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
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 superfamiliy 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.

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.

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs</td>
<td>Eukarya</td>
<td>22</td>
<td>27</td>
<td>10</td>
<td>7</td>
<td>Homo</td>
</tr>
<tr>
<td></td>
<td>Metazoa</td>
<td>11</td>
<td>15</td>
<td>4</td>
<td>3</td>
<td>Hs</td>
</tr>
<tr>
<td>Mm</td>
<td>Eukarya</td>
<td>22</td>
<td>27</td>
<td>10</td>
<td>7</td>
<td>Homo</td>
</tr>
<tr>
<td></td>
<td>Metazoa</td>
<td>11</td>
<td>15</td>
<td>4</td>
<td>3</td>
<td>Mm</td>
</tr>
<tr>
<td>At</td>
<td>Viridiplantae</td>
<td>9</td>
<td>15</td>
<td>4</td>
<td>3</td>
<td>At</td>
</tr>
<tr>
<td></td>
<td>Embryophyta</td>
<td>(15,112)</td>
<td>(33,938)</td>
<td>(27,680)</td>
<td>(17,739)</td>
<td>(16,792)</td>
</tr>
</tbody>
</table>

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.
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://syster.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 (A1), 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 \( y = a \times (x + i)^m \) 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
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).

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 $\Delta S$ is calculated as the difference between obtained and calculated values using the fitting parameters $a$, $i$, $m$ for the generalised power-law function, $y = a \cdot (x + i)^m$.

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 ($\sim -2.5$) 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 $\Delta S$, see Table 2. Taxon-related singleton differences $\Delta S$ 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 $\sim 3.75$ 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


