Structural Genomics Analysis: Phylogenetic Patterns of Unique, Shared, and Common Folds in 20 Genomes

Hedi Hegyi,

Jimmy Lin

&

Mark Gerstein

Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (203) 432-5175
Mark.Gerstein@yale.edu

(Revised version, 010326)

ABSTRACT

We carried out a structural-genomics analysis of the folds in the first 20 completely sequenced genomes, focusing on the patterns of fold usage. We assigned folds to sequences using PSI-blast, run with a systematic protocol to reduce the amount of computational overhead. On average, folds could be assigned to about a fourth of the ORFs in the genomes and about a fifth of the amino acids in the proteomes. More than 80% of all the folds in the scop structural classification were identified in one of the 20 organisms, with worm and *E. coli* having the largest number of distinct folds. Folds are particularly effective at comprehensively measuring levels of gene duplication, as they group together even very remote homologues. Using folds, we find the average level of duplication varies depending on the complexity of the organism, ranging from 2.4 in *M. genitalium* to 32 for the worm -- values significantly higher than those observed based purely on sequence similarity. We rank the common folds in the 20 organisms, finding that the top three folds are the P-loop NTP hydrolase, the ferrodoxin fold, and the TIM-barrel. We also discuss in detail the many factors that affect and bias these rankings. From the overall patterns of shared folds, we were able to group the 20 organisms into a whole-genome tree, which is similar but not identical to the classic ribosomal tree. We also focus on specific patterns of fold (and fold pair) occurrence in the genomes, associating some of them with instances of horizontal transfer and others with gene loss. In particular, we find three possible examples of transfer between archaea and bacteria and six between eukarya and bacteria. We make available our detailed results at the following URL: http://bioinfo.mbb.yale.edu/genome/20.

INTRODUCTION

Structural genomics, which combines structural biology with genomics, is emerging as a new sub-discipline. It has a central concept of mapping the whole protein structure space – i.e. determining the complete protein-fold "parts list." Estimates for the number of naturally occurring folds run somewhere between 1,000 and 10,000 [1-3] and the current structural classifications divide the known structures into ~500 known folds [4-6].

Large-scale sequence analysis of structural domains in completely sequenced microbial and eukaryotic genomes will affect both the set of proteins to be selected for experimental high-throughput structure determination and the biological conclusions we eventually draw from the massive amount of experimental work. It is timely, therefore, to perform such an analysis by comparing the sequences of the currently completed genomes to those of the already resolved and classified structural domains. Here, we survey the patterns of fold usage in the first 20 completely sequenced genomes, in the manner of a demographic census. This enables us to identify unique folds, which are potentially antibiotic targets in pathogens; shared folds, which provide information on evolutionary relatedness; common folds, which may be generic scaffolds; and overall patterns of fold usage, which may reveal aspects of protein structure and evolution beyond that found by sequence similarity. We also survey the level of gene duplication implied by the sharing of the same fold by many genes, finding that it varies greatly between genomes.

Our work follows upon previous (mostly smaller-scale) surveys of the occurrence of folds in genomes [7-11] and much work on assigning folds to genomes as comprehensively as possible [12-19].

It also relates to a number of previous analyses in more general areas of genomics. One goal of large-scale genome analysis is to study the evolution of completely sequenced organisms by deciphering their genetic makeup through identifying orthologs and paralogs in their genomes [20]. These studies also provide information about the conserved core of the genomes, which are necessary to the basic cellular functions of all bacteria, archea and eukaryotes. Another interesting aspect of evolution is the relatively high frequency with which these primitive organisms incorporate foreign genes into their genomes, i.e. horizontal gene transfer [21]. These horizontally transferred genes can represent new folds in the organism and provides a possible mechanism for an organism to acquire a new "part". Analyzing a large number of closely related genomes helps to clarify this issue with greater certainty than in the past [22]. Large-scale genome comparison has also provided a glimpse into the evolutionary process of genome degradation in parasitic microorganisms [23].

