Power law tails in phylogenetic systems

Chongli Qin* and Lucy J. Colwell†§

*Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

†To whom correspondence should be addressed. Email: ljc37@cam.ac.uk.

Associate Editor: Michael Levitt, Stanford University, Stanford, CA

Covariance analysis of protein sequence alignments uses coevolving pairs of sequence positions to predict features of protein structure and function. However, current methods ignore the phylogenetic relationships between sequences, potentially corrupting the identification of covarying positions. Here, we use random matrix theory to demonstrate the existence of a power law tail that distinguishes the spectrum of covariance caused by phylogeny from that caused by structural interactions. The power law is essentially independent of the phylogenetic tree topology, depending on just two parameters—the sequence length and the average branch length. We demonstrate that these power law tails are ubiquitous in the large protein sequence alignments used to predict contacts in 3D structure, as predicted by our theory. This suggests that to decouple phylogenetic effects from the interactions between sequence distal sites that control biological function, it is necessary to remove or down-weight the eigenvectors of the covariance matrix with largest eigenvalues. We confirm that truncating these eigenvectors improves contact prediction.

Keywords: power law | sequence covariance | phylogeny | protein | structure prediction

Although some approaches to biological sequence analysis typically assume that mutations at different sites are independent of each other, this approximation is clearly limited. Indeed, covariance between sequence distal positions is important for predicting RNA secondary structure (1), where Watson–Crick base-pairing rules create strong covariance signals that can be detected by straightforward methods. In contrast, for proteins, the signal is less strong, and for many years it was unclear whether any remnant of molecular phenotypes such as protein structure is imprinted on covarying sequence positions (2–4). Recently, with the growth of protein sequence databases (5) and the introduction of sophisticated analyses (6–8), it has become clear that covariance analysis of protein sequences can yield exciting biological insights in a wide range of contexts (9–27). In general a set of homologous protein sequences is constrained by protein structure and function, and with sufficient data it is possible to tease out the nature of these constraints and make biologically relevant predictions (12, 13, 16, 28–32).

An important consideration that limits our ability to infer sets of covarying residues is sequence phylogeny, i.e., the relatedness structure of the data samples (33–35). If some population subgroups are more closely related, then part of the covariation observed in the data will be of purely phylogenetic origin, unrelated to molecular phenotypes such as structure or function (36–41). In population and medical genetics features such as geographical population structure are known to affect the degree of covariance observed between sequences (42–44).

This raises the question of whether, given a set of aligned protein sequences of length \( p \), it is possible to distinguish covariance due to phylogeny from that caused by molecular phenotypes (36–41). Here, we analyze a simple theoretical model of molecular evolution and use the tools of random matrix theory (RMT) to develop a theory for the covariance when both phylogeny and structural constraints are present. We show that phylogenetic covariance is distinguished by a power law tail of large eigenvalues, which is essentially independent of phylogenetic details, depending only on the average branch length \( m/p \) and the number \( b \) of branching events or generations.

Thus motivated, we turn to data and find that the eigenvalue distributions of covariance matrices from large protein sequence alignments (MSAs) have power law tails. This suggests a strategy for cleaning the covariance matrix that at least partly controls for confounding phylogenetic effects: removing the power law tail representing those modes that are most strongly corrupted by phylogeny. For several protein families, we show that contact prediction accuracy improves by excluding those eigenvectors that correspond to the largest eigenvalues. It is interesting to note that the commonly used method of inverting the sample covariance matrix similarly down-weights the largest eigenvalues and up-weights the smallest ones. Our analysis therefore gives an alternative rationalization for why direct coupling analysis (DCA) has proved so successful at inferring true contacts in proteins from sequence data alone. More generally, this eigenvalue power law will occur in any dataset where the samples have a similar hierarchical relationship.

Results

Molecular phenotypes cause covariance between sequence positions (columns) of the MSA matrix \( X \), while phylogeny causes covariance between sequences (rows) of \( X \). Covariance from either source will appear in both the residue covariance matrix \( C_R = X^T X / p \) and the sequence covariance matrix \( C_S = \frac{XX^T}{p} \). It is clear that \( C_R \) and \( C_S \) contain the same information; they have the same nonzero eigenvalues, and their eigenvectors \( V_R \) and \( V_S \) are related by \( V_R = X^T V_S \) and \( V_S = XV_R \). Analyses of protein sequence data typically attribute the detected covariance signal to interactions between sequence positions. This can be misleading: Fig. 1A and B shows \( C_R \) and \( C_S \) for a simulated dataset where phylogeny is the only source of covariance. Note that \( C_R \) contains isolated high-scoring residue

