5 Analysis of Coordinated Substitutions in Proteins

David Campo

CONTENTS

> I mean by this expression (Correlation of growth) that the whole organization is so tied together during its growth and development, that when slight variations in any one part occur, and are accumulated through natural selection, other parts become modified.
>
> **Darwin. C, *The Origin of Species* [1]**

5.1 INTRODUCTION

One way of studying protein structure and function is to carry out site-directed mutagenesis where specific residues within a protein are altered, and then to examine the effects of these changes on protein characteristics. Changes in the amino acid (aa) properties (e.g., hydrophobicity, volume, and charge) of the mutated sites can then be correlated with changes in protein characteristics [2]. Another approach is to analyze large families of naturally occurring proteins or protein domains. During divergent evolution, protein sequences change through genetic drift, while the biochemical function of the protein is substantially retained. It is known that the number of sequences exceeds the number of structures by several orders of magnitude and, therefore, the number of three-dimensional protein structures corresponding to a given function is small, from one (like the hand-shaped structure of nucleic acid polymerases) to a small number (for example the four families of endoproteases) [3,4]. The core conformation of homologous proteins persists long after the statistically significant sequence similarities have vanished [5] and this persistence underlies all

tools where a function is predicted or a three-dimensional model of a protein is built by extrapolation from an experimental structure of a homologue sequence [6].

By examining patterns of sequence diversity, one can explore how naturally occurring sequence variability and aa properties are important in maintaining protein structure. Analysis of the variation in aa at different sites allows the understanding of the structural–functional role of residues at these positions and to predict protein structure [7]. Some functionally important protein sites are easily detected since they correspond to conserved columns in a multiple-sequence alignment (MSA) but non-conserved sites are also interesting as they may be functionally or structurally important, or possibly key sites of interaction between the protein and its substrate [8]. Experimental and quantitative analyses of proteins often assume that the protein sites are independent, i.e., the presence of a residue at one site is assumed to be independent of residues at other sites. However, the activities and properties of proteins are the result of interactions among their constitutive aa and this leads to the hypothesis that in the course of evolution, substitutions which tend to destabilize a particular structure are probably compensated by other substitutions which confer stability on that structure [9]. Interactions among aa sites include salt bridges between charged residues, hydrogen bonds between electron acceptors and donors, size constraints reflecting structural interactions between large and small side chains, electrostatic interactions, hydrophobic effects, van der Waal's forces, and similar phenomena [2]. It is reasonable to suppose that sites that can compensate for a destabilizing substitution at another site are likely to be close to this site in the three-dimensional structure of the protein. For example, if a salt bond were important to structure and function, a substitution of the positively charged residue with a neutral residue would need to be compensated by a nearby residue substituting from a neutral to a positive residue (Figure 5.1). Similarly, a substitution involving a reduction of volume in the protein core might cause a destabilizing pocket which only one or a few adjacent residues would be capable of filling. Thus, if structural compensation is a general phenomenon, sites which are close together in the three-dimensional structure will tend to evolve in a correlated fashion due to the compensation process [9].

There is experimental evidence indicating that proteins contain pairs of covariant sites, which were found both by analysis of the families of natural proteins with known structures [10–15] and in proteins into which point mutations have been introduced by site-directed mutagenesis [16–18]. In these examples, sites distant in the sequence but near in three-dimensional space in the folded structure have been observed to undergo simultaneous compensatory variation to conserve the overall volume, charge, or hydrophobicity [6]. Several experiments also have provided strong evidence of compensatory mutations in the fast evolution of RNA viruses [15,19–24].

Independent mutations among functionally linked sites would be disadvantageous but simultaneous or sequential compensating mutations may allow the protein to retain function [25]. Furthermore, there are constraints on aa replacements that arise for functional reasons, such as aa bias at recognition sites related to DNA binding in transcriptional regulators. Evolutionarily related sequences should contain the vestiges of these effects in the form of covariant pairs of sites [26] and these interactions can be manifested in covariation between substitutions at pairs of alignment positions in a MSA. The analysis of covariation has been used in protein-engineering approaches [27], sequence-function correlations [2,28], protein structure prediction methods [13,26,29–39], and in finding important motifs in viral proteins [40–43]. However, early studies did not optimally discriminate the three different sources of covariation (1) chance, (2) common ancestry, and (3) structural or functional constraints. Effectively discriminating among these underlying causes is a difficult task with many statistical and computational difficulties. Improved analyses that discriminate the different sources of covariation confirmed that highly coordinated sites are often functionally related or spatially coupled [2,6,8,22,44,45].

