Global view of the protein universe

Sergey Nepomnyachiy[a], Nir Ben-Tal[b,1], and Rachel Kolodny[c,1]

[a]Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201; [b]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel; and [c]Department of Computer Science, University of Haifa, Mount Carmel 31905, Israel

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved July 2, 2014 (received for review February 24, 2014)

To explore protein space from a global perspective, we consider 9,710 SCOP (Structural Classification of Proteins) domains with up to 70% sequence identity and present all similarities among them as networks: In the "domain network," nodes represent domains, and edges connect domains that share "motifs," i.e., significantly sized segments of similar sequence and structure. We explore the dependence of the network on the thresholds that define the evolutionary relatedness of the domains. At excessively strict thresholds the network falls apart completely; for very lax thresholds, there are network paths between virtually all domains. Interestingly, at intermediate thresholds the network constitutes two regions that can be described as "continuous" versus "discrete." The continuous region comprises a large connected component, dominated by domains with alternating alpha and beta elements, and the discrete region includes the rest of the domains in isolated islands, each generally corresponding to a fold. We also construct the "motif network," in which nodes represent recurring motifs, and edges connect motifs that appear in the same domain. This network also features a large and highly connected component of motifs that originate from domains with alternating alpha/beta elements (and some all-alpha domains), and smaller isolated islands. Indeed, the motif network suggests that nature reuses such motifs extensively. The networks suggest evolutionary paths between domains and give hints about protein evolution and the underlying biophysics. They provide natural means of organizing protein space, and could be useful for the development of strategies for protein search and design.

protein cooccurrence networks | protein similarity networks

How are proteins related to each other? Which physicochemical considerations affect protein evolution and how? A global view of the protein universe may shed light on these fundamental questions. It could also suggest new strategies for protein search and design (1–3). However, forming a global picture of the protein universe is difficult because we have to piece it together from the many local glimpses that our empirical data and computational tools provide. In other words, a global picture needs to portray the relationships among all proteins, yet we only have evidence of such relationships among several proteins, based on the similarity between their sequences, structures, and functions. The considerable size of the Protein Data Bank (4) also complicates this task.

In particular, an intensely debated question is whether protein space is "discrete" or "continuous" (2, 3, 5–10). These terms are loosely defined. Discrete implies that the global picture consists of separate, island-like, structural entities. In the hierarchical protein domains Structural Classification of Proteins (SCOP) (11) these entities are termed "folds," and in the CATH database (12) they are called "topologies." Alternatively, "continuous" implies that the space between these entities is generally populated by cross-fold similarities (e.g., refs. 2, 5, 6, 9, 13–15). If such similarities are abundant, then one must account for them when organizing and searching proteins (5, 8, 16). In support of the abundance of such similarities is the remarkable success of structure prediction methods that piece together predictions of protein fragments or larger protein segments (e.g., ref. 17).

There are different approaches to forming a global view of the protein universe (18). The most significant efforts are the ones embodied in the hierarchical classifications CATH and SCOP. However, a hierarchy implicitly assumes that there are isolated regions in protein space. An alternative approach is to study the protein universe via maps—where domains are represented by points in two or three dimensions, placed so that the distances between them depend on the dissimilarity between their corresponding domains (e.g., refs. 19–21). By coloring the points according to domain characteristics, one can visually identify global properties of the protein universe (19, 20). However, a map representation in low-dimensional Euclidean space implicitly suggests that similarity among domains is transitive (i.e., that similarity within the pairs AB and BC implies that AC is similar too); we know that this is often not the case (6). Finally, a third approach to study protein space is via similarity and cooccurrence networks. In similarity networks, nodes typically represent protein domains and edges connect similar domains. Several successful studies of protein space capitalize on such networks (22, 23). Cooccurrence networks of protein domains, in which nodes represent domains and edges connect cooccurring domains, were also studied to better understand protein evolution (24–26).

Here, we study the global nature of the protein universe using domain and motif networks (Fig. 1). To construct these networks, we identify evolutionary relationships among a representative set of SCOP domains; we relate two domains if they share a significantly sized part (denoted motif) with similar structure and sequence. Our analysis reveals that protein space is both discrete and continuous: SCOP domains of the all-alpha, all-beta, and alpha + beta classes, in which alpha and beta elements do not mix, mostly populate the discrete parts, whereas alpha/beta

Significance

To globally explore protein space, we use networks to present similarities among a representative set of all known domains. In the "domain network" edges connect domains that share "motifs," i.e., significantly sized segments of similar sequence and structure, and in the "motif network" edges connect recurring motifs that appear in the same domain. The networks offer a way to organize protein space, and examine how the organization changes upon changing the definition of "evolutionary relatedness" among domains. For example, we use them to highlight and characterize the uniqueness of a class of domains called alpha/beta, in which the alpha and beta elements alternate. The networks can also suggest evolutionary paths between domains, and be used for protein search and design.

