Biomolecular Structure Prediction via Coevolutionary Analysis: A Guide to the Statistical Framework

M. B. Zerihun, A. Schug

published in

NIC Symposium 2018
K. Binder, M. Müller, A. Trautmann (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 49,
ISBN 978-3-95806-285-6, pp. 15.
http://hdl.handle.net/2128/17544

Biomolecular Structure Prediction via Coevolutionary Analysis: A Guide to the Statistical Framework

Mehari B. Zerihun[1] **and Alexander Schug**[2]

[1] Steinbuch Centre for Computing & Department of Physics,
Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen, Germany

[2] John von Neumann Institute for Computing and Jülich Supercomputing Centre,
Institute for Advanced Simulation, Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: al.schug@fz-juelich.de

On the molecular level, life is orchestrated through an interplay of many biomolecules. To gain any detailed understanding of biomolecular function, one needs to know their structure. Yet despite incredible progress in experimental structure determination techniques, many important biomolecules are still not structurally resolved. An orthogonal theoretical approach are structure prediction techniques which take advantage of constantly growing computational resources. Mostly untapped information of evolutionarily closely related sequences in the exponentially growing genomic databases can be statistically analysed to: (i) accurately infer pairs of residues in spatial contact within biomolecules and (ii) guide the prediction of biomolecular structures when used in combination with molecular modelling techniques. By now, this approach has revolutionised the field of protein structure prediction by providing highly accurate models. The same mathematical framework can also go beyond structure prediction by analysing evolutionary fitness landscapes and inferring biomolecular interactions or epistasis.

1 Introduction

The exponentially growing genomic databases, the "Biomolecular Big Data", provide a constantly growing wealth of data. Hidden within this treasure is nothing less than the encoding of all aspects of life ranging from the molecular level to entire organisms. An successful exploitation of this resource requires bringing together powerful algorithms and fully harnessing the capabilities of the, also exponentially growing, high-performance computing resources. An outstanding success can be found in the field of protein structure prediction, where the combination of innovative statistical analysis of large sequential databases and raw computing power has been successfully leveraged.

This breakthrough is crucial, as biomolecular interactions influence all facets of life including a wide variety of tasks such as oxygen transport, enzymatic function, genetic regulation, signal transduction, and muscle function. Yet to gain any detailed understanding of the function of a specific biomolecular systems, one needs to know its structure. Many biologically important systems, however, are stabilised by weak or transient interactions, which makes their experimental characterisation challenging and frequently of uncertain outcome. So how can one still elucidate these experimentally poorly accessible structures?

The core idea is that biomolecules are not random heteropolymers but sequences that have evolved over long timescales. The evolutionary process of mutation and selection tends to conserve properties such as function, stability and foldability. The biomolecular sequences found in databases therefore do not have random changes in sequences, but will express mutational patterns that maintain these properties to ensure survival of the entire

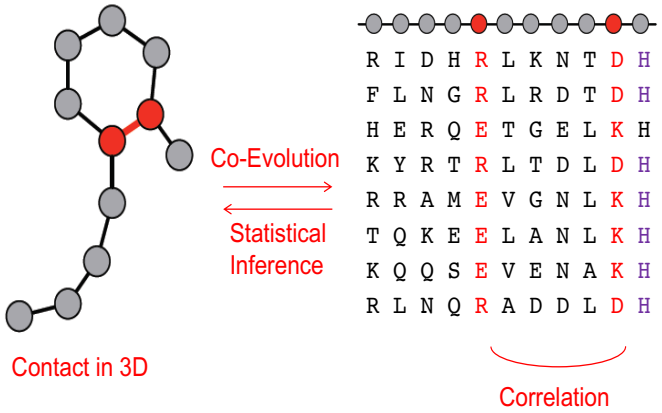

Figure 1. Relationship coevolution and statistical inference of 3d contacts. A three-dimensional contact (red) stabilising a biomolecule leads to correlated coevolving mutations in sequences (each line is the same protein in different organisms; letters on top of each other represent residues at the same position). Given such a sequence alignment, one can infer from correlations found between two positions that these sequence positions are spatially adjacent – they are contacts. Some sequence positions are functionally constrained and do not mutate (purple). Accordingly, these highly conserved residues do not express coevolutionary patterns.

organism (see Fig. 1). For a evolutionary related family of biomolecules, such sequences are gathered in databases such as RFAM or PFAM. The statistical analysis of these sequence families reveals coevolving residue pairs. What is the reason of this coevolution?

