

Chapter 1

Correlated co-evolving mutations at protein-protein interfaces.

Alexander Schug^{1,2*}

¹ *Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany*

² *Department of Biology, University of Duisburg/ Essen, Germany*

*Corresponding Author: Author; al.schug@fz-juelich.de

1.1. Abstract

Interactions of many biomolecules drive life at the molecular level. The incredible advances in sequencing techniques have led to a vast and rapidly growing treasure trove of genomic data. This wealth of genomic data can be analyzed by advanced statistical methods to accurately infer pairs of residues or contacts which have co-evolved. One such method, direct-coupling analysis (DCA), is based on statistical physics by describing co-evolution via an inverse Potts model. DCA predict inter-residue contacts which can be used as spatial constraints in structure prediction tools to predict, e.g., proteins, RNA, and their

complexes. In this chapter, we will introduce the mathematical framework of DCA, investigate its limits and exemplify applications with a focus on protein complexes.

1.2. Introduction

Life is organized hierarchically from the molecular level, where life is organized by biomolecules such as DNA, proteins, or RNA and their interactions, over cells and their compartments [36] up to organs, organisms or even whole ecosystems. At the molecular level are biomolecules the key players. Despite their simple buildup by nucleic acids or amino acids biomolecules realize an incredible diversity of functions in living organisms. Examples include the storage, handling, and readout of genetic information in DNA, enzymatic function, molecular sensing and signaling, motion (e.g. muscles, cell motion) or structural stability (e.g. collagen, hair, or spider silk). To mechanistically understand biomolecular function, however, one must know the unique biomolecular structure, i.e. the three-dimensional arrangement of all atoms inside the biomolecule. Common experimental techniques used in structure determination have made incredible progress but also have their limits and are often quite involved. One of the oldest and best known method, X-ray crystallography, requires the growth of crystals of the investigated biomolecules and subsequent interpretation of scattering data. Nuclear magnetic resonance (NMR), in contrast, does not require such crystals and can directly be applied to biomolecules in solution but relies on the correct assignment of NMR shifts, which gets increasingly difficult for larger systems. The use of cryo election microscopy (CryoEM) has skyrocketed in the last decade, but still relies on automatized interpretation

of large data sets thus involved highly optimized workflows[63]. Small angle x-ray scattering (SAXS) is experimentally quite simple, but only provides low-resolution information which has to be carefully interpreted[77, 38, 62]. So are there are possible theoretical complements to these experimental approaches?

In-silico protein structure prediction has a long history [44, 33, 34, 65, 55, 10, 83, 45, 43, 21, 53, 1, 29, 81, 70] and can complement experimental work. Commonly summed up under "structure prediction tools", there are many *Ansätze* tackling the challenge of providing biomolecular structures from their sequence alone. Homologue Modelling tools rely on the structural similarity of evolutionary related biomolecules and use experimentally resolved known structures as templates upon which unknown structures can be build. If no evolutionary similar structures are known, one could predict a biomolecular structure from its sequence alone by, e.g., identifying the global free-energy minimum in a suitable physics-based force fields. Such a global search is challenging due to the gigantic search space. Any guidance towards the global minimum would support the search by reducing this search space. In 2009, a methods coined Direct Coupling Analysis (DCA) provided such guidance by investigating the mutational patterns of co-evolution[78] and applied this approach to blind prediction of a protein complex[68]. As highlighted in Fig. 1.2, co-evolving residue pairs are considered spatially adjacent or contacts, as evolution put constraints on mutations by the need to maintaining structure and function. While the general idea was already proposed in the 1990's, earlier methods[31, 54, 47] based on Mutual Information were plagued by high number of false positive contact predictions due to only accounting for strictly pairwise correlations while disregarding the the global context of other residues. DCA[78, 68, 67, 50]

considers this global context and is based on inverse problems in statistical physics, so-called inverse Potts Models. In short, DCA mimics fitness landscapes of proteins and drastically improves signal-to-noise ratios [19, 52, 28, 46]. DCA has inspired similar approaches[48, 3, 40, 25, 49, 57] for tracing co-evolution. In a typical interpretation, such co-evolving residue pairs are considered spatially adjacent contacts and exploited as structural constraints in molecular modeling tools for proteins [68, 48, 71, 17, 39, 56, 73, 57, 58, 75] but also for RNA [20, 79, 60]. Remarkably, the Hamiltonian can be considered a fitness landscape and thus infer biomolecular function such as biological signaling[15], antibiotics resistance [28], or protein/ protein interactions [32, 7].

1.3. Short Introduction into Biomolecular Modeling

A realistic theoretical description of biomolecules based on quantum mechanical (QM) *ab-initio* approaches to accurately model electronic properties and atomic interactions is computationally extremely demanding. Therefore, the most common atomistic description of biomolecules is based on classical or Newtonian force fields and simplifies the QM interactions coarsely into *molecular mechanics*. Typical energy terms are divided into short ranged bonded and long-ranged non-bonded interactions.

