Power Law Tails In Phylogenetic Systems

Chongli Qin[a] **and Lucy J. Colwell**[a]

[a] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

This manuscript was compiled on October 26, 2017

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 downweight the eigenvectors of the covariance matrix with largest eigenvalues. We confirm that truncating these eigenvectors improves contact prediction.

Protein sequence covariance analysis | Co-evolution | Sequencing data | Maximum entropy | Direct coupling analysis

Approaches to biological sequence analysis typically assume that mutations at different sites are independent of each other, though this approximation is clearly limited. Indeed, covariation 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 n 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 analyse 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 rationalisation for why direct coupling analysis (DCA) has proven 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 $C_R = X^T X/n$, and the sequence covariance matrix $C_S = XX^T/p$. This is because C_R and C_S contain the

Significance Statement

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 which 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 towards decoupling phylogenetic effects from biologically meaningful interactions.

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

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. E-mail: ljc37@cam.ac.uk

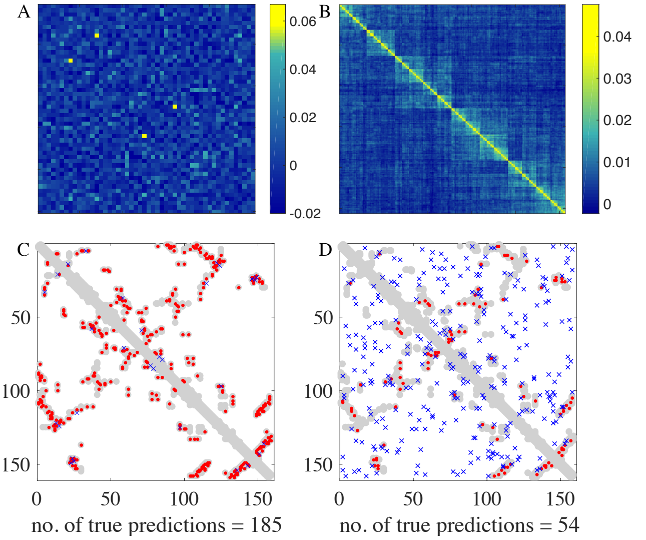

Figure 1: The covariance matrices A) C_R and B) C_S for sequences simulated with just phylogeny, note that C_R has isolated large entries that could be interpreted to indicate interactions between pairs of sequence positions, though none exist in the simulation. C) In simulations where the contact map of DHFR is used to generate interactions (grey), causal interactions are detected well by the largest 200 off-diagonal pairs of C_R in the absence of phylogeny (red = true interaction, blue = false positive). D) The addition of phylogeny to these simulations confounds the signal.

same information; they have the same non-zero eigenvalues, and their eigenvectors V_R and V_S are related by $V_R = X^T V_S$ and $V_S = X V_R$. Analyses of protein sequence data typically attribute the detected covariance signal to interactions between sequence positions. This can be misleading: Figs. ??A, B show 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 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. ??C, D we compare the true interactions (grey) with the top 200 scoring pairs from covariance matrices for sequences simulated (C) without and (D) with phylogeny. Without phylogenetic corruption, 185/200 predictions are correct; whereas with phylogeny this reduces to 54/200. The essential question is to find a way to disentangle phylogenetic and phenotypic (e.g. structural) covariance from matrices that contain a superposition of both (e.g. Fig. ??D). To address this, we first analyse the covariance signal produced by sequences for which the only source of covariance is phylogeny, and then ask if we can distinguish this signal when both phylogenetic and structural correlations are present.