Covariation analyses can also be important in identifying sites that may change the phenotype of a protein, and they could be used as a tentative map for researchers attempting to define functional domains in the protein through mutational analysis. For instance, covariant sites could be used as a

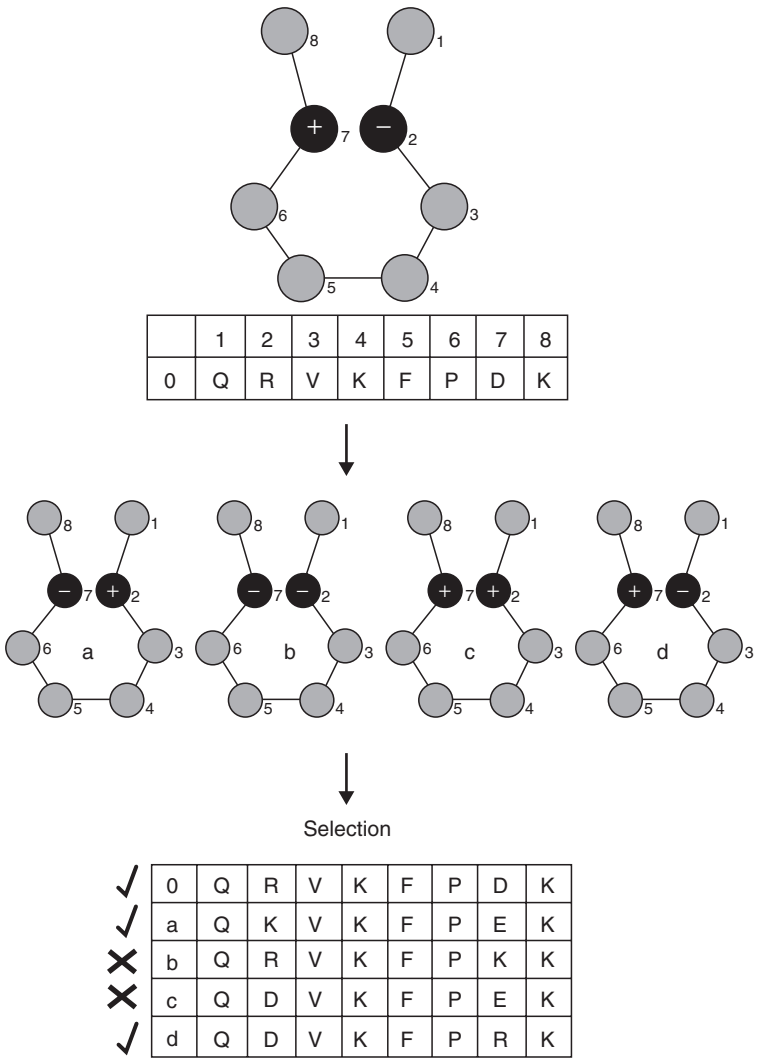

FIGURE 5.1 Schematic representation of coordinated substitutions in a pair of aa sites forming a salt bond in a protein domain. Sequences that contain residues of the same charge at positions 2 and 7 are unstable (b and c) and are eliminated during natural selection. Sequences containing residues of different charges that are stable (o, a, and d) can occur in a multiple-sequence alignment.

guide for reasoned selection of sets of sequences for inclusion in a mixture of peptides for vaccine design. Therefore, by selecting sequences which include pairs of aa that are highly predictive of each other, one may be covering important classes of sequences that are structurally or functionally related. Thus inclusion of peptides with highly covariant aa may be a useful strategy for designing broadly reactive vaccines [41].

5.2 METHODS

Bioinformatics methods for detecting correlated mutations consist of two main steps (1) alignment of homologous sequences and (2) identification of pairs of columns in the alignment in which there is a statistically significant tendency for mutations in one column to be accompanied by

corresponding and usually different mutations in the other column [46]. There are many algorithms for covariation analysis [9,29,36,41,47–49]. Here we explain the two principal algorithms and then discuss some important problems of covariation analysis.

5.2.1 Mutual Information

Statistical analyses of biological sequences present difficulties because these sequences are represented by symbols that have no natural ordering or underlying metric [50]. Consequently, conventional statistical estimates of variability and covariability are difficult to apply. Several authors have suggested the use of the concepts of entropy and mutual information [2,26,39,41,44,50,51]. Entropy (H) is a measure of uncertainty derived from thermodynamics and statistical physics that has considerable utility for studies of protein structure. The entropy $H(X)$ for a discrete random variable X is defined as follows:

$$H(X) = -\sum_{i=1}^{k} p(x_i) \log_b p(x_i)$$

where
$k = 20$ (aa residues)
$p(x_i)$ the probability of an aa being of the ith kind

$H = 0$ when all elements are in the same category (the same aa at a particular site). H increases if the number of categories (residues at a site) increases or if the categories have similar probabilities. Thus, the minimum entropy or uncertainty value will be zero when only a single residue occurs at a particular site in all included proteins. The choice of logarithm base b serves to scale the entropy, if $b = k$, then the maximum entropy is 1, when all 20 residues are present in equal frequencies at a given site. The concept of entropy can be easily extended to the case of two random variables with ordered pairs (x_i,y_i). In this instance, it is helpful to think of the pairs as elements of an extended alphabet, whose elements are all possible distinct pairs. If we have a pair of random variables, then pair entropy is defined as follows:

$$H(X,Y) = -\sum_{i=1}^{k} \sum_{j=1}^{l} p(x_i,y_i) \log_b p(x_i,y_i)$$

The relative information content of Y contained in X is termed the mutual information (MI) and is calculated as follows:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$

