COMPUTATIONAL TOOLS

# Emerging methods in protein co-evolution

*David de Juan[1], Florencio Pazos[2] and Alfonso Valencia[1]*

Abstract | Co-evolution is a fundamental component of the theory of evolution and is essential for understanding the relationships between species in complex ecological networks. A wide range of co-evolution-inspired computational methods has been designed to predict molecular interactions, but it is only recently that important advances have been made. Breakthroughs in the handling of phylogenetic information and in disentangling indirect relationships have resulted in an improved capacity to predict interactions between proteins and contacts between different protein residues. Here, we review the main co-evolution-based computational approaches, their theoretical basis, potential applications and foreseeable developments.

Molecular phylogenetics
The study of evolutionary phenomena using biomolecular data, generally in the form of sequences of nucleic acids or proteins.

Covarion model
A phylogenetic model in which the evolutionary rate of different codons are interdependent.

Protein family
A set of homologous proteins defined according to a given threshold of sequence similarity.

[1]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.*
[2]*Computational Systems Biology Group, National Centre for Biotechnology (CNB–CSIC), Madrid, Spain.*
*Correspondence to A.V.*
*e-mail:* avalencia@cnio.es
doi:10.1038/nrg3414
Published online 5 March 2013

As a part of evolutionary theory, co-evolution is essential for understanding living systems. In its simplest definition, co-evolution refers to the coordinated changes that occur in pairs of organisms or biomolecules, typically to maintain or to refine functional interactions between those pairs. Darwin himself initiated the study of co-evolution, and his observation on the relationship between the size of orchids' corollae and the length of the proboscis of pollinators led him to predict successfully the existence of a new species that was able to suck from the large spur of Darwin's orchid. The studies of Dobzhansky[1] and others[2] contributed to the establishment of this concept in genetic terms, although the term co-evolution is usually attributed to Ehrlich[3], and it is commonly defined as 'reciprocal evolutionary change in interacting species'[4].

For the past 20 years, much effort has been dedicated to investigating co-evolution at the molecular level. In a classical study, coordinated sequence changes among genes (and their protein products) were proposed to be essential to optimize physiological performance and reproductive success[5], thus indicating that molecular co-evolution could be an important and widespread determinant of fitness.

Much progress has been made in the development of computational tools for the detection of molecular co-evolution. Although co-evolution can potentially occur between various biomolecules, most recent tools focus on protein co-evolution. These can be broadly divided into those methods that search for co-evolution at the amino acid residue level and those that search at the protein level, all of which are based on principles of molecular phylogenetics.

Tools at the residue level were inspired by the existence of interdependent changes in groups of variable amino acids, as formulated for the first time by the covarion model[6], and they typically use the multiple sequence alignment (MSA) for a protein family of homologues to search for correlated mutations. Such correlated mutations are suggestive of compensatory changes that occur between entangled residues (for example, those in proximity, direct contact or acting together in catalytic or binding sites) to maintain protein stability, function or folding[7–10]. Furthermore, extending these methods to search for correlated mutations between pairs of interacting proteins can identify sites of inter-protein interaction[11–17]. In parallel, related methods have been developed to search for larger groups of residues that are specifically co-conserved within particular protein subfamilies. These methods can identify residues that define functional properties of that subfamily, such as substrate binding specificity of a given enzyme[18,19].

Methods for detecting co-evolution at the protein level often mine phylogenetic trees that have been built for many protein families using entire protein sequences[20,21]. The co-evolution between interacting species, such as parasites–hosts, predators–prey and symbionts–hosts, is in many cases manifested as a similarity of the phylogenetic trees of these co-evolving species. Likewise, molecular co-evolution caused by physical or functional protein interactions frequently results in similarities of the corresponding protein family trees. Consequently, approaches based on protein family tree similarity can successfully identify interaction partners for a given protein, such as ligand–receptor pairs[21,22].

Homologues
Genes and proteins arisen from a common ancestor. In most cases, this common origin is traceable at the sequence level, albeit the sequence similarity can be very low and difficult to detect.

Correlated mutations
Relationship between two positions of a multiple sequence alignment in which the amino acid changes in one of the positions (mutational pattern) parallels that in the other.

The increasing availability of protein sequences and key new methodological advances aimed at disentangling direct and indirect co-evolutionary signals have renewed the interest in co-evolution-based prediction strategies, leading to possible applications in a wider range of biological problems, such as the prediction of contacts in protein structures, sites of specific protein interactions and the predictions of protein interaction partners at the genomic scale.

In this Review, we present the most prominent computational methods based on co-evolution (FIG. 1). We progress through methods analysing pairs, and then groups, of co-evolving residues through to methods at the whole-protein level, and we describe the biological problems to which they have been applied and briefly explain the concepts and algorithms behind them. Key technical details (including the underlying algorithms) that differentiate the methods are described in Supplementary information S1 (boxes), and examples of the connection between computational and experimental approaches are given in BOXES 1,2,3. For a summary of practical information about these computational methods and tools, see TABLE 1. Finally, we discuss the problems relating to the interpretation of the results in evolutionary terms and present what, in our opinion, will be the most promising future developments.

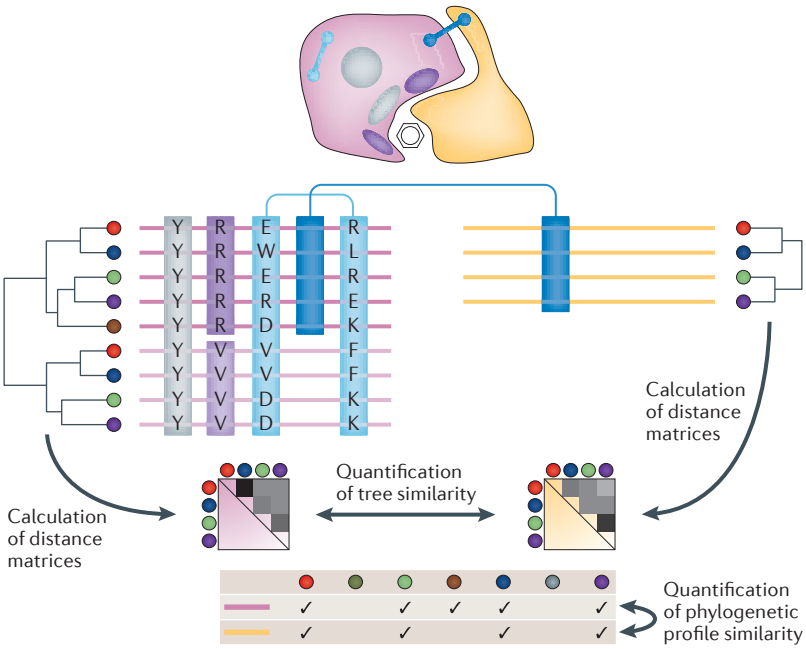



Figure 1 | **Co-evolutionary features extracted from protein multiple sequence alignments.** The three-dimensional structures of two interacting proteins (purple and yellow) are schematized as well as their multiple sequence alignments (MSAs) and phylogenetic trees based on orthologous sequences from a number of organisms. Circles of different colours represent different species from which the protein sequences are derived. The purple MSA includes a family of paralogues that are present in some of these organisms (sequences in dark purple). A number of evolutionary and co-evolutionary features are schematized in the alignments and the corresponding three-dimensional structures. Fully conserved positions (grey) tend to form a part of the protein core and are also in functional regions (such as protein interaction sites and catalytic sites). Specificity-determining positions (SDPs; purple) tend to be in functional sites conferring specificity. Intra-protein correlated mutations (light blue) are related to residue spatial proximity, whereas inter-protein correlated mutations (dark blue) reflect in many cases proximity between residues in different protein chains. Additionally, the protein–protein co-evolution can be evaluated: the phylogenetic trees of the two families can be compared in different ways, as shown below the MSAs. For example, the trees can be converted to distance matrices to quantify the tree similarity. Alternatively, the similarity of the patterns of presence or absence of the two proteins in a set of genomes of different species (phylogenetic profiles) is also an indication of co-dependence.

