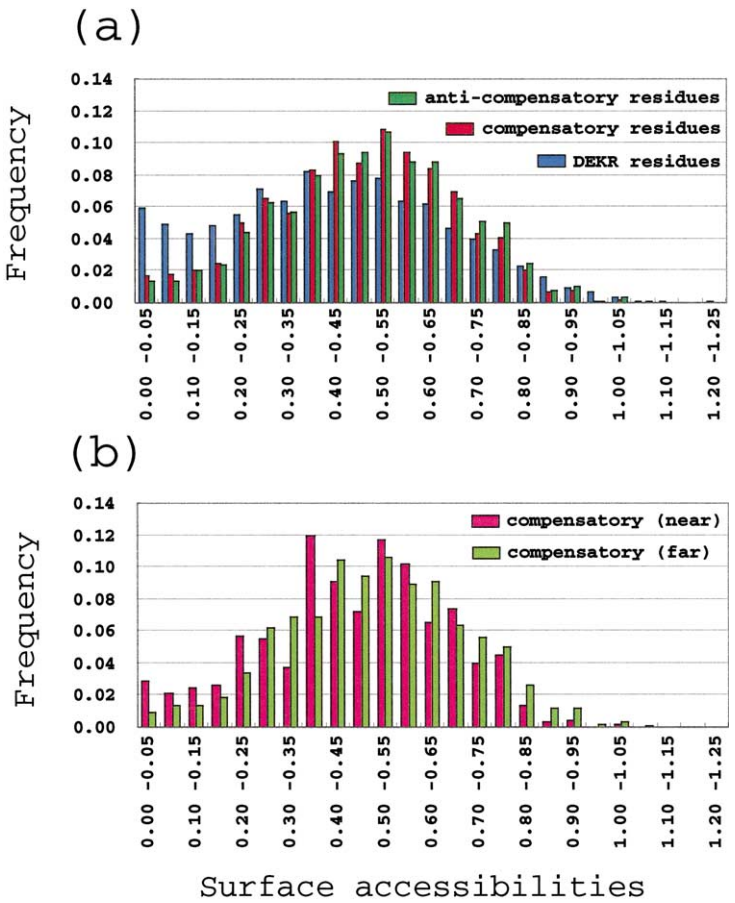


Figure 6. Surface accessibility of charged residues, and charged residues participating in a charge compensatory event. (a) A frequency *versus* accessibility distribution for all Asp, Glu, Lys, and Arg (DEKR) residues (blue), those participating in a charge compensatory (red) and anti-compensatory (green) events. Bars of each color sum to unity. Buried DEKR residues are less likely than average to suffer charge reversal substitution. Compensated charge reversal substitutions are slightly more likely than anti-compensated charge reversal substitutions. The total number of observations in sample: anti, 2812.4; comp, 2792.2; DEKR, 3530.0. (b) A frequency *versus* accessibility distribution showing the frequency of compensated charge reversals for proximal pairs near in space (less than 12 Å distant, red) in the folded structure, and distal pairs distant in the folded structure (more than 12 Å distant, green). A pair of charged compensatory substitutions is slightly more likely to be near in space if the side-chains involved are more buried. The total number of observations in sample: near, 367.1; far, 795.7.

and both together are neutral. Thus, while the two compensatory substitutions taken together might be regarded as collectively neutral, the second mutation in a truly compensatory pair is viewed as positively adaptive in this model.

These results also suggest that as the database grows, the charge compensatory signal will become more perceptible as more sequences are added to each family. More sequences mean more highly articulated evolutionary trees. This, in turn, means that compensatory events will become better isolated on specific branches, preventing the "spurious" signals that arise when more than one pair of compensatory events occurs along a specific branch on a tree.

The stronger signal will undoubtedly find use. Predicting secondary structure using contiguous pairs of compensatory changes is one. It remains to be seen, however, how much data in a family is required for the signal to be useful to support *de novo* assembly of a protein fold in a prediction setting.

Accounting for the stronger signal from node–node comparison

The tools that we have presented make the compensatory covariation signal more perceptible. Isolation of truly compensatory pairs on shorter branches of a tree away from other changes is, we believe, sufficient explanation for this effect. Shorter branches implies a smaller n, the total number of differences between the two protein sequences being compared, diminishing the number of $n(n-1)/2$ pairs behind which the compensatory signal might be obscured.

