**BRIEFINGS IN BIOINFORMATICS.** VOL 9. NO 1. 46–56
Advance Access publication November 13, 2007

doi:10.1093/bib/bbm052

# Correlated substitution analysis and the prediction of amino acid structural contacts

*David S. Horner, Walter Pirovano and Graziano Pesole*

Submitted: 31st July 2007; Received (in revised form): 8th October 2007

## Abstract

It has long been suspected that analysis of correlated amino acid substitutions should uncover pairs or clusters of sites that are spatially proximal in mature protein structures. Accordingly, methods based on different mathematical principles such as information theory, correlation coefficients and maximum likelihood have been developed to identify co-evolving amino acids from multiple sequence alignments. Sets of pairs of sites whose behaviour is identified by these methods as correlated are often significantly enriched in pairs of spatially proximal residues. However, relatively high levels of false-positive predictions typically render such methods, in isolation, of little use in the *ab initio* prediction of protein structure. Misleading signal (or problems with the estimation of significance levels) can be caused by phylogenetic correlations between homologous sequences and from correlation due to factors other than spatial proximity (for example, correlation of sites which are not spatially close but which are involved in common functional properties of the protein). In recent years, several workers have suggested that information from correlated substitutions should be combined with other sources of information (secondary structure, solvent accessibility, evolutionary rates) in an attempt to reduce the proportion of false-positive predictions. We review methods for the detection of correlated amino acid substitutions, compare their relative performance in contact prediction and predict future directions in the field.

**Keywords:** *correlated mutation analysis; amino acid contacts; functional correlation; phylogeny*

## INTRODUCTION

Models of molecular evolution used in bioinformatics and phylogenetic studies of both nucleotide and protein sequences have become increasingly sophisticated in recent years. Modern nucleotide substitution models allow different frequencies of bases and substitution types (reviewed in [1]) and permit the estimation and incorporation of site-specific substitution rates (reviewed in [2]). However, such stochastic substitution models assume that each site in a molecular sequence evolves independently of all other sites. While this assumption greatly facilitates probabilistic calculations such as the scoring of alignments or the construction of phylogenetic trees, it is naïve for a number of reasons. For example, in coding sequences where strong evolutionary pressures are likely to act on the protein sequence encoded, the nature of the genetic code ensures that multiple substitutions within a given codon are unlikely to be accepted in an independent manner. Furthermore, selective constraints requiring the maintenance of secondary structure elements in non-coding RNAs or elements involved in post-transcriptional regulation of gene expression (often found in untranslated regions of

Corresponding author. David S. Horner, Department of Biomolecular Sciences and Biotechnology, University of Milan, via Celoria 26, 20133 Milano, Italy. Tel: +39 50314880; Fax: +39 50314912; E-mail: David.horner@unimi.it

**David S. Horner** is a researcher in molecular biology at the University of Milan. His research interests include comparative genomics and molecular phylogenetics.

**Walter Pirovano** is currently a PhD student in the Centre for Integrative Bioinformatics VU at the Free University of Amsterdam. His research interests include multiple sequence analysis and the development of new sequence alignment tools.

**Graziano Pesole** is full professor of molecular biology at the University of Bari and leads a research team in 'Bioinformatics and Comparative Genomics' at the Institute of Biomedical Technology (National Research Council). His research interests include bioinformatics, the development of tools for genome annotation, comparative genomics and molecular evolution.

mRNAs) would be expected to violate the assumption of independence between sites.

The evolutionary behaviour of codons has been modelled with codon level substitution matrices [3]. Covariance models combined with sequence-based heuristics can also be used to efficiently detect known secondary structure elements in genome sequences [4] and coordinated substitutions have been used in comparative sequence analysis to predict secondary structure [5].

In the case of amino acid sequences, empirically derived probabilistic substitution models specific to diverse taxonomic groups and sub-cellular localisations are commonly used for sequence alignment and phylogenetic analyses (reviewed in [1]). Such models typically allow the incorporation of dataset-specific amino acid frequencies [6] and variability in site substitution rates. Such models capture only the average tendencies of amino acids to undergo interconversions and do not necessarily describe perfectly the context-specific evolutionary tendencies of individual sites. Analogously to the situation in RNA sequences, physical and functional interactions between sites in protein sequences might lead to non-independence of their evolutionary behaviour. Such behaviour is an explicit prediction of the COVARION hypothesis [7] and forms the basis of heterotachy [8] and evolutionary trace [9] analyses. Recently, Lin *et al*. [10] introduced contact-mutation matrices derived from structural information. It has been shown that substitution scores reflecting contact information can also improve alignment accuracy [11]. However, failure to accommodate non-independence of substitutions is also likely to compromise phylogenetic analyses, ancestral sequence reconstruction and molecular dating strategies. The detection of pairs or groups of sites undergoing correlated evolution should also allow the prediction of structurally or functionally interacting amino acids [12–14].

