Sequence analysis

The properties of protein family space depend on experimental design

Victor Kunin1,*,†, Sarah A. Teichmann2, Martijn A. Huynen3 and Christos A. Ouzounis1

1Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, 2MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and 3Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, University of Nijmegen, Nijmegen, The Netherlands

Received on November 4, 2004; revised on February 16, 2005; accepted on March 9, 2005
Advance Access publication March 15, 2005

ABSTRACT

Motivation: Databases of protein families often exhibit drastically different properties of the protein family space.

Results: We compared the properties of protein family space as reflected by exhaustive protein family databases and databases with predefined families. We used TRIBES, Protomap, ProDom and COGs as representatives of the exhaustive databases, and Pfam-A and Superfamily as databases that predefine families. We observe a power-law distribution of family sizes in all these databases, albeit in predefined databases the power-law line collapses before reaching smaller sized families. We discuss the future trends of this power-law distribution and suggest that saturation in the sampling of protein family space will result in a distortion of the power law in small family sizes. For larger genome sizes, predefined databases show logarithmic growth of the number of families per genome, whereas exhaustive databases exhibit a virtually linear relationship. All databases consistently differ in the proportion of protein families shared between taxa. Predefined databases have a larger number of protein families shared between the three domains of life, while exhaustive databases show a much more fragmented distribution. We argue that these discrepancies reflect alternative approaches to the trade-off issue of sensitivity versus specificity in the detection of homologous proteins. We conclude that these properties are complementary rather than contradictory, while describing the protein universe from different perspectives.

Contact: vkunin@lbl.gov

INTRODUCTION

In an attempt to consolidate the data constantly being produced by sequencing projects, recent years have seen a rise in the number of protein family databases (refer Ouzounis et al., 2003 for a recent review). Some of these databases are based on exhaustive, completely automatic, unsupervised clustering from pairwise similarities of protein sequences. These databases include SYSTERS (Krause et al., 2002), ClustR (Kriventseva et al., 2001), Protomap (Yona et al., 2000), ProDom (Bru et al., 2005), COGs (Tatusov et al., 2001) and TRIBES (Enright et al., 2003), with the last two dedicated exclusively to clustering sequences derived from genome projects.

All these databases aim to cover the complete protein universe and do not assume any prior knowledge about these families. The other type of protein family databases is based on existing knowledge of related groups of proteins and use profile or hidden Markov Model (HMM)-based searches to draw more sequences into these families. Examples of such databases are Pfam (Bateman et al., 2004), TIGRFAMS (Haft et al., 2003) and SMART (Letunic et al., 2002). These databases often report better sensitivity of searches than databases based on unsupervised clustering. Databases that assign structural domains to proteins with profiles or HMMs make use of the evolutionary information inherent in the protein structures to assign domains to predefined structural families. Examples are the Superfamily database, Gene3D (Gough et al., 2001), 3D-PSSM (Bates et al., 1998) and others. In these databases, several distinct HMMs can represent the same protein family.

In our previous analysis of the TRIBES database, we have observed three key properties of protein family space: (1) a power-law distribution of protein family sizes, (2) constant paralogy levels across microbial genomes and (3) a small number of families common to the three domains of life (Enright et al., 2003). In this study, we examine the key properties of protein family space in a number of databases dedicated to describing protein families in complete genomes. We thus used COGs, a database of groups of orthologous proteins constructed automatically with some manual intervention (Tatusov et al., 2001), ProDom-CG, a database of protein domains that are automatically generated from complete genome sequences (Bru et al., 2005) and the Superfamily database, which contains HMM-based assignments of structural superfamilies to proteins from completely sequenced genomes (Gough et al., 2001). To resolve issues that do not depend directly on the completeness of the genomic data we also used Protomat, a completely automatic exhaustive database of protein families (Yona et al., 2000) and Pfam, a manually curated database of protein families and HMM assignments (Bateman et al., 2004). A brief summary of the databases we used and their principal features are presented in Table 1.

RESULTS

Distribution of protein family sizes in protein family databases

The size distributions of protein families, superfamilies and folds was shown to follow a power-law in many individual genomes
Table 1. Overview of the databases used in the study