Significance

Covariance analysis of protein sequence alignments can predict structure and function from sequence alignments alone. Current methodologies typically assume that sequences are independent, notwithstanding their phylogenetic relationships. This corruption constrains the alignments for which covariance analysis can be used. It is critically important to control for phylogeny and understand how phylogeny contaminates signal. This paper presents a mathematical analysis that argues that there is a distinctive signature of phylogeny in the covariance matrix, allowing us to identify modes that are corrupted by phylogeny. This signature is present in large protein sequence alignments, explaining recent covariance analyses, and provides an important step toward decoupling phylogenetic effects from biologically meaningful interactions.

Author contributions: C.Q. and L.J.C. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

1To whom correspondence should be addressed. Email: ljc37@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711913115/-/DCSupplemental.
pairs caused by phylogeny, which could be erroneously interpreted to be caused by molecular phenotypes.

What happens if there are structural interactions between specific residue pairs in the simulation? In Fig. 1C and D we compare the true interactions (gray) with the top 200 off-diagonal pairs of \( \Sigma \) in the absence of phylogeny (red, true interaction; blue, false positive). (D) The addition of phylogeny to these simulations confounds the signal.

Phylogenetic Covariance. To understand the signature of phylogenetic covariance, we consider a Markov model where mutations occur at random and different sites evolve independently. The process starts with a random sequence of length \( p \), drawn from a \( q \)-letter alphabet, which undergoes a series of mutation and duplication events dictated by a user-imposed phylogeny with \( b \) branching events. This generates an alignment of \( n = 2^b \) simulated sequences. Population structure changes the eigenvalue spectrum of the resulting covariance matrix. To see this, consider the simplest phylogeny, consisting of a single branching event with equal-length branches. The true covariance matrix \( \Sigma_S \), i.e., the covariance matrix of the distribution the samples are drawn from, follows by calculating the covariance between the resulting sequences \( x_i \) and \( x_j \). Since this is a stationary Markov process, the covariance between two sequences separated by \( 2m \) mutations, which we denote \( \alpha(m) \), is \( \mathbb{E}(x_i(2m)x_j(0)) \), which yields

\[
\alpha(m) = \exp \left[-\frac{2qm}{(q-1)p}\right] = \exp[-4m/p], \tag{1}
\]

where the last equality specializes to a binary alphabet. A phylogeny with a single branching event has the true covariance matrix

\[
\Sigma_S = \frac{1}{\alpha} \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}. \tag{2}
\]

As the mutation rate \( m \to \infty \), note that \( \alpha \to 0 \). This means that \( \Sigma_S \to I \), i.e., the sequences are uncorrelated, and phylogenetic influence is negligible. More generally, as the number of branching events or generations \( b \) increases, we find that \( \Sigma_S \) is composed of nested squares that correspond to each branching event. This yields \( b + 1 \) distinct eigenvalues \( \lambda_i \), with \( P(\lambda = \lambda_i) = p_i \propto 2^{-b} \), except for the two largest eigenvalues, which have \( p_i \propto 2^{-b} \). (Supporting Information). These relationships imply that the eigenvalues follow the power law

\[
\lambda \sim r^\beta, \tag{3}
\]

where \( r \) is the rank, and \( \beta \propto \log 2n/p \) is a function of \( m/p \). Under the influence of phylogeny, the maximum eigenvalue increases exponentially with the number of branching events \( b \). Note that there is a precise threshold at \( 2n = 1 \), which given Eq. 1 for \( \alpha \) implies \( 2qm/p( q-1) = \ln(2) \), above which this power law behavior occurs.

Finite-Sampling Effects. We have thus seen that phylogeny produces a striking signature in the covariance matrix. However, because the number of MSA sequences is limited, this signature will be affected by finite sampling—the sample covariance matrix will contain large entries purely by chance. We use RMT to develop a quantitative characterization of the effect on the corresponding eigenvalue distribution. Consider \( n \) independent sequences of length \( p \), with amino acids drawn uniformly at random. The probability distribution of the sample eigenvalues follows the Marcenko–Pastur (MP) distribution

\[
f(\lambda) = \frac{\sqrt{(b_+ - \lambda)(\lambda - b_-)}}{2\pi \lambda c}, \quad b_{\pm} = (1 \pm \sqrt{c})^2, \tag{4}
\]