Reliable prediction of pairs of amino acids that are physically proximal in protein tertiary structures would be of immense value in the *ab initio* prediction of protein structure through the provision of a priori structural constraints. A priori knowledge of relatively few contacts can allow discrimination between structure models or between solutions suggested by threading [15]. However, correlations between sites that are physically distant in protein structures might also be attributable to networks of sites, which influence protein domain architecture, substrate specificity or folding [16].

Correlations between amino acid sites do not follow simple rules analogous to those in RNA sequences and different physico-chemical considerations may tend to be more important in different secondary structure contexts and at different evolutionary distances [17]. Other considerations, such as the phylogenetic trees describing relationships between homologous sequences also complicate the identification of meaningfully correlated pairs or groups of sites [18].

Here, we provide an overview of different approaches that have been considered for the detection of correlated mutations from multiple sequence alignments and where possible, we try to compare the relative performance of these approaches. We briefly consider machine learning and other methods that incorporate correlated substitutions and other types of information in the prediction of amino acid contacts. Finally, we speculate on future directions in the analysis of correlated amino acid substitutions.

## SOURCES OF APPARENT CORRELATION BETWEEN AMINO ACID SITES

Atchley *et al*. [18] formalized a simple linear model to explain correlation (C) between two sites in a sequence alignment.

$$C = C_{\text{structure}} + C_{\text{function}} + C_{\text{phylogeny}} + C_{\text{interaction}} + C_{\text{stochastic}} \quad (1)$$

Where, $C_{\text{phylogeny}}$ is correlation due to phylogenetic relationships between homologous sequences that are related by a tree-like evolutionary structure and therefore cannot be considered to be statistically independent observations. Thus, we expect that the outcome of compensatory substitutions that occurred in a sequence ancestral to a group of sequences under consideration will be manifest in the descendent sequences—and that simple pairwise comparisons between sequences will not be sufficient to provide an accurate account of evolutionary events.

$C_{\text{structure}}$ and $C_{\text{function}}$ signify correlation due to structural and functional constraints, effectively the signal that correlated substitution analyses attempt to uncover. However, these sources of correlation may not be independent from one another or indeed from phylogenetic correlation. $C_{\text{interaction}}$ describes

**Table 1:** Software and webservers for correlated amino acid substitution analyses

| Software/Reference | Website | Notes |
|---|---|---|
| CRASP [11] | wwwmgs.bionet.nsc.ru/mgs/programs/crasp/pcrasp/index.html | Webserver, physiochemical correlations |
| LnLCorr [30] | www.evolutionarygenomics.com/LnLCorr.html | Software, two state maximum likelihood |
| PDGcon [14] | www.pdg.cnb.uam.es:8081/pdg.contact.pred.html | Webserver, McLachlan correlation |
| Dependency [22] | www.uhnres.utoronto.ca/tillier/depend2/dependency.html | Software, Mutual Interdependency |
| P2P [36] | ignmtest.ccbb.pitt.edu/cgi-bin/p2p/p2p0.cgi | Webserver, P2P matrix methods |
| MI with z-score [23, 24] | www.biochem.uwo.ca/cgi-bin/CDD/index.cgi | Software for MI method with z-scores |
| Various [29, 39] | www.afodor.net/ | Java code for various methods |

interactions between the aforementioned sources of correlation. Finally, random effects from uneven or incomplete sequence sampling, casual co-variation and other stochastic factors are represented by $C_{\text{stochastic}}$.

In practice, most bioinformatics approaches to correlated mutation analysis do not discriminate between structural and functional correlations, attempting to filter only stochastic and potentially phylogenetic noise. This may be no simple task as an elegantly designed recent study provides compelling evidence that the strength of correlations due to phylogenetic factors are, at least in many sequence alignments, of at least the same order of magnitude as those due to structure and function [19]. These authors used a method known to find relatively well-characterized correlations due to secondary structure and showed that distributions of levels of apparent correlation between pairs of sites within proteins were similar to those between pairs of sites from sequence matched non-interacting protein alignments.

For convenience, the majority of published correlated mutation analysis (CMA) algorithms can be grouped into various categories: methods using the Pearson correlation coefficient, methods employing information theory, methods based on observed and expected patterns of data distribution, methods based on alignment perturbation, methods based on maximum likelihood, methods using empirically derived contact probabilities and machine learning approaches. We will discuss each of these categories of methods independently and will describe elements common to different implementations. Representative software and server implementations of these methods are described in Table 1.

## Methods based on correlation coefficients

The simplest correlation coefficient-based methods do not consider substitutions but correlations between physiochemical properties of amino acids found in pairs of sites in individual sequences. In particular, the program CRASP [20–21] calculates correlation coefficients based on an array of amino acid physiochemical indices, weighting the contribution of individual sequences according to their degree of evolutionary relatedness with other homologs under consideration.

