Review

Israel Journal of Chemistry

DOI: 10.1002/ijch.201400001

# Representation of the Protein Universe using Classifications, Maps, and Networks

Nir Ben-Tal*[a] and Rachel Kolodny*[b]

**Abstract**: A meaningful and coherent global picture of the protein universe is needed to better understand protein evolution and the underlying biophysics. We survey the studies that tackled this fundamental challenge, providing a glimpse of the protein space. A global picture represents all known local relationships among proteins, and needs to do so in a comprehensive and accurate manner. Three types of global representations can be used: classifications, maps, and networks. In these, the local relationships are derived, based on the similarity of the proteins' sequences, structures, or functions (or a combination of these). Alternatively, the local relationships can be co-occurrences of elements in the protein universe. The representations can be based on different objects: full polypeptide chains, fragments, such as structural domains, or even smaller motifs. Different protein qualities were revealed in each study; many point out the uniqueness of domains of the alpha/beta SCOP (structural classification of proteins) class.

**Keywords:** classification of proteins · protein evolution · protein maps · protein structures · proteins

## 1. Introduction

A global view of the protein universe is an abstraction that allows one to formulate, quantify, and hold up to scrutiny general observations regarding protein evolution.[1–7] A major difficulty in forming a global picture of the protein universe is that it must be pieced together from many local observations that experimental techniques and computational tools can provide. In other words, a global picture needs to portray relationships among all proteins, yet we only have evidence of such relationships among several proteins. The considerable size of the protein databases also complicates this task. Thus, building a meaningful and coherent global picture of protein space remains a significant challenge on the path to understanding protein evolution and the biophysical principles that underlie it.

Focusing on different local similarities among proteins results in different global perspectives of protein space. In particular, the local relationships can be derived, based on the proteins' amino acid sequence, three-dimensional structure, function (i.e., phenotype), or linkage to diseases.[2] The sequence databases hold the largest number of proteins: TrEMBL currently holds almost 50 million sequences and SWISS-PROT holds approximately 500,000 sequences.[8] The Protein Data Bank (PDB), in which structures are stored, holds over 100,000 structures.[9]

The similarities among pairs of proteins are typically based on the similarity of their parts, i.e., of two subsections that have similar sequence and/or structure. Scientists consider such similarities significant because we assume that insertions and deletions are common evolutionary events. However, one must be aware of the pitfalls when relying on such similarities to form a global picture. Most importantly, such similarities are not transitive.[10–12] That is, it may be that there is significant similarity (or evidence of evolutionary relationship) between proteins A and B, and between proteins B and C, yet there is no detectable similarity between proteins A and C. This can happen, for example, if the similarity lies in non-overlapping subsections of protein B (See Figure 3 in Ref. [2]). Cases of missing transitivity are widespread when considering protein chains, as these are composed of structural domains that are reused by nature.[13] Focusing on structural domains as the basic units can alleviate the problem, but does not solve it altogether, as there are many cases of missing transitivity, even when comparing domains.[14] In addition, it is not trivial to predict domain boundaries based on a protein's sequence, raising techni-

[a] N. Ben-Tal
Department of Biochemistry and Molecular Biology
George S. Wise Faculty of Life Sciences
Tel Aviv University
Ramat Aviv 69978 (Israel)
e-mail: bental@ashtoret.tau.ac.il
Homepage: http://ashtoret.tau.ac.il

[b] R. Kolodny
Department of Computer Science
University of Haifa
Mount Carmel 31905 (Israel)
e-mail: trachel@cs.haifa.ac.il
Homepage: http://cs.haifa.ac.il/~trachel

cal problems for broad applications of domain-based classifications to proteins of unknown structure;[15] even when the structure is known, identifying domains, and in particular, the exact boundaries, is challenging.[16]

Three alternative representations of protein space are classifications, maps, and networks. Classifications organize objects into categories, often hierarchically. This is, by far, the predominant way to study the global nature of protein space, and it rests on a long tradition of studies that classified objects in biological archives, dating back to Darwin and earlier. Maps visualize the abstract protein space as a collection of points in low-dimensional Euclidean space (namely, in two or three dimensions). In maps, the distance between the points approximates the distance between the objects they represent, and color encodes the objects' properties. Networks represent protein space as a set of nodes connected by edges. Scholars study the properties of these networks, including visualizing the networks in two dimensions, to better understand the global nature of protein space. To construct a classification, a map, or a network, one must decide on the unit objects (i.e., the object to classify, to represent by points or by nodes), and on what relates two (or more) objects. Indeed, different decisions reveal different global perspectives of protein space.

