Sequence-related human proteins cluster by degree of evolutionary conservation

Ralf Mrowka* and Andreas Patzak
Systems Biology Group, Department of Physiology, Charité Universitätsmedizin Berlin, and Gemeinsame Einrichtung von Freier Univeristät Berlin und Humboldt-Universität zu Berlin, Tucholskystrasse 2, 10117 Berlin, Germany

Hanspeter Herzel
Fachinstitut für Theoretische Biologie, Humboldt-Universität zu Berlin, 10115 Berlin, Germany

Dirk Holste†
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 15 June 2004; published 17 November 2004)

Gene duplication followed by adaptive evolution is thought to be a central mechanism for the emergence of novel genes. To illuminate the contribution of duplicated protein-coding sequences to the complexity of the human genome, we study the connectivity of pairwise sequence-related human proteins and construct a network ($\mathcal{N}$) of linked protein sequences with shared similarities. We find that (i) the connectivity distribution $P(k)$ for k sequence-related proteins decays as a power law $P(k) \sim k^{-\gamma}$ with $\gamma \approx 1.2$, (ii) the top rank of $\mathcal{N}$ consists of a single large cluster of proteins ($\approx 70\%$), while bottom ranks consist of multiple isolated clusters, and (iii) structural characteristics of $\mathcal{N}$ show both a high degree of clustering and an intermediate connectivity ("small-world" features). We gain further insight into structural properties of $\mathcal{N}$ by studying the relationship between the connectivity distribution and the phylogenetic conservation of proteins in bacteria, plants, invertebrates, and vertebrates. We find that (iv) the proportion of sequence-related proteins increases with increasing extent of evolutionary conservation. Our results support that small-world network properties constitute a footprint of an evolutionary mechanism and extend the traditional interpretation of protein families.

DOI: 10.1103/PhysRevE.70.051908 PACS number(s): 87.15.Cc, 87.23.Kg, 89.75.Hc

I. INTRODUCTION

The analysis of statistical patterns in protein sequences is of interest, since correlations in and relationship between protein sequences may reflect biologically significant features of primary structures. For instance, the primary structure of proteins, which is constrained by encoding secondary and higher-order structural information, carries a high information content (low redundancy in the order of 1%) [1], and amino acid correlations of protein sequences affect predominantly base-base DNA sequence periodicities at distances below about 35 bases, while longer-ranging correlations up to 100 bases found in yeast DNA sequences reflect primarily DNA folding properties [2]. The secondary structure of proteins has been linked to a 10-11 bp correlation in DNA sequences [3,4], with several other lines of evidence suggesting that this correlation is associated with DNA bendability and nucleosome formation [2,5,6].

The availability of the complete and qualitatively improved draft sequence of the human genome in conjunction with sequences derived from other species also permits the illumination of historical patterns of genome evolution. An interesting feature of the euchromatic portion of all human chromosomes consists in its repeat content ($\sim 50\%$) and repetitive complexity [7]. Notably large paralogous regions (intra- or interchromosomal segmental duplications) of genomic DNA involve the translocation of 1–200 kb blocks to one or more locations within the genome [7], and a number of such large-scale segmental duplications have been identified within the genome of *Homo sapiens* [8,9], as well as in the genomes of other species [10]. In *H. sapiens*, it has been estimated that 5–7 % of all human DNA sequences may have duplicated in the last 30 Myr of evolution [11]. The fact that most duplications occur in blocks of size >10 kb distinguishes the human genome from other sequenced genomes, including the genomes of the fly *Drosophila melanogaster*, the nematode *Caenorhabditis elegans*, or the plant *Arabidopsis thaliana* [10].

