Computational challenges in genome wide association studies: data processing, variant annotation and epistasis

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

CHAPTER 1
Introduction

How does one's DNA influence their risk of getting a disease? Contrary to popular belief, your future health is not "hard wired" in your DNA. Only in a few diseases, referred as "Mendelian diseases", are there well known, almost certain, links between genetic mutations and disease susceptibility. For the majority of what are known as "complex traits", such as cancer or diabetes, genomic predisposition is subtle and, so far, not fully understood.

With the rapid decrease in the cost of DNA sequencing, the complete genome sequence of large cohorts of individuals can now be routinely obtained. This wealth of sequencing information is expected to ease the identification of genetic variations linked to complex traits. In this work, I investigate the analysis of genomic data in relation to complex diseases, which offers a number of important computational and statistical challenges. We tackle several steps necessary for the analysis of sequencing data and the identification of links to disease. Each step, which corresponds to a chapter in my thesis, is characterized by very different problems that need to be addressed.

i) The first step is to analyze large amounts of information generated by DNA sequencers to obtain a set of "genomic variants" present n each each individual. To address these big data processing problems, Chapter **??** shows how we designed a programming language (BigDataScript [4]), that simplifies the creation robust, scalable data pipelines.

ii) Once genomic variants are obtained, we need to prioritize and filter them to discern which variants should be considered "important" and which ones are likely to be less relevant. We created the SnpEff & SnpSift

[2, 3] packages that, using optimized algorithms, solve several annotation problems: a) standardizing the annotation process, b) calculating putative genetic effects, c) estimating genetic impact, d) adding several sources of genetic information, and e) facilitating variant filtering.

iii) Finally, we address the problem of finding associations between interacting genetic loci and disease. One of the main problems in GWAS, known as "missing heritability", is that most of the phenotypic variance attributed to genetic causes remains unexplained. Since interacting genetic loci (epistasis) have been pointed out as one of the possible causes of missing heritability, finding links between such interactions and disease has great significance in the field. We propose a methodology to increase the statistical power of this type of approaches by combining population-level genetic information with evolutionary information.

In a nutshell, this thesis addresses computational, analytical, algorithmic and methodological problems of transforming raw sequencing data into biological insight in the aetiology of complex disease. In the rest of this introduction we give the background that provides motivation for our research.

1.1 Coevolution

1.1.1 Definition

Distinct combinations of alleles in coevolving genes interact differently, conferring varying degrees of fitness. If this fitness differential is adequately large, the resulting selection for allele matching could maintain allelic association, even between physically unlinked loci. [21] Coevolving genes are expected to undergo compensatory mutations to maintain their interaction. [21] Most cases of selective advantage for specific allele pairing would be resolved with fixation of the optimal allele pair.7 [21]

COEVOLUTION FIRST PAPER: Coevolution of any type has its origin in the covarion hypothesis proposed first by Fitch and Markowitz (1970). This hypothesis states that, at any given time, some sites are invariable due to their functional or structural constraints but, as mutations are fixed elsewhere in the sequence, these constraints may change. [9]

The reason is that, while interaction would necessarily involve coevolution, coevolution does not imply physical interaction. [9] The identification of genes showing particular amino acid residues that have undergone adaptive evolution is key in determining functionally or structurally important protein regions. [9]

Many proteins have evolved to form specific molecular complexes and the specificity of this interaction is essential for their function. [18] The network of the necessary inter-residue contacts must consequently constrain the protein sequences to some extent. [18] In other words, the sequence of an interacting protein must react the consequence of this process of adaptation. It is reasonable to assume that the sequence changes accumulated during the evolution of one of the interacting proteins must be compensated by changes in the other. [18]

Predicting protein structure from primary sequence is one of the ultimate challenges in computational biology. [1] Given the large amount of available sequence data, the analysis of co-evolution, i.e., statistical dependency, between columns in multiple alignments of protein domain sequences remains one of the most promising avenues for predicting residues that are contacting in the structure. [1] [1]

It has long been suggested that the resulting correlations among amino acid compositions at different sequence positions can be exploited to infer spatial contacts within the tertiary protein structure. [16]

In its simplest definition, coevolution refers to the coordinated changes that occur in pairs of organisms or biomolecules, typically to main tain or to refine functional interactions between those pairs. [6] Darwin himself initiated the study of coevolution, and his observation on the relationship between the size of orchids' corollae and the length of the proboscis of pol linators led him to predict successfully the existence of a new species that was able to suck from the large spur of Darwin's orchid. [6] The studies of Dobzhansky1 and others2 contributed to the establishment of this concept in genetic terms, although the term coevolution is usually attrib uted to Ehrlich3, and it is commonly defined as 'reciprocal evolutionary change in interacting species'4. [6] For the past 20 years, much effort has been dedicated to investigating coevolution 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 success5, thus indicating that molecular coevolution could be an important and widespread determinant of fitness. [6] Although coevolution can potentially occur between various biomolecules, most recent tools focus on protein coevolution. [6] 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 model6, and they typically use the multiple sequence alignment (MSA) for a protein family of homologues to search for correlated mutations. [6] Such correlated muta tions are suggestive of compensatory changes that occur between entangled residues (for example, those in prox imity, direct contact or acting together in catalytic or binding sites) to maintain protein stability, function or folding7-10. Furthermore, extending these methods to search for correlated mutations between pairs of interacting proteins can identify sites of interprotein interaction11-17. [6] In parallel, related methods have been developed to search for larger groups of residues that are specifically coconserved within particular protein subfamilies. [6] The coevolution between interacting species, such as parasites-hosts, predators-prey and symbionts-hosts, is in many cases manifested as a similarity of the phy logenetic trees of these coevolving species [6] Likewise, molecular coevolution 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 [6] Coevolution at the residue level: Substantial effort has been invested in studying the coevolution of pairs of positions in MSAs of protein families (that is, residue coevolution). These pairs of coevolving positions were often found to correspond to spatially proximal residues in the protein structure, and such putative interresidue contacts have aided protein structure prediction [6] Furthermore, coevolution between residues in different proteins has been used as predic tors of the interacting surfaces (protein interfaces) in pro tein complexes as well as in the search for interacting partners of a given protein, as discussed later in ‘Hybrid residue-protein methods’. [6]

Protein folding: This information can be efficiently mined to detect evolutionary couplings between residues in proteins and address the long-standing challenge to compute protein three-dimensional structures from amino acid sequences. [15] We expect computation of covariation patterns to complement experimental structural biology in elucidating the full spectrum of protein structures, their functional interactions and evolutionary dynamics. [15] In the past 50 years, there has been tremendous progress in experimental determination of protein three-dimensional structures, but this has not kept pace with the explosive growth of sequence information that results from massively parallel sequencing technology. [15] Computational prediction of protein structures, which has been a long-standing challenge in molecular biology for more than 40 years, may be able to fill this gap, if done with sufficient accuracy. [15] However, correct de novo predictions from sequence, when not a single structure in a protein family is known, have been hard to achieve, [15] Clearly, and unfortunately, the de novo structure prediction problem does not scale13, the conformational search space increases exponentially as the size of the protein increases, presenting a fundamental computational challenge, even for fragment-based methods14. In this sense, the general problem of de novo three-dimensional structure prediction has remained unsolved. [15] A substantial step forward in protein-structure prediction is now on the horizon based on the power of evolutionary information found in patterns of correlated mutations in protein sequences [15] Several groups have demonstrated that extracting covariation information from sequences is sufficient not only to estimate which pairs of residues are close in three-dimensional space15-21 but also to fold a protein to reasonable accuracy15,22-25 [15]