The objects described by the classifications, maps, and networks can be short protein segments, protein motifs, structural domains, full chains, or multi-chain complexes. The relationships between the objects can be based on the similarities of their sequence, structure, function, or a combination of these. Alternatively, two objects can be considered related if they co-occur in nature, or if there is evidence of an evolutionary relationship between them. A global description can include all objects which can be related, yet only those contained in the various databases. For example, if the relationship depends on sequence similarity, then all chains in the sequence databases SWISS-PROT and TrEMBL can be represented. However, if the relationship depends on structure similarity, we are restricted to the chains in the PDB. Similarly, if the relationship depends on functional similarity, we are restricted even further to the proteins with annotated functions.

## 2. Classifications

The first and most significant efforts to understand the global nature of the protein universe were based on classifications. There are many classifications of protein space; for reviews, see Refs. [17–20]. Here, we focus on classifications that aim at providing a global view of the protein universe. Notice that protein databases that organize known proteins can also be interpreted as classifications; these databases are beyond the scope of this review (e.g., Refs. [21–27]). Classifications cluster all proteins, typically based on comparisons of protein pairs.

Several classifications rely only on the similarity of the protein sequences. As such, these classifications can describe all sequences in the large and comprehensive sequence databases SWISS-PROT and TrEMBL. Because these databases are so large, it is necessary to use algorithms that automatically cluster the proteins into meaningful groups. Different classifications use different algorithms. When considering protein sequences, the unit object is typically a protein chain, partly because it is difficult to identify domain boundaries. Consequently, the clustering of sequences has to accommodate (the many) cases of missing transitivity. Some examples are provided herein.

CluSTr uses a single-linkage hierarchical clustering, and relies on pairwise Smith-Waterman comparisons.[28,29] Protonet is a bottom-up hierarchical clustering that relies on pairwise BLAST (basic local alignment search tool) comparisons;[30–32] Kaplan *et al.* showed that Protonet clusters capture functional and structural aspects of the protein world.[33] TRIBE-MCL is also based on BLAST comparisons. TRIBE-MCL uses the stochastic matrix that encodes a similarity graph for the set of proteins and analyze the graph to detect clusters by manipulating this matrix (using expansion and inflation operators).[34] SYSTERS (systemic re-searching) groups the sequences hierarchically at the family and the superfamily level.[35] Another sequence-based classification is the clusters of orthologous groups (COG) database, which is a phylogenetic classification of proteins encoded in complete genomes.[36,37] The COGs are groups of proteins that are tightly related (and thought to be orthologous), these are clustered into superfamilies using PSI-BLAST. COGs are classified into 17 broad functional categories, and some of the COGs with known functions are also organized so as to represent specific cellular systems and biochemical pathways.

There are also classifications that consider the protein structures. Chothia and Levitt first observed that proteins structures can be classified into four classes,[1] and Richardson constructed the first general classification scheme.[38] Today, the most widely-used classifications are SCOP (structural classification of proteins)[39] and CATH (class, architecture, topology, homologous superfamily):[40] both are hierarchical, and group proteins based initially on the similarity of their sequences, and then based on the similarity of their structures. Additional structure-based classifications are the Dali Domain Dictionary (DDD),[41] ECOD (evolutionary classification of protein domains),[42a] the new SCOP2[42b], and COPS (classification of protein structures).[43] Because the structures of the classified proteins are known, they can be reliably partitioned to domains. Indeed, SCOP, CATH, and ECOD classify structural domains, rather than full polypeptide chains. These classifications offer different and complementary views of the protein universe. This is true, both in terms of the set of classified domains (e.g., SCOP and CATH domains are not the same), and the clustering itself, which is different (e.g., when considering a subset of domains that are similar between SCOP and CATH).[44,45]