While it is possible in principle to construct a statistical model that permits double replacements to be evaluated, this requires a high degree of empirical parameterization. For example, nearly a decade ago, we collected the parameters needed to build a statistical model that concerned double replacements at adjacent sites.[11] Some 220 parameters are formally required in this exercise, a large number by most measures.

For the purpose of this work, a signal is considered to be significant if it lies two standard deviations outside of the fluctuation expected for n sites at a specified distance. This can be estimated by the square root of the number of sites. By this measure, all of the results reported here are strongly significant, with the exception of those reported in Figures 3(e),(f), and 4.

A model-independent method to evaluate an evolutionary tree

The strength of the compensatory covariation signal undoubtedly depends on the degree to which the trees and the reconstructed ancestral

sequences accurately reflect the history of the family. If the branching of the tree or the reconstructed sequences themselves are not correct, a pair of charge compensatory replacements that are coincident, in fact, may not be assigned to the same branch of a tree. In this case, the signal from this pair will be lost.

Getting the branching correct in an evolutionary tree is a difficult problem. Part of the difficulty arises because of the trade-off between the accuracy of the tree and the cost of generating it. For example, the Clustal-W[26] and Fitch parsimony tools used here are relatively inexpensive methods for reconstructing trees and ancestral sequences. Clustal-W uses a neighbor joining tool[34] based on estimates of the distances between sequence pairs derived from the Kimura empirical formula.[14] Ancestral sequences reconstructed by parsimony are well known to be sensitive to incorrect branching topology. This is the principal error associated with the choice of this inexpensive reconstruction tool.

More sophisticated methods, including maximum likelihood methods, are expected to provide better trees, at least given the first-order stochastic models. These are expected to generate ancestral reconstructions that are more robust to errors in tree topology. They are, however, more expensive.

Even the more expensive tools do not guarantee a correct tree, of course. In practice, the approximations made in the model (see Introduction) may create systematic error larger than fluctuation error. To date, the only way to benchmark a tree requires knowledge of the evolutionary history of the sequences in question,[35] or a reconstruction of a simulated evolutionary process.[36] The first is difficult to get for sequences emerging from natural history. The second requires a mathematical model for evolution, which is often the same one that is used to construct the tree in the first place.

Here, the compensatory covariation signal, extracted from reconstructed ancestral sequences, may provide a metric for the quality of a tree based on organic chemistry, independent of any mathematical model for evolution. Hypothetically, the best tree should be the tree that places compensatory replacements truly driven by natural selection on the same branch. This requires the construction of a tree that reflects the actual evolutionary history. This, in turn, implies that the tree that has the most compensatory covariation is the tree that is most likely to reflect the actual history.

To illustrate this application, consider four hypothetical proteins, just four amino acids in length, having the sequences ALKD, MVKD, ALER, and MVER. Exactly two topologies exist for unrooted trees that relate these four sequences (Figure 7). Both reconstructions have two ambiguous sites in both ancestors. In Topology I, the first two positions are ambiguous, and in Topology II, the last two positions are ambiguous. Both trees require four "homoplastic" events (independent mutations that cause sequence convergence). Both trees require exactly six changes. Classical parsimony therefore ranks these two topologies as equally likely.

The two topologies are different, however, with respect to the extent to which charge changes are compensated. In Topology I, a charge altering replacement is 100% likely to be compensated. In Topology II, however, a charge altering replacement is only 50% likely to be compensated. This is illustrated in Figure 7 by writing out four trees, each equally likely, that carry reconstructions that the ambiguities require. If we postulate that compensatory covariation is maximized, then Topology I is preferred to Topology II.

Conversely, an analogous logic can be used to assign preferred ancestral states involving charged residues. For Topology I, the ancestral states involving charged residues are fixed. For Topology II, the preferred ancestral sequences are in reconstructions IIa and IIb.