It has long been observed that with sufficient correct information about a protein's residue-residue contacts, it is possible to elucidate the fold of the protein (Gobel et al., 1994 [12] The underlying rationale rests on the fact that any given contact critical for maintaining the fold of a protein will constrain the physicochemical properties of the amino acids involved. Should a given contacting residue mutate and potentially perturb the properties of the contact, then its contacting partner will be more likely to mutate to a physicochemically complementary amino acid residue, to ensure the native fold of the protein remains stabilized. [12] Turning this observation around, pairs of residues seen to coevolve in tandem and thus preserving their relative physiochemical properties, are likely candidates to form contacts. [12]

1.1.2 Co-evolution examples

HLA and KIR are well established as interacting immune-response loci under intense diversifying selection. Although these genes are on different chromosomes, their allele frequencies are significantly correlated within human populations, as one would expect under intense selection for allele matching.15 [21]

Potential coevolution between functionally related pro tein 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 coevolution was proposed to be required for the maintenance of their specific interactions2 [6]

Coevolution analysis and functional data for heat-shock proteins, Hsp90 and GroEL, highlight that almost all detected coevolving sites are functionally or structurally important. [9]

In this paper, we explore the ramifications of coevolution between the genes mediating sperm-ZP binding in humans. Specifically, the ZP-located protein ZP3 (MIM 182889) has been shown to mediate sperm binding to the ZP19 [21] Because ZP3 and ZP3R are putative interactors mediating gamete recognition, are polymorphic among humans, and are located on different chromosomes, they are excellent candidates for coevolution-induced allelic association [21] WTCCC, Affymetrix 500K SNP genotyping platform for 1504 individuals in the 1958 Birth. [21] General allelic association between a pair of SNPs is quantified by CLD. An estimate of CLD has been previously given35.. [21] To address this possibility, we use a standard contingency table for independence between the two genotypes (Table 1), resulting in the chi-square distributed test statistic with four degrees of freedom: [21] The CLD and GA test statistics measure allelic association, but they are also dependent on marginal one-locus genotype counts [21] To control for the one-locus genotype counts, X12 and X42 are used as test statistics in permutation tests. [21] Permutation p values approximate exact p values, which are the probabilities of an allelic association at least as strong as that observed, given the marginal genotypes at each locus. [21]

CANCER: Helicobacter pylori is the principal cause of gastric cancer, the second leading cause of cancer mortality worldwide. However, H. pylori prevalence generally does not predict cancer incidence. [13] DATA: To determine whether coevolution between host and pathogen influences disease risk, we examined the association between the severity of gastric lesions and patterns of genomic variation in matched human and H. pylori samples. Patients were recruited from two geographically distinct Colombian populations with significantly different incidences of gastric cancer, but virtually identical prevalence of H. pylori infection. [13] All H. pylori isolates contained the genetic signatures of multiple ancestries, with an ancestral African cluster predominating in a low-risk, coastal population and a European cluster in a high-risk, mountain population. The human ancestry of the biopsied individuals also varied with geography, with mostly African ancestry in the coastal region (58%), and mostly Amerindian ancestry in the mountain region (67%). [13] The interaction between the host and pathogen ancestries completely accounted for the difference in the severity of gastric lesions in the two regions of Colombia. In particular, African H. pylori ancestry was relatively benign in humans of African ancestry but was deleterious in individuals with substantial Amerindian ancestry. [13] Thus, coevolution likely modulated disease risk, and the disruption of coevolved human and H. pylori genomes can explain the high incidence of gastric disease in the mountain population. [13]

The field has yet to identify a gene pair that is certainly coevolving in which both genes are polymorphic. In the absence of a clear positive control, [21]

The divergent evolution of proteins in cellular signaling pathways requires ligands and their receptors to co-evolve, creating new pathways when a new receptor is activated by a new ligand. [11] We have used phosphoglycerate kinase (PGK), an enzyme that forms its active site between its two domains, to develop a standard for measuring the co-evolution of interacting proteins. [11] The N-terminal and C-terminal domains of PGK form the active site at their interface and are covalently linked. Therefore, they must have co-evolved to preserve enzyme function. [11] The analysis is extended to ligands and their receptors, using the chemokines as a model. [11] The chemokine family of protein ligands and their G-protein coupled receptors have coevolved so that each subgroup of chemokine ligands has a matching subgroup of chemokine receptors. [11] Protein-protein binding is a subset of these interactions which is of primary importance in metabolic and signaling pathways. [11] Proteins and their interaction partners must coevolve so that any divergent changes in one partner's binding surface are complemented at the interface by their interaction partner (Atwell et al., 1997; Jespers et al., 1999; Moyle et al., 1994; Pazos et al., 1997). [11]

GroESL is a heat-shock protein ubiquitous in bacteria and eukaryotic organelles. This evolutionarily conserved protein is involved in the folding of a wide variety of other proteins in the cytosol, being essential to the cell. [22] The folding activity proceeds through strong conformational changes mediated by the co-chaperonin GroES and ATP: [22] We hypothesize that different overlapping sets of amino acids coevolve within GroEL, GroES and between both these proteins [22] METHOD: CAPS [22]

Genome-wide scans for signals of natural selection in human populations have identified a large number of candidate loci that underlie local adaptations. This is surprising given the relatively short evolutionary time since the divergence of the human population. Approximately 50,000-100,000 years ago, the anatomically modern human migrated from Africa to the rest of the globe [19] One hypothesis that has not been formally examined is whether and how the recent human evolution may have been shaped by coselection in the context of complex molecular interactome [19] In this study, genome-wide signals of selection were scanned in East Asians, Europeans, and Africans using 1000 Genome data, and subsequently mapped onto the protein-protein interaction (PPI) network [19] We found that the candidate genes of recent positive selection localized significantly closer to each other on the PPI network than expected [19] Furthermore, gene pairs of shorter PPI network distances showed higher similarities of their recent evolutionary paths than those further apart. [19] Several hundred to more than one thousand candidate regions, which may have undergone recent positive selection, have been reported in these studies (Sabeti et al. 2002; Voight et al. 2006; Tang et al. 2007; Akey 2009; Pickrell et al. 2009). The abundant selection signals are in sheer contrast with the relatively short period of time since humans migrated from Africa. [19] EXAMPLES: A number of positively selected candidate genes in the same pathways or interaction subnetworks have also been identified. Known examples are EGLN1 and EPAS1 in the hypoxia-response pathway playing key roles in genetic adaptation to high-altitude regions (Xu et al. 2011) as well as multiple genes in the NRG-ERBB4 developmental pathway (Pickrell et al. 2009) [19] Coevolution of interacting proteins, in large time frames, has been intensively studied and is typically based on the evolutionary distances across different species [19] an alternative mechanism notes that epistatic interaction is not compulsory to explain the associated evolutionary patterns. If selection pressures act on an entire pathway or a functional subnetwork, multiple genes in the same pathway/subnetwork may change in the same fitness direction, and at a same evolutionary rate and time to achieve a common phenotypic outcome. [19] the association in evolutionary patterns may simply reflect parallel selection of different genes in the same pathway of shared functionality [19] Nonetheless, both hypothetic mechanisms would lead to a set of similar predictions: first, the genes of positive selection would cluster closer to each other in the PPI network than predicted under null hypothesis; second, the clustered genes of selection may share more similar evolutionary paths than genes unrelated on the PPI network. [19] Numerous studies reported that proteins located closer to the center of PPI network evolved more slowly than those at the periphery of the network, consistent with the view that central proteins are more essential and receive greater evolutionary constraints [19] DATA: 1000 Genomes Project interim phase 1 data set, including Europeans (CEU; 85 samples), Yorubans (YRI; 88 samples), and EAS (186 samples CHB+JPT). T [19] METHOD: [19] A modified CMS method was applied to scan for genomewide signals of recent positive selection, as previously described (Grossman et al.