<table>
<thead>
<tr>
<th>Database</th>
<th>Type</th>
<th>Level</th>
<th>Version</th>
<th>Family definition method</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIBES</td>
<td>Exhaustive</td>
<td>Family</td>
<td>84</td>
<td>Blast and MCL clustering</td>
</tr>
<tr>
<td>Protomap</td>
<td>Exhaustive</td>
<td>Family</td>
<td>3.0</td>
<td>Blast and hierarchical clustering</td>
</tr>
<tr>
<td>COGs</td>
<td>Exhaustive</td>
<td>Ortholog</td>
<td>Updated</td>
<td>Blast and bidirectional best hits and triplet clustering</td>
</tr>
<tr>
<td>ProDom</td>
<td>Exhaustive</td>
<td>Domain family</td>
<td>CG120</td>
<td>Blast and extension with PSI-Blast searches</td>
</tr>
<tr>
<td>Pfam-A</td>
<td>Predefined</td>
<td>Domain family</td>
<td>Pfam13</td>
<td>Manual family definition and extension with HMM searches</td>
</tr>
<tr>
<td>Superfamily</td>
<td>Predefined</td>
<td>Domain superfamily</td>
<td>1.65</td>
<td>Structural definition of superfamily and extension with HMM searches</td>
</tr>
</tbody>
</table>

Fig. 1. Distribution of family sizes in various databases. The x-axis represents protein family size, and the y-axis represents the frequency or the count of protein families for each given size. Family sizes were extracted from each database as available in summer 2004. The power law in ProDom database is shown with a dotted line, to demonstrate the difference in distribution of smaller versus larger domain families. Due to the small size, the data from Superfamily database was binned in exponentially increasing bins of powers of 2.

(Huynen and van Nimwegen, 1998; Harrison and Gerstein, 2002; Koonin et al., 2002) and globally for all genomes in an analysis of the protein families in the TRIBES database (Enright et al., 2003). In this paper, we compare the global distribution of family sizes over all genomes for a range of other databases. All databases in the analysis exhibit a power-law distribution of protein family sizes in the part of the graph that represents larger families (Fig. 1). The power-law line holds up to the mid-size protein families. Thus, for all protein families with large enough coverage, a clear power-law cluster size distribution can be observed.

For databases that determine protein families exhaustively based on pairwise sequence similarity (TRIBES and Protomap), the power-law line holds all the way to family size one (singletons). Though the COGs database is somewhat similar in its clustering procedure,
The properties of protein family space depend on the number of genes and the number of families in prokaryotes using the TRIBES database.

We recently observed a constant relationship between the number of protein families and the number of genes. Some reports suggested that paralogy increases with expanding genome size (Pushker et al., 2004; Chothia et al., 2003; Muller et al., 2002). However, we recently observed a constant relationship between the number of genes and the number of families in prokaryotes using the TRIBES database.

Future trends for protein family size distribution

As protein sequence space is explored and gaps closed, will the distribution of protein family sizes in exhaustive databases continue to follow a power law? Assuming that there is a finite number of protein families in nature, what would be an indication that most of these families are covered by known sequences? It is likely, in the foreseeable future, that sequencing of distantly related species will give way to sequencing of closely related species, strains and different individuals within a population. This will lead to a decrease in the number of unique sequences, followed by a decrease in counts of small protein families. We aim to anticipate how this saturation will be detectable and the resulting patterns of size distribution for protein family space.

To test the distribution of protein sequence space after saturation, we modelled it on a data sample where phylogenetic proximity of sequenced organisms is high. An ideal model is provided where many closely related strains of the same species are sequenced, thus providing a dense coverage of a phylogenetic group. We used five strains of *Staphylococcus aureus* published genomes to test our hypothesis (Holden et al., 2004; Baba et al., 2002). When proteins are clustered into families by TRIBE-MCL, each genome individually produces a power-law distribution of family sizes (Fig. 2A), consistent with previous reports. However, when all five strains of *S. aureus* are considered together (Fig. 2B), the power-law line is broken, as predicted. We thus expect the collapse of the power law for size distribution at a certain critical value to be a signal of saturation in the sampling of protein families.

Genomic paralogy

We will refer to genomic paralogy as the number of protein families per genome compared to the number of genes. Some reports suggested that paralogy increases with expanding genome size (Pushker et al., 2004; Chothia et al., 2003; Muller et al., 2002). However, we recently observed a constant relationship between the number of genes and the number of families in prokaryotes using the TRIBES database and taking into account a minimal number of families per genome (Enright et al., 2003).

Growth in the number of protein families with increasing genome size is database-dependent (Fig. 3). Databases that use predefined families exhibit logarithmic growth in the number of protein families with growing genome size (Superfamily). The data from exhaustive databases (TRIBES and ProDom) fits a linear trend best, though a logarithmic trend has only a slightly worse fit to the data (TRIBES). Interestingly, COGs are again in between the exhaustive and predefined databases, with linear and logarithmic patterns fitting the data almost equally well. This results from the hybrid nature of the COGs database, which has the characteristics of both types of protein family databases, as discussed above.