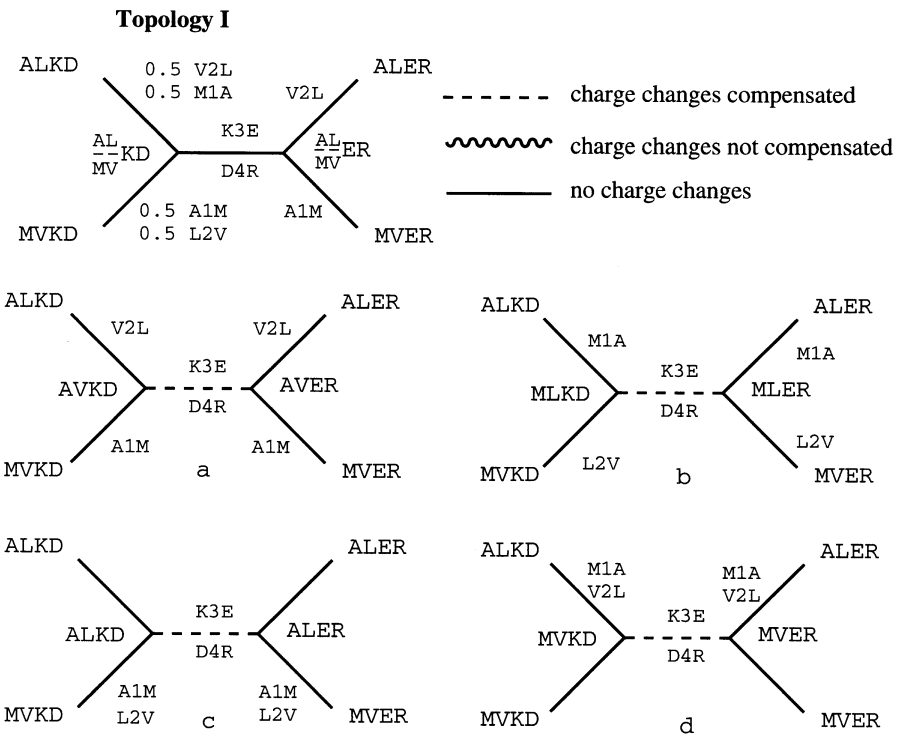
This metric can be applied even if no crystal structure is available for a protein family. If, however, a crystal structure is available, then (as a practical matter) one would maximize the number of proximal charge compensatory changes when identifying the preferred tree.

It will require much future work with many families to determine how useful the metric will be. Worth noting at this point, however, is that this metric is rooted in principles of structural biology (that is, organic chemistry), not in a mathematical formalism. Further, the proposed metric values changes at position i in light of changes at position j. Thus, this metric for evaluating the quality of a tree is fundamentally different from any metric based on first-order stochastic analyses of protein sequences, which treat replacements at site i and site j as independent.

Darwinian requirements for compensatory covariation

Even given these results, and the evidence that the charge compensatory substitution signal can only become stronger as the database grows, it remains inescapable that the charge compensatory signal is weak, perhaps even weaker "than expected". What might be the scientific implications of this observation?

Charge compensatory covariation might be weak because the coulombic interactions being sought may themselves be largely unimportant to the selective fitness of proteins. Gaining or losing them, in this view, has insufficient impact on fitness to ensure that natural selection will require compensation, and thereby prevent uncompensated charge reversals from entering the global proteome. This implies a limit to the tool generally, one imposed by the physical organic chemistry of folded protein sequences.


Topology I
ALKD
0.5 V2L
0.5 M1A
V2L
ALER
K3E
AL MV KD
AL MV ER
D4R
0.5 A1M
0.5 L2V
A1M
MVKD
MVER
charge changes compensated
charge changes not compensated
no charge changes
ALKD
V2L
V2L
ALER
K3E
AVKD
AVER
D4R
A1M
A1M
MVKD
a
MVER
ALKD
M1A
ALER
K3E
M1A
MLKD
MLER
D4R
L2V
MVKD
L2V
b
MVER
ALKD
ALER
K3E
ALKD
ALER
D4R
A1M
A1M
MVKD
L2V
c
L2V
MVER
ALKD
M1A
V2L
M1A
V2L
ALER
K3E
MVKD
MVER
D4R
MVKD
d
MVER
Topology II
ALKD
ALER
0.5 E3K
0.5 R4D
0.5 K3E
0.5 D4R
AL KD ER
A1M
V2L
MV KD ER
0.5 E3K
0.5 R4D
0.5 K3E
0.5 D4R
MVKD
MVER
charge changes compensated
charge changes not compensated
no charge changes
ALKD
ALER
K3E
D4R
ALKD
A1M
V2L
MVKD
K3E
D4R
a
MVKD
MVER
ALKD
ALER
E3K
R4D
ALER
A1M
V2L
MVER
E3K
R4D
b
MVKD
MVER
ALKD
ALER
R4D
K3E
ALKR
A1M
V2L
MVKR
R4D
K3E
MVKD
c
MVER
ALKD
ALER
E3K
D4R
ALED
A1M
V2L
MVED
E3K
D4R
MVKD
d
MVER

An alternative explanation should be considered, however. Observation of a compensatory pair of substitutions implies, under the neutral theory, that natural selection preserved some global feature of a protein during the episode represented by the branch between two nodes. This, in turn, implies some degree of constancy in the behavior of the protein before and after the episode where compensatory change has occurred.