Bonded interactions are named by counting the involved number of atoms as 1 – 2 or bond, 1 – 3 or angle, and 1 – 4 or dihedral interactions. The 1 – 2 interaction is a harmonic potential $V_B = \epsilon_B(x - x_0)^2$ (bond constant ϵ_B , distance x

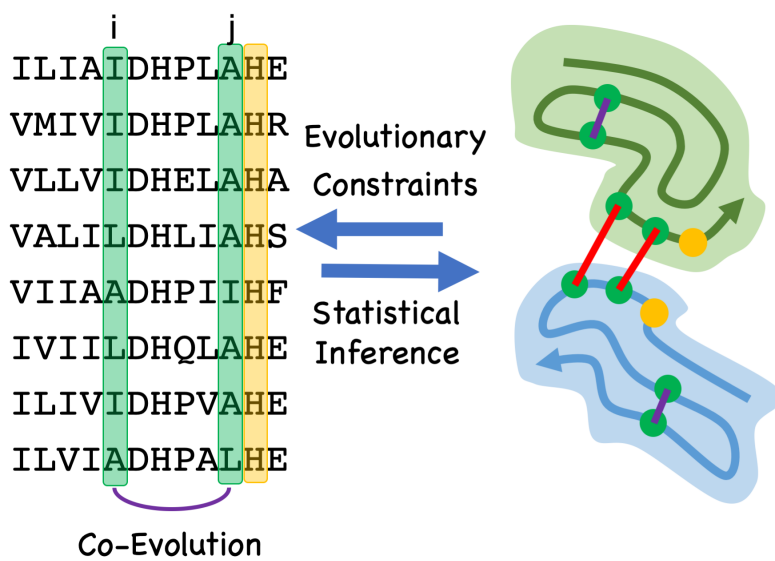


Figure 1.1: Co-Evolutionary Analysis builds on the premise that a biomolecular 3d structure leaves an evolutionary imprint on the sequences of a protein family or in both protein families that form a complex, as a mutation at site i affects mutations at the spatially adjacent sites j . Statistical analysis can therefore infer such pairs of co-evolving residues, both intra-molecular (purple) or inter-molecular (red). Specific functionally relevant residues (orange) are conserved across a protein family and will not show co-evolutionary signals.

of involved atoms 1, 2 and their equilibrium distance x_0). The 1 – 3-interaction is also harmonic $V_A = \epsilon_A(\theta - \theta_0)^2$ (angle constant ϵ_A , angle between bonds of atoms (1,2) and (2,3), θ_0 equilibrium angle). The 1 – 4-interaction is provided by $V_D = \sum_{z=1,3} \epsilon_{z,D} (1 - \cos n(\phi - \phi_0))$ (dihedral constant $\epsilon_{z,D}$, ϕ the angle or dihedral between the respective planes formed by atoms (1,2,3) and (2,3,4), equilibrium dihedral θ_0 and the multiplicity n).

In addition, there are typically two types of *non-bonded interactions*. The short-ranged *Lennard-Jones* potential can be written as $V_{LJ} = \epsilon_{LJ} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma}{r_{ij}} \right)^6 \right]$ (ϵ_{LJ} the potential strength, σ the equilibrium distance, r_{ij} inter-atomic distance of atoms i and j). Finally, the *electrostatics* term represents interactions resulting from two point charges $V_{ES} = \epsilon_{ES} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_R r_{ij}}$ (ϵ_{ES} potential strength, q_i point charge of atom i , ϵ_0 electric constant, ϵ_R dielectric constant, r_{ij} distance between charged atoms i and j).

The total sum of all terms for all atoms is the molecular mechanics potential or *force field*. Some common force fields for biomolecular simulations are AMBER[12] or CHARMM[8]. Given the importance of water to biomolecules [35] and their interactions[27, 2], the solvent interactions have to be modeled as well, either explicitly or implicitly [76, 22, 84, 59, 16, 13, 14, 30]. In structure prediction, the global minimum of the potential should represent the native fold and can be identified by, e.g., stochastic global optimization methods such as simulated annealing and its variants for this task[9, 37, 65, 64, 66, 18].

1.4. Statistical Inference of Coevolution

1.4.1. Limitations of local statistical inference

Protein interactions are the main actuator in biological signaling. Proteins need to interact specifically to prevent unwanted cross-talk, interact sufficiently strong to accommodate transfer of signaling molecules, and, at the same time, interact sufficiently weak or transient to allow disassociation after functional interactions. The main elements of protein interactions is the interaction interface being stabilized by the properties of the involved amino acids. If specific amino acids enable chemical functions (e.g. Kinases), these amino acids tend to be conserved in evolution. All other involved amino acids can more freely mutate in evolution but are still constraint by the need to maintain the overall interacting interface.