A straightforward interpretation is that of spatial adjacency of the involved residues. It turns out, however, that mere correlation measured by, e.g., Mutual Information (MI) does not always mean that the involved residues are in direct (spatial) contact. Instead, such correlations can result both from direct or indirect effects. Direct means that the two residues are proximal to each another, often as a result from physical interactions such as electrostatics. Indirect effects result from transitivity, where spatially distant residues show coevolution resulting from a network of intermediate residues. Direct interactions (DI) are highly useful spatial constraints for biomolecular structure prediction (see Fig. 2). In the context of structure prediction, the quality of the statistical analysis method of residue coevolution therefore lies in its ability to distinguish these two cases. Although these ideas have already been pursued in 1990's[1–3] the first method to reliably identify DI or contacts, Direct coupling analysis (DCA)[4–6], was only recently developed.

Since then, the integration of coevolutionary contact information into molecular modelling has revolutionised the field of protein structure prediction with its long history[7–11] by providing highly accurate models with resolutions comparable to experiments[5,12] in the prediction of tertiary protein structure[13–16], protein complexes[5,17–19], membrane proteins[20,21], conformational transitions[22–25], and, most recently, first RNA structures[26,27]. Apart from the structural insight, the model Hamiltonian can also be interpreted in terms of fitness landscapes which allows to make, e.g., predictions on antibiotics resistance[28], drug design[29], biological signalling[30,31] or epistatic effects[32]. An overview of DCA applications can be found in Ref. 33. We will here explain the basic concepts of the statistical analysis.