Another goal of genomics is to study biological function on a large scale in terms of the functions of many proteins. Recent success in assigning a function to a novel protein based merely on its structure (i.e. guessing what a part does from its shape) suggests that structural genomics might be useful in this endeavor. For example, Stawiski et al. identified several novel proteases based purely on their unique structural features [24], and Eisenstein et al. outlined a strategy to characterize 65 novel *H. influenzae* proteins through high-throughput crystallography [25]. In terms of functional assignment, there has been much recent progress based on comparing phylogenetic profiles of different gene products. These studies predict the function of an uncharacterized protein based on its consistent appearance with a protein of known function in the same genomes. Eisenberg and co-workers studied correlated evolution using phylogenetic profiles derived from 16 completely sequenced genomes, and used these, in addition to patterns of domain fusion, to identify functionally related proteins [26, 27]. Enright *et al.* followed a similar approach and identified several unique fusion events by comparing the complete genomes of two bacteria and an archaea [28]. Reflecting the great amount of experimental functional information available for E. coli, this organism's genome been studied in rather great detail in terms of functional prediction and structure-function relationships [29-32].

Finally, genomics is also driven by practical goals, such as the need to discover new antibiotics to treat emerging antibiotics-resistant bacteria. Genes that are conserved in several microbial genomes but are missing from eukaryotic and archaeal genomes would be ideal targets for broad-spectrum antibiotics [33]. Another approach is to identify species-specific genes with unique structures to reveal organism-specific biochemical pathways. Such genes are suspected to play a role in the pathogenicity of the bacteria [34] and could be used to develop antibiotics against specific pathogens.

Materials and Methods

Specific Databases Used in the Sequence Comparisons

Table 1A shows a list of 20 genomes we analyzed, their phylogenetic classifications, and their sizes. They represent all three domains of life (Archaea, Bacteria and Eukaryota). 19 of the 20 are single-cell organisms, and one is a eukaryote (yeast), with genome size varying from 479 (*M.genitalium*) to 6218 ORFs (yeast). The only metazoan of the twenty, *C.elegans*, has ~19000 ORFs, and the average genome size, which we denote by G below is 2179.

Insert Table 1A -- "Organisms + ORF AA Coverage"

We compared the amino acid sequences of the structural domains in the SCOP classification of protein structures [4] to the sequences of the 20 genomes. (Specifically, we used a clustered version of the scop database 1.39, called pdb95d, as queries. This contains 3266 distinct representative sequences, which we denote as P.) For the PSI-blast runs we also used a 90% non-redundant protein database NRDB90 [35] in our comparisons. The version we used is from December 1999 and contains 195,866 sequences (denoted as N). Both the databases (NRDB and the genome sequences) and the query sequences (scop domain) were masked with the SEG program using standard parameters to mask low-complexity regions [36, 37].

Fold assignment by PSI-BLAST, Development of a Fast Hybrid Protocol

One of the goals of this work was to develop a simple, robust approach for automatically using PSI-blast [38] to do fold assignments to genomes in bulk.

For all our PSI-blast runs we used an inclusion threshold (h) of 10^{-5}, a number of iterations (j) of 10, and a final match threshold of 10^{-4}. These parameters, considerably more conservative than in a number of recent analyses [14, 15, 39-41]. We used these parameters because we intended that our fold assignments run in a highly automated fashion and we wanted to guard against false positives that would not be caught by manual checking. Furthermore, while PSI-blast, with proper masking for low-complexity regions, is known to be quite robust, the iterations occasionally do go out of control with fairly liberal parameter choices (particularly the inclusion threshold h) and we wished to specifically guard against this. Moreover, since we varied the size of the databases (see below) used in a variety of the runs, we wanted to try to ensure that our parameter choices resulted in significant matches in any of the databases used. We performed our PSI-blast comparisons in a number of ways:

(i) Default Protocol

We concatenated the sequences of a genome onto NRDB and used PSI-blast to run the scop domains as queries against them. This is the "default" way to run PSI-blast. However, it has the drawback that every time one adds a new genome to the analysis, even a small one, one has to re-run each scop domain against the new genome and all of NRDB, a computationally intensive process. That is, each genome requires approximately (N+G)PK pairwise comparisons, where K is the average number of iterations required by a PSI-blast comparison. (K obviously depends on many factors, including various biases both in the target database and the query, but for rough reckoning we can estimate it at j/2 = 5.) This is a very rough number, which we plan to use below for illustrative purposes. Using the values above it comes out to ~3.2 billion (3,234,074,850).

(ii) NRDB PSI-blast Profiles

We ran each scop query against NRDB to generate a PSI-blast profile, giving us a profile for each scop fold and superfamily. Then we re-ran these against the genomes without iteration, using a match threshold of 10^{-4}. (Note that because we use very conservative choices for the inclusion threshold in building up the original PSI-blast profiles, at this stage we can confidently assume that the final match threshold of 10^{-4} is selecting truly similar sequences to our original scop domain queries.) Note also that this is potentially a much more efficient process, since when one analyzes a new genome one only need run the profiles against each genome sequence once. That is, each new genome requires GP comparisons. (There is no K factor since there is no iteration.) Plugging in the numbers above, we get ~7.1 million (7,116,614).

(iii) Intra-genome Profiles

A problem with the above approach is that often the proteins that contribute most to the PSI-blast profile for a given query are in the same organism as the query. This could result, for instance, if one is searching for a protein in a family that is highly duplicated in one organism but otherwise does not have wide phylogenetic distribution. Thus, given a new genome with a highly duplicated family, one could potentially compromise sensitivity using solely NRDB generated profiles. (This would not be a problem in the default approach since one would include the genome with NRDB in the making up the of the profiles.) To get around this, while still retaining some computational efficiency for each new genome, we tried running each scop domain query against the genome with PSI-blast. For this protocol, for each new genome, we will require GKP comparisons, which evaluates to ~36 million (35,583,070) -- of course, assuming the same value for K as above, which is only approximately true.

(iv) Hybrid Protocol

For a number of select genomes, in particular *m. genitalium*, yeast and worm, we carefully compared the matches resulting from the above three protocols. We found that for the larger genomes, such as worm, use of the intra-genome profiles (protocol iii) generated quite a few additional matches beyond those found by the straight NRDB profiles (ii). In particular, using the intra-genome protocol for the worm we found 501 extra matches that were not found by the NRDB profiles (while the NRDB profiles found 576 matches that the intra-genome protocol did not find).

Combining the matches from the NRDB profiles and the intra-genome profiles (protocols ii and iii) into a new hybrid protocol resulted in essentially the same set of matches as the default PSI-blast protocol (i). For instance, for *m. genitalium*, the hybrid protocol produced at least one match for 163 different ORFs of the 483 total ORFs, whereas the default protocol produced matches for 161 different ORFs. These numbers are very similar to the values found in other PSI-blast analyses. [14, 15, 21, 39] different ORFs. Moreover, for a new genome this was considerably more efficient than the default method, 7.1 + 3.6 vs. 3,234 million comparisons, about 75 times more comparisons using the numbers above. To make the results of the various protocols completely clear, we make available on the web sets of matches resulting from running with the three protocols. See http://bioinfo.mbb.yale.edu/genomes/20. Note also that since in our hybrid protocol we are "mixing" databases for the comparisons, the precise e-values for each comparison are not exactly comparable. This is another reason for the very conservative choices we made above for our PSI-blast thresholds.