These general considerations led to the development of statistical methods to infer such mutational constraints, e.g. by scoring substitution patterns [31] or comparing single $f_i(\alpha)$ and pairwise occurring amino acid frequencies $f_{ij}(\alpha, \beta)$ ($\alpha, \beta \in \{1, \dots, q\}$) are typically the q naturally occurring amino acids plus gap), [47, 42, 80]. One can calculate the f_i, f_{ij} out of a multiple sequence alignment (MSA) for a protein family or out of a joint MSA for a complex. The sequences of such a protein family in a MSA are assumed undergoing selective pressure. Commonly, Mutual Information is then used to quantify co-evolution of sites i, j

$$\text{MI}_{ij} = \sum_{\alpha, \beta \in \{1, \dots, q\}} f_{ij}(\alpha, \beta) \ln \left(\frac{f_{ij}(\alpha, \beta)}{f_i(\alpha) f_j(\beta)} \right) \quad (1.1)$$

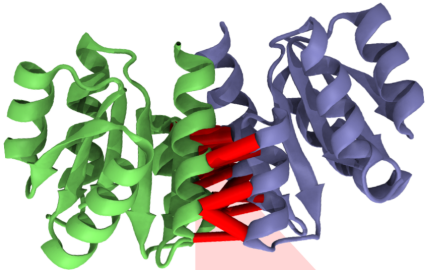
with the the sum running over all the possible amino acids. Here, high values of MI correlate with biological function, but MI is plagued by high numbers of false positive signals when interpreted, e.g., as spatial adjacency when above a threshold. How can be improve this statistical analysis?

1.4.2. Direct Coupling Analysis- a Potts model based on multiple sequence alignments

Direct Coupling Analysis [78, 68, 50] frames co-evolution as an inverse problem based on statistical mechanics (cf. Fig 1.4.2. As above, the sequences in a protein family as found in the MSA are assumed undergoing selective pressure. Thus, a MSA should allow inferring the evolutionary dynamics based on the the marginal distributions of single sites and pairs by, e.g., a maximum-entropy approach to derive a Boltzmann-type distribution.

$[q] = \{i \in \mathbb{N} | 1 \leq i \leq q\}$ is a q -letter alphabet of amino acids (or nucleic acids for RNA) plus the gap position in a MSA. L -tuples formed from $[q]$ provide protein sequences $\sigma = \{\sigma_\nu\}_{\nu=1}^L$, where σ_ν is the amino acid in position ν for a protein of length L . An MSA is viewed as a random sampling of possible sequencess σ of the entire protein family Γ (i.e. Γ are the set of possible L -tuples formed from $[q]$) and we want to infer the probability distribution. According to the maximum-entropy principle, the distribution P that best represents the data given prior knowledge maximizes the entropy function

$$S(P) = - \sum_{\sigma \in \Gamma} P(\sigma) \ln P(\sigma) \quad (1.2)$$



$$H(\sigma) = \sum_{1 \leq i < j \leq L} e_{ij}(\sigma_i \sigma_j) + \sum_{1 \leq i \leq L} h_i(\sigma_i)$$

Fitness and mutational landscapes

Calculation of Direct Contacts

structure prediction

Figure 1.2: Once the inverse problem is solved, the DCA Hamiltonian can be interpreted. In the context of structure prediction, typically the coupling parameters e_{ij} are projected on a scalar such as the direction interaction score and high values interpreted as spatial adjacency of the involved residues i and j . The entire Hamiltonian can also be interpreted as a fitness landscape.

The distribution maximizing the entropy is of the form

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp \{-\mathcal{H}(\boldsymbol{\sigma})\} \quad (1.3)$$

with the Hamiltonian function $\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{k=1}^N \lambda_k g_k(\boldsymbol{\sigma})$, the Lagrange multipliers $\{\lambda_k\}_{k=1}^N$, and the partition function $Z = \sum_{\boldsymbol{\sigma}} e^{-\mathcal{H}(\boldsymbol{\sigma})}$. We now need to find the Lagrange multipliers best describing our data.

In DCA[78, 68, 50], we assume the pair-wise coupling i, j (e.g. stabilizing interactions) and single-site i behavior (e.g. active sites) to be reflected in the MSA:

$$\langle \delta_{\sigma_i, \alpha} \rangle = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma} | \sigma_i = \alpha) \quad (1.4)$$

$$\langle \delta_{\sigma_i, \alpha} \delta_{\sigma_j, \beta} \rangle = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma} | \sigma_i = \alpha, \sigma_j = \beta) \quad (1.5)$$

Ignoring numerical stability[78, 68, 50], marginal probabilities can be estimated from the MSA by direct frequency counts of the single sites $f_i(\alpha)$ and pairs $f_{ij}(\alpha, \beta)$ and we arrive at the Hamiltonian:

$$\mathcal{H}(\boldsymbol{\sigma}) = - \sum_{1 \leq i \leq L} h_i(\sigma_i) - \sum_{1 \leq i < j \leq L} e_{ij}(\sigma_i, \sigma_j) \quad (1.6)$$

The matrix e of pairwise interactions is called the couplings matrix and the single site components h_i coupling. In statistical physics, this mathematical description is called a Potts model, a generalized model of the Ising model. The Potts model has $\binom{N}{2}q^2 + Nq$ inferred parameters, or $\binom{N}{2}(q-1)^2 + N(q-1)$ non-redundant constraints upon normalization, i.e. couplings and local fields are not uniquely defined. A common constraint is to impose gauge-fixation to reduce the parameter space.