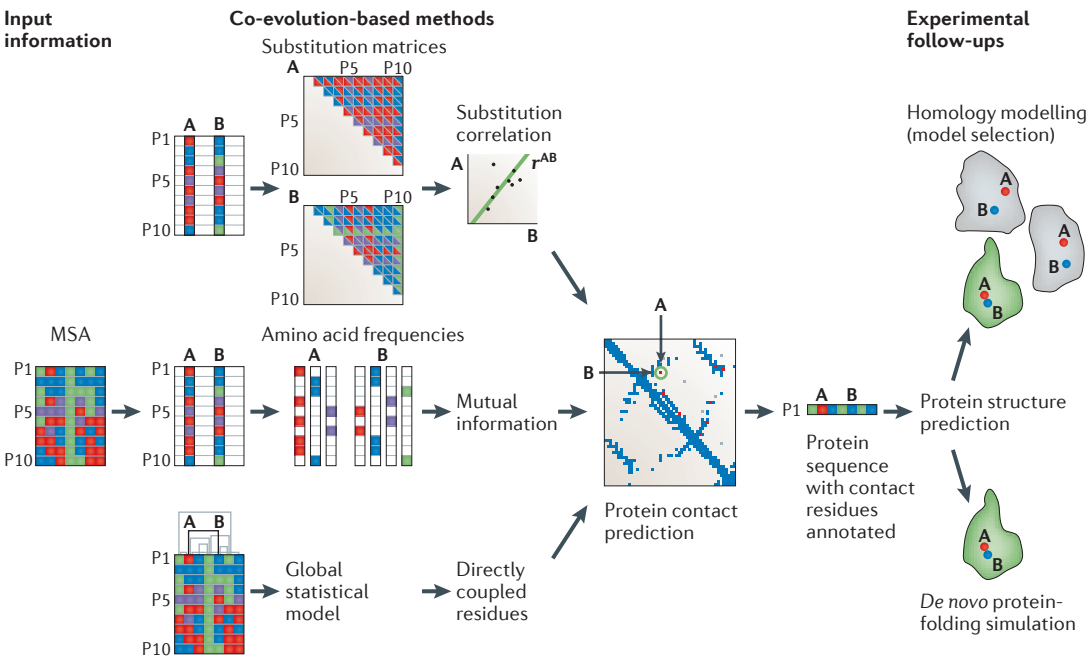
## Co-evolution at the residue level

Substantial effort has been invested in studying the co-evolution of pairs of positions in MSAs of protein families (that is, residue co-evolution). These pairs of co-evolving positions were often found to correspond to spatially proximal residues in the protein structure, and such putative inter-residue contacts have aided protein structure prediction (BOX 1). Recent methodological developments have improved the accuracy of protein contact prediction by disentangling the direct pairwise couplings from the background network of coordinately mutating positions. Furthermore, co-evolution between residues in different proteins has been used as predictors of the interacting surfaces (protein interfaces) in protein complexes as well as in the search for interacting partners of a given protein, as discussed later in 'Hybrid residue–protein methods'.

***Detecting correlated amino acid changes in pairs of positions.*** Residue co-evolution was originally assessed through detecting pairs of positions (two columns of the MSA) that have interdependent amino acid frequencies[23] or similar patterns of amino acid substitutions[7,9,10]. In particular, these substitution patterns can be built according to a pre-calculated amino acid substitution matrix, and their similarity can be assessed by a linear correlation. This approach is often called the McLachlan-based substitution correlation (McBASC). This method has been extensively tested and compared with newer methods and shows a small but significant capability to recover pairs of positions in physical contact[24] and still serves as a baseline to benchmark the performance of new methods[25].

Estimation of inter-position co-variation using substitution matrices imposes a demanding scenario for highly variable positions to be assigned as co-varying. Consequently, McBASC tends to detect co-evolving positions preferentially that are fairly well-conserved during evolution[24]. This approach provides a rough estimate of the magnitude of the correlation between positions without assuming a specific compensatory biochemical nature. The initial proposal lacked a defined confidence threshold to extrapolate between cases. In addition, it did not address problems associated with the alignment quality, such as the inclusion of divergent or redundant sequences[26]. Inspired by this approach, co-evolution analysis using protein sequences (CAPS)[27] dampens the influence of background phylogenetic divergence by requiring the detected correlations to still be detected after particular clades are removed from the MSA. It also corrects the amino acid substitution matrix so as to consider the actual divergence among the sequences. Owing to the high computational

Box 1 | **From pairwise residue co-evolution to protein structure prediction**




Different co-evolution-based methodologies are used for predicting residue contacts on the basis of information extracted from an input multiple sequence alignment (MSA). The three main approaches for evaluating the co-evolution between two residues involve substitution correlations, mutual information of amino acid frequencies or — as in direct coupling analysis (DCA) or protein sparse inverse covariance (PSICOV) — a global statistical model of the MSA. Actual or inferred protein contacts are traditionally represented as contact maps, where every point represents a contact between two residues (for example, A and B in the figure)[91].

Predicted contact maps have been used to help the prediction of three-dimensional protein structures in different methodological contexts. For example, homology modelling or fold recognition methods can suggest various structural models based on known structures of related proteins. These can then be filtered into the most likely models that contain the residue contacts predicted by correlated mutations, mutual-information-based[7–10,92] or DCA-like approaches[93]. By contrast, *de novo* protein modelling attempts to predict protein structures without prior structural information from related proteins. This often involves computationally intense protein-folding simulations. Pairwise constraints can be an effective way of reducing the vast space that has to be explored by the *de novo* protein-folding simulations. Although the accuracy of substitution correlation and mutual information approaches for predicting residue contacts is not enough to produce systematically reliable *de novo* protein models, new prediction methods (such as PSICOV and DCA)[38,39] that disentangle directly from indirectly coupled positions have improved the prediction of protein structures by *de novo* protein-folding simulations[94–97]. The application of these approaches produced accurate protein contact predictions for two sets of ~150 large protein families (both with more than 1,000 members)[38,39]. In fact, in one of these analyses DCA showed an average true-positive rate of up to 0.8 for the first 20 predictions[38]. Further use of these constraints in *de novo* structure prediction, with the help of predicted secondary structure, resulted in high-quality protein models for short protein sequences[94,97], and their combination with other topological features retrieved equally good results for two sets of transmembrane proteins[95,96]. Although these methods have been used on proteins belonging to large families, a recent evaluation of the performance of one of them (PSICOV)[43] in a more general predictive scenario shows that its accuracy can drop to 20% of correct predictions in typical protein families that have a fairly small number of members (that is, proteins of unknown structure submitted to the Critical Assessment of Techniques for Protein Structure Prediction (CASP) evaluation of protein structure prediction methods).

demands of CAPS, this methodology has been tested in specific cases but not yet in a large-scale data set.

Mutual information has been also used to detect co-varying positions. Whereas correlation-based methods explore inter-sequence amino acid substitutions, mutual information considers the distribution of each amino acid in the different sequences for a position. In fact, mutual information quantifies whether the presence of an amino acid in a given sequence for a position is a 'good prediction' of the presence of any given amino acid in the same sequence for a second position[23]. In this sense, mutual information does not account for which particular amino acids are present in the same sequences in both positions but relies on the statistical significance of the observed co-variations. Therefore, the different amino acids are treated as different symbols that are not related by similarity relationships, and the magnitude of the biochemical changes is not taken into account when assessing the similarity of mutational patterns. The initial formulations of this approach were vulnerable to large variations in sequence conservations in the MSAs[24] as well as to the effect of the phylogenetic

Phylogenetic trees
Representations of the evolutionary relationships between a set of biological entities (such as proteins, genes or organisms).

Protein interfaces
Regions of the surface of a protein involved in the interaction with others.

Amino acid substitution matrix
A matrix containing, for every possible pair between the 20 canonical amino acids, a quantification of the 'interchangeability' of one by the other in the same protein site, as a proxy of the evolutionary feasibility of the corresponding change (mutation). They are often derived from curated sets of MSAs assumed to contain real representations of the amino acids allowed at a given protein site.

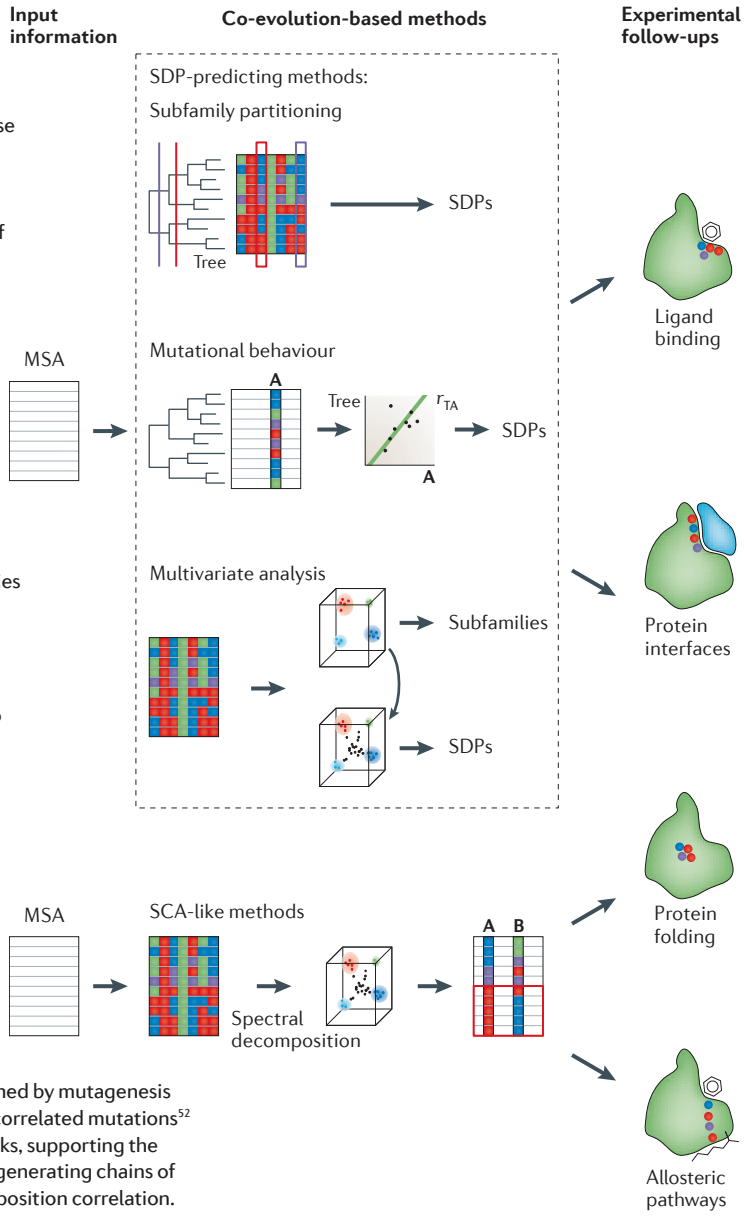
Benchmark
In bioinformatics, this term describes the assessment of the performance of a method using a set of examples of known outcome (the 'gold standard'), particularly by testing its predictive power relative to current best practice tools.

