Title

Combined approaches from physics, statistics, and computer science for ab initio protein structure prediction: ex unitate vires (unity is strength)?

Authors

Delarue, Marc
Koehl, Patrice

Publication Date

2018

DOI

10.12688/f1000research.14870.1

Peer reviewed

REVIEW

Combined approaches from physics, statistics, and computer science for *ab initio* protein structure prediction: *ex unitate vires* (unity is strength)? [version 1; referees: 2 approved]

Marc Delarue[1], Patrice Koehl[2]

[1]Unité Dynamique Structurale des Macromolécules, Institut Pasteur, and UMR 3528 du CNRS, Paris, France
[2]Department of Computer Science, Genome Center, University of California, Davis, Davis, California, USA

v1 **First published:** 24 Jul 2018, **7**(F1000 Faculty Rev):1125 (doi: 10.12688/f1000research.14870.1)
Latest published: 24 Jul 2018, **7**(F1000 Faculty Rev):1125 (doi: 10.12688/f1000research.14870.1)

Abstract

Connecting the dots among the amino acid sequence of a protein, its structure, and its function remains a central theme in molecular biology, as it would have many applications in the treatment of illnesses related to misfolding or protein instability. As a result of high-throughput sequencing methods, biologists currently live in a protein sequence-rich world. However, our knowledge of protein structure based on experimental data remains comparatively limited. As a consequence, protein structure prediction has established itself as a very active field of research to fill in this gap. This field, once thought to be reserved for theoretical biophysicists, is constantly reinventing itself, borrowing ideas informed by an ever-increasing assembly of scientific domains, from biology, chemistry, (statistical) physics, mathematics, computer science, statistics, bioinformatics, and more recently data sciences. We review the recent progress arising from this integration of knowledge, from the development of specific computer architecture to allow for longer timescales in physics-based simulations of protein folding to the recent advances in predicting contacts in proteins based on detection of coevolution using very large data sets of aligned protein sequences.

Keywords

Protein structure prediction, template-free prediction, secondary structure prediction, co-variation

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 24 Jul 2018	✓	✓

F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty. In order to make these reviews as comprehensive and accessible as possible, peer review takes place before publication; the referees are listed below, but their reports are not formally published.

1 **Amarda Shehu**, George Mason University, USA

2 **Dong Xu**, University of Missouri, USA
University of Missouri, USA

Discuss this article

Comments (0)

Corresponding author: Patrice Koehl (koehl@cs.ucdavis.edu)

Author roles: Delarue M: Conceptualization, Writing – Original Draft Preparation; **Koehl P**: Conceptualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Delarue M and Koehl P. **Combined approaches from physics, statistics, and computer science for *ab initio* protein structure prediction: *ex unitate vires* (unity is strength)? [version 1; referees: 2 approved]** *F1000Research* 2018, **7**(F1000 Faculty Rev):1125 (doi: 10.12688/f1000research.14870.1)

First published: 24 Jul 2018, **7**(F1000 Faculty Rev):1125 (doi: 10.12688/f1000research.14870.1)

Introduction

Proteins are essential macromolecular biomolecules to all organisms, as they participate in nearly all processes within cells, thereby sustaining life. They catalyze most biochemical reactions and are involved in maintaining the integrity of the genomic information. They control and perform self-replication, transport molecules (including nutrients) in and out of the cell, play a role in immune responses, and control cell cycles; those are only a small subset of the activities associated with proteins within a cell. The amino acid sequence of a protein is defined by the nucleotide sequence of its gene. This relationship is well known and captured by the genetic code. The function of a protein then is defined by the geometry adopted by the corresponding chain of amino acids, the so-called tertiary structure of the protein. However, understanding the relationships between the sequence and structure of a protein, the second piece in the puzzle of understanding how proteins function, remains elusive. This review is focused on the computational approaches developed to unravel these relationships.

Although much progress has been achieved in connecting protein sequence to their function, there remains a high proportion of genes with unknown function, especially in bacteriophages and, even more dramatically, in archaeal viruses, in which 75 to 90% of their genomes remain largely unknown[1]. Determination of the function of a protein experimentally is a highly resource-intensive process. As a consequence, much hope is put into alternate approaches based on computational methods. Sequence analysis is usually the first step, as significant sequence similarity is still the most reliable way of inferring function. There are, however, some exceptions to this inference, from proteins with nearly identical sequences but different functions[2] to proteins with different sequences but similar functions[3]. When sequence is uninformative when it comes to predicting function or when there are no detectable homologues in the databases of annotated protein sequences, structure can often provide further insight. Therefore, significant efforts are put into predicting function from structure[4]. These efforts rely obviously on the availability of structural information.

Our current knowledge of protein structure comes mostly from decades of experimental studies, using X-ray crystallography, nuclear magnetic resonance spectroscopy, or more recently cryo-electron microscopy. The first protein structures to be solved were those of hemoglobin and myoglobin more than 50 years ago[5–7]. As of April 2018, there are close to 140,000 protein structures in the database of biological macromolecular structures (http://www.rcsb.org). However, this number drops to about 50,000 if one keeps sequences with less than 90% sequence identity (http://www.rcsb.org/pdb/statistics/clusterStatistics.do). This number is obviously small compared with the number of existing proteins. Although the incentive to significantly increase the number of experimentally defined protein structures is evident (for instance, for drug-design purposes), how to achieve such a drastic increase is less clear, as the cost (mostly in man-months) to solve a new protein structure is very high. Unfortunately, this high cost led to the ending of the Protein Structure Initiative by the National Institutes of Health in 2015, long before it had reached its goal. As a reminder, the main ideas behind this initiative were to progressively document a full library of possible natural protein folds and to use the structure of new proteins to annotate them functionally[4,8].