Author contributions: N.B.-T. and R.K. designed research; S.N. and R.K. performed research; N.B.-T. and R.K. analyzed data; and N.B.-T. and R.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: bental@ashtoret.tau.ac.il or trachel@cs.haifa.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1403395111/-/DCSupplemental.

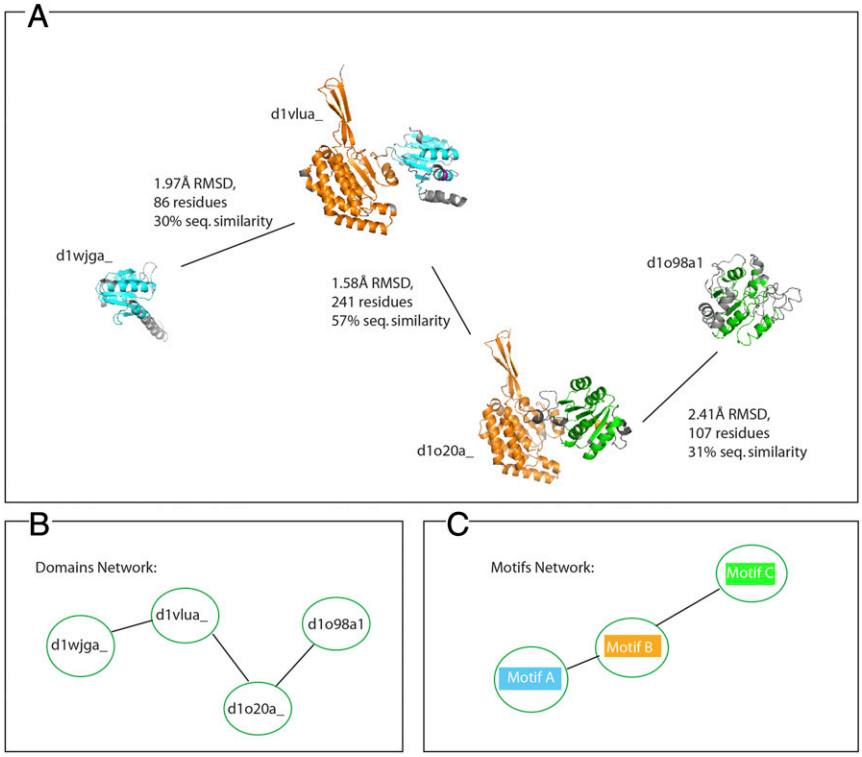

Fig. 1. Constructing the domain and motif networks. (*A*) The aligned protein segments, marked in colors, are the motifs. (*B*) In the domain network, edges connect domains that share similar motifs (e.g., domain d1wjga_ and d1vlua_ that share the cyan motif). (*C*) In the motif network, edges connect cooccurring motifs (e.g., the orange and cyan motifs cooccur in the d1vlua_ domain).

domains, with alternating alpha and beta segments, mostly populate the continuous ones. We also find that recurring motifs are very abundant; the motifs from the all-alpha and alpha/beta domains are the more abundant, and the more gregarious ones.

Results

We align all-versus-all in a set of 70% sequence nonredundant SCOP v.1.72 domains (11) using the structural aligner SSM (27). For each pair of aligned domains, we calculate the length of the aligned region, the percent sequence similarity of aligned residues (using the BLOSUM62 substitution matrix), and the root-mean-square deviation (rmsd) of these residues. Then, we define cutoffs for these values and use them to filter the alignments. From the filtered alignments, we construct the domain network (Fig. 1*B*) and the motif network (Fig. 1*C*). In the domain network, nodes are the SCOP domains in the dataset, and edges connect pairs of domains that share a similar motif. In the motif network, nodes are motifs, and the edges connect pairs of motifs that cooccur in a domain. We consider length thresholds of 55 and 75 residues, percent similarity of aligned residues thresholds of 30%, 40%, and 50%, and rmsd thresholds of 2, 2.5, and 3 Å. We explore how well threshold combinations reproduce SCOP segregation into folds, i.e., optimally including all domains from the same fold in a connected component, whereas excluding from it domains of other folds.

Protein Space Includes Continuous and Discrete Regions. The connectivity of the domain network varies depending on the thresholds used to define the evolutionary relationships (Fig. 2 and *SI Appendix*, Figs. S1–S4). If we consider the relatively lax thresholds of 50 residues, 30% sequence similarity, and 3-Å rmsd, then the resulting domain network is virtually a single connected component (including 9,385 or 97% of the domains). For more stringent thresholds, which we consider to represent evolutionary relationships more faithfully, the network reveals both continuous and discrete regions of protein space (Fig. 2 and *SI Appendix*, Figs. S2 and S3). At even more stringent length and similarity thresholds the network falls apart completely (e.g., *SI Appendix*, Fig. S4). *SI Appendix*, Fig. S5 shows the stacked histograms of sizes of the connected components, of representative networks. Indeed, using longer length, higher percent similarity, or lower rmsd thresholds results in a more disconnected network, and places more domains in smaller components. Importantly, in all these cases, we see a single exceptionally large connected component.