1.5. Solving the Inverse Potts Model

Inferring the Hamiltonian from available sequence data requires solving an inverse problem, which was for biological sequence data first solved by Weigt and co-workers[78, 68] by DCA. Due to the finite nature of any MSA and the fact, that sequences in the MSA are not a random subsample of possible sequences¹, we can only approximately solve the inverse problem. It is also common to improve numerical robustness by including pseudo-count corrections λ in the restrictions of the marginals[78, 68]:

$$f_\nu(\alpha) = \frac{1}{\lambda q + M} \left[\lambda + \sum_{a=1}^M \delta_{\sigma_\nu, \alpha} \right] \quad (1.7)$$

$$f_{\nu\xi}(\alpha, \beta) = \frac{1}{\lambda q + M} \left[\frac{\lambda}{q} + \sum_{a=1}^M \delta_{\sigma_\nu, \alpha} \delta_{\sigma_\xi, \beta} \right] \quad (1.8)$$

(M is the size in sequences of the MSA) with minimal impact for large sample size ($M \gg \lambda q$), sampling re-weighting [50] or other refinements of input data. Several Approaches have been developed to solve the inverse statistical problem. The original Message-Passing DCA [78, 68] is a based on susceptibility propagation and computationally quite expensive as it scales as $O(L^4 q^2)$. An improvement was Mean-Field DCA (mfDCA) [50], which considerably lowered computational cost. In mfDCA the DCA Hamiltonian is decomposed into a non-interacting part $\mathcal{H}_0(\boldsymbol{\sigma}) = -\sum_{1 \leq i \leq L} h_i(\sigma_i)$ and a couplings sector $\mathcal{H}(\boldsymbol{\sigma}) = \mathcal{H}_0(\boldsymbol{\sigma}) + \Delta\mathcal{H}(\boldsymbol{\sigma})$. Then, one introduces a trial non-interacting

¹Databases tend to focus on sequences which are either experimentally easy to access as they are from common bacteria or are of particular medical relevance. This leads to phylogenetic and other biases in the sequence data.

Hamiltonian $\mathcal{H}_0(\boldsymbol{\sigma}) + \langle \Delta \mathcal{H}(\boldsymbol{\sigma}) \rangle_0$ (here $\langle X \rangle$ stands for the average of X over the canonical ensemble defined by the non-interacting part). Due to the Bogoliubov inequality $\mathcal{F} \leq \mathcal{F}_0 + \langle \Delta \mathcal{H}(\boldsymbol{\sigma}) \rangle_0$ mfDCA optimizes the local fields to ensure that the trial non-interacting model approximates the closest free energy to the actual system. In the first mfDCA [50] *Ursell functions* are calculated from the empirical frequencies and the corresponding matrix are inverted to recovered the mean-field couplings.

While mfDCA is computationally quite efficient, a subsequent approach is based on Pseudo-Likelihood Maximization (plmDCA)[6, 3]. To infer the values of couplings and of single site fields, the likelihood is substituted by the product of conditional probabilities of observing the variable σ_i^n given the ensemble of all the others $(\sigma_1^n \dots \sigma_{i-1}^n \sigma_{i+1}^n \dots \sigma_L^n)$. In plmDCA, a maximization step proves the computational bottleneck and different gradient descend algorithms are able to tackle this challenge. The most common one is the limited-memory BFGS [11] used as default by plm-DCA implementations [79, 69, 41, 26, 82]. The large redundancy of parameters is solved by regularization [4]. A l_1 -block regularization has been firstly employed in [5]. Many plm-DCA implementation use a l_2 -regularization by adding $l_2 = \lambda_h \sum_{i=1}^N \|\mathbf{h}_i\|^2 + \lambda_J \sum_{1 \leq i \leq j \leq N} \|\mathbf{e}_{ij}\|^2$ to the pseudo-likelihood, which leads to a Ising-type gauge [26].

Finally, one typically takes the coupling matrices e_{ij} and condenses them into a scalar for scoring, e.g. by the Frobenius norm [25, 69, 79, 41]:

$$FN_{ij} = \|e_{ij}\| = \sqrt{\sum_{k,l=1}^q e_{ij}(\alpha, \beta)^2} \quad (1.9)$$

often combined with a Averaged Product Correction [24] $APC_{ij} = FN_{ij} - \frac{\sum_i FN_{ij} \sum_j FN_{ij}}{\sum_i \sum_j FN_{ij}}$. The highest scores pairs of residues are then assumed to be

spatially adjacent. Alternative scores such as Direct Information [78] exist. Typically, for proteins plmDCA provides higher accuracies than mfDCA at elevated computational costs, while for RNA both plmDCA and mfDCA provide similar results[60, 61].

1.6. Contact guided protein and RNA structure prediction

Experimental measurement of protein and RNA 3d structures is often quite involved while the sequence databases grow exponentially and can be exploited by DCA. Here, one typically condenses the coupling matrices into a scalar S_{ij} (see above) and ranks or sorts them by value. These top ranked site-pairs are then inserted as distance constraints into molecular modeling tools to predict protein [68, 48, 71, 39, 17, 51, 72] or RNA [20, 79, 61] systems. Specific examples include all-atom models of globular proteins [48], membrane proteins [39, 56], proteins with multiple conformations [17, 39, 51], structural pattern in disordered proteins [74], or combined with nuclear magnetic resonance (NMR) data [72]. For RNA, DCA has improved both secondary and tertiary structure prediction [20], which was quickly corroborated [79].