where \( c = n/p \). Our simulations confirm that the histogram of eigenvalues of sequences simulated without phylogeny or structural interactions is well described by this analytical formula (Fig. 2A). As \( n \) increases, Eq. 4 implies that this distribution sharpens around \( \lambda = 1 \). RMT further predicts how Eq. 4 generalizes to describe the eigenvalue distribution of the sample covariance matrix \( \Sigma \) for any true covariance matrix \( \Sigma \), such as those caused by phylogeny. We start with the Stieltjes transform,

\[
G(z, c) = \int_{-\infty}^{\infty} \frac{dF(\lambda)}{z - \lambda}, \tag{5}
\]

where \( F(\lambda) \) is the cumulative distribution function of \( f(\lambda) \), the limiting eigenvalue distribution of \( C \). Marcenko and Pastur (45) used the method of characteristics to relate \( G(z, c) \) to \( T(\lambda) \), the cumulative eigenvalue distribution of the true covariance matrix \( \Sigma \), yielding

\[
G(z, c) = -1 \left( z - c \int_{-\infty}^{\infty} \frac{\lambda dT(\lambda)}{1 + \lambda G(z, c)} \right)^{-1}. \tag{6}
\]

This equation describes the effects of finite sampling. If the true eigenvalues cluster at/near unity, this will result in the MP distribution of Eq. 4. For phylogeny, the eigenvalues of \( \Sigma \) are drawn from a discrete distribution, so \( dT(\lambda) = \sum \delta(\lambda - \lambda_i) \delta d\lambda \), which follows the power law of Eq. 3. Eq. 6 describes how finite sampling smoothens out this discrete distribution.

Fig. 2B shows how the eigenvalue distribution changes if the sequences follow our simplest phylogeny, where \( \Sigma_S \) (Eq. 2) has
This is used to plot the red curve in Fig. 2. Applying Eq. 6 we find

\[ z - c \sum_{i=0}^{\infty} \frac{\rho_i (1 + \alpha_i)}{1 + (1 + \alpha_i) G} = - \frac{G}{1} . \]

Examining the summands, we note that

\[ \sum_{i=0}^{\infty} \frac{\rho_i (1 + \alpha_i)}{1 + (1 + \alpha_i) G} = \frac{1}{G} \left[ 1 + \sum_{\alpha_i \neq 1} \rho_i \frac{1 + \alpha_i}{1 + (1 + \alpha_i) G} \right] . \]

In the limit of large \( p \) the dependence on the tree parameters \( \rho_i \) and \( \alpha_i \) simplifies, so that

\[ \sum_{i=0}^{\infty} \rho_i (1 + \alpha_i)^2 = \exp \left( \frac{2 \mu (e^{-q/p} - 1)}{1 + (e^{-2q/p} - 1)} \right) e^{-2\mu q/p} . \]

This approximation, valid for large \( p \), allows us to write

\[ \sum_{i=0}^{\infty} \frac{\rho_i (1 + \alpha_i)}{1 + (1 + \alpha_i) G} \approx \frac{1 + e^{-2\mu q/p} - 1)}{1 + (e^{-2q/p} - 1)} G . \]

Hence the Stieltjes transform for the inhomogeneous tree is equal to the Stieltjes transform for a homogeneous tree with \( m = \mu \) the mean of the distribution the branch lengths are drawn from. This result can be generalized for any arbitrary distribution and phylogenetic tree topology (Supporting Information).

This result about inhomogeneous phylogenies is important as it extends our analysis methods to more realistic phylogenies, implying that the power law tail of large eigenvalues described above is general.