SI Appendix, Fig. S6 shows the percent of domain pairs with the same SCOP fold that are in the same connected component in a domain network (x axis), versus the percent of pairs that have a different SCOP fold and that are not connected (y axis). We consider all pairs among the all-alpha, all-beta, alpha/beta, and alpha + beta domains (*SI Appendix*, Fig. S6*A*), and all pairs among the 61% domains that are not alpha/beta (*SI Appendix*, Fig. S6*B*). Notice that when considering the region of protein space that does not include the alpha/beta domains (*SI Appendix*, Fig. S6*B*), the domain network captures the notion of fold far better and fairly well overall. As expected, lax thresholds generate a network with larger connected components, and consequently the percent of domain pairs with the same fold that are connected is greater (higher values along the x axis), but also, there are more domain pairs of different folds that are (inappropriately) connected (lower values along the y axis). The thresholds that generate domain networks that overall best agree with SCOP fold assignments are either (*i*) alignments longer than 75 residues, with percent similarity greater than 30%, and rmsd smaller than 2.5Å, or (*ii*) alignments longer than 55 residues, percent similarity greater than 30%, and rmsd smaller than 2 Å.

SI Appendix, Fig. S7 shows the same analysis per SCOP class. We see that in the all-beta class, and to a lesser extent in the alpha + beta class, our optimal thresholds can generally identify SCOP folds and place domains of the same fold in the same connected component, while still being disconnected from the domains that are not in that fold (high values along the x and y axes). In the alpha/beta class, and to a lesser extent in the all-alpha class, if we want to successfully connect domains that are in the same fold (i.e., achieve high values along the x axis), we inevitably connect to domains that are not in the same fold

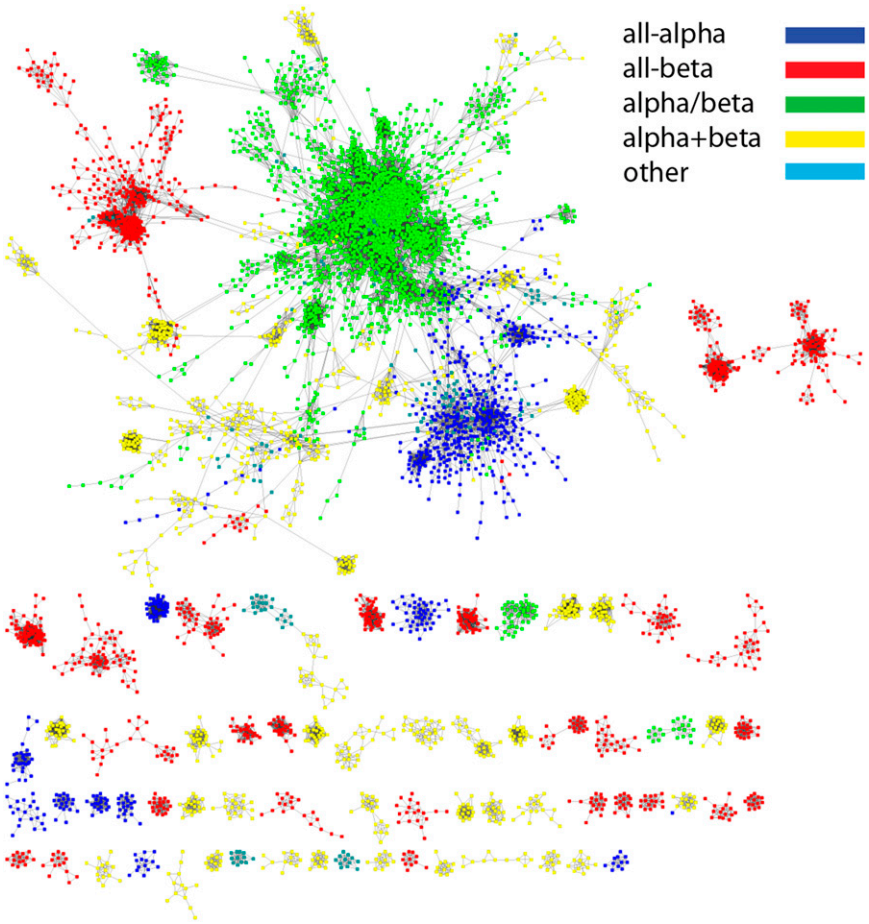

Fig. 2. Global view of protein space via the domain network. The nodes represent the set of 70% sequence nonredundant SCOP domains, colored by their SCOP class (see color legend); edges connect between domains that share a motif. Here, two domains are connected if we found a similarity of at least 75 residues, with at least 25% sequence similarity, and at most 2.5 Å rmsd. We see that there are two regions: one is very connected, or continuous, and populated mostly by (green) alpha/beta domains in which the alpha and beta elements alternate; the other is discrete, composed of many disconnected components, and populated by the all-alpha, all-beta, and alpha + beta domains. Only components with more than 10 domains are shown.

