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

1.1 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.

1.2 Coevolution

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"*, first published in 1862 [7] Darwing further explored this concept and providing more detailed examples. By observing the relationship between the size of orchids' corolla 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 originally referred to the coordinated changes occurring in pairs of organisms to improve or refine interactions. Thes concept was extended to pairs 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].

Distinct allele combinations in co-evolving genes interact to confer different 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].

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 some well known examples of co-evolution in humans are worth mentioning:

- HLA and KIR are two genes located in different chromosomes conforming 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 completely accounted for 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]. This is in an example of co-evolution modulating disease risk.

1.2.1 Detecting co-evolution: Independent models

Correlated phylogenetic trees. 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]. Proteins and their interaction partners co-evolve so that divergent changes in one are complemented at the interface by their interaction partner [16] creating similar evolutionary trees. Thus tree similarity approaches can successfully be extended for protein-protein coevolution assumed to be caused by physical interactions. This kind of methods have been shown to be capable of identifying interaction partners, such as ligand-receptor pairs [8].

Evolutionary relationships within protein families can be mined to predict physical interaction specificities [25]. Duplicate genes (paralogs) often diverge in a way such that new binding specificities are evolved, thus the underlying hypothesis is that interacting proteins exhibit coordinated evolution and tend to have similar phylogenetic trees. This was first demonstrated in a study of chemokines and their receptors showing very similar phylogenetic trees [**?**]. Using similarity of phylogenetic trees as a proxy for the co-evolution of interacting proteins [25], a computational method based on matrix alignment can find an optimal alignment between protein family similarity matrices (conceptually equivalent to superimposing phylogenetic trees from the two protein families) [25]. One matrix is shuffled using stochastic simulated annealing-based to make the two matrices maximally agree by minimizing the root mean square difference. Interactions can be predicted by observing equivalent columns proteins heading in the two matrices. [25]

Although some methods based on phylogenetic tree similarity exists, the majority of co-evolutionary methods focuses on analysis of multiple sequence alignment [26].

Correlated mutations. Proteins have evolved to interact or function in specific molecular complexes and the specificity of these interactions is essential for their function, 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]. Furthermore, 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 [14, 21, 1, 8]. 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 protein sequences [1, 1], which has been a fruitful 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]. Statistical methods to find correlated mutations loci between pairs of proteins can identify putative interaction sites in protein pairs [8], but we should keep in mind that correlated mutations suggesting compensatory changes between residues can be due to several factors different than direct contact, such as physical proximity, catalytic action, binding sites, or even maintaining folding stability.

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... [sequence alignments] 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 compare their results with experimentally validated contact maps determined by crystallography, showing that prediction accuracy up to 68%.

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 several 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 (MI) 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 while some studies claim that these require alignments of at least 125 sequences to remove stochastic noise [**?**], other studies suggest that these methods may require in the order of $1,000$ homologous yet variable protein sequences to achieve correct predictions [21].

Phylogenetic correction. Mutual information measures the reduction of uncertainty about one position given information about the other. When used in as a measurement for co-evolution, MI can be confounded by several factors such as: i) structural and functional constraints, and ii) the background sum of contributions from random noise and shared ancestry. In an attempt to MI's "improve signal to noise ratio" by eliminating or minimizing the second factor, a model postulated by [11] tries to factorize these terms in order to estimate a correction. They propose that each amino acid position in the MSA has a propensity toward the background MI (related to its entropy and phylogenetic history) and estimate the joint background MI as the product of their propensities. It follows that a joint background correction term can be approximated as product of the average background MI divided by the average overall MI of all positions in the MSA, they call this term average product correction (APC) [11]. They show that APC is a metric than can accurately estimate MI in the absence of structural or functional relationships (i.e. the null model) [11], which is assumed to be normally distributed thus a p-value can be inferred using a Z-score.

Another method, 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].

Evolutionary timespan. 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 was also able to identify 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]. This shows that sequences from shorter time spans can also be mined for co-evolution.

MSA quality influences predictions. Since many co-evolutionary 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 co-evolutionary methods were applied to different alignments of the same protein family, giving rise to distinct 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 differed significantly [9]. The authors of this analysis demonstrated 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].

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* protein structure prediction does not scale since

the conformational space grows exponentially with the protein length. Contact information can constrain the fold thus significantly reducing the search space. Since covariation patterns can complement experimental structural biology thus helping to elucidate functional interactions [20], information of co-evolutionary couplings between residues are 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]. It is not surprising to know that the vast majority of methods for finding protein co-evolution are designed with the specific aim to generate results useful in the context of protein folding.

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]. Evolutionary selective pressures on function 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]. Similarly, computational protein design predicts energetically optimal sequences based on protein structure, so it is expected that highly covarying amino acids pairs in both designed and natural sequences have covaried to maintain optimal protein structure [22]. A study [22] using computational protein

design to quantify protein structure constraints form amino acid covariation for 40 diverse protein domains, shows that structural constraints imposed by covariation play a dominant role in protein architecture. Computational protein design methods could make use of knowledge form natural co-evolution effects [22].

1.2.2 Detecting co-evolution: Global 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 [28]

As opposed to models using the independence assumption, 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 an intensive area of 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].