Finally, there are classifications that consider the protein functions. Gene ontology (GO) offers a common language, formulated as a directed graph, to describe the biological process of a protein, its molecular function, and its cellular component.[46] Hence, the set of all GO annotations for PDB proteins form a functional classification. These annotations are of protein chains, but there are also GO classifications for structural domains, which were derived from the chain annotations, including references.[47,48] Similarly, the EC (Enzyme Commission) numbering offers a functional classification of the region in the protein universe that hosts the enzymes.[49]

## 3. Maps

A map depicts protein space as a set of points in two- or three-dimensional space. Each point represents a protein, and the distance between points approximates the distance between their corresponding proteins. The coordinates for the $N$ points, which represent $N$ proteins, are calculated from the $N \times N$ proteins' distance matrix, so that the distances between the points best preserve the distances in the matrix. The quality of the embedding in the low-dimensional map is measured via stress functions, which compare the distances in Euclidean space to those in the input matrix; there are different stress functions, with some being easier to optimize than others.[50] A common solution to calculating the low-dimensional coordinates is called multidimensional scaling (MDS);[51] it is optimal in terms of a specific stress function (denoted strain) and requires calculating the eigenvalue decomposition of the $N \times N$ input matrix. Alternatively, one can heuristically optimize the stress function (e.g., using gradient descent),[10,50,52,53] and thus compute maps even for large $N$s, i.e., for datasets of significant size.

Most maps of protein space represent proteins with known structure. The maps by Orengo *et al.*,[54] Holm and Sander,[55] and Kim and colleagues[56–58] represent structural distances among protein domains. Rogen and Fain represented CATH domains by vectors of 30 numbers inspired by Vassiliev knot invariants, and projected this representation to two dimensions.[59] The maps of Yona and Levitt represent distances that are based on both the sequence and structure similarity,[10] while those of Grishin and Grishin[60] represent evolutionary distances among domains. Farnum *et al.*[52] and Stanberry *et al.*[61] heuristically calculated maps that represent the distance between protein sequences, and in the latter case, the map describes very large datasets.

Osadchy and Kolodny suggested using the FragBag representation of protein structures[62] to calculate maps more efficiently.[53] In FragBag, the structure of a protein domain is represented as a point in an $L$-dimensional space (with $L \leq 400$); Budowski *et al.* showed that the similarity between the FragBag vectors, or the points in the $L$-dimensional space, can identify near structural neighbors as accurately as the state-of-the-art structural aligners, STRUCTAL and CE (combinatorial extension).[62] Using FragBag's fixed-size vector descriptors of protein domains, Osadchy and Kolodny replaced MDS with the more efficient computational procedure, principal component analysis (PCA).[63] PCA generates the same map (up to a reflection and rotation of the entire space) as the one generated by MDS, if the distances in the MDS matrix were the Euclidean distances between the vectors in the PCA matrix. The difference is in efficiency: PCA calculates the eigenvalue decomposition of an $L \times L$ matrix, where $L$ is the length of the vector describing a protein (i.e., $L \leq 400$), whereas MDS calculates the eigenvalues decomposition of an $N \times N$ matrix, where $N$ is the size of the dataset. Using this technology, Osadchy and Kolodny calculated a map for a larger dataset with $N = 31{,}155$ SCOP domains.

Maps provide a comprehensive view of protein space, which is unconstrained by the implied "boundaries" between objects that are classified differently (e.g., with a different SCOP fold, or with a different CATH, classification). In maps of protein space, one can visually identify patterns, formulate these observations to hypotheses, and test them quantitatively. When formulating such hypotheses, one must verify that the low-dimensional projection of the data did not introduce artifacts (see, for example, Ref. [53]). It is also noticeable that the representation of objects as points in low-dimensional Euclidean space implicitly assumes that transitivity holds, which is often not the case (see above). Thus, domains, rather than full chains, are more appropriate as the objects represented in maps.