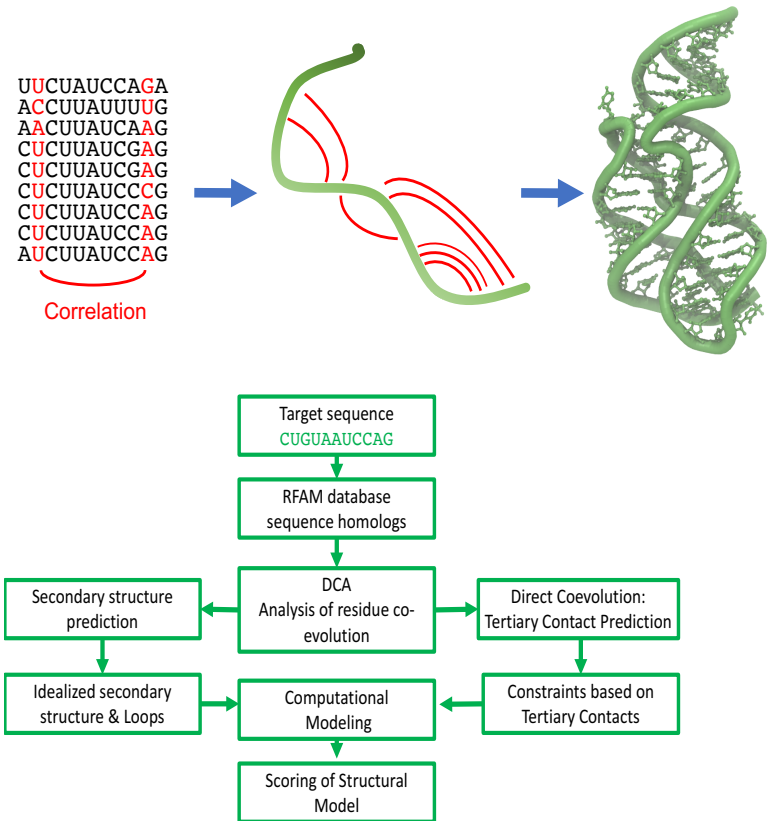

Figure 2. Structure Prediction Workflow on the Examples of RNA. (top) Correlated mutations are identified by aligning the sequences of a biomolecular family and building a global statistical model. The top-scoring contacts indicate spatial proximity of the respective residue pairs. This information is used as constraints to build a 3D model of a biomolecule. (bottom) The detailed structure prediction workflow for a particular target sequence. First, one identifies appropriate sequence homologs in RFAM. The corresponding multiple sequence alignment is analysed by DCA and secondary structure as well as tertiary contacts are predicted. Fragments of the secondary structure elements and tertiary constraints are used to construct an initial prediction model, which then scored.

2 Mathematical Framework of Direct Coupling Analysis (DCA)

Global Probability from Maximum Entropy

The mathematical framework of DCA is based on entropy maximisation subject to constraints[34]. It describes the probability P of sampling a sequence $\sigma = (a_1 a_2 a_3 ... a_L)$ of length L containing a residue or gap a_i at site i. Formally, it is given by

$$\begin{aligned} P(\sigma) &= \frac{1}{Z} \exp\{H(\sigma)\}, \\ H(\sigma) &= \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i), \end{aligned} \tag{1}$$

where H is the sequence dependent dimensionless Hamiltonian, J_{ij} is the *coupling* between sites i and j, h_i is the *local field* or *bias* at site i, and Z is the normalisation constant (i.e. partition function)[4,6]. A site in a sequence can assume any of q possibilities where q is the number of residue types plus a gap. For instance, q is 21 for proteins, and 5 for RNAs.

The model has a total number of $\frac{1}{2}L(L-1)q^2$ parameters for the couplings and Lq for the fields. Not all of them are independent. The requirement that the model to be consistent with single-site marginals, $\sum_{\{a_k,k\neq i\}} P(\sigma) = P_i(a_i)$, and pair-site marginals $\sum_{\{a_k,k\neq i,j\}} P(\sigma) = P_{ij}(a_i,a_j)$ together with the normalisation condition $\sum_{\{a_k\}} P(\sigma) = 1$ reveals the total number of unique parameters to be $\frac{1}{2}L(L-1)(q-1)^2 + L(q-1)$. To overcome over-parametrisation, one can use a *gauge fixation* in either of two forms. One is to use *lattice-gas* gauge which is achieved by setting all the coupling and fields to zero when the last residue q (assuming that the residue types plus gap are numbered in sequence as 1, 2, 3, ..., q) is involved; $J_{ij}(a,q) = J_{ij}(q,b) = h_i(q) = 0$. The other is to use the *zero-sum* gauge which is done by setting the sum of the couplings and fields over all possible states of a site in a sequence to zero; $\sum_{a=1}^{q} J_{ij}(a,b) = \sum_{b=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} h_i(a) = 0$.

Inverse Inference and Parameter Estimation

The parameters of Eq. 1 can be inferred from the given input data. Typically, the input data is a multiple sequence alignment (MSA) of N sequences for a protein or RNA family. From the aligned sequences, single-site and pair-site frequency counts are computed as,

$$\begin{aligned} f_i(a) &= \frac{1}{N}\sum_{n=1}^{N} \delta_{a_i^n,a}, \\ f_{ij}(a,b) &= \frac{1}{N}\sum_{n=1}^{N} (\delta_{a_i^n,a})(\delta_{b_j^n,b}) \end{aligned} \tag{2}$$

with $\delta_{x,y} = 1$ when $x = y$ and 0 otherwise. Then, the marginal probabilities, $P_i(a)$ and $P_{ij}(a,b)$ are forced to be consistent with the respective empirical frequency counts in Eq. 2. However, the computational complexity to estimate the partition function Z scales as $O(q^L)$ which is too complex even for a sequence of average length.

