Elements of Coevolution in Biological Sequences

Olivier Rivoire
CNRS/UJF-Grenoble 1, LIPhy UMR 5588, Grenoble F-38402, France
(Received 13 December 2012; published 23 April 2013)

Studies of coevolution of amino acids within and between proteins have revealed two types of coevolving units: coevolving contacts, which are pairs of amino acids distant along the sequence but in contact in the three-dimensional structure, and sectors, which are larger groups of structurally connected amino acids that underlie the biochemical properties of proteins. By reconciling two approaches for analyzing correlations in multiple sequence alignments, we link these two findings together and with coevolving units of intermediate size, called "sectons," which are shown to provide additional information. By extending the analysis to the co-occurrence of orthologous genes in bacterial genomes, we also show that the methods and results are general and relevant beyond protein structures.

DOI: 10.1103/PhysRevLett.110.178102 PACS numbers: 87.15.Qt, 02.50.Sk, 87.14.E−, 87.18.Wd

The structural and functional properties of proteins emerge from interactions between their amino acids. During evolution, these interactions constrain the substitutions of amino acids that may happen. Sequences resulting from multiple independent evolutionary trajectories reflect these constraints and, therefore, contain information about the organization of interactions within proteins. Such sequences are now made available by DNA sequencing technology, which provides thousands of protein sequences that have diverged independently and under similar selective pressures from a common ancestral sequence.

These protein sequences are commonly collected into multisequence alignments on the basis of their sequence similarity. An alignment is formally an $M \times L$ array X, where X_{si} indicates which of the $A = 20$ natural amino acids is present at position i in sequence s; some positions contain a gap, inserted to ensure an optimal alignment and represented as a 21st amino acid. Typical numbers are $M \sim 10^2$–10^4 for the number of sequences and $L \sim 10^2$–10^3 for the length of the alignment.

The pattern of functional couplings between amino acids may be inferred from the statistical correlations between pairs of positions in the alignment. Analyses of these correlations are complicated by several factors: (i) proteins are gathered in an alignment based on sequence similarity, with no guarantee to have been subject to common selective constraints; (ii) sequences are not sampled independently during evolution but through a branching process, which introduces a sampling bias; (iii) the information content of the alignment, $\sim ML\log_2 A \sim 10^5$–$10^7$ bits, is small compared to the number $\sim A^2L^2/2 \sim 10^6$–10^8 of continuous parameters defining the correlations between every pair of amino acids, which implies a severe undersampling; (iv) two positions may be correlated while not directly interacting, reflecting a fundamental difference between interactions and correlations.

Standard statistical analyses identify the observed samples to an asymptotically large number of independently and identically distributed random variables. Points (i)–(iii) violate each of these assumptions, while point (iv) suggests that, even in the absence of bias, further processing is required to infer interactions from correlations.

Many approaches have been proposed to tackle these challenges [1]. Recently, two methods have been developed, each rooted in a different concept of statistical mechanics. In an extension of an approach called statistical coupling analysis (SCA) [2], an application of concepts from random matrix theory [3] to address (iii) has revealed collective modes of coevolution named sectors [4]. A protein sector is a group of structurally contacting positions, and experiments indicate that each sector controls independently a biochemical property of the protein [4]. In a different approach called direct coupling analysis (DCA) [5], the problem (iv) of inferring interactions from correlations was formulated and solved as a problem of inverse statistical mechanics, leading to the inference of a large number of pairs of positions in contact in the folded structure [6].

The two approaches, SCA and DCA, differ in their principles as well as in their results. Using the Pfam alignment for the trypsin family [7] as an illustrative example, we show here how they can be connected at different levels. Specifically, we show that (1) their respective measures of coevolution rely on distinct parts of the spectrum of a same covariance matrix, (2) a parallel analysis of the two measures of coevolution reveals different types of coevolving units—previously identified sectors and smaller units, which we call "sectons," and (3) these coevolving units, and the contacting pairs from DCA, stem from different aspects of the data but are interrelated, with sectons and contacting pairs respecting the overarching decomposition into independent sectors.