(low values along the *y* axis). We fail to find threshold combinations that are successful along both axes.

The Continuous Region of the Protein Universe Contains the Alpha/Beta Domains; the All-Alpha, All-Beta, and Alpha + Beta Domains Are in the More Discrete Region. Fig. 2 shows the domain network that best reproduces SCOP's classification into folds. It is balanced in that it connects a significant amount of domains to each other even though it was obtained using conservative thresholds to represent evolutionary relationships. As the networks obtained using all other (reasonable) thresholds, the network features both discrete and continuous regions. For the most part, SCOP domains of the all-alpha, all-beta, and alpha + beta classes, in which alpha and beta elements do not mix, populate the discrete parts (Fig. 2); roughly speaking, in this region the connected components correspond to SCOP folds (*SI Appendix*, Fig. S8). In contrast, the alpha/beta domains, with alternating alpha and beta elements, populate the large connected component. This continuous region includes domains from most alpha/beta folds, including the TIM barrels, NAD(P) binding Rossmanns, FAD/NAD(P) binding, and many more (*SI Appendix*, Fig. S8); the domains of each fold are often found in very close vicinity to each other in the main connected component. It is known that individual folds (e.g., TIM barrels) have undergone circular mutations and splicing within their respective folds (28). Our analysis indicates that the evolutionary relationships extend beyond the individual fold, covering, in essence, the entire alpha/beta SCOP class.

If we consider the similarity of sequence and structure as indicative of evolutionary relationships among proteins, Fig. 2 can be interpreted as collections of evolutionary paths in protein space. For example, Fig. 3 shows a path passing between domains from the FAD/NAD(P) binding, TIM barrel, Rossmann NAD(P) binding, nucleotide binding, and Flavodoxin folds. In this network, there is a path between 77% of the alpha/beta domain pairs, whereas paths between a pair of domains from the all-alpha, all-beta, and alpha + beta classes are found only in 10%, 6%, and 8% of the cases, respectively. This is not an artifact of the different number of domains in the four SCOP classes: we see similar numbers when we randomly sample 1,000 domains from each SCOP class. The large amount of paths within the alpha/beta SCOP class suggests that it is particularly easy to add and delete motifs among them without impeding structural stability. It could be that evolution took advantage of this property to design new proteins with novel functions.

Fig. 4 shows the network of 8,219 recurring motifs, obtained using the same parameters used for the domain network of Fig. 2. Of these, 994 are nonsingletons. The number of different domains which are present in a motif is, by definition, greater than 2, and almost always less than 50 (983, or 99%, of the motifs that are nonsingletons). In 82% of these cases (810 motifs) all of the domains in the motif have the same SCOP class; only in 31% of these cases (311 motifs) do they have the same SCOP fold.

Recurring motifs are very common. We see that the (green) alpha/beta are the most abundant: the percent (number) of nonsingleton motifs that are all-alpha, all-beta, alpha + beta, and alpha/beta is 22%, 4%, 4%, and 55%, (223, 43, 44, and 547), respectively; 28% (279) of the nonsingleton motifs have an equal part of two classes. The weak connection between motifs taken from domains of the alpha/beta and all-alpha classes is mediated by superimpositions of small domains on excessively large domains. Had these larger domains been divided into smaller

domains, the vast majority of motifs from the all-alpha domains would disintegrate from the main connected component.

Discussion

The Domain Network Reveals the Continuous–Discrete Nature of Protein Space. The question if protein space is continuous or discrete has been extensively debated (2, 3, 5–10), and is interesting both fundamentally and for its implications on how to organize and search protein databases (5, 16). The domain network allows us to describe "continuous" and "discrete" more concretely based on the sizes and number of connected components. We find that protein space has both discrete and continuous regions, in agreement with Sadreyev et al. (7), and that the distinction largely depends on the domains' SCOP class: continuity is most prevalent among the alpha/beta domains whereas the region of the all-alpha, all-beta, and alpha + beta domains is mostly discrete. Skolnick et al. attributed the continuity to physical properties of proteins and to backbone hydrogen bonds in particular (15). That alpha/beta domains are more interconnected than other SCOP classes suggests that the domains in this class share unique physicochemical qualities that are yet to be discovered.

Edges in the domain network are determined using specific thresholds. More lax thresholds imply more edges and hence a more connected network; at the extreme case all protein space is a single connected component. Stricter thresholds imply fewer edges and hence a less connected network. Also, using a more sensitive method to identify similarity among domains will reveal a more connected network. Indeed, the method and the thresholds for inferring the relationships among domain pairs should fit the question at hand. We consider "local" relationships that represent domains closer and further apart in evolution and combine them into a "global" view of protein space to study its properties.