2010). [19] Degree centrality (DC) and betweenness centrality (BC) were used to measure both local and global topological positions of candidate genes on the human PPI network (Freeman 1977; Kim et al. 2007). Degree is defined as the number of connections a node has with its neighbors whereas betweenness quantifies the number of times a node acts as a bridge along the shortest path between any other pairwise nodes. [19] The Mann-Whitney U test (also called Wilcoxon rank sum test) was applied to compare the two centrality measures, BC and DC, between selection signals and nonselection signals, to determine whether network position influences recent positive selection [19] RESULTS [19] we found a moderate pattern that recent positive selections tended to occur more on the subcentral region of the PPI network, [19] However, it differs from studies done on a macroevolutionary timescale, which have consistently reported that accelerated evolutionary rates tend to happen at the periphery of the protein interaction network, [19]

Amino acid covariation, where the identities of amino acids at different sequence positions are correlated, is a hallmark of naturally occurring proteins. [17] This covariation can arise from multiple factors, including selective pressures for maintaining protein structure, requirements imposed by a specific function, or from phylogenetic sampling bias. [17] Here we employed flexible backbone computational protein design to quantify the extent to which protein structure has constrained amino acid covariation for 40 diverse protein domains. [17] We find significant similarities between the amino acid covariation in alignments of natural protein sequences and sequences optimized for their structures by computational protein design methods. [17] These results indicate that the structural constraints imposed by protein architecture play a dominant role in shaping amino acid covariation and that computational protein design methods can capture these effects. [17] Evolutionary selective pressures on protein structure and function have shaped the sequences of today's naturally occurring proteins [1-3]. As a result of these pressures, sequences of natural proteins are close to optimal for their structures [17] Natural protein sequences therefore provide an excellent test for computational protein design methods, where the goal is to predict protein sequences that are optimal for a desired protein structure and function [5]. [17] Beyond simply recovering the native sequence, a further challenge in computational protein design is to predict the set of tolerated sequences that are compatible with a given protein fold and function [9-13]. Predicting sequence tolerance is important for applications such as characterizing mutational robustness [14,15], predicting the specificity of molecular interactions [16- 20], and designing libraries of proteins with altered functions [21,22]. [17] Previous work has indicated that networks of covarying amino acids play a role in allosterically linking distant functional sites, suggesting that amino acid covariation is driven by protein functional constraints [30,31]. [17] In this paper, we use computational protein design to measure the extent to which protein structure has shaped amino acid covariation in a diverse set of 40 protein domains. [17] Since computational protein design predicts sequences that are energetically optimal based on protein structure alone, we expect that pairs of amino acids that highly covary in both designed and natural sequences to have likely covaried to maintain protein structure [17] We find significant overlap in the sets of highly covarying amino acid pairs between designed and natural sequences for all 40 domains examined, suggesting that maintenance of protein structure is a dominant selective pressure that constrains the evolution of amino acid interactions in proteins. [17] METHOD: The mutual information (MI) between each pair of columns , Z-score resepct to the mean MI. This normalization of MIp was demonstrated to reduce the sensitivity to misaligned regions in multiple sequence alignments, which otherwise result in artificially high mutual information scores [28]. [17]

1.1.3 Detecting co-evolution

SAMINAL PAPER (could not download): The maintenance of protein function and structure constrains the evolution of amino acid sequences. This fact can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions. Here we present a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. We then use these correlations to predict contact maps for each of 11 protein families and compare the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranges from 37 to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. Predicted contact maps can be used as input for the calculation of protein tertiary structure, either from sequence information alone or in combination with experimental information. [10] METHOD (from another paper): Correlated mutations were calculated as described (GoEbel et al., 1994). Each position in the alignment is coded by a distance matrix. This position-specific matrix contains the distances between all pairs of sequences at that position. Distances are defined by the scoring matrix of McLachlan (1971). The association between each pair of positions is calculated as the average of the correlation for each corresponding bin of the position-specific matrices. Positions with more than 10% gaps or completely conserved were not included in the calculation. [10]

Here we apply a method for detecting correlated changes in multiple sequence alignments to a set of interacting protein domains and show that positions where changes occur in a correlated fashion in the two interacting molecules tend to be close to the proteinprotein interfaces. [18] This leads to the possibility of developing a method for predicting contacting pairs of residues from the sequence alone. Such a method would not need the knowledge of the structure of the interacting proteins, and hence would be both radically different and more widely applicable than traditional docking methods. [18] We indeed demonstrate here that the information about correlated sequence changes is sufficient to single out the right inter-domain docking solution amongst many wrong alternatives of two-domain proteins. [18] We propose here a new and completely different approach to the study and prediction of protein protein interaction. Instead of considering the structural nature of the interactions, we try to detect the sequence traces that evolution may have left on the interacting sequences during the process of preserving the proteinprotein interaction sites. [18] Therefore, our approach is not restricted to the cases in which the structures of the proteins to be docked are known and is applicable to any family of interacting proteins for which a large enough sequence family is available. [18] Over time, amino acid substitution may stabilise an interface that does not exist in the closed monomer ... stabilising mutations in these interfaces would be favoured in natural selection" (Bennett et al., 1995); [18] We propose that it is possible to detect this signal by studying compensatory mutations. In order to do so, we have appropriately modified our previously published method for the calculation of correlated mutations in multiple-sequence alignments (GoEbel et al., 1994; Pazos et al., 1997). [18]

Methods have been developed for detecting coevolution by testing for high correlation of phylogenetic distance matrices between gene families, genes, or gene domains.1-6 [21]

CAPS: Methods designed to detect adaptive evolution can be based on Bayesian approaches (Yang et al. 2000) or maximum parsimony (Suzuki and Gojobori 1999; Fares et al. 2002a). None of these methods takes into account the evolutionary interdependence between protein residues [9] Sites constraints are hence dependent on the interactions with other residues of the molecule [9] Mutations at either nearby sites or functionally related distant sites in the structure will change the selective constraints. [9] For instance, linear sliding-window methods are one-dimensional based and assume independence between different window regions irrespective of their three-dimensional proximity [9] Conversely, classification of amino acids in the same group of evolution based on their three-dimensional proximity (three-dimensional sliding window) will ignore the coevolution between functional regions that are spatially distant Various reports state that residues can form a physically connected network that links distant functional sites in the tertiary protein structure (Su el et al. 2003) [9] Coevolution between clusters of sites, which are not in contact, has also been shown (Pritchard and Dufton 2000) [9] Coevolution between distant sites has been observed in sites proximal to regions with critical functions, where coevolution occurs to maintain the structural characteristics around these regions and consequently to maintain the protein conformational and functional stability (Gloor et al. 2005). [9] [various methods coevolution exist...] =¿ The main limitation of many of these methods has been their inability to separate phylogenetic linkage from functional and structural coevolution. [9] Gloor et al. (2005) partially corrected these effects although their method requires alignments of at least 125 sequences to remove stochastic covariation.