**Phenotypic Covariance.** The eigenvalue spectrum for phenotypic covariance depends on how phenotype couples the residues to each other. While this will differ for different phenotypes, recent work has focused on using covariance analysis to predict contacts in tertiary protein structure (11, 13, 14, 16–18). If we consider interactions drawn from a protein contact map, what covariance is caused? For an alphabet with \( q = 2 \), the correlation between two residues that interact with strength \( j \) is given by \( \tanh(j) \), which saturates as \( j \) increases so that the resulting correlation does not exceed unity. With multiple interactions and a larger alphabet, the situation is more complex; however, we can use simulations to characterize the sample covariance matrix and corresponding eigenvalue distribution. We first simulate sequences without phylogeny, using a simple Markov model with nonzero residue couplings at locations dictated by protein contact maps. These couplings were chosen uniformly from the interval \([-5, 5]\). With the 784 interactions of Fig. 3A, the eigenvalue distribution of the sample covariance matrix is described well by the MP distribution (Fig. 3B). This empirical observation suggests that the eigenvalues of the true covariance matrix are all of similar size, suggesting that structural interactions do not lead to an eigenvalue power law. While real proteins will also have other phenotypic interactions, this model provides a relevant starting point.
The point at which the eigenvalues deviate drops off before the yellow curve. Importantly, A shows a log–log plot of the eigenvalues as a function of rank $\beta$ vol. 115, $\lambda$ A–C 1 Top January 23, 2018 Top $\geq \cdots \geq$ shows that the phylogenetic parameters extracted B B $\cdots$ = 1 are each found using the phylogenetic parameters from A–C. With inter-

and 693 and 4. Fig 4 D shows that the interaction matrix, and $B$ is the spectrum of the covariance matrix of the resulting sequence alignment. $B$, Inset shows the upper edge of the eigenvalue distribution in more detail; compare with Fig. 2D.

Phylogenetic vs. Structural Covariance. Crucially, this model suggests that there are strikingly different signatures between the covariance matrix expected from phylogeny and that expected for interactions caused by residue contacts. If only structural interactions are present, the limiting behavior of the maximum eigenvalue saturates logarithmically as a function of the number of interactions (Fig. 4A). In contrast, Fig. 4B shows that the maximum eigenvalue caused by phylogeny increases exponentially as the sequences undergo more duplication events. Moreover, Fig. 4C shows a log–log plot of the eigenvalues as a function of rank for our simulations with just phenotypic interactions (Fig. 3); the data are well fitted by a line of slope zero reflecting the absence of the power law.

To probe these signatures further, we use simulations with a controlled mix of phylogeny and structural interactions. Fig 4D shows that the spectra for simulations with just phylogeny and simulations with both phylogeny and 200 random structural interactions obey the same power law. In both cases the upper power law tail follows $\beta = \log(2\alpha) / \log(2)$ (red line). With interactions, the lower extent of the power law is diminished; the blue curve in Fig. 4D drops off before the yellow curve. Importantly, these two spectra diverge only outside the power law regime, implying that phylogeny dominates those modes that follow the power law.

These simulations therefore suggest that interactions between residues affect the smallest eigenvalues, while phylogeny affects the largest eigenvalues, giving a potential mechanism for distinguishing the effects of phylogeny. Intuitively, this could arise because interactions between residues make it less likely that mutations at those sites will be accepted, reducing the effective mutation rate of these residues and hence affecting eigenvectors with low eigenvalues. In Fig 5 we simulate sets of sequences with both phylogeny and structural interactions from two different protein contact maps and obtain similar results to those in Fig 4D. In contrast to Fig. 3, we find that the eigenvalue distributions of the resulting sequence alignments are not MP, but are well fitted by our analytic approach. The red curves in Fig. 5A and B are each found using the phylogenetic parameters from the power law fits in Fig. 5C and D, respectively.

Eigenvalue Spectra of Protein Sequence Data. Given the vastly different signatures in the eigenvalue distributions expected from phylogeny and structural interactions, it is of great interest to see whether such signatures arise in protein sequence data. To probe this, we choose three representative protein families for which covariance analysis has been shown to yield accurate contact predictions. In Fig. 6 A–C, Top we show that the eigenvalue distributions follow a power law in each case, as predicted by our theory. Furthermore, as for the simulated data, Fig. 6 A–C, Middle shows that the phylogenetic parameters extracted from the power fitted in each case provide a closer fit (red curves) to the eigenvalue distribution than to the MP distribution (blue curves).

Cleaning Protein Spectra. The analysis of simulated data suggests that the effects of phylogeny can be diminished by removing large modes of the covariance matrix and enforcing the constraint that the remaining eigenvalues are all of the same size. Namely, instead of the full covariance matrix from the sequence alignments, we propose truncating the highest modes, $C(t) = \nu_1\nu_1^T + \cdots + \nu_t\nu_t^T$, $\lambda_1 \geq \cdots \geq \lambda_t \geq \cdots \geq \lambda_r$, where $r = p(q - 1)$. Fig. 6 shows the results of this approach for contact prediction. For each protein, the slope of the power law fit in Fig. 6 A–C, Top is used to estimate the phylogenetic parameters required for the analytical solution in Fig. 6 A–C, Middle (red curve). The point at which the eigenvalues deviate from the power law fit in Fig. 6 A–C, Top (purple dashed line) is used to determine which modes are dominated by phylogeny and should be truncated from the outer product expansion of the sample covariance matrix. Fig. 6 A–C, Bottom shows how well different truncations do at contact prediction; the purple dashed line reflects the threshold found from the power law fit and is near optimal in all cases. This phenomenon is entirely consistent with the notion that the modes corresponding to the largest eigenvalues reflect the phylogenetic relatedness of the aligned sequences.