Fundamental insights regarding the nature of protein space emerged from studying low-dimensional maps. Importantly, the domains from the four major SCOP classes, namely, from the all-alpha, all-beta, alpha/beta, and alpha + beta classes, are generally located in different regions of space.[10, 53, 54, 57, 58] Choi and Kim used maps to study the evolution of protein folds and concluded that not all present-day proteins evolved from a single set of proteins in the last common ancestor, and new common ancestral proteins were "born" at different evolutionary times.[56] Osadchy and Kolodny showed that the density of protein structure space is uneven, i.e., certain regions have more domains per "unit volume" than others. More significantly, they showed that functional diversity also varies considerably across structure space; structure space has a region of high functional diversity, and diversity abates when moving away from it. The domains in this high-diversity region are mostly alpha/beta structures, which are also known to be the most ancient proteins.[56,64] As expected, the high functional diversity region includes domains of the TIM-barrel fold, famous for its functional diversity, but also many other alpha/beta folds. The evolutionary and biophysical reasons for this remain to be revealed.

## 4. Networks

Networks represent protein space as a set of nodes connected by edges. The network properties can be studied directly, e.g., by analyzing the distribution of the number of neighbors in the network nodes, or the distribution of the sizes of the connected components in the network. Alternatively, networks can be visualized,[65,66] alongside with properties of the proteins, to gain novel insights, e.g., by coloring the nodes based on the SCOP class/fold of the domains they represent. Networks are widely-used to study proteins: for example, to study protein interactions (e.g., Refs. [67,68]), or phylogeny.[64] We focus here on two specific types of networks: "similarity networks" and "co-occurrence networks".

In similarity networks, edges connect nodes that represent similar objects; in co-occurrence networks, edges connect nodes that co-occur in nature. The nodes represent the unit objects, which, in both networks, can be chains, domains, or motifs. The similarity can be derived from the sequences, structures, or functions of these unit objects. These two networks are dual to each other and thus offer a complementary view of protein space.[69,70] In dual networks, the roles of nodes and edges are exchanged: the nodes of the primary network are represented by edges in its dual network, and the edges in the primary network are represented by nodes in its dual. If we consider, for example, similarity networks, in which the nodes represent chains (that may include more than a single domain), then the edges represent their recurring subparts, namely the domains. Thus, co-occurrence networks in which the nodes represent domains are their dual: these nodes represent the recurring subparts of the chains, namely the domains, and the edges represent the chains that include more than a single domain.

### 4.1 Similarity Networks

Studies investigating similarity networks rely on sequence or structure similarity. Two significant studies that relate proteins, based on sequence similarity, are Protomap by Yona *et al.*,[12] and "A Galaxy of Folds" by Alva *et al.*[71] In Protomap, nodes are SWISS-PROT proteins, and edges connect proteins that the local aligners Smith-Waterman, FASTA, and BLAST identified as similar. In "A Galaxy of Folds", Alva *et al.* visualized networks that represent a set of 20% sequence, non-redundant, SCOP domains: the nodes were the domains, colored by their SCOP class, fold, or superfamily, and edges connected domains identified as similar by HHSearch, a sensitive sequence aligner. Others studied networks that were based on structure similarity. Skolnick *et al.* studied a network, in which edges connect domains that have sufficiently similar substructures, as quantified using TMScore.[7] Dokholyan *et al.*[72,73] designed the protein domain universe graph (PDUG), in which the nodes represent a set of 25% sequence, non-redundant, FSSP domains,[74] and edges connect domains that the structure aligner DALI identified as similar. They studied the connected components in the PDUG, which generally correspond to SCOP folds. Sun *et al.*[75] created a PDUG-like network with a sparser set of domains, so that nodes represent SCOP folds, rather than families. There are also networks in which the similarity is identified by both sequence and structure aligners, including Valavanis *et al.*,[76] Camoglu *et al.*,[77] and Yona and Levitt.[10] Finally, Fragnostics by Friedberg and Godzik[78] is a similarity network, in which the nodes represent SCOP folds, and edges connect nodes that represent folds with similar fragment composition; the FragBag study suggests that this is similar to connecting folds with similar global structures.[62]