Clades
Groups of entities (such as genes or organisms) in a phylogenetic tree that have all arisen from a common ancestor.

Homology modelling
Protein structure prediction technique that, on the basis of the proven relationship between sequence similarity and structural similarity, models the three-dimensional structure of a protein based on the (experimentally determined) structure of a homologue (known as a 'template' in this context). Also known as 'comparative modelling'.

*De novo* protein modelling
Any approach for predicting protein structure that does not make use of information on other existing protein structures (such as those of homologues). Also known as 'ab initio modelling'.

Box 2 | **Groups of co-evolving residues are implicated in functional specificity and structure–function coordination**




Groups of co-evolving residues are expected to reflect the coordinated action of these residues in a functional or structural context. Specificity-determining positions (SDPs) are groups of positions that coordinately mutate in the context of subfamily divergence. They can be detected by various methods, including those that partition the evolutionary tree into subfamilies, mutational behaviour and multivariate analyses (in the dashed box in the figure). SDPs tend to form three-dimensional clusters[46,47] at ligand- and/or protein-binding sites[18,19,47,48] to determine the functional specificity of the protein. Therefore, SDPs are often used to predict ligand- and protein-binding sites, which can help to interpret disease-associated mutations[98,99] or to design direct mutagenesis experiments to alter protein function[100–104]. The most common experimental approaches have been to mutate SDP positions to impede partner binding[102,104] or to exchange the residues (or regions) between members of different subfamilies to switch the corresponding binding partners[100,101,103].

Other examples include the study of co-evolution between proliferating cell nuclear antigen (PCNA) and its interaction partners across species of fungi[105]. PCNA orthologues clearly segregate into two subfamilies, from which a SequenceSpace-like analysis identified SDPs in PCNA that distinguish these two subfamilies and that represent partner-binding sites. Experimentally switching these sites with those from the other subfamily disrupted PCNA–partner interactions and caused cell death. This suggests that co-evolution of protein interaction networks (in this case, PCNA and its interaction partners) could contribute to hybrid incompatibility to promote and to stabilize speciation. Similarly, co-evolution between components of signalling cascades can retain evolutionarily constrained interactions while reducing pathway crosstalk[106]. In this case, an evolutionary-trace-like analysis of the bacterial PhoR kinases and their PhoB substrates detected SDPs that were specific to the α-proteobacterial clade. Inter-clade amino acid exchange showed that these residues have a role in avoiding crosstalk to another signalling cascade that arose specifically in α-proteobacteria.

Statistical coupling analysis (SCA)-like approaches have successfully been used to explore the implication of networks of co-evolving residues in protein folding[50] and allosteric communication[51,53] (see the lower panel of the figure), and the functional importance of these sites has been experimentally confirmed by mutagenesis or chimaera generation in several cases[54]. Interestingly, networks of correlated mutations[52] and SDPs[49] have been also anecdotally related with allosteric networks, supporting the recruitment of networks of co-evolving residues as a mechanism for generating chains of allosteric transmission. MSA, multiple sequence alignment; $r_{TA}$, tree–position correlation.

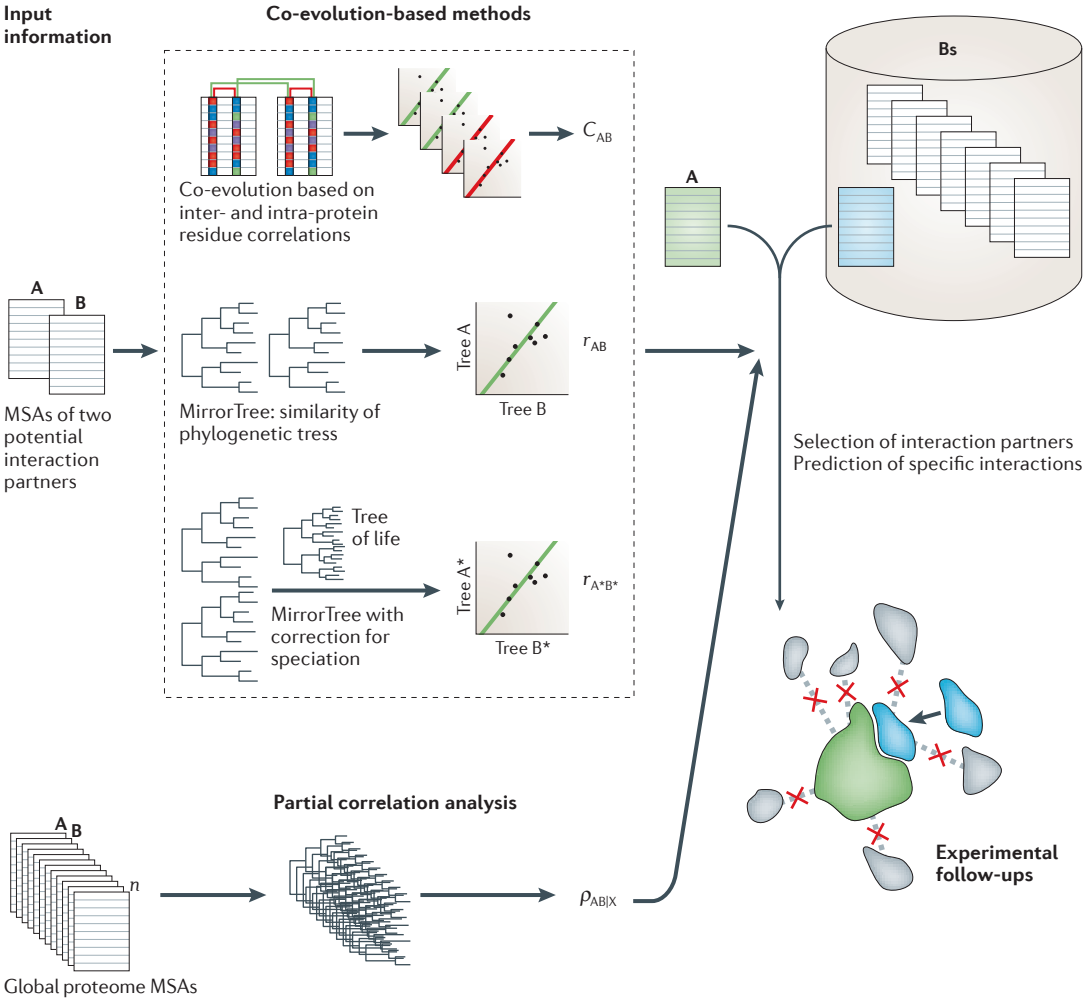
background. These problems have been addressed in subsequent versions[25,28,29] that improved contact prediction performance[25] (represented as 'Mutual information corrected' in FIG. 2), providing important information for understanding protein structures[30].

McBASC and mutual information approaches are probably the most commonly used approaches to study residue co-evolution, although many others have been developed over recent years, such as methods addressing the similarities of vectors that represent the presence or absence of the 20 different amino acids in each position of the MSAs[31,32]. Other methods use phylogenetic approaches to start from a reconstructed ancestral state and then to characterize the sequence of evolutionary changes that have occurred over time to detect patterns of simultaneous substitution[13,33,34] or explicitly to contrast independent evolution and co-evolution models[35,36]. In this case, the use of an enhanced continuous-time Markov process model for sequence co-evolution represented an important step forwards[13]. These approaches are suitable for small-scale studies of co-evolution in small protein families, but the evaluation of their performance in large-scale studies remains excessively demanding in computational terms.

***Disentangling directly coupled residues from the network of indirectly correlated positions.*** An important obstacle in the detection of co-evolving positions is the apparent co-variation or indirect coupling that can occur when more than two positions show coordinated substitution patterns. In these cases, the apparent co-variation between two positions is the consequence of

Mutual information
In information theory, this is entropy-based formulation for quantifying the interdependence between the values of two random categorical variables.

Box 3 | **Protein–protein co-evolution has an important role in many biological systems**




Many methodologies for quantifying the co-evolution between two proteins use multiple sequence alignments (MSAs) as their input. These can be used to generate phylogenetic trees to infer co-evolution on the basis of tree similarity (in the dashed box of the figure). Some methods, such as Tol-MirrorTree, use the global evolutionary relationships between species (the 'tree of life') to correct the background tree similarity due to speciation. Another group of approaches, such as ContextMirror, corrects this and other factors affecting observed protein tree similarities using information from large collections of trees (for example, those derived from the whole proteome of interest; lower panel of the figure).

To maintain interactions during evolution, binding partners could remain conserved, or they could co-evolve. Co-evolution has been found in systems that must evolve quickly or when proteins acquire new functions while keeping the interactions between the involved partners. For example, some nuclear-encoded members of the mitochondrial NADH–ubiquinone reductase complex have accelerated evolutionary rates, possibly to accommodate the intrinsic accelerated evolution of their mitochondrial counterparts. In one study, a method based on the accumulation of inter-protein correlated mutations (see 'Hybrid residue–protein methods' in the main text) was used to uncover the direct physical interactions within this large complex of 45 subunits[107]; interactions were then experimentally confirmed using a yeast two-hybrid approach.

In another application, the MirrorTree approach[21,108] was used to demonstrate co-evolution between some pairs of proteins involved in redox homeostasis and cellular timekeeping[109]. The functional interconnectivity of these seemingly disparate processes was shown by the oxidation–reduction cycles of peroxiredoxin proteins being universal markers for circadian rhythms across bacteria, archaea and eukaryotes, despite the large mechanistic differences[109].