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, a single branching event and equal length branches. The true covariance matrix Σ_S, i.e. the covariance matrix of the distribution the samples are drawn from, follows by calculating the covariance between the resulting sequences $\mathbf{x}_i$ and $\mathbf{x}_j$. Since this is a stationary Markov process, the covariance between two sequences separated by $2m$ mutations, which we denote $\alpha(m)$, is $\mathbf{E}(\mathbf{x}(2m)\mathbf{x}(0))$:

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

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

$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}. \qquad [2]$$

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

$$\lambda \sim r^{\beta} \qquad [3]$$

where r is the rank, and $\beta \propto \log 2\alpha$ 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 $2\alpha = 1$, which given Eq. (1) for α implies $2qm/p(q-1) = ln(2)$, above which this power law behaviour 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 random matrix theory 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 Marčenko-Pastur (MP) distribution:

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

where $c = n/p$ [45]. Our simulations confirm that the histogram of eigenvalues of sequences simulated without phylogeny or structural interactions is well described by this analytical formula (fig. ??A). As n increases, Eq. (4) implies that this distribution sharpens around $\lambda = 1$. Random Matrix theory further predicts how Eq. (4) generalises to describe the eigenvalue distribution of the sample covariance matrix C for any true covariance matrix Σ, such as those caused by phylogeny. We start with the Stieltjes transform:

$$G(z, c) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}F(\lambda)}{z - \lambda}, \qquad [5]$$

where $F(\lambda)$ is the cumulative distribution function of $f(\lambda)$, the limiting eigenvalue distribution of C. Marčenko and Pastur [45] used the method of characteristics to relate $G(z, c)$

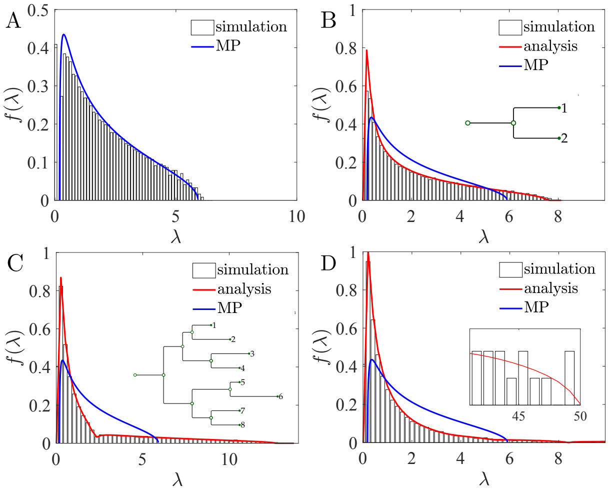

Figure 2: Eigenvalue distributions of (A) n = 4096 sequences with p = 100 residues, drawn from a model with $\Sigma = I$. The eigenvalues of the sequence covariance matrix (bars) match the classical MP distribution (blue curve). (B) Here 2^{11} initial sequences are simulated along a tree with equal branch lengths and a single branching event, with m = 10 mutations per branch. (C) 2^9 initial sequences are evolved with b = 3 branching events (inset), branch lengths drawn from a Poisson distribution with mean 10. (D) 2^5 initial sequences are evolved with b = 7 branching events and branch lengths drawn from a half-normal distribution with mean 10 (Inset: eigenvalue tail). In each case, the histogram of eigenvalues (averaged over 200 runs) matches the analytical solution (red curve), not the classical MP distribution (blue curve).

to $T(\lambda)$, the cumulative eigenvalue distribution of the true covariance matrix Σ, yielding

$$G(z,c) = -1\left(z - c\int_{-\infty}^{\infty}\frac{\lambda \mathrm{d}T(\lambda)}{1+\lambda G(z,c)}\right)^{-1} \qquad [6]$$

This equation describes the effects of finite sampling. If the true eigenvalues cluster at/near unity, this will result in the Marčenko-Pastur (MP) distribution of Eq. (4). For phylogeny, the eigenvalues of Σ are drawn from a discrete distribution, so $dT(\lambda) = \sum_{\mathrm{i}} p_{\mathrm{i}}\delta(\lambda-\lambda_{\mathrm{i}})\mathrm{d}\lambda$ [46], where $p_i = \mathbb{P}(\lambda = \lambda_i)$ follows the power law of Eq. (3). Eq. (6) describes how finite sampling smooths out this discrete distribution.