[9] METHOD: [9] The method instead compares the transition probability scores between two sequences at these particular sites, using the blocks substitution matrix (BLOSUM) [9] For each protein alignment the correspondent BLOSUM matrix is applied, depending on the average sequence identity. [9] an alignment including two highly divergent sequence groups (for example, gene duplication predating speciation) could show an unrealistic pairwise average identity level. In this respect, sequences that diverged a long time ago are more likely to fix correlated mutations at two sites by chance [9] BLOSUM values should be hence normalized by the time of divergence between sequences. BLOSUM values (Bek) are thus weighted for the transition between amino acids e and k using the time (t) since the divergence between sequences i and j: [9] The mean variability for the corrected BLOSUM transition is... [9] The coevolution between amino acid sites (A and B) is estimated thereafter by measuring the correlation in the pairwise amino acid variability, relative to the mean pairwise variability per site, between them. [9] LIMITATIONS: [9] For example, saturation of synonymous sites can lead to underestimates of the divergence times, although data sets used in this study did not show such effects. [9] The number of sequences in the alignment also poses a problem when sequences are too divergent, although the sensitivity is improved compared to that of previous methods. [9] Further, constant amino acid sites that are very likely to be functionally important cannot be tested for coevolution using CAPS, although this limitation affects all the methods so far. [9]

Different alignments of the same protein family give different results demonstrating that covariation depends on the quality of the sequence alignment. [7] We show that current criteria are insufficient to build alignments for use with covariation analyses as systematic sequence alignment errors are present even in hand-curated structure-based alignment datasets like those from the Conserved Domain Database. [7] We demonstrate that removing alignment errors due to 1) improper structure alignment, 2) the presence of paralogous sequences, and 3) partial or otherwise erroneous sequences, improves contact prediction by covariation analysis [7] Standard benchmarks for covariation accuracy measure the fraction of covarying amino acid pairs that are in contact. [7] First, the sequence alignments must contain sufficient sequences with enough variation for the signal to exceed the noise. Estimates of the required number of sequences needed in the alignments for this to be true vary from *30 [6] to w125 [4,8,15,16]. [7] Secondly, all positions in a protein appear to covary because of their shared ancestry, and this signal is the only systematic source of covariation for the vast majority of position pairs [6,14,17]. [7] As one example, structure-based alignment algorithms are susceptible to shift error [18], meaning that positions in the structure alignment are not orthologous despite the fact that much of the secondary structural elements seem to overlap between aligned structures. [7] We observed that the same protein family often gave different numbers of covarying positions when alignments were from different sources even if the alignments contained comparable numbers of sequences. [7] We also found that alignments generated without structural information identified fewer pairs in contact in the folded protein compared to alignments generated with structural information. [7] Here we show that a strong covariation signal can be caused by alignment error, potentially leading to false positive predictions. [7]

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 dis tributed range from distant to similar sequences [6] In general, optimal performance is obtained when protein subfamilies (branches of the tree) are spaced at regular intervals, [6] 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 rear rangements have occurred. [6]

1.1.4 Co-evolution algorithm complexity

Calculating the power of these exact tests can be prohibitively slow with a large sample size. As an alternative, we quickly estimate power by using theoretical test statistic distributions under the alternative hypothesis. Under the alternative hypothesis with genotype frequency matrix F, X 2 is approximately chi-square 1 distributed with one degree of freedom and noncentrality parameter [21] We ran a similar analysis on a secondary candidate gene pair implicated in maternal-fetal interactions: GHR (MIM 600946) and GH2 (MIM 139240).37 [21] Power: because of computational limitations, we were unable to perform the exact test for larger value of n; [21] Asymptotic Analysis: For a high but biologically reasonable s of 0.1,38 with a sample size of n 14 1480, the asymptotic CLD test has a power of 0.525 and the asymptotic GA test has a power of 0.327 [21]

Population: Population structure could also cause allelic association between physically unlinked loci. [21] Allelic association would be observed if the alleles at each locus have different frequencies in different populations and those populations are pooled together. In this analysis, ZP3 and ZP3R are associated as compared to other genes in the same individuals. It is not likely that population structure would cause allelic association in our candidate gene pair but not in other gene pairs in the same population. It is possible that ZP3 and ZP3R are statistical outliers that we expect under no selection and are associated simply by chance. However, given our limited single-hypothesis candidate gene approach, we find that unlikely. [21]

Predicting interaction specificity, such as matching members of a ligand family to specific members of a receptor family, is largely an unsolved problem. Here we show that by using evolutionary relationships within such families, it is possible to predict their physical interaction specificities. [20] We introduce the computational method of matrix alignment for finding the optimal alignment between protein family similarity matrices [20] Binding specificities of duplicate genes (paralogs) often diverge, such that new binding specificities are evolved [20] the use of phylogenetic trees to account for the co-evolution of interacting proteins [20] the hypothesis underlying these approaches is that interacting proteins often exhibit coordinated evolution, and therefore tend to have similar phylogenetic trees. Goh et al.17 demonstrated this by showing that chemokines and their receptors have very similar phylogenetic trees [20] In order to exploit the evolutionary information contained in such interacting protein families, we developed an algorithm that is conceptually equivalent to superimposing the phylogenetic trees of the two protein families. [20] The matrix alignment method for predicting protein interaction specificity. Proteins in family A interact with those in family B. In each family, a similarity matrix summarizes the proteins' evolutionary relationships. The algorithm uses the similarity matrices to pair up the genes in the two families. Columns of matrix B are re-ordered (along with their corresponding rows in the matrix) such that the B matrix agrees maximally with matrix A, judged by minimizing the root mean square difference (r.m.s.d.) between elements in the two matrices. Interactions are then predicted between proteins heading equivalent columns of the two matrices. [20] One matrix is shuffled, maintaining the correct relationships between proteins but simply re-ordering them in the matrix, until the two matrices maximally agree, minimizing the root mean square difference between elements of the two matrices. Interactions are then predicted between proteins heading equivalent columns of the two matrices. [20] For matrix alignment, MATRIX currently applies a stochastic simulated annealing-based algorithm. [20]

Detecting correlated amino acid changes in pairs of positions. Residue coevolution was originally assessed through detecting pairs of positions (two columns of the MSA) that have interdependent amino acid frequencies23 or similar patterns of amino acid substitutions7,9,10 [6] ...can be assessed by a linear correlation. 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 contact24 and still serves as a baseline to benchmark the performance of new methods25. [6] CAPS dampens the influence of background phylo genetic divergence by requiring the detected correlations to still be detected after particular clades are removed from the MSA. It also corrects the amino acid substi tution matrix so as to consider the actual divergence among the sequences [6] MI: Mutual information has been also used to detect covarying positions. Whereas correlationbased meth ods explore intersequence 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 pres ence of an amino acid in a given sequence for a posi tion is a ‘good prediction’ of the presence of any given amino acid in the same sequence for a second position. 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 covariations. Therefore, the different amino acids are treated as different sym bols 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. [6] MARKOV MODELS: In this case, the use of an enhanced continuous-time Markov process model for sequence coevolution represented an important step forwards13. These approaches are suit able for smallscale studies of coevolution in small pro tein families, but the evaluation of their performance in largescale studies remains excessively demanding in computational terms. [6]

Indirect correlations: [15] if residues A and B contact each other, as do residues B and C, then there is in general, a transitive influence observed between residues A and C ('chaining effect'17,27). [15] As residues can contact many other residues (not just one), transitive effects occur across the network, and pairs of residues that are correlated as computed using a 'local' statistical model, such as mutual information scores, are not necessarily functionally constrained or close in space [15] LOCAL MODELS: [15] Local statistical models (below referred to as local models or local methods) assume that pairs of residue positions are statistically independent of other pairs of residues [15] Other confounding effects that have prevented high-accuracy prediction of residue contacts include uneven representation of family members in sequence space, statisticalnoise as the result of an inadequate number of sequences in the family as well as phylogenetic effects. [15]

1.1.5 Complex models

Disentangling directly coupled residues from the net work of indirectly correlated positions. [6]

An important obstacle in the detection of coevolving positions is the apparent covariation 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 the evolutionary interdependence of both positions with one or more additional positions. The aggrega tion of these indirect couplings can make it difficult to recognize the directly interdependent positions. As the direct couplings are more reliable for predict ing physically proximal residues in protein structures, approaches are needed to distinguish direct from indi rect couplings [6] 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 covariations. This first approach used a Monte Carlo algorithm to infer the sim plest probabilistic model that was able to account for the whole network of covariations in a simulated sce nario. [6] 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 positionspecific variability and interposition coupling [6] Alternatively, Burger and van Nimwegen's41 method uses a Bayesian network model that includes pair wise conditional dependencies, and the regularized multinomial regressionbased correlated mutations (RMRCM) approach42 takes into account the whole network of dependencies and not only the individual pairwise dependencies. [6] For MSAs with more than 1,000 sequences, DCA and PSICOV seem to be superior to Burger and van Nimwegen's method38,39. [6] 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 [6]