Segmental duplications often harbor protein-coding genes, as well as repeats with high copy numbers [12,13]. The presene and distribution of such segments may foster domain and coding sequence (exon) shuffling [14], and thus contribute to the protein diversity. Across species, the human genome exhibits comparatively greater numbers of gene/protein families and multidomain proteins, as well as paralogous genes, indicating that the greater complexity of its proteome is a consequence not simply of its larger size, but also of large-scale protein innovation [7]. Specifically, in the human genome, between one- and two-tenths of known protein sequences are possibly related to each other via paralogous relationships [7,16]. For instance, the genome-wide evolution of the complete repertoire of human olfactory receptor proteins has shown that intrachromosomal duplications and gene cluster expansion may have led to the creation of one of the largest gene superfamilies in vertebrates that comprises about 1% of the total human genome [15]. Provided that gene duplication, followed by adaptive evolution, constituted

*Electronic address: ralf.mrowka@charite.de
†Electronic address: holste@mit.edu

a major source for the emergence of new genes [12,17], one would expect a considerable degree of similarity and relationships between human proteins.

Here, we study the sequence similarity and diversity of human proteins at the amino acid sequence level. The paper is organized as follows. In Sec. II, we describe the primary data, and introduce the notation and quantities used in this study; in Sec. III and IV, we examine structural network properties of clusters of sequence-related human proteins; and in Sec. V, we investigate the extent to which characteristic cluster features are related to their ancestral states.

II. DATA AND DEFINITIONS

The primary structure of proteins is polymeric and can be considered as a symbolic sequence of $\lambda=20$ symbols (amino acids). Protein sequences for $N_{\text{total}}=21\,787$ known (i.e., ENSEMBL-annotated) human protein-coding genes were obtained from the ENSEMBL genome database (see Ref. [51]). Protein sequence information for the invertebrate *C. elegans*, the plant *A. thaliana*, and the bacterium *M. genitalium* were obtained from the National Center for Biotechnology Information (NCBI) database (see Ref. [52]). Amino acid sequences for the yeast *Saccharomyces cerevisiae* were obtained from the Munich Information Center for Protein Sequences (MIPS) database (see Ref. [53]). The taxonomic classification follows that of the NCBI (see Ref. [54]).

Protein sequence similarities were detected in an all-by-all comparison using the computer program BLASTP [18], versions of which have been used in a variety of related studies, and a number of advanced algorithms have been implemented for downstream analysis (e.g., hierarchically protein cluster) [19–22]. BLASTP matches between a pair of protein sequences were evaluated by the expectation value (e) and assumed to be statistically significant at $e<e_{\text{cr}}=0.001$ (using the BLOSUM62 amino acid substitution matrix). In order to avoid matches of "low-complexity" sequences (sequences of unusual amino acid composition, e.g., repetitive sequences), queries were filtered prior to the analysis [18]. Proteins with significant sequence similarities are linked and joined into clusters.

Three standard statistical quantities of cluster properties are provided by the average cluster coefficient C, the average shortest path length L, and the connectivity distribution $P(k)$. Definitions of C, L, and $P(k)$ can be found, e.g., in [23,24].

First, consider a network $\mathcal{N}(N,E)$ comprised of N nodes and E edges between nodes. Existing links or edges between a pair of nodes E_n and E_m are indicated as $\xi_{nm}\equiv 1$ and zero otherwise. Then the subset of nearest neighbors of node E_n is given by $\{E_n\}\equiv\{\xi_{nm}=1 \,|\, \forall\, m\}$. One can calculate the number of direct connections between the k neighbors $\{E_n\}$ of node E_n as

$$\Xi_n=\sum_{m=1}^{N}\xi_{nm}\left[\sum_{m<l;l\in\{E_n\}}\xi_{ml}\right] \tag{1}$$

and the clustering coefficient for E_n is defined by $C(n)=2\Xi_n/k(k-1)$, where the quantity $k(k-1)/2$ is the maximum number of possible different edges between the neighbors of node E_n. The overall clustering coefficient C is then obtained by averaging over all nodes E_n,

$$C\equiv\langle C(n)\rangle_n=\frac{1}{N}\sum_{n=1}^{N}C(n), \tag{2}$$

and measures the average fraction of pairs of neighbors of a node that are themselves neighbors of each other.

Next, consider two nodes (E_n, E_m; $n\neq m$) and choose $\ell_n^{(m)}$ as the shortest path lengths between E_n and E_m. Then the node-average shortest path length associated with E_n is

$$L(n)=\frac{1}{(N-1)}\sum_{m=1}^{N}\ell_n^{(m)} \tag{3}$$

and the mean over all nodes E_n ($n=1,2,\ldots,N$) defines the average shortest path lengths $L=\langle L(n)\rangle_n$.

