The SYSTERS Protein Family Database: Taxon-related Protein Family Size Distributions and Singleton Frequencies

Thomas Meinel, Martin Vingron, and Antje Krause

Max Planck Institute for Molecular Genetics
Dept. Computational Molecular Biology
Ihnestrasse 73, D-14195 Berlin

E-mail: {Thomas.Meinel|Martin.Vingron|Antje.Krause}@molgen.mpg.de

Keywords: protein family; large scale clustering; taxonomy; taxon-related; cluster size distribution

Abstract

Based on the SYSTERS protein family database, we present taxon-related protein family frequencies and distributions. A set of *taxon-related protein families* is a subset of the whole family set with respect to one taxon, where *taxon* is not restricted to the species level but may be any rank in the taxonomy. We examine eight ranks in the lineages of seven organisms. A strong linear correlation is observed between the total number of different families and the number of sequences in the data set under consideration. We fitted the generalised power-law function to protein family distributions in a least-squares sense excluding singleton frequencies. Taxon-related family distributions tend to have the same shape and a negative slope being not larger than -2.1 for large data sets. For smaller data sets, the slope is decreasing down to -3.7. Slopes of family distributions are found to be slowly increasing towards higher taxonomic ranks. Our observations lead to a new estimation of single sequence cluster frequencies. Data sets of various species are studied with respect to being complete or incomplete.

Introduction

The determination of protein families has been of interest since scientists began to analyse proteins. Several sequence based clustering methods have been proposed. A review of these methods is presented by Heger and Holm in [6]. Other concepts consider protein architecture and structural features, i.e., folds or domains. Some databases, like the SCOP [13] and the CATH [15] classification systems, provide a combination of both sequence-based and structure-based approaches.

Estimates of family frequencies have been reported for structure based data sets [3, 14]. A recent enumeration of protein domain families close to both clustering concepts is published in [7]. Discrete protein family distributions are described presenting power-law or generalised power-law function fits, reviewed in [9]. Analysing distributions resulting from sequence-based methods, small data sets are examined by Huynen and von Nimwegen [8], while Unger *et al.* [17] compare large scale data sets.

Several approaches focus on the (complete) genome or proteome of organisms. Others include (incomplete) data sets from various sources and different species. Accordingly, protein

rank	superkingdom	kingdom	phylum	class	order	family	genus	species
incomplete data sets								
	Eukaryota	Metazoa	Chordata	Mammalia	Primates	Hominidae	Homo	Hs
Hs	41,757	26,207	11,892	10,372	7,537	7,442	7,418	7,418
	(156,833)	(96,404)	(52,406)	(39,902)	(18,793)	(17,692)	(17,399)	(17,399)
	Eukaryota	Metazoa	Chordata	Mammalia	Rodentia	Muridae	Mus	Mm
Mm	41,757	26,207	11,892	10,372	5,197	5,129	4,132	4,064
	(156,833)	(96,404)	(52,406)	(39,902)	(15,627)	(14,975)	(9,716)	(9,539)
	Eukaryota	Viridiplantae	Embryophyta	(eudicotyledons)	Brassicales	Brassicaceae	Arabidopsis	At
At	41,757	9,717	9,344	8,100	6,862	6,854	6,756	6,756
	(156,833)	(35,114)	(33,938)	(27,680)	(17,739)	(17,692)	(16,891)	(16,883)
complete data sets								
	Eukaryota	Metazoa	Arthropoda	Insecta	Diptera	Drosophilidae	Drosophila	Dm
Dm	41,757	26,207	10,245	10,015	9,480	9,263	9,261	9,179
	(156,833)	(96,404)	(20,928)	(19,851)	(16,686)	(15,690)	(15,636)	(14,896)
	Eukaryota	Metazoa	Nematoda	Chromadorea	Rhabditida	Rhabditidae	Caenorhabditis	Ce
Ce	41,757	26,207	9,790	9,770	9,646	9,611	9,610	9,607
	(156,833)	(96,404)	(19,521)	(19,478)	(18,959)	(18,701)	(18,695)	(18,642)
	Eukaryota	Fungi	Ascomycota	Saccharo-mycetes	Saccharo-mycetales	Saccharo-mycetaceae	Saccharomyces	Sc
Sc	41,757	7,751	7,547	5,141	5,141	5,009	4,907	4,884
	(156,833)	(17,060)	(16,088)	(7,990)	(7,990)	(7,213)	(6,672)	(6,577)
	Bacteria	–	Proteo-bacteria	Gammaproteo-bacteria	Entero-bacteriales	Entero-bacteriaceae	Escherichia	Ec
Ec	28,360	–	12,650	7,753	5,391	5,377	4,262	4,260
	(83,132)	–	(35,816)	(19,020)	(10,626)	(10,346)	(6,516)	(6,508)