Glass spin systems. Global interaction models are well understood 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. [**?**] One of the first global

models for co-evolution was proposed by Lapedes in 2002 [19], who 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 theory addressing co-operative effects in interacting positions in proteins assuming that a sequence of length L is a global state of an *L-site* spin system of twenty states (for twenty amino acids). Then he solved the global statistical formalism 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) was also based on spin-glass physics [28]. In their implementation a generalized message-passing techniques is used to massively parallelize the algorithm implementation [28]. 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 three or more positions can be included, however dataset size does not allow for inferring beyond [28] two-residue model parameters. Determining parameters which is the most computationally expensive task

is achieved by using a two-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 heuristic which reduces the computational complexity to $O(21^2N^4)$ [28]. 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 [28]. 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 two-variable model) [28] Even after all optimizations and parallelizations, the method could not be applied to more than 60 positions in the protein alignment simultaneously [28]. 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. [28] In their test dataset, all the top 10 candidate interactions identified were shown to be true contacts, and were used to create interacting protein complex quite accurately (3 Årmsd) [21].

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]. In a method published by [21], they propose a "mean field" approximation to DCA. They first 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 approximation method also tries to disentangle direct and indirect couplings by inferring a global statistical and least-constrained model which, 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 in disordered Ising spinglass models with binary variables [21]. A well known result is that the first derivative of the Gibbs potential, the Legendre transform of the free energy $F = -ln(Z)$, equals the average of the coupling term in the Hamiltonian. This simplifies this average calculation since the joint distribution of all variables becomes factorized over the single sites [21]. This algorithm speeds up the original DCA implementataion 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, 28].

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 can indeed be 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 a simple extension of mutual information was proposed by considering "additional information channels" corresponding to indirect amino acid dependencies [5]. This is achieved by defining 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. The effect of the indirect input (X_3) on the transmission between X_1 and X_2 can then be marginalized simply by summing mutual information for each possible value X_3 weighted by the probability of occurrence [5]. Similarly a four variable model extension can be defined, in which case the marginalization would be done over two variables (X_3 and X_4). The authors 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 for sequences longer than 200 residues [5].

Bayesian network model. Another attempt to disentangle direct from indirect statistical dependencies between residues assumes that the sequences in a MSA are drawn from unknown joint probability distribution. The model considers pairwise conditional dependencies and factorizes the joint probability by a single other position which the residue depends on, using the conditional probabilities as nuisance parameters that are integrated out when in calculating the likelihood of the alignment. Most notably, the model does not consider only 'the best' way of choosing the dependent position, but rather sums over all possible ways in which dependencies could be chosen [1]. This sum over all spanning trees is a generalization of Kirchhoff's matrix-tree theorem and can be efficiently computed by the Laplacian of the dependency matrix [1].

1.2.3 Algorithm limitations

Residue coevolution was originally detected using correlated amino acid changes in pairs of positions represented by two columns of the MSA. Under the assumption of interdependent amino acid frequencies or similar patterns of amino acid substitutions it can be assessed by a linear correlation, a method that shows a small but significant capability to recover pairs of positions in physical contact [8].

Mutual information was one of the first proposed methods used to detect covarying positions. As opposed to correlationbased methods, mutual information considers

the distribution of each amino acid in the different sequences for a position quantifying whether presence of an amino acid one position can be used to predict presence of an amino acid in the other position. Mutual information does not take into account which amino acids are present, therefore different amino acids are treated just as symbols [8]. MI is an attractive and simple metric because it explicitly measures the dependence of one position on another, but it is limited by factors such as [11]: i) positions with higher entropy (variability), tend to have higher MI than positions of lower entropy (due to both levels of both random and nonrandom factors) even though the latter are more constrained and would seem more likely to be co-evolving [11]; and random MI arises when alignments do not contain enough sequences to reduce noise to signal ratio, it was shown that alignments should contain at least 125 sequences to reduce this effect [**?**, **?**].

Influence of background phylogenetic relationship between sequences in the MSA confounds results and some efforts try to address this by removed certain problematic clades from the MSA. For instance, it has been shown that the effect may be limited to some degree by excluding highly similar sequences (from closely related species) from the alignment [**?**, **?**, **?**, **?**, **?**]. Continuous-time Markov process model for sequence coevolution can model this explicitly and some approaches have been implemented for smallscale studies of coevolution in small protein families, but computational limitations have hindered their usage int largescale studies [8]. Other confounding effect is an uneven representation of protein sequence members and leading to statistical noise as the result of a low number of sequences in the alignment [20].

Indirect correlations arise because if A correlates with and B are in contact with each other and B and C are in contact as well, there is an observed indirect correlation between A and C [20]. Since amino acids often contact many not just one, but many others, these transitive effects tend to form a network. Thus pairs of residues analysed using a simple statistical model (such as correlation or mutual information) may not necessarily be close in space or functionally constrained [20]. Algorithms to overcome this limitation exists, but they are based in global probabilistic models which require parameter estimation of complex distributions, such as the Bolzmann distribution, as well as marginalizing over all indirect variables. This makes global models computational prohibitively for all but very small datasets and impossible to apply to genome wide scale analysis.

Usually co-evolutionary methods are tested with high quality MSAs containing large number of sequences varying form $5L$ up to $25L$ (where L is sequence length). This is not a realistic case since a proteins that have such large MSAs are often well known and might already have a crystallized structure, thus analysis of amino acids in contact are not needed to infer the 3-D structure. Often investigators study not-so-well-known proteins having MSA of less than L sequences, and low alignment quality due to the presence of many gaps, [5].

Finally it should be mentioned that results from different models usually do not agree, even for complex global models. In a recent study a comparison of several methods shows that while all methods detected similar numbers of covarying pairs (when taking into account residues separated by ≤ 8 Åin reference X-ray structures),

there is less than 65% overlap between the top scoring pairs by methods that based on different principles [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 intramolecular 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] 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.