Finally, define the connectivity distribution $P(k)$, the overall normalized frequency of all nodes E_n having exactly k edges, as

$$P(k)=\frac{1}{\Omega}\sum_{n=1}^{N}\delta\left(k-\sum_{m=1}^{N}\xi_{nm}\right), \tag{4}$$

where $\delta(0)=1$ and 0 otherwise, and Ω is a normalization factor such that $\Sigma_k P(k)=1$.

In numerical simulations of random networks with an average number of links per node $\langle k\rangle$, the quantities L and C, as well as $P(k)$, can be approximately derived by $L_{\text{rand}}\approx\log N/\log\langle k\rangle$ and $C_{\text{rand}}\approx\langle k\rangle/N$, and for large N the connectivity can be approximated by a Poisson distribution $P(k)_{\text{rand}}=e^{-\langle k\rangle}(\langle k\rangle^k/k!)$ [24]. For a fixed value of $\langle k\rangle$, small worlds show an average shortest path length $L\approx L_{\text{rand}}$ and a clustering coefficient $C\gg C_{\text{rand}}$ [30].

In this study, we introduce a quantity termed *degree of evolutionary conservation* that is obtained in the following way. (i) In addition to a reference set of human protein sequences, choose sets of known genes and corresponding protein sequences from completed genomes of species with different evolutionary distances. Here, we chose the four species *M. genitalium* (MG), *S. cerevisiae* (SC), *A. thaliana* (AT), and *C. elegans* (CE) from the taxonomic classes bacteria, yeast, plants, and invertebrates, respectively. (ii) Compare each human protein sequence to the set of all protein sequences of other species and detect in which species the human protein shares a significant sequence similarity (hence being evolutionary "conserved"), by using BLASTP matches $e<e_{\text{cr}}$. (iii) Define the degree of the conservation by the number of species in which the protein is conserved.

Note that both the detection of sequence similarities and the computation of the degree of conservation deliberately do not incorporate prior information about ancestral (e.g., orthologous or paralogous) relationships between protein sequences. Instead, this study expands the search space to the set of all known human proteins and focuses on partial sequence-related proteins. In order to reduce the complexity of this search, only the following combinations of sequence sets have been tested and sorted by their overall evolutionary relatedness to *H. sapiens* (HS): HS-CE; HS-CE-SC; HS-CE-

SC-AT; HS-CE-SC-AT-MG. The statistical analysis of the connectivity of conserved and nonconserved protein sequences was based on the nonparametric Wilcoxon rank sum test [25].

III. SEQUENCE-RELATED HUMAN PROTEINS CLUSTER HETEROGENEOUSLY

In this section, we compare the sequences of all known human proteins and study the extent to which protein sequences cluster into groups of shared sequence similarities.

We examine pairwise relationships between the protein sequences of known human genes, using algorithm BLASTP [18] (see Sec. II). When BLASTP detected a significant sequence similarity between two human proteins, these proteins were linked and joined into a cluster [26]. We cluster proteins based upon single linkage [27], so two proteins belong to one cluster if there is at least one direct or indirect link via one protein to the other. In the constructed network $\mathcal{N}$ of links between human protein sequences with shared similarities, we find that $N_{\mathrm{sim}}=17\,532$ out of $N_{\mathrm{total}}=21\,787$ human proteins give rise to clusters of various sizes, ranging from one large group of ~70% of proteins (12 281/17 532). The next cluster comprises ~0.9% of proteins (164/17 532) that, in turn, is followed by 1284 cluster of smaller sizes.

In order to test whether inferred sequence similarities at the amino acid level by using BLASTP are affected by chance alignments (false positives), we randomize a set of $N_{\mathrm{total}}=21\,787$ human protein sequences, while maintaining the genome-wide amino acid composition as well as the sequence length. In numerical simulations of random networks $\mathcal{N}_{\mathrm{rand}}$, fewer than five protein sequence pairs were false-positively detected as "significant" at a threshold expectation value of e_{cr}.