GLOBAL MODELS [15] In contrast, a 'global' modeling approach treats correlated pairs of residues as dependent on each other, rather than as statistically independent, thereby minimizing the effects of transitivity and spurious noise. [15] This approach also uses globally consistent single-residue marginals, which takes into account effects from conservation of single residue positions. Global approaches yield high coupling scores only for pairs or residue positions that are likely to be causative of all the observed correlations. [15] Noncausal correlation is well understood in statistical physics; it includes, for instance, long-range order observed in spin systems, where in fact the spins only have short-range direct interactions, and is called 'chained covariation'27,34. [15] One global statistical approach is known as entropy maximization under data constraints, a classic inference method connecting information theory and Boltzmann statistics 35 [15] Maximizing entropy under constraints36 has been successfully used in statistical physics and other areas of statistical inference37-39, and the conditional mutual information derived from correlations between positions in a protein sequence is a discrete, nonlinear analog of partial correlation analysis40 [15] In contrast to simple mutual information, the conditional mutual information can be thought of as the degree of covariation between residues at positions a and b that is due solely to direct effects of a on b, factoring out contributions to the correlation that are caused by interaction of both a and b with the rest of the network of residues. [15] CORRELATION (PSICOV) the covariance matrix (the observed minus expected pair counts) of dimension (20L)2, where L is the length of the protein sequence, by counting how often a given pair of the 20 amino acids, say alanine and lysine, occurs in a particular pair of positions, say position 15 and 67, in any one sequence, summing over all sequences in the multiple-sequence alignment. This large matrix contains the raw data capturing all residue pair relationships across evolution up to second order (pairs, not triplets or higher). One can then compute a measure of causative correlations, the conditional mutual information, in the global statistical approaches by taking the inverse of the covariance matrix. That such a matrix inversion results in a measure of causative correlations is well known in the statistical theory of Gaussian multivariate distributions of continuous variables40. [15] MEAN FIELD APPROX: An analogous derivation for discrete-state biological sequence analysis is, for example, based on a mean-field expansion in analogy to statistical physics16. The resulting explicit probability model for a sequence in the particular protein family resulting from inversion of the covariation matrix contains numerical estimates of direct pair interactions. These are directly and simply computed from the raw data in the covariation matrix, in contradistinction to machine-learning methods that rely on parameter fitting in learning sets and cross-validation in test sets. The pair interaction terms can also be interpreted as residue-residue pair energies, in analogy to pair terms in a Hamiltonian energy expression in statistical physics. The conditional mutual information between a pair of positions derived using the global statistical approach becomes a useful predictor of residue-residue contacts. [15] The maximum-entropy approach to potentially solving the problem of protein structure prediction from residue covariation patterns was first described by Lapedes and collaborators17,27. However, instead of inversion of the covariance matrix, they used a more computationally demanding Monte Carlo method (that is, iterative exploration of the best set of pair interactions values) to derive the probability terms in conditional mutual information. Although Lapedes and Jarzynski did not compute three-dimensional structures, they reached a first breakthrough in contact prediction in 2002 for 11 small proteins and reported 50-70% accuracy for top 20 contact predictions, in contrast to 35-45% accuracy with the previous best methods available17. [15]

However, because the coevolving genes are not necessarily in physical linkage, this is not an appropriate measure of coevolution-induced allelic association [21]

SEMINAL PAPER (Lapedes 2002) We present a sequence-based probabilistic formalism that directly addresses co-operative effects in networks of interacting positions in proteins, providing significantly improved contact prediction [14] Each sequence of length L of a given family can be viewed as a different global state of an L-site, twenty-state (for twenty amino acids) spin system, with spinspin (i.e. residue-residue) interactions determined by (1) the (unknown) structure of the associated fold, and (2) the physico-chemical characteristics of the residues [14] Solving the inverse problem to determine the underlying physical interactions addresses "correlation at a distance", in which correlations between locally connected sites in an interacting network such as a spin system [14] Previous computational work on abstract models of proteins [4], as well as a statistical analysis of the frequency of ion-pairs in crystal structures of real proteins [5], provided early hints that Boltzmann-like statistics are associated with aspects of protein architecture. [14] The Boltzmann network method presented here does not treat each individual pair of sites of interest as isolated from other residues. Instead, we construct a probability distribution describing full length sequences of length L for each protein sequence family. [14] Any given sequence alignment typically contains enough data to estimate only single and pairwise amino acid frequencies with reasonable accuracy. [14] The maximum entropy distribution whose moments match a given set of single and pairwise amino acid frequencies may be written in the following form [23], reminiscent of thermal Boltzmann statistics [14] It can be shown [25] that matching the moments of the maximum entropy distribution to the given sequence data is equivalent to maximizing the log-likelihood of the given sequence data given the parametric form [14] we use the probability distribution over all L sites, Eqns. (1,2), to resolve issues of correlation at a distance (network effects) in proteins, resulting in significantly improved contact prediction from sequence information [14] LIMITATIONS: Limiting factors in application of the Boltzmann network algorithm include (1) the amount of naturally evolved sequence data currently available per family (size of the sequence alignment), and (2) the phylogenetic relatedness (and associated selection artifacts) of these sequences. Modifications to the algorithm presented here, e.g. (1) consideration of statistical significance of the fitted parameters, and (2) addressing phylogenetic relationships of sequences in an alignment, have the potential to further increase accuracy using naturally evolved sequence sets. [14]

Applied to a set of ¿2,500 representatives of the bacterial two-component signal transduction system, the combination of covariance with global inference successfully and robustly identified residue pairs that are proximal in space without resorting to ad hoc tuning parameters, both for heterointeractions between sensor kinase (SK) and response regulator (RR) proteins and for homointeractions between RR proteins. [24] The spectacular success of this approach illustrates the effectiveness of the global inference approach in identifying direct interaction based on sequence information alone. [24] Experimental approaches to identify surfaces of interaction between proteins such as surface-scanning mutagenesis and cocrystal structure generation are arduous and/or serendipitous. [24] Covariance methods rely on the premise that amino acid substitution patterns between interacting residues are constrained and hence correlated. To maintain protein function, the acceptance of a deleterious substitution at 1 position must be compensated for by substitution(s) in the residue(s) interacting with it (14) [24] However, the covariance approach has a number of shortcomings that may significantly affect its predictive power (15). One important problem stems from the fact that correlation in amino acid substitution may arise from direct as well as indirect interactions [24] A formidable technical challenge with this approach is to work out the expected statistical correlation generated by a given set of trial direct interactions, because this itself is a very difficult global optimization problem [as exemplified by the notorious “spinglass” problem (16)]. This challenge is dealt with here by applying a message-passing approach (17, 18). In recent years, insights from spin-glass physics have led to the development of generalized message-passing techniques, which have been applied successfully to a number of hard combinatorial problems such as K-SAT (19 -21). [24] The statistically correlated pairs are candidates for positions in contact at the protein-protein interface. However, statistical correlation does not automatically imply strong direct interaction. Imagine that position i is coupled directly to j, and j to k. Then i and k will also show correlation, without being directly coupled. [24] To circumvent this problem, we infer a global statistical model [24] Note that in principle higher correlations of 3 or more positions can be included in a similar way. However, the size of the available dataset does not allow for going beyond 2-residue correlations. The 21 21 elements of fij(Ai, Aj) have to be estimated from the M 2,546 sequences in the database; frequency counts for 2 positions would be very imprecise because of insufficient sample size. [24] Application of the maximumentropy principle yields the simplest possible [Boltzman distribution] [24] Determining these parameters to meet Eq. 1 is an algorithmically hard task, and can be achieved by using a 2-step procedure. [24] Givenacandidatesetofmodelparameters,single-and2-residue distributions Pi(Ai) and Pij(Ai, Aj) are estimated from Eq. 2. This is computationally expensive, the exact summation over all possible protein sequences would require $O(21^{N-2}N^2)$ steps. Approximations can be achieved by MCMC sampling-which is expected to be very slow for 21-state variables-or more efficiently by a semiheuristic message-passing approach (31). We use the latter approach; it reduces the computational complexity to $O(21^2N^4)$. [24] Once all Pij(Ai, Aj) are estimated, we can use gradient descent to adjust the coupling strengths eij(Ai, Aj) [24] This equation can be derived variationally within a Bayesian approach, it maximizes the joint probability of the data under model 2 (compare SI Text). Because this probability is convex, it is guaranteed to converge to a single global maximum. [24] a quantity called direct information (DI) is introduced. It measures the part of the mutual information of a position pair, which is induced by the direct coupling. Intuitively, it can be understood as the mutual information in a 2-variable model for positions i and j only, which has the correct statistics of the amino acid occupancy of single positions, and coupling eij(Ai, Aj) in between. [24] Because of the scaling of the algorithmic complexity, the method cannot be applied simultaneously to all 212 positions of the protein alignment. Therefore, the 60 positions of the protein alignment being involved in the 140 highest MI-ranking pairs (containing the 32 candidates for contact pairs identified before) are selected. [24]