Some of the similarity networks were constructed to gain insights regarding the evolution and biophysics of the protein universe, or as an intermediate step when classifying proteins automatically (e.g., Protomap[12]). Skolnick *et al.* deduced from their network that structure space is highly connected, and that the distance between any two domains is only a few edges. The PDUG is a scale-free network and has 'small-world' characteristics, unlike random graphs with similar distribution of the number of edges per node; Dokholyan *et al.* theorized that this suggests that all proteins originated from a single fold, or a few precursor folds – a scenario akin to that of the origin of the universe from the Big Bang.[72,75,76] In the PDUG, there is also a correlation between domain structure and function (as described by functional fingerprints), which may suggest that divergent evolution is more dominant than convergent evolution.[73] From the global view of the sequence-based similarity network, Alva *et al.* showed incidences of homologous connections that transcend both superfamily and fold levels.[71] Friedberg *et al.* demonstrated with the Fragnostics network that functional similarity (as measured using GO annotations) is correlated with structural similarity.[78]

### 4.2 Co-occurrence Networks

Studying co-occurrence networks can lead to a better understanding of processes that create new proteins, i.e., duplication, recombination, fusion, and fission of their respective genes.[79,80] Wuchty[81] studied co-occurrence networks of the ProDom, Pfam, and Prosite domains, with edges between domains that co-occur in at least one protein. He showed that the resulting network does not have random graph characteristics; rather, it is scale-free. He noted in this context that the network generation model of Barabási and Albert,[82] which preferentially attaches newly-added vertices to already well-connected ones, also generates scale-free networks. Koonin *et al.* studied co-occurrence of the domain sequences in genomes to gain insights into protein evolution.[83] Apic *et al.* studied the co-occurrence of SCOP domains in different genomes and connected two domains if they lie sufficiently close to each other in a genome.[84,85] In particular, they studied tandem domains in which the two domains are from the same SCOP family, as these may have evolved via a mechanism of internal duplication; they showed that tandem domains are a relatively rare event. They also showed that the co-occurrence networks are scale-free. Finally, Kummerfeld and Teichman studied a directed version of the co-occurrence network, in which they took into account the order of the domains along the protein chain.[86]

## 5. Conclusions

Understanding the global nature of the protein universe, and forming abstractions that will represent all protein data coherently and meaningfully, is a fundamental challenge. It is interesting theoretically, as it can help to better characterize protein biophysics and evolution. It also has practical implications, as it may lead to better designing protein-related tools, e.g., to organize and search protein databases.[5,87] The rapid growth of the protein databases renders the challenge of representing protein space more technically complicated, yet holds the promise of developing a comprehensive view of the protein universe.

We surveyed three alternative global representations of protein space: classifications, maps, and networks. Classifications are the most commonly used representations, and are the most informative when studying a specific protein or protein family. However, classifications are less amenable to visualizing the whole of protein space. Maps and networks, on the other hand, are easy to visualize, and thus offer a complementary way to study the protein universe. Maps are more intuitive, in that the distance between points is the same as our intuitive notion of distance; this is also their weakness as it suggests that similarity among proteins is transitive, which is often not the case. Similarity networks are less intuitive, but do not suffer from this weakness, as they describe similarity explicitly. Indeed, studying maps and network representations has revealed novel properties of the protein universe.

It is desirable to combine the three representations under the same umbrella in a way that enables going back and forth between them. It requires designing interactive and up-to-date visualization tools of abstract representations of the protein universe for studying protein, both globally and locally. The desired computational tool should be based on the available structures for accuracy, but should also cover proteins for which only the sequence is known. It should also include automated means to detect domains in a reliable way. Studying the global nature of protein space will hopefully allow us to gain insights, to raise new hypotheses, and to better understand the relationships between protein sequences, structures, and functions at the global level. At the same time, such tools can be adapted to study specific proteins and protein families, as the local environment in which a protein lies (i.e., its neighboring proteins), and the location of a protein within the protein universe, can help in the investigation of the protein at hand.