IV. SEQUENCE-RELATED PROTEINS FORM A SMALL-WORLD NETWORK

In this section, we treat the sequence relationships between human proteins obtained in the previous section as a network of sequence-linked proteins and we study the structural network properties, including the shortest path length and degree of clustering.

The above relationships between human proteins can be considered as a network $\mathcal{N}(N,E)$, with constituents of N nodes or proteins and $E=k$ edges or links between sequence-related proteins. The structure of such networks $\mathcal{N}(N,E)$ has been studied for several decades [28,29]. It is common to quantify local and global network properties by means of their characteristic path length and clustering coefficient [30,31]. The path length $L\in[1,N-1]$ measures the average number of edges in the shortest path linking two proteins. The clustering coefficient $C\in[0,1]$ quantifies the fraction of pairs of neighbors of a protein that are, in turn, neighbors, and is measured over all actual pairs of neighbors (see Sec. II).

Within the largest cluster (~70% of N_{total}), we find the average shortest path length $L\approx 4.45$. Consequently, about two-thirds of all pairs of proteins are sequence-related by traversing over four to five edges on average. When we compute over all clusters the clustering coefficient C, we find that $C\approx 0.75$. A comparison between C and $C_{\mathrm{rand}}\approx 0.008$, which is the value obtained for random networks (see Sec. II), shows that C is about two orders of magnitude larger and hence indicates a markedly higher degree of clustering as compared with randomly linked networks.

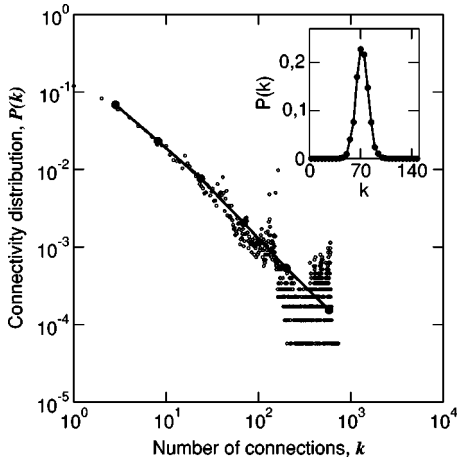

FIG. 1. Connectivity distribution of human protein sequences. Shown is a double logarithmic representation of the probability $P(k)$ that a given human protein is sequence-related with k other proteins. $P(k)$ is shown with $k=1$ linear (small circles) as well as with logarithmic bin sizes (full circles) to reduce noise. The decay of $P(k)$ can be approximated by a power law, $P(k)\sim k^{-\gamma}$, with a decay exponent $\gamma\approx 1.2$ (best-fit regression on double-logarithmic scales). The inset shows $P(k)$ for a numerical simulation of a random network $\mathcal{N}_{\mathrm{rand}}$, with the same overall number of proteins and connections as observed in the original empirical network. The connectivity distribution for $\mathcal{N}_{\mathrm{rand}}$ can be approximated by $P(k)_{\mathrm{rand}}\sim e^{-\langle k\rangle}(\langle k\rangle^k/k!)$, with the average connectivity $\langle k\rangle\approx 70$.

While L and C are average quantities that characterize the shortest path and cluster distribution, respectively, the full connectivity distribution $P(k)$ quantifies the probability that a protein has k neighboring or linked proteins. We find that $P(k)$ follows a power-law over approximately three orders of magnitude, $P(k)\sim k^{-\gamma}$, and decays with an exponent $\gamma\approx 1.2$ (see Fig. 1).

According to these measures, intermediate connectivity (relatively small value of L, common for random networks), a persistence for a high degree of clustering (relatively large value of C, common for regular networks), and a power-law connectivity distribution $P(k)$, means the network $\mathcal{N}(N,E)$ of human sequence-related proteins can be characterized as a so-called small world [30]. Structural properties of small worlds are widely shared by a number of other network types, including metabolic constituents [32], protein-protein interactions [33], or computer networks [23,34]. In the above network $\mathcal{N}(N,E)$, the clustering property reflects evolutionary patterns of gene duplications. The decay property of the connectivity distribution $P(k)\sim k^{-\gamma}$ has been related to the mechanisms of network expansion (in the number of nodes, N) and link-attachment preference (in the number of edges, E) [35,36].

