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

In a book published in 1859 entitled *"On the origin of species by means of natural selection"* [6], Charles Darwin introduced the concept of co-evolution referring to the coordinated changes occurring in pairs of organisms. In another of his books *"On the various contrivances by which British and foreign orchids are fertilised by insects"*, which was first published in 1862 [7] Darwing further wrote about this concept of providing more detailed examples. By observing the relationship between the size of orchids' corollae and the length of the proboscis of pollinators, Darwin predicted the existence of a new species able to suck from a large spur [8].

Coevolution refers to the coordinated changes occurring in pairs of organisms to improve or refine interactions, this concept was extended to proteins or more generically, any pair of biomolecules which can be within the same organism [8]. The modern use of co-evolution in genetics is often attributed to Dobzhansky's [10] and Elrich's [12] seminal works that were published in 1950 and 1964 respectively. In recent years, much effort has been dedicated to research of coordinated sequence changes in proteins (and genes) were coevolution could be an important and widespread catalyst of fitness optimizitaion [8].

Different allele combinations in co-evolving genes interact confer distinct degrees of fitness. If this fitness difference is large, selection for alleles could maintain allelic association even between unlinked loci [26], thus co-evolving genes are expected to maintain their interaction by pressures favouring compensatory mutations. [26] Under this hypothesis, genetic loci may be invariable due to their functional or structural constraints but these constraints may change subject to mutations in their functional counterpart. [13] In many

cases, selective advantages for a specific allele pair could fixate the optimal allele pair in the population [26].

Proteins have evolved to interact or function in specific molecular complexes and the specificity of these interactions is essential for their function [23]. Consequently, residue contacts constrain the protein sequences to some extent [23]. In other words, sequences form interacting proteins react as a consequence of adaptation, thus it is reasonable to assume that evolution of sequence changes on one of the interacting proteins must be compensated by mutations in the other [23] It should be noted that this relationship between co-evolution and interaction is not symmetrical. While interaction would involve coevolution, coevolution does not imply physical interaction [13]. Co-evolution between clusters of sites not in contact has also been shown [**?**].

Identification of genes showing signs of adaptive evolution can be used in determining functional regions in proteins [13] It has long been suggested that correlations in amino acid changes can be used to infer protein contact, thus aiding to predict tertiary protein structure [21, 1]. A large number of genomes and protein sequences have become available in recent years enabling the analysis of co-evolution by means of statistical inference between columns in multiple sequence alignments of proteins [1, 1], which is a promising technique for predicting contacting residues in the structure. This interdependent changes in amino acids was formulated for the first time by the "covarion model" [14] and applied in multiple sequence alignments of a family of homologue proteins[8].

Correlated mutations suggest compensatory changes between residues likely to be due to direct contact, physical proximity, catalytic action, binding sites, or even maintaining folding stability. Extending these statistical

methods to correlated mutations between pairs of proteins can identify sites of interaction in protein pairs[8].

Coevolution of interacting species, such as symbionts-hosts, predators-prey, and parasites-hosts, is assumed to be manifested by similarities in the phylogenetic trees [8]. This approach has been extended to protein coevolution assumed to be caused by physical interactions and approaches based on protein tree similarity can identify interaction partners, such as ligand-receptor pairs [8]. Proteins and their interaction partners coevolve so that divergent changes in one are complemented at the interface by their interaction partner [16].

1.1.2 Co-evolution examples

In the absence of a clear positive control, identifying gene pairs that is certainly co-evolving is a difficult task [26]. Here we mentions some accepted examples of co-evolution in humans:

- HLA and KIR are two genes located in different chromosomes are a well established interacting immune-response loci their allele frequencies are highly correlated in human populations as one expects under intense allele matching selection [**?**].
- A remarkable phylogenetic trees similarity was observed between ligands (such as insulins and interleukins) and their corresponding receptors. This coevolution is proposed to be required for maintenaning their specific interactions [**?**].
- An alternative method for ligands-receptors co-evolution is based on the N-terminal and C-terminal phosphoglycerate kinase (PGK) which are covalently linked and form an active site at their interface, therefore, they must are inferred to have co-evolved to preserve enzyme function. [16]. Researchers found that chemokines 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. [16]