A key impediment to this approach is that strong statistical dependencies are also observed for many residue pairs that are distal in the structure. [1] Using a comprehensive analysis of protein domains with available three-dimensional structures we show that co-evolving contacts very commonly form chains that percolate through the protein structure, inducing indirect statistical dependencies between many distal pairs of residues [1] The identification of functionally and structurally important elements in DNA, RNA and proteins from their sequences has been a major focus of computational biology for several decades. A common approach is to create a multiple alignment of homologous sequences, which places ‘equivalent’ residues into the same column and as such gives a hint of the evolutionary constraints that are acting on related sequences. [1] Markov models [1] of protein families and domains have been highly successful in identifying sequences that have similar function and fold into a common structure, [1] These hidden Markov models typically assume that the residues occurring at a given position are probabilistically independent of the residues occurring at other positions. At the time at which these models were developed, it was entirely reasonable to ignore dependencies between residues at different positions, since the amount of available sequence data was generally insufficient to estimate joint probabilities of multiple residues. [1] As the functionality of biomolecules crucially depends on their three-dimensional structures, whose stabilities depend on interactions between residues that are near to each other in space, it is of course to be expected that significant dependencies between residues at different positions will exist. [1] CAPS and MI SUCK: [1] We collected a comprehensive set of 2009 multiple alignments of protein domains from the Pfam database [19] for which a three dimensional structure was available (see Materials and Methods) and calculated, for each pair (ij) of columns in each alignment, the statistical dependency using a measure, log (Rij), which is a finite-size corrected version of mutual information (see Materials and Methods). Since the distribution of log (R) values for an alignment depends strongly on the number of sequences in the alignment, their phylogenetic relationship, and the length of the alignment, log (R) values cannot be directly compared across different alignments. Therefore, we calculated the mean and variance of log (R) values for each alignment and transformed the log (R) values to Z-values (number of standard deviations from the mean). Finally, for each alignment, we divided all pairs of residues into those that are contacting in the three-dimensional structure, and those that are distant in the structure, and calculated the distribution of Z-values for these two sets of residue pairs. As in previous work (e.g. [10,20]) and as defined for CASP [21], two residues were considered in [....] [1] [...] indeed, a higher fraction of contacting residues shows strong statistical dependencies than distal residues. However, we also see that the difference in the Z-distribution of close and distal pairs is only moderate. [1] Since there are generally many more distal pairs than close pairs, this implies that, even at high Z-values, the majority of residuepairs are in fact distal in the structure [1] This result shows that simple measures of statistical dependency, such as mutual information, are poor at predicting which pairs of residues are directly contacting in the structure. [1] WHY DO MI AND CAPS SUCK? [1] The main question is why so many structurally distal pairs show statistical dependencies in their amino-acid distributions that are stronger than those between directly contacting residues. [1] 1) First, whereas measures such as mutual information treat the sequences in the multiple alignments as statistically independent, in reality many of the sequences are phylogenetically closely related [1] 2) Some of these distant dependencies have been suggested to be caused by homooligomeric interactions [14,22]. Thus, in this interpretation, some of the 'distal' pairs with strong statistical dependencies are in fact contacting in the homo-oligomer. [1] 3) dependencies are induced by indirect interactions that are mediated either by intermediate molecules [15,23] or by chains of directly interacting residue pairs that run through the protein and connect distal pairs [23-25] [1] METHOD: [1] We show that a Bayesian network model which we recently developed to predict protein-protein interactions [27] can be adapted to rigorously disentangle direct from indirect statistical dependencies between residues [1] Briefly, our model assumes that the sequences in a multiple alignment D (the data) are drawn from an (unknown) underlying joint probability distribution P(x1,x2,...,xl) with l the width of the alignment and xi the amino acid at position i. Profile hidden Markov models typically assume that the amino acids at different positions are independent [1] Any model that considers only pairwise conditional dependencies factorizes the joint probability...where $\pi(i)$ is the single other position which the residue at position i depends on [1] In particular, we do not attempt to estimate the conditional probabilities P(xi jxj) but rather treat these conditional probabilities as nuisance parameters that we integrate out in calculating the likelihood of the alignment. [1] In addition, and importantly, we do not consider only a single 'best' way of choosing which other position p(i) each position i depends on, but rather we sum over all ways in which the dependencies can be chosen. [1] The sum over spanning trees in (9) can be calculated using a generalization of Kirchhoff's matrix-tree theorem. For this we need to calculate the Laplacian of the matrix... [1]