Fold assignment by FASTA, a Benchmark

As a further benchmark comparison, we ran the scop domains directly against the genomes using fasta with a standard .01 e-value cutoff [42-44]. It is known that simple pairwise comparison with either fasta or blastp is considerably less sensitive than profile-search with PSI-blast, so we did not expect this to add substantially to the number of matches that we found. However, we elected to perform the fasta searches because for certain small compositionally biased proteins, the PSI-blast profiles may not be effective [39, 41]. Also, we felt that these would be a useful benchmark for comparison against PSI-blast. As expected, we only found a very small number of additional matches with fasta. For instance, for the worm, the combination of the PSI-blast approaches produced at least one match for 4556 ORFs of the 19099. Fasta only added in 30 additional matches to these, considerably less than 1%, and it, of course, it missed 1553 of the matches.

Tabulation in terms of Scop Folds and Superfamilies

Using the SCOP scheme we tabulated our results in terms of distinct folds and structural superfamilies. In scop, for structures to have the same fold it is necessary for them to have the same overall core topology and geometric disposition of secondary structures. In contrast, a superfamily is a subset of the fold, denoting groups of proteins that have closer structural similarity and consequently probably share an evolutionary relationship [4]. We will report our specific results here separately in terms of both scop folds and structural superfamilies; however, in the text it is awkward to constantly refer to "scop folds and structural superfamilies" so sometimes we will loosely use the term "fold" to stand for both scop fold and superfamily. For instance, we will use the terms "fold assignment" and patterns of "fold occurrence" to refer to general ideas that are equally as applicable to scop structural superfamilies as to scop folds.

RESULTS

Coverage of the Genome by Known Structures

Table 1A also lists the number of the ORFs in the 20 genomes that have at least one match with one of the scop domains, along with the ratio of these numbers and the total number of ORFs for each genome. (For a complete list of occurrences of all the folds and all the superfamilies in the 20 genomes, please see the website http://bioinfo.mbb.yale.edu/genome/20).

The ratio of at least partially matching ORFs varies between about 18% (for the Lyme-disease agent *B. burgdorferi*) and 34% (for *A. aeolicus* and *M. genitalium*). *M. genitalium* has often been used to benchmark the degree of fold assignment [10, 14, 18, 39, 45]. The numbers we list for this organism are consistent with those reported in previous analyses.

Table 1A also lists the total number of amino acids in the genome "covered" by the matches and the fraction of the proteome this corresponds to (the ratio of matched and total number of amino acids). This value is surprisingly low, only about 14% for yeast and worm. Even the 'most covered' organisms, *A. aeolicus* and *H. influenzae,* have only slightly less than a quarter of their amino acids covered by known folds, leaving much room for either improvement in the structure prediction methods or discovery of new protein structures.

Overall Level of Duplication

The last section of Table 1A shows the level of duplication for the 20 organisms both in terms of folds (dividing the total number of domain matches by the number of different folds identified in each organism) and superfamilies (matches per superfamily). The worm has by far the highest level of fold duplication (~32), with yeast coming second with a significantly lower level, followed by *M. tuberculosis* and *E.coli*, with a fold duplication level of about 7.

Not too surprisingly, the largest number of different folds is present in the worm, followed by the most-studied microorganism, *E.coli,* while yeast is ranked only third, despite its considerably larger genome size. As for the superfamilies, *E.coli* has nearly as many as the worm (303 and 304, respectively), perhaps due to (i) a systematic bias in the structural databases, (ii) gene loss in the worm, or (iii) folds in *E.coli* acquired by horizontal transfer from its host or other bacteria. However, the two organisms share only about two thirds (196) of their superfamilies (see the website for details).

Fold-class Specific Duplication

Table 1B also shows the total number of superfamilies and their average duplication level in the different structural classes for *A.fulgidus*, *E.coli*, yeast and worm -- representative organisms of archea, bacteria, single-celled eukaryotes, and metazoa. One can look at this table as a subdivision of the data in Table 1 by structural class. There are clear-cut differences among the structural classes for the four organisms. While in *E.coli* the most enriched structural class is the alpha/beta one, in the worm a reverse tendency is present, rendering the Multidomain and especially the Small proteins the most duplicated, with a striking ~ 64X duplication level in the latter class. In yeast a similar trend can be observed, although to a lesser extent. This observation comports with the fact that the majority of the Small domains appear in extracellular proteins, which are required in increasing proportions to carry out the complex intercellular functions found in metazoa.