Fig. ??B shows how the eigenvalue distribution changes if the sequences follow our simplest phylogeny, where Σ_S (Eq. (2)) has eigenvalues $\lambda_{\pm} = 1 \pm \alpha$. Alignments of $n_0 = 2^{11}$ sequences were simulated with $m = 10$ mutations per branch. The shape of the eigenvalue distribution differs significantly from the MP distribution (blue curve). RMT allows us to predict this spectrum using Eq. (6), which becomes

$$z - \frac{c}{2}\left(\frac{1+\alpha}{1+(1+\alpha)G}\right) - \frac{c}{2}\left(\frac{1-\alpha}{1+(1-\alpha)G}\right) = -\frac{1}{G}\,.$$

The inverse Stieltjes transform, given by the positive imaginary part of $G(z,c)$, analytically describes the expected eigenvalue distribution of C_S. This is used to plot the red curve in Fig. ??B, which shows excellent quantitative agreement with the simulation, unlike the MP distribution shown in blue. As the number of branching events increases we simply use our exact formulas (see S.I.) for the true eigenvalue distributions in Eq. (6) to compute the expected distribution.

Analysis of Inhomogeneous Phylogenies. Real phylogenetic trees are inhomogeneous, with branches of different lengths. Our framework naturally extends to this setting. Figs. ??C, D show the eigenvalue distributions of trees drawn from different distributions; Fig. ??C has three branching events with branch lengths drawn from a Poisson distribution, while Fig. ??D has seven branching events with branch lengths drawn from a half normal distribution. Note that the eigenvalue distribution broadens as the number of branching events b increases, reflecting that the maximum true eigenvalue is $\propto \alpha^b$.

For inhomogeneous phylogenies we discovered that analytical solutions follow a simple rule. Consider a phylogeny with branch lengths drawn from a distribution with mean $\langle m\rangle$ and bounded variance, the eigenvalue distribution is then well approximated by the eigenvalue distribution for the tree with all branch lengths equal to $\langle m\rangle$, and the same number of branching events. The red curves in fig. ??C, D show that this prediction fits the simulated data closely. To derive the result, we consider a phylogeny with $b = 1$ and branch lengths m_1, m_2 drawn from a Poisson distribution with mean $\langle m\rangle = \mu$, so that $\rho_i := \mathbf{P}(m_1 + m_2 = i) = (2\mu)^i e^{-2\mu}/i!$. If $\alpha_i = \exp(-qi/p(q-1))$, then the eigenvalues of the true covariance matrix are $\lambda = 1 \pm \alpha_i$. Applying Eq. (6) we find:

$$z - \frac{c}{2}\sum_{i=0}^{\infty}\frac{\rho_i(1+\alpha_i)}{1+(1+\alpha_i)G} - \frac{c}{2}\sum_{i=0}^{\infty}\frac{\rho_i(1-\alpha_i)}{1+(1-\alpha_i)G} = -\frac{1}{G}$$

Examining the summands, we note that

$$\sum_{i=0}^{\infty}\frac{\rho_i(1+\alpha_i)}{1+(1+\alpha_i)G} = \frac{1}{G} - \frac{1}{G(1+G)}\sum_{i=0}^{\infty}\frac{\rho_i}{1+\alpha_i\frac{G}{1+G}}$$

where

$$\sum_{i=0}^{\infty}\frac{\rho_i}{1+\alpha_i\frac{G}{1+G}} = \sum_{i=0}^{\infty}\rho_i\left[1 - \frac{G}{1+G}\alpha_i + \left(\frac{G}{1+G}\right)^2\alpha_i^2 + \cdots\right]$$

In the limit of large p the dependence on the tree parameters, ρ_i and α_i simplifies, so that:

$$\sum_{i=0}^{\infty}\rho_i\,(\alpha_i)^j = \exp\left\{2\mu(e^{-qj/p(q-1)} - 1)\right\} \sim e^{-2\mu qj/p(q-1)}$$

This approximation, valid for large p allows us to write

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

Hence the Stieltjes transform for the inhomogenous 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 generalised for any arbitrary distribution and phylogenetic tree topology (see S.I.). 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.

