Supplementary Note

An information theoretic covariation framework reveals a tunable allosteric network in the highly conserved group II chaperonins

Lopez, Tom tomlopez@stanford.edu
Dalton, Kevin kmdalton@stanford.edu
Frydman, Judith jfrydman@stanford.edu

Contents

1 Description of Average Product Corrected Mutual Information

1.1 The Mathematical Description of a Protein Multiple Sequence Alignment

A protein multiple sequence alignment (MSA) can be described by a matrix, $\{M \in R^{nxt} : m_{ij} \in \{1, 2, \ldots, 21\}\}$, with n residues and t taxa. Each vector, M_i, represents a residue in the MSA. Each entry m_{ij} takes one of 21 values corresponding to each of the amino acids plus the gap character. The marginal probability distribution of residues, $P(M_i) \in R^{21}$ is the vector of marginal probabilities for each amino acid.

$$p_k(M_i) = \frac{1}{t} \sum_{s=1}^{t} \delta(M_{is}, k)$$

Where $k \in \{1, 2, \ldots, 21\}$ is an amino acid and the Kronecker delta, $\delta(a, b)$ is defined as:

$$\delta(a, b) = \begin{cases} a = b & 1 \\ a \neq b & 0 \end{cases}$$

Likewise, the joint probability density of a residue pair, M_i, M_j can be described by $P(M_i, M_j) \in R^{21x21}$, wherein

$$p_{kl}(M_i, M_j) = \frac{1}{t} \sum_{s=1}^{t} \delta(M_{is}, k)\delta(M_{js}, l)$$

1.2 Shannon Entropy

Shannon entropy quantifies the amount of information required to encode a probability distribution. It can be applied to marginal or joint probability distributions. The Shannon entropy of a particular residue, M_i, can be computed as follows:

$$H(M_i) = -\sum_{k=1}^{21} p_k(M_i) \log p_k(M_i)$$

In this context, the Shannon entropy describes the degree of conservation of a given residue. Larger values correspond to less conservation. The base of the logarithm determines the units of the resulting entropy. It is most common in computational disciplines to use $\log_2$ such that the corresponding entropy is measured in *bits*. It is also common to employ the natural logarithm in place of $\log_2$. In which case, the resulting entropy is expressed in units called *nats*.

Shannon entropy can also be computed for the joint probability distribution governing two residues M_i, M_j:

$$H(M_i, M_j) = -\sum_{k=1}^{21} \sum_{l=1}^{21} p_{kl}(M_i, M_j) \log p_{kl}(M_i, M_j)$$

We refer to this quantity as the joint entropy.

1.3 Mutual Information

Mutual information measures the amount of information shared between two variables. It describes how much can be learned about the second variable by observing the first. The mutual information between two residues M_i, M_j in the MSA is given by:

$$I(M_i; M_j) = \sum_{k=1}^{21} \sum_{l=1}^{21} p_{kl}(M_i, M_j) \log \frac{p_{kl}(M_i, M_j)}{p_k(M_i) p_l(M_j)}$$

Mutual information has an equivalent representation in terms of entropies:

$$I(M_i; M_j) = H(M_i) + H(M_j) - H(M_i, M_j)$$

1.4 Normalized Mutual Information

Normalized mutual information was introduced as a technology for multiple sequence alignment analysis in 2005 [7]. It is computed for a residue pair, M_i, M_J, by dividing the mutual information by the joint entropy:

$$NMI(M_i; M_j) = \frac{I(M_i; M_j)}{H(M_i, M_j)}$$

NMI is an intriguing modification of mutual information. Unlike raw mutual information, it is restricted to values between zero and one, $NMI(M_i; M_j) \in (0, 1]$. As we have shown using synthetic alignments, NMI (Figure 1,S1) has very little dependence upon the entropy of residues in the alignment. Compared to mutual information, it is less sensitive to artifacts introduced by residue conservation[4].

Additionally, NMI may be used to construct a normalized distance metric on the space of MSA residues:

$$D(M_i, M_j) = 1 - NMI(M_i; M_j)$$

This quantity has very desirable qualities. It is still normalized in the sense that $D(M_i, M_j) \in [0, 1)$. Furthermore it satisfies the criteria of a metric [11]:

$$D(a, b) \geq 0 \tag{1}$$
$$D(a, b) = 0 \iff a = b \tag{2}$$
$$D(a, b) = D(b, a) \tag{3}$$
$$D(a, b) \leq D(a, c) + D(b, c) \tag{4}$$

1.5 The Average Product Correction

The average product correction was introduced in 2008 [2].

$$APC(M_i, M_j) = \frac{1}{n^2} \sum_{r=1}^{n} I(M_i, M_r) I(M_j, M_r)$$

Average product corrected mutual information (MIp) is currently the best performing information theoretic measure for amino acid covariation as judged by contact prediction [6]. More accurate contact prediction algorithms have emerged since the beginning of our study. Such methods rely on optimization routines involving psuedo-likelihoods and/or sparse priors [1, 8, 5]. They are computationally costly as compared to simpler information theoretic measures. Simple measures like MIp offer a nice compromise with intermediate accuracy and ease of computation.

$$MIp(M_i, M_j) = I(M_i, M_j) - APC(M_i, M_j)$$

Despite its favorable performance, MIp has a similar entropy dependence to uncorrected MI (Figure 1, S1). We show here that this entropy dependence can be partially mitigated by normalizing to the joint entropy yielding the following measure:

$$APC\text{-}NMI(M_i, M_j) = \frac{I(M_i, M_j) - APC(M_i, M_j)}{H(M_i, M_j)}$$

This simple measure is readily computed and has less dependence on entropy than MIp (Figure 1, S1).