To connect domain pairs that are likely evolutionarily related, we verified that the domains share similar structure and sequence over a significant number of residues. Skolnick et al. (15) showed that when relating domain pairs based solely on the similarity of their structures (and a minimal TM_Score threshold of 0.4), protein space is essentially a single connected component. Our work deals with what happens when we "raise the metaphorical bar" for relating two domains, and enforce that the domain pairs are likely evolutionarily related (using a range of thresholds). Indeed, even in this stricter setting, if the thresholds are sufficiently lax (namely, at least 50 residues with more than 25% sequence identity and rmsd less than 3 Å) virtually all of protein space is connected, suggesting that protein space is evolutionarily (not only structurally) connected. However, if we consider stricter thresholds, and specifically ones which were calibrated to best capture the connectivity of SCOP folds, then protein space disintegrates, and this disintegration is generally in the region of non–alpha/beta domains.

One could argue that all of fold space is discrete; only each SCOP class requires different thresholds to disintegrate. Our data show that this is not the case. To learn this, we focused on each of the four SCOP classes, and searched for optimal thresholds resulting in networks that capture SCOP fold connectivity. Recall that a successful network simultaneously keeps same-fold domains connected, and disconnects them from domains in different folds. The success stems directly from the properties of the class of domains: If a class has a more discrete nature, that is, if its intrafold similarities are greater than its interfold similarities, then we can find appropriate thresholds. If, on the other hand, it has a more continuous nature, then by using increasingly strict thresholds to relate domain pairs, the domain network will disintegrate, but it will do so altogether, and lose the property that same-fold domains remain in the same connected component. Indeed, we see that the SCOP classes vary in how well the best thresholds capture their fold connectivity: the all-beta domains have the most discrete nature, followed by the alpha + beta domains, the all-alpha domains, and finally the alpha/beta domains that have the most continuous nature (*SI Appendix*, Fig. S4).

We construct the dataset of likely evolutionary relationships using two steps: (*i*) searching for candidate domain pairs, and then (*ii*) verifying that their corresponding subparts satisfy predefined length, sequence similarity, and structure similarity criteria. For the first step, we used the structural aligner SSM (27). However, structural aligners vary in the relationships that they identify: some are more sensitive than others (29, 30). Here, we chose SSM because it was shown to be particularly sensitive (30). The search procedure can be augmented using additional structural

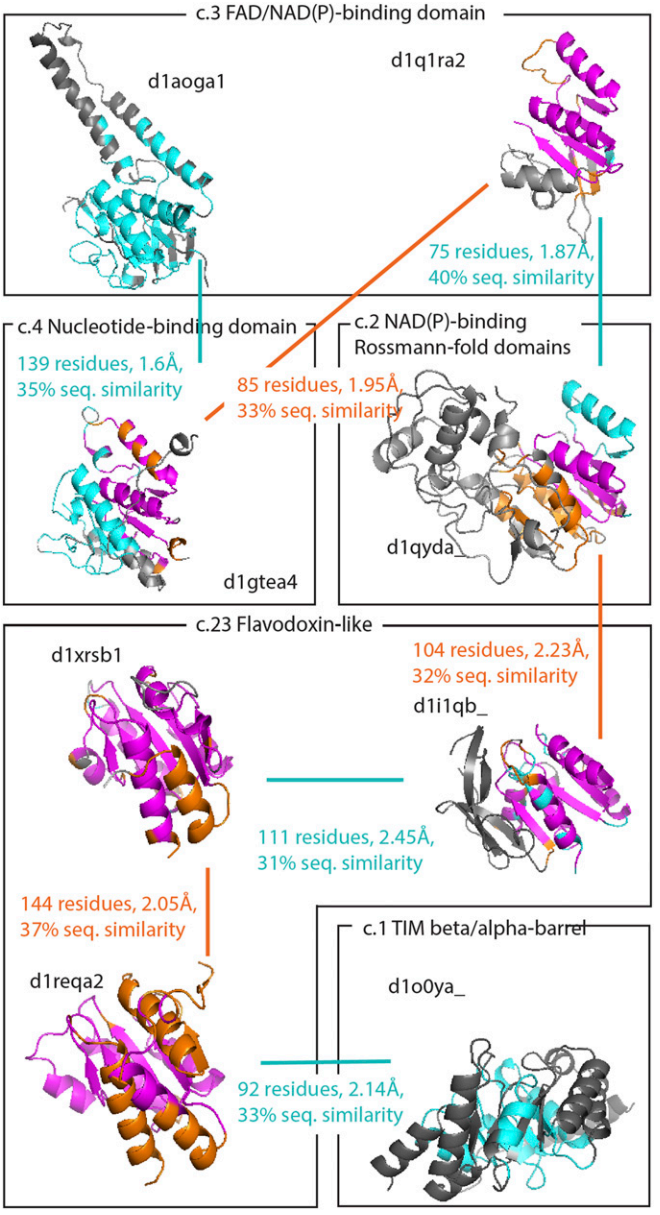

Fig. 3. "Walking" in the domain network. A putative evolutionary path, to demonstrate the relationships between connected domains. The path, taken from the major connected component, passes through eight domains from five different SCOP folds of the alpha/beta class. The aligned motifs are marked in orange or cyan; residues shared by the motifs in both directions along the path are in magenta. The number of residues, rmsd, and percent sequence similarity (using BLOSUM62) of the aligned motifs are indicated.