Distribution of protein families across domains of life

There have been several conflicting reports about the extent to which protein families are common to all three domains of life ([Enright et al., 2003; Kyrpides et al., 1999; Tatusov et al., 1997]). Therefore, we compare percentages of protein families that are shared according to the different databases in a uniform and consistent manner. We find that the proportion of protein families reported as shared by the three domains of life is strongly dependent on the nature of the database (Table 2). Exhaustive databases derived from unsupervised clustering of pairwise similarities tend to estimate the sharing of common families as being very low (1.2% for TRIBES and 0.6% for Superfamily) compared to the number of genes.
Properties of the protein family space

Fig. 3. Genomic paralogy in prokaryotes reported by various databases. The \( x \)-axis represents the number of genes and the \( y \)-axis represents the number of families as it appears in each database. The trend lines are shown, and the formulas describe the lines and their fit to the data. For TRIBES and COGs two possible trend lines are shown: linear as a bold line with the expression in bold font for the fit at the top left corner and logarithmic as a dotted line with the expression in regular font at the bottom right corner. Pfam and Protomap are not considered as they do not have sufficiently exhaustive coverage of genome data.

Table 2. Distribution of protein families across the three domains of life in various databases

<table>
<thead>
<tr>
<th>Domains of life</th>
<th>TRIBES</th>
<th>COGs</th>
<th>ProDom</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABE</td>
<td>1.2</td>
<td>15.0</td>
<td>0.6</td>
<td>63.8</td>
</tr>
<tr>
<td>AB</td>
<td>2.0</td>
<td>22.9</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>AE</td>
<td>0.2</td>
<td>2.3</td>
<td>0.1</td>
<td>1.4</td>
</tr>
<tr>
<td>BE</td>
<td>1.5</td>
<td>5.4</td>
<td>0.7</td>
<td>23.5</td>
</tr>
<tr>
<td>A</td>
<td>13.2</td>
<td>6.7</td>
<td>7.3</td>
<td>0.2</td>
</tr>
<tr>
<td>B</td>
<td>47.6</td>
<td>41.8</td>
<td>48.8</td>
<td>2.2</td>
</tr>
<tr>
<td>E</td>
<td>34.2</td>
<td>5.8</td>
<td>41.5</td>
<td>7.6</td>
</tr>
</tbody>
</table>

The numbers show the percentage of protein families shared in all three domains of life in the shaded cell, as well as the percentages unique to each domain of life. Domain names are abbreviated as following: A for Archaea, B for Bacteria and E for Eukaryota. Pfam and Protomap are not considered (Fig. 3).

for ProDom). The Superfamily database estimates that 62.8% of superfamilies are shared among all domains of life. The estimates based on the COGs database are in between these two extremes (15.0%). This number may be an overestimate, as the COGs database does not include multicellular eukaryotes.

DISCUSSION

The debate about the structure of the protein universe, as exemplified by the differences in the properties of the protein family databases characterized here, arises from differences in sensitivity and specificity of the methods used to construct the databases. By providing wider coverage, exhaustive databases can identify families that are not detectable by approaches that use predefined families. However, since exhaustive databases typically use pairwise sequence comparison methods, these are likely to detect only smaller groups of closely related proteins, and are not designed to cluster distantly related proteins belonging to the same structural superfamily or fold. On the other hand, databases that assign proteins to predefined families use more sensitive search tools, e.g. multiple sequence comparison methods such as HMM. These are very powerful for detecting homologues of known families, but fail to identify unknown families, because they typically require a multiple sequence alignment in the first place.

The assignment of folds and superfamilies to individual genomes, rather than across a large group of genomes, follows a power-law up to family size one (Gough et al., 2001). So why is the number of unique folds so small on the global scale across all genomes? The collapse of the power-law line might indicate the saturation of sampling of structural superfamilies. Since structure-defined families are broader, we expect them to be saturated by sampling earlier than sequence-defined families. Another possible explanation for the collapse of the power law is that there is a preference for solving structures from larger families rather than unique and obscure proteins. This is reinforced by the observation that Archaea have a surprisingly small number of unique superfamilies—over ten times fewer than Bacteria, which have been characterized more extensively and are frequently of medical or industrial relevance. The Pfam database also has a bias for large families, though not as extreme as the Superfamily database. In Pfam, this could be inherent to the process of creating new families by manual curators, who are more likely to be alerted to larger families. Since both Pfam and Superfamily databases are based on HMMs, the bias against small families in genomes could also be influenced by HMMs built from few sequences. Such
HMMs are likely to represent small families, and will be less effective in detecting distant homologues.

In conclusion, we demonstrate that different results are obtained from databases built by exhaustive all-against-all comparison and predefined protein families. These results complement rather than contradict each other, describing protein diversity from different perspectives and fulfilling different user requirements.

REFERENCES