There is hope for an alternate solution to complement the experimental efforts to characterize the protein structure space. From the seminal work of Anfinsen[9], we know that the sequence fully determines the three-dimensional structure of a protein. In addition, the information on protein sequences is growing exponentially: as of April 2018, there were more than 557,000 protein sequences deposited in SwissProt-Uniprot version 2018-03, the fully annotated repository of protein sequences (note that these form a small subset of all known proteins; the RefSeq database of non-redundant proteins currently contains more than 110 million sequences). As a consequence, there is a lot of effort put into predicting the structure of a protein directly from the knowledge of its sequence (and its relatives). This has been referred to as one of the "holy grails" in molecular biology, also called the protein structure prediction problem. This problem has been around for decades[10,11] but clearly remains a very active area of research. This is illustrated, for example, by the following: (i) a PubMed search with the keywords "protein structure prediction" finds more than 4,000 hits since 2014; (ii) this field has its own dedicated biannual conference, the Critical Assessment of Structure Prediction (CASP) meetings[12–14]; (iii) it has its own section in F1000; and (iv) it is even the topic of a popular video game[15,16]. In this article, we review the latest advances related to solving this problem and put a special focus on highlighting the specific scientific fields that are involved in those advances.

Protein structure prediction is in fact part of a larger problem, the protein-folding problem. The latter can be seen as an investigation over two questions: (i) understanding the kinetic folding mechanism, namely unraveling the temporal sequence of events describing how a polypeptide chain finds its way from an unfolded structure into a compact, globular conformation in a seemingly unreasonably fast time given the number of possible conformations it may adopt (the Levinthal paradox[17,18]), and (ii) understanding the physical folding code, namely deciphering how the physicochemical properties of the amino acids along the polypeptide chain that form a protein sequence are uniquely responsible for the tertiary structure of a protein and its stability. We focus on the second question and recent computer-based solutions to this problem. We note first that most successful structure prediction algorithms rely on the assumption that similar sequences lead to similar structures[19]. This has led to the development of comparative modeling techniques, which have seen considerable improvements over the years, as reported in the overview of the most recent CASP meeting[14,20]. Based on those reports, homology modeling methods are estimated to produce reasonable atomic models provided that the alignment has a sequence identity above 50%, with high coverage (>80%). We will not cover this topic that has been reviewed extensively elsewhere[21–23]. Instead, we review results in the more challenging problem of template-free (*ab initio*) modeling. More specifically, we limit our presentation to recent techniques

that have been developed with physics-based and machine-learning perspectives. We focus on free energy-based methods, secondary structure prediction, and contact predictions. We provide our own perspective on recent progress in each of these topics; recent technical reviews can be found elsewhere[10,24–35].

Physics-based approaches to protein structure prediction

The stability of a protein structure is defined by an interplay of multiple factors, from the local geometry imposed by structural chemistry (bond lengths, bond angles, and dihedral angle preferences) to close-range physical interactions that govern the structures of all liquids and solids. Among those, van der Waals interactions lead to tight packing in the protein core, and electrostatic interactions, such as salt bridges and hydrogen bonds, define short- or long-range couplings that underlie, for example, allostery[36,37] and active transport[38,39]. In addition, a protein interacts with its environment, namely water, which creates a polar environment, forcing hydrophobic residues to pack into a hydrophobic core, and ions (Na^+, Cl^-, Mg^{2+}, etc.), which interact with charged residues at the surface of the protein. All of those interactions are captured by semi-empirical "force fields", with different levels of approximation, from implicit or explicit solvent, presence or absence of terms accounting for polarization, to complete treatment of classic electrostatics[40,41]. Combined with sampling methods such as molecular dynamics (MD) or Monte Carlo (MC) simulations, these force fields can be used to generate ensembles of conformations for any protein structure, from which the native tertiary structure hopefully can be selected[42]. The input to such sampling techniques can be an initial guess for the conformation of the protein, obtained, for example, from comparative modeling; this procedure is referred to as model refinement. Although progress associated with this problem is always optimistically conveyed in the reports from the successive CASP meetings (for example,[43]), there is a growing consensus that the current techniques have reached a plateau and that a combination of better force fields and improved sampling is needed for further improvements[44,45].

The real challenge is to start from any random conformation for the protein and hopefully induce, through *in silico* molecular simulation in the computer, the folding to the native structure of the protein. These simulations have the advantage of directly sampling the free energy surface spanned by the protein. In addition, if successful, they solve the structure prediction problem and provide kinetic and thermodynamic information about folding. As the size of the conformation space is overwhelmingly large, the general understanding in the structure prediction community for a long time was that such an approach would be limited to predicting the structure and folding of small peptides[46] unless it could be supplemented with additional experimental[47–49] or data-based[50–52] information. This limited optimism was fortunately proven wrong on at least a few occasions, mostly as a consequence of changes in computing power and in how computing is performed. One example is Folding@home[53,54], developed by Vijay Pande at Stanford University, in which computer users all over the world donate their idle computer time to perform the physical molecular simulations required for protein folding. Using this form of "social distributed computing", the Pande group demonstrated that molecular dynamics can generate accurate protein-folding rates[55]. More recent results with this distributed computing approach include the analyses over "long" (over the millisecond) time ranges of the transitions between inactive and active forms of G-protein-coupled receptors (GPCRs) (proteins with 500 amino acids)[56] and modeling of the activation pathways of Src kinases (250 residues included in the simulations)[57]. The success of Folding@home has led the computational biology community nowadays to expect "routine" simulations of protein dynamics in the millisecond timescale and longer. We also note an interesting trend within the field of MD simulations in terms of methods development. Such simulations request constant development of the algorithms implemented for performing the dynamics; the community has responded well to this need, and new versions of the software packages appear regularly; see, for example, the recent developments around OpenMM[58], the software package behind Folding@home. These new software packages push the limits of what can be achieved with those simulations, enabling the study of larger and more complex systems, which in turn foster the development of new methods for analyzing the results. Indeed, these large simulations generate massive amounts of data, raising new challenges in data sciences. Sophisticated statistical machineries have been proposed to analyze those data, such as the Markov state models[59–61], which are constantly updated[62], to the point of switching "from [being] an Art to [becoming] a Science", paraphrasing a recent review by Husic and Pande[63]. Interestingly, the specificity of the data generated by molecular simulations has led to the natural adaptation of machine-learning techniques to support their analyses; for a recent review on this topic, see Mittal and Shukla[64].