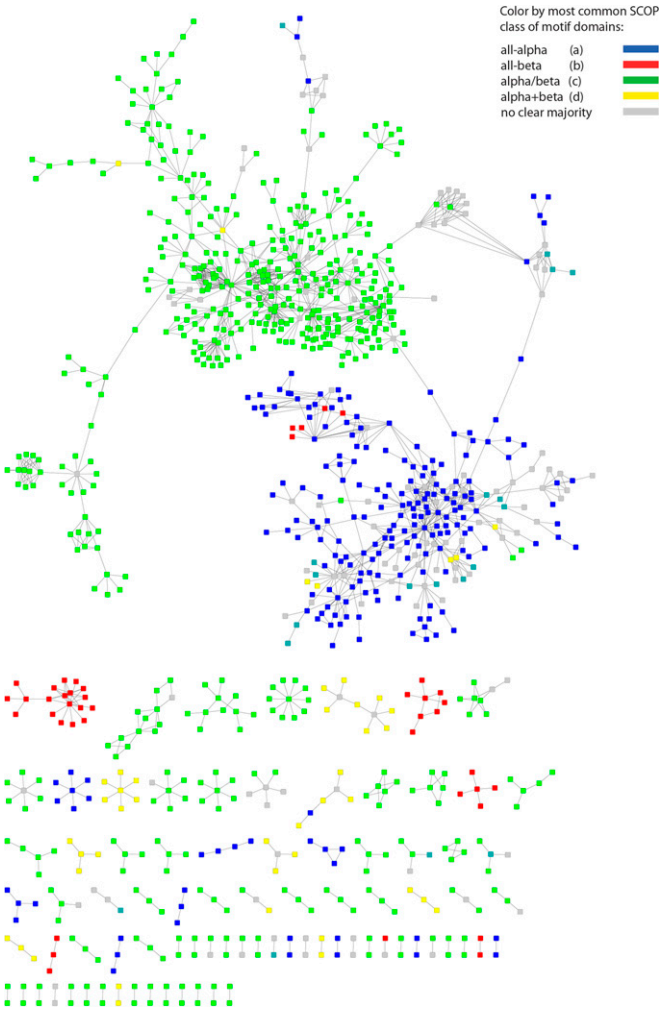

Fig. 4. Global view of protein space via the motif network. The nodes represent the set of 8,219 identified motifs, colored by the SCOP class of the majority of their domains (see color legend; white represents cases where no SCOP class is the majority); edges connect between motifs that cooccur in a domain. The motif network was constructed using the set of alignments that are longer than 75 residues, with more than 25% sequence similarity, and less than 2.5 Å rmsd (*Methods*). We see that the alpha/beta (and the all-alpha) motifs are more common, more gregarious, and form the largest connected component.

aligners [e.g., Matt (29), STRUCTAL (31), or TM_align (15)]. Hopefully, these can identify additional candidate evolutionary relationships, which we can subsequently subject to the similarity filters in step (*ii*).

The Motif Network Reveals the Ubiquitous Reuse of Motifs in Nature. Previous studies of cooccurrence networks use domains as the unit element (24, 26, 32, 33). In those networks, nodes represent domains, and edges connect between cooccurring domains. Our motif network is similar, only we represent motifs that are smaller than domains. The distributions of the number of neighbors in the domain cooccurrence and our motif network are similar (24). Also, the alpha/beta motifs and domains tend to have more partners (or a higher rank) in their respective networks (24).

Importantly, we derive the unit element (or nodes) in the cooccurrence network from the data rather than relying on predefined (e.g., SCOP–CATH) domains. Domains were used because they are considered the basic unit of protein evolution (34). It is assumed that there is only a limited set of them, and domains from this set are combined to form the set of proteins in the proteome using genetic mechanisms (24). For example, genetic recombination can cause loss or duplication of parts of genes, entire genes, or even longer chromosomal regions; mobile genetic elements (DNA transposons and retro-transposons) can lead to duplication or deletions (26). It may be, however, as suggested by Lupas et al., that the basic unit is actually smaller than a domain (35). Our tools offer a way to further investigate this idea and demonstrate the abundance of mix-and-join events. In this respect it is noteworthy that whereas domains are considered to be autonomous structural units, which are stable on their own, it may well be that the motifs are not, and that despite their ability to "hop between domains" they are stable only within the context of the intact domain. Note that we have used the same thresholds in the motif and domain networks. These thresholds are not necessarily the best ones to highlight all significant similarity at the subdomain level. Future in-depth study is required to better understand the properties of motifs and their networks with more lax thresholds. Regardless of the actual evolutionary scenario underlying the motif network, the network lends itself naturally to protein engineering efforts by suggesting which substructures can replace one another while maintaining protein foldability. Just like evolution has recycled such motifs so could protein engineers, enriching the topologies of engineered proteins and their likelihood of performing new functions.