Table 1. Total cluster frequencies and total numbers of sequences from 46 taxa. Following the lineages of seven organisms, *H. sapiens* (*Hs*), *M. musculus* (*Mm*), *A. thaliana* (*At*), *D. melanogaster* (*Dm*), *C. elegans* (*Ce*), *S. cerevisiae* (*Sc*), and *E. coli* (*Ec*), for up to eight ranks. A kingdom rank for Ec is not available; the taxon *eudicotyledons* corresponds to a class rank for the plant At and is therefore in parantheses. Sequence numbers are in parentheses.

families are either built within one species or covering data from various species. The SYSTERS protein family database is an automatically generated partitioning of all publicly available protein sequences into family and superfamily clusters [11]. Including all sequences from various species opens the opportunity to analyse also higher taxonomic levels (ranks) and allows for an extended view on the evolution of protein families. We compared the taxon-related protein family frequencies and distributions for eight ranks in the lineages of seven organisms. Singleton protein families are found to form the most abundant family size in all distributions. With respect to their biological relevance they have to be surveyed carefully [4], mostly being an artefact of the underlying clustering method. The SYSTERS Release 3 database consists of 82,449 disjoint family clusters with 55,181 of them being single sequence clusters. Sequences ending up in these clusters are mostly fragmental and of minor length. Although database search methods take the length of a sequence into account, shorter sequences in average result in worse E-values than longer sequences when used as query sequence. We present a new estimation of single sequence cluster frequencies based on protein family distributions.
The paper is organised as follows: We start with a short recapitulation of the methods and data sets used in the SYSTERS database. The next paragraph covers the taxonomic basics and the data selected for our approach, followed by a description of the methods used to calculate family frequencies and distributions. We will report on our results obtained by the analysis of 46 individual taxa at different taxonomic levels.

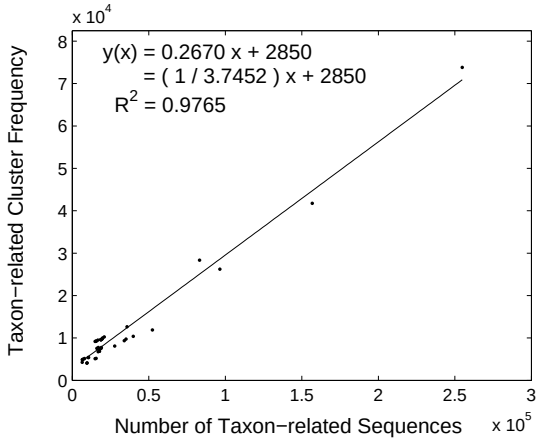

Figure 1. Total cluster frequency and total number of sequences related to 46 taxa. Frequencies, numbers and related taxa as given in Table 1. Slope, intercept and correlation coefficient R^2 according to a least square fit.

Methods

SYSTERS Protein Family Database. The SYSTERS database provides an automatically generated grouping of all publicly available protein sequences into disjoint superfamily and family clusters [10]. The underlying redundant sequence set contains sequences from the SWISS-PROT/TrEMBL [2] and the PIR [20] databases as well as of several completely sequenced organisms, e.g., worm [5], fly [16], and yeast [18]. The data set was made up in July 2000 and contains $583,448$ sequences. Sequences which are identical or nearly identical (at least 99 % identity) to other sequences over at least 95 % of their entire length were considered redundant, and were removed from the initial sequence set. All results in this paper refer to the non-redundant data set of $290,809$ sequences. The classification of this set of sequences into the SYSTERS cluster set is mainly based on a traditional database search tool [1] and done in two steps, a similarity searching step and a clustering step. First, each sequence in the database is searched against the whole sequence database down to a weak E-value of 0.05. Then, a series of graph-based clustering methods is applied to these pairwise E-values.