Given a multiple sequence alignment, SCA and DCA use as input the same basic statistical quantities: the frequency f_i^a of amino acid a at position i and the joint frequency f_{ij}^{ab} of the pair of amino acids a, b at the pair

of positions i, j. Prior to defining these frequencies, some steps must be taken to clean the alignment from positions with excessive gaps and mitigate the effects of (i) and (ii) by weighting differentially the contributions of the various sequences. These steps are straightforward but essential and may be common for both approaches (all details are provided as Supplemental Material [8]).

The frequencies f_{ij}^{ab} and f_i^a define a covariance matrix $C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_i^b$. SCA combines this matrix with a measure of amino acid conservation to define a matrix of conserved correlations $\mathcal{C}_{ij}$, while DCA relies on the inverse $J = -\bar{C}^{-1}$ of a regularized variant of C_{ij}^{ab} (see below) to define a matrix of direct information $\mathcal{D}_{ij}$ [8]. Inspired by previous applications of random matrix theory to the study of covariance matrices [3,9], we analyze here these two matrices by a common method: (1) we compute the eigenvectors associated with the top k_{top} eigenvalues; (2) we rotate these eigenvectors into maximally independent components, $V^{(1)}, \ldots, V^{(k_{\text{top}})}$, using independent component analysis (ICA) [10]; (3) we define coevolving units as sets of positions making largest contributions to a component, $\mathcal{S}_k = \{i\colon V_i^{(k)} > \epsilon\}$. The analysis involves two cutoffs: the number k_{top} of modes that is retained and a threshold $\epsilon > 0$ of significance for the contribution of positions to the components. The results, however, will be shown to be insensitive to the exact values of these cutoffs.

For the SCA matrix $\mathcal{C}_{ij}$ [2], this analysis leads to coevolving units called protein sectors [4]. They are represented in Fig. 1 for the alignment of the trypsin family, using $k_{\text{top}} = 4$ and $\epsilon = 0.1$ (for simplicity, these sectors do not include the positions i with $V_i^{(k)} > \epsilon$ for multiple k; see Fig. S1 [8]). Each sector forms a contacting group of positions on the three-dimensional structure, despite not necessarily consisting of consecutive positions along the sequence. Sectors have no sharp boundaries but are typically organized into an onionlike hierarchy, with the core of sector k consisting of positions i with largest $V_i^{(k)}$ and layers associated with decreasing values of $V_i^{(k)}$, as revealed by varying ϵ and k_{top} (Figs. S2–S6 [8]). Three sectors were previously inferred for the same protein family using an alignment about 10 times smaller [4]: two, the same green and red sectors, correspond to enzymatic activity and specificity respectively (Table SI and Fig. S7 [8]); the third one, which had the peculiarity of a disconnected core, and which correlated experimentally with stability, is now partly spread over two new sectors, whose functional role remains to be characterized.

DCA leads to a matrix $\mathcal{D}_{ij}$ of direct information, previously analyzed by ranking its entries [5]: In a number of protein families, these top entries have been shown to consist of pairs of positions in physical contact in the three-dimensional structure [6] (contacts are defined here by a distance <8 Å). Most of these top pairs are, however,