In this view, compensatory replacement should be observed only in protein families whose behavior must remain largely constant during this branch. This, in turn, implies that compensatory covariation should be observed only during episodes where function, defined as the behavior that contributes to fitness, is largely conserved. In the language of the neutral theory, the demand for compensation arises because the protein is optimized at the beginning of the episode for fitness, the same behaviors are optimal at the end of the episode, and any replacements occurring during the episode must have the net (and, if necessary, combined) impact of being neutral with respect to their impact on selected behavior.

This implies, of course, that when functional behavior is changing, there may be no need to compensate individual replacements in a sequence, even those that reverse charge. Indeed, an uncompensated change may be more likely to generate a protein with different behaviors, whose (now) different behaviors contribute most to the (now different) requirements for fitness. In this view, compensatory covariation should not be observed, or should be observed less frequently, whenever functional behavior is changing.

In this view, compensatory covariation is scarce because branches of an evolutionary tree where functional behavior is rigorously conserved are scarce. This is, of course, a controversial suggestion. Many computational biologists treat homologous proteins in distinct organisms as if they were "the same protein", and neutral theory remains the majority view of protein sequence evolution. In contrast, recent work in these laboratories and elsewhere has suggested that functionally significant divergence in behavior is frequent, and may be the rule more than the exception. For example, it is almost certainly observed in elongation factors, regarded as some of the most functionally conserved proteins in the biosphere.[5]

Given this observation, compensatory replacements may become a powerful tool in functional genomics to detect episodes where function is, and is not, conserved. A branch that has more compensatory replacement is more likely to represent an episode where functional behavior is constant than one with less compensatory replacement.

This is relevant to the issue of "annotation transfer" in comparative proteomics. Annotation transfer assigns the function of a new protein by identifying in the database a homologous protein for which the function is known, and transferring annotation describing that function to the annotation for the new protein. Annotation transfer assumes that function does not change within a set of homologous protein sequences as they diverge.[37,38] This assumption has long been known to be poor in many proteins, including many characterized before the dawn of the age of the genome.[39]

To date, several tools have been suggested to detect functional change. One of these is to measure K_a/K_s values for branches of an evolutionary tree between reconstructed ancestral sequences at the nodes of the tree.[28,40] Here, compensatory changes would indicate functional constancy, while uncompensated changes would indicate functional change. Because compensatory analysis rests on protein sequences, while the K_a/K_s value requires measurement of silent substitution rates, and because silent substitution rates are frequently rather high, this metric for functional recruitment may ultimately prove to be more valuable than K_a/K_s ratios, especially for deeply branching sequences.

Methods

The core of this study exploited the PDB "select 25" subset of proteins.[41] Each protein in this database was matched against the proteins contained in SWISS-PROT (version 33).[42] The older sequence dataset was chosen to avoid under-annotated sequences, in particular, pseudogenes that might be divergently evolving without functional constraints. Families that contained at least

Figure 7. A schematic illustration of the use of compensatory covariation to select a preferred tree from two equally parsimonious trees. The two tree topologies relating the four sequences (ALKD, MVKD, ALER, and MVER) each require six changes. The changes are marked on individual branches, with fractional changes arising from the ambiguity in the ancestral sequences. The ancestral sequences are placed at the nodes in the tree, with ambiguous sites (by parsimony) noted by placing the two possible residues above and below a horizontal line. For each topology, identical trees holding all four possible ancestral sequences are shown. Each, by parsimony, has equal likelihood (0.25 for each). In Topology I, the ancestral sequences are ambiguous at the first two positions and in Topology II, at the last two positions. Both trees require the same amount of homoplasy (convergence). Classical parsimony analysis is indifferent with respect to the two topologies. In Topology I, however, the likelihood that a charge reversal is compensated is unity. In Topology II, it is only 0.5. Thus, Topology I is preferred if compensatory covariation is maximized. This criterion is independent of mathematical formalisms used to construct the tree. Further, the metric weights changes at position i depending on events at position j, making this metric for evaluating a tree fundamentally different from any metric based on a first-order stochastic analysis of protein sequences.

