Inferring macromolecular structure from evolutionary couplings

Kee Myoung Nam
Department of Systems Biology
Harvard Medical School
Boston, MA, USA

(Dated: May 13, 2016)

This document reviews the method of inferring macromolecular structure from sequence covariation, which infers three-dimensional contacts between residues based on the assumption that correlations between residues may provide information about the spatial proximity of residues. This method thus casts the problem of structure prediction as one of "inverse statistical mechanics," in which correlations between residues are defined as parameters of a maximum entropy model of a multiple sequence alignment.

I. INTRODUCTION

The prediction of the three-dimensional structures of proteins is a fundamental open problem in molecular biology, with vast implications in both basic and pharmacological research on myriad diseases. As a result, a host of different strategies for predicting protein structures have been proposed, many of which rely on some kind of preliminary structural information regarding a homologous protein (homology mapping).

A more recent strategy leverages amino acid coevolution in protein families to infer 3D contacts. The reasoning is as follows: the maintenance of energetically favorable interactions between residues, as well as overall protein function, may require spatially proximal residues to coevolve across a protein family. One might hypothesize, for instance, that the alteration or loss of a critical residue in the binding pocket of an enzyme may affect the evolutionary trajectory of other residues in the binding pocket, thus giving rise to a distinct pattern of coevolution. This further implies that correlations between residues may provide information about 3D contacts within a protein [4, 5], as well as contacts that may arise through oligomerization, protein-protein interactions, and other protein-substrate interactions. This provides an insight into a possible strategy for inferring structure from sequence: given an amino acid sequence, compile its evolutionary family (from a public repository, such as Pfam [7]), compute a multiple sequence alignment, and use correlated pairs of residues to, in a principled manner, infer 3D contacts.

This last step is particularly challenging, in that simply measuring correlations between residues is insufficient to accurately infer 3D contacts: "indirect" or "transitive" correlations between noninteracting residues, say i and j, may arise as a result of "direct" correlations (i.e., correlations that indeed constitute 3D contacts) between each of i and j with a third residue k. In fact, it has been demonstrated that "local" measures of covariation, such as mutual information (MI):

$$\mathrm{MI}(i,j) = \sum_{\sigma,\tau\in\Sigma} p(i=\sigma,\, j=\tau)\log\left(\frac{p(i=\sigma,\, j=\tau)}{p(i=\sigma)\,p(j=\tau)}\right),$$

where Σ is the set of possible symbols in an alignment (i.e., the 20 proteinogenic amino acids and the gap symbol), fails the distinguish between direct and transitive correlations [4]. MI is a local measure of covariation in the sense that its value for a pair of sites depends only on the two sites and not on the rest of the alignment. On the other hand, disentangling direct and transitive correlations requires a global probabilistic model of the alignment; this is precisely what the method of *evolutionary couplings*, otherwise known as *direct coupling analysis*, entails. In short, the method of evolutionary couplings develops a global probabilistic model, based on the principle of maximum entropy, of a multiple sequence alignment built from a protein family that successfully disentangles direct and transitive correlations. These pairs of directly correlated residues, termed "evolutionary couplings" (Fig. 1) have often been found to correspond to 3D contacts, and are capable of identifying residue-residue interactions crucial to the overall protein function, and—more broadly—can lead to the design of accurate all-atom models of proteins.

II. EVOLUTIONARY COUPLINGS

The method of evolutionary couplings seeks to develop a global probabilistic model $p(\boldsymbol{\sigma})$ for any sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$, where L is the length of the alignment. We also require that the marginal distributions $p(\sigma_i)$ and $p(\sigma_i, \sigma_j)$ generate the empirical frequencies $f(\sigma_i)$ and $f(\sigma_i, \sigma_j)$:

$$p(\sigma_i) = \sum_{\sigma_k\in\Sigma|k\neq i} p(\boldsymbol{\sigma}) = f(\sigma_i)$$
$$p(\sigma_i, \sigma_j) = \sum_{\sigma_k\in\Sigma|k\neq i,j} p(\boldsymbol{\sigma}) = f(\sigma_i, \sigma_j).$$

(We could also, in principle, constrain marginals over K sites for $K > 2$; this, however, leads to an enormous blowup in the number of parameters to be inferred.) The principle of maximum entropy implies that the joint distribution with the greatest entropy, given the above constraints, is a Boltzmann distribution of the form

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L} h_i(\sigma_i) + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)\right),$$

where Z is the partition function, h_i is a single-site bias for site i, and J_{ij} is a "coupling" term between residues i and j. It is immediately clear that this expression is analogous to one for the probability of a configuration in a spin system: more specifically, this expression is analogous to a 21-state Potts model, in which the sum in the exponent is a Hamiltonian comprised of spin-spin interaction energies J_{ij} and single-spin energies h_i due to an external magnetic field.