1.6 Limitations of Covariation Measures in Inferring Co-evolution

The underlying phylogeny of biological samples have long been known to corrupt covariates between taxa [3]. This is a serious consideration when performing covariation analysis on multiple sequence alignments. Put succinctly, the underlying phylogeny of a sample may impact the observed pattern of covariation. The average product correction is purported to account for this bias to some extent. However, recent work from the molecular evolution community[9] suggests that covariation signal recovered from MIp strongly depends on the rate of evolution. Particularly, this study reports that high scoring residue pairs correspond to slowly evolving residues. The authors raise the concern that residues in the protein core tend to evolve more slowly and are thereby more likely to be selected by MIp. Such arguments suggest that the favorable performance of MIp in the task of contact prediction may be artifactual, because core residues are close in 3D space.

There are further issues confounding the inference of proper co-evolution from covariation. Namely, the sampling bias inherent in extant sequences renders covariation estimates inaccurate irrespective of the underlying phylogeny.

Covariation relies on the assumption that sequences are drawn at random from the underlying distribution of sequences comprising a protein family. This assumption will be violated any time that an MSA is constructed from extant sequences owing to the bias of sequence databases toward model organisms.

The final issue confounding the inference of co-evolution from multiple sequence alignments is the issue of conservation. Residue covariation measures including SCA5 and MI have been shown to depend strongly on the entropy of residues in the MSA [10, 4]. We and others [7] have shown that this bias can be mitigated by normalizing to the joint entropy.

1.7 Rationale for choosing Met47 for further analysis

We obtained 14 subnetworks in our APC-NMI analysis (Fig. 1). When these 14 subnetworks were mapped on the structure, it was clear that most of them could be explained by covariation due to structural considerations. The Supplemental Figure Fig. S4C, showing 3 of the 14 subnetworks mapped onto the Cpn structure, highlights several important points (also clearly observable in pymol session included as supplemental information):

- Most subnetworks are small and correspond to spatially contiguous interactors, which probably is the reason why these residues covary
- The fact that for most subnetworks mapping onto the structure shows that they engage in structural interactions between residues non-contiguous in sequence but spatially coupled is in itself quite nice and validates our approach reveals meaningful interactions. However, these subnetworks are less interesting in terms of understanding allostery.
- The largest network corresponds to residues centered around the nucleotide: clearly this is the network of interest to understand the allostery of ATP. Validating the functional significance of this subnetwork for ATPase function, many of its residues have been functionally linked to the ATPase function of Cpns. Other residues in the network that are not in contact with the neighboring subunits may indeed have interesting properties of their own, but would not be involved in the complex-wide phenomena we were looking to study. Further, we did not want to introduce the complication of potentially destabilizing the structure of individual subunits by mutating core residues. This left the interfacial residue Met 47 as our best candidate for further analysis, since it connected the networks in two adjacent subunits. Considering how laborious it is to study allostery and understand its mechanism, the integration of structural and computational insights is essential.
- We note that all residues in the network (Fig 1F) are given high scores along the first component of the APC-NMI matrix indicating that they are contributing to the same underlying process. We consider the spectral decomposition of the network as an equally important indicator of covariation to the raw pairwise scores.

References

[1] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins*, 79(4):1061–1078, April 2011.

[2] S D Dunn, L M Wahl, and G B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 1 February 2008.

[3] Joseph Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, 125(1):1–15, 1985.

[4] Anthony A Fodor and Richard W Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–221, 1 August 2004.

[5] David T Jones, Daniel W A Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 15 January 2012.

[6] Wenzhi Mao, Cihan Kaya, Anindita Dutta, Amnon Horovitz, and Ivet Bahar. Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics*, 31(12):1929–1937, 15 June 2015.

[7] L C Martin, G B Gloor, S D Dunn, and L M Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, 15 November 2005.

[8] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.*, 108(49):E1293–301, 6 December 2011.

[9] David Talavera, Simon C Lovell, and Simon Whelan. Covariation is a poor measure of molecular coevolution. *Mol. Biol. Evol.*, 32(9):2456–2468, September 2015.

[10] Tiberiu Teşileanu, Lucy J Colwell, and Stanislas Leibler. Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput. Biol.*, 11(2):e1004091, February 2015.

[11] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11(Oct):2837–2854, 2010.