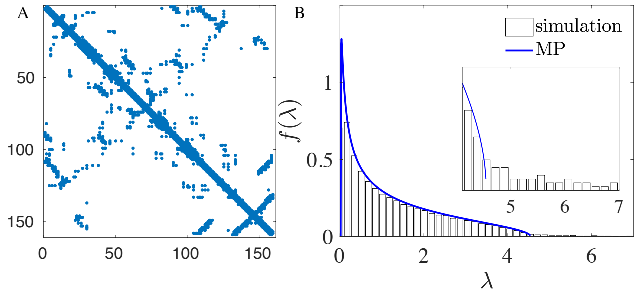

Figure 3: **Simulations with just structural interactions.** Here 4096 sequences are simulated without phylogeny, with structural interactions taken from the contact map of DHFR with strengths uniformly distributed on [-5, 5]. Left panel shows the interaction matrix, right panel is the spectrum of the covariance matrix of the resulting sequence alignment. Inset shows the upper edge of the eigenvalue distribution in more detail, compare with Fig. ??D.

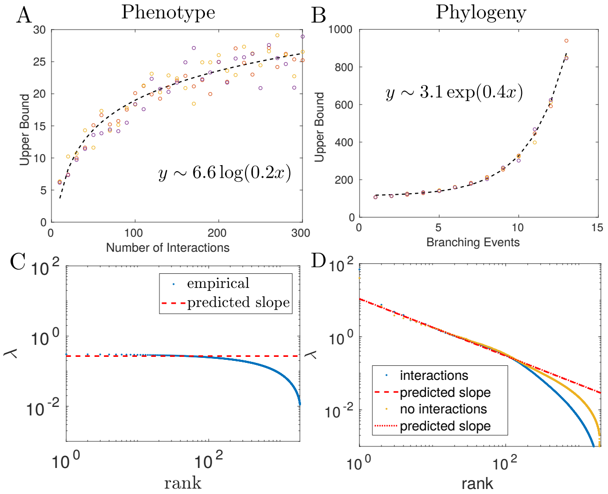

Figure 4: **Largest eigenvalue obtained in simulations** with A) only structural interactions with strength 1.0, and B) only phylogeny. Lines of best fit shown. C) Eigenvalue distribution obtained in simulations from Fig. ?? with just structural interactions, the predicted slope is zero. D) Comparison of eigenvalue distributions obtained in simulations with (yellow) phylogeny, and (blue) phylogeny and structural interactions. The presence of interactions does not alter the power law, but does affect the small eigenvalue behaviour. The lines of best fit are constrained using $\lambda \propto (r)^{\log(2\alpha)/\log(2)}$, where α is from Eq. (1)

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 characterise the sample covariance matrix and corresponding eigenvalue distribution. We first simulate sequences without phylogeny, using a simple Markov model with non-zero 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. ??A, the eigenvalue distribution of the sample covariance matrix is described well by the Marčenko-Pastur distribution (Fig. ??B). 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.

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 behaviour of the maximum eigenvalue saturates logarithmically as a function of the number of interactions (Fig. ??A). In contrast, Fig. ??B shows that the maximum eigenvalue caused by phylogeny increases exponentially as the sequences undergo more duplication events. Moreover, fig. ??C shows a log-log plot of the eigenvalues as a function of rank for our simulations with just phenotypic interactions (see Fig. ??); the data are well fit 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 ??D 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 that 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. ??D drops off before the yellow curve. Importantly, these two spectra only diverge 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 makes 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 ?? we simulate sets of sequences with both phylogeny and structural interactions from two different protein contact maps, and obtain similar results to Fig ??D. In contrast to Fig. ??, we find that the eigenvalue distributions of the resulting sequence alignments are not MP, but are well fit by our analytic approach. The red curves in Fig. ??A, B are each found using the phylogenetic parameters from the power law fits in Figs. ??C, 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 if 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 the top row of Fig. ?? we show that the eigenvalue distributions follow a power law in each case, as predicted by our theory. Furthermore, as for the simulated data, the middle row of Fig. ?? shows that the phylogenetic parameters extracted from the power fitted in each case provide a closer fit (red curves) to the eigenvalue distribution than the MP distribution (blue curves).