The "inverse statistical mechanics" problem of inferring J_{ij} from a given multiple sequence alignment may be solved using any one of a number of approximate schemes; here, we describe one such scheme, based on the Plefka expansion for disordered Ising spin glasses [2, 4, 5]. First, note that

$$Z = \sum_{\boldsymbol{\sigma}} \exp\left(\sum_{i} h_i(\sigma_i) + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)\right).$$

Introducing a small perturbative parameter α and taking the Legendre transform of the free energy $-\log Z(\alpha)$, we get the Gibbs potential

$$\Gamma(\alpha) = \log Z(\alpha) - \sum_{i=1}^{L} \sum_{\sigma_i \in \Sigma} h_i(\sigma_i) P_i(\sigma_i).$$

Expanding Γ about $\alpha = 0$ up to first order, we get

$$\Gamma(\alpha) = \Gamma(0) + \left.\frac{\partial \Gamma(\alpha)}{\partial \alpha}\right|_{\alpha=0} + \mathcal{O}(\alpha^2),$$

from which it can be shown that the couplings J_{ij} may be obtained by inverting the covariance matrix $C_{ij} = f_{ij}(\sigma_i, \sigma_j) - f_i(\sigma_i) f_j(\sigma_j)$ [4, 5].

Other methods for inferring the coupling terms include directly maximizing the likelihood function corresponding to $p(\boldsymbol{\sigma})$ via gradient descent, in a manner more commonplace in the machine learning literature [8]. Furthermore, additional molecular modeling is necessary to convert the coupling terms into predicted 3D contacts; possible procedures are explained in [4, 8].

III. DIVERSITY AND STRUCTURE

If, in principle, it is possible to predict a protein's tertiary contacts, especially those relevant to the protein's function, from multiple sequence alignments, it is natural to ask how the sequence diversity relates to the accuracy of predicted contacts. One can imagine, for instance, that all of the sequences are very similar (low diversity), then every pair of sites would be highly correlated, the evolutionary couplings would not be meaningful, and the structure prediction would suffer. On the other hand, if all of the sequences are highly dissimilar in the sense that all 20 amino acids seem to occur at each site uniformly at random (high diversity, at least in the sense of Shannon entropy), then every pair of sites would be close to uncorrelated, and so the evolutionary couplings would again not be meaningful and the structure prediction would suffer. This implies that there may be a finite, optimal level of sequence diversity at which the predicted protein structure is most accurate.

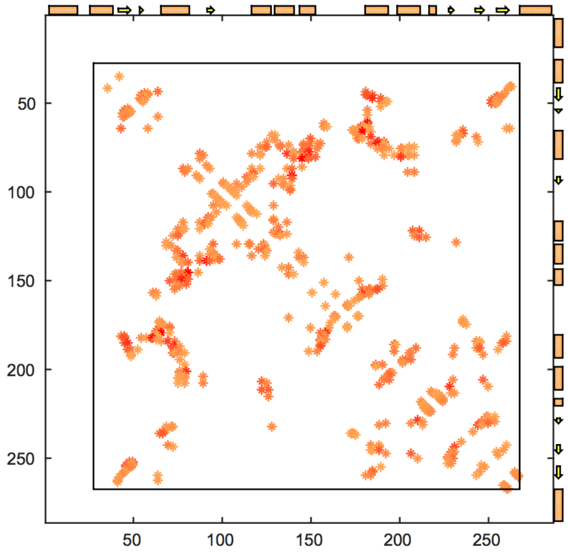

FIG. 1. The top 290 evolutionary couplings for a family of 10030 bacterial beta-lactamases (Pfam ID: PF13354), computed at `evfold.org`.

In this rudimentary analysis, we take two measures of diversity, mean site-wise normalized richness R and mean site-wise normalized Shannon entropy S, to quantify the diversity of a multiple sequence alignment. Namely, we define the mean normalized richness R_i and mean normalized Shannon entropy S_i for a site i as

$$R_i = \frac{1}{21} \sum_{\sigma_i \in \Sigma} f(\sigma_i)$$

and

$$S_i = \frac{1}{21} \sum_{\sigma_i \in \Sigma} p(\sigma_i) \log p(\sigma_i),$$

and $R = \langle R_i \rangle$ and $S = \langle S_i \rangle$. We note that richness and Shannon entropy are the two simplest examples of the diversity indices introduced in [1, 3], and this analysis may be generalized to other measures of diversity.

Fig. 2 shows the relationship between sequence diversity, as formulated via R and S, and the structure prediction accuracy, here quantified as the mean root-mean-square deviation (RMSD) between α-carbons in the predicted structure and those in a crystal structure, for 15