In biological sequences, MI describes the extent of association between residues at aa sites X and Y that might arise from evolutionary, functional, or structural constraints [2]. Note that $\mathrm{MI}(X,Y) = \mathrm{MI}(Y,X)$, and if X and Y are independent, then $\mathrm{MI}(X,Y) = 0$, corresponding to the fact that no information is obtained regarding Y by finding out about X. MI is always nonnegative and achieves its maximum value if there is complete covariation. The minimum value of 0 is obtained either when X and Y vary independently or when there is no variation [41]. MI is always lower than the minimum entropy of X and Y and therefore, it is convenient to normalize the MI to compare pairs of different entropy. Martin et al. [8] assessed the performance of various normalizations of MI in enhancing the detection of covariation and found that normalizing MI by the pair entropy optimized the ability to detect coevolving sites over a large range of mutation rates.

5.2.2 Physicochemical Correlation

The information theoretic approach to sequence data has serious shortcomings. For example, it is difficult to describe inverse (negative) relationships among sequence sites, such as those found with compensatory variation associated with aa charge or size. Furthermore, this approach provides little information about the underlying causal complexity of observed covariation [52]. Functionally, significant coordinated substitutions of residues in proteins must result from interactions dependent on the physicochemical property values of the residues [7]. One approach for the analysis of covariation is based on estimation of the correlation coefficient between the values of a physicochemical parameter at a pair of positions of sequence alignment. Let us consider a sample of N aligned sequences of length L. Then, we consider a certain physicochemical aa property f. A value of this property is attributed to every aa in the alignment. As a result, we obtain a matrix whose element f_{ki} is the f value at the ith position of the kth sequence. In the case of evolutionary unrelated sequences, the covariance s_{ij} (if $i \neq j$) and variance (if $i = j$) are equal to

$$S_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (f_{ki} - \bar{f}_i)(f_{kj} - \bar{f}_j)$$

To estimate the relation between the pair of variables f_i and f_j, the linear correlation coefficient is calculated as

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} \cdot S_{jj}}}$$

When the correlation coefficient between two sites is negative, an increase in the value of a property at position i will make more likely a substitution at position j that will result in a decrease in the value of the property (a net value compensatory substitution) [53]. When the correlation coefficient is positive, it may be assumed that substitutions are compensatory for the difference between the property values of two residues (the amount of the difference is conserved) [53]. There are other correlation methods that do not assume a linear relationship between variables such as $\chi 2$, Spearman's ρ, and Kendall's τ.

5.2.3 Physicochemical Properties of Amino Acids

A study by Chelvanayagam et al. [6] found that the analysis of covariation involving different physicochemical characteristics improves the number of truly covariant pairs. However, there are many reported aa properties and the selection of the right ones is a difficult choice. Interestingly, Atchley et al. [52] used multivariate statistical analyses on 494 aa properties [54] to produce a small set of highly interpretable numeric patterns of aa variability that can be used in a wide variety of analyses directed toward understanding the evolutionary, structural, and functional aspects of protein variability. Factor analysis was used to produce a subset of numerical descriptors to summarize the set of aa physiochemical properties. Factor analysis simplifies high-dimensional data by generating a smaller number of factors that describe the structure of highly correlated variables. The resultant factors are linear functions of the original data, fewer in number than the original, and reflect clusters of covarying traits that describe the underlying structure of the variables [52].

Factor analysis of the aa attributes resulted in five factors or patterns of highly intercorrelated physiochemical variables, a reduction in dimensionality of two orders of magnitude from the original 494 properties [52] (Table 5.1). POLARF1 reflects polarity and simultaneous covariation in portion of exposed residues versus buried residues, nonbonded energy versus free energy, number of hydrogen bond donors, polarity versus nonpolarity, and hydrophobicity versus hydrophilicity. HELIXF2 is a secondary structure factor. There is an inverse relationship of

TABLE 5.1
Highest Factor Coefficients of AA Properties

POLARF1	
Average nonbonded energy per atom	1.028
Percentage of exposed residues	1.024
Percentage of buried residues	−1.017
Average accessible surface area	1.005
Transfer free energy	−1.003
Residue accessible surface area in folded protein	0.95
Average interactions per side chain atom	−0.928
Average side chain orientation angle	−0.896
Eisenberg hydrophobic index	−0.864
Hydropathy index	−0.856
HELIXF2	
Average relative probability of helix	−1.004
Relative frequency in α-helix	−0.987
α-Helix indices	−0.939
Normalized frequency of coil	0.863
Free energy in α-helical region	0.858
Normalized frequency of turn	0.831
Information measure for loop	0.786
Chou–Fasman parameter of coil conformation	0.78
Helix–coil equilibrium constant	−0.724
Conformational parameter of β-turn	0.693
SIZEF3	
Bulkiness	0.988
Hydrophobicity factor	0.833
Size	0.811
Residue volume	0.794
Average volume of buried residue	0.766
Side chain volume	0.754
Normalized frequency of extended structure	0.706
Molecular weight	0.657
Normalized frequency of left-handed α-helix	−0.641
Normalized frequency of β-sheet, unweighted	0.611
CODONF4	
aa composition of total proteins	0.963
Relative frequency of occurrence	0.931
Number of codons	0.867
aa composition	0.852
Heat capacity	−0.656
Refractivity	−0.621
Average nonbonded energy per residue	−0.507
Molecular weight	−0.504
Conformational parameter of β-turn	−0.439
Normalized frequency of turn	−0.393