12 members, where the maximum evolutionary distance between any pair of sequences in the family was between 50 and 120 PAM units, and where the family had at least two subfamilies defined at PAM 20 with four or more members, were retained. These criteria, made to ensure a balanced tree, were satisfied by 71 families.

The sequences within each family were aligned using the MultAlign package[43] with the option PROB from Darwin system.[44]†[45] The gap shifting heuristic was applied iteratively until the overall alignment score ceased to improve. Secondary structure assignments were extracted from the crystallographic data using DSSP.[46]

An evolutionary distance matrix and a phylogenetic tree were computed for each family using Clustal-W,[26] which employs a neighbor joining method using distances derived from Kimura empirical formula.[14] Branches with negative lengths were ignored.

Probabilistic reconstructed ancestral sequences at nodes in the tree were then built using the Fitch parsimony method.[27] For those sites where parsimony did not generate a single assigned residue in an ancestral sequence, fractional probabilities were assigned to each of the contender amino acids by the statistical method described by Fitch.[27] The statistical method was used to assign probabilities to each possible path between the residue at a site at an ancestral node residue (either a single amino acid or a set of possible amino acids each with an assigned probability) and its two descendant residues (either a residue in an extant species' sequence (a node–leaf comparison) or a residue (or set of possible amino acids) in another ancestral species (a node–node comparison)).

Using probabilistic reconstructed ancestral sequences computed in this way, charge compensation was sought on individual branches of the evolutionary trees. For each branch in each tree, the sequences at the flanking nodes were compared (residue by residue) to identify single substitutions that reversed charge. A position was retained if and only if one of the following transition probabilities ($K \rightarrow E$, $K \rightarrow D$, $R \rightarrow E$, $R \rightarrow D$, $E \rightarrow K$, $E \rightarrow R$, $D \rightarrow K$, $D \rightarrow R$) was >0.2. If more than one probability was >0.2, then both transition probabilities were retained for that position.

Coordinates locating the position of C-β were then extracted from the reference crystal structure with the PDB. When the residue was Gly, the position of the β carbon atom was inferred from the position of the backbone atoms. A Perl script was written to permit calculation of the distances between the β carbon atoms for each pair of amino acids retained. This generated the distances reported in the Figures. The accessible surface area (ASA) was computed using the DSSP program.[46]

A list was then made of all pairs of positions having compensatory charge reversals for each branch of the tree. These were defined to be "position pairs" undergoing charge compensatory covariation. The distance between the β carbon atoms of the position pairs was inferred from the positions of those two carbon atoms in the reference crystal structure. The pairwise distance was then calculated for all amino acids in the reference crystal structure, and then for the distance between pairs of residues on the surface of the reference structure.

† The alignment program implements a rigorous measure of the quality of the alignment, and a gap shifting heuristic that is consistent with the known behavior of gapping in natural proteins.

Acknowledgments

We are indebted to the National Institutes of Health (grant MH 55479) and the NASA Astrobiology Institute for partial support of this work. This work was also supported in part by a Grant-in-Aid for Science Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. We are grateful to Drs Eric Gaucher and Mike Chang for reading the final version of this manuscript.

