Correspondence

Myriads of protein families, and still counting

Victor Kunin*, Ildefonso Cases*, Anton J Enright*, Victor de Lorenzo*† and Christos A Ouzounis*

Addresses: *Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK. †Centro Nacional de Biotecnología CSIC, Campus de Cantoblanco 28049 Madrid, Spain.

Correspondence: Christos A Ouzounis. E-mail: ouzounis@ebi.ac.uk

Published: 28 January 2003

Genome ***Biology*** 2003, **4**:401

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/2/401

© 2003 BioMed Central Ltd

Abstract

From the historical record of genome sequencing, we show that the rate of discovery of new families has remained constant over time, indicating that our knowledge of sequence space is far from complete.

With the advent of genome projects, the number of proteins has increased exponentially. We have analyzed the historical record of the discovery of 56,667 protein families encompassing 311,256 proteins from 83 complete genomes (available as of 28 May 2002). Our findings show that the rate of discovery of new families has remained constant over time, indicating that our knowledge of sequence space is far from complete.

A decade ago, it was proposed that there might be a limited number of protein families and folds [1]. Ever since, the expectation has been that the discovery of new proteins will eventually slow down with better sampling of protein space through genome sequencing [2,3]. With a multitude of complete genomes, it is now possible to assess the extent of this notion by examining the rate of protein family discovery.

To achieve this, we have clustered all protein sequences from all 83 complete genomes, using the TRIBE-MCL algorithm [4]. The resulting clusters represent sequence families with common functional properties and are tighter than structure-defined families or folds [4]. For each family, we recorded the first sequenced genome in which it appears for the first time (the 'founder' genome). We then counted the number of new families each genome sequence has contributed at the moment of its release.

Remarkably, the number of families is increasing steadily with each new sequenced genome (Figure 1). This result contradicts the belief that the exploration of protein space is reaching saturation. In fact, it reflects the consistent reporting of unique proteins in almost every publication of a new genome [5].

According to our data, the rate of protein family discovery continues to be constant over time (correlation coefficient with the genome sequencing order is $R^2 > 0.98$). Although the major leaps have been produced by eukaryotic genomes, which contributed a third of new protein families, diversity

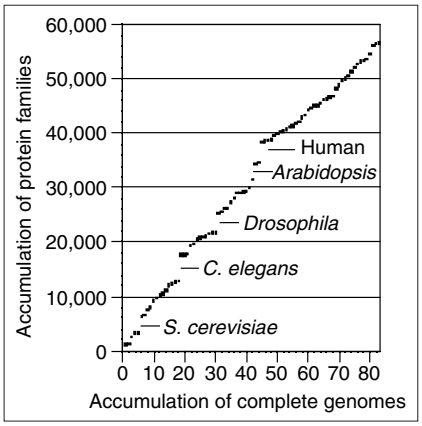

Figure 1
The number of unique protein families accumulated from genome projects. Families were obtained by clustering proteins from complete genomes with the TRIBE-MCL algorithm (inflation value 1.1). Species with the largest contributions are indicated. All data and supplementary information are available at [9].

cannot only be attributed to eukaryotes. When only the Bacteria and the Archaea are considered, the trait of a constant rate for novel families is even more pronounced (correlation coefficient $R^2 > 0.99$), suggesting that the exploration of prokaryotic diversity using genome sequencing is far from reaching completion.

What are the major contributors of novel protein families? When normalized by the number of families per genome, the phylogenetic position of the corresponding organism is crucial. The leading contributions come from *Haemophilus influenzae*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, representing the first bacterial, eukaryotic and metazoan genomes sequenced, respectively. In contrast, the smallest contributions come from multiple strains of already sequenced species.

While a thorough analysis of all these families is ongoing, some general trends can already be inferred. We have observed a great variability of family sizes, ranging from one to over 2,000 members. Although most of the largest families were already found in the very first genomes that had been sequenced, the reverse, that the earliest-found families have proved to be the largest, is not true (Figure 2). Some of the earliest families have remained very small. Of the 1,175 families founded by the first cellular genome sequenced, that of *Haemophilus influenzae*, 215 contain fewer than 10 members, seven years and 82 sequenced genomes later. On the contrary, newer families are generally smaller, with the exception of families founded by large genomes with a significant degree of paralogy, such as that of *Arabidopsis thaliana*. Thus, it is very difficult to predict how new families will develop as more complete genomes become available. In fact, family size is more related to phylogenetic distribution than to the time of the family's discovery, supporting the notion that the phylogenetic distribution of proteins ranges from universal to strain-specific [6]. While the size of universal families increases with every new genome available, taxon-specific families might grow only when a close relative to the founder genome is sequenced.