- An analysis of Hsp90 and GroEL are heat-shock proteins highlighted sites are functionally or structurally important in almost all cases where co-evolution was detected [13]. GroESL is involved in the folding of a wide variety of other proteins with the folding activity mediated by the co-chaperonin GroES [27]. It was recently shown that different overlapping sets of amino acids co-evolve within GroEL and GroES [27].
- Putative interaction in genes mediating sperm-ZP binding in humans (ZP3 and ZP19) mediating gamete recognition are polymorphic among humans and located on different chromosomes was observed [26]
- Helicobacter pylori is the main cause of gastric cancer. Host-pathogen co-evoutionary interaction was completely accounted most of the difference in the severity of gastric lesions in the populations analysed. For instance African H. pylori ancestry was relatively benign in population of African ancestry but was deleterious in individuals with substantial Amerindian ancestry [18], in an example of co-evolution modulating disease risk [18].

Coevolution of interacting proteins is often analysed in large time frames typically based on the evolutionary analysis across different species [24] Genome-wide scans have identified a several candidate loci that underlie local adaptations, which seems surprising given the short evolutionary time since the human divergence which is estimated have happened around $50,000$ to $100,000$ years ago when humans migrated out of Africa[24]. In light of this, it may make sense to analyse co-evolution within human population since within a pathway or a functional subnetwork, multiple genes may change in the same fitness direction at a same evolutionary rate to achieve a common phenotypic

outcome [24]. A study using 1000 Genome [**?**] project data form East Asians, Europeans, and Africans populaitons, researchers found candidate genes having signals of recent positive selection are significantly closer to each other than expected when the information is mapped onto protein-protein interaction (PPI) networks [24]. The methodology also identified known examples such as EGLN1 and EPAS1 (hypoxia-response pathway playing key roles in adaptation to high-altitude) as well as multiple genes in the NRG-ERBB4 (developmental pathway) [24].

1.1.3 Co-Evolution and protein structure

Protein structure prediction from amino acid sequence is one of the ultimate goals in computational biology [1], despite significant efforts the general problem of de novo three-dimensional structure prediction has remained one of the most challenging problems in computational biology [20]. Unfortunately, de-novo proptein structure prediction does not scale since the conformational space grows exponentially with the protein length. It is expect computation of covariation patterns to complement experimental structural biology thus helping to elucidate functional interactions [20] Information of coevolutionary couplings between residues is often used to compute protein three-dimensional structures from amino acid sequences [20]. It has been observed that using information about a protein residue contacts, it is possible to elucidate the fold of the protein [17] Several researchers demonstrated that using co-evolutionary information from multiple sequence alignments greatly helps to deduce which amino acid pairs are close (or in contact) in the three-dimensional structure thus allowing to calculate protein fold with a reasonable accuracy [20]. The underlying idea is that contact information constrains the fold thus reducing the search space.

Protein design.. It has recently been proposed to use co-evolutionary theory in computational protein design methods. Significant similarities were found between the amino acid covariation in natural protein sequences and sequences structures optimized by computational protein design methods [22]. A study [22] using computational protein design to quantify protein structure constraints form amino acid covariation for 40 diverse protein domains show that structural constraints imposed by protein architecture play a dominant role and computational protein design methods could capture these effects [22]. Evolutionary selective pressures on funstion and structure shaped the sequences to be close to optimal for their structures, natural protein sequences provide an excellent test for computational protein design methods [22]. Computational protein design predicts energetically optimal sequences based on protein structure it is expected that highly covarying amino acids pairs in both designed and natural sequences have likely covaried to maintain protein structure [22].

1.1.4 Detecting co-evolution

One of the first attempts to statistical inference of co-evolving pairs was performed by Gobel et. al. in 1994. In their seminal paper they point out that the fact that *"maintenance of protein function and structure constrains the evolution of amino acid sequences[...] can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions"* [15]. They analysed correlations between different positions in a multiple sequence alignment and used such correlations to predict contact maps. In their study of 11 protein families they compares the result with experimentally validated contact maps determined by crystallography, showing that prediction accuracy up to 68%.

Although some methods were base on testing for correlation of phylogenetic distance matrices between gene families [26] the majority of method have focused on extracting co-evolutionary information from multiple sequence alignment.

The promise of developing methods for predicting contacting pairs from sequence information alone was radically different and more applicable than traditional docking methods [23] This lead to the development of multiple methods for detecting correlated changes in multiple sequence alignments with the primary intention of using them to detect protein interfaces in interacting molecules [23], thus facilitating protein structure prediction. It was demonstrated that the correlated sequence information was enough to select the right inter-domain docking solution amongst many alternatives [23].