## Acknowledgements

We congratulate the 2013 Nobel laureates in Chemistry. RK acknowledges the fantastic mentorship of M. Levitt. NBT acknowledges the financial support of grant No. 1775/12 of the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation.

## References

[1] M. Levitt, C. Chothia, *Nature* **1976**, *261*, 552–558.
[2] R. Kolodny, L. Pereyaslavets, A. O. Samson, M. Levitt, *Annu. Rev. Biophys.* **2013**, *42*, 559–582.
[3] R. A. Goldstein, *Curr. Opin. Struct. Biol.* **2008**, *18*, 170–177.
[4] W. R. Taylor, *Curr. Opin. Struct. Biol.* **2007**, *17*, 354–361.
[5] R. Kolodny, D. Petrey, B. Honig, *Curr. Opin. Struct. Biol.* **2006**, *16*, 393–398.
[6] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, W. S. Noble, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6559–6563.
[7] J. Skolnick, A. K. Arakaki, S. Y. Lee, M. Brylinski *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 15690–15695.
[8] The UniProt Consortium, *Nucleic Acids Res.* **2013**, *41*, D43–D47.
[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
[10] G. Yona, M. Levitt, in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology,* (Eds.: R. Altman, T. L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I. N. Shindyalov, L. F. Ten Eyck, H. Weissig), **2000**, pp. 395–406.
[11] M. J. Sippl, *Bioinformatics* **2008**, *24*, 872–883.
[12] G. Yona, N. Linial, M. Linial, *Proteins: Struct. Funct. Genet.* **1999**, *37*, 360–378.
[13] M. Levitt, *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11079–11084.
[14] A. Pascual-García, D. Abia, Á. R. Ortiz, U. Bastolla, *PLoS Comput. Biol.* **2009**, *5*, e1000331.
[15] C. H. Tai, W. J. Lee, J. J. Vincent, B. Lee, *Proteins Struct. Funct. Bioinf.* **2005**, *61*, 183–192.
[16] T. A. Holland, S. Veretnik, I. N. Shindyalov, P. E. Bourne, *J. Mol. Biol.* **2006**, *361*, 562–590.
[17] C. A. Ouzounis, R. M. Coulson, A. J. Enright, V. Kunin, J. B. Pereira-Leal, *Nat. Rev. Genet.* **2003**, *4*, 508–19.

[18] P. Koehl, in *Reviews in Computational Chemistry, Vol. 22*, (Eds.: K. B. Lipkowitz, T. R. Cundari, V. J. Gillet, D. B. Boyd), John Wiley & Sons, Inc., Hoboken, New Jersey, **2006**, pp. 1–55.

[19] J. Liu, B. Rost, *Curr. Opin. Chem. Biol.* **2003**, *7*, 5–11.

[20] A. Andreeva, in *Homology Modeling, Methods in Molecular Biology, Vol. 857*, (Eds.: A. J. W. Orry, R. Abagyan), Springer, New York, **2012**, pp. 1–31.

[21] T. K. Attwood, M. E. Beck, D. R. Flower, P. Scordis, J. Selley, *Nucleic Acids Res.* **1998**, *26*, 304–308.

[22] P. Vanhee, E. Verschueren, L. Baeten, F. Stricher, L. Serrano, F. Rousseau, J. Schymkowitz, *Nucleic Acids Res.* **2011**, *39*, D435–D442.

[23] D. H. Haft, J. D. Selengut, O. White, *Nucleic Acids Res.* **2003**, *31*, 371–373.

[24] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, C. Yeats, *Nucleic Acids Res.* **2009**, *37*, D211–D215.

[25] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, P. Bork, *Nucleic Acids Res.* **2004**, *32*, D142–D144.

[26] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, *Nucleic Acids Res.* **2004**, *32*, D138–D141.