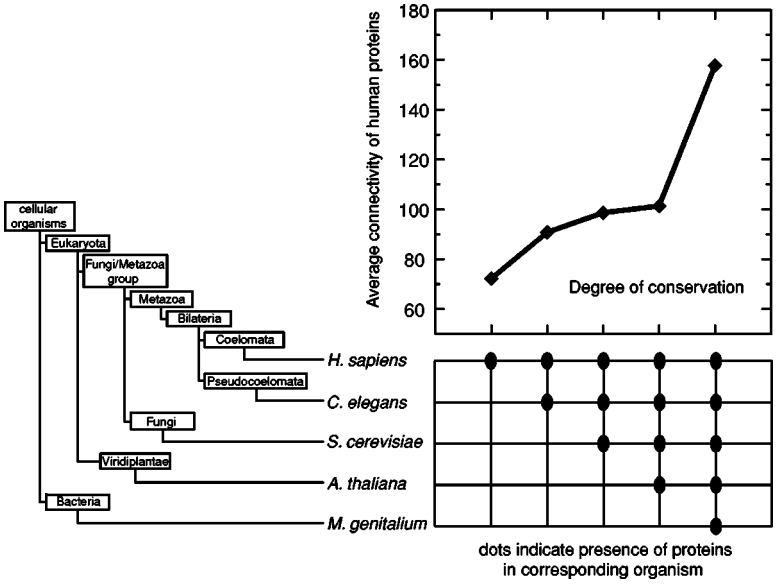

FIG. 2. Highly sequence-related human proteins are increasingly conserved throughout evolution. Left-hand side: schematically a phylogenetic tree with the taxonomic classes of the species *C. elegans*, *S. cerevisiae*, *A. thaliana*, and *M. genitalium* as compared to *H. sapiens*. Right-hand side: the average of the connectivity distribution of human proteins and the corresponding degree of evolutionary conservation, defined as the number of species in which the protein is sequence-related. Statistical analysis shows that each subset is significantly different ($p < 2.2 \times 10^{-16}$) from the corresponding nonconserved set (data not shown).

V. HIGHLY SEQUENCE-RELATED PROTEINS SHOW A HIGHER DEGREE OF CONSERVATION

In this section, we study the interrelationship between sequence-related human proteins and their degree of conservation throughout evolutionary history.

To this end, we test whether a human protein is phylogenetically *sequence-conserved* across other species, by showing a sufficiently high sequence similarity in a given species (cf. Secs. II and III). We define the degree of conservation as the number of species in which this protein is detectable. In a first comparison, we chose protein sequences of known genes from representative completely sequenced genomes of four species across four taxonomic classes, including bacteria, yeasts, plants, and invertebrates.

The central result of this study is shown in Fig. 2. For the network of human proteins, we find that the average connectivity is about $\langle k \rangle \approx 70$. Considering the set of human proteins that is conserved in *C. elegans*, we find that $\langle k \rangle$ increases to about 90 in HS-CE. Using further distantly related species, we find that the average connectivity $\langle k \rangle$ of sequence-related human proteins increases consistently with increasing evolutionary conservation in HS-CE-SC, HS-CE-SC-AT, and HS-CE-SC-AT-MG.

Figure 2 is in accord with a scenario in which gene duplication is coupled to partial conservation of sequence structure and allows for a generalization of the traditional concept of protein families.

VI. DISCUSSION

We study the network properties of sequence-related human proteins $\mathcal{N}(N,E)$ using the average shortest path length L, the clustering coefficient C, and the connectivity distribution $P(k)$.

We find that sequence-related human proteins cluster heterogeneously, with one single cluster comprising about 70% of the total number of $N_{\text{total}} = 21\,787$ known human proteins used in this study. Clearly, the size of clusters, as well as the total number of sequence relationships, depends on the choice of the threshold expectation value e_{cr}. The finding of a single, dominating group is in accord with previous studies [17]. The outcome that about one-third of human proteins remains in rather isolated clusters is fewer than what had been found in other estimates [17] and could possibly originate from differences in the detection of sequence similarities (see Sec. II) to construct the according protein groups [17,37].