An alternate approach to distributing computing is to adapt the computing hardware to the equations and the force field implemented in molecular simulations. The Anton computer from DE Shaw Research, which was custom-designed for molecular simulations[65,66], gives at least one order of magnitude better performance than conventional computers[67]. Using this new technology, this group was able to study, for example, the dynamics in human ubiquitin in the picosecond to millisecond time range using unbiased dynamics, revealing that conformations visited in the simulations are very similar to those found in crystal structures of ubiquitin and detecting correlated motions in the protein that are consistent with experimental observations[68]. Of equal significance, simulations performed on Anton are helping to identify the structural origins of slow diffusion during protein folding[69,70].

Finally, it should be mentioned that advances in computing models (either cloud-based or with a specific architecture) are not the only options to improve simulations of protein folding. Dill *et al.* recently illustrated that the addition of some semi-reliable external information to the potential that steers the molecular simulation enables accurate prediction of the native conformations for small protein structures within the context of CASP[52,71]. This information is provided in the form of binary residue contacts deduced from the protein sequence itself. This will be discussed in more detail below.

Secondary structure prediction: significant improvements from machine learning

Ultimately, the structure of a protein is defined by the laws of chemistry and physics. In the section above, we described attempts to derive a physics-based solution to the structure prediction problem; our tone was meant to be positive and optimistic, but we cannot deny that those attempts have their limitations. At the same time, the structure prediction community is exploring methods outside the realm of physics-based first-principles equations. Secondary structure prediction, a subproblem of protein structure prediction, is a good example of how methodologies in this whole field have evolved over the last 50 years.

Here, we are concerned with the prediction of the local conformation of the backbone of a protein, namely its organization into helices, strands, and coils (that is, loosely structured regions). As discussed in recent reviews of secondary structure prediction, this problem is important in its own right, as it impacts many areas of structural bioinformatics and functional genomics[25,28]. Linus Pauling can be considered the father of this field, as he correctly predicted the presence of helices and strands as stable substructures in proteins[72,73]. In the nearly 80 years that followed his intuition, secondary structure prediction has struggled to come up with a definite solution through at least three phases, which Burkhard Rost referred to as generations[74]. Briefly, the starting idea was to derive statistical preferences for amino acids to be within a specific secondary structure based on known protein structures and use those preferences for predictions. This inference, also called the inverse problem, proved harder than expected. Chou and Fasman, for example, derived empirical rules for both helices and strands[75]. However, refining those rules proved to be a game with moving targets, and the method remained limited in scope. The second generation of methods focused on either the definition of the propensities themselves, making them locally context-dependent to include neighborhood effects[76], or how to derive the rules for inference, relying more and more on learning those rules from the data instead of enforcing a specific model. The latter led to the introduction of neural networks in the field[77]. This second generation unfortunately led to only modest improvement. However, once the door was open to learning from the data, it was then only natural to increase the amount and diversity of the data that are included to generate better models. The most obvious choice was to introduce information from homologous sequences[78], which led to the third generation of secondary structure prediction methods, including the significantly improved neural network solutions implemented in PhD[79] and PSIPRED[80]. Those methods were introduced in the 1990s. Since then, we have seen the development of more and more sophisticated machine-learning methods with more involved neural networks, such as the recent use of deep neural networks[81] and deep convolutional neural fields[82] for secondary structure prediction (for a comprehensive review of the different types of networks and machine-learning techniques that have been applied to solve the secondary structure prediction problem, see 28). This recently led Yang *et al.* to ask whether we are in "the final stretch" for protein structure prediction[25]. Although we appreciate their optimism, we raise a small word of caution. The goal of machine-learning methods is to identify recurrent patterns in data that can be used to solve the inference problem: that is, connecting features (here amino acid sequences) to underlying variables (here secondary structure conformation). Optimizing a model to a generalized inference model usually prevents prediction of atypical behavior. However, we note that this is a problem for machine learning in general.

Predicting contacts in proteins: (statistical) physics or machine learning?

Protein structures are more conserved than their sequences. It is therefore legitimate to expect that correlated mutations occur between contacting residues: in the event that one residue of such an interacting pair mutates, the effect of this mutation is likely to be accommodated by a corresponding mutation of the contacting residue. This was discussed in detail in the recent analysis of coevolution between residues by Baker *et al.*[83]. This basic idea implies that if we can detect co-variation in sequences, we should be able to predict contacts in protein.

The possibility to detect co-variations has significantly increased with the increase in the number of protein sequences that are available. The related search for an inference between those co-variations and actual contacts in protein has been the focus of decades of research, and recent breakthroughs are highlighted below. We will limit our presentation to general concepts (especially the mean-field approach) and refer readers to excellent recent reviews for more technical presentations[34,35,84].

Let us consider a multiple sequence alignment (MSA) for a protein family P. This MSA can be considered as a table T, consisting of L rows corresponding to the L proteins included in the MSA and of M columns, where M is the length of the proteins. The value $T(l,m)$ at row l and column m is the type of amino acid observed at the m-th position of protein l. $T(l,m)$ is assumed to be represented by an integer number in $[1,q]$, where $q = 21$, representing the 20 amino acids and a gap. The information in this MSA is first summarized in terms of frequencies,

$$f_i(a) = \frac{1}{L}\sum_{l=1}^{L}\delta_{a,T(l,i)} \qquad f_{ij}(a,b) = \frac{1}{L}\sum_{l=1}^{L}\delta_{a,T(l,i)}\delta_{b,T(l,j)}, \quad (1)$$

with i and j in $[1,M]$ and a and b in $[1,q]$ and δ the Dirac function such that $\delta_{xy} = 1$ if $x = y$, and 0 otherwise. In practice, weights and pseudo-counts are introduced to account for possible biases in the distribution of sequences in the MSA[85]; we ignore them here for the sake of simplicity. In these definitions, $f_i(a)$ is the frequency of the occurrence of amino acid a at column i in the MSA, and $f_{ij}(a,b)$ is the frequency of co-occurrence of amino acid types a and b at columns i and j. If the two columns i and j are independent, the joint distribution $f_{ij}(a,b)$ would simply be equal to the product $f_i(a)\, f_j(b)$. Deviations from this independence can be either quantified by using a covariance