References

1. Benner, S. A., Cannarozzi, G., Chelvanayagam, G. & Turcotte, M. (1997). *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* **97**, 2725–2843.
2. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443–453.
3. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
4. Thorne, J. L., Kishino, H. & Felsenstein, J. (1992). Inching toward reality. An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**, 3–16.
5. Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. (2001). Functional genomics using covarion-based evolutionary analysis. *Proc. Natl Acad. Sci. USA*, **98**, 548–552.
6. Rost, B., Sander, C. & Schneider, R. (1994). PHD. An automatic server for protein secondary structure prediction. *CABIOS*, **10**, 53–60.
7. Benner, S. A., Trabesinger-Ruef, N. & Schreiber, D. R. (1998). Post-genomic science. Converting primary structure into physiological function. *Advan. Enzyme Regul.* **38**, 155–180.
8. Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S. & Knecht, L. (2000). Functional inferences from reconstructed evolutionary biology involving rectified databases. An evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**, 97–106.
9. Liberles, D. A., Schreiber, D. R., Govindarajan, S., Chamberlin, S. G. & Benner, S. A. (2001). The adaptive evolution database (TAED). *Genome Biol.* **2**, 3.1–3.18.
10. Miyamoto, M. M. & Fitch, W. M. (1995). Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**, 503–513.
11. Cohen, M. A., Benner, S. A. & Gonnet, G. H. (1994). Analysis of mutation during divergent evolution. The 400 by 400 dipeptide mutation matrix. *Biochem. Biophys. Res. Commun.* **199**, 489–496.
12. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. (1987). Correlation of coordinated amino-acid substitutions with function in tobamoviruses. *Protein Eng.* **1**, 228–236.
13. Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino-acid changes in homologous protein families. *Protein Eng.* **2**, 193–199.
14. Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, New York.
15. Benner, S. A. & Gerloff, D. L. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure. The catalytic

domain of protein kinases. *Advan. Enzyme Regul.* **31**, 121–181.
16. Sternberg, M. J. E. & Taylor, W. R. (1984). Modeling the ATP binding site of oncogene products, the epidermal growth-factor receptor and related proteins. *FEBS Letters*, **175**, 387–392.
17. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293.
18. Olmea, O., Rost, B. & Valencia, A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**, 1221–1239.
19. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H. & Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**, 307–316.
20. Chelvanayagam, G., Knecht, L., Jenny, T. F., Benner, S. A. & Gonnet, G. H. (1998). A combinatorial distance constraint approach to predicting protein tertiary models from known secondary structure. *Fold. Design*, **3**, 149–160.
21. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358.
22. Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.
23. Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98–102.
24. Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348.
25. Dayhoff, M. O., Schwartz, R. M. & Orcott, B. C. (1978). *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, p. 345, Nat. Biomed. Res. Found., Washington, DC.
26. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Clustal-W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
27. Fitch, W. M. (1971). Toward defining the course of evolution. Minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416.
28. Messier, W. & Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, **385**, 151–155.
29. Trabesinger-Ruef, N., Jermann, T. M., Zankel, T. R., Durrant, B., Frank, G. & Benner, S. A. (1996). Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Letters*, **382**, 319–322.
30. Benner, S. A. (1988). Reconstructing the evolution of proteins. In *Redesigning the Molecules of Life* (Benner, S. A., ed.), pp. 115–175, Springer, Heidelberg.
31. Malcolm, B. A., Wilson, K. P., Mathews, B. W., Kirsch, J. F. & Wilson, A. C. (1990). Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, **345**, 86–89.
32. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. (1990). The ribonuclease from an extinct boivd. *FEBS Letters*, **262**, 104–106.
33. Jermann, T. M., Opitz, J. G., Stackhouse, J. & Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, **374**, 57–59.
34. Saitou, N. & Nei, M. (1987). The neighbor-joining method. A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
35. Hillis, D. M., Huelsenbeck, J. P. & Cunningham, C. W. (1994). Application and accuracy of molecular phylogenies. *Science*, **264**, 671–677.
36. Takahashi, K. & Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol.* **17**, 1251–1258.
37. Hegyi, H. & Gerstein, M. (2001). Divergence in multidomain proteins. *Genome Res.* **11**, 1632–1640.
38. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T-1 ribonucleases. *J. Mol. Biol.* **281**, 949–968.
39. Benner, S. A. & Ellington, A. D. (1988). Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* **23**, 369–426.
40. Li, W. H., Wu, C. I. & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* **2**, 150–174.
41. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
42. Bairoch, A. & Boeckmann, B. (1991). The Swiss-Prot protein sequence data bank. *Nucl. Acids Res.* **19**, 2247–2250.
43. Korostensky, C. (2000). *Algorithms for Building Multiple Sequence Alignments and Evolutionary Trees*, Diss. no. 13550, ETH, Zurich.
44. Gonnet, G. H. & Benner, S. A. (1991). *Computational Biochemistry Research at ETH*, Technical report 154, Department Informatik, Swiss Federal Institute of Technology, Zurich.
45. Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **229**, 1065–1082.
46. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Edited by F. E. Cohen

(Received 30 May 2001; received in revised form 25 January 2002; accepted 13 March 2002)