TABLE 5.1 (continued)
Highest Factor Coefficients of AA Properties

CHARGEF5

Eisenberg hydrophobic index	−0.864
Number of hydrogen bond donors	0.809
Negative charge	0.451
Positive charge	0.442
Relative mutability	0.337
Isoelectric point	0.224
Number of codons	0.079
Normalized frequency of left-handed α-helix	−0.079
Net charge	0.078
Average nonbonded energy per residue	0.042

Source: From Atchley, W., Zhao, J., Fernandes, A., and Druke, T., *Proc. Natl. Acad. Sci. U S A*, 102, 6395, 2005.

Note: The values of each AA Index and the source reference can be found at the AA Index On-Line Database (Kawashima, S. and Kanehisa, M., *Nucleic Acids Res*., 28, 374, 2000). The names of the factors (e.g., POLARF1) reflect the most important physicochemical property of each factor but the reader must keep in mind that each factor reflects a set of related properties.

relative propensity for various aa in various secondary structural configurations, such as a coil, a turn, or a bend versus the frequency in an α-helix. SIZEF3 relates to molecular size or volume with high factor coefficients for bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight. CODONF4 reflects relative aa composition in various proteins and the number of codons for each aa. These attributes vary inversely with refractivity and heat capacity. CHARGEF5 refers to electrostatic charge with high coefficients on isoelectric point and net charge. Atchley et al. [52] showed how the transformation into one of the five multidimensional factors of physicochemical properties was useful in the analysis of bHLH proteins that bind DNA.

5.2.4 Invariance of Protein Physicochemical Characteristics

An important feature of coordinated substitutions is their additional contribution to the invariance of the integral physicochemical characteristics of a protein, such as the total volume and net charge. Invariance of a physicochemical characteristic may result from the pressure of selection either on the entire protein or on its functionally or structurally significant parts. For example, Afonnikov et al. [36] analyzed the DNA-binding domain of the homeodomain class, finding two conservative physicochemical characteristics preserved due to coordinated substitutions at certain groups of positions in the protein sequence. Integral characteristics of proteins have also been found in Zinc-finger domains, DNAJ domains of heat-shock proteins, DNA-binding domains of CREB, and AP-1 and the Btk PH domain, where information on these characteristics facilitated predictions of their functional motifs [36,53,55]. An integral characteristic (F) of a protein is described as the sum of the values of a physicochemical property at protein positions, with variance D(F) [36]. A permutation procedure can be used to test the hypothesis that the observed sample variance is lower than the expected if correlation between sites is absent.

5.3 SOME PROBLEMS AND THEIR SOLUTIONS

5.3.1 Sequence Conservation

All the methods for detecting correlated mutations are sensitive to the degree of sequence conservation in the alignment [56]. A covariation analysis is supposed to detect how the changes in column *i* effect column *j* and, therefore, if there are no changes in a column, the algorithm must choose to report no score, a perfectly high score, or a perfectly low score. Each algorithm for covariation analysis works on certain level of sequence conservation and within that level chooses the residue pairs that truly covary. The MI approach will tend towards a low covariance score for highly conserved pairs of columns in an alignment. The physicochemical correlation approach will tend toward a high covariance score for highly conserved columns. Simulations of protein coevolution have showed that it is very difficult to separate sites which are coevolving from those that are not if either site is highly conserved [8,44]. A first step in the covariation analysis must be the removal of perfectly conserved columns (entropy equal to zero) and depending of the variability of the dataset, a polymorphism cutoff is also recommended. The polymorphism cutoff depends on the diversity of the dataset but protein simulations performed by Martin et al. [8] showed that the ability of MI to filter out false positives is optimized for MSA with mean entropy of 0.3 for alignments of approximately 100 sequences, suggesting a natural entropy cutoff when analyzing real protein MSA.

5.3.2 Sequence Weights

In the analysis of covariation, it is important to include weighting functions that correct for the different numbers of proteins in different branches of an evolutionary tree [6]. It is known that over-representation of some homologous sequences in the sample may cause biases in statistical estimates. In the context of a MSA, this is important because alignments frequently contain very similar (even duplicated) sequences; these can bias the construction of the alignment itself or make some trends (merely due to nonrandom sampling) appear strong [57]. Equally problematic is the low representation of interesting but rare data. Scores averaged over alignment columns are vulnerable to over- or under-representation of certain sequences. A remedy is to assign weights to the sequences in an alignment before calculating any average value. To avoid such biases, different schemes of sequence weighting have been proposed [57–59]. These approaches reduce the weights of over-represented sequences and imply that the distribution of sequences in the sequence space is expected to be homogeneous [36,53].