$$C_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b) \quad (2)$$

or summarized by computing the mutual information between columns i and j[86]

$$M_{ij} = \sum f_{ij}(a,b) \ln \frac{f_{ij}(a,b)}{f_i(a) f_j(b)}. \quad (3)$$

It is tempting to infer spatial proximity from those covariance or mutual information measures (or any variation of). This idea was pursued as early as in the 1980s for predicting virus functions[87] and in the 1990s to detect contacts in proteins[88,89]. However, it met with little success, for a reason that can be described as follows[30,85]: when a residue *i* is in contact with a residue *j*, and *j* is in contact with a residue *k*, then *i* and *k* will exhibit correlation, even if they are not in contact. Correct analyses of the observed co-variations (computed by using equation 1 and equation 2) require that direct interactions be distinguished from possible indirect correlations. Currently, two approaches are being developed to disentangle those direct and indirect variations and infer contacts from observed co-variations: those that construct a statistical model for full-length protein sequences by using methods from statistical physics and those that learn this model from the data by using machine-learning techniques.

Among the statistical methods developed for co-variation analyses, sequence-based probabilistic formalisms have been proposed as early as 2002[90], followed by message-passing algorithms[91], mean-field methods[85], and Gaussian[92] or pseudo-likelihood[93,94] approximations. More precise methods, based on adaptive cluster expansion[95] or Boltzmann learning using MC sampling[96,97], have been proposed recently. In the following, we briefly review the mean-field methods.

From a statistical physics point of view, the protein sequence space can be equipped with a spin-glass Hamiltonian model such that the Hamiltonian of a sequence *S* of length *M* is given by

$$H(S) = -\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} J_{ij}(s_i, s_j) - \sum_{i=1}^{M} h_i(s_i), \quad (4)$$

where s_i is the amino acid type at position *i* in *S*, $h_i(s_i)$ are single-site "fields", and $J_{ij}(s_i,s_j)$ are pair-site "couplings" between *i* and *j*. When the size of the alphabet of amino acid types is 2, equation 4 corresponds to an Ising model. In the more general case of an alphabet of size 21, equation 4 is referred to as a Potts Hamiltonian model. Given this model, the probability of observing a sequence *S* follows a Boltzmann distribution,

$$P(S) = \frac{e^{-\beta H(S)}}{\sum_C e^{-\beta H(C)}}, \quad (5)$$

where $\beta = 1/kT$ is a normalization parameter, and the sum at the denominator runs over all sequences *C* of length *M*. Note that this denominator is the partition function *Z* over the sequence space, from which a free energy can be derived:

$$\beta F = -\log(Z). \quad (6)$$

The knowledge of *Z* (or *F*) is enough to fully describe the thermodynamics of the system. In particular, the mean state at position *i* and the correlation between positions *i* and *j* can be computed as

$$< s_i >= \frac{1}{Z}\frac{dZ}{dh_i} \qquad < s_i s_j >= \frac{1}{Z}\frac{d^2 Z}{dh_i dh_j}. \quad (7)$$

The parameters $h_i(s_i)$ and $J_{ij}(s_i,s_j)$ need to be adjusted such that these model-based state and correlation values are consistent with the observed values given in equation 1. This is the basic concept behind the direct coupling analysis (DCA)[98,99]. Several versions of DCA, differing in the method used to compute $h_i(s_i)$ and $J_{ij}(s_i,s_j)$, have been proposed. In the mean-field approximation, there is a simple relationship between J_{ij} and the covariance C_{ij} defined in equation 2:

$$J_{ij}(a,b) = -(C^{-1})_{ij}(a,b). \quad (8)$$

Recent versions of DCA typically reach accurate prediction of about 85% true positives and only 15% false positives at the 8 Å distance (typically used as the definition of "contact" in DCA studies)[35]. This success has led to DCA being used outside of the protein structure prediction problem, from the prediction of the structure of protein complexes[34,100] and protein conformational transitions[96,101], RNA structure prediction[102–104], and prediction of ordered states for disordered proteins[105] to the prediction of mutation effects in proteins[106].

The Potts Hamiltonian model for defining the stability of a protein sequence summarizes all forms of interactions between residues into two simple sets of parameters: single-site and pair-sites. As the information content of an MSA goes beyond those simple interactions, it was natural to see attempts to combine such additional information to co-variation measures by using machine-learning techniques, much akin to what has been done for secondary structure prediction (see above). In the popular MetaPSICOV[107] approach, for example, a first-stage classifier is trained to predict whether residues are in contact by considering 672 input features for each pair of positions *i* and *j* in an MSA. Three windows are defined: one nine-residue window centered at *i*, one nine-residue window centered at *j*, and a central five-residue window centered at *(i+j)/2*. Each of those 23 positions then is characterized by its amino acid composition in the MSA (21 values), followed by information on secondary structures and solvent accessibility, both predicted from the MSA itself. These position-specific features then are complemented with coevolution information, such as mutual information (computed with equation 3) and DCA-predicted contact information, as computed with equation 8. The complete set of 672 features then is used to predict whether the positions *i* and *j* define a contact, using different cutoff values for defining such a contact. The outputs of the first-stage classifier then are used as input to a second-stage classifier, again combining information over windows around the residues under scrutiny. MetaPSICOV is a hybrid method: it uses both the results of the statistical physics DCA method and additional information derived from the MSA to predict contact. It has been shown to achieve higher accuracy than DCA alone[107]. The success of MetaPSICOV is not isolated. It is interesting, for example, that out of the 23 methods that have been used for contact predictions in CASP12, at least 21 are clearly

based on machine learning, using different versions of deep learning methods or combinations of such methods[108]. Those methods were reported to be strikingly successful, and precisions above 90% were achieved by the best predictors in more than half the targets in CASP12, a result considered to be a highly significant improvement compared with the results obtained during CASP11[108]. In this respect, it will be particularly important to watch the result of CASP13 (2018), which will fully integrate mature contact prediction methods (http://predictioncenter.org).