Although there are important reasons why a genome might be sequenced other than just to cover protein-sequence space [7], the contribution of new protein families can also be normalized by genome size, roughly representing the corresponding sequencing effort. From this viewpoint, the human genome has added only 1.3 new families per megabase, although this may be an underestimate given the uncertainty surrounding the total number of genes [8]. This number compares unfavorably to an average of 172 new families per megabase over all organisms, or to species such as *Xylella fastidiosa* and *Borrelia burgdorferi* with 380 new families per megabase each.

In conclusion, the constant growth rate for new protein families suggests that protein-sequence space remains largely unexplored. Sampling biological diversity through genome sequencing will continue to produce vast amounts of novel protein families with interesting biochemical properties.

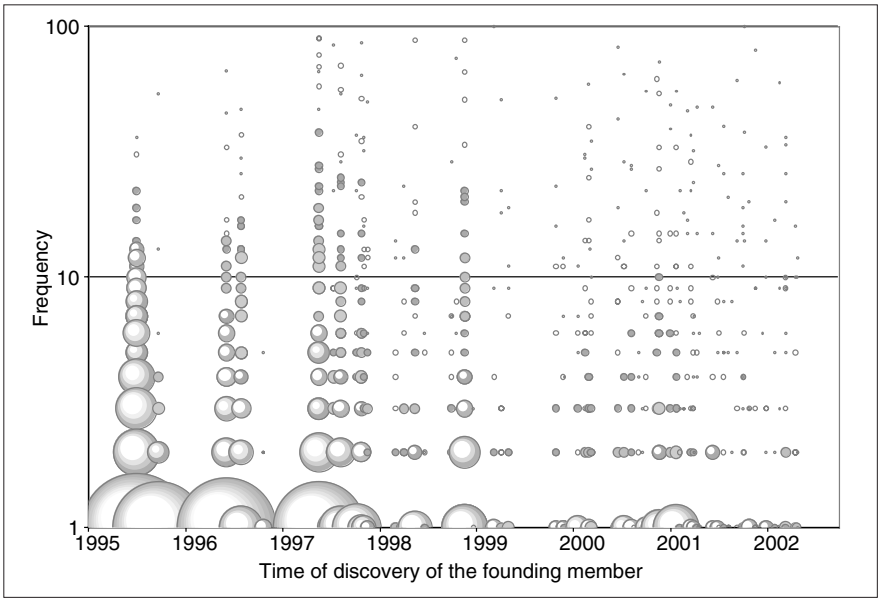

Figure 2
Size distribution of protein families in relation to the time of their discovery. The *x*-axis represents the time of discovery of the founding member of a family; the *y*-axis represents frequency (on a logarithmic scale); each circle represents the number of protein families corresponding to the value on the *y*-axis; and the area of each circle corresponds to family size. It is notable that some of the largest families were founded early, but large families are still being discovered. Recently discovered small families (upper right) are expected to grow with better sampling of protein space.

References

1. Chothia C: **One thousand families for the molecular biologist.** *Nature* 1992, **357:**543-544.
2. Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics.** *Nat Struct Biol* 2001, **8:**559-566.
3. Fischer D, Eisenberg D: **Finding families for genomic ORFans.** *Bioinformatics* 1999, **15:**759-762.
4. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30:**1575-1584.
5. Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A, Valencia A, Leroy C, Sander C, Ouzounis CA: **Genome sequences and great expectations.** *Genome Biol* 2001, **2:**interactions0001.1-0001.3.
6. Boucher Y, Nesbø CL, Doolittle WF: **Microbial genomes: dealing with diversity.** *Curr Opin Microbiol* 2001, **4:**285-299.
7. Doolittle RF: **Microbial genomes multiply.** *Nature* 2002, **416:**697-700.
8. Daly MJ: **Estimating the human gene count.** *Cell* 2002, **109:**283-284.
9. **The European Bioinformatics Institute Computational Genomics Group** [http://www.ebi.ac.uk/research/cgg/seqspace]