Discussion
This paper was motivated by recent advances (9–14, 16, 21) in predicting protein structure and function from the covariance of sequences, a strategy that has been successful for predicting RNA secondary structure for some time (1, 35). A major confounding effect in both situations is the effect of phylogeny,
Fig. 5. Simulations with phylogeny and interactions. Here sequences are simulated with phylogeny and interactions taken from the contact map of (A) DHFR, using $m/p = 0.068$, and (B) Trypsin, using $m/p = 0.059$. A and B, Top show the histograms of eigenvalues, compared with the MP distribution (blue curve); Insets show the contact maps. These $m/p$ values are used to compute the analytical distributions (red curves) which match the data well. A and B, Bottom show log–log plots of the eigenvalues as a function of rank. The predicted slope is calculated from the value of $\alpha(m/p)$ using Eq. 1 in each case and provides an excellent fit.

which introduces correlations between residues (30, 36, 38). The correlations due to structure/function and phylogeny must be disentangled for accurate prediction.

The primary accomplishment of this paper is to identify a feature of the eigenvalue distribution of protein covariance matrices (the power law tail) that distinguishes covariance due to phylogeny from that caused by structural interactions. The presence of power law tails in the data from diverse protein families allows us to develop an initial approach to deconvolving structural interactions from the covariance that results from sequence phylogeny alone. Our finding that the largest modes of the covariance matrix are dominated by phylogeny suggests an alternative rationalization for the matrix inversion step that enabled features of protein structure and function to be predicted from covariance analysis of large protein sequence alignments. Furthermore the resulting cleaned covariance matrix can be used as input for other inference approaches (9–12, 18, 19, 21).

A further result is a general understanding of how phylogenetic effects impact sequence covariation in different regions of parameter space. Depending on the sequence length $p$ and the average branch length $m$, there is a parameter regime where the covariance matrix does not feature a power law tail of large eigenvalues, and hence a different approach to disentangling phenotypic interactions from phylogenetic correlations is required. Specifically, as the eigenvalues of the true covariance matrix for phylogenetic interactions are $\approx (2\alpha)^k$, we expect large eigenvalues when $2\alpha > 1$. Given Eq. 1 for $\alpha$, this is equivalent to $2q/(q-1)m/p < \ln(2)$.

We have focused on the eigenvalue distribution; however, information about the phylogeny will also be imprinted in the eigenvectors of the covariance matrix. In the phylogenetic regime, the eigenvectors will have structure that reflects the relationship between the different sequences (43, 44), providing additional information about which modes should be removed.

Fig. 6. Protein sequence alignments follow the power law, and moreover spectral deviation from the power law can be used to deconvolve the influence of phylogeny from the covariance matrix, and facilitate contact prediction. A–C show analysis of protein sequence data from (A) Trypsin, (B) DHFR, and (C) TRML-HAEIN, a knotted tRNA-methyltransferase. In A–C, Top we show that the eigenvalues of each protein sequence alignment follow a power law. The purple dashed line indicates the point at which the spectrum deviates from this power law, indicating a threshold above which phylogeny dominates the spectrum. The parameter $m$ is inferred from this power law using the equation $\lambda \sim r^{m/p}$, where $m = \log 2\alpha/\log 2$ and $\alpha(m/p)$ is given by Eq. 1. The inferred values of $m$ are used to plot the red lines in A–C, Middle, which provide a good fit to the empirical spectral distributions. A–C, Bottom show that the phylogenetic threshold, derived from A–C, Top, provides an excellent indication of which modes should be removed from the covariance matrix to deconvolve the influence of phylogeny and dramatically improve contact prediction using just the covariance matrix.

which introduces correlations between residues (30, 36, 38). The correlations due to structure/function and phylogeny must be disentangled for accurate prediction.
for better inference of phenotypic interactions. Understanding the extent to which the effects of phylogeny and structural/functional interactions can be disentangled is an important direction for future research. Is it possible to separate the effects of phylogeny from those of interaction in parameter regimes with no power law tail? Under what circumstances can we accurately infer the strength of interactions? The approach outlined here provides a mathematical framework that future work can exploit to definitively answer these questions.