There is a general depletion of the all-beta folds in the Archaea. As shown for *A.fulgidus*, only 18 superfamilies are represented, with an average duplication rate of 2.1 in this category, a relatively low value. A similar tendency can be observed in the other three archaeal genomes, which might indicate a lesser thermostability for the all-beta structures in general, or simply reflect a lesser presence of the all-beta fold types in the last common ancestor of these organisms.

Overall Occurrence Matrix

Figure 1A shows an overview of the "occurrence matrix", the number of folds and superfamilies occurring in the six soluble fold classes for each of the 20 genomes. Each row represents a fold, each column a genome grouped by the traditional phylogenetic tree, and each cell represents the occurrence of a particular fold in a genome. The complete matrix is available in an interactive clickable form from the website. This represents the basic data from which all our fold pattern analysis is derived.

As expected, the mixed helix and sheet classes (alpha/beta and alpha+beta) have the most universally present folds and superfamilies. The two eukaryotic genomes contain proportionately more all-alpha and all-beta folds and superfamilies than the prokaryotic ones. As previously noted, the large majority of the Small folds are present only in eukaryotes, many of them only in the metazoa worm.

Insert Fig. 2 -- "Overview"

Most Common Folds

Figure 1B shows a close-up of the occurrence matrix, focusing on the most frequently occurring folds and superfamilies. Two specific aspects are discussed here – the ranking biases and the top folds and superfamilies.

Factors Affecting the Ranking

In Figure 1B, to produce the ranking of the folds in terms of frequency of occurrence for the 20 genomes, we were faced with the task of arranging the folds in the occurrence matrix. There is no unique way of doing this and any method chosen introduces some form of bias. For instance, the simplest method would just order the table in terms of the raw number of matches to each fold, but these would strongly favor the large genomes, such as *C. elegans*, over the small ones, such as *M. genitalium*. Alternatively, one could rank the table purely in terms of the degree of phylogenetic conservation -- i.e. the more organisms in which a fold occurs, the higher it is in the table. However, here the ranking would be affected by the phylogenetic biases in the genomes chosen. There are many more bacterial (especially pathogen) genomes than eukaryotes. This means that folds prevalent in bacteria will tend to rank higher than those common in eukaryotes. We have developed a ranking scheme that balances a variety of factors and corrects for some obvious biases. Our scheme, described in detail in the caption to the figure, tries to rank folds in terms of their average frequency in the main groupings of organisms (Eukaryotes, Bacteria, and Archaea), where occurrence is defined in terms of the fraction of total domains in an organism matched by a fold. (The focus on fraction of domains instead of ORFs takes into account the fact that some organisms, particularly yeast, have considerably longer ORFs than others.)

Figure 1B also shows how the highly ranked folds are connected to specific highly ranked superfamilies. When a fold is composed of many superfamilies (e.g. the TIM barrel), it sometimes will rank highly, whereas the associated superfamilies will not. This shows how the structure of the SCOP classification itself potentially introduces a bias into the rankings. If a superfamily associated with a highly ranked fold is sufficiently different from the other members of the fold, one could potentially "split it off" and consider it as a separate fold. Doing this will decrease the ranking of the original, highly ranked fold and introduce another, lower ranking fold.

The Top-ranked Folds and Superfamilies

Based on this ranking scheme, the most abundant fold (and superfamily) in the majority of the genomes is the universally present P-loop containing NTP-hydrolase. The second-ranking Ferredoxin-fold is also present in all 20 genomes; however, its most frequently occurring superfamily, 4Fe-4S Ferredoxin, is missing from several bacterial genomes. In each of the 20 genomes, at least one of the 19 superfamilies in the Ferrodoxin fold is present, performing a large number of various functions, both enzymatic and non-enzymatic as explored in detail previously [46]. The third-ranking fold is the TIM-barrel, also breaking down into numerous different superfamilies. This explains why even the most abundant of the TIM-barrel's superfamilies, the NAD(P)-linked oxidoreductase, ranks only 9th in the superfamily rankings. Many of the most frequent folds correlated well with those identified as superfolds, i.e. folds that accommodate many distinctly different sequence families [47].