Finally, sex-related molecular systems are another prototypic case of rapidly evolving systems in which co-evolution has an important role, because they have to differentiate and to acquire specificity quickly so as to avoid cross-fertilization while maintaining the specific interactions. Studies of co-evolution in these systems include the sequencing of 14 alleles of *Brassica campestris* genes, which found co-evolution between the (male) SCR and the (female) S receptor kinase[110]. This system forms the basis of the pollen discrimination mechanism. More recently, deep sequencing was used to study the co-evolution between male and female fertilization proteins in abalone snails[111]. $\rho_{AB|X}$, partial correlation between proteins A and B given any X protein; $C_{AB}$, interaction index; $r_{AB}$, tree–tree correlation; $r_{A*B*}$, corrected tree–tree correlation.

Continuous-time Markov process
Process in which a system explores along time different states of a finite 'state space' in such a way that the Markov property is satisfied. This property means that the probability distribution of the system at a time point given the whole history of the process up to a previous time depends only on the state of the system at that previous time.

the evolutionary interdependence of both positions with one or more additional positions. The aggregation of these indirect couplings can make it difficult to recognize the directly interdependent positions.

As the direct couplings are more reliable for predicting physically proximal residues in protein structures, approaches are needed to distinguish direct from indirect couplings. Such methods are varied, but all consider that a set of direct pairwise couplings between positions seeds a larger set of indirectly coupled positions to form the whole network of coordinately mutating positions. A first basic model was proposed by Lapedes *et al.*[37], who assumed that indirect couplings do not represent evolutionary interdependence and can be considered to be uninformative pairwise co-variations. This first approach used a Monte Carlo algorithm to infer the simplest probabilistic model that was able to account for the whole network of co-variations in a simulated scenario. The importance of this first approach remained unacknowledged until a number of recent publications revisited the problem with different strategies[15,38,39]. Direct coupling analysis (DCA)[15–17,38] and protein sparse inverse covariance (PSICOV)[39] establish a global statistical model of the MSA in terms of position-specific variability and inter-position coupling[38,39]. Heuristic approaches are used to resolve the model, obtaining the estimated values of direct interposition couplings that, in the case of DCA, can finally be transformed into a mutual-information-based formulation. It is interesting that a related approach was also useful to carry out sequence homology searches[40].

Alternatively, Burger and van Nimwegen's[41] method uses a Bayesian network model that includes pairwise conditional dependencies, and the regularized multinomial regression-based correlated mutations (RMRCM) approach[42] takes into account the whole network of dependencies and not only the individual pairwise dependencies.

It is still too early to compare these methodologies fully, but they represent an important advance in the field. For MSAs with more than 1,000 sequences, DCA and PSICOV seem to be superior to Burger and van Nimwegen's method[38,39]. In fact, some of these methods are able to predict contacts between residues far apart in the linear sequence with sufficient accuracy as to be useful for guiding *in silico* folding experiments (BOX 1). Nevertheless, such clear improvements are obtained only for protein families with thousands of members[43]. Direct coupling approaches are expected to remove unspecific influences, such as the previously discussed phylogenetic background[44,45], although this attractive possibility needs further investigation. It will be interesting to see how the application of these new methods progresses in the hands of the scientific community, as the results in the recent Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition remain largely inconclusive.

## Groups of co-evolving residues

Although various methods that are focused on pairwise interactions actively exclude larger groups of co-varying residues, such groups of residues can provide useful information albeit that is not always related to direct contacts. In many protein families, such as kinases, their phylogenetic trees reveal various subfamilies, which often represent proteins that, in the framework of the common function of the whole family, have different functional specificities, such as binding of different substrates or effectors. Whereas some positions are conserved in all sequences — and thus might represent residues with important structural or catalytic roles across the whole protein family — other positions may be conserved only within particular subfamilies (FIG. 1). These subfamily-specific residues are likely to define the specific functionality of that subfamily, such as forming three-dimensional clusters[46,47] that make-up ligand- and/or protein-binding sites[18,19,47,48] or allosteric chains of residues[49]. This family-dependent conservation pattern results in these positions showing correlated mutational patterns, making this phenomenon an essential part of the study of molecular co-evolution (BOX 2).

These positions were originally termed 'tree-determinant' positions to reflect their relation with the structure of the phylogenetic trees and were more recently renamed as specificity-determining positions (SDPs) to highlight their potential functional role. As discussed below, various tools are being developed to detect SDPs. Additionally, related methods based on statistical coupling analysis (SCA) explicitly search for groups of co-evolving residues that can contribute to processes such as protein folding[50] or allosteric interactions[51–54] (BOX 2); these methods are distinguished from SDP-detecting methods by their less stringent requirement for identified residues to be specific to particular protein subfamilies.

***Methods using phylogenetic trees.*** Analysing the conservation in different branches of phylogenetic trees is perhaps the most obvious approach for detecting SDPs. The methods that implement this idea start by building a tree from the MSA before establishing branching points that partition the tree into subfamilies in which there is a significant concentration of specifically conserved positions (that is, SDPs for that subfamily). This idea was first implemented in the evolutionary trace method[19], which explores the hierarchical organization of a protein family by following the similarity of the sequences that split in each tree branch and then scores these positions as a function of when they became conserved (how close they are to the origin of the tree). Evolutionary trace avoids the need to determine an optimal partition of the MSA into subfamilies. However, other approaches, such as the approach proposed by Hannenhalli and Russell[55] and the S method[47], explicitly look for the optimal partition of subfamilies by comparing the distributions of intra- and inter-group residue entropy for every possible split in the tree.

Evolutionary trace has been improved by calculating a score on the basis of the weighted entropy of a position in all the possible subfamilies defined by the nodes of the corresponding phylogenetic tree[56] (a method called rvET, for 'real value evolutionary trace'). The evolutionary trace approach has also been independently improved by detecting SDPs using the mutual information of

Monte Carlo algorithm
An algorithm based on simulated repeated random sampling to obtain approximate solutions to complex mathematical and statistical problems.

Heuristic approaches
Methods that makes use of approximations or assumptions so as to reduce the search space but that consequently do not ensure the exact solution to be found.

Bayesian network
Probabilistic model in which a set of random variables (nodes) and their conditional dependencies (directed edges) are arranged in a network representation.

Residue entropy
Quantification of the evolutionary variability of the position of a multiple sequence alignment corresponding to a given protein residue based on the 'entropy' parameter of information theory.

Orthologues
Homologous genes or proteins split in a speciation event, ending up in different organisms.

Table 1 | **Representative protein co-evolution methods**

| Method | Analysis | Main application | Servers and databases | Refs |
|---|---|---|---|---|
| ***Inter-residue co-evolution*** | | | | |
| Mutual information | Simple inter-position co-evolution | Protein contacts (model selection for homology modelling) | Co-evolution analysis server (http://coevolution.gersteinlab.org/coevolution) | 23 |
| Mutual information corrected (Mlp) | Inter-position co-evolution without phylogenetic contribution | Protein contacts (model selection for homology modelling) | | 25 |
| McBASC | Simple inter-position co-evolution | Protein contacts (model selection for homology modelling) | Co-evolution analysis server (http://coevolution.gersteinlab.org/coevolution) | 7 |
| CAPS | Inter-position co-evolution without phylogenetic contribution | Protein contacts (model selection for homology modelling) | CAPS server (http://bioinf.gen.tcd.ie/caps/home.html) | 27 |
| DCA or DCA optimized | Pair-specific inter-position co-evolution | Protein contacts (ab initio protein structure prediction) | | 15, 38 |
| PSICOV | Pair-specific inter-position co-evolution | Protein contacts (ab initio protein structure prediction) | | 39 |
| ***SDPs*** | | | | |
| Evolutionary trace | SDPs | Ligand and protein interaction specificity | Evolutionary Trace Server (http://mammoth.bcm.tmc.edu/ETserver.html) | 56 |
| SDPsite | SDPs and subfamilies | Ligand and protein interaction specificity | SDPsite (http://bioinf.fbb.msu.ru/SDPsite/index.jsp) | 58 |
| Mutational behaviour | SDPs | Ligand and protein interaction specificity | TreeDet (http://treedetv2.bioinfo.cnio.es/treedet/index.html) | 47 |
| SequenceSpace | SDPs and subfamilies by visual inspection | Ligand and protein interaction specificity | | 18 |
| S3det | SDPs and subfamilies | Ligand and protein interaction specificity | TreeDet (http://treedetv2.bioinfo.cnio.es/treedet/index.html) | 48 |
| ***SCA-like*** | | | | |
| SCAold | Conditioned conservation | Intra-protein pathways (allostery) | | 51 |
| SCAnew | Subfamily-specific conservation | Intra-protein pathways (allostery) | | 54 |
| ***Inter-protein co-evolution*** | | | | |
| MirrorTree | Simple inter-protein co-evolution | Physical and functional interactions | MirrorTree Server (http://csbg.cnb.csic.es/mtserver) | 21 |
| i2h | Simple inter-protein co-evolution | Physical and functional interactions | | 88 |
| Tol-MirrorTree | Inter-protein co-evolution without phylogenetic contribution | Physical and functional interactions | | 67 |
| ContextMirror | Pair-specific inter-protein co-evolution | Physical and functional interactions | EcID database (http://ecid.bioinfo.cnio.es) | 71 |
| MMM | Inter-protein co-evolution of the strongest co-evolving sequence in the alignments | Physical and functional interactions | MatrixMatchMaker Web interface (http://www.uhnresearch.ca/labs/tillier/MMMWEBvII/MMMWEBvII.php); MMM-D database of co-evolving proteins (http://tillier.uhnres.utoronto.ca/MMMD.php) | 74 |
| Phylogenetic profiles | Sequence presence- or absence-associated inter-protein co-evolution | Physical and functional interactions | STRING database (http://www.string-db.org) | 80 |

A more comprehensive version of this table is available in Supplementary information S2 (table). CAPS, co-evolution analysis using protein sequences; DCA, direct coupling analysis; i2h, *in silico* two-hybrid; McBASC, McLachlan-based substitution correlation; MMM, MatrixMatchMaker; PSICOV, protein sparse inverse covariance; SCA, statistical coupling analysis; SDP, specificity-determining positions; STRING, search tool for the retrieval of interacting genes/proteins.