The Pearson coefficient [Equation (2)] has also been widely used to quantify correlations between inferred substitutions at pair of sites.

$$r_{ij} = \frac{1}{N}\sum_{kl}\frac{(s_{ikl}-\bar{s}_i)(s_{jkl}-\bar{s}_j)}{\sigma_i\sigma_j} \qquad (2)$$

Where, referring to a multiple sequence alignment (MSA), $r_{ij}$ is the correlation coefficient between sites $i$ and $j$, $N$ is the number of comparisons made between sequences, $s_{ikl}$ is a weight for a substitution inferred between two sequences $k$ and $l$ at position $i$, $\bar{s}$ is the mean weight for substitutions at site $i$ and $\sigma_i$ is the standard deviation of values of $s_{ikl}$. Such methods evaluate correlations between the weights for substitutions that are inferred to have occurred at each of two sites during the course of evolution. Substitutions can be weighted equally (a binary classification) or differently, using a metric of substitution probabilities or differences in biochemical properties between pairs of amino acids (charge, volume, etc). The best known implementation of this method [12] used all pairwise comparisons between sequences and a binary scheme for scoring substitutions (perfectly conserved sequence positions and positions with >10% gaps are excluded from the analysis). The component of the correlation coefficient derived from each pairwise comparison was also weighted according to the degree of overall identity between the pair of sequences considered. Workers from the same research group subsequently proposed the use of the McLachlan amino acid similarity matrix [22] to provide a general measure of the severity of amino acid substitutions [23]. Other related implementations have been proposed by Neher [13], Taylor and Hatrick [14] and Vicatos *et al.* [24]. These differ from the method of Olmea and Valencia [23], principally in the choice of substitution weight measures, the choice of which site pairs are evaluated and the treatment of negative correlations (for a comprehensive summary see [25]).

One limitation of the methods described above is that they take no explicit account of the phylogenetic tree linking the sequences under study. In tree-based correlation coefficient methods, hypothetical ancestral sequences corresponding to nodes on a phylogenetic tree are reconstructed and only changes observed along single branches are used in the calculation of correlation coefficients. Fukami-Kobayashi *et al.* [26] used maximum parsimony to reconstruct ancestral sequences on a neighbour joining tree and considered only charge reversing substitutions scored in a binary manner, while Fleishman *et al.* [27] reconstructed trees and ancestral sequences with probabilistic methods and considered all substitutions—using weights derived from the Miyata amino acid replacement matrix. We have also implemented a method similar to that of Fleishman, but using the McLachlan matrix and a different measure of significance (see Discussion section). Such tree-based methods may better account for phylogenetic correlations, but inevitably use less comparisons than pairwise methods as the number of internal branches in a bifurcating tree is smaller than the number of possible pairwise comparisons between sequences.

## Methods using information theory

Mutual information (MI) [Equation (3)] is a measure of the mutual dependence of two variables, in the case of two positions *i* and *j* in a sequence alignment it can be considered a measure of how much information about site *j* in a given sequence is given by knowledge of site *i*.

$$\mathrm{MI}(i,j)=\sum_{x=1}^{n}\sum_{y=1}^{m}P_{xy}\log_2\frac{p_{xy}}{p_x p_y} \qquad (3)$$

Where MI*(i, j)* is the mutual information between sites *i* and *j*, the indices refer to the 20 possible amino acid states and $p_{xy}$ reflects the probability of observing amino acid *x* at site *i* and amino acid *y* at site *j*, $p_x$ and $p_y$ are the respective independent probabilities of these events.

Unlike the correlation coefficient-based methods described earlier, mutual information-based approaches do not consider substitutions, but rather the relative frequencies of different amino acids and combinations of amino acids observed at two positions in the MSA. As for basic correlation coefficient methods, mutual information used in isolation might be expected to be sensitive to correlation derived from phylogeny and various methods have been proposed to account for this signal.

Wollenberg and Atchley [18, 28] suggested a parametric bootstrap approach, whereby a phylogenetic tree was used to repeatedly generate simulated sequences under a standard amino acid substitution matrix. Datasets generated in this way would be expected to show correlation only due to phylogeny and stochastic considerations—and allow the estimation of levels of phylogenetic correlation that would be expected to result from the tree linking the sequences.

Tillier and Lui [29] reasoned that the majority of sites in an alignment would follow the general phylogenetic trend, while pairs of sites which were undergoing correlated evolution for structural or functional reasons would share high mutual information between each other, but low mutual information with other sites in the alignment. They used the measure of multiple interdependency, effectively, the ratio of the mutual information between a pair of sites and the sum of the mutual information of each of these sites with all other sites.

Others have found that normalizing the mutual information score by pair entropy reduces the impact of phylogenetic correlations [30–31]. Pair entropy is a measure of the relative frequency with which a given combination of character states is observed at a pair of sites. Reasoning that all sites in the distribution follow the same phylogenetic history and related correlation trends, these authors use the distribution of pair normalized mutual information scores to convert individual pair scores into *z*-scores.