Weigt *et al.*[4] infer the parameters by approximating the marginal probabilities and computing the parameters using an iterative numerical scheme until a desired convergence is achieved. The iterative procedure is consistent with zero-sum gauge. This method called *message passing Direct-Coupling Analysis* (mpDCA) is computationally very costly[4].

Later, a more efficient parameter inference was achieved based on *mean-field* approximation[6]. This method called *mean-field Direct-Coupling Analysis* (mfDCA) computes the couplings by inverting a matrix C whose elements are constructed from the empirical connected correlations $C_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b)$. In mfDCA, lattice-gas gauge is used when a gap is involved. This results in C to become a $L(q-1)$ by $L(q-1)$ matrix in dimension. The couplings are obtained from C; $J_{ij}(a,b) = -(C^{-1})_{ij}(a,b)$. The fields are computed in a self-consistent manner[6].

Another method to learn the parameters of the global probability model is using maximum likelihood estimation. Instead of taking the global probability in Eq. 1, an approxima-

tion is required to make the procedure computationally feasible. The approximate probability is obtained by considering a particular site in a sequence within the MSA in the presence of other sites[35]. This method named *pseudo-likelihood maximisation Direct-Coupling Analysis* (plmDCA) was shown to outperform mfDCA if the dataset in the MSA is sufficiently large[35]. Many competing methods have been developed since, such as Boltzmann-machine learning[36], Gaussian approximation[37,38], adaptive cluster expansion[39], minimal probability flow[40] or Bayesian networks[41] (for a recent review please cf. Ref. 42).

Still there are many challenges[43]. The number of inferred parameters is $O(L^2q^2)$ but many sequence families have insufficient sequence data to infer these parameters reliably. A common strategy to alleviate this problem is regularisation via pseudo-counts in Eq. 2 or a penalised pseudo-likelihood approximation[4,6,35]. Another challenge stems from the origin of sequences. Ideally, they are assumed to be drawn randomly from all possible sequences for a given biomolecular family by Eq. 1. The sequences in genomic databases, however, often overrepresent sequences from biologically important or experimentally easily accessible organisms. This requires reweighting sequences which are too similar[4–6,35].

Mapping Sequence Family Residue-Residue Contacts

The couplings $J_{ij}(a,b)$ quantify the coupling strength between two sites i and j in a sequence when they are occupied by residues a and b. Since there are q^2 couplings associated with any pair of sites in a sequence family, one maps these quantities into a single parameter that quantifies the sequence family's residue-residue contact strength.

One way of mapping the parameters[4] is to define *direct-information*, DI_{ij} between two sites i and j from the *direct-probability*, P_{ij}^{dir} as

$$\begin{aligned} DI_{ij} &= \sum_{a=1}^{q}\sum_{b=1}^{q} P_{ij}^{dir}(a,b)\log\left\{\frac{P_{ij}^{dir}(a,b)}{f_i(a)f_j(b)}\right\}, \\ P_{ij}^{dir}(a,b) &= \frac{1}{Z_{ij}}\exp\left\{J_{ij}(a,b)+\tilde{h}_i(a)+\tilde{h}_j(b)\right\}. \end{aligned} \tag{3}$$

In Eq. 3, the new fields $\tilde{h}_i$ and $\tilde{h}_j$ are computed by requiring consistency with the empirical frequencies as $\sum_b P_{ij}^{dir}(a,b) = f_i(a)$ and $\sum_a P_{ij}^{dir}(a,b) = f_j(b)$ respectively. Z_{ij} is the partition function for the direct-probability. Another way is to compute the Frobenius norm from the q by q matrix formed from the couplings

$$F_{ij} = \sqrt{\sum_{a=1}^{q}\sum_{b=1}^{q}|J_{ij}(a,b)|^2}. \tag{4}$$

The advantage of DI over the Frobenius norm is that the former is independent of type of gauge chosen.