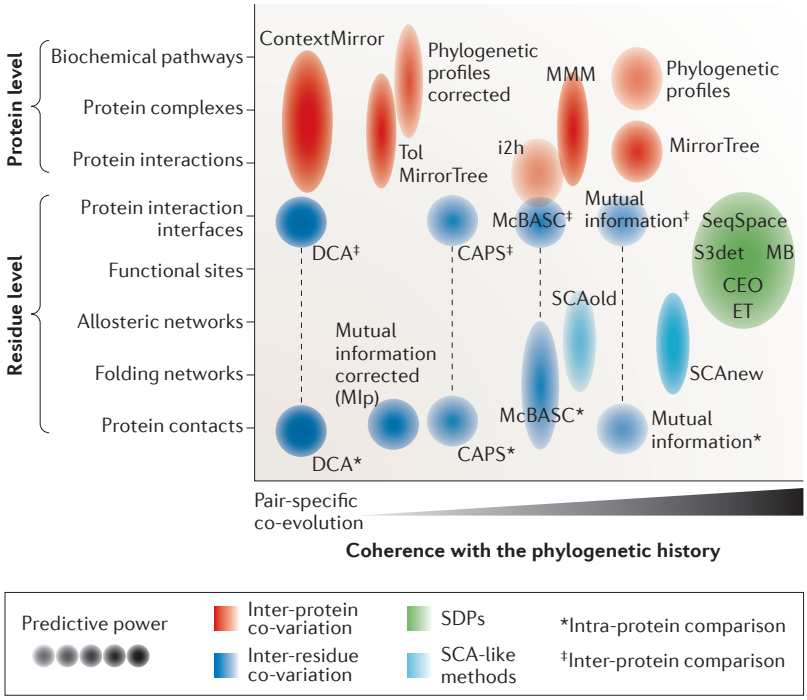



Figure 2 | **Influence of phylogenetic history in the association of co-evolution and different types of molecular interactions.** This schema shows the most representative methods according to the molecular interactions with which they are associated and the coherence of their results to the phylogenetic history of the family. Methods focused on detecting pair-specific co-evolution are also independent of the phylogenetic history. Methods are coloured according to the categories discussed in the main text. Methodologies presenting a higher predictive power are represented with more intense colour gradients. In this figure, it is evident that methods based on inter-residue and inter-protein co-evolution tend to present higher predictive powers when they become less dependent on the phylogenetic history and become more pair-specific, whereas statistical coupling analysis (SCA)-like and specificity-determining positions (SDPs) that are strongly associated to the phylogenetic history are more informative for functional sites and networks of residues. CAPS, co-evolution analysis using protein sequences; CEO, combinatorial entropy optimization; DCA, direct coupling analysis; ET, evolutionary trace; i2h, *in silico* two-hybrid; McBASC, McLachlan-based substitution correlation; MB, mutational behaviour; MMM, MatrixMatchMaker.

every position and different groups of orthologues[57] or tree-based partitions[58] (for SDPsite). In the case of rvET, it provides a score that reflects both the global level of conservation and the coherence of this conservation (or variation) with the phylogenetic tree. In this way, it provides a gradual transition from complete conservation to specific SDPs. These methods detect clusters of residues that are structural or functionally important[56], and they have been used in combination with structural information, such as solvent accessibility and spatial proximity of the inferred SDPs (for example, see REF. 58), to improve the detection of functional sites.

Evolutionary-trace-related methods consider only identical amino acids as being conserved and do not take into account amino acid similarities. This categorical approach might miss more subtle conservation patterns that involve similar amino acids. However, the evolutionary trace strategy avoids being restricted to a single subfamily partition, and it provides a good estimate of the evolutionary relevance of positions. A potential difficulty is that it requires manual analysis of the positions detected, which may be associated with very different activities, such as ligand binding[19], allosteric regulation[49] or structural roles[56].

***Methods based on the detection of positions that are representative of the global variability.*** The premise of these methods is that groups of positions that better reflect the evolution of the family are more likely to be responsible for the functional specificity of the protein. The variation among these positions would simultaneously tend to be correlated. The mutational behaviour method[47] compares the pattern of mutations of every position in the MSA with the variation of the complete sequences. Importantly, both position and sequence divergences are calculated on the basis of amino acid substitution matrices, and thus this method accounts for the biochemical relevance of the amino acid change. Accordingly, mutational behaviour tends to detect clear shifts in amino acid properties and their influence on sequence divergences. This methodology is particularly useful to detect SDPs that are correlated with the divergence of the protein families, and these tend to correspond to ligand- or protein-binding interfaces[47]. This type of approach has also been used in combination with structural information to improve the detection of functional sites in known protein structures[59] as well as to improve the selection of docking models[12]. Like evolutionary trace, this methodology does not require protein subfamilies to have been detected previously, thus facilitating the implementation but complicating the interpretation of the subfamily–SDP relationship. An additional drawback of the method is that it does not explicitly remove uninformative signals (such as redundant sequences) or outliers (such as sequences introduced by horizontal gene transfer).

***Methods based on multivariate analyses.*** One of the first methods that aimed to detect SDPs was SequenceSpace[18]. This method was based on the idea of detecting the main sources of variability in an MSA and the positions responsible for this, as achieved by carrying out a principal component analysis (PCA) of the MSA. This analysis projects the sequences onto simplified multidimensional spaces in which the subfamilies that are equivalent to main branches in the tree are represented as clusters of proteins. The key issue is that this approach makes it possible to represent the residue–position associations that are characteristic of each subfamily in an equivalent multidimensional space, making the detection of SDPs straightforward. The elegant approach implemented in SequenceSpace has been useful to guide experimental studies (BOX 2). However, it requires interactive manual inspection of the PCA spaces by experts to identify the protein subfamilies and their corresponding SDPs.

More recently, a fully automated method has been developed, called S3det, that is based on a related mathematical approach called multiple correspondence analysis[48] (MCA), which is conceptually equivalent to PCA but is better suited to dealing with categorical data, such as amino acid identities. S3det automatically selects the most informative dimensions, carrying out a robust

Horizontal gene transfer (HGT). Transmission of genetic material between organisms different from that which occurs between the parents and the offspring ('vertical transfer'). Also known as 'lateral gene transfer'.

clustering analysis of the sequence space and detecting those residues associated with the subfamilies obtained in the equivalent position space. SDPs are detected as those that are differentially conserved among any possible combination of subfamilies. S3det has been tested on a large set of protein families and is robust in detecting protein subfamilies associated with distinct enzymatic or protein binding specificities. Moreover, SDPs identified by S3det can represent ligand- and protein-binding sites[48]. However, as is the case for the evolutionary trace, the main limitation of S3det is that it does not consider similarities between amino acids, such that it might miss SDPs with subtle conservation patterns. When compared with other methods[48], such as rvET[56], mutational behaviour[47], combinatorial entropy optimization (CEO; see below)[60] and an automated version of SequenceSpace[48], S3det provides better information about subfamily classification, whereas rvET and S3det are the most effective tools for detecting binding sites that functionally distinguish the protein subfamilies.

***Other approaches to detect SDPs.*** Other methods to detect SDPs do not use phylogenetic trees for detecting subfamilies. For example, kPax is based on a Bayesian model that allows the simultaneous detection of subfamilies and SDPs through the optimization of the subset of residues that are considered informative for such purpose[61]. In this approach, only this subset of residues is used to calculate sequence clusters. In an alternative approach, CEO carries out a combinatorial exploration of the possible subfamily partitions from which an optimal split is selected, and SDPs are assigned on the basis of residue entropies[60]. These conceptually interesting methods have not been sufficiently evaluated in large-scale comparisons, although a recent comparison that included CEO suggests that both rvET and S3det more reliably identify ligand- and protein-binding specificities[48].

***Detection of residue co-evolution through statistical coupling analysis strategies.*** Statistical coupling analysis (SCA) methods are challenging to classify and can be considered as a combination of approaches based on residue co-variation and SDPs. SCA is intended to detect positions with similar patterns of amino acids, but it usually focuses on functionally associated patterns that actually define groups of co-evolving residues. The first implementation of SCA[51] (here called SCAold) proposed that a change in the amino acid frequency of one position produces a statistical perturbation in the amino acid frequency of evolutionarily coupled positions. In practice, this method detects, for a given amino acid in a site of reference, other positions that are conserved in the sequences containing that particular amino acid in the reference site. This reference site was defined manually to focus on functionally relevant amino acids.