But what are the challenges? One big challenge is building a high quality MSA, as one needs to account for phylogenetic bias, non-random sampling of sequence space, etc. Also, while the top scoring contacts tend to be correct or true positive (TP), lower scoring contacts are more likely to be false negatives and only a fraction of all contacts can be predicted with good signal-to-noise ratios. Typically, the top L or $2L$ contacts are used. Lastly, the integration of

predicted contacts into molecular modeling software is not unique and needs to be error tolerant. Also, the used modeling force fields are not perfect, i.e. the lowest energy might not be the native state of a protein.

Intermonomer interaction and signaling

Residue pair coevolution occurs also at inter-protein interfaces. Here, one typically performs a DCA analysis of possible contacts between the interacting proteins. These contacts are then used as constraints in docking the interacting proteins. The abundance of sequence data makes two-component signal transduction system (TCS) a common target of coevolutionary analysis [78, 68, 50, 15, 7]. In fact, the first application of DCA was a blind prediction of a specific TCS [68, 67]. As predicting TCS exemplifies the general approach for predicting protein complexes nicely, I will quickly go over the crucial steps in the last study. TCS are ubiquitous signal transduction systems in bacteria, hence even in 2009 there were many sequences available. To study this heterodimer, it was necessary to build a concatenated MSA data of the interacting protein partners. Due to the possible presence of paralogs, the identification of the correct non-crosstalking interacting partner is challenging. Luckily for TCS, the two interacting partners sensor histidine kinase (HK) and response regulator (RR), can be found adjacently within the same operon- which greatly simplifies building the common MSA. The HK receives an extracellular signal which affect its autophosphorylation rate. The chemical signal, i.e. the phosphoryl group, is then transduced between a highly conserved His of the HK and an Asp of the RR. This conserved His-Asp pair is invisible to DCA due to its inmutatbility but provides an additional spatial constraint for docking

HK and RR. Taking the DCA contacts at the HK-RR interfaces and this additional contacts, it was possible to blindly predict the TCS complex within about 3.5\AA of an independently measured crystal structure[68].

For other classes of protein complex there are other challenges. The difficulty to build a joint MSA is greatly diminished for homodimeric complexes as a protein interacts with itself. Here, the challenge is to distinguish between inter-monomeric and intra-monomeric contacts, as it is unknown how co-evolving contact pairs interact and could be formed within each monomer, between the copies of the monomer or even both within and between. Also, contacts could only be formed in additional conformations, e.g. in conformational transitions or even in domain-swapping. One possibility to address this challenge is assuming that intra-monomeric contacts are not formed at the interface. Thus, one can simply rank all contacts by their score and exclude all contacts already formed in the (typically known) monomer. The remaining contacts can then be formed at the interface [57, 23, 75]. A problem with this approach is the large number of false positive contact predictions at the interface[75], which needs to be addressed by the modeling tools. A large scale study of ≈ 2000 homodimers [75] systematically identified several main results for homodimeric interfaces.

- Higher quality MSAs lead to significantly improved signal-to-noise ratios. Large protein families from the database could contain sub-families with different binding modes, strongly distorting the statistical analysis.
- Larger interacting surface regions are better detected by DCA. Smaller interacting surface regions are more difficult to detect. This is not trivial, as one could assume that smaller interacting surfaces have stronger co-evolutionary signals.
- The majority of predicted false positive (FP) contacts in the monomeric

structure are in fact true positive (TP) contacts of the homodimeric interface. Most predicted contacts are thus formed intra, inter, or both supporting the thesis of spatial adjacency contributing strongly to co-evolution.

1.7. Summary

Co-evolutionary analysis is a powerful toolkit to quantify evolutionary effects on biomolecular structures. Physics-driven methods such as DCA can be directly integrated into molecular modeling tools to predict a large variety of structures, ranging from globular proteins to complexes and RNA. Considering the ongoing growth of both sequence data and raw computational power, these and similar methods based on machine learning will continue to impact structural biology and complement advances in the experimental techniques.

Bibliography

- [1] Badri Adhikari. Deepcon: protein contact prediction using dilated convolutional neural networks with dropout. *Bioinformatics*, 36(2):470–477, 2020.
- [2] Mazen Ahmad, Wei Gu, Tihamér Geyer, and Volkhard Helms. Adhesive water networks facilitate binding of protein interfaces. *Nature communications*, 2(1):1–7, 2011.
- [3] Erik Aurell and Magnus Ekeberg. Inverse ising inference using all the data. *Physical review letters*, 108(9):090201, 2012.
- [4] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozin-

- ski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, January 2012. ISSN 1935-8237. doi: 10.1561/22000000015. URL <http://dx.doi.org/10.1561/22000000015>.
- [5] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011. doi: 10.1002/prot.22934. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22934>.
- [6] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195, 1975. ISSN 00390526, 14679884. URL <http://www.jstor.org/stable/2987782>.
- [7] Anne-Florence Bitbol, Robert S Dwyer, Lucy J Colwell, and Ned S Wingreen. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, 113(43):12180–12185, 2016.
- [8] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem*, 4(2):187–217, 1983. ISSN 0192-8651. URL [GotoISI://A1983QP42300010](http://www.intl.jstor.org/stable/41983QP42300010).
- [9] S. P. Brooks and B. J. T. Morgan. Optimization using simulated annealing. *The Statistician*, pages 241–257, 1995.
- [10] Janusz M Bujnicki. Protein-structure prediction by recombination of fragments. *Chembiochem*, 7(1):19–27, 2006.
- [11] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyu Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal*