Conclusions

Is the protein structure prediction problem "solved"? In the past, many have claimed that to be the case or have at least described progress and remaining challenges[109,110]. In line with a recent review by Dill and MacCallum[10], it is unclear to us whether this question remains relevant, given the diversity associated with structure prediction, both in terms of the methods that have been developed as attempts to solve it and in terms of their applications, from predicting the structures of small globular proteins, of membrane proteins, and the conformational space accessible to intrinsically disordered proteins. We also note the difficulties that come with assessing what "success" means, and this is illustrated by the need to have a series of conferences dedicated to that problem, the CASP conferences[12].

The prediction of contacts in proteins, described in this review, is a good example that highlights the constant evolution of the field of protein structure prediction. The concept of correlating co-variations between residues in sequence alignments to spatial proximity was well known for RNA structure prediction. As highlighted above, its application to proteins was delayed because of the difficulties of dealing with indirect variations. Independent applications of statistical physics methods and machine-learning methods have, to some extent, solved those difficulties, opening the door to predicting geometric contacts in proteins from sequence only. This progress leads, however, to the next challenge: using those contacts to predict the overall geometry, that is, the structure of a protein. This is by no means an easy problem, as those contacts are usually given as binary, noisy information. Significant methodological developments are expected to solve this new challenge, which in turn will lead to yet another challenge.

The protein structure prediction problem intrinsically relates to basic science. However, it has led to important methodological and technical developments, from computer hardware for physics-based simulations to statistical methods and new machine-learning technologies. In practice, in biology, it will have an impact on many diseases related to protein misfolding, from neurodegenerative diseases to diabetes through the prediction of stability of mutants (EVmutation website: http://marks.hms.harvard.edu) and perhaps even to personalized medicine.

Attempts to solve the protein structure prediction problem have involved scientists from a large panel of disciplines, including biology, physics, statistics, and computer science. Even if the divide between physics-based approaches and data-driven methods still exists, this divide has been constructive and not disruptive. Data scientists involved in structure prediction are regularly using results from physics-based models, and reciprocally physicists are adding more and more data coming from data mining in their simulations. Therefore, the question we have raised in the title of this review may be as irrelevant as the question of whether the problem is solved: the future of protein structure prediction is interdisciplinary and will remain so.

Abbreviations

CASP, Critical Assessment of Structure Prediction; DCA, direct coupling analysis; MC, Monte Carlo; MD, molecular dynamics; MSA, multiple sequence alignment

Competing interests

The authors declare that they have no competing interests.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

Some of the ideas discussed here originated from a workshop organized by the Institute for Mathematical Sciences, National University of Singapore during the summer in 2017. We thank them for their hospitality and financial support.

References

F1000 recommended

1. F Krupovic M, Cvirkaite-Krupovic V, Iranzo J, *et al.*: **Viruses of archaea: Structural, functional, environmental and evolutionary genomics.** *Virus Res.* 2018; **244**: 181–93.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

2. F Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol.* 2002; **318**(2): 595–608.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

3. Whisstock JC, Lesk AM: **Prediction of protein function from protein sequence and structure.** *Q Rev Biophys.* 2003; **36**(3): 307–40.
PubMed Abstract | Publisher Full Text

4. Watson JD, Laskowski RA, Thornton JM: **Case studies: function predictions of structural genomics results.** In *From protein structure to function with bioinformatics.* Edited by Ridgen D: Springer; 2017.
Publisher Full Text

5. Kendrew JC, Dickerson RE, Strandberg BE, *et al.*: **Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution.** *Nature.* 1960; **185**(4711): 422–7.
PubMed Abstract | Publisher Full Text

6. Kendrew JC, Bodo G, Dintzis HM, *et al.*: **A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.** *Nature.* 1958; **181**(4610): 662–6.
PubMed Abstract | Publisher Full Text

7. Perutz MF, Rossmann MG, Cullis AF, *et al.*: **Structure of haemoglobin: a**

three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature.* 1960; **185**(4711): 416–22.
PubMed Abstract | Publisher Full Text

8. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science.* 2001; **294**(5540): 93–6.
PubMed Abstract | Publisher Full Text

9. F Anfinsen CB: **Principles that govern the folding of protein chains.** *Science.* 1973; **181**(4096): 223–30.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

10. F Dill KA, MacCallum JL: **The protein-folding problem, 50 years on.** *Science.* 2012; **338**(6110): 1042–6.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

11. Krokhotin A, Dokholyan NV: **Protein folding: Over half a century lasting quest: Comment on "There and back again: Two views on the protein folding puzzle" by Alexei V. Finkelstein et al.** *Phys Life Rev.* 2017; **21**: 72–4.
PubMed Abstract | Publisher Full Text

12. Moult J, Pedersen JT, Judson R, *et al.*: **A large-scale experiment to assess protein structure prediction methods.** *Proteins.* 1995; **23**(3): ii–v.
PubMed Abstract | Publisher Full Text

13. Moult J: **A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.** *Curr Opin Struct Biol.* 2005; **15**(3): 285–9.
PubMed Abstract | Publisher Full Text

14. Moult J, Fidelis K, Kryshtafovych A, *et al.*: **Critical assessment of methods of protein structure prediction (CASP)-Round XII.** *Proteins.* 2018; **86 Suppl 1**: 7–15.
PubMed Abstract | Publisher Full Text | Free Full Text

15. F Cooper S, Khatib F, Treuille A, *et al.*: **Predicting protein structures with a multiplayer online game.** *Nature.* 2010; **466**(7307): 756–60.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

16. Kleffner R, Flatten J, Leaver-Fay A, *et al.*: **Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta.** *Bioinformatics.* 2017; **33**(17): 2765–7.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Levinthal C: **How to fold graciously.** In *Mossbauer Spectroscopy in Biological Systems; Allerton House, Monticelle, Illinois.* Edited by Debrunner P, Tsibris JCM, Munck E: University of Illinois Press; 1969; 22–24.
Reference Source

18. Karplus M: **The Levinthal paradox: yesterday and today.** *Fold Des.* 1997; **2**(4): S69–75.
PubMed Abstract | Publisher Full Text

19. Kaczanowski S, Zielenkiewicz P: **Why similar protein sequences encode similar three-dimensional structures?** *Theor Chem Acc.* 2010; **125**(3–6): S69–75.
Publisher Full Text