SCAold has been widely reworked to the point that it is no longer appropriate to consider its latest versions as mere follow-ups of the original approach; rather, they represent a very different method based on a similar rationale. The most recent version of SCA (here called SCAnew)[54] is, like S3det[48], based on a multivariate analysis. In this case, between-position correlations are weighted according to their conservation, and the weighted correlation matrix is reduced by spectral decomposition. The 'principal components' obtained are expected to indicate subfamily-associated conservations. Thus, groups of co-varying positions are detected as those that significantly contribute to these components. The groups of positions identified by SCAnew were called 'protein sectors'[62], and in several cases they have been shown to form structurally independent clusters of amino acids with distinct roles. In practice, these positions tend to be specifically conserved within protein subfamilies but are not necessarily differentially conserved in the different subfamilies (unlike the SDPs).

SCA-like approaches have been used in interesting small-scale studies to identify networks of co-evolving residues implicated in protein folding[50] and allosteric communication[53] (BOX 2). However, SCAnew has not been evaluated on a large scale, and comparisons based on SCAold and some of its subsequent developments have shown that this approach is not particularly competitive for predicting protein contacts[24,63]. A lack of benchmarking standards is a major limitation for the general application of SCA-based approaches.

## Co-evolution at the protein level

Potential co-evolution between functionally related protein families was initially observed in sporadic cases. For example, remarkable similarity was detected between the phylogenetic trees of ligands (such as insulins and interleukins) and their receptors; this co-evolution was proposed to be required for the maintenance of their specific interactions[22]. These initial observations of protein co-evolution remained anecdotal until the genomics revolution prompted the development of methods to infer co-evolution automatically between proteins using MSAs. Indeed, co-evolution-based approaches can be regarded as part of the methods to detect interactions and functional relationships between proteins using genomic information, jointly known as 'context-based' methods[64,65]. These methods provide an orthogonal alternative to the more traditional prediction of function based on homology[66] (BOX 3).

***Family tree similarities.*** The first methods to quantify tree similarities implemented a simple linear correlation between the distance matrices of the two protein families, as a proxy of their phylogenetic trees[20,21]. This approach, which is called MirrorTree, made it possible to evaluate the relationship between tree similarities and physical or functional interaction and to predict potential protein–protein interactions on a genomic scale[21]. For the two protein families for which co-evolution is to be evaluated, MSAs are generated using orthologues from a set of reference genomes. MirrorTree then extracts inter-orthologue distance matrices from the MSA-derived trees or from the MSAs themselves (for example, as percentages of identities). Finally, the similarity between these distance matrices, as a proxy of the similarity of the corresponding trees, is evaluated with a linear correlation criterion.

Principal component analysis
(PCA). Multivariate data analysis technique that consists of calculating a lower dimensionality space in which the axes explain most of the variability of the original data. The rationale is that such lower dimensionality space is easy to handle and to visualize, whereas most of the information of the original data (for example, in terms of relative distances) is retained and some contributions of noise are removed.

Multiple correspondence analysis
(MCA). Multivariate data analysis technique similar to principal component analysis but more suitable for categorical data.

Spectral decomposition
Decomposition of a squared matrix ($A$) as the product of its eigenvectors ($V$) times the diagonal matrix of its eigenvalues ($D$) times the inverse of its eigenvectors: $A = V \cdot D \cdot V^{-1}$. Also known as 'eigendecomposition'.

One difficulty of approaches based on family tree similarities is that the phylogenetic trees of all protein families retain a degree of similarity to the archetypal 'tree of life', which represents the global evolutionary relationships between organisms. Owing to this shared phylogenetic background, any pair of family trees has a 'basal' level of similarity, regardless of whether the corresponding proteins are functionally interacting. Some approaches to correct for this organismal speciation use the evolutionary distances between the corresponding species to normalize each protein–protein distance[67–69]. Another alternative is to infer the background similarity not from the general species tree but from the main tendencies of actual data from many protein families[70,71]. Similar challenges are faced by residue-level tools that analyse MSAs, and methods such as CAPS and mutual information corrected (see above) also apply corrections for the phylogenetic background.

An important improvement in the co-evolution-based detection of protein–protein interactions came from the assembly of genome-wide co-evolutionary networks from the pairwise co-evolution of individual protein pairs[71]. This approach, which is called ContextMirror, increases the accuracy of interaction predictions by correcting numerous problems related to the global tendencies of the data, including the background tree similarity due to speciation. Additionally, it is able to separate co-evolution that is specific to a given pair of proteins from general co-evolutionary trends that involve many families concomitantly, as it evaluates pair-specific protein co-evolution by considering the contribution of every other protein tree. This disentanglement of direct and indirect correlations is closely related to the DCA-like methodologies of residue co-evolution, which were independently developed (see above).

The application of MirrorTree-related methods benefits from a careful selection of the species used to construct the trees. Different criteria for the selection of species have been shown to influence the results, probably because of both taxonomy-specific biases and differences in the age of the co-evolutionary relationship are associated with different types of interactions[72]. To circumvent these problems, some methods automatically look for subsets of species in which co-evolution is particularly strong[73,74].

Other MirrorTree-derived approaches predict specific pairs of interacting proteins within two large protein families that include paralogues. These methods are distinct from the previously described applications of family-tree similarity that predict a general interaction between protein families; instead, the inter-family interaction is already known, and these methods analyse the correspondence between different members of the two families to predict which paralogues within one family interact with those in the other. The basic assumption is that the right 'mapping' (set of links) between the two families will render the strongest correlation of the corresponding trees[75–77].

In addition to the use of whole-protein sequences, phylogenetic trees derived from particular protein domains can be used to identify the interacting domains of two interacting proteins on the basis that these domains exhibit stronger co-evolution signals than other non-interacting domains within the same proteins[78]. Similarly, trees restricted to the interfaces showed that these more intensely co-evolve than the rest of the protein[79]. Both of these observations indicate that in many cases the co-evolution between interacting proteins is a local phenomenon that can be circumscribed to certain regions or even particular residues. This would be in line with the use of co-evolution signatures to identify interacting residues within proteins, as discussed earlier, and points to the possibility of using those co-evolutionary signatures at the residue level to look for protein–protein co-evolution (see 'Hybrid residue–protein methods' below).

***Similarity of phylogenetic profiles.*** The tendency of proteins that carry out common functions to be present or absent from the same organisms can be considered to be an extreme case of co-evolution, and this phenomenon is exploited by some methods to predict interacting or functionally related families of proteins[80]. The pattern of presence or absence of the orthologues of a protein across a set of genomes is known as a 'phylogenetic profile'[81], which is most simply represented as a binary vector denoting the presence or absence of an orthologue in each genome[80]. However, more recent versions replaced this binary representation by quantifying the sequence similarity of the orthologues[82] or by assessing the number of copies (paralogues) of the gene in each genome[83].

As for the MirrorTree methods, the underlying speciation process introduces an important bias, in this case imposing strong constraints on the set of genes that are present in the genomes. For example, many proteins will simultaneously be present in bacterial and archaeal sequences but absent in eukaryotic sequences without implying that all of them develop a common function. To tackle this problem, some variants of this approach incorporate more sophisticated evolutionary models to weight the presence or absence of genes, depending on the species involved[84,85]. Moreover, the selection of the organisms for the construction of the profiles again has a drastic effect on the performance of these methods[86] and on the type of function that can be predicted[87]. As in the case of tree-similarity-based methods, different sets of genomes should be used to build the phylogenetic profiles, depending on the type of interactions or functional relationships that are being predicted.

## Hybrid residue–protein methods

So far, we have separated co-evolution-based methods into two classes: those that use the MSA of a single protein family to detect (intra-protein) residue co-evolution and those that use MSAs of two families to detect inter-protein co-evolution, including the search for the interaction region of proteins known to interact and the search in complete genomes for the interaction partners of a given protein. However, intra- and inter-protein co-evolution are two related phenomena that involve similar physical interaction forces and evolutionary constraints, thus the global co-evolution observed

Paralogues
Homologous genes or proteins split in a gene duplication event, resulting in two copies of the parental gene in the same organism that latter diverge in sequence and function.

Protein domains
Pieces of a protein defined according to given criteria: for example, structural domains or functional domains.

between two proteins could be (at least partially) explained by a set of co-evolving inter-protein pairs of residues. Indeed, various methods for detecting residue co-evolution have also been used to detect residues at the interfaces of inter-protein interactions[11–17]. Therefore, although historically separated, inter- and intra-protein co-evolution-based methods and their applications are closely related and share conceptual and methodological similarities, such as the idea of disentangling direct from indirect co-evolution.

Furthermore, these two classes of methods can be combined: for example, by methods that infer protein co-evolution on the basis of the accumulation of inter-protein residue co-evolutions. Of note, the *in silico* two-hybrid (i2h) system[88] uses the balance between inter- and intra-protein residue correlations to estimate the interaction potential of two protein families. This idea has been developed into more sophisticated methods, such as one developed by Burger and van Nimwegen[14], which more effectively deal with paralogous and orthologous relationships to improve the results. Similarly, a method developed by Yeang and Haussler[13] simultaneously detects interacting proteins with co-evolving residues that participate in their interfaces. The accumulation of correlated mutations has also been used to locate the best mapping between the members of two interacting families, taking the set of links that maximizes the number of inter-protein correlations[89].

## Discussion

We have reviewed the main classes of co-evolution-based methods and have highlighted those that have shaped the field, notwithstanding the many other implementations that have been important in developing what is still an open area. This overview highlights how co-evolutionary signals are influenced by the available phylogenetic information associated with the protein families and how the different co-evolution-based methods deal with this problem (FIG. 2). Whereas for the methods that study pairwise co-evolutions (both at the residue and protein levels) it is essential to correct the phylogenetic background to identify the specific signals related with direct protein or residue interactions, the approaches aimed at detecting groups of co-evolving residues (for example, SDPs) use this phylogenetic information to detect groups of residues that are concomitantly involved in a concerted role. These different strategies provide alternative views of the role of co-evolution in defining the actual evolutionary landscape of intra- and inter-protein interactions. Co-evolution contributes to the functional and evolutionary divergence by permitting the maintenance of interactions while simultaneously allowing the variation of the interacting partners. For example, this can avoid undesired crosstalk with recently appeared signalling pathways, or it can maintain the interaction with a rapidly evolving protein.