5.3.3 Statistical Significance

The shape of the distribution of covariation values depends on the particular method used. However, most distributions are skewed, with most pairs having a very low value. To define the pairs of position with a high probability of being structurally or functionally linked, some studies chose the pairs of positions with the highest covariation values, usually using an arbitrary cutoff based on Z scores (number of standard deviations from the mean covariation value) [8,26,44]. Recent studies have opted to assess the significance of the covariation values using a permutation procedure. The aa at each site in the sequence alignment are vertically shuffled, creating 10,000 or more random alignments that simulate the distribution of the covariation values under the null hypothesis that substitutions of aa at two sites are statistically independent. If the sequences are effectively unrelated then the pairs of positions with a significant covariation must have structural or functional links. Sequences are unrelated if the relationships by descent have been lost and there is no longer a significant phylogenetic signal or the sequences were obtained by in vitro selection [53,60]. However, in the case of related sequences it is necessary to test whether a correlation value reflects a significant association (possibly due to structural and functional constraints), or, instead, results from evolutionary history and stochastic events [2].

5.3.4 Multiple-Comparisons Problem

The multiple-comparisons problem occurs when one considers a set of statistical inferences simultaneously. In the covariation analysis, there are usually a large number of pairs of sites tested. Technically, the problem of multiple comparisons can be described as the potential increase in false positives that occurs when statistical tests are used repeatedly [61]. If m independent comparisons are performed with a given allowable error (α_i) the experiment-wide significance level α_g is given by

$$\alpha_g = 1 - (1 - \alpha_i)^m$$

To retain the same overall rate of false positives (rather than a higher rate) in a test involving more than one comparison, the standards for each comparison must be more stringent [61]. The Bonferroni correction states that the statistical significance level that should be used for each hypothesis separately is $1/n$ times of what it would be if only one hypothesis were tested [61]. However, many biological applications require a less conservative approach with greater power to detect true positives, at a cost of increasing the likelihood of obtaining false positives. The false discovery rate (FDR) is currently the most popular approach, which controls the expected proportion of false positives instead of the chance of any false positives. Consider testing $H_1, H_2, \ldots, H_m$ based on the corresponding ordered p-values $P_1, P_2, \ldots, P_m$, where $P_1 \leq P_2 \leq, \ldots, \leq P_m$ and H_i denotes the null hypothesis corresponding to P_i. Let k be the largest i for which

$$P_i \leq \frac{i}{m} q$$

Then all $H_i \leq k$ are rejected. This procedure controls the FDR at a proportion q which can be adjusted as low as possible [61].

5.3.5 Background Sequence Covariation

An obvious source of covariation among residues at different sites is common evolutionary history. Felsenstein (1985) showed that related sequences are part of a hierarchically structured phylogeny and, therefore, for statistical purposes, cannot be regarded as being drawn independently from the same distribution. In estimating the significance of the correlation coefficients, homologous sequences cannot be considered statistically independent because they share common evolutionary ancestry. Tree structure generally imparts extreme non-normality in the form of kurtosis to the correlation distribution, and this invalidates significance statistics based on the assumption of the normal distribution [9]. A large number of residue pairs with high correlation are expected simply due to background noise in the presence of phylogenetic structure and all the methods that do not incorporate the effect of this tree structure have many false positives [9]. Recent methods have been developed that address these issues but there is a great lack of agreement between all methods. The following five methods incorporate different solutions to the background sequence covariation problem and have different sensitivity and specificity but a definitive comparison between them is still absent.

1. Removal of multiple covariation [62]. This procedure claims to detect statistical correlations stemming from functional interaction by removing the strong phylogenetic signal that leads to the correlations of each site with many others in the sequence. The method assumes that all sites in a sequence have followed the same phylogeny resulting in a consistent pattern of substitution throughout the sequences thus creating many correlations between sites. Their analysis is biased towards those positions that covary with at most a few others and excludes coevolving groups of positions. However, the amount of coevolution in proteins is unknown and therefore this assumption is highly conservative and lacks experimental or theoretical support.