20. F Kryshtafovych A, Monastyrskyy B, Fidelis K, *et al.*: **Evaluation of the template-based modeling in CASP12.** *Proteins.* 2018; **86 Suppl 1**: 321–34.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

21. Webb B, Sali A: **Protein structure modeling with *MODELLER*.** *Methods Mol Biol.* 2014; **1137**: 1–15.
PubMed Abstract | Publisher Full Text

22. Fiser A: **Comparative protein structure modelling.** In *From protein structure to function with bioinformatics.* Edited by Ridgen J: Springer; 2017; 57–90.
Publisher Full Text

23. Lam SD, Das S, Sillitoe I, *et al.*: **An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences.** *Acta Crystallogr D Struct Biol.* 2017; **73**(Pt 8): 628–40.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Dror RO, Dirks RM, Grossman JP, *et al.*: **Biomolecular simulation: a computational microscope for molecular biology.** *Annu Rev Biophys.* 2012; **41**: 429–52.
PubMed Abstract | Publisher Full Text

25. Yang Y, Gao J, Wang J, *et al.*: **Sixty-five years of the long march in protein secondary structure prediction: the final stretch?** *Brief Bioinform.* 2018; **19**(3): 482–94.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Lee J, Freddolino PL, Zhang Y: ***Ab initio* protein structure prediction.** In *From protein structure to function with bioinformatics.* Edited by Ridgen J: Springer; 2017; 3–25.
Publisher Full Text

27. Finkelstein AV, Badretdin AJ, Galzitskaya OV, *et al.*: **There and back again: Two views on the protein folding puzzle.** *Phys Life Rev.* 2017; **21**: 56–71.
PubMed Abstract | Publisher Full Text

28. Jiang Q, Jin X, Lee SJ, *et al.*: **Protein secondary structure prediction: A survey of the state of the art.** *J Mol Graph Model.* 2017; **76**: 379–402.
PubMed Abstract | Publisher Full Text

29. de Oliveira S, Deane C: **Co-evolution techniques are reshaping the way we do structural bioinformatics [version 1; referees: 2 approved].** *F1000Res.* 2017; **6**: 1224.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Cocco S, Feinauer C, Figliuzzi M, *et al.*: **Inverse statistical physics of protein sequences: a key issues review.** *Rep Prog Phys.* 2018; **81**(3): 32601.
PubMed Abstract | Publisher Full Text

31. Li B, Fooksa M, Heinze S, *et al.*: **Finding the needle in the haystack: towards solving the protein-folding problem computationally.** *Crit Rev Biochem Mol Biol.* 2018; **53**(1): 1–28.
PubMed Abstract | Publisher Full Text

32. Maldonado-Nava FG, Frausto-Solis J, Sanchez-Hernandez JP, *et al.*: **Comparative study of computational strategies for protein structure prediction.** In *Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications.* Edited by Castillo O, Melin P, Kacprzyk J: Springer; 2018; **749**: 449–459.
Publisher Full Text

33. F Adhikari B, Hou J, Cheng J: **Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning.** *Proteins.* 2018; **86 Suppl 1**: 84–96.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

34. Szurmant H, Weigt M: **Inter-residue, inter-protein and inter-family coevolution: bridging the scales.** *Curr Opin Struct Biol.* 2017; **50**: 26–32.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Figliuzzi M, Barrat-Charlaix P, Weigt M: **How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?** *Mol Biol Evol.* 2018; **35**(4): 1018–27.
PubMed Abstract | Publisher Full Text

36. Chakravorty DK, Merz KM Jr: **Studying allosteric regulation in metal sensor proteins using computational methods.** *Adv Protein Chem Struct Biol.* 2014; **96**: 181–218.
PubMed Abstract | Publisher Full Text

37. Simonson T, Aleksandrov A, Satpati P: **Electrostatic free energies in translational GTPases: Classic allostery and the rest.** *Biochim Biophys Acta.* 2015; **1850**(5): 1006–16.
PubMed Abstract | Publisher Full Text

38. F Newstead S: **Recent advances in understanding proton coupled peptide transport via the POT family.** *Curr Opin Struct Biol.* 2017; **45**: 17–24.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

39. F Kaila VRI: **Long-range proton-coupled electron transfer in biological energy conversion: towards mechanistic understanding of respiratory complex I.** *J R Soc Interface.* 2018; pii: 20170916.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

40. F Nerenberg PS, Head-Gordon T: **New developments in force fields for biomolecular simulations.** *Curr Opin Struct Biol.* 2018; **49**: 129–38.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

41. F Huang J, MacKerell AD Jr: **Force field development and simulations of intrinsically disordered proteins.** *Curr Opin Struct Biol.* 2018; **48**: 40–8.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

42. F Piana S, Klepeis JL, Shaw DE: **Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.** *Curr Opin Struct Biol.* 2014; **24**: 98–105.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

43. Hovan L, Oleinikovas V, Yalinca H, *et al.*: **Assessment of the model refinement category in CASP12.** *Proteins.* 2018; **86 Suppl 1**: 152–67.
PubMed Abstract | Publisher Full Text

44. F Heo L, Feig M: **What makes it difficult to refine protein models further via molecular dynamics simulations?** *Proteins.* 2018; **86 Suppl 1**: 177–88.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

45. F Park H, Ovchinnikov S, Kim DE, *et al.*: **Protein homology model refinement by large-scale energy optimization.** *Proc Natl Acad Sci U S A.* 2018; **115**(12): 3054–9.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

46. Hardin C, Pogorelov TV, Luthey-Schulten Z: ***Ab initio* protein structure prediction.** *Curr Opin Struct Biol.* 2002; **12**(2): 176–81.
PubMed Abstract | Publisher Full Text

47. Bowers PM, Strauss CE, Baker D: ***De novo* protein structure determination using sparse NMR data.** *J Biomol NMR.* 2000; **18**(4): 311–8.
Publisher Full Text

48. Lindert S, McCammon JA: **Improved cryoEM-Guided Iterative Molecular Dynamics--Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction.** *J Chem Theory Comput.* 2015; **11**(3): 1337–46.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Tang Y, Huang YJ, Hopf TA, *et al.*: **Protein structure determination by combining sparse NMR data with evolutionary couplings.** *Nat Methods.* 2015; **12**(8): 751–4.
PubMed Abstract | Publisher Full Text | Free Full Text