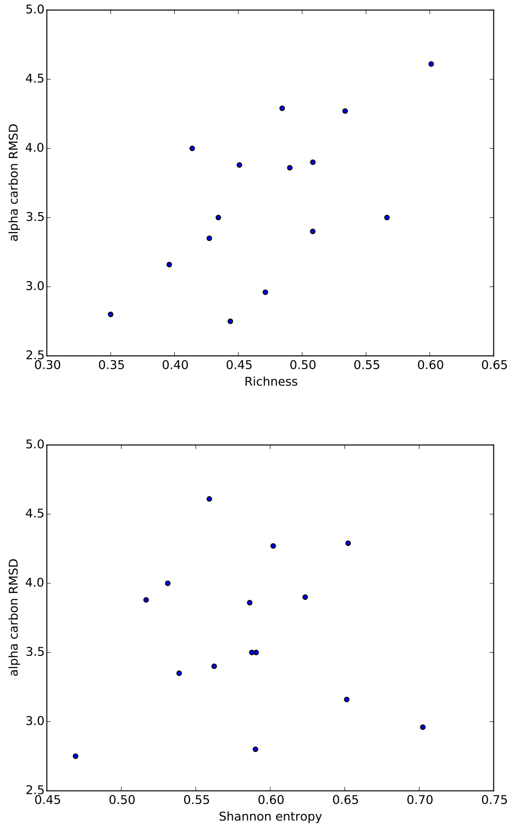

FIG. 2. Distribution of root-mean-square deviations (RMSDs) between the α-carbons of the predicted structure and a reference crystal structure for 15 distinct protein families tested in [4] as functions of richness R and Shannon entropy S.

distinct protein families [4]. (Note that high prediction accuracy implies low RMSD.) We note that, as of current writing, the relationship between diversity and predicted structure accuracy is unclear: in the case of richness, it seems like there exists a *negative* correlation between richness and prediction accuracy, somewhat contrary to expectations. In the case of entropy, it seems that the consideration of other protein families must be used to find a possible correlation between entropy and prediction accuracy. It is at least clear that sequence covariation is nonrandom, as expected: all the families seem to limit their richness and entropy to a range between 0.35 and 0.70, which is somewhat striking in light of the presumed variability in selective pressures to which these proteins are subjected.

IV. DISCUSSION

Here, we have introduced a method for inferring 3D contacts from the evolutionary record of a protein, without the aid of information directly concerning the protein's structure. This method, based on the insight that spatially proximal residues may exhibit correlations, uses a multiple sequence alignment to learn the parameters of a global probabilistic model analogous to a 21-state Potts model with single-spin and spin-spin interaction energies. We have also provided the beginnings of an analysis of the relationship between sequence diversity and structure prediction accuracy across 15 protein families, finding that richness and Shannon entropy may not correlate well with prediction accuracy.

The analysis presented in this document, being quite rudimentary, opens the door to further work regarding the link between diversity, structure, and function. A more thorough investigation into the relationship between sequence diversity and the evolutionary couplings may involve the generation of artificial protein families with arbitrary levels of diversity by simulating protein evolution. However, this would require a site-dependent model of protein evolution that also accounts for epistatic interactions, and moreover it would be impossible to validate conclusions about the structures of these simulated proteins.

In another interesting avenue of research, one could combine phylogenetics and evolutionary couplings by computing residue correlations for chronologically nested subsets of the alignment, and thereby identifying the emergence of correlated residues over evolutionary timescales. This would allow us to trace an evolutionary history of residues with possible functional importance. Furthermore, one could ask if there is a lower or upper bound on the amount of diversity in a family of proteins, and how the interplay of various selective pressures may allow a family to remain within a given range in the level of diversity.

Finally, recent work has demonstrated the success of evolutionary couplings in predicting the tertiary structures of RNA molecules [8] and the effects of mutations on protein function [6]. The tertiary structures of RNA molecules may, in particular, elucidate hitherto unknown mechanisms in the regulatory mechanisms that underlie their function. One could also imagine applying variations of this model to investigate the coevolution of other features of macromolecules, such as the DNA binding sites of transcription factors.

[1] M O Hill, Diversity and evenness: a unifying notation and its consequences. *Ecology* (1973) 54, 2, 427–432.

[2] A Georges, J Yedidia, How to expand around mean-field theory using high-temperature expansions. *J Phys A Math*

Gen 24, 2173–2192.

[3] W S J Valdar, Scoring residue conservation. *Proteins Struct Funct Genet* (2002) 48, 227–241.

[4] D S Marks, L J Colwell, R Sheridan, T A Hopf, A Pagnani, R Zecchina, C Sander, Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE* (2011) 6, 12, e28766.

[5] F Morcos, A Pagnani, B Lunt, A Bertolino, D S Marks, C Sander, R Zecchina, J N Onuchic, T Hwa, M Weight, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* (2011) 108, 49, E1293–E1301.

[6] T A Hopf, J B Ingraham, F J Poelwijk, M Springer, C Sander, D S Marks, Quantification of the effect of mutations using a global probability model of natural sequence variation. (2015) arXiv:1510.04612.

[7] R D Finn, P Coggill, R Y Eberhardt, S R Eddy, J Mistry, A L Mitchell, S C Potter, M Punta, M Qureshi, A Sangrador–Vegas, G A Salazar, J Tate, A Bateman, The Pfam protein families database: towards a more sustainable future. *Nucl Acids Res* (2016) 44, D279–D285.

[8] C Weinreb, A J Riesselman, J B Ingraham, T Gross, C Sander, D S Marks, 3D RNA and functional interactions from evolutionary couplings. *Cell* (2016) 165, 963–975.