ACKNOWLEDGMENTS. We thank M. P. Brenner and A. W. Murray for comments on a draft of this paper. This work was supported by a Next Generation fellowship (to L.J.C.), a Marie Curie Career Integration Grant [Evo-Couplings, Grant 631609], and an Engineering and Physical Sciences Research Council PhD studentship (to C.Q.).

1. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge Univ Press, Cambridge, UK).
2. Altschul S, Leuk A, Bloomer A, Kugl A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 193:693–707.
3. Shindyalov I, Kolchanov N, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7:349–358.
4. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299.
5. Finn RD, et al. (2016) The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res 44:D279–D285.
6. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630.
7. Lapedes AS, Giraud BG, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: Phylogenetic and structural effects. Statistics in Molecular Biology and Genetics, IMS Lecture Notes - Monograph Series, ed Seiller-Moiseiwitsch F (Institute of Mathematical Statistics, Hayward, CA), pp 236–256.
8. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions. arXiv:0712.4397.
9. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 4:165.
10. Skerker JM, et al. (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133:1043–1054.
11. Weigt M, White RA, Szymant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106:67–72.
12. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. Cell 138:774–786.
13. Marks DS, et al. (2011) Protein 3d structure computed from evolutionary sequence variation. PloS One 6:e28766.
14. Dahirel V, et al. (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci USA 108:11520–11525.
15. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108:E1293–E1301.
16. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149:1607–1621.
17. Šulcoveška JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci USA 109:10340–10345.
18. Jones DT, Buchan DW, Cozzetto D, Ponti M (2012) Piccov: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184–190.
19. Ekeberg M, Lückvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys Rev E 87:021707.
20. Ferguson AL, et al. (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunology 38:606–617.
21. Ochvinitnikov S, Kaminetsky H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife 3:e02030.
22. De Leonardi E, et al. (2015) Protein and RNA structure prediction by integration of co-evolutionary information into molecular simulation. Biophys J 108:13a–14a.
23. Tang Y, et al. (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods 12:751–754.
24. Barton JP, Kardar M, Chakraborty AK (2015) Scaling laws describe memories of host-pathogen riposte in the HIV population. Proc Natl Acad Sci USA 112:1965–1970.
25. Weinreb C, et al. (2016) 3D RNA and functional interactions from evolutionary couplings. Cell 165:963–975.
26. Sung YM, Wilkins AD, Rodriguez GJ, Wensel TG, Lichtarge O (2016) Intramolecular allosteric communication in dopamine d2 receptor revealed by evolutionary amino acid coevolution. Proc Natl Acad Sci USA 113:3539–3544.
27. Bitbol AF, Ouyver RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA 113:12180–12185.
28. Shakhnovich EI, Gutin AM (1993) Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 90:7195–7199.
29. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358.
30. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. Mol Biol Evol 17:164–178.
31. Cocco S, Monasson R, Weigt M (2013) From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. PloS Comput Biol 9:e1003176.
32. Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson R (2015) Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. PLoS Comput Biol 12:e1004889.
33. Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15.
34. Rivers E (2005) Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics 6:63.
35. Rivers E, Clements J, Eddy SR (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in incRNAs. Nat Methods 14:45–48.
36. Altschul SF, Carroll RJ, Lipman DJ (1989) Weights for data related by a tree. J Mol Biol 207:647–653.
37. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA 97:3288–3291.
38. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340.
39. Duthée JY (2012) Detecting coevolving positions in a molecule: Why and how to account for phylogeny. Brief Bioinform 13:228–243.
40. Obermayer B, Levine E (2014) Inverse Ising inference with correlated samples. New J Phys 16:123017.
41. Barton JP, Chakraborty AK, Cocco S, Jacquin H, Monasson R (2016) On the entropy of protein families. J Stat Phys 162:1267–1293.
42. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PloS Genet 2:e190.
43. Price AL, et al. (2009) The impact of divergent time on the nature of population structure: An example from Iceland. PloS Genet 5:e1000505.
44. McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5:e1000868.
45. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. Mat Sb 114:507–536.
46. Rao NR, Edelman A (2008) The polynomial method for random matrices. Found Comput Math 8:649–702.