50. MacCallum JL, Perez A, Dill KA: **Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference.** *Proc Natl Acad Sci U S A.* 2015; **112**(22): 6985–90.
PubMed Abstract | Publisher Full Text | Free Full Text

51. F Perez A, MacCallum JL, Dill KA: **Accelerating molecular simulations of proteins using Bayesian inference on weak information.** *Proc Natl Acad Sci U S A.* 2015; **112**(38): 11846–51.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

52. Perez A, Morrone JA, Brini E, *et al.*: **Blind protein structure prediction using accelerated free-energy simulations.** *Sci Adv.* 2016; **2**(11): e1601274.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Shirts M, Pande VS: **COMPUTING: Screen Savers of the World Unite!** *Science.* 2000; **290**(5498): 1903–4.
PubMed Abstract | Publisher Full Text

54. Beberg AL, Ensign DL, Jayachandran G, *et al.*: **Folding@home: Lessons from**

eight years of volunteer distributed computing. In:; Rome, Italy. 1–8.
Publisher Full Text

55. Snow CD, Nguyen H, Pande VS, *et al.*: **Absolute comparison of simulated and experimental protein-folding dynamics.** *Nature.* 2002; **420**(6911): 102–6.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

56. Kohlhoff KJ, Shukla D, Lawrenz M, *et al.*: **Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways.** *Nat Chem.* 2014; **6**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

57. Shukla D, Meng Y, Roux B, *et al.*: **Activation pathway of Src kinase reveals intermediate states as targets for drug design.** *Nat Commun.* 2014; **5**: 3397.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

58. Eastman P, Swails J, Chodera JD, *et al.*: **OpenMM 7: Rapid development of high performance algorithms for molecular dynamics.** *PLoS Comput Biol.* 2017; **13**(7): e1005659.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

59. Chodera JD, Noé F: **Markov state models of biomolecular conformational dynamics.** *Curr Opin Struct Biol.* 2014; **25**: 135–44.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Shukla D, Hernández CX, Weber JK, *et al.*: **Markov state models provide insights into dynamic modulation of protein function.** *Acc Chem Res.* 2015; **48**(2): 414–22.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Wang W, Cao S, Zhu L, *et al.*: **Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules.** *WIREs Comput Mol Sci.* 2018; **8**: e1343.
Publisher Full Text | F1000 Recommendation

62. Harrigan MP, Sultan MM, Hernández CX, *et al.*: **MSMBuilder: Statistical Models for Biomolecular Dynamics.** *Biophys J.* 2017; **112**(1): 10–15.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Husic BE, Pande VS: **Markov State Models: From an Art to a Science.** *J Am Chem Soc.* 2018; **140**(7): 2386–96.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

64. Mittal S, Shukla D: **Recruiting machine learning methods for molecular simulations of proteins.** *Mol Simul.* 2018; **44**(11): 891–904.
Publisher Full Text | F1000 Recommendation

65. Shaw DE, Chao JC, Eastwood MP, *et al.*: **Anton, a special-purpose machine for molecular dynamics simulation.** *Commun ACM.* 2008; **51**(7): 91–97.
Publisher Full Text

66. Shaw DE, Grossman JP, Bank JA, *et al.*: **Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer.** In *International Conference for High Performance Computing, Networking, Storage and Analysis.* Piscataway, NJ: IEEE; 2014; 41–53.
Publisher Full Text

67. Pan AC, Weinreich TM, Piana S, *et al.*: **Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems.** *J Chem Theory Comput.* 2016; **12**(3): 1360–7.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

68. Lindorff-Larsen K, Maragakis P, Piana S, *et al.*: **Picosecond to Millisecond Structural Dynamics in Human Ubiquitin.** *J Phys Chem B.* 2016; **120**(33): 8313–20.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

69. Lindorff-Larsen K, Piana S, Dror RO, *et al.*: **How fast-folding proteins fold.** *Science.* 2011; **334**(6055): 517–20.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

70. Chung HS, Piana-Agostinetti S, Shaw DE, *et al.*: **Structural origin of slow diffusion in protein folding.** *Science.* 2015; **349**(6255): 1504–10.
PubMed Abstract | Publisher Full Text

71. Shell MS, Ozkan SB, Voelz V, *et al.*: **Blind test of physics-based prediction of protein structures.** *Biophys J.* 2009; **96**(3): 917–24.
PubMed Abstract | Publisher Full Text | Free Full Text

72. PAULING L, COREY RB, BRANSON HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.** *Proc Natl Acad Sci U S A.* 1951; **37**(4): 205–11.
PubMed Abstract | Publisher Full Text | Free Full Text

73. Pauling L, Corey RB: **Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets.** *Proc Natl Acad Sci U S A.* 1951; **37**(11): 729–40.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Rost B: **Review: protein secondary structure prediction continues to rise.** *J Struct Biol.* 2001; **134**(2–3): 204–18.
PubMed Abstract | Publisher Full Text

75. Chou PY, Fasman GD: **Prediction of protein conformation.** *Biochemistry.* 1974; **13**(2): 222–45.
PubMed Abstract | Publisher Full Text

76. Garnier J, Osguthorpe DJ, Robson B: **Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.** *J Mol Biol.* 1978; **120**(1): 97–120.
PubMed Abstract | Publisher Full Text

77. Holley LH, Karplus M: **Protein secondary structure prediction with a neural network.** *Proc Natl Acad Sci U S A.* 1989; **86**(1): 152–6.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Zvelebil MJ, Barton GJ, Taylor WR, *et al.*: **Prediction of protein secondary structure and active sites using the alignment of homologous sequences.** *J Mol Biol.* 1987; **195**(4): 957–61.
PubMed Abstract | Publisher Full Text

79. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol.* 1993; **232**(2): 584–99.
PubMed Abstract | Publisher Full Text

80. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol.* 1999; **292**(2): 195–202.
PubMed Abstract | Publisher Full Text