- of Scientific Computing*, 16:1190–1208, 9 1995. ISSN 1064-8275. doi: 10.1137/0916069.
- [12] D. A. Case, 3rd Cheatham, T. E., T. Darden, H. Gohlke, R. Luo, Jr. Merz, K. M., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16): 1668–88, 2005. ISSN 0192-8651 (Print) 0192-8651 (Linking). doi: 10.1002/jcc.20290. URL <http://www.ncbi.nlm.nih.gov/pubmed/16200636>.
- [13] J. Chen and 3rd Brooks, C. L. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys*, 10(4):471–81, 2008. ISSN 1463-9076 (Print) 1463-9076 (Linking). doi: 10.1039/b714141f. URL <http://www.ncbi.nlm.nih.gov/pubmed/18183310>.
- [14] J. Chen, 3rd Brooks, C. L., and J. Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr Opin Struct Biol*, 18(2):140–8, 2008. ISSN 0959-440X (Print) 0959-440X (Linking). doi: S0959-440X(08)00007-9[pil]10.1016/j.sbi.2008.01.003. URL <http://www.ncbi.nlm.nih.gov/pubmed/18304802>.
- [15] Ryan R Cheng, Faruck Morcos, Herbert Levine, and José N Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, 111(5):E563–E571, 2014.
- [16] M. S. Cheung, A. E. Garcia, and J. N. Onuchic. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proceedings of the National Academy of*

- Sciences of the United States of America*, 99(2):685–690, 2002. ISSN 0027-8424. URL [GotoISI://000173450100029](http://www.ncbi.nlm.nih.gov/pubmed/11911111).
- [17] Angel E Dago, Alexander Schug, Andrea Procaccini, James A Hoch, Martin Weigt, and Hendrik Szurmant. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences*, 109(26):E1733–E1742, 2012.
 - [18] R. Das and D. Baker. Macromolecular modeling with rosetta. *Annu Rev Biochem*, 77:363–82, 2008. ISSN 0066-4154 (Print) 0066-4154 (Linking). doi: 10.1146/annurev.biochem.77.062906.171838. URL <http://www.ncbi.nlm.nih.gov/pubmed/18410248>.
 - [19] David De Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249, 2013.
 - [20] Eleonora De Leonardis, Benjamin Lutz, Sebastian Ratz, Simona Cocco, Rémi Monasson, Alexander Schug, and Martin Weigt. Direct-coupling analysis of nucleotide coevolution facilitates rna secondary and tertiary structure prediction. *Nucleic acids research*, 43(21):10444–10455, 2015.
 - [21] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.
 - [22] S. Donnini, F. Tegeler, G. Groenhof, and H. Grubmuller. Constant ph molecular dynamics in explicit solvent with lambda-dynamics. *J Chem Theory Comput*, 7(6):1962–1978, 2011. ISSN 1549-9626 (Electronic) 1549-9618 (Linking). doi: 10.1021/ct200061r. URL <https://www.ncbi.nlm.nih.gov/pubmed/21687785>.

- [23] Ricardo N Dos Santos, Faruck Morcos, Biman Jana, Adriano D Andri-
copulo, and José N Onuchic. Dimeric interactions and complex formation
using direct coevolutionary couplings. *Scientific reports*, 5:13652, 2015.
- [24] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without
the influence of phylogeny or entropy dramatically improves residue con-
tact prediction. *Bioinformatics*, 24(3):333–340, 12 2007. ISSN 1367-
4803. doi: 10.1093/bioinformatics/btm604. URL [https://doi.org/10.1093/
bioinformatics/btm604](https://doi.org/10.1093/bioinformatics/btm604).
- [25] Magnus Ekeberg, Cecilia Lökvist, Yueheng Lan, Martin Weigt, and Erik
Aurell. Improved contact prediction in proteins: using pseudolikelihoods
to infer potts models. *Physical Review E*, 87(1):012707, 2013.
- [26] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. Fast pseudolike-
lihood maximization for direct-coupling analysis of protein structure
from many homologous amino-acid sequences. *Journal of Computational
Physics*, 276:341 – 356, 2014. ISSN 0021-9991. doi: [https://doi.org/
10.1016/j.jcp.2014.07.024](https://doi.org/10.1016/j.jcp.2014.07.024). URL [http://www.sciencedirect.com/science/
article/pii/S0021999114005178](http://www.sciencedirect.com/science/article/pii/S0021999114005178).
- [27] Susanne Eyrisch and Volkhard Helms. Transient pockets on protein sur-
faces involved in protein- protein interaction. *Journal of medicinal chem-
istry*, 50(15):3457–3464, 2007.
- [28] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenailon,
and Martin Weigt. Coevolutionary landscape inference and the context-
dependence of mutations in beta-lactamase tem-1. *Molecular Biology and
Evolution*, 33(1):268–280, 2015.