There are scientific and technical limitations that must be overcome by the methods. The quality of MSAs is obviously essential as they serve as the initial input to most of the methods. Furthermore, the methods work better on large protein families for which the degree of sequence similarity has a wide but homogenously distributed range from distant to similar sequences. In particular, some of the more recent approaches require densely populated alignments with thousands of sequences. In general, optimal performance is obtained when protein subfamilies (branches of the tree) are spaced at regular intervals, whereas most approaches (particularly SDP-related methods) tend to fail in the extreme case of protein families with a star-like family tree in which all sequences are equidistant. Furthermore, predicting protein interactions has additional practical constraints in that the MSAs of the two potential interactors have to include orthologous sequences coming from the same set of species (see Box 4 in Supplementary information S1 (boxes)). Together with the alignments, another basic requirement for many of the methods is the adequate treatment of phylogenetic information. For example, assembling phylogenetic trees is confounded by complex evolutionary scenarios, such as sequences acquired by horizontal gene transfer, genetic saturation or the difficulties in identifying the correct orthologous sequences when genome duplication and domain rearrangements have occurred. Another clear limitation to the progress in this field is the need to compare methods systematically, particularly for large-scale protein interaction predictions. Blind tests with hidden gold standards will be the best way to facilitate the appropriate use of the methods.

An additional difficulty in the field is the confusion in the terminology used for describing what is observable in an MSA versus the underlying evolutionary phenomenon. We have proposed to differentiate between the well-documented and clearly observable phenomenon of 'co-evolution' and the more elusive concept of 'co-adaptation' between co-evolving components[44,45]. In this definition, co-evolution would be confined to the observation of concerted patterns of co-variation derived from MSAs without implying reciprocal evolutionary events. By contrast, co-adaptation implies evolutionary reciprocity as the causal phenomenon behind the observed co-evolution. In this sense, co-adaptation would be referring to, for example, the compensatory changes required for maintaining residue or protein interactions. A more detailed discussion of the contribution of co-adaptation to the observed patterns of co-evolution can be found elsewhere[44,45].

Another fundamental problem in the field is to differentiate the direct relationships from the indirect couplings caused by them, a problem tackled by a number of approaches[38,39,41,71]. This disentanglement helps the functional interpretation of the relationships: direct couplings suggest physical interactions between residues or proteins, whereas pairings regarded as indirect couplings could indicate clusters of interacting residues, chains of residues participating in allosteric transmission or (at the protein level) signalling pathways, protein complexes and functional clusters. The methodological advances have set the scene for the characterization of co-adaptive relationships that take place in complex molecular systems and that might only be truly understood at a network level, such as networks of residues participating in

Genetic saturation
Apparent reduction with time of the observed divergence between two genes owing to factors such as reversed or convergent mutations.

intramolecular communication[54], or proteins intimately working together in complexes or functional pathways[71]. It will also be important to consider the dynamic aspects of these interactions. This situation is reminiscent of ecological networks in which the interaction between species occurs in environmental conditions that change in space and time and in which co-evolution will be a prevalent phenomenon[90]. It seems that in the future, the field of molecular co-evolution might again have to look for inspiration in the study of species co-evolution.


1. Dobzhansky, T. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. *Genetics* **35**, 288–302 (1950).
2. Wallace, B. On coadaptation in *Drosophila*. *Am. Nat.* **87**, 343–358 (1953).
3. Ehrlich, P. & Raven, P. Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608 (1964).
4. Thompson, J. N. *The Coevolutionary Process* (Univ. Chicago Press, 1994).
5. Burton, R. & Rawson, P. Genetic architecture of physiological phenotypes: empirical evidence for coadapted gene complexes. *Amer. Zool.* **39**, 451–462 (1999).
6. Fitch, W. M. & 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.* **4**, 579–593 (1970).
7. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317 (1994).
**This paper describes one of the first automatic approaches for extracting correlated patterns of amino acid replacements between positions of MSAs with the goal of predicting residues close in three-dimensional structures.**
8. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358 (1994).
9. Taylor, W. R. & Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348 (1994).
10. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA* **91**, 98–102 (1994).
11. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511–523 (1997).
12. Tress, M. *et al.* Scoring docking models with evolutionary information. *Proteins* **60**, 275–280 (2005).
13. Yeang, C.-H. & Haussler, D. Detecting coevolution in and among protein domains. *PLoS Comp. Biol.* **3**, e211 (2007).
14. Burger, L. & van Nimwegen, E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* **4**, 165 (2008).
**Here, the authors present a parameter-free Bayesian method for predicting interaction partners from MSAs (eventually including paralogues) based on co-evolution between multiple positions of potential interacting partners.**
15. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
16. Schug, A., Weigt, M., Onuchic, J. N., Hwa, T. & Szurmant, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl Acad. Sci. USA* **106**, 22124–22129 (2009).
17. Dago, A. E. *et al.* Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl Acad. Sci. USA* **109**, E1733–E1742 (2012).
18. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178 (1995).
**This is one of the original approaches detecting SDPs in MSAs. It is the basis for a family of methodologies that use PCA-related vectorial representations of the alignments to detect amino acid patterns associated with the corresponding protein subfamilies.**
19. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
**This is the initial proposal of the evolutionary trace methodology. A simple analysis of differential sequence conservation at different levels of the family phylogenetic tree is used to locate protein-binding surfaces.**
20. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293 (2000).
21. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614 (2001).
**This is the initial publication of the 'MirrorTree' approach for the quantification of similarities of phylogenetic trees (represented by their distance matrices) to predict potential protein interactions.**
22. Fryxell, K. J. The coevolution of gene family trees. *Trends Genet.* **12**, 364–369 (1996).
23. Korber, B. T., Farber, R. M., Wolpert, D. H. & Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl Acad. Sci. USA* **90**, 7176–7180 (1993).
**This is one of the initial publications in the field of protein co-evolution. In this work, a mutual information method is used to detect co-evolving positions in a particular biological case.**
24. Fodor, A. A. & Aldrich, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221 (2004).
25. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).
26. Olmea, O. & Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25–S32 (1997).
27. Fares, M. A. & Travers, S. A. A. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* **173**, 9–23 (2006).
28. Tillier, E. R. M. & Lui, T. W. H. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **19**, 750–755 (2003).
29. Martin, L. C., Gloor, G. B., Dunn, S. D. & Wahl, L. M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116–4124 (2005).
30. Fairman, J. W. *et al.* Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic *E. coli* and enteropathogenic *Y. pseudotuberculosis*. *Structure* **20**, 1233–1243 (2012).
31. Altschuh, D., Lesk, A. M., Bloomer, A. C. & 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).
32. Oliveira, L., Paiva, A. C. M. & Vriend, G. Correlated mutation analyses on very large sequence families. *Chembiochem* **3**, 1010–1017 (2002).
33. Fleishman, S. J., Yifrach, O. & Ben-Tal, N. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J. Mol. Biol.* **340**, 307–318 (2004).
34. Dutheil, J., Pupko, T., Jean-Marie, A. & Galtier, N. A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.* **22**, 1919–1928 (2005).
35. Pollock, D. D., Taylor, W. R. & Goldman, N. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287**, 187–198 (1999).
36. Barker, D. & Pagel, M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comp. Biol.* **1**, e3 (2005).
37. Lapedes, A. S., Giraud, B. G., Liu, L. C. & Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Stat. Mol. Biol. Genet.* **33**, 236–256 (1999).
38. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
**This is an efficient methodology based on reference 15 to extract direct couplings between positions in MSAs that can obtain accurate predictions of physical contacts for many very large MSAs.**
39. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
**This article presents an innovative methodology using sparse inverse covariance estimation techniques to remove indirect couplings between residues in very large MSAs.**
40. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
41. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comp. Biol.* **6**, e1000633 (2010).
42. Sreekumar, J., Braak, ter, C. J. F., van Ham, R. C. H. J. & van Dijk, A. D. J. Correlated mutations via regularized multinomial regression. *BMC Bioinformatics* **12**, 444 (2011).
43. Di Lena, P., Nagata, K. & Baldi, P. Deep architectures for protein contact map prediction. *Bioinformatics* **28**, 2449–2457 (2012).
44. Juan, D., Pazos, F. & Valencia, A. Co-evolution and co-adaptation in protein networks. *FEBS Lett.* **582**, 1225–1230 (2008).
45. Pazos, F. & Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J.* **27**, 2648–2655 (2008).
46. Madabushi, S. *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154 (2002).
47. del Sol Mesa, A., Pazos, F. & Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302 (2003).
48. Rausell, A., Juan, D., Pazos, F. & Valencia, A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl Acad. Sci. USA* **107**, 1995–2000 (2010).
**This is a recent methodology for the automatic detection of subfamilies and SDPs in MSAs. The application of this method to a large set of protein families demonstrates the relation between SDPs and regions of functional importance for binding to specific interactors and substrates.**
49. Rodriguez, G. J., Yao, R., Lichtarge, O. & Wensel, T. G. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Natl Acad. Sci. USA* **107**, 7787–7792 (2010).
50. Socolich, M. *et al.* Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
51. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
52. Kass, I. & Horovitz, A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* **48**, 611–617 (2002).
53. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10**, 59–69 (2003).

54. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).
**This work demonstrates that mutations at surface residues predicted by SCAnew (a method based on reference 51) modify the activity of the active site of selected proteins by altering the chain of allosteric interactions.**
55. Hannenhalli, S. S. & Russell, R. B. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76 (2000).
56. Mihalek, I., Res, I. & Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265–1282 (2004).
**An improved version of the evolutionary trace methodology (reference 19) that incorporates an entropy-based quantification of the conservation of each position in a MSA for the different partitions of the corresponding family phylogenetic tree.**
57. Mirny, L. A. & Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **321**, 7–20 (2002).
58. Kalinina, O. V., Gelfand, M. S. & Russell, R. B. Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics* **10**, 174 (2009).
59. Landgraf, R., Xenarios, I. & Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502 (2001).
60. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8**, R232 (2007).
61. Marttinen, P., Corander, J., Törönen, P. & Holm, L. Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics* **22**, 2466–2474 (2006).
62. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
**This work shows that SCAnew (reference 54) can detect 'protein sectors' (that is, pseudo-independent groups of correlated positions of the MSA) that are related to the structural and functional organization of proteins in a selected number of examples.**
63. Brown, C. A. & Brown, K. S. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS ONE* **5**, e10779 (2010).
64. Harrington, E. D., Jensen, L. J. & Bork, P. Predicting biological networks from genomic data. *FEBS Lett.* **582**, 1251–1258 (2008).
65. Wass, M. N., David, A. & Sternberg, M. J. Challenges for the prediction of macromolecular interactions. *Curr. Opin. Struct. Biol.* **21**, 382–390 (2011).
66. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
67. Pazos, F., Ranea, J. A. G., Juan, D. & Sternberg, M. J. E. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–1015 (2005).
68. Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**, 3482–3489 (2005).
69. Kann, M. G., Jothi, R., Cherukuri, P. F. & Przytycka, T. M. Predicting protein domain interactions from coevolution of conserved regions. *Proteins* **67**, 811–820 (2007).
70. Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M. & Toh, H. Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. *Bioinformatics* **22**, 2488–2492 (2006).
71. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl Acad. Sci. USA* **105**, 934–939 (2008).
**This methodology relies on the whole set of pairwise similarities between phylogenetic trees within a given proteome (co-evolutionary network) to reassess the co-evolutionary signal of every pair of proteins. The method predicts interactions at the level of macromolecular complexes and functional units for fully sequenced genomes.**
72. Herman, D. *et al.* Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics* **12**, 363 (2011).
73. Choi, K. & Gomez, S. M. Comparison of phylogenetic trees through alignment of embedded evolutionary distances. *BMC Bioinformatics* **10**, 423 (2009).
74. Tillier, E. R. M. & Charlebois, R. L. The human protein coevolution network. *Genome Res.* **19**, 1861–1871 (2009).
75. Ramani, A. K. & Marcotte, E. M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284 (2003).
76. Jothi, R., Kann, M. G. & Przytycka, T. M. Predicting protein–protein interaction by searching evolutionary tree automorphism space. *Bioinformatics* **21** (Suppl. 1), i241–i250 (2005).
77. Izarzugaza, J. M., Juan, D., Pons, C., Pazos, F. & Valencia, A. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics* **9**, 35 (2008).
78. Jothi, R., Cherukuri, P. F., Tasneem, A. & Przytycka, T. M. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J. Mol. Biol.* **362**, 861–875 (2006).
79. Kann, M. G., Shoemaker, B. A., Panchenko, A. R. & Przytycka, T. M. Correlated evolution of interacting proteins: looking behind the MirrorTree. *J. Mol. Biol.* **385**, 91–98 (2009).
80. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
81. Gaasterland, T. & Ragan, M. A. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genom.* **3**, 199–217 (1998).
82. Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotech.* **21**, 1055–1062 (2003).
83. Ranea, J. A. G., Yeats, C., Grant, A. & Orengo, C. A. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comp. Biol.* **3**, e237 (2007).
84. Zhou, Y., Wang, R., Li, L., Xia, X. & Sun, Z. Inferring functional linkages between proteins from evolutionary scenarios. *J. Mol. Biol.* **359**, 1150–1159 (2006).
85. Ta, H. X., Koskinen, P. & Holm, L. A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics* **27**, 700–706 (2011).
86. Sun, J. *et al.* Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics* **21**, 3409–3415 (2005).
87. Jothi, R., Przytycka, T. M. & Aravind, L. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics* **8**, 173 (2007).
88. Pazos, F. & Valencia, A. *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227 (2002).
89. Tillier, E. R. M., Biro, L., Li, G. & Tillo, D. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins* **63**, 822–831 (2006).
90. Thompson, J. N. The coevolving web of life. *Am. Nat.* **173**, 125–140 (2009).
91. Graña, O. *et al.* CASP6 assessment of contact prediction. *Proteins* **61** (Suppl. 7), 214–224 (2005).
92. Tress, M. L. & Valencia, A. Predicted residue–residue contacts can help the scoring of 3D models. *Proteins* **78**, 1980–1991 (2010).
93. Sadowski, M. I., Maksimiak, K. & Taylor, W. R. Direct correlation analysis improves fold recognition. *Comput. Biol. Chem.* **35**, 323–332 (2011).
94. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
95. Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
**This publication presents a new methodology for obtaining high quality *de novo* models of transmembrane proteins by integrating DCA (reference 38) predictions with various topological constraints.**
96. Nugent, T. & Jones, D. T. Accurate *de novo* structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA* **109**, E1540–E1547 (2012).
97. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA* **109**, 10340–10345 (2012).
98. Izarzugaza, J. M. G. *et al.* Characterization of pathogenic germline mutations in human protein kinases. *BMC Bioinformatics* **12** (Suppl. 4), S1 (2011).
99. Izarzugaza, J. M. G., del Pozo, A., Vazquez, M. & Valencia, A. Prioritization of pathogenic mutations in the protein kinase superfamily. *BMC Genomics* **13** (Suppl. 4), S3 (2012).
100. Bauer, B. *et al.* Effector recognition by the small GTP-binding proteins Ras and Ral. *J. Biol. Chem.* **274**, 17763–17770 (1999).
101. Morillas, M. *et al.* Identification of conserved amino acid residues in rat liver carnitine palmitoyltransferase I critical for malonyl-CoA inhibition. Mutation of methionine 593 abolishes malonyl-CoA inhibition. *J. Biol. Chem.* **278**, 9058–9063 (2003).
102. Hernanz-Falcón, P. *et al.* Identification of amino acid residues crucial for chemokine receptor dimerization. *Nature Immunol.* **5**, 216–223 (2004).
103. Shenoy, S. K. *et al.* β-arrestin-dependent, G protein-independent ERK1/2 activation by the β2 adrenergic receptor. *J. Biol. Chem.* **281**, 1261–1273 (2006).
104. Ribes-Zamora, A., Mihalek, I., Lichtarge, O. & Bertuch, A. A. Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. *Nature Struct. Mol. Biol.* **14**, 301–307 (2007).
105. Zamir, L. *et al.* Tight coevolution of proliferating cell nuclear antigen (PCNA)-partner interaction networks in fungi leads to interspecies network incompatibility. *Proc. Natl Acad. Sci. USA* **109**, E406–E414 (2012).
106. Capra, E. J., Perchuk, B. S., Skerker, J. M. & Laub, M. T. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222–232 (2012).
107. Gershoni, M. *et al.* Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex i. *J. Mol. Biol.* **404**, 158–171 (2010).
108. Ochoa, D. & Pazos, F. Studying the co-evolution of protein families with the MirrorTree web server. *Bioinformatics* **26**, 1370–1371 (2010).
109. Edgar, R. S. *et al.* Peroxiredoxins are conserved markers of circadian rhythms. *Nature* **485**, 459–464 (2012).
110. Watanabe, M. *et al.* Highly divergent sequences of the pollen self-incompatibility (*S*) gene in class-I S haplotypes of *Brassica campestris* (syn. *rapa*) L. *FEBS Lett.* **473**, 139–144 (2000).
111. Clark, N. L. *et al.* Coevolution of interacting fertilization proteins. *PLoS Genet.* **5**, e1000570 (2009).



### Acknowledgements
We thank J. Onuchic from the University of California, San Diego, USA, D. Jones from the University College London, UK, C. Sander from the Computational Biology Center at the Memorial Sloan–Kettering Cancer Center, New York, USA, E. van Nimwegen from Biozentrum at the University of Basel, Switzerland, F. Gervasio and S. Marsili from the Computational Biophysics Group at CNIO, A. Rausell from the Swiss Institute of Bioinformatics Vital-IT & Institute of Microbiology of the University of Lausanne and D. Ochoa from the Computational Systems Biology Group at CNB–CSIC for interesting discussions, as well as the many authors and collaborators with important contributions to the field of molecular co-evolution in the past 20 years, many of which could not be included in this Review.


### Competing interests statement
The authors declare no competing financial interests.

### FURTHER INFORMATION
**Florencio Pazos's homepage:** http://csbg.cnb.csic.es
**Alfonso Valencia's homepage:** http://www.cnio.es/ing/grupos/plantillas/presentacion.asp?grupo=50004294
**Home — CASP10:** http://www.predictioncenter.org/casp10/index.cgi
**Home — Prediction Center:** http://predictioncenter.org

### SUPPLEMENTARY INFORMATION
**See online article:** S1 (boxes) | S2 (table)
**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**