We examine characteristic properties of $\mathcal{N}(N,E)$, and find that the average shortest path length $L \approx 4.45$ (largest cluster) and cluster coefficient $C \approx 0.75$ are such that $\mathcal{N}(N,E)$ qualifies as a small-world network. The connectivity distribution decays as $P(k) \sim k^{-1.2}$ across about three orders of magnitude. Previous empirical studies conducted on a variety of other network types, including the worldwide web, social, linguistic, citation, ecological, or cellular networks, have shown that $1 < \gamma < 3$ [24]. In particular, scale-free (power-law) network properties have been observed and/or predicted for gene-family sizes [40–42], as well as for the protein interaction network of *S. cerevisiae* [38]. In this study, the decay exponent $\gamma \approx 1.2$ ranks $\mathcal{N}(N,E)$ as a network with relatively gradual reduction in the node connectivity.

In order to corroborate these findings, we used two modified threshold expectation values (by increasing and then decreasing e_{cr} by a factor of 10). Using these different thresholds did not qualitatively change the power-law distribution of $P(k)$ nor the dependence of the connectivity on the degree of conservation (data not shown), so our results remain largely unaffected in this range of expectation values. The presence of a single large cluster may explain to some degree the bioinformatic difficult recognition of evolutionary "young" as well as "distant" genes/proteins, as compared with the recognition of well-conserved ones, by *ab initio* gene-finding algorithms that are based on statistical features of DNA sequences and protein-homology search [43–45].

We study the dependence between the connectivity distribution and the conservation of human proteins, and we find that the average connectivity increases within sets of proteins

with increasing degree of conservation. The feature that highly sequence-related proteins show a higher conservation relates the small-world network structure to the evolution of genomes, and substantiates the close relationship of genes as a common principle of molecular evolution [14,42,46]. Using protein-protein interaction data derived from studies of *S. cerevisiae*, pairwise interactions of yeast proteins have been found to be evolutionary conserved across several species [47,48]. In addition, it is interesting to note that conserved yeast proteins exhibiting a relatively high connectivity tend to evolve more slowly (reduced number of substitutions/amino acid site) [39].

Our results integrate with several further lines of evidence that are indicative of evolutionary mechanisms that are footprinted in genomic and proteomic network properties. An initial statistical modeling approach relates empirically observed network properties to network growth and preferential attachment of new connections [24,35]. While this model assumes that all nodes possess indistinguishable properties except for their connectivity, further comparisons of model predictions with biological networks (genes, proteins) will be of interest in ongoing works [49,50].

ACKNOWLEDGMENTS

We thank A. "Elmo" Esser (MIT, Cambridge, MA) for valuable comments, and the DFG (Deutsche Forschungsgemeinschaft) and NGFN (Nationales Genomforschungsnetz) for financial support.