Alpha/Beta Domains Are Unique. Previous work showed that the alpha/beta domains are older (19), more stable (36), more frequently involved in domain fusion events (32), and are associated with high functional diversity (20). Our analysis shows two additional unique features of these domains: they lie in a tightly connected region of protein space and their motifs mix-and-join with a wider range of motifs. The tendency of the alpha/beta domains to easily mix-and-join could explain their functional diversity.

Two alternative explanations for these properties of the alpha/beta domains and motifs are (*i*) they existed in ancient evolutionary history, and were mixed from these entities (35, 37), or (*ii*) their biophysical properties give them a selection advantage. Our observations do not help in determining which of the two explanations is more likely, and this remains a significant challenge.

We provide tools for navigation in the domain and motif networks by integrating Cytoscape (38) and PyMOL (39). To visualize our networks, download the Cytoscape files describing them at http://cs.haifa.ac.il/~trachel/domain_motif_networks/. The networks could be used to theorize about protein evolution, suggest evolutionary pathways between domains, and hence maybe suggest strategies for protein design.

Methods

Dataset. Our dataset consists of 9,710 domains that are 70% sequence nonredundant from the SCOP database. We filtered away domains whose structures were not accurately determined [a Summary PDB ASTRAL Check Index score (40) lower than 0.2]. We aligned all-versus-all domains using the structural alignment method SSM (27). We parsed the alignments, measured their length (i.e., number of aligned residues), and calculated the percent of identical residues, and the percent of similar residues (using the BLOSUM62 matrix). From these data, we constructed and visualized the domain and the motif networks using Cytoscape (38).

The Domain Network. The nodes in the domain network represent the domains in the dataset (Fig. 1*A*); a single edge connects two nodes if we found a significant alignment of sufficiently many residues, sufficiently low rmsd, and sufficiently high percent sequence similarity (Fig. 1*B*). We considered different thresholds of alignment length (55 and 75 residues), rmsd (2, 2.5, and 3 Å), and percent sequence similarity (30%, 40%, and 50%).

The Motif Network. The motif graphs offer an alternative representation of the same alignment data (Fig. 1*C*). The first step is to identify the nodes of the motif graph. An alignment, A, matches a set of residues in protein *P*1 with a set of residues in protein *P*2; here, we denote these subsections $P1_A$

and $P2_A$. As evidenced by the alignment itself, $P1_A$ and $P2_A$ are two names of a similar subsection. This subsection can have additional names: consider another alignment, B, which matches subsections $P1_B$ and $P3_B$. If the residues in subsection $P1_B$ are actually the same ones as those in $P1_A$, then $P1_B$ and $P3_B$ are also names of this subsection. Thus, we need to identify the different names (in the example given here: $P1_A$,$P2_A$,$P1_B$,$P3_B$) that describe similar subsections. To do this, we constructed an auxiliary graph, in which the nodes are the raw subsections extracted directly from the set of significant alignments (two per alignment); in the example described the nodes in the auxiliary graph will include the nodes $P1_A$,$P2_A$,$P1_B$,$P3_B$. In the auxiliary graph we connect pairs of subsections associated with each alignment (one edge per alignment); in the example these will be the edges between $P1_A$ and $P2_A$, and between $P1_B$ and $P3_B$. In the auxiliary graph we also connect (almost) similar subsections of the same domain; in the example given above this is an edge between $P1_A$ and $P1_B$. For this, we used a threshold of 90% overlap (e.g., we connected the motifs that represent residues 1–100 and residues 2–101 of the same domain). Each connected component in the auxiliary graph is a node in the motif network. In other words, each node in the motif graph is a set of recurring subsections.

To generate a clearer motif network, we added a few more steps. First, even when using the 90% overlap threshold, we may suffer from a "dragging" effect, where we start with one subsection, and then via a series of intermediate subsections that are 90% similar to each other, we reach another subsection of vastly different size. To circumvent this problem, we greedily split motifs in which the ratio between the longest and shortest subsection is greater than 1.5. Also, we remove motifs that we identify as supermotifs of other motifs in the dataset: if motif1 includes subsection P_A and motif2 includes subsection P_B, and all residues in subsection P_B are also subsection P_A, then we consider motif1 a supermotif of motif2, and remove it. The edges in the motif network connect motif pairs for which there are subsections of that domain in both motifs.