Taxonomy and chosen Taxa. Organisms are systematically sorted into biologically meaningful groups by the taxonomy. A continually curated data set is provided by the NCBI taxonomy [19]. Scientific names, taxonomic identification numbers (TaxIDs), lineages, and ranks are obtained from this source and are used in the SYSTERS database. A taxon is the systematic entity of the taxonomy covering organisms as leaves and groups of organisms as internal nodes of the taxonomic tree. A lineage is the consecutive listing of all taxa an organism belongs to.
From the SYSTERS taxonomy web interface (`http://systers.molgen.mpg.de`), cluster sets were obtained for 46 taxa. Eight ranks (superkingdom, kingdom, phylum, class, order, family, genus, and species) were chosen along the lineages of the organisms *Homo sapiens* (*Hs*), *Mus musculus* (*Mm*), *Arabidopsis thaliana* (*At*), *Drosophila melanogaster* (*Dm*; complete data set), *Caenorhabditis elegans* (*Ce*; complete), *Saccharomyces cerevisiae* (*Sc*; complete), and *Escherichia coli* (*Ec*; complete).

Taxon-related Cluster Sets and Cluster Frequencies. Taxon-related cluster sets are extracted as follows: From the whole SYSTERS cluster set, we remove all sequences not belonging to the taxon under consideration. The remaining set of non-empty clusters builds a cluster set only related to this taxon. If we choose for example the taxon *chordata*, we subtract all sequence entries of non-chordate organisms from the whole SYSTERS cluster set. In this