Once the contact information between pairs of sites are computed, they are ranked according to score, with high scoring pairs implying strong coevolutionary signal. Typically, the accuracy of the predicted contacts is evaluated by comparing the proximity of pairs of residues within a known protein data bank (PDB) structure. Similar methods that disentangle direct contacts using sequence information alone are included in Refs. 37, 44–51. Since all the methods are still in their infancy, we can expect new developments, based on rigorous approaches as well as on heuristics, to be proposed soon.

3 Summary and Outlook

Exploiting genomic data by large scale analysis on HPC resources has revolutionised biomolecular structure prediction in the last decade. Invented to improve structure prediction, the integration of coevolutionary contact information into molecular modelling is now used wide-spread and provides accurate models. In addition, novel applications have been found, such as predictions on antibiotics resistance, drug design, or epistatic effects. Currently, major efforts go into complementing existing sequential data by metagenomic data[52] or pushing the frontier from a few "anecdotal" cases to 1000's of systems[53]. Further, biomolecular simulations[5,54] or other efficient techniques[55] can also integrate experimental measurements such as cryoEM[56], sparse NMR data[57] or SHAPE-data[58] to suggest structural models up to atomic resolution. Hybrid approaches using multiple techniques including coevolutionary information could provide structural models not accessible to a single technique. All these applications demand effective use of high performance computing resources and hybrid data integration techniques. Given the current trajectory of the field, both basic biomolecular research and more applied fields such as medicine or pharmacology will benefit from this transdisciplinary research.

References

1. U. Göbel, C. Sander, R. Schneider, and A. Valencia, *Correlated mutations and residue contacts in proteins*, Prot: Struct, Funct, and Bioinf, **18**, no. 4, 309–317, 1994.
2. E. Neher, *How frequent are correlated changes in families of protein sequences?*, Proc Nat Acad Sci USA, **91**, no. 1, 98–102, 1994.
3. S. W. Lockless and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families*, Science, **286**, no. 5438, 295–299, 1999.
4. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, *Identification of direct residue contacts in protein-protein interaction by message passing*, Proc Nat Acad Sci USA, **106**, no. 1, 67–72, 2009.
5. A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, *High-resolution protein complexes from integrating genomic information with molecular simulation*, Proc Nat Acad Sci USA, **106**, no. 52, 22124–22129, 2009.
6. M. Faruck, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, *Direct-coupling analysis of residue coevolution captures native contacts across many protein families*, Proc Nat Acad Sci USA, **108**, no. 49, E1293–301, 2011.
7. U. H. Hansmann and Y. Okamoto, *Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem*, J Comp Chem, **14**, no. 11, 1333–1338, 1993.
8. A. Schug, T. Herges, and W. Wenzel, *Reproducible protein folding with the stochastic tunneling method*, Phys Rev Lett, **91**, no. 15, 158102, 2003.
9. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, *The protein folding problem*, Annu Rev Biophys, **37**, 289–316, 2008.
10. A. Schug and J. N. Onuchic, *From protein folding to protein function and biomolecular binding by energy landscape theory*, Curr Op Pharm, **10**, no. 6, 709–714, 2010.