Data Visualization. We added an interface to viewing structural information using PyMOL (39). In the domain network we visualize the domains that correspond to the nodes, as well as the domain superimpositions that correspond to the edges; the aligned residues are highlighted. In the motif network an edge is a domain that includes both motifs at its end nodes: we show the two motifs in cyan and in orange, with the overlapping residues in magenta; if there is more than one possible domain, the user needs to choose the one to visualize. For the nodes in the motif network, we visualize two domains with these motifs superimposed on one another.

ACKNOWLEDGMENTS. We thank Yonatan Bilu, Sarel Fleishman, and Dan Tawfik for insightful discussions, Varda Wexler for graphics consulting, and the anonymous reviewers for helpful comments. N.B.-T. acknowledges the financial support of Grant 1775/12 of the Israeli Centers of Research Excellence Program of the Planning and Budgeting Committee and the Israel Science Foundation.

1. Levitt M, Chothia C (1976) Structural patterns in globular proteins. *Nature* 261(5561): 552–558.
2. Kolodny R, Pereyaslavets L, Samson AO, Levitt M (2013) On the universe of protein folds. *Annu Rev Biophys* 42:559–582.
3. Taylor WR (2007) Evolutionary transitions in protein fold space. *Curr Opin Struct Biol* 17(3):354–361.
4. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
5. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction. *Curr Opin Struct Biol* 16(3):393–398.
6. Pascual-García A, Abia D, Ortiz ÁR, Bastolla U (2009) Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLOS Comput Biol* 5(3):e1000331.
7. Sadreyev RI, Kim B-H, Grishin NV (2009) Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol* 19(3):321–328.
8. Sadowski MI, Taylor WR (2010) On the evolutionary origins of "Fold Space Continuity": A study of topological convergence and divergence in mixed alpha-beta domains. *J Struct Biol* 172(3):244–252.
9. Harrison A, Pearl F, Mott R, Thornton J, Orengo C (2002) Quantifying the similarities within fold space. *J Mol Biol* 323(5):909–926.
10. Valas RE, Yang S, Bourne PE (2009) Nothing about protein structure classification makes sense except in the light of evolution. *Curr Opin Struct Biol* 19(3):329–334.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536–540.
12. Orengo CA, et al. (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108.
13. Andreeva A, Prlić A, Hubbard TJP, Murzin AG (2007) SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* 35(Database issue, suppl 1): D253–D259.
14. Shindyalov IN, Bourne PE (2000) An alternative view of protein fold space. *Proteins* 38(3):247–260.
15. Skolnick J, Arakaki AK, Lee SY, Brylinski M (2009) The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci USA* 106(37):15690–15695.
16. Petrey D, Honig B (2009) Is protein classification necessary? Toward alternative approaches to function annotation. *Curr Opin Struct Biol* 19(3):363–368.
17. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294(5540):93–96.
18. Ben-Tal N, Kolodny R (2014) Representation of the protein universe using classifications, maps, and networks. *Isr J Chem*, in press.
19. Choi IG, Kim SH (2006) Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci USA* 103(38):14056–14061.
20. Osadchy M, Kolodny R (2011) Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc Natl Acad Sci USA* 108(30): 12301–12306.
21. Holm L, Sander C (1996) Mapping the protein universe. *Science* 273(5275):595–603.
22. Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002) Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci USA* 99(22): 14132–14136.
23. Alva V, Remmert M, Biegert A, Lupas AN, Söding J (2010) A galaxy of folds. *Protein Sci* 19(1):124–130.
24. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310(2):311–325.
25. Wuchty S (2001) Scale-free behavior in protein domain networks. *Mol Biol Evol* 18(9): 1694–1702.
26. Forslund K, Sonnhammer EL (2012) *Evolution of Protein Domain Architectures. Evolutionary Genomics* (Springer, Berlin), pp 187–216.
27. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2256–2268.
28. Söding J, Lupas AN (2003) More than the sum of their parts: On the evolution of proteins from peptides. *BioEssays* 25(9):837–846.
29. Daniels NM, Kumar A, Cowen LJ, Menke M (2012) Touring protein space with Matt. *IEEE/ACM Trans Comput Biol Bioinformatics* 9(1):286–293.
30. Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. *J Mol Biol* 346(4):1173–1188.
31. Subbiah S, Laurents DV, Levitt M (1993) Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr Biol* 3(3):141–148.
32. Hua S, Guo T, Gough J, Sun Z (2002) Proteins with class α/β fold have high-level participation in fusion events. *J Mol Biol* 320(4):713–719.
33. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Res* 18(3):449–461.
34. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300(5626):1701–1703.
35. Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* 134(2-3):191–203.
36. Minary P, Levitt M (2008) Probing protein fold space with a simplified model. *J Mol Biol* 375(4):920–933.
37. Ponting CP, Russell RR (2002) The natural history of protein domains. *Annu Rev Biophys Biomol Struct* 31(1):45–71.
38. Saito R, et al. (2012) A travel guide to Cytoscape plugins. *Nat Methods* 9(11): 1069–1076.
39. Schrodinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1. Available at www.pymol.org.
40. Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 28(1):254–256.