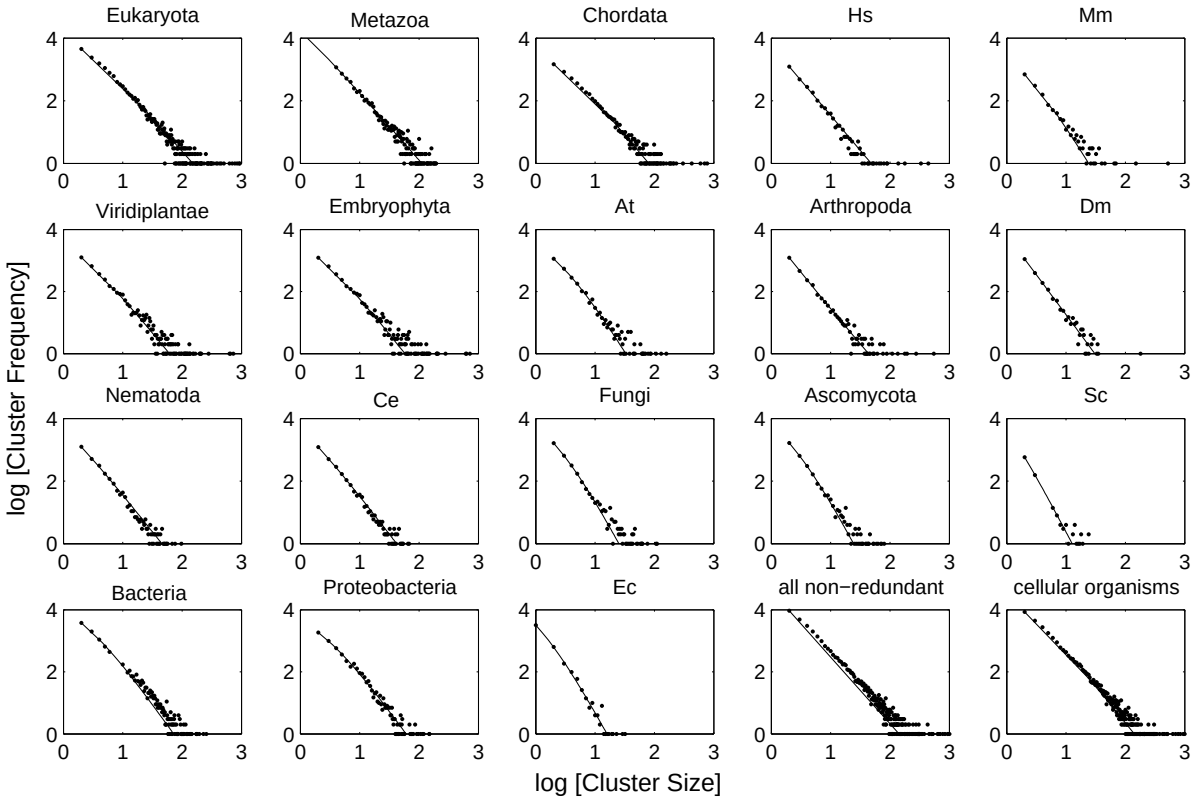

Figure 2. Taxon-related cluster distributions of 19 taxa and of the whole SYSTERS data set. Taxa in the lineage of the seven organisms *H. sapiens, M. musculus, A. thaliana, D. melanogaster, C. elegans, S. cerevisiae, E. coli*, covering the ranks superkingdom, kingdom, phylum, species. The taxon *cellular organisms* includes all other taxa. Repeated ranks are shown only once. Distributions are fitted excluding singleton frequencies (exception: Ec). For detailed results see Table 2.

case, the remaining set of chordate clusters contains $11,892$ clusters (*cluster frequency*) with $52,406$ sequences, see Table 1.

Taxon-related Cluster Size Distributions and Singleton Frequencies. The generalised power-law function $y = a * (x + i)^{m}$ can be fitted to protein family frequencies and sizes (distribution) in a least-squares sense [12]. The fitting procedure minimises the sum error of the three parameters $a,\ i$ and m with the slope m as a significant parameter. Distribution curves were fitted excluding singletons (cluster size $= 1$) knowing the fact that the SYSTERS procedure sorts unclustered sequences into single sequence clusters. Extrapolating the fit curves to cluster size 1 leads to a difference between the extrapolated and the obtained singleton frequency in the non-redundant data set. Such a difference gives furthermore an estimation for the singleton over-formation in taxon-related sets. We examined the following ranks for the seven organisms: superkingdom, kingdom, phylum, and species.

Results

Taxon-related Cluster Frequencies. The reduction from the whole SYSTERS set to taxon-related data sets going along the lineage down to the species rank is shown in Table 1. Some

Taxon	Cluster Frequency	Singleton Cluster Frequency obtained from the database set	ΔS	Slope m	Increment i	Intercept a 10^5
all non-redundant entries	82,449	55,181	32,436	−2.39	1.22	1.52
cellular organisms	73,824	48,960	28,261	−2.38	1.21	1.37
Eukaryota	41,757	27,491	17,227	−2.17	1.14	0.53
Metazoa	26,207	16,716	8,716	−2.20	1.00	0.368
Chordata	11,892	7,094	3,944	−2.33	1.59	0.291
H. sapiens	7,418	4,784	− 745	−2.25	0.04	0.061
M. musculus	1,477	2,587	421	−3.24	1.38	0.36
Viridiplantae	9,717	6,155	2,871	−2.35	0.99	0.166
Embryophyta	9,344	5,848	2,757	−2.44	1.19	0.208
A. thaliana	6,756	4,104	1,013	−3.15	1.69	0.701
Arthropoda	12,245	7,710	524	−2.26	−0.15	0.049
D. melanogaster	9,179	7,098	712	−2.59	0.04	0.070
Nematoda	9,790	7,037	2,333	−2.39	0.34	0.096
C. elegans	9,607	6,966	2,335	−2.47	0.39	0.105
Fungi	7,742	4,574	−1,143	−3.61	1.42	1.40
Ascomycota	7,547	4,431	−1,551	−3.64	1.35	1.34
S. cerevisiae	4,884	3,990	− 688	−3.69	0.32	0.129
Bacteria	28,360	18,297	5,883	−2.76	1.62	1.31
Proteobacteria	12,650	7,554	3,167	−2.64	1.61	0.554
E. coli	4,260	3,201	(*n.d.*)	−3.56	0.72	0.222