11. S. Mohanty, J. H. Meinke, and O. Zimmermann, *Folding of Top7 in unbiased all-atom Monte Carlo simulations*, Prot: Struct, Funct, Bioinf, **81**, no. 8, 1446–1456, 2013.
12. P. Casino, V. Rubio, and A. Marina, *Structural Insight into Partner Specificity and Phosphoryl Transfer in Two-Component Signal Transduction*, Cell, **139**, no. 2, 325–336, 2009.
13. D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, *Protein 3D structure computed from evolutionary sequence variation*, PloS one, **6**, no. 12, e28766, 2011.
14. J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, *Genomics-aided structure prediction*, Proc Nat Acad of Sci USA, **109**, no. 26, 10340–10345, 2012.
15. P. Tian, W. Boomsma, Y. Wang, D. E. Otzen, M. H. Jensen, and K. Lindorff-Larsen, *Structure of a functional amyloid protein subunit computed using sequence variation*, J Amer Chem Soc, **137**, no. 1, 22–25, 2014.
16. S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, and D. Baker, *Large-scale determination of previously unsolved protein structures using evolutionary information*, Elife, **4**, e09248, 2015.
17. A. Schug, M. Weigt, J. A. Hoch, J. N. Onuchic, T. Hwa, H. Szurmant *Computational modeling of phosphotransfer complexes in two-component signaling*, Meth Enzym, **471**, 43–58, 2010.
18. T. A. Hopf, C. P. Schärfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, and D. S. Marks, *Sequence co-evolution gives 3D contacts and structures of protein complexes*, Elife, **3**, e03430, 2014.
19. R. N. dos Santos, F. Morcos, B. Jana, A. D. Andricopulo, and J. N. Onuchic, *Dimeric interactions and complex formation using direct coevolutionary couplings*, Sci Rep, **5**, no. 1, 13652, 2015.
20. T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, *Three-dimensional structures of membrane proteins from genomic sequencing*, Cell, **149**, no. 7, 1607–1621, 2012.
21. S. Hayat, C. Sander, D. S. Marks, and A. Elofsson, *All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences*, Proc Nat Acad Sci USA, **112**, no. 17, 5413–5418, 2015.
22. A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, and H. Szurmant, *Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis*, Proc Nat Acad Sci USA, **109**, no. 26, E1733–42, 2012.
23. F. Morcos, B. Jana, T. Hwa, and J. N. Onuchic, *Coevolutionary signals across protein lineages help capture multiple protein conformations*, Proc Nat Acad Sci USA, **110**, no. 51, 20533–20538, 2013.
24. L. Sutto, S. Marsili, A. Valencia, and F. L. Gervasio, *From residue coevolution to protein conformational ensembles and functional dynamics*, Proc Nat Acad Sci USA, **112**, no. 44, 13567–13572, 2015.
25. A. Toth-Petroczy, P. Palmedo, J. Ingraham, T. A. Hopf, B. Berger, C. Sander, and D. S. Marks, *Structured states of disordered proteins from genomic sequences*, Cell, **167**, no. 1, 158–170, 2016.
26. E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt, *Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction*, Nuclc Ac Res, **43**, no. 21, 10444–10455, 2015.

27. C. Weinreb, A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander, and D. S. Marks, *3D RNA and Functional Interactions from Evolutionary Couplings*, Cell, **165**, no. 4, 963–975, 2016.
28. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, *Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1*, Mol Bio Evol, **33**, no. 1, 268–280, 2016.
29. F. Bai, F. Morcos, R. R. Cheng, H. Jiang, and J. N. Onuchic, *Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis*, Proc Nat Acad Sci USA, **113**, no. 50, E8051–E8058, 2016.
30. T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani, *Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis*, Proc Nat Acad Sci USA, **113**, no. 43, 12186–12191, 2016.
31. A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, *Inferring interaction partners from protein sequences*, Proc Nat Acad Sci USA, **113**, no. 43, 12180–12185, 2016.
32. T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. Schärfe, M. Springer, C. Sander, and D. S. Marks, *Mutation effects predicted from sequence co-variation*, Nat Biotech, **35**, no. 2, 128–135, 2017.
33. M. B. Zerihun and A. Schug, *Biomolecular coevolution and its applications: going from structure prediction toward signaling, epistasis, and function*, Biochl Soc Trans, accepted, 2017.
34. L. R. Mead and N. Papanicolaou, *Maximum entropy in the problem of moments*, J Math Phys, **25**, no. 8, 2404–2417, 1984.
35. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, *Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models*, Phys Rev E - Stat, Nonl, Soft Matt Phys, **87**, no. 1, 1–16, 2013.
36. D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cogn Sci, **9**, no. 1, 147–169, 1985.
37. D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, *PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments*, Bioinf, **28**, no. 2, 184–190, 2012.
38. C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, *Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners*, PLoS ONE, **9**, no. 3, 1–12, 2014.
39. S. Cocco and R. Monasson, *Adaptive cluster expansion for inferring Boltzmann machines with noisy data*, Phys Rev Lett, **106**, no. 9, 090601, 2011.
40. J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, *New method for parameter estimation in probabilistic models: minimum probability flow*, Phys Rev Lett, **107**, no. 22, 220601, 2011.
41. L. Burger and E. Van Nimwegen, *Disentangling direct from indirect co-evolution of residues in protein alignments*, PLoS Comp Bio, **6**, no. 1, 2010.
42. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, *Inverse Statistical Physics of Protein Sequences: A Key Issues Review*, 2017, preprint arXiv:1703.01222.
43. C. Feinauer, M. J. Skwark, A. Pagnani, and E. Aurell, *Improving contact prediction along three dimensions*, PLoS Comp Bio, **10**, no. 10, e1003847, 2014.