MEAN FIELD Crucial to this inference is the ability to disentangle direct and indirect correlations, as accomplished by the recently introduced direct-coupling analysis (DCA). Here we develop a computationally efficient implementation of DCA, which allows us to evaluate the accuracy of contact prediction by DCA for a large number of protein domains, based purely on sequence information. [16] DCA is shown to yield a large number of correctly predicted contacts, recapitulating the global structure of the contact map for the majority of the protein domains examined. [16] Furthermore, our analysis captures clear signals beyond intradomain residue contacts, arising, e.g., from alternative protein conformations, ligand-mediated residue couplings, and interdomain interactions in protein oligomers. [16] Correlated substitution patterns between residues of a protein family have been exploited to reveal information on the structures of proteins (1-10). [16] However, such studies require a large number (e.g., the order of 1,000) of homologous yet variable protein sequences. [16] If two residues of a protein or a pair of interacting proteins form a contact, a destabilizing amino acid substitution at one position is expected to be compensated by a substitution of the other position over the evolutionary timescale, in order for the residue pair to maintain attractive interaction. [16] A major shortcoming of covariance analysis is that correlations between substitution patterns of interacting residues induce secondary correlations between noninteracting residues [16] This problem was subsequently overcome by the direct-coupling analysis (DCA) (16, 17), which aims at disentangling direct from indirect correlations. [16] The top 10 residue pairs identified by DCA were all shown to be true contacts between the TCS proteins, and they were used to guide the accurate prediction (3-A rmsd) of the interacting TCS protein complex (18, 19) [16] Previously, a message-passing algorithm was used to implement DCA (16). This approach, here referred to as mpDCA, was rather costly computationally because it is based on a slowly converging iterative scheme. This cost makes it unfeasible to apply mpDCA to large-scale analysis across many protein families. [16] Here we will introduce mfDCA, an algorithm based on the meanfield approximation of DCA. The mfDCA is 10^3 to 10^4 times faster than mpDCA [16] Starting with a multiple-sequence alignment (MSA) of a large number of sequences of a given protein domain, extracted using Pfam's hidden Markov models (HMMs) (21, 22), the basic quantities in this context are the frequency count f i A for a single MSA column i, characterizing the relative frequency of finding amino acid A in this column, and the frequency count f ijA;B for pairs of MSA columns i and j, characterizing the frequency that amino acids A and B coappear in the same protein sequence in MSA columns i and j. Alignment gaps are considered as the 21st amino acid. Mathematical definitions of these counts are provided in Methods. [16] The raw statistical correlation obtained above suffers from a sampling bias, resulting from phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. The problem has been discussed extensively in the literature (10, 23-26). [16] In this study, we implemented a simple sampling correction, by counting sequences with more than 80% identity and reweighting them in the frequency counts. [16] A simple measure of correlation between these two columns is the mutual information (MI), defined by Eq. 3 in Methods. As we will show, the MI turns out to be an unreliable predictor of spatial proximity. [16] Central to our approach is the disentanglement of direct and indirect correlations, which is attempted via DCA, [16] This algorithm, termed mfDCA, is able to perform DCA for alignments of up to about 500 amino acids per row, as compared to 60-70 amino acids in the message-passing approach. [16] METHOD [mean field calculation] [16] To disentangle direct and indirect couplings, we aim at inferring a statistical model P(A1;...;AL) for entire protein sequences (A1 ;...;AL). [16] Besides this constraint, we aim at the most general, least-constrained model PA1;:::;AL. This model can be achieved by applying the maximum-entropy principle (45, 46), and it leads to an explicit mathematical form of PA1;:::;AL as a Boltzmann distribution with pairwise couplings eij A;B and local biases [16] The exponential of [the partition function] is expanded into a Taylor series. Keeping only the linear order of this expansion, we obtain the well-known mean-field equations [16] For later convenience, we also introduce the Hamiltonian [exponential of negative Hamiltonian is the partition function] [16] It is important to note that the partition function itself contains all necessary information on the marginals, in particular we have.... [16] The algorithmic approach is based on a systematic small-coupling expansion, i.e., on a Taylor expansion around zero coupling. This expansion was introduced in [12] by Plefka for disordered Ising models (Ising spinglasses, corresponding to binary variables with q = 2). [16] Furthermore we introduce the so-called Gibbs potential ... as the Legendre transform of the free energy F = ln Z . [16] The first derivative of the Gibbs potential with respect to equals thus the average of the coupling term in the Hamiltonian. At = 0, this average can be done easily, since the joint distribution of all variables becomes factorized over the single sites [.......] we find the firstorder approximation of the Gibbs potential [16]

The starting point of our method is to consider an alignment with m columns and n rows, [12] where each row represents a different homologous sequence and each column a set of equivalent amino acids across the evolutionary tree, with gaps considered as an additional amino acid type. We can compute a 21m by 21m sample covariance matrix as follows: [12] Any individual element of this matrix gives the covariance of amino acid type a at position i with amino acid type b at position j. [12] By calculating the matrix inverse of the covariance matrix, the precision or concentration matrix () is obtained, from which a matrix of partial correlation coefficients for all pairs of variables can be calculated as follows [12] In the simplest case, a partial correlation coefficient can be calculated between two random variables with the controlling effect of a third random variable taken into account. The partial correlation matrix above, however, gives the correlations between all pairs of variables with the controlling effects of all other variables taken into account [12] Thus, assuming the sample covariance matrix can in fact be inverted, the inverse covariance matrix provides information on the degree of direct coupling between pairs of sites in the given MSA. Off-diagonal elements of the inverse covariance matrix which are significantly different from zero are indicative of pairs of sites which have strong direct coupling (and are likely to be in direct physical contact in the native structure). [12] Unfortunately, the empirical covariance matrices produced in this application are guaranteed to be singular due to the fact that not every amino acid will be observed at every site [12] Although different approaches have been proposed to allow inverse covariance estimation where the sample covariance matrix cannot be directly inverted, one of the most powerful techniques is that of sparse inverse covariance estimation. [12] In general terms, where an inverse covariance estimate is constrained to be sparse, the non-zero terms tend to more accurately relate to correct positive correlations in the true inverse covariance matrix [12] The graphical Lasso is a statistical method which estimates the inverse covariance of the data by minimizing the objective function:... [12] For 44% of the targets, contact prediction was excellent with a precision ¿0.5 for the longest-range top-L/2 predicted contacts (i.e. ¿50% correctly predicted long-range contacts per residue). [12]

(multidimensional) extension of traditional mutual information (MI) can be an additional tool to study covariation [5] as tested with a set of 9 MSAs each containing ¡400 sequences, and was shown to be comparable to that of the newest methods based on maximum entropy/pseudolikelyhood statistical models of protein sequences. [5] METHOD COMPARISSON: However, while all the methods tested detected a similar number of covarying pairs among the residues separated by ¡ 8 A in the reference X-ray structures, there was on average less than 65% overlap between the top scoring pairs detected by methods that are based on different principles. [5] Unfortunately, the reliability of covariation data can be diminished by the existence of correlations originating not just from the direct interactions (physical or functional) between two residues, but also from their shared interaction with one or more other residues, and by the shared phylogenetic history of several homologous proteins in the

MSA. [5] While the performance of these methods has been tested primarily with high quality MSAs containing a very large number of sequences (between 5L and 25L, with L=sequence length), very often investigators are interested in studying the covarying positions of proteins for which the available MSA contains less than L sequences, and whose alignment quality is not optimal due to the presence of many (or large) gaps, [5] METHOD: [5] We can consider a more complicated case including a third channel (column). In this case, I(X1;X3;X2) between the three variables represents the 'interaction information' for a channel with two discrete inputs X1 and X3 and a single discrete output X2 (a 2-way channel). [5] If we are interested in 'explaining out' the effect of X3 on the transmission between X1 and X2, we can take a sum of the mutual information I(X1;X2) for each possible value x3 of X3, weighted by the probability of occurrence (px3) of each of those values: [5] Averaging over all values of X3 (a 3rd column) in an MSA we obtain for the 3-dimensional MI between any two columns (X1 and X2): [5] LIMITATIONS: Due to the long execution times and large memory requirements (growing with the 4th power of the sequence length) of $4D_MI$ only the removal of 3rd order indirect coupling ($3D_MI$) is practical with desktop computers for MSAs of sequences longer than 200 residues. [5] COMPARISSON: [5] We have evaluated the performance of standard MI ($2D_MI$), $3D_MI$, $4D_MI$, PSICOV [14], plmDCA [17], GREMLIN [18], and Hopfield-Potts DCA with Principal Component Analysis [19] (called here hpPCA) with the MSAs of 9 protein families [5] These MSAs contain less than 400 sequences with ratios of sequence number to sequence length (called here the 'L ratio') between 0.4 and 2.0, and thus represent a particularly sensitive test for the performance of the different methods with less than optimal size MSAs. [5] all the methods tested produced covariation maps that closely resembled the contact maps derived from the representative X-ray structures of each family [5] While all the methods used in this study performed quite well in terms of percentage of close contacts recognized among the top covarying pairs, they did not necessarily recognize the same close contacts, as no more than 50% of all the pairs were shared between the MI/mdMI based methods and the other methods [5] Finally, since there is ¡ 65% overlap among the sets of covarying residues identified by algorithms based on different principles, further improvement in accuracy is likely to be obtained by selecting only the shared pairs or by averaging the results from different methods. [5]