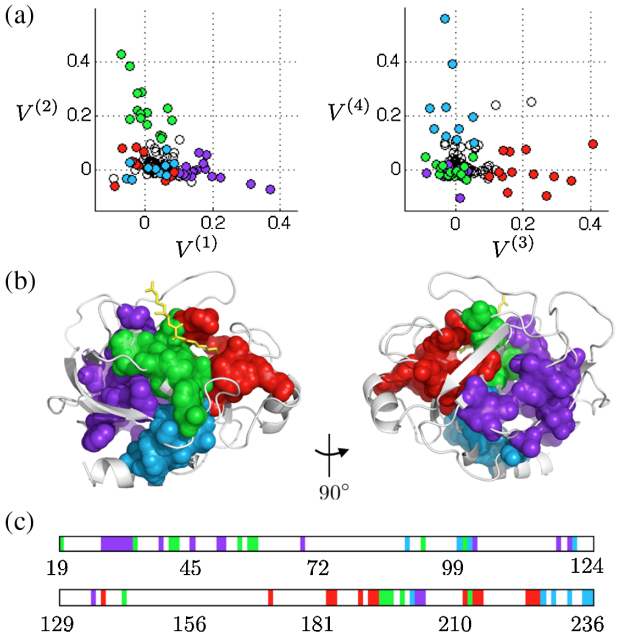

FIG. 1 (color online). Protein sectors in the trypsin family, as inferred from the Pfam alignment PF00089 [7]—(a) Projections of the positions i along the vectors $V^{(k)}$ obtained by rotating by ICA the top $k_{\text{top}} = 4$ eigenvectors of the SCA matrix $\mathcal{C}_{ij}$ [2]: Each dot corresponds to a position i, with coordinates $(V_i^{(1)}, V_i^{(2)})$ in the first graph and $(V_i^{(3)}, V_i^{(4)})$ in the second. Sector k is defined by the positions i with $V_i^{(k)} > \epsilon$ and $V_i^{(\ell)} < \epsilon$ for $\ell \neq k$, with $\epsilon = 0.1$. The positions of each sector are represented with a different color: purple ($k = 1$), green ($k = 2$), red ($k = 3$), and cyan ($k = 4$). (b) Location of the sectors on a three-dimensional structure of trypsin [21]. (c) Location of the sectors along the sequence (cut in two for readability), with nonsector positions in white (numbering system of bovine chymotrypsin).

consecutive along the protein chain, due to the presence of stretches of gaps in the alignment. To discard these trivial contacts, we consider here a truncated matrix $\tilde{\mathcal{D}}_{ij}$, where $\tilde{\mathcal{D}}_{ij} = \mathcal{D}_{ij}$ if $|i - j| > \Delta$, and 0 otherwise, with $\Delta = 5$ (other values give consistent results; Figs. S8–S9 [8]). For the trypsin alignment that serves here as illustration, the top 79 entries of this matrix are found to be in physical contact [Fig. 2(a) and Table SII [8]). The same figure shows that these contacts are not unrelated to sectors but respect the decomposition into independent sectors, with top pairs of $\tilde{\mathcal{D}}_{ij}$ found within sectors, outside sectors, or at the edge of sectors, but almost never intersecting two sectors.

Instead of considering the top elements of $\tilde{\mathcal{D}}_{ij}$, we also analyze here its spectral properties, following the method used to infer sectors from the SCA matrix $\mathcal{C}_{ij}$. This analysis leads to a large number ($\sim$100) of independent components, each localized on a small group of 2–4 positions, which we call protein "sectons" (Table SIII and Fig. S10 [8]). Figure 3 shows the first eight sectons, using $k_{\text{top}} = 120$ and $\epsilon = 0.2$, but similar results are obtained for a