[1] O. Weiss, M. Jiménez-Montano, and H. Herzel, J. Theor. Biol. **206**, 379 (2000).
[2] H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999).
[3] V.B. Zhurkin, J. Biomol. Struct. Dyn. **4**, 785 (1981).
[4] O. Weiss and H. Herzel, J. Theor. Biol. **190**, 341 (1998).
[5] E.N. Trifonov and J.L. Sussman, Proc. Natl. Acad. Sci. U.S.A. **77**, 3816 (1980).
[6] D. Holste *et al.*, Phys. Rev. E **67**, 061913 (2003).
[7] E.S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[8] J.C. Venter *et al.*, Science **291**, 1304 (2001).
[9] J.A. Bailey *et al.*, Genome Res. **11**, 1005 (2001).
[10] The Arabidopsis Genome Initiative, Nature (London) **408**, 796 (2000).
[11] M.E. Johnson *et al.*, Nature (London) **413**, 514 (2001).
[12] J.A. Bailey *et al.*, Science **297**, 1003 (2002).
[13] M. Lynch, Science **297**, 945 (2002).
[14] W.H. Li, *Molecular Evolution* (Sunderland, Sinnauer Associates, Sunderland, MA, 1997).
[15] G. Glusman *et al.*, Genome Res. **11**, 685 (2001).
[16] M. Leveugle *et al.*, Nucleic Acids Res. **31**, 63 (2003).
[17] W.H. Li *et al.*, Nature (London) **409**, 847 (2001).
[18] S.F. Altschul *et al.*, Nucleic Acids Res. **25**, 3389 (1997).
[19] BLASTP is part of the software package BLASTALL, the distribution of which is publicly available from the NCBI (http://www.ncbi.nlm.nih.gov). In this study, we use version 2.2.6. Searching a database of a particular size, the expectation value (e) is a parameter that incorporates the background noise that exists for random matches between sequences and estimates the number of hits one can expect by chance in a given set of sequences.
[20] G. Yona, N. Linial, and M. Linial, Nucleic Acids Res. **28**, 49 (2000).
[21] A. Krause, J. Stoye, and M. Vingron, Nucleic Acids Res. **28**, 270 (2000).
[22] A.J. Enright, V. Kunin, and C.A. Ouzounis, Nucleic Acids Res. **31**, 4623 (2003).
[23] R.F. Cancho, C. Janssen, and R.V. Sole, Phys. Rev. E **64**, 046119 (2001).
[24] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[25] R.R. Sokal and F.J. Rohlof, *Biometry* (W.H. Freeman and Co., New York, 1981).
[26] For the set of matches within the $N_{\text{total}}=21\,787$ human proteins, the median of the match length distribution is 197 amino acids, and the 25.0 and 75.0 percentiles of that distribution are 134 and 382 amino acids, respectively. We consider the sequence relation to be significant when the e value of the BLASTP match of two sequences is below $e_{\text{cr}}=0.001$. The shortest sequence length at which sequences align at the threshold of e_{cr} is 85 amino acids for our dataset.
[27] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification* (John Wiley & Sons, New York, 2001).
[28] S.A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).
[29] D.J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).
[30] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[31] H. Herzel, Fractals **6**, 301 (1998).
[32] H. Jeong *et al.*, Nature (London) **407**, 651 (2000).
[33] H. Jeong *et al.*, Nature (London) **411**, 41 (2001).
[34] R. Albert, H. Jeong, and A.-L. Barabasi, Nature (London) **401**, 130 (2001).
[35] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).
[36] A. Vazquez, Phys. Rev. E **67**, 056104 (2003).
[37] B. Rost, Protein Eng. **12**, 85 (1999).
[38] S.H. Yook, Z.N. Oltvai, and A.-L. Barabasi, Proteomics **4**, 928 (2004).
[39] H.B. Fraser *et al.*, Science **296**, 750 (2002).
[40] M.A. Huynen and E. van Nimwegen, Mol. Biol. Evol. **15**, 583 (1998).
[41] I. Yanai, C.J. Camacho, and C. DeLisi, Phys. Rev. Lett. **85**, 2641 (2000).
[42] J. Qian, N.M. Luscombe, and M. Gerstein, J. Mol. Biol. **313**, 673 (2001).
[43] R. Guigo *et al.*, Genome Res. **10**, 1631 (2000).
[44] R.F. Yeh, L.P. Lim, and C.B. Burge, Genome Res. **11**, 803 (2001).

[45] M.Q. Zhang, Nat. Rev. Genet. **3**, 698 (2002).
[46] A. Wagner, Proc. Natl. Acad. Sci. U.S.A. **91**, 4387 (1994).
[47] H. Qin *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 12 820 (1994).
[48] S. Wuchty, Z.N. Oltvai, and A.-L. Barabasi, Nat. Genet. **35**, 176 (2003).
[49] E. Eisenberg and E.Y. Levanon, Phys. Rev. Lett. **91**, 138701 (2003).
[50] V. Kunin, J.B. Pereira-Leal, and C.A. Ouzounis, Mol. Biol. Evol. **21**, 1171 (2004).
[51] Accessible at http://www.ensembl.org (temporal freeze of December, 2003).
[52] Accessible at ftp://ftp.ncbi.nlm.nih.gov/genomes/.
[53] Accessible at ftp://ftpmips.gsf.de/yeast/catalogues/.
[54] Accessible at http://www.ncbi.nlm.nih.gov/Taxonomy.