81. Spencer M, Eickholt J, Jianlin Cheng: **A Deep Learning Network Approach to *ab initio* Protein Secondary Structure Prediction.** *IEEE/ACM Trans Comput Biol Bioinform.* 2015; **12**(1): 103–12.
PubMed Abstract | Publisher Full Text | Free Full Text

82. Wang S, Peng J, Ma J, *et al.*: **Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields.** *Sci Rep.* 2016; **6**: 18962.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

83. Anishchenko I, Ovchinnikov S, Kamisetty H, *et al.*: **Origins of coevolution between residues distant in protein 3D structures.** *Proc Natl Acad Sci U S A.* 2017; **114**(34): 9122–7.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

84. Levy RM, Haldane A, Flynn WF: **Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness.** *Curr Opin Struct Biol.* 2017; **43**: 55–62.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

85. Morcos F, Pagnani A, Lunt B, *et al.*: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proc Natl Acad Sci U S A.* 2011; **108**(49): E1293–301.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

86. Cline MS, Karplus K, Lathrop RH, *et al.*: **Information-theoretic dissection of pairwise contact potentials.** *Proteins.* 2002; **49**(1): 7–14.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

87. Altschuh D, Lesk AM, Bloomer AC, *et al.*: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol.* 1987; **193**(4): 693–707.
PubMed Abstract | Publisher Full Text

88. Göbel U, Sander C, Schneider R, *et al.*: **Correlated mutations and residue contacts in proteins.** *Proteins.* 1994; **18**(4): 309–17.
PubMed Abstract | Publisher Full Text

89. Shindyalov IN, Kolchanov NA, Sander C: **Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?** *Protein Eng.* 1994; **7**(3): 349–58.
PubMed Abstract | Publisher Full Text

90. Lapedes A, Giraud B, Jarzynski C: **Using sequence alignments to predict protein structure and stability with high accuracy.** *arXiv.* 2002; **29**.
Reference Source

91. Weigt M, White RA, Szurmant H, *et al.*: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proc Natl Acad Sci U S A.* 2009; **106**(1): 67–72.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

92. Jones DT, Buchan DW, Cozzetto D, *et al.*: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.** *Bioinformatics.* 2012; **28**(2): 184–90.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

93. Balakrishnan S, Kamisetty H, Carbonell JG, *et al.*: **Learning generative models for protein fold families.** *Proteins.* 2011; **79**(4): 1061–78.
PubMed Abstract | Publisher Full Text

94. Ekeberg M, Lövkvist C, Lan Y, *et al.*: **Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2013; **87**(1): 12707.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

95. Barton JP, De Leonardis E, Coucke A, *et al.*: **ACE: adaptive cluster expansion for maximum entropy graphical model inference.** *Bioinformatics.* 2016; **32**(20): 3089–97.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

96. Sutto L, Marsili S, Valencia A, *et al.*: **From residue coevolution to protein conformational ensembles and functional dynamics.** *Proc Natl Acad Sci U S A.* 2015; **112**(44): 13567–72.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

97. Haldane A, Flynn WF, He P, *et al.*: **Structural propensities of kinase family proteins from a Potts model of residue co-variation.** *Protein Sci.* 2016; **25**(8): 1378–84.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

98. Marks DS, Colwell LJ, Sheridan R, *et al.*: **Protein 3D structure computed from evolutionary sequence variation.** *PLoS One.* 2011; **6**(12): e28766.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

99. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation.** *Nat Biotechnol.* 2012; **30**(11): 1072–80.
PubMed Abstract | Publisher Full Text | Free Full Text

100. F Hopf TA, Schärfe CPI, Rodrigues JP, *et al.*: **Sequence co-evolution gives 3D contacts and structures of protein complexes.** *eLife.* 2014; **3**: e03430.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

101. F Morcos F, Jana B, Hwa T, *et al.*: **Coevolutionary signals across protein lineages help capture multiple protein conformations.** *Proc Natl Acad Sci U S A.* 2013; **110**(51): 20533–8.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

102. F De Leonardis E, Lutz B, Ratz S, *et al.*: **Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction.** *Nucleic Acids Res.* 2015; **43**(21): 10444–55.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

103. F Weinreb C, Riesselman AJ, Ingraham JB, *et al.*: **3D RNA and Functional Interactions from Evolutionary Couplings.** *Cell.* 2016; **165**(4): 963–75.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

104. F Miao Z, Westhof E: **RNA Structure: Advances and Assessment of 3D Structure Prediction.** *Annu Rev Biophys.* 2017; **46**: 483–503.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

105. F Toth-Petroczy A, Palmedo P, Ingraham J, *et al.*: **Structured States of Disordered Proteins from Genomic Sequences.** *Cell.* 2016; **167**(1): 158–170.e12.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

106. F Hopf TA, Ingraham JB, Poelwijk FJ, *et al.*: **Mutation effects predicted from sequence co-variation.** *Nat Biotechnol.* 2017; **35**(2): 128–35.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

107. F Jones DT, Singh T, Kosciolek T, *et al.*: **MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins.** *Bioinformatics.* 2015; **31**(7): 999–1006.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

108. F Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, *et al.*: **Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age.** *Proteins.* 2018; **86 Suppl 1**: 51–66.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

109. Bonneau R, Baker D: **Ab initio protein structure prediction: progress and prospects.** *Annu Rev Biophys Biomol Struct.* 2001; **30**: 173–89.
PubMed Abstract | Publisher Full Text

110. Zhang Y: **Progress and challenges in protein structure prediction.** *Curr Opin Struct Biol.* 2008; **18**(3): 342–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Editorial Note on the Review Process

F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty and are edited as a service to readers. In order to make these reviews as comprehensive and accessible as possible, the referees provide input before publication and only the final, revised version is published. The referees who approved the final version are listed with their names and affiliations but without their reports on earlier versions (any comments will already have been addressed in the published version).

The referees who approved this article are:

Version 1

1 **Dong Xu** [1,2] [1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA
[2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, USA
Competing Interests: No competing interests were disclosed.

1 **Amarda Shehu** Department of Computer Science, George Mason University, Fairfax, VA, USA
Competing Interests: No competing interests were disclosed.