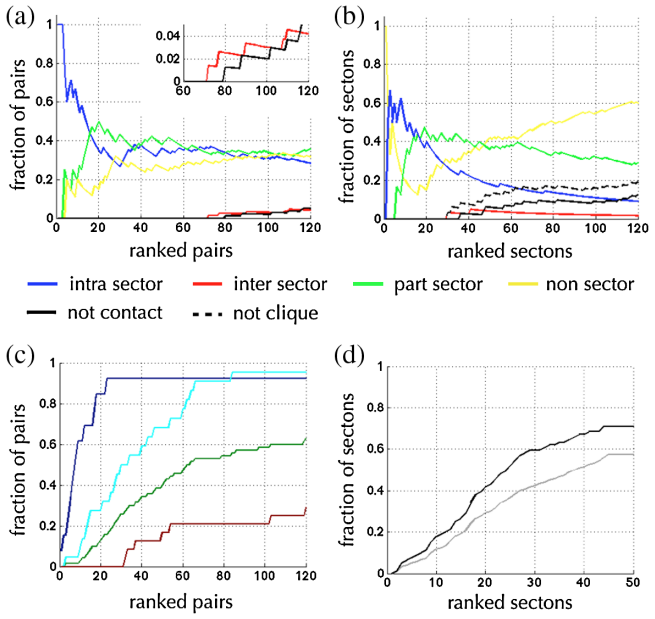

FIG. 2 (color online). Relations between top pairs of $\tilde{\mathcal{D}}_{ij}$, sectors and sectons—(a) Fraction of top pairs ij of $\tilde{\mathcal{D}}_{ij}$, ranked by decreasing value of $\tilde{\mathcal{D}}_{ij}$, that are within a sector (blue curve), across two sectors (red), partly in a sector (green), and outside sectors (yellow). The fraction of pairs not in contact (black) becomes nonzero at rank 80 (zoom in inset). (b) Similar to (a), but for sectons instead of top pairs, and with an extra curve (dotted line) for the fraction of sectons that are not cliques, i.e., with two positions not directly in physical contact, but possibly contacting through other positions in the secton. As top pairs of $\tilde{\mathcal{D}}_{ij}$, sectons respect the decomposition into sectors. (c) Fraction of contacting pairs within sectons of size 2 (blue) or size ≥ 3 (green) that are top pairs of $\tilde{\mathcal{D}}_{ij}$, for the top 35 sectons that are structurally connected. Contacts in sectons of size ≥ 3 can be partitioned into contacts associated with the 2 positions contributing most to the secton (cyan), which are nearly all top pairs of $\tilde{\mathcal{D}}_{ij}$, and other contacts (red), of which only $\sim$20% are top pairs of $\tilde{\mathcal{D}}_{ij}$. (d) Fraction of the top 79 (black) or 120 (gray) pairs of $\tilde{\mathcal{D}}_{ij}$ contained in a secton: $\sim$30% of these top pairs are not in a secton.

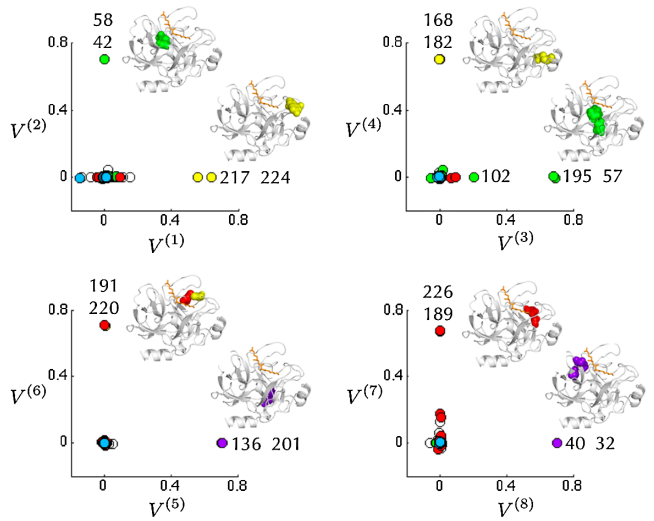

FIG. 3 (color online). Top protein sectons in the trypsin family—each graph is a projection of the positions along ($V^{(k)}$, $V^{(k+1)}$), the components of order k and $k+1$ obtained by rotating by ICA the top eigenvectors of the truncated matrix of direct information $\tilde{\mathcal{D}}_{ij}$. Sectons are defined by $s_k = \{i: V_i^{(k)} > \epsilon\}$, with $\epsilon = 0.2$. The labeling of positions follows the numbering system of bovine chymotrypsin (in several instances, positions appear as superimposed), and the colors reflect the sectors as in Fig. 1, with yellow for nonsector positions. The location of the sectons on the three-dimensional structure is also indicated (more sectons are shown in Fig. S10 [8]). Sectons s_2, s_4, s_5, and s_6 are disulfide bonds, and s_3 is the catalytic triad.

range of values of k_{top} and ϵ (Figs. S11–S12 [8]). As indicated in Fig. 2(b), the first 35 sectons are structurally connected. Out of these 35 sectons, 13 have size 2, 20 size 3, and 2 size 4 (Fig. S13 [8]). Figure 2(c) shows that sectons of size 2 are top pairs of $\tilde{\mathcal{D}}_{ij}$ (for technical reasons, an exception is the first secton; see Fig. S12 [8]), but sectons of size ≥ 3 include contacting pairs that are not top pairs of $\tilde{\mathcal{D}}_{ij}$. Reciprocally, Fig. 2(d) shows that $\sim$60% of the top pairs of $\tilde{\mathcal{D}}_{ij}$ are not in the top sectons. Thus, sectons and top pairs reveal different aspects of the correlations (see also Table SIV [8]). Finally, Fig. 2(b) shows that sectons are also consistent with the decomposition into sectors, with almost no secton intersecting two sectors.