[27] W. Li, A. Godzik, *Bioinformatics* **2006**, *22*, 1658–1659.

[28] E. V. Kriventseva, F. Servant, R. Apweiler, *Nucleic Acids Res.* **2003**, *31*, 388–389.

[29] R. Petryszak, E. Kretschmann, D. Wieser, R. Apweiler, *Bioinformatics* **2005**, *21*, 3604–3609.

[30] O. Sasson, A. Vaaknin, H. Fleischer, E. Portugaly, Y. Bilu, N. Linial, M. Linial, *Nucleic Acids Res.* **2003**, *31*, 348–352.

[31] N. Kaplan, O. Sasson, U. Inbar, M. Friedlich, M. Fromer, H. Fleischer, E. Portugaly, N. Linial, M. Linial, *Nucleic Acids Res.* **2005**, *33*, D216–D218.

[32] N. Rappoport, S. Karsenty, A. Stern, N. Linial, M. Linial, *Nucleic Acids Res.* **2012**, *40*, D313–D320.

[33] N. Kaplan, M. Friedlich, M. Fromer, M. Linial, *BMC Bioinf.* **2004**, *5*, 196.

[34] A. J. Enright, S. Van Dongen, C. A. Ouzounis, *Nucleic Acids Res.* **2002**, *30*, 1575–1584.

[35] A. Krause, J. Stoye, M. Vingron, *BMC Bioinf.* **2005**, *6*, 15.

[36] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, E. V. Koonin, *Nucleic Acids Res.* **2001**, *29*, 22–28.

[37] R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **1997**, *278*, 631–637.

[38] J. S. Richardson, *Adv. Protein Chem.* **1981**, *34*, 167–339.

[39] T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, C. Chothia, *Nucleic Acids Res.* **1999**, *27*, 254–256.

[40] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, J. Thornton, *Structure* **1997**, *5*, 1093–1108.

[41] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, L. Holm, *Nucleic Acids Res.* **2001**, *29*, 55–57.

[42] a) Evolutionary Classification of Protein Domain Structures website, last modified May 31, 2014, http://prodata.swmed.edu/ecod/; b) A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A. G. Murzin, *Nucleic Acids Res.* **2014**, *42*, D310–D314.

[43] S. J. Suhrer, M. Wiederstein, M. Gruber, M. J. Sippl, *Nucleic Acids Res.* **2009**, *37*, 539–544.

[44] P. Koehl, *Curr. Opin. Struct. Biol.* **2001**, *11*, 348–353.

[45] R. Day, D. A. C. Beck, R. S. Armen, V. Daggett, *Protein Sci.* **2003**, *12*, 2150–2160.

[46] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, *Nucleic Acids Res.* **2004**, *32*, D258–D261.

[47] D. Lopez, F. Pazos, *Proteins: Struct. Funct. Bioinf.* **2009**, *76*, 598–607.

[48] H. Fang, J. Gough, *Nucleic Acids Res.* **2013**, *41*, D536–D544.

[49] A. Bairoch, *Nucleic Acids Res.* **2000**, *28*, 304–305.

[50] R. Duda, P. Hart, D. Stork, *Pattern Classification, 2nd Edition*, John Wiley & Sons, Inc., New York, **2001**.

[51] T. F. Cox, M. A. A. Cox, *Multidimensional Scaling, 2nd Edition*, CRC Press LLC, Boca Raton, **2000**.

[52] M. A. Farnum, H. Xu, D. K. Agrafiotis, *Protein Sci.* **2003**, *12*, 1604–1612.

[53] M. Osadchy, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12301–12306.

[54] C. A. Orengo, T. P. Flores, W. R. Taylor, J. M. Thornton, *Protein Eng.* **1993**, *6*, 485–500.

[55] L. Holm, C. Sander, *Science* **1996**, *273*, 595–603.

[56] I. G. Choi, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 14056–14061.

[57] J. Hou, G. E. Sims, C. Zhang, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2386–2390.