## Methods based on observed and expected patterns of data distribution

Several groups have considered the observed patterns of amino acid occurrence at two positions and compared them to the frequencies of data expected under an independent sites model, generating a chi-squared observed minus expected squared (OMES) statistic according to Equation (4):

$$\chi^2(i, j) = \sum_n \frac{(N_{n,OBS} - N_{n,EX})^2}{N_{n,EX}} \tag{4}$$

Where, $n$ runs over all possible combinations of amino acid pairs at positions $i$ and $j$, $N_{n,EX}$ is the expected number of occurrences of $n$ if positions $i$ and $j$ are independent and $N_{n,OBS}$ is the observed number of pairs of $n$. Initial work following this approach assumed that the test statistic indeed followed the chi-squared distribution [32–33], while a later modification introduced reshuffling of data columns to estimate the distribution of the test statistic. One proposed implementation of the reshuffling uses evolutionary distances to alter the probabilities that sites from particular sequences will be reshuffled to simulate tree-like evolution during the estimation of significance intervals [19].

## Methods based on alignment perturbation

Alignment perturbation methods compare the relative composition of columns of the entire alignment with that of a sub-alignment, whose sequence content is defined by the occurrence of a pre-specified amino acid at a given site. In statistical coupling analysis (SCA) [16, 34] a sub-alignment containing all sequences with a pre-defined amino acid at a given site is created. For the second site, the sum of the squares of the ratio of probabilities of observing each amino acid in the perturbed and complete alignments is calculated.

A second perturbation-based method, explicit likelihood of subset co-variation (ELSC) [35], selects sub-alignments in the same way as SCA, but calculates the probability of the observed partition of character states at a given site by chance given the entire alignment. This value is normalized to give a probability of the observed data relative to the most probable random partition of the data.

## Probabilistic methods

A true tree-based maximum likelihood method to uncover pairs of coevolving sites was proposed by Pollock *et al.* [36]. In this method, the probability of the observed data given a pre-calculated phylogenetic tree is calculated under independent and non-independent models, and a likelihood ratio test performed to evaluate the significance of the difference of likelihoods. However, in order both to remain computationally tractable and to allow adequate parameterisation of the substitution models from individual datasets, the method uses a simple two-state evolutionary model allowing, for example, discrimination only between positively and negatively charged amino acids or large and small amino acids. The method performed extremely well with simulated data, but has received limited testing with real data.

Bayesian mutational mapping to analyse distributions of substitutions over phylogenetic trees has been used with coding sequence data under codon substitution models [37]. Dimmic *et al.* implemented a variety of tests to identify correlations between codons. However, analogously to the maximum likelihood method, parametric tests implemented in this study were based on simple models of favourable or non-favourable interaction states between pairs of amino acids. With simulated data this method detected between 60 and 99% of pairs evolving in a correlated manner (depending on the strength of correlation imposed in the simulation) and only around 3% of other pairs of sites were recovered as significantly correlated. In limited testing on real data, 13% of contact pairs and 3% of non-contact pairs were recovered as significantly correlated (meaning that 40% of predictions made were true contacts—given the numbers of contact and non-contact pairs considered). However, these numbers are difficult to compare with other methodologies given the restricted size of the dataset (one protein) and the fact that the authors did not consider pairs of sites with inter-amino acid distances between 8 and 16 Å.

## Empirical matrix-based methods

Singer *et al.* [38] developed a log-odds matrix describing the relative probability that a given pair of amino acids should be involved in inter-residue contacts or not taking a normalized sum of log-odds scores for pairs of sites observed to change simultaneously through pairwise comparisons of sequences. Significance levels were established through a data randomization test.

Lin *et al.* [10] developed a series of contact-based substitution matrices. The Markov model underlying the contact accepted mutations takes into account the ‘interchanging of structurally defined side-chain contacts’ and was build from 6912 pairs of domains of the CATH database [39]. While this matrix was not aimed at the detection of contacts, Eyal *et al.* [40]

used pairwise comparisons of sequences from Blocks protein alignments [41] that were associated with protein crystal structures to infer the frequencies of all possible pair substitutions for contacting and non-contacting pairs. They then created a log-odds matrix of 400 × 400 entries depicting the relative probabilities of all pair substitutions in contact and non-contact pairs. Analogously to the approach of Singer *et al.*, the probability that a pair of sites constitute a structural contact can then be estimated simply by summing the matrix entries for pair substitutions inferred from pairwise comparisons in a sequence alignment [42].

### Machine learning approaches

Machine learning approaches incorporate many different sources of information—often including but not restricted to correlated substitution data—to generate predictions of amino acid contacts. Such methods use large training sets to identify patterns of observations that tend to be associated with contacts.