Only few sectons are well-recognized structural or functional units: For the trypsin family, the top six sectons thus include four disulfide bonds [11], and the catalytic triad, a group of three residues mediating peptide bond hydrolysis and shared among several other protein families [12]. Characterizing the structural and/or functional roles of other sectons is an open experimental challenge. Sectons are found in other protein families [13], thus raising the question of whether different families sharing a common fold also share common sectons [14]. Sectons and sectors are in any case distinct from previously recognized structural units such as secondary structures or "foldons" [15], which consist of consecutive positions along the chain.

Formally, sectors and sectons originate from exclusive parts of the spectrum of a common covariance matrix, $\bar{C}_{ij}^{ab} = \bar{f}_{ij}^{ab} - \bar{f}_i^a \bar{f}_j^b$, defined from the regularized frequencies $\bar{f}_i^a = (1-\mu) f_i^a + \mu/(A+1)$ and $\bar{f}_{ij}^{ab} = (1-\mu) f_{ij}^{ab} + \mu/(A+1)^2$, where $A = 20$ is the number of amino acids. A parameter $\mu = 1/2$ is introduced by DCA to define the coupling matrix $J = -\bar{C}^{-1}$ on which $\mathcal{D}_{ij} = \mathcal{D}_{ij}[J]$ relies [6]. This regularization is not required for SCA, but using $\mathcal{C}_{ij} = \mathcal{C}_{ij}[\bar{C}]$ with $\mu = 1/2$ instead of $\mu = 0$, which amounts to adding random sequences to the alignment, does not alter significantly sector identification (Fig. S14 [8]). If $\bar{C} = \sum_k |k\rangle \lambda_k \langle k|$ denotes the spectral decomposition of $\bar{C}$ in the bra-ket notation, with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_L$, we can decompose $\bar{C}$ as $\bar{C} = \bar{C}^+ + \bar{C}^-$, where

$$\bar{C}^{+} = \sum_{k \le k^*} |k\rangle \lambda_k \langle k| \quad \text{and} \quad \bar{C}^{-} = \sum_{k > k^*} |k\rangle \lambda_k \langle k|.$$

With, for instance, $k^* = 100$, the sectors inferred by SCA from $\mathcal{C}_{ij}[\bar{C}^{+}]$ are indiscernible from those from $\mathcal{C}_{ij}[\bar{C}]$ (Fig. S14 [8]). On the other hand, using $\mathcal{D}_{ij}[J^{-}]$ with $J^{-} = -\sum_{k>k^*} |k\rangle \lambda_k^{-1} \langle k|$ instead of $\mathcal{D}_{ij}[J]$ in DCA [16], not only do we recover the same contacts and sectons (Fig. S15 [8]), but an additional ~20% of them are found to be structurally connected (Fig. S16 [8]). The association of sectors and sectons to different parts of the spectrum of $\bar{C}$ relates to random matrix theory, which indicates that both ends of the spectra of undersampled empirical covariance matrices are statistically significant [9]. The spectral decomposition, however, does not account *per se* for the relation between sectors, contacts, and sectons, and it may ultimately not be the most relevant decomposition of the correlations.

The concepts and methods exposed thus far are not limited to protein structures. Another example involving biological sequences is the inference of functional couplings between genes in a genome. A first-order approach to this problem is to study the co-occurrence of genes in a large number genomes, also known as their phylogenetic profile [17]. The raw data are an $M \times L$ binary array x_{si}, where $x_{si} = 1$ indicates that gene i is present in the genome of species s and 0 that it is absent ($A = 1$ in this case). Building such a data set requires mapping corresponding genes across genomes: here we rely on the partition of bacterial genes into clusters of orthologous genes (COGs) [18], to obtain a data set consisting of $M \simeq 10^3$ genomes and $L \simeq 1.5 \times 10^3$ orthologous classes [8].

