
Mehari B. Zerihun¹ and Alexander Schug²

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
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’s1–3, 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 history7–11 by providing highly accurate models with resolutions comparable to experiments5,12; in the prediction of tertiary protein structure13–16, protein complexes5,17–19, membrane proteins20,21, conformational transitions22–25, and, most recently, first RNA structures26,27. 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 resistance28, drug design29, biological signalling30,31 or epistatic effects32. An overview of DCA applications can be found in Ref. 33. We will here explain the basic concepts of the statistical analysis.
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\(^3^4\). It describes the probability \( P \) of sampling a sequence \( \sigma = (a_1a_2a_3...a_L) \) of length \( L \) containing a residue or gap \( a_i \) at site \( i \). Formally, it is given by

\[
P(\sigma) = \frac{1}{Z} \exp \{ H(\sigma) \},
\]

\[
H(\sigma) = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i),
\]

(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). 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, $\sum_{\{a_k,k \neq i\}} P(\sigma) = P_i(\sigma_i)$, and pair-site marginals $\sum_{\{a_k,k \neq i,j\}} P(\sigma) = P_{ij}(\sigma_{i,j})$ together with the normalisation condition $\sum_a P(\sigma) = 1$ 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 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,

$$f_i(a) = \frac{1}{N} \sum_{n=1}^N \delta_{a,a_i},$$

$$f_{ij}(a,b) = \frac{1}{N} \sum_{n=1}^N (\delta_{a,a_i})(\delta_{b,b_j})$$

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 $O(q^4)$ which is too complex even for a sequence of average length.

Weigt et al. 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.

Later, a more efficient parameter inference was achieved based on mean-field approximation. 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.

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. This method named pseudo-likelihood maximisation Direct-Coupling Analysis (plmDCA) was shown to outperform mfDCA if the dataset in the MSA is sufficiently large. Many competing methods have been developed since, such as Boltzmann-machine learning, Gaussian approximation, adaptive cluster expansion, minimal probability flow, or Bayesian networks (for a recent review please cf. Ref. 42).

Still there are many challenges. 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.

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.

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 is to define direct-information, \( DI_{ij} \), between two sites \( i \) and \( j \) from the direct-probability, \( P_{ij}^{dir} \) as

\[
DI_{ij} = \sum_{a=1}^{q} \sum_{b=1}^{q} P_{ij}^{dir}(a,b) \log \left( \frac{P_{ij}^{dir}(a,b)}{f_i(a)f_j(b)} \right),
\]

\[
P_{ij}^{dir}(a,b) = \frac{1}{Z_{ij}} \exp \left\{ J_{ij}(a,b) + \tilde{h}_i(a) + \tilde{h}_j(b) \right\}.
\]

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

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

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 or pushing the frontier from a few “anecdotal” cases to 1000’s of systems. Further, biomolecular simulations or other efficient techniques can also integrate experimental measurements such as cryoEM, sparse NMR data or SHAPE-data 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


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.