Fariselli *et al.* used multiple sequence alignments associated with known protein structures to train a neural network which used, for each pair of sites, the McLachlan correlation score, information concerning site conservation, secondary structure context of each of the sites in question, the separation of the sites in the primary sequence and the frequencies of different amino acid pairs observed in the alignment (as well as frequencies of neighbouring amino acids in the primary sequence) [43]. More recently, Cheng and Baldi [44] implemented a support vector machine called SVMcon that used several measures of correlated substitutions between sites among other features including secondary structure data, site composition probabilities for a nine residue windows centred around the sites of interest and other measures of contact potential to predict amino acid contacts.

Another neural network approach that is perhaps more explicitly dependent of correlated substitutions reasons that correlations should not be restricted to sites making contacts, but should also be manifest between nearby sites in the primary sequence [45]. Pairwise correlations are thus calculated between all pairs of sites in two windows of length five amino acids, centred on the two positions of interest in the alignment. Other information used as input to the neural network includes the biophysical character of the amino acids on which the window was centred (non-polar, polar, acidic or basic), the secondary structure types in which the sites of interest were predicted to participate, the mean empirical frequency with which amino acids observed at the tested sites in the alignment were seen to constitute contacts in a training set, the length of the sequence alignment and the separation between tested sites in the alignment. The neural network of Hamilton *et al.* [45] outperformed that of Fariselli *et al.* [43] in direct comparisons [45].

## RELATIVE PERFORMANCE OF DIFFERENT METHODS

Evaluation of the performance of different correlated substitution algorithms is complicated by a series of factors. Using real data, it is almost impossible to know which pairs indeed undergo non-independent substitutions. Nevertheless, it is possible, given a crystal structure, to identify true positive contact predictions, and on occasion functional correlations not depending on spatial proximity can be confirmed through reference to the literature. However, false-positive prediction rates are hard to ascertain as it is difficult to be sure that an inferred correlation between two sites is functionally spurious.

Several authors [31, 36] have used simulated data to evaluate the behaviour of various CMA algorithms. The usual approach here is to allow the majority of sites to evolve in the usual, mutually independent, manner and to specify a subset of sites as ‘drivers’ which, when they undergo substitutions increase the probability of substitution at defined ‘acceptor’ sites. Such experiments can be informative, although it is important to recognize that the mechanism of correlation imposed is unlikely to be representative of biological evolution and indeed may more closely reflect the expectations of the methods used to identify correlated evolution.

Accordingly, most algorithms have been evaluated by their capacity to predict known contacts (analysing alignments of proteins for which at least one sequence has a known 3D structure). By informal convention, structural contacts are almost always defined as pairs of amino acids with β-carbon to β-carbon distances of <8 Å (α-carbon distances in the case of glycine residues). Constant sites and site with multiple gaps in sequence alignments are routinely excluded from all calculations as are potential interactions between sites closer than around 10 positions in the primary sequence (interactions due to secondary structure or primary sequence proximity being of little interest). The sensitivity (percentage of true contacts that are detected) of CMA algorithms is not typically considered, while accuracy (the percentage of predictions that are correct) is considered as a key performance indicator. Even when powerful methods of estimating significance of results have been implemented, it is not uncommon for the most significant (or highest scoring) $L$, $L/2$, $L/5$ or $L/10$ predictions (where $L$ is the length of a given protein) to be used and many researchers express their results in terms of improvement over random prediction as the accuracy of methods can be greatly dependent on the length of proteins considered. This is easy to appreciate as the number of contacts in a protein structure tends to increase approximately linearly with the length of the protein, while the number of possible interactions between sites increases quadratically with the length of the protein.

Relatively few studies have evaluated different correlated substitution methods on identical datasets and considering similar numbers of predictions. Fodor and Aldrich [46] performed a thorough analysis of the influence of site conservation on the behaviour of the McLachlan pairwise correlation, OMES, SCA and MI algorithms using 224 datasets derived from PFAM [47] entries with corresponding crystal structures. They concluded that the McLachlan correlation and OMES methods significantly outperformed SCA and MI and that the performance of these methods is not entirely due to their tendency to ascribe high scores to highly conserved pairs of sites, which also tend to be clustered in protein structures. It has also been shown that ELSC outperforms SCA in the prediction of contacts in the analysis of alignments corresponding to 143 PFAM entries [35].

Eyal *et al.* [40] compared the performance of their pair-to-pair substitution matrix with the contact propensity matrix method of Singer [38] and the pairwise correlation method [12]. They concluded that in the prediction of contacts in protein cores (where both residues have low solvent accessibility) their matrix outperforms other methods in terms of accuracy (up to 24% when few predictions are made), while the correlation method tends to outperform the pair-to-pair matrix method when solvent accessible regions were considered.

Halperin and coworkers [48] recently compared the performance of several algorithms to predict pairs of residues that participate in intermolecular interactions between domains of Cohesin and Dockerin proteins and in interactions between domains of fusion proteins. While this problem is distinct from the prediction of intramolecular contacts, it is of interest to note that these authors also concluded that Pearson correlation methods (such as the correlated McLachlan substitution weights) and OMES methods outperformed MI, SCA and ELSC.