2. Parametric bootstrap [2,63]. This approach compares the distribution of inter-site mutual information for an alignment of naturally occurring sequences with the distribution of mutual information for artificial sequence data generated using the parametric bootstrap from a random ancestral sequence, a given substitution matrix, and the same tree. Correlated mutations in the set of artificial sequences can arise solely from common ancestry and, therefore, this comparison enables the calculation of the probability that a pair of covarying sites with a certain value of the mutual information statistic did not result from common ancestry [46]. This approach depends crucially on measures of aa similarity or distance. There are many such distance substitution matrices, the most used being those based on the propensity for evolutionary change from one aa to another [64]. However, the exchangeability of aa in a protein context is problematic because phylogenetic relations, mutational bias, and physicochemical effects are difficult to separate [65].
3. Background MI [8,44]. This method makes the assumption that each position in a MSA is affected equally by phylogenetic linkage, and that the majority of positions in the alignment covary only because of linkage. On the basis of these assumptions, each alignment is used as its own null model for the identification of covarying positions. The average MI of all possible pairs and its standard deviation are evaluated, allowing the identification of pairs with high MI values (z scores >4). Gloor et al. [44] used this method to identify non-conserved coevolving sites in MSA from a variety of protein families, finding that coevolving sites in these alignments fall into two general categories. One set is composed of sites that coevolve with only one or two other sites, often displaying direct aa side chain interactions with their coevolving partner. The other set comprises sites that coevolve with many others and are frequently located in regions critical for protein function, such as active sites and surfaces involved in molecular interactions and recognition. Gloor et al. [44] also found that coevolving positions are more likely to change protein function when mutated than are positions showing little coevolution. These results imply that these coevolving positions compose an important subset of the positions in an alignment, and may be as important to the structure and function of the protein family as are highly conserved positions. Interestingly, the analysis of the homeodomain mutations associated with human disease showed that those positions with high levels of covariation are more likely to be associated with a mutant phenotype when mutated [44]. In addition, the *E. coli* ATP synthase e subunit has been extensively mutated in vitro, and Gloor et al. [44] found that positions with high level of covariation are more likely to change the activity of the protein upon in vitro mutagenesis than those with low levels of covariation.
4. Genetic linkage on synonymous (S) and non-synonymous (A) sites [66]. This interesting new method systematically separates the covariation induced by selective interactions between aa from background sequence covariation, using silent (S) versus aa replacing (A) mutations. Covariation between two aa mutations, (A,A), can be affected by selective interactions between aa, whereas covariation within (A,S) pairs or (S,S) pairs cannot. This study performed an analysis of the pol gene in HIV, revealing that (A,A) covariation levels are enormously higher than for either (A,S) or (S,S), and thus cannot be attributed to phylogenetic effects. Inspection of the most prominent (A,A) interactions in the HIV pol gene showed that they are known sites of independently identified drug resistance mutations, and physically cluster around the drug-binding site.
5. Removal of phylogenetic clades [67–69]. To remove the effects of background sequence covariation, their analysis is applied to the complete alignment and to subalignments where specific phylogenetic clades are removed from the tree. Coevolving aa sites that are no longer detected following removal of one of the clades are classified as phylogenetically related sites as they occur in specific branches of the tree. The clades chosen for removal are identified before the covariation analysis is applied and they include

sequences that form a well-defined biological cluster or a cluster with high statistical support [68]. This method was successfully applied to heat-shock protein GroEL, ATPase Hsp90, the Gag protein from HIV-1, and the env gene of the HIV-1 group M subtype, highlighting that almost all detected coevolving sites are functionally or structurally important [68,69]. A possible weakness of this approach is the removal of pairs of positions that are truly linked only in the genomic context of certain sequences that formed a clade due to selective pressure. Some changes could be negatively selected and maintained in the population depending on the particular patterns of epistasis that apply in that genomic context. The positions where these type of changes occur are clade-specific and may define different evolutionary paths of clades. Thus, these covariable changes may still be functionally and structurally important but only in the context of the specific clade.

6. Markov model for sequence coevolution [45]. Recently, Yeang and Haussler [45] proposed a continuous-time Markov process model for sequence coevolution under two hypotheses and testing the most likely for each pair of sites. The null (independent) model hypothesizes that two sites evolve independently. The alternative (coevolutionary) model is obtained from the null model by re-weighting the independent substitution rate matrix to favor double over single changes. This method was applied to all the inter- and intra-domain position pairs in all the known protein domain families in Pfam database [70]. The majority of the inferred coevolving pairs of positions are functionally related or spatially coupled. Many of the coevolving positions are located at functionally important sites of proteins/protein complexes, such as the subunit linkers of superoxide dismutase, the tRNA-binding sites of ribosomes, the DNA-binding region of RNA polymerase, and the active- and ligand-binding sites of various enzymes.

5.4 AVAILABLE SOFTWARE

The following are publicly available programs that are currently used in covariation analysis:

DEPENDENCY [62]. http://www.uhnres.utoronto.ca/tillier/depend2/dependency.html
PCOAT [71]. ftp://iole.swmed.edu/pub/PCOAT/
CRASP [53]. http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/
CAPS [68]. http://bioinf.gen.tcd.ie/~faresm/software/caps/.