[58] J. Hou, S. R. Jun, C. Zhang, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3651–3656.

[59] P. Rogen, B. Fain, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 119–24.

[60] V. N. Grishin, N. V. Grishin, *Bioinformatics* **2002**, *18*, 1523–1534.

[61] L. Stanberry, R. Higdon, W. Haynes, N. Kolker, W. Broomall, S. Ekanayake, A. Hughes, Y. Ruan, J. Qiu, E. Kolker, G. Fox, in *Proceedings of the 3rd International Workshop on Emerging Computational Methods for the Life Sciences*, ACM, New York, **2012**, pp. 13–22.

[62] I. Budowski-Tal, Y. Nov, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 3481–3486.

[63] V. deSilva, J. B. Tenenbaum, *Sparse Multidimensional Scaling Using Landmark Points*, Stanford University, Stanford, **2004**.

[64] H. F. Winstanley, S. Abeln, C. M. Deane, *Bioinformatics* **2005**, *21*, 449–458.

[65] R. Saito, M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, *et al.*, *Nat. Methods* **2012**, *9*, 1069–1076.

[66] A. T. Adai, S. V. Date, S. Wieland, E. M. Marcotte, *J. Mol. Biol.* **2004**, *340*, 179–190.

[67] R. Singh, J. Xu, B. Berger, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 12763–12768.

[68] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, C. von Mering, *Nucleic Acids Res.* **2011**, *39*, D561–D568.

[69] T. Przytycka, G. Davis, N. Song, D. Durand, *J. Comput. Biol.* **2006**, *13*, 351–363.

[70] S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.*, doi:10.1073/pnas.1403395111.

[71] V. Alva, M. Remmert, A. Biegert, A. N. Lupas, J. Söding, *Protein Sci.* **2010**, *19*, 124–130.

[72] N. V. Dokholyan, B. Shakhnovich, E. I. Shakhnovich, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14132–14136.

[73] B. E. Shakhnovich, N. V. Dokholyan, C. DeLisi, EI. Shakhnovich, *J. Mol. Biol.* **2003**, *326*, 1–9.

[74] L. Holm, C. Sander, *Nucleic Acids Res.* **1996**, *24*, 206–209.

[75] Z.-B. Sun, X.-W. Zou, W. Guan, Z.-Z. Jin, *EPJ B* **2006**, *49*, 127–134.

[76] I. Valavanis, G. Spyrou, K. Nikita, *J. Biomed. Inf.* **2010**, *43*, 257–267.

[77] O. Çamoğlu, T. Can, A. K. Singh, *Bioinformatics* **2006**, *22*, 1585–1592.

[78] I. Friedberg, A. Godzik, *Structure* **2005**, *13*, 1213–1224.

[79] K. Forslund, E. L. Sonnhammer, in *Evolutionary Genomics, Methods in Molecular Biology, Vol. 856*, (Ed: M. Anisimova), Springer, New York, **2012**, pp. 187–216.

[80] E. Bornberg-Bauer, F. Beaussart, S. Kummerfeld, S. Teichmann, J. Weiner III, *Cell. Mol. Life Sci.* **2005**, *62*, 435–445.

[81] S. Wuchty, *Mol. Biol. Evol.* **2001**, *18*, 1694–1702.

[82] A.-L. Barabási, R. Albert, *Science* **1999**, *286*, 509–512.

[83] E. V. Koonin, Y. I. Wolf, G. P. Karev, *Nature* **2002**, *420*, 218–223.

[84] G. Apic, J. Gough, S. A. Teichmann, *J. Mol. Biol.* **2001**, *310*, 311–325.

[85] G. Apic, J. Gough, S. A. Teichmann, *Bioinformatics* **2001**, *17*, S83–S89.

[86] S. Kummerfeld, S. Teichmann, *BMC Bioinf.* **2009**, *10*, 39.

[87] D. Petrey, B. Honig, *Curr. Opin. Struct. Biol.* **2009**, *19*, 363–368.

Received: January 2, 2014
Accepted: April 27, 2014
Published online: ■■ ■■, 0000