- [29] Hiroyuki Fukuda and Kentaro Tomii. Deepeca: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC bioinformatics*, 21(1):1–15, 2020.
- [30] E. Gallicchio and R. M. Levy. Agbnp: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J Comput Chem*, 25(4):479–99, 2004. ISSN 0192-8651 (Print) 0192-8651 (Linking). doi: 10.1002/jcc.10400. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14735568.
- [31] 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.
- [32] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 113(43):12186–12191, 2016.
- [33] Ulrich HE Hansmann, Yuko Okamoto, and Frank Eisenmenger. Molecular dynamics, langevin and hybrid monte carlo simulations in a multicanonical ensemble. *Chemical physics letters*, 259(3-4):321–330, 1996.
- [34] Corey Hardin, Michael P Eastwood, Zaida Luthey-Schulten, and Peter G Wolynes. Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins. *Proceedings of the National Academy of Sciences*, 97(26):14235–14240, 2000.
- [35] Volkhard Helms. Protein dynamics tightly connected to the dynamics of

- surrounding and internal water molecules. *ChemPhysChem*, 8(1):23–33, 2007.
- [36] Volkhard Helms. *Principles of computational cell biology: from protein complexes to cellular networks*. John Wiley & Sons, 2018.
- [37] T. Herges, A. Schug, H. Merlitz, and W. Wenzel. Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. *Nanotechnology*, 14(11):1161–1167, 2003. URL [⟨GotoISI⟩://WOS:000187038400003](#).
- [38] Markus R Hermann and Jochen S Hub. Saxs-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *Journal of chemical theory and computation*, 15(9):5103–5115, 2019.
- [39] Thomas A Hopf, Lucy J Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.
- [40] 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, 2011.
- [41] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

- [42] Itamar Kass and Amnon Horovitz. Mapping pathways of allosteric communication in groel by analysis of correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, 48(4):611–617, 2002. doi: 10.1002/prot.10180. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10180>.
- [43] Michal A Kurowski and Janusz M Bujnicki. Genesilico protein structure prediction meta-server. *Nucleic acids research*, 31(13):3305–3307, 2003.
- [44] Kit Fun Lau and Ken A Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.
- [45] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, volume 487, pages 545–574. Elsevier, 2011.
- [46] Ronald M Levy, Allan Haldane, and William F Flynn. Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current opinion in structural biology*, 43:55–62, 2017.
- [47] Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.
- [48] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure

- computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.
- [49] Mirco Michel, Sikander Hayat, Marcin J Skwark, Chris Sander, Debora S Marks, and Arne Elofsson. Pconsfold: improved contact predictions improve protein models. *Bioinformatics*, 30(17):i482–i488, 2014.
- [50] 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.
- [51] Faruck Morcos, Biman Jana, Terence Hwa, and José N Onuchic. Co-evolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, 2013.
- [52] Faruck Morcos, Nicholas P Schafer, Ryan R Cheng, José N Onuchic, and Peter G Wolynes. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences*, 111(34):12408–12413, 2014.
- [53] John Moult, Krzysztof Fidelis, Andriy Kryshchuk, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction: Progress and new directions in round xi. *Proteins: Structure, Function, and Bioinformatics*, 84:4–14, 2016.
- [54] Erwin Neher. How frequent are correlated changes in families of protein

- sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102, 1994.
- [55] Marilisa Neri, Claudio Anselmi, Michele Cascella, Amos Maritan, and Paolo Carloni. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Physical review letters*, 95(21):218102, 2005.
- [56] Timothy Nugent and David T Jones. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences*, 109(24):E1540–E1547, 2012.
- [57] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030, 2014.
- [58] Sergey Ovchinnikov, Lisa Kinch, Hahnbeom Park, Yuxing Liao, Jimin Pei, David E Kim, Hetunandan Kamisetty, Nick V Grishin, and David Baker. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*, 4:e09248, 2015.
- [59] D. Paschek, H. Nymeyer, and A. E. Garcia. Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J Struct Biol*, 157(3):524–33, 2007. ISSN 1047-8477 (Print) 1047-8477 (Linking). doi: S1047-8477(06)00329-7[pri]10.1016/j.jsb.2006.10.031. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17293125.
- [60] Fabrizio Pucci and Alexander Schug. Shedding light on the dark mat-