5.5 CONCLUSION

During protein evolution, certain substitutions at different sites may occur in a coordinated manner due to interactions between aa residues. The detection of these interactions among separate aa sites is fundamental for understanding protein structure and evolution. Covariation analyses can also be important in identifying sites that may change the phenotype of a protein or be located at functionally important sites of proteins/protein complexes, a useful compass for researchers attempting to define functional domains through experimental analysis. In sequence alignments of homologous proteins, these interactions between sites can be manifested in correlation between substitutions at pairs of alignment positions. However, it is necessary to discriminate the three different sources of covariation: (1) chance, (2) common ancestry, and (3) structural or functional constraints. Effectively discriminating among these underlying causes is a difficult task with many statistical and computational difficulties, which are addressed in very different ways by recent methods. Although there is not a consensus about the best way of discriminating the sources of covariation, all recent methods have confirmed that the detection of coordinated substitutions is a very important tool for protein analyses, predictions of spatial structure, inter-residue contacts, function, and protein–protein interactions.

REFERENCES

1. Darwin, C. *The Origin of Species*, Penguin, Middlesex, United Kingdom, 1859.
2. Atchley, W., Wollenberg, K., Fitch, W., Terhalle, W., and Dress, A. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol Biol Evol* 17, 164–178, 2000.
3. Tolou, H., Nicoli, J., and Chastel, C. Viral evolution and emerging viral infections: What future for the viruses? A theoretical evaluation based on informational spaces and quasispecies. *Virus Genes* 24, 267–274, 2002.
4. Schuster, P. Evolution at molecular resolution. *Nolinear Cooperative Phenomena in Biological Systems*, pp. 86–112, 1998. AQ1
5. Lesk, A. and Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol* 136, 225–270, 1980.
6. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G., and Benner, S. An analysis of simultaneous variation in protein structures. *Protein Eng* 10, 307–316, 1997.
7. Tomii, K. and Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng* 9, 27–36, 1996.
8. Martin, L., Gloor, G., Dunn, S., and Wahl, L. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21, 4116–4124, 2005.
9. Pollock, D. and Taylor, W. Effectiveness of correlation analysis in identifying protein residues. *Protein Eng* 10, 647–657, 1997.
10. Chothia, C. and Lesk, A. Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. *J Mol Biol* 160, 309–323, 1982.
11. Lesk, A. and Chothia, C. Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. *J Mol Biol* 160, 325–342, 1982.
12. Oosawa, K. and Simon, M. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in *Escherichia coli. Proc Natl Acad Sci U S A* 83, 6930–6934, 1986.
13. Altschuh, D., Vernet, T., Berti, P., Moras, D., and Nagai, K. Coordinated amino acid changes in homologous protein families. *Protein Eng* 2, 193–199, 1988.
14. Bordo, D. and Argos, P. Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. *J Mol Biol* 211, 975–988, 1990.
15. Mateu, M. and Fersht, A. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization. *Proc Natl Acad Sci U S A* 96, 3595–3599, 1999.
16. Lim, W. and Sauer, R. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* 339, 31–36, 1989.
17. Lim, W., Farruggio, D., and Sauer, R. Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry* 31, 4324–4333, 1992.
18. Baldwin, E., Hajiseyedjavadi, O., Baase, W., and Matthews, B. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* 262, 1715–1718, 1993.
19. Burch, C. and Chao, L. Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics* 167, 559–567, 2004.
20. Bonhoeffer, S., Chappey, C., Parkin, N., Whitcomb, J., and Petropoulos, C. Evidence for positive epistasis in HIV-1. *Science* 306, 1547–1550, 2004.
21. Sanjuan, R., Moya, A., and Elena, S. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc Natl Acad Sci U S A* 101, 15376–15379, 2004.
22. Poon, A. and Chao, L. The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics* 170, 989–999, 2005.
23. Mateo, R. and Mateu, M. Deterministic, compensatory mutational events in the capsid of foot-and-mouth disease virus in response to the introduction of mutations found in viruses from persistent infections. *J Virol* 81, 1879–1887, 2007.
24. Garriga, C., Pérez-Elías, M.J., Delgado, R., Ruiz, L., Nájera, R., Pumarola, T., Alonso-Socas, M., García-Bujalance, S., and Menéndez-Arias, L. Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. *J Med Virol* 79, 1617–1628, 2007.