Table 2. Taxon-related cluster frequencies, taxon-related singleton frequencies and regression parameters obtained from distribution curve fits. Cluster frequencies for the non-redundant cluster set (first row) and taxon-related cluster sets for seven organisms, covering the ranks superkingdom, kingdom, phylum, and species. The taxon *cellular organisms* is the union of all other taxa (second row). Sub-levels are indented; repeated ranks are shown only once. Distribution plots are shown in Figure 2. With exception of *Ec*, the singleton difference ΔS is calculated as the difference between obtained and calculated values using the fitting parameters $a,\ i,\ m$ for the generalised power-law function, $y = a * (x + i)^{\ m}$.

organisms like *Dm*, *Ce* or *Sc* share a wide range of the same lineage with a small number of other organisms. This results in quite constant frequency levels in the concerning ranks. For other organisms like *Hs* or *Mm*, ranks close to the species level have significantly raised frequency levels, maybe because of the greater interest in research. Comparing the kingdoms, it is remarkable that *Metazoa* have a two- or threefold larger number of protein families than *Fungi* or plants (*Viridiplantae*). For the bacterium *Ec*, the phylum rank (*Proteobacteria*) cluster frequency is comparable to that of *Chordata*. The cluster set covering *cellular organisms* is found to be in the same order of magnitude as the non-redundant data set, in total cluster frequency (see Table 2) as well as in the total number of sequences (*cellular organisms*: $254,869$).
Taxon-related cluster frequencies are linearly correlated to the total number of non-redundant sequences (correlation coefficient, $R^2 = 0.9765$; Figure 1) over the whole set of the examined taxa, suggesting an average general cluster size between 3.5 and 4 and an intercept at $2,850$.

Cluster Size Distributions for Taxon-related Data Sets. Discrete protein family size distributions have the same shape in all examined data sets and occur in a power-law like manner. Slopes of -2.1 are observed for large cluster sets, i.e., the whole non-redundant set or the set for *cellular organisms*. They are strongly steeper (~ -3.7) for data sets of completely sequenced organisms (*Ec*, *Sc*), having a small sequence space and resulting in $4,000$ to $5,000$

protein families. Slopes are moderate (~ -2.5) for organisms (*Dm*, *Ce*) having a two-fold higher number of clusters, while slopes for incomplete data sets (*At*, *Mm*) with low or medium cluster frequencies are again found to be steeper than -3.0. (The incomplete data set of *Hs* is apparently large enough to produce a moderate slope.) For the whole lineage of *Sc*, slopes remain in the same range of the species level until to the kingdom level, together with an average cluster number of around $7,700$ clusters.

A large number of singletons are formed by the SYSTERS clustering procedure: $55,181$ out of 290.809 non-redundant sequences. If singletons are included into the fitting procedure for the datasets of 19 taxa, only plots of five species show a negative curvature in log-log-plots corresponding to a positive fit parameter i, (data not shown). Thus, singletons were excluded from the calculation, resulting in a negative curvature in most plots of Figure 2. As a consequence, singleton frequencies can be calculated as the y-axis intercept using the generalised power-law fit parameters. The difference between the calculated and the obtained singleton frequency is the singleton difference ΔS, see Table 2. Taxon-related singleton differences ΔS are linearly dependent on both, the number of taxon-related sequences and on total taxon-related cluster frequencies ($R^2 > 0.98$; cf. Table 2).

Discussion

The procedure presented in this paper relies on the reduction of an existing protein family (cluster) set using the taxon criterion. The SYSTERS cluster structure is not affected, larger sets contain all sequences of smaller sets. Going along the lineage of an organism towards the species level, cluster sizes as well as cluster frequencies decrease. Total taxon-related cluster frequencies and total numbers of (taxon-related) sequences are linearly correlated ($R^2 > 0.97$), an average cluster size of ~ 3.75 is obtained.