Correlation and mutual information have been used to asses co-evolution but they do not take into account the evolutionary interdependence between protein residues [13]. Phylogenetic relationships can inflate these co-evolutionary measures, thus one of main limitations of these methods has been their inability to separate phylogenetic linkage from functional and structural co-evolution [13]. Some methods partially correct these effects but requires alignments of at least 125 sequences to remove stochastic noise [**?**].

CAPS [13] compares transition probability scores from blocks substitution matrix (BLOSUM) between two sequences at the sites being analysed for interaction. An alignment-specific BLOSUM matrix is applied depending on the average sequence identity.

Co-evolution between protein sites is estimated by the correlation in the pairwise variability respect to the mean pairwise variability per site [13]. A limitation of this method is that the number of sequences in the alignment may be problematic when sequences are too divergent, since an alignment including

highly divergent sequence groups could show unrealistic pairwise identity level (BLOSUM values are normalized by the time of divergence between sequences to reduce the impact). Another problem common to many MSA-based co-evolutionary methods is that constant amino acid sites, which are very likely to be functionally important, cannot be tested for [13].

Since many of these methods rely so heavily on multiple sequence alignments, it should not be surprising to know that the quality of the input alignment may affect the results. As one example, it is well known that structure-based alignment algorithms may be susceptible to shift error and other systematic errors, thus strong covariation signal can be caused by alignment errors leading to false positive predictions [9]. Phylogeny of the sequences also affects perfromance, since methods work better on large protein families having a wide but homogenously distributed degree of sequence similarity ranging from distant to similar sequences [8]. In a recent study different co-evolutionary methods were applied to different alignments of the same protein family, giving rise to different results and demonstrating that covariation may greatly depend on the quality of the sequence alignment [9]. Even when alignments for the same protein family contained comparable numbers of sequences the number of estimated covarying positions differ significantly [9]. The authors of this analysis demonstrate that contact prediction can be improved by removing alignment errors due to several factors such as partial or otherwise erroneous sequences, the presence of paralogous sequences, and improper structure alignment [9].

1.1.5 Complex models

An important problem when inferring co-evolution is indirect coupling typically occurring when more than two positions show coordinated substitution patterns. Apparent covariation between two positions is the consequence of the evolutionary interdependence and these indirect couplings can make it

difficult to recognize the directly interdependent positions. Imagine a protein sequence of length L with amino acids label by their positions, amino acid at position i (aa_i) is coupled directly with aa_j, and aa_j to aa_k, then aa_i and aa_k will show correlation despite not being directly coupled [29] As opposed to the independence assumption, in a 'global' model treats correlated pairs of residues as dependent on each other thereby minimizing effects of transitivity [20] Since direct couplings are more reliable predictions of physical interactions, approaches that can distinguish direct from indirect couplings have been a topic of intensive study [8]. Global approaches are designed to reach high scores only for amino acid pairs that are likely to be causative of the observed correlations [20] This is well understood phenomena in statistical physics, a typical example of it are long-range order observed in spin systems, where the spins only have short-range direct interactions. [**?**]

First global model. The first model was proposed by Lapedes in 2002 [19] used a Monte Carlo algorithm to infer the simplest probabilistic distribution able to account for the whole network of covariations [8]. He presented a sequence-based probabilistic formalism addressing co-operative effects in interacting positions in proteins assuming that a sequence of length L is a global state of an $L - site$ spint system of twenty states (for twenty amino acids). A global statistical approach based on maximizing entropy under constraints which is known to lead to Boltzmann statistics [20]. Finally the conditional mutual information is calculated using this Boltzmann model which leads to the degree of covariation between residues at two positions factoring out contributions by interaction with the rest of the residues [20]. The amount sequence data is a limiting factors when performing inference of Boltzmann distribution parameters, thus it is usually infeasible to use more than first order distributions [19]. Another limitation is the phylogenetic relatedness of these

sequences, which are not addressed in this algorithm and have the potential to increase accuracy [19].