25. Govindarajan, S. et al. Systematic variation of Amino acid substitutions for stringent assessment of pairwise covariation. *J Mol Biol* 328, 1061–1069, 2003.
26. Clarke, N. Covariation of residues in the homeodomain sequence family. *Protein Sci* 4, 2269–2278, 1995.
27. Voigt, C., Mayo, S., Arnold, F., and Wang, Z. Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A* 98, 3778–3783, 2001.
28. Fukami-Kobayashi, K., Schreiber, D., and Benner, S. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J Mol Biol* 319, 729–743, 2002.
29. Göbel, U., Sander, C., Schneider, R., and Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* 18, 309–317, 1994.
30. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A* 91, 98–102, 1994.
31. Shindyalov, I., kolchanov, N., and Sander, C. Can three dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7, 349–358, 1994.
32. Taylor, W. and Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Eng* 7, 341–348, 1994.
33. Benner, S., Cannarozzi, G., Gerloff, D., Turcotte, M., and Chelvanayagam, G. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem Rev* 97, 2725–2844, 1997.
34. Nagl, S., Freeman, J., and Smith, T. Evolutionary constraint networks in ligand-binding domains: An information-theoretic approach. *Pac Symp Biocomput*, 90–101, 1999.
35. Larson, S., Di Nardo, A., and Davidson, A. Analysis of covariation in an SH3 domain sequence alignment: Applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol* 303, 433–446, 2000.
36. Afonnikov, D., Oshchepkov, D., and Kolchanov, N. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics* 17, 1035–1046, 2001.
37. Nemoto, W., Imai, T., Takahashi, T., Kikuchi, T., and Fujita, N. Detection of pairwise residue proximity by covariation analysis for 3D-structure prediction of G-protein-coupled receptors. *Protein J* 23, 427–435, 2004.
38. Wang, L. Covariation analysis of local amino acid sequences in recurrent protein local structures. *J Bioinform Comput Biol* 3, 1391–1409, 2005.
39. Shackelford, G. and Karplus, K. Contact prediction using mutual information and neural nets. *Proteins* 69, 159–164, 2007.
40. Altschuh, D., Lesk, A., Bloomer, A., and Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 193, 693–707, 1987.
41. Korber, B., Farber, R., Wolpert, D. and Lapedes, A. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc Natl Acad Sci U S A* 90, 7176–7180, 1993.
42. Gilbert, P., Novitsky, V., and Essex, M. Covariability of selected amino acid positions for HIV type 1 subtypes C and B. *AIDS Res Hum Retroviruses* 21, 1016–1030, 2005.
43. Kolli, M., Lastere, S., and Schiffer, C. Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. *Virology* 347, 405–409, 2006.
44. Gloor, G., Martin, L., Wahl, L., and Dunn, S. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44, 156–165, 2005.
45. Yeang, C. and Haussler, D. Detecting coevolution in and among protein domains. *PLoS Comput Biol* 3, e211, 2007.
46. Noivirt, O., Eisenstein, M., and Horovitz, A. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng* 18, 247–253, 2005.
47. Olmea, O., Rost, B., and Valencia, A. Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol* 293, 1221–1239, 1999.
48. Lockless, S. and Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286, 295–299, 1999.
49. Kass, I. and Horovitz, A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 48, 611–617, 2002.
50. Atchley, W., Terhalle, W., and Dress, A. Positional dependance, cliques and predictive motifs in the bHLH protein domain. *J Mol Evol* 48, 501–506, 1999.

51. Crooks, G. and Brenner, S. Protein secondary structure: Entropy, correlations and prediction. *Bioinformatics* 20, 1603–1611, 2004.
52. Atchley, W., Zhao, J., Fernandes, A., and Druke, T. Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A* 102, 6395–6400, 2005.
53. Afonnikov, D. and Kolchanov, N. CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res* 32, W64–W68, 2004.
54. Kawashima, S. and Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res* 28, 374, 2000.
55. Shen, B. and Vihinen, M. Conservation and covariance in PH domain sequences: Physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng Des Sel* 17, 267–276, 2004.
56. Fodor, A. and Aldrich, R. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56, 211–221, 2004.
57. Vingron, M. and Sibbald, P. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci U S A* 90, 8777–8781, 1993.
58. Altschul, S., Carroll, R., and Lipman, D. Weights for data related by a tree. *J Mol Biol* 207, 647–653, 1989.
59. Sibbald, P. and Argos, P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol* 216, 813–818, 1990.
60. Segal, M., Cummings, M., and Hubbard, A. Relating amino acid sequence to phenotype: Analysis of peptide-binding data. *Biometrics* 57, 632–642, 2001.
61. Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc, Series B* 57, 289–300, 1995.
62. Tillier, E. and Lui, T. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19, 750–755, 2003.
63. Wollenberg, K. and Atchley, W. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci U S A* 97, 3288–3291, 2000.
64. Jones, D., Taylor, W., and Thornton, J. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8, 275–282, 1992.
65. Yampolsky, L. and Stoltzfus, A. The exchangeability of amino acids in proteins. *Genetics* 170, 1459–1472, 2005.
66. Wang, Q. and Lee, C. Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase. *Plos ONE* 2, e814, 2007.
67. Fares, M. and Travers, S. A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analysis. *Genetics* 173, 9–23, 2006.
68. Fares, M. and McNally, D. CAPS: Coevolution analysis using protein sequences. *Bioinformatics* 22, 2821–2822, 2006.
69. Travers, S., Tully, D., McCormack, G. and Fares, M. A study of the coevolution patterns operating within the env gene of the HIV-1 group M subtypes. *Mol Biol Evol* 24, 2787–2801, 2007.
70. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E. The Pfam protein families database. *Nucleic Acids Res* 30, 276–280, 2002.
71. Qi, Y. and Grishin, N. PCOAT: Positional correlation analysis using multiple methods. *Bioinformatics* 20, 3697–3699, 2004.

AUTHOR QUERY

[AQ1] Please the publishers name and location for the reference ''Schuster, 1998.''