We fitted the generalised power-law function to cluster distributions by excluding singletons. We included the extra parameter i to take into account the curvature of the generalised power-law function. For cluster distribution curvature, we found equal shapes in the plots of large data sets as well as of small data sets. Distributions of all cluster sets are found to be power-law like. The slope m as the most cited fitting parameter follows general dependencies: It is not larger than -2.1 (whole non-redundant cluster set), suggesting a general saturation is reached at this point. The slope is not smaller than -3.7 for small data sets like that of *Sc* ($4,900$ clusters) or *Ec* ($4,200$ clusters). These slopes are close to that of comparably small and complete data sets presented in [8]. The slope is decreasing along the lineage from the superkingdom level to the species level. As a general observation, slopes decrease if data sets are reduced.

Comparing total taxon-related clusters and sequences, we find $2,850$ clusters more than expected, which can be interpreted as being generally over-formed by the clustering procedure. Furthermore, we presented a method to re-calculate singleton frequencies using generalised power-law fitting parameters combined with successive reduction of the SYSTERS data set along independent lines, which are the lineages of the seven organisms. We observe an over-formation of singletons between the calculated and the obtained singleton frequencies, which is correlated to the underlying sequence or cluster data. Sometimes denoted as ORFans [4], singletons often lack biological relevance or are a by-product of a sequence-based clustering method.

We will continue our analyses on the growing number of upcoming data sets of completely sequenced organisms. Alternatively to our concept to reduce a whole data set to taxon-related data sets, we plan to cluster taxon-related sequence sets with our SYSTERS algorithm separately.

Acknowledgements

We would like to thank Christine Steinhoff, Thomas Manke and Tobias Müller for fruitful discussions, helpful support and instructions in fitting power-law plots.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.

[2] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–70, Jan 2003.

[3] C. Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544, Jun 1992.

[4] D. Fischer and D. Eisenberg. Finding families for genomic ORFans. *Bioinformatics*, 15(9):759–62, Sep 1999.

[5] T. W. Harris, R. Lee, E. Schwarz, K. Bradnam, D. Lawson, W. Chen, D. Blasier, E. Kenny, F. Cunningham, R. Kishore, J. Chan, H. M. Muller, A. Petcherski, G. Thorisson, A. Day, T. Bieri, A. Rogers, C. K. Chen, J. Spieth, P. Sternberg, R. Durbin, and L. D. Stein. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res*, 31(1):133–137, Jan 2003.

[6] A. Heger and L. Holm. Towards a covering set of protein family profiles. *Prog Biophys Mol Biol*, 73(5):321–37, 2000.

[7] A. Heger and L. Holm. Exhaustive enumeration of protein domain families. *J Mol Biol*, 328(3):749–67, May 2003.

[8] M. A. Huynen and E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol*, 15(5):583–589, May 1998.

[9] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–23, Nov 2002.

[10] A. Krause. *Large Scale Clustering of Protein Sequences*. PhD thesis, University of Bielefeld, 2002.

[11] A. Krause, S. A. Haas, E. Coward, and M. Vingron. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res*, 30(1):299–300, Jan 2002.

[12] V. A. Kuznetsov. Statistics of the number of transcripts and protein sequences encoded in the genome. In W. Zhang and I. Shmulevich, editors, *Computational and Statistical Approaches to Genomics*, pages 125–171. Kluver, Boston, 2002.

[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, Apr 1995.

[14] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634, Dec 1994.

[15] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–108, Aug 1997.

[16] The FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, 31(1):172–175, Jan 2003.

[17] R. Unger, S. Uliel, and S. Havlin. Scaling law in sizes of protein sequence families: From superfamilies to orphan genes. *Proteins*, 51(4):569–76, Jun 2003.

[18] S. Weng, Q. Dong, R. Balakrishnan, K. Christie, M. Costanzo, K. Dolinski, S. S. Dwight, S. Engel, D. G. Fisk, E. Hong, L. Issel-Tarver, A. Sethuraman, C. Theesfeld, R. Andrada, G. Binkley, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res*, 31(1):216–218, Jan 2003.

[19] D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, and L. Wagner. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, 31(1):28–33, Jan 2003.

[20] C. H. Wu, L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang, and W. C. Barker. The Protein Information Resource. *Nucleic Acids Res*, 31(1):345–347, Jan 2003.