- ter of the biomolecular structural universe: Progress in rna 3d structure prediction. *Methods*, 2019.
- [61] Fabrizio Pucci, Mehari B Zerihun, Emanuel K Peter, and Alexander Schug. Evaluating dca-based method performances for rna contact prediction by a well-curated data set. *RNA*, 26(7):794–802, 2020.
- [62] Ines Reinartz, Marie Weiel, and Alexander Schug. Fret dyes significantly affect saxs intensities of proteins. *Israel Journal of Chemistry*, 60(7):725–734, 2020.
- [63] Christine Röder, Tatsiana Kupreichyk, Lothar Gremer, Luisa U Schäfer, Karunakar R Pothula, Raimond BG Ravelli, Dieter Willbold, Wolfgang Hoyer, and Gunnar F Schröder. Cryo-em structure of islet amyloid polypeptide fibrils reveals similarities with amyloid- β fibrils. *Nature structural & molecular biology*, 27(7):660–667, 2020.
- [64] A. Schug and W. Wenzel. Predictive in silico all-atom folding of a four-helix protein with a free-energy model. *Journal of the American Chemical Society*, 126(51):16736–16737, 2004. doi: 10.1021/ja0453681|ISSN0002-7863. URL \langle GotoISI \rangle ://WOS:000225910400026.
- [65] A Schug, T Herges, and W Wenzel. Reproducible protein folding with the stochastic tunneling method. *Physical review letters*, 91(15):158102, 2003.
- [66] A. Schug, W. Wenzel, and U. H. E. Hansmann. Energy landscape paving simulations of the trp-cage protein. *Journal of Chemical Physics*, 122(19), 2005. ISSN 0021-9606. doi: 194711Artn194711. URL \langle GotoISI \rangle ://000229743500056.

- [67] Alexander Schug and José N Onuchic. From protein folding to protein function and biomolecular binding by energy landscape theory. *Current opinion in pharmacology*, 10(6):709–714, 2010.
- [68] Alexander Schug, Martin Weigt, José N Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, 2009.
- [69] Stefan Seemayer, Markus Gruber, and Johannes Söding. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, 07 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu500. URL <https://doi.org/10.1093/bioinformatics/btu500>.
- [70] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Židek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
- [71] Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, and José N Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26):10340–10345, 2012.
- [72] Yuefeng Tang, Yuanpeng Janet Huang, Thomas A Hopf, Chris Sander, Debora S Marks, and Gaetano T Montelione. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nature Methods*, 12(8):751–754, 2015.
- [73] Pengfei Tian, Wouter Boomsma, Yong Wang, Daniel E Otzen, Mogens H

- Jensen, and Kresten Lindorff-Larsen. Structure of a functional amyloid protein subunit computed using sequence variation. *Journal of the American Chemical Society*, 137(1):22–25, 2014.
- [74] Agnes Toth-Petroczy, Perry Palmedo, John Ingraham, Thomas A Hopf, Bonnie Berger, Chris Sander, and Debora S Marks. Structured states of disordered proteins from genomic sequences. *Cell*, 167(1):158–170, 2016.
- [75] Guido Uguzzoni, Shalini John Lovis, Francesco Oteri, Alexander Schug, Hendrik Szurmant, and Martin Weigt. Large-scale identification of co-evolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 114(13):E2662–E2671, 2017.
- [76] J. Wagoner and N. A. Baker. Solvation forces on biomolecular structures: a comparison of explicit solvent and poisson-boltzmann models. *J Comput Chem*, 25(13):1623–9, 2004. ISSN 0192-8651 (Print) 0192-8651 (Linking). doi: 10.1002/jcc.20089. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15264256.
- [77] Marie Weiel, Ines Reinartz, and Alexander Schug. Rapid interpretation of small-angle x-ray scattering data. *PLoS computational biology*, 15(3):e1006900, 2019.
- [78] 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.
- [79] Caleb Weinreb, Adam J Riesselman, John B Ingraham, Torsten Gross,

- Chris Sander, and Debora S Marks. 3d rna and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975, 2016.
- [80] Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. [4] - features of protein–protein interactions in two-component signaling deduced from genomic libraries. In Melvin I. Simon, Brian R. Crane, and Alexandrine Crane, editors, *Two-Component Signaling Systems, Part A*, volume 422 of *Methods in Enzymology*, pages 75 – 101. Academic Press, 2007. doi: [https://doi.org/10.1016/S0076-6879\(06\)22004-4](https://doi.org/10.1016/S0076-6879(06)22004-4). URL <http://www.sciencedirect.com/science/article/pii/S0076687906220044>.
- [81] Qi Wu, Zhenling Peng, Ivan Anishchenko, Qian Cong, David Baker, and Jianyi Yang. Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics*, 36(1):41–48, 06 2019. doi: 10.1093/bioinformatics/btz477. URL <https://doi.org/10.1093/bioinformatics/btz477>.
- [82] Mehari B Zerihun, Fabrizio Pucci, Emanuel K Peter, and Alexander Schug. pydca v1. 0: a comprehensive software for direct coupling analysis of rna and protein sequences. *Bioinformatics*, 36(7):2264–2265, 2020.
- [83] Weihua Zheng, Nicholas P Schafer, Aram Davtyan, Garegin A Papoian, and Peter G Wolynes. Predictive energy landscapes for protein–protein association. *Proceedings of the National Academy of Sciences*, 109(47):19244–19249, 2012.
- [84] P. I. Zhuravlev, S. Wu, D. A. Potoyan, M. Rubinstein, and G. A. Papoian. Computing free energies of protein conformations from explicit solvent simulations. *Methods*, 52(1):115–21, 2010. ISSN 1095-9130 (Electronic)

1046-2023 (Linking). doi: S1046-2023(10)00138-6[pii]10.1016/j.ymeth.
2010.05.003. URL <http://www.ncbi.nlm.nih.gov/pubmed/20493264>.