Direct coupling analysis. A similar approach called direct-coupling analysis (DCA) is based on spin-glass physics uses generalized message-passing techniques to massively parallelize the algorithm implementation [29]. They manage to apply the method to a set consisting of over $2,500$ bacterial genes from a two-component signal transduction system. With global inference robustly identifying residue pairs proximal in space without between sensor kinase (SK) and response regulator (RR) proteins and for homo-interactions in RR proteins. [29] As in [19] an application of the maximum entropy principle yields the Boltzmann distribution which is used to estimate the second order interaction model. In principle higher correlations of 3 or more positions can be included, however dataset size does not allow for going beyond [29] 2-residue correlations. Determining parameters which is the most computationally expensive task is achieved by using a 2-step procedure: i) given a candidate set of model parameters, single and two residue distributions are estimated; ii) the summation over all possible protein sequences would require $O(21^{N-2}N^2)$ steps, so an approximation is performed using MCMC sampling. This last step is the most expensive step and is expected to be very slow for 21-state variables. An message-passing approach is implemented using an efficient semiheuristic approach which reduces the computational complexity to $O(21^2N^4)$ [29]. Once all probability distributions are estimated, gradient descent is used to adjust the coupling strengths maximizing the joint probability of the data since the model is convex, it is guaranteed to converge to a single global maximum [29]. Finally, a quantity called direct information (DI) measures the part of the mutual information of a position pair induced by the direct coupling (intuitively similar to mutual information in a 2-variable

model) [29] Even after all optimizations and parallelizations, the method could not be applied to more than 60 positions in the protein alignment simultaneously [29].

Mean field approximation. DCA has been shown to yield a large number of correctly predicted contacts based on its ability to disentangle direct and indirect correlations, unfortunately the method is computationally expensive and does not scale [21]. Starting with multiple-sequence alignments of a large number of sequences they attempt to mitigate phylogenetic tree biases using a simple sampling correction based on re-weighting sequences with more than 80% [21]. In a nutshell, the method tries to disentangle direct and indirect couplings by inferring a global statistical and least-constrained model whic, as discussed before, is achieved using a maximum-entropy principle leading to a Boltzmann distribution of couplings [21]. The partition function (Z) is then approximated by keeping only linear order term in a Taylor series expansion, obtaining thus the mean-field equations [21]. This approach is based on small-coupling expansion, thus a Taylor expansion around zero, a technique introduced disordered Ising spinglass models with binary variables [21]. A well known result is that the first derivative of the Gibbs potential (i.e. the Legendre transform of the free energy $F = lnZ$) equals thus the average of the coupling term in the Hamiltonian, which simplifies this average calculation since the joint distribution of all variables becomes factorized over the single sites [21] This algorithm based on the mean field approximation speeds up the original DCA algorithm by 10^3 to 10^4 times [21], and can run on alignments up to 500 amino acids per row which is an order of magnitude larger the previous version of DCA based on message passing [21, 29].

PSI-COV. As other methods, PSI-COV starts from a multiple sequence alignment [17], a covariance matrix is calculated by counting how often a given

pair of the 20 amino acids occurs in a particular pair of positions summing over all sequences in the MSA. Since this matrix contains the raw data capturing all residue pair relationships, one can then compute a measure of causative correlations, in the global statistical approaches by taking the inverse of the covariance matrix (a.k.a. precision or concentration matrix) [17, 20]. Assuming that this covariance matrix inverted, the inverse matrix relates to the degree of direct coupling, a well known fact in statistical theory under the assumption of continuous variables Gaussian multivariate distributions [20]. Elements significantly different from zero (off-diagonal) indicate pairs of sites which have strong direct coupling, thus likely to be in direct physical contact [17]. The empirical covariance matrices are actually almost always singular simply because it is unlikely that every amino acid is observed at every site, one of the most powerful techniques to overcome this problem is sparse inverse covariance estimation under Lasso constraints. The authors claim that the non-zero terms tend to more accurately relate to correct correlations in the true inverse covariance matrix [17].

Multidimensional mutual information. In a recent study, and simple extension of mutual information was proposed by considering "additional information channels" corresponding to indirect amino acid dependencies [5]. They define the information $I(X_1; X_3; X_2)$ representing an 'interaction information' for a channel with two inputs X_1 and X_3 and a single output X_2. They then marginalize the effect of the indirect input (X_3) on the transmission between X_1 and X_2 simply by summing mutual information for each possible value X_3 weighted by the probability of occurrence [5]. Similarly they define a four variable model extension in which case the marginalization would be done over two variables (X_3 and X_4). They test and compare their results using a set of 9 MSAs consisting of less than 400 sequences, showing that their

simple extension is comparable to other maximum entropy statistical models [5]. Even thought the method is simple, the marginalization sums impose a heavy computational burden requiring long execution times and large memory footprints making the method impractical only for sequences longer than 200 residues [5].