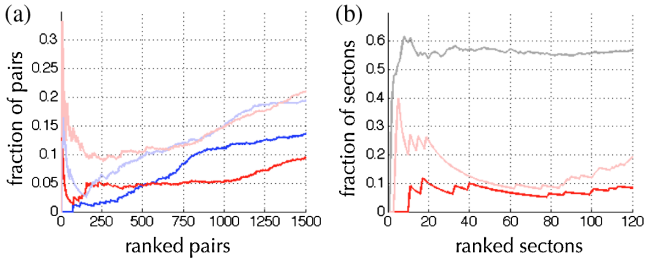

FIG. 4 (color online). Correlated pairs and sectons in bacterial genomes—(a) Fraction of pairs of genes from different functional categories in the top pairs of $\mathcal{D}_{ij}$ (dark red curve, with lowest values for high ranks) or $C_{ij} = f_{ij} - f_i f_j$ (dark blue): Up to rank 1000, a plateau around 5% is observed in the first case, and a continuous increase in the second. The two matrices share some of their top pairs (Fig. S20 [8]), but, for instance, top pairs of $\mathcal{D}_{ij}$ initially contain more poorly characterized genes (light red) than top pairs of C_{ij} (light blue). (b) Fraction of sectons with two or more genes from different functional categories (dark red), and at least one poorly characterized gene (light red). In gray, fraction with genes from different functional categories after randomizing the content of the sectons, showing that finding less than 10% of functionally mixed sectons is significantly lower.

No structural data are available for comparison in this case, but the classification of COGs into three broad, nonexclusive, functional classes [18] (metabolism, cellular processes, and information processing, with a fourth class for poorly characterized genes [19]) indicates that the top pairs of the matrix of direct information $\mathcal{D}_{ij}$ are dominantly composed of genes from a same functional class [Fig. 4(a)]; these results are consistent with those previously derived from a similar approach [20]. As for protein alignments, sectons can be defined that consist here of small clusters, typically of 2–6 genes (Figs. S17–S20 and Table SV [8]). These sectons are mostly composed of functionally related genes [Fig. 4(b)]; many sectons in fact consist of different subunits of a same protein complex (Table SVI [8]). Genomic sectors, involving larger groups of correlated genes, may be defined as well, although their significance is more difficult to assess (Fig. S21 and Table SVII [8]).

In conclusion, we provided evidence that the contacting pairs inferred by DCA [6] and the sectors inferred by SCA [4] are two interrelated features of a common pattern of coevolution, with coevolving units of intermediate size, called sectons, providing additional information. A fully unified mathematical framework for representing the hierarchy of correlations in biomolecules remains to be developed. Characterizing the structural, functional, and evolutionary roles of patterns of coevolution is more generally a problem that extends beyond the scope of statistical studies of sequence data; in particular, experiments are needed to assess the extent to which statistical patterns of coevolution, inferred from a collection of sequences, are reflected in individual biomolecules.

I thank L. Colwell, B. Houchmandzadeh, I. Junier, S. Kuehn, S. Leibler, R. Ranganathan, K. Reynolds, and T. Tesileanu for discussions and comments. This work is supported by ANR grant "CoevolInterProt."

[1] D. S. Horner, W. Pirovano, and G. Pesole, Brief. Bioinform. **9**, 46 (2007).
[2] S. Lockless and R. Ranganathan, Science **286**, 295 (1999).
[3] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).
[4] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[5] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).
[6] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[7] M. Punta *et al.*, Nucleic Acids Res. **40**, D290 (2011).
[8] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.178102 for details.

[9] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).
[10] A. Bell and T. Sejnowski, Neural Comput. **7**, 1129 (1995).
[11] Rat trypsin has a total of six disulfide bonds, but the other two are not conserved in the family, with frequencies $<5\%$.
[12] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry* (Freeman, San Francisco, 2007), Chap. 9, 6th ed.
[13] O. Rivoire and R. Ranganathan (to be published).
[14] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **291**, 177 (1999).
[15] A. R. Panchenko, Z. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **93**, 2008 (1996).
[16] J^- is the inverse of $-C^-$ on its nonzero eigenspace, spanned by $\{|k\rangle\}_{k>k^*}$.
[17] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, Proc. Natl. Acad. Sci. U.S.A. **96**, 4285 (1999).
[18] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, Nucleic Acids Res. **28**, 33 (2000).
[19] Gene annotation may, however, be incomplete and serve here only as a proxy for biological significance. Note also that COGs are themselves only proxies for orthologous classes, and more elaborated definitions of orthology may lead to more informative results.
[20] P.-J. Kim and N. D. Price, PLoS Comput. Biol. **7**, e1002340 (2011).
[21] A. Pasternak, D. Ringe, and L. Hedstrom, Protein Sci. **8**, 253 (1999).