We have compared the 8 Å contact prediction accuracy of four methods using a set 25 manually curated alignments of randomly selected, non-redundant, protein families whose high resolution crystal structures are available (with a mean length of 182 amino acids and an average of 93 sequences per alignment). Where possible, as in many other studies, the best $L/2$ predictions are considered. The first method considered was that of [12], using the Mclachlan amino acid similarity matrix and with other details as described in [23]. We have also implemented a mutual information parametric bootstrap analysis, similar to the one described in [28], where site substitution rate variability and root sequence reconstruction were incorporated in the parametric bootstrap. We also used the program *Dependency* [29], which implements the multiple interdependency algorithm of Tillier and Lui, and generated on average less than half the number of predictions of methods using the $L/2$ predictions per dataset rule.

Finally, we implemented a tree-based correlation (*TreeCorr*) measure similar to that of Fleishman *et al.* [27]. *TreeCorr* calculates Pearson correlation coefficients from McLachlan matrix scores from branch-specific comparisons between extant and reconstructed ancestral sequences. A $z$-score measure of significance is derived from reshuffling of substitutions on branches of the phylogenetic tree 1000 times. Maximum likelihood trees were estimated using *Phyml* [49] with the JTT substitution matrix [50] and allowing for site rate variation. Ancestral sequences were inferred using the maximum likelihood method [51] and the $L/2$ scores with the highest $z$-scores were selected.

Our data are consistent with other studies in suggesting that pairwise correlation methods outperform MI (Table 2). Although *Dependency* has an accuracy comparable to the correlation methods, considerably fewer predictions were considered. The tree-based correlation measure marginally outperformed the pairwise correlation measure suggesting

**Table 2:** Accuracies of intramolecular contact predictions by various methods. The table reports accuracies of predictions (number of true positives/number of predictions made) suggested where various methods have been directly compared or where it is straight forward to extract comparable data. Note, experimental methodologies and characteristics of alignments vary greatly between publications. In some cases the accuracy scores have been inferred from graphs in the primary publication

| **Method** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Ref Hamilton** | **Pearson-pairwise** | **Tree-correlation** | **OMES** | **SCA** | **ELSC** | **Dependency** | **MI** | **P2P** | **Singer** | **Fariselli** |
| This article[a] | 0.13 | 0.14 | | | | 0.1 | 0.02 | | | |
| [39][b] | ∼0.12 | | ∼0.11 | ∼0.05 | | | ∼0.05 | | | |
| [29][c] | | | | ∼0.07 | ∼0.09 | | | | | |
| [34][d] | ∼0.14 | | | | | | ∼0.18 | ∼0.15 | | |
| [17][e] | 0.14–0.27 | | | | | | | | | |
| [32][f] | | | | | | | | | ∼0.30 | |
| [37][g] | | | | | | | | | 0.21 | |
| [38][h] | | | | | | | | | 0.14 | 0.205 |

[a] For experimental conditions see discussion.
[b] 224 PFAM alignments, each method considers best 75 predictions.
[c] 138 PFAM alignments, each method considers best 75 predictions.
[d] 59 PFAM alignments, best $L/2$ predictions.
[e] 127 PFAM alignments, cutoff chosen for maximum accuracy in each case, range of mean accuracies for different SCOP fold categories is shown.
[f] 118 hssp alignments, best $L/10$ predictions.
[g] 173 alignments of at least 15 sequences, best $L/2$ predictions.
[h] 29 alignments of at least 15 sequences, best $L/2$ predictions.

that explicit consideration of the phylogenetic structure inherent in the data might help to increase accuracy. In agreement with other authors [46], we find that accuracy is strongly linked to the number of predictions made by both the pairwise and tree-based correlation methods. Considering $L/10$ predictions, both the tree-based and pairwise methods yielded accuracies close to 0.18.

Many methods for contact prediction have also been assessed in the CASP (critical assessment of structure prediction) experiments [52–53]. Some of these—such as SVMcon [44]—explicitly use correlated substitution data while others, for example PROFcon [54], benefit implicitly from information contained in correlated substitutions through the way that input data are encoded. A common thread linking the best performing methods is the use of many different types of data in the generation of predictions (see section on Machine learning) and it is difficult to estimate the relative contribution of correlated substitution information to their performance. The criteria used to evaluate prediction accuracy in the recent CASP7 experiment are somewhat different to those that have classically been employed in correlated substitution analyses (predictions between sites at least 24 amino acids apart in the primary sequence) [53]. However, as suggested by the relative performance of machine learning approaches that incorporate correlated substitutions and other sources of information (Table 2), the best performing contact prediction pipelines appear to consistently outperform methods based exclusively on correlated substitutions. However, it is encouraging that methods such PROFcon [54] and SVMcon [44] that use correlated substitution information—either implicitly or explicitly—have produced highly competitive contact predictions [52–53].