1.1.6 Co-evolution algorithm complexity and limitations

1.2 REWRITE

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 [26] We ran a similar analysis on a secondary candidate gene pair implicated in maternal-fetal interactions: GHR (MIM 600946) and GH2 (MIM 139240).37 [26] Power: because of computational limitations, we were unable to perform the exact test for larger value of n; [26] 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 [26]

Population: Population structure could also cause allelic association between physically unlinked loci. [26] 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. [26]

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. [25] We introduce the computational method of matrix alignment for finding the optimal alignment between protein family similarity matrices [25] Binding specificities of duplicate genes (paralogs) often diverge, such that new binding specificities are evolved [25] the use of phylogenetic trees to account for the co-evolution of interacting proteins [25] 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 [25] 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. [25] 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. [25] 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. [25] For matrix alignment, MATRIX currently applies a stochastic simulated annealing-based algorithm. [25]

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 [8] ...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. [8] 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 [8] 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. [8] 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. [8]

Indirect correlations: [20] 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). [20] 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 [20] LOCAL MODELS: [20] 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 [20] 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. [20]

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). [11] 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. [11] 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. [11] 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). [11] 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. [11] METHOD: [11] MI measures the reduction of uncertainty about one position given information about the other [11] 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. [11] 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), [11] 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 [11] 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 [11] 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; [11] 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. [11] We have also mathematically demonstrated the validity of the APC correction. [11]

1.2.1 Method LIMITATIONS

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. [28] A major shortcoming of covariance analysis is that correlations between substitution patterns of interacting residues induce secondary correlations between noninteracting residues [21] This problem was subsequently overcome by the direct-coupling analysis (DCA) (16, 17), which aims at disentangling direct from indirect correlations. [21]

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) [21] 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. [21]

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. [21] 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). [21]

Starting with multiple-sequence alignments of a large number of sequences In this study, we implemented a simple sampling correction, by counting sequences with more than 80% identity and reweighting them in the frequency counts. [21]

This algorithm, termed mfDCA [MEAN FIELD], 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. [21]

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]

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 hp-PCA) 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]

1.2.2 Method comparisson

A major shortcoming of covariance analysis is that correlations between substitution patterns of interacting residues induce secondary correlations between noninteracting residues [21]

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. [28] 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 [28] Three clear signals of evolution are: conservation, conservation within specific groups of sequences sharing a common function, and coevolution between residues (see Figure 1) [28] -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. [28] -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]. [28] -3) The degree of co-evolution between pairs of residues is commonly estimated using a measure of mutual information (MI) [28] 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. [28] Here, we aim at addressing this question by comparing the ability to identify catalytic residues (CR) in enzymatic proteins of different information-based methods [28] DATA: The analysis is based on a set of 424 enzymatic Pfam families earlier described by Marino Buslje (2010) [28] 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. [28] 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. [28] 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. [28] We next analyzed the correlation between methods aimed to rank the residues by functional importance [28] From

our results, we find that the methods included in the benchmark can be divided in three groups with limited mutual overlap. [28] -1) One group consists of methods which predictive signal is strongly correlated to sequence conservation (rvET, and sequence conservation itself), [28] -2) one group consists of the methods whose predictive signal is derived from mutual information (cMI), [28] -3) and the last group consists of the methods developed for prediction of specificity determining positions (SDPfox, XDET and ivET). [28] 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. [28]

Correlated substitution patterns between residues of a protein family have been exploited to reveal information on the structures of proteins (1-10). [21] However, such studies require a large number (e.g., the order of 1,000) of homologous yet variable protein sequences. [21]

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]

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] Charles Darwin. On the origin of species by means of natural selection, or. *The Preservation of Favoured Races in the Struggle for Life, London/Die Entstehung der Arten durch natürliche Zuchtwahl, Leipzig oJ*, 1859.

[7] Charles Darwin. *On the various contrivances by which British and foreign orchids are fertilised by insects*. 1877.

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

[9] 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.

[10] Theodosius Dobzhansky. Genetics of natural populations. xix. origin of heterosis through natural selection in populations of drosophila pseudoobscura. *Genetics*, 35(3):288, 1950.

[11] 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.

[12] Paul R Ehrlich and Peter H Raven. Butterflies and plants: a study in coevolution. *Evolution*, pages 586–608, 1964.

[13] 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.

[14] Walter M Fitch and Etan Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical genetics*, 4(5):579–593, 1970.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.