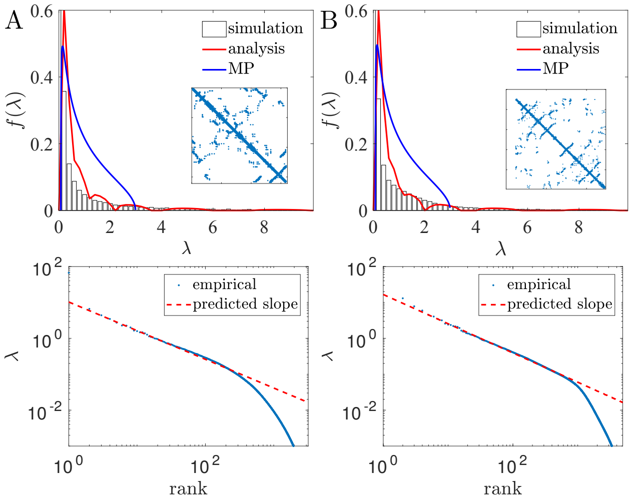

Figure 5: **Simulations with phylogeny and interactions.** Here sequences are simulated with phylogeny and interactions taken from the contact map of A) DHFR, using m/p = 0.068 and B) Trypsin, using m/p = 0.059. Top row shows the histograms of eigenvalues, compare each with the MP distribution (blue curve), insets show the contact maps. These m/p values are used to compute the analytical distributions (red curves) which match the data well. The bottom row shows 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.

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 the same size. Namely, instead of the full covariance matrix from the sequence alignments, we propose truncating the highest modes:

$$C(t) = \mathbf{v}_t\mathbf{v}_t^T + \cdots + \mathbf{v}_r\mathbf{v}_r^T, \quad \lambda_1 \geq \cdots \geq \lambda_t \geq \cdots \geq \lambda_r \, ,$$

where $r = p(q-1)$. Fig. **??** shows the results of this approach for contact prediction. For each protein, the slope of the power law fit in the top row is used to estimate the phylogenetic parameters required for the analytical solution in the middle panel (red curve). The point at which the eigenvalues deviate from the power law fit in the top row (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. The bottom panels show 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 phenomenology is entirely consistent with the notion that the modes corresponding to the large 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 covariation 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, 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 manuscript 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 rationalisation 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 α, 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 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 manuscript. This work was supported by a Next Generation fellowship (LJC), a Marie Curie CIG [Evo-Couplings, Grant 631609], and an Engineering and Physical Sciences Research Council PhD studentship (CQ).

1. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* (Cambridge university press).
2. Altschuh D, Lesk A, Bloomer A, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology* 193(4):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 Engineering* 7(3):349–358.
4. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295–299.
5. Finn RD et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44(D1):D279–285.
6. Jaynes ET (1957) Information Theory and Statistical Mechanics. *Phys. Rev.* 106(4):620–630.
7. Lapedes AS, Giraud BG, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: phylogenetic and structural effects in *Statistics in molecular biology and genetics - IMS Lecture Notes - Monograph Series.* Vol. 33, pp. 236–256.
8. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions. *arXiv preprint 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(6):1043–1054.
11. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue

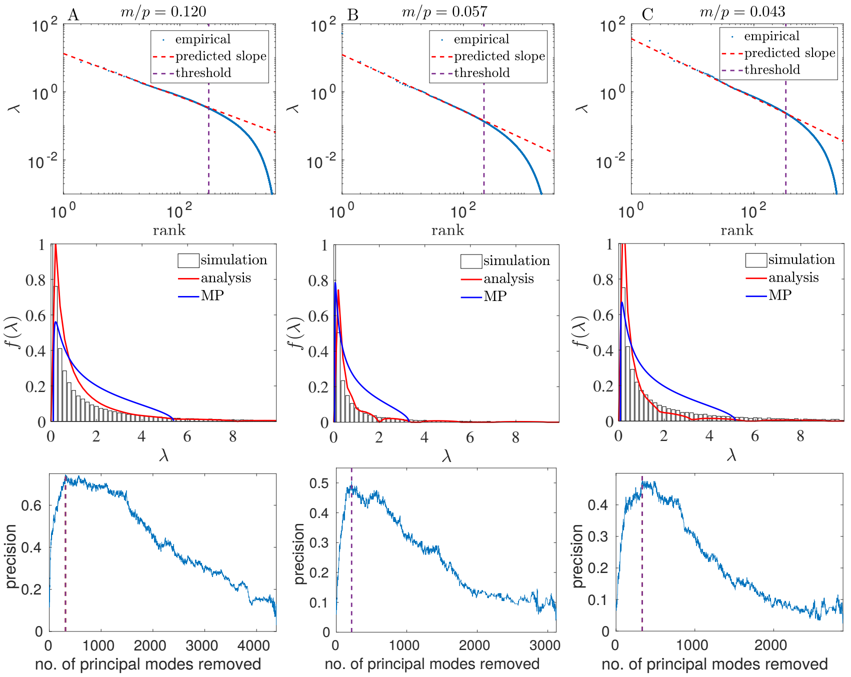

Figure 6: **Protein sequence alignments follow the power law, and moreover spectral deviation from the power law can be used to deconvolve the influence of phylogeny and facilitate contact prediction.** The three panels show analysis of protein sequence data from A) Trypsin, B) DHFR, and C) TRML-HAEIN, a knotted tRNA-methyltransferase. In the first row 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 $\lambda \sim r^{-\beta}$, where $\beta = \log 2\alpha/\log 2$ and $\alpha(m)$ is given by Eq. (1). The inferred values of m are used to plot the red lines in the second row, which provide a good fit to the empirical spectral distributions. The third row of plots show that the phylogenetic threshold, derived from the first row of plots, 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.

contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* 106(1):67–72.

12. N.Halabi, 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(12):e28766.
14. Dahirel V et al. (2011) Coordinate linkage of hiv evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences* 108(28):11530–11535.
15. Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* 108(49):E1293–E1301.
16. Hopf TA et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
17. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* 109(26):10340–10345.
18. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
19. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E* 87(1):012707.
20. Ferguson AL et al. (2013) Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38(3):606–617.
21. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 3:e02030.
22. De Leonardis E et al. (2015) Protein and rna structure prediction by integration of co-evolutionary information into molecular simulation. *Biophysical Journal* 108(2):13a–14a.
23. Tang Y et al. (2015) Protein structure determination by combining sparse nmr data with evolutionary couplings. *Nature methods* 12(8):751–754.
24. Barton JP, Kardar M, Chakraborty AK (2015) Scaling laws describe memories of host–pathogen riposte in the hiv population. *Proceedings of the National Academy of Sciences* 112(7):1965–1970.
25. Weinreb C et al. (2016) 3d rna and functional interactions from evolutionary couplings. *Cell* 165(4):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 covariation. *Proceedings of the National Academy of Sciences* 113(13):3539–3544.
27. Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences* 113(43):12180–12185.
28. Shakhnovich EI, Gutin AM (1993) Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences* 90(15):7195–7199.
29. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology* 257(2):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(1):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(8):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. *bioRxiv* p. 028936.
33. Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist* pp. 1–15.
34. Rivas E (2005) Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC bioinformatics* 6(1):1.
35. Rivas E, Clements J, Eddy SR (2017) A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. *Nature methods* 14(1):45–48.
36. Altschul SF, Carroll RJ, Lipman DJ (1989) Weights for data related by a tree. *Journal of molecular biology* 207(4):647–653.
37. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences* 97(7):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(3):333–340.
39. Dutheil JY (2012) Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Briefings in bioinformatics* 13(2):228–243.
40. Obermayer B, Levine E (2014) Inverse ising inference with correlated samples. *New Journal of Physics* 16(12):123017.
41. Barton JP, Chakraborty AK, Cocco S, Jacquin H, Monasson R (2015) On the entropy of protein families. *Journal of Statistical Physics* pp. 1–27.
42. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS genet* 2(12):e190.
43. Price AL et al. (2009) The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. *PLoS Genetics* 5(6):e1000505.
44. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics* 5(10):e1000686.
45. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik* 114(4):507–536.
46. Rao NR, Edelman A (2008) The polynomial method for random matrices. *Foundations of Computational Mathematics* 8(6):649–702.