44. E. Aurell and M. Ekeberg, *Inverse ising inference using all the data*, Phys Rev Lett, **108**, no. 9, 090201, 2012.
45. H. Kamisetty, S. Ovchinnikov, and D. Baker, *Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era*, Proc Nat Acad Sci USA, **110**, no. 39, 15674–15679, 2013.
46. M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson, *Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns*, PLoS Comp Bio, **10**, no. 11, 2014.
47. M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, and A. Elofsson, *PconsFold: Improved contact predictions improve protein models*, Bioinf, **30**, no. 17, 482–488, 2014.
48. M. Källberg, G. Margaryan, S. Wang, J. Ma, and J. Xu, *RaptorX server: a resource for template-based protein structure modeling*, Prot Struct Pred, 17–27, 2014.
49. R. Sheridan, R. J. Fieldhouse, S. Hayat, Y. Sun, Y. Antipin, L. Yang, T. Hopf, D. S. Marks, and C. Sander, *EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction*, biorxiv, 021022, 2015.
50. D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, *MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins*, Bioinf, **31**, no. 7, 999–1006, 2015.
51. K. Uziela, D. Menéndez Hurtado, N. Shu, B. Wallner, and A. Elofsson, *ProQ3D: improved model quality assessments using deep learning*, Bioinf, **33**, no. 10, 1578–1580, 2017.
52. S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, *Protein structure determination using metagenome sequence data*, Science, **355**, no. 6322, 294–298, 2017.
53. G. Uguzzoni, S. J. Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt, *Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis*, Proc Nat Acad Sci USA, **114**, no. 13, E2662–E2671, 2017.
54. H. Lammert, J. K. Noel, E. Haglund, A. Schug, and J. N. Onuchic, *Constructing a folding model for protein S6 guided by native fluctuations deduced from NMR structures*, J Chem Phys, **143**, no. 24, 243141, 2015.
55. M. Wegner, O. Taubert, A. Schug, and H. Meyerhenke, *Maxent-Stress Optimization of 3D Biomolecular Models*, 2017, preprint arXiv:1706.06805.
56. L. Gremer, D. Schölzel, C. Schenk, E. Reinartz, J. Labahn, R. B. Ravelli, M. Tusche, C. Lopez-Iglesias, W. Hoyer, H. Heise, et al., *Fibril structure of amyloid-ß (1-42) by cryoelectron microscopy*, Science, eaao2825, 2017.
57. Y. Tang, Y. J. Huang, T. A. Hopf, C. Sander, D. S. Marks, and G. T. Montelione, *Protein structure determination by combining sparse NMR data with evolutionary couplings*, Nat Meth, **12**, no. 8, 751–754, 2015.
58. S. Kirmizialtin, S. P. Hennelly, A. Schug, J. N. Onuchic, and K. Y. Sanbonmatsu, *Chapter Nine-Integrating Molecular Dynamics Simulations with Chemical Probing Experiments Using SHAPE-FIT*, Meth Enzym, **553**, 215–234, 2015.