## CONCLUSIONS AND DISCUSSION

To date, no single method has proved itself vastly superior to others in the prediction of intramolecular contacts from correlated substitution analysis. While Pearson correlation measures and OMES seem to outperform MI, SCA and ELSC, it must be recognized that even their performance is inconsistent, particularly when more than $L/2$ predictions are made. For all methods it is difficult to know a priori whether conclusions are likely to be reliable. Additional estimates of significance that take phylogenetic correlation into account and more widespread testing of probabilistic methods are also needed.

The strong performance of the P2P matrix in direct comparison to correlation methods is striking and suggests that with additional consideration of phylogenetic issues inherent in the pairwise comparison method used in matrix construction, such approaches could provide powerful alternatives. This is perhaps not surprising as the P2P matrix considers patterns of co-substitution that actually occur at amino acid contacts. Such methods would however not be expected to perform well in the detection of functional rather than structural correlations.

One surprising aspect is the virtual neglect of DNA data as a source of information [37]. We feel that new methods based on codon comparisons are also warranted, given the capacity of DNA sequences to provide insight into site-specific selective pressures [55].

It is arguably unfair to directly compare the performance of algorithmic and machine-learning approaches—the latter make use of more data than that available to pure correlated mutation algorithms. However, at this time, the greatest potential for improvement probably comes from machine learning-based approaches such as those of Hamilton [45] and Fariselli [43] as well as the further development of empirical matrix-based approaches. It would be of interest to incorporate diverse types of correlation measures in single machine learning systems or in other types of contact prediction pipelines that at present appear to outperform pure correlated substitution methods [53].

The value of correlated substitution analysis does not lie solely with its capacity to predict intramolecular contacts. The objective of using contact maps in the *ab initio* prediction of protein structure or the selection of correct threading solutions is commendable, but with the advent of 'structural genomics' and (relatively) high-throughput crystal structure resolution [56–58], the practical need for *ab initio* structure prediction is arguably lessening. On the other hand, prediction of protein–protein interactions or the nature of interactions between proteins known to interact (docking) presents a related problem to that of intra-protein contact prediction, while protein design and customization of protein properties require a profound understanding of the functional interactions occurring within proteins. There is evidence that a significant proportion of non-contact predictions are likely to be biologically relevant. For example, it has been noted that pairs of sites recovered as correlated tend to be relatively close in protein tertiary structures (within the same functional domain at least) and thus are more likely to be functionally related even if not constituting contacts [35] and many authors have provided compelling evidence for functional long distance correlations in individual proteins [27, 30, 32, 36, 59]. Therefore, it is likely that many 'false positive' contact predictions are likely to constitute real functional correlations.

Finally, from an evolutionary and bioinformatics perspective, continuing study of correlated substitutions is of great importance. Many bioinformatics and molecular phylogenetic approaches rely on parametric bootstraps and data simulations. Failure to adequately account for significant evolutionary processes in simulations and phylogenetic analyses will inevitably lead to reconstruction of erroneous phylogenies, divergence time estimates and ancestral sequence estimations.

### Key Points

- Correlated amino acid substitutions are expected to reflect physical interactions between sites and networks of sites sharing functional constraints.
- Phylogenetic structure underlying alignments of homologous protein sequences introduces strong apparent correlations into the data.
- Various conceptually distinct methods allow significant enrichment of contact sites with correlated mutation analysis, but discrimination of functional and structural correlations remains an unsolved problem.

**Acknowledgements**
This work was in part funded by a European Community Marie Curie category 30 individual fellowship (MCFI-2001-00634) to D.S.H. and by FIRB (MIUR, Italy) to G.P.

## *References*

1. Wheelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 2001;**17**:262–72.
2. Yang Z. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol Evol* 1996;**11**:367–72.
3. Mayrose I, Doron-Faigenboim A, Bacharach E, *et al*. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* 2007;**23**:i319–27.
4. Griffiths-Jones S, Bateman A, Marshall M, *et al*. Rfam: an RNA family database. *Nucleic Acids Res* 2003;**31**:439–41.
5. Dutheil J, Pupko T, Jean-Marie A, *et al*. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* 2005;**22**:1919–28.
6. Cao Y, Adachi J, Janke A, *et al*. Phylogenetic relationships among Eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J Mol Evol* 1994;**39**:519–27.

7. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 1970;**4**:579–93.
8. Lopez P, Casane D, Philippe H. Heterotachy, an important process of protein evolution. *Mol Biol Evol* 2002;**19**:1–7.
9. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;**257**:342–58.
10. Lin K, Kleinjung J, Taylor WR, *et al*. Testing homology with contact accepted mutatiOn (CAO): a contact-based Markov model of protein evolution. *Comput Biol Chem* 2003;**27**:93–102.
11. Kleinjung J, Romein J, Lin K, *et al*. Contact-based sequence alignment. *Nucleic Acids Res* 2004;**32**:2464–73.
12. Gobel U, Sander C, Schneider R, *et al*. Correlated mutations and residue contacts in proteins. *Proteins* 1994; **18**:309–17.
13. Neher E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA* 1994;**91**: 98–102.
14. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. *Protein Eng* 1994;**7**: 341–48.
15. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins* 1998;**32**:475–94.
16. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999;**286**:295–9.
17. Chelvanayagam G, Eggenschwiler A, Knecht L, *et al*. An analysis of simultaneous variation in protein structures. *Protein Eng* 1997;**4**:307–16.
18. Atchley WR, Wollenberg KR, Fitch WM, *et al*. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol* 2000;**17**:164–78.
19. Noivirt O, Eisenstein M, Horovitz A. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng* 2005;**18**:247–53.
20. Afonnikov DA, Oshchepkov DY, Kolchanov NA. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics* 2001;**17**:1035–46.
21. Afonnikov DA. CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res* 2004;**32**:W64–8.
22. McLachlan AD. Test for comparing related amino acid sequences. *J Mol Biol* 1971;**61**:409–24.
23. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 1997;**2**: S25–32.
24. Vicatos S, Reddy BVB, Kaznessis Y. Prediction of distant residue contacts with the use of evolutionary information. *Proteins* 2005;**58**:935–49.
25. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng* 1997;**10**:647–57.
26. Fukami-Kobayashi K, Schreiber DR, Benner SA. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J Mol Biol* 2002;**319**:729–43.
27. Fleishman SJ, Yifrech O, Ben-Tal N. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J Mol Biol* 2004;**340**: 307–18.
28. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci USA* 2000;**97**: 3288–91.
29. Tillier ERM, Lui TWH. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 2003;**19**:750–5.
30. Gloor GB, Martin LC, Wahl LM, *et al*. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 2005;**44**: 7156–65.
31. Martin LC, Gloor GB, Dunn SD, *et al*. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 2005;**22**:4116–24.
32. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 2002;**48**:611–7.
33. Larson SM, Di-Nardo AA, Davidson AR. 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* 2000;**303**:433–46.
34. Suel GM, Lockless SW, Wall MA, *et al*. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 2003; **10**:59–69.
35. Dekker JP, Fodor A, Aldrich RA, *et al*. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics* 2004;**20**:1565–72.
36. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* 1999;**287**: 187–98.
37. Dimmic MW, Hubisz MJ, Bustamante CD, *et al*. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* 2005;**21**:i126–35.
38. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng* 2002;**15**:721–5.
39. Orengo CA, Michie AD, Jones S, *et al*. CATH–a hierarchic classification of protein domain structures. *Structure* 1997;**5**: 1093–108.
40. Eyal E, Frenkel-Morgenstern M, Sobolev V, *et al*. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins* 2007; **67**:142–53.
41. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;**89**: 10915–19.
42. Eyal E, Pietrokovski S, Bahar I. Rapid assessment of correlated amino acids from pair-to-pair (P2P) substitution matrices. *Bioinformatics* 2007;**23**:1837–9.
43. Fariselli P, Olmea O, Valencia A, *et al*. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng* 2001;**14**:835–43.

44. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007;**8**:113–22.

45. Hamilton N, Burrage K, Ragan MA, *et al*. Protein contact prediction using patterns of correlation. *Proteins* 2004;**56**: 679–84.

46. Fodor A, Aldrich RA. Influence of conservation on calculations of amino acid covariance in multple sequence alignments. *Proteins* 2004;**56**:211–21.

47. Bateman A, Birney E, Cerruti L, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2002;**30**:276–80.

48. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins* 2006;**63**:832–45.

49. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003;**52**:696–704.

50. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;**8**:275–82.

51. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997;**13**:555–6.

52. Graña O, Baker D, MacCallum RM, *et al*. CASP6 assessment of contact prediction. *Proteins* 2005;**61**:214–24.

53. Izarzugaza JMG, Graña O, Tress ML, *et al*. Assessment of intramolecular contact predictions for CASP7. *Proteins* 2007;**69**:152–8.

54. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *Bioinformatics* 2005;**21**:2960–8.

55. Creevey CJ, McInerney JO. An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. *Gene* 2002;**300**:43–51.

56. Burley SK, Almo SC, Bonanno JB, *et al*. Structural genomics: beyond the human genome project. *Nat Genet* 1999;**23**:151–7.

57. Kim S. Shining a light on structural genomics. *Nat Struct Biol* 1998;**5**(Suppl):643–5.

58. Montelione G, Anderson S. Structural genomics: keystone for a human proteome project. *Nat Struct Biol* 1999;**6**:11–12.

59. Fares MA, Travers SAA. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* 2006;**173**:9–23.