How, then, do non-conserved positions change during evolution? It is believed that mutations in these positions can occur because they are either accompanied or preceded by compensatory changes in other variable positions (Fitch et al., 1970; Yanofsky et al., 1964). [8] In the context of multiple sequence alignments, MI is an attractive metric because it explicitly measures the dependence of one position on another, but its usefulness has been limited by three factors. [8] 1) First, positions with higher variability, or entropy, will tend to have higher levels of both random and nonrandom MI than positions of lower entropy (Fodor and Aldrich, 2004a; Martin et al., 2005), even though the latter are more constrained and would seem more likely to depend on neighboring positions. [8] 2) Second, random MI arises because the alignments do not contain enough sequences for background noise to be negligible; our previous modeling studies showed that alignments should contain at least 125 sequences before the random signal begins to subside relative to non-random MI (Martin et al., 2005). [8] 3) A third complicating factor is that all position pairs have MI due to the phylogenetic relationships of the organisms represented in the alignment (Wollenberg and Atchley, 2000). This latter source may be limited to some degree by excluding highly similar sequences from closely related species from the alignment, but cannot be eliminated (Martin et al., 2005; Tillier and Lui, 2003). Each of these sources of MI will tend to obscure the desired signal based on the structural or functional relationships of positions. [8] METHOD: [8] MI measures the reduction of uncertainty about one position given information about the other [8] Thus the challenge is to separate the signal caused by structural and functional constraints, MIsf, from the background, MIb, which is the sum of contributions from random noise and shared ancestry. [8] we postulated that each position in a multiple sequence alignment may have a particular propensity toward MIb, that is related to its entropy and phylogenetic history, and that the MIb between any two positions is the product of their propensities. It then follows that MIb for positions a and b may be expressed as the product of the average MIb values of positions a and b with all other positions in the set, divided by the average MIb of all positions in the set. We call this term the average product correction, (APC), [8] We determined how different a given covariance value was relative to all other values in the data set. The mean and SD of the values determined by each of the algorithms were calculated for all pairs of positions. The number of SD from the mean, i.e. the Z-score, was determined for each value or for each corrected value in a given data set [8] A number of obstacles, including random noise, the influence of entropy, the phylogenetic history and the number of sequences required, complicate the identification of coevolving positions in multiple sequence alignments when using MI [8] We have taken a different approach and developed a correction that rapidly and accurately estimates the background MI found in protein family multiple sequence alignments. Our method was initially based on the assumptions that the coevolution signal between pairs of unrelated positions is derived from random noise or from shared ancestry but not from structural or functional constraints; [8] We have shown that the APC accurately estimates MI in the absence of structural or functional relationships. Furthermore, in real protein alignments the subtraction of the APC from MI results in a metric, MIp, that is independent of the entropy of the positions, and that provides a significant improvement over previously published methods in identifying co-evolving positions that are proximal in protein structure. [8] We have also mathematically demonstrated the validity of the APC correction. [8]

1.1.6 Method comparisson

However, it is not clear to what extent these different methods overlap, and if any of the methods have higher predictive potential compared to others when it comes to, in particular, the identification of catalytic residues (CR) in proteins. [23] The importance of a particular residue in a protein can be due to many different factors, including structural stability, proteinprotein interaction, protein-DNA/RNA interaction, ligand binding site and maintenance of protein functions. [23] In most cases, it is difficult to assign a particular function to a particular residue or group of residues, as function is determined by a subtle interplay between multiple residues and mutation to any of them might impact the protein function and/or structure [23] Three clear signals of evolution are: conservation, conservation within specific groups of sequences sharing a common function, and coevolution between residues (see Figure 1) [23] -1) Conservation is straightforward to calculate and interpret. A change in a conserved position (even when proteins are highly diverse) should have a deleterious effect on the protein function. [23] -2) Specificity determining positions (SDPs) are those positions within multiple sequence alignments (MSAs) that are conserved within groups of proteins that perform the same function (specificity groups) and varying between groups with different functions/specificities. These sites generally determine protein specificity either by binding specific substrate/inhibitor or through interaction with other protein [2-4]. [23] -3) The degree of co-evolution between pairs of residues is commonly estimated using a measure of mutual information (MI) [23] Several methods to predict specificity-determining positions have been developed. Many of these require a previous classification of the proteins into functional groups [3,5,6], which is a problematic limitation since the specificity of a given protein is unavailable in the great majority of cases and is non-trivial to calculate and validate. [23] Here, we aim at addressing this question by comparing the ability to identify catalytic residues (CR) in enzymatic proteins of different information-based methods [23] DATA: The analysis is based on a set of 424 enzymatic Pfam families earlier described by Marino Buslje (2010) [23] Given this data set, we calculated measures related to evolution for the different methods included in the benchmark, and next analyzed the overlap/correlation between these measures and their predictive potential for identification of CR in proteins. [23] RESULTS: Methods for prediction of SDPs aim at estimating a score that correlates with the functional importance of a given residue in terms of protein specificity. [23] From Figure 2, it is clear that the methods for SDP identification (ivET, SDPfox and XDET) show limited mutual overlap. The correlations values are low for all comparisons, with the highest value of 0.34 being between SDPfox and XDET. [23] We next analyzed the correlation between methods aimed to rank the residues by functional importance [23] From our results, we find that the methods included in the benchmark can be divided in three groups with limited mutual overlap. [23] -1) One group consists of methods which predictive signal is strongly correlated to sequence conservation (rvET, and sequence conservation itself), [23] -2) one group consists of the methods whose predictive signal is derived from mutual information (cMI), [23] -3) and the last group consists of the methods developed for prediction of specificity determining positions (SDPfox, XDET and ivET). [23] CONCLUSION: we find that only methods from the first two of the above three groups displayed a reliable predictive performance (mean AUC value above 0.8), indicating that the methods from the SDP group has limited value for the identification of residues critical for protein function. [23]

References

[1] Lukas Burger and Erik van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, 2010.

[2] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):0–1, 2012.

[3] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Toxicogenomics in non-mammalian species*, page 92, 2012.

[4] Pablo Cingolani, Rob Sladek, and Mathieu Blanchette. Bigdatascript: a scripting language for data pipelines. *Bioinformatics*, 31(1):10–16, 2015.

[5] Greg W Clark, Sharon H Ackerman, Elisabeth R Tillier, and Domenico L Gatti. Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. *BMC bioinformatics*, 15(1):157, 2014.

[6] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[7] Russell J Dickson, Lindi M Wahl, Andrew D Fernandes, and Gregory B Gloor. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *PloS one*, 5(6):e11082, 2010.

[8] Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[9] Mario A Fares and Simon AA Travers. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, 173(1):9–23, 2006.

[10] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[11] Chern-Sing Goh, Andrew A Bogan, Marcin Joachimiak, Dirk Walther, and Fred E Cohen. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2):283–293, 2000.

[12] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[13] Nuri Kodaman, Alvaro Pazos, Barbara G Schneider, M Blanca Piazuelo, Robertino Mera, Rafal S Sobota, Liviu A Sicinschi, Carrie L Shaffer, Judith Romero-Gallo, Thibaut de Sablet, et al. Human and helicobacter pylori coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences*, 111(4):1455–1460, 2014.

[14] Alan Lapedes, Bertrand Giraud, and Christopher Jarzynski. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv preprint arXiv:1207.2484*, 2012.

[15] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[16] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[17] Noah Ollikainen and Tanja Kortemme. Computational protein design quantifies structural constraints on amino acid covariation. *PLoS computational biology*, 9(11):e1003313, 2013.

[18] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511–523, 1997.

[19] Wei Qian, Hang Zhou, and Kun Tang. Recent coselection in human populations revealed by protein–protein interaction network. *Genome biology and evolution*, 7(1):136–153, 2015.

[20] Arun K Ramani and Edward M Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of molecular biology*, 327(1):273–284, 2003.

[21] Rori V Rohlfs, Willie J Swanson, and Bruce S Weir. Detecting coevolution through allelic association between physically unlinked loci. *The American Journal of Human Genetics*, 86(5):674–685, 2010.

[22] Mario X Ruiz-González and Mario A Fares. Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system groes-l. *BMC evolutionary biology*, 13(1):156, 2013.

[23] Elin Teppa, Angela D Wilkins, Morten Nielsen, and Cristina M Buslje. Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. implication for catalytic residue prediction. *BMC bioinformatics*, 13(1):235, 2012.

[24] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.