It is clear from the table that the most frequent folds and superfamilies in worm and yeast are quite different from those in the bacterial and archaeal genomes. The most abundant fold in the worm is the immunoglobulin fold, while the most abundant superfamily is the EGF/Laminin, both mostly present in extracellular, often highly repetitious proteins, providing for different functions of multicellular life.

Overall Patterns of Fold Sharing

Another interesting avenue of study follows from the phylogenetic patterns of the folds, where only the presence or absence of a particular fold (or superfamily, family, etc.) in the 20 genomes is taken into consideration, and the patterns are analyzed subsequently from several viewpoints.

Fold Tree

Clustering the genomes based on how many folds or superfamilies they share leads to a type of whole-genome tree. We have previously presented whole-genome fold trees based on structure assignments to 8 genomes [11]. The pairwise distance between two genomes used here for constructing the tree is defined in terms of the fraction of shared folds, i.e. the number of shared folds divided by the total number of folds in every pair of genomes out of the 20. Figure 2A shows the resulting tree, together with the traditional ribosomal tree in Figure 2B, which is based on the sequence similarities of small subunit ribosomal RNAs. Although the fold tree is not completely identical to the traditional tree, it correctly partitions the major kingdoms, Eukarya, Archaea, and Bacteria, and preserves many of the clusters in the tree.

```
Insert Fig. 3 -- "Fold Tree"
```

Overall Distribution of Fold Conservation

Another type of overall analysis of occurrence patterns is shown in Figure 3, which lists the number of superfamilies present in a given number of genomes in the six different structural classes. As expected, the alpha/beta structural class appears to be the most conserved, having 14 superfamilies common to all 20 genomes. What is more, there are only a few superfamilies in this class that appear only in one or two genomes (4 and 5, respectively). On the other hand, the all-beta, all-alpha and alpha+beta classes have many superfamilies that appear only in one or two genomes (values in these categories vary between 12 and 19, as shown in the Figure). The main reason for this, especially in the all-alpha and alpha+beta categories, is that there are many new superfamilies in these classes that appear in eukaryotes (yeast and worm here). In the Small class the large majority of the superfamilies (17) appear only in one of the 20 genomes, mostly in the worm.

A most interesting feature in this table is that the distribution in five of the six fold classes (with the exception of the Small class) does not have a "smooth tail" at the end. That is, by increasing the number of genomes, the number of conserved superfamilies does not continuously fall off; instead all have an increased value at 20 – highlighting the importance of the 38 superfamilies that are absolutely conserved throughout evolution, despite the large evolutionary diversity these 20 genomes represent. These superfamilies tend to have a disproportionately high presence in the genomes; on average about one third of all the matches in the 20 genomes belong to one of these 38 ‘universal’ superfamilies. (However, this number varies considerably among the different genomes; in the smallest genome, *M.genitalium,* more than half the matches occurred within one of these universal superfamilies, while in *C.elegans* only about one eighth of all the matches fall into this category.) An earlier analysis we performed [8] also indicated that many of the folds encompassing these highly conserved superfamilies tend to be superfolds [47].

Insert Fig. 4 -- “Total Sfam Distribution”

Fold Pair Sharing

Figure 1C shows a similar view to that of figure 2A and 2B, focusing on the patterns of occurrence of the pairs of folds in the genomes. A fold pair is two different folds occurring in order, in tandem. Clearly, some pairs are greatly over-represented. However, the patterns of over-representation are very different from those applicable to single folds (e.g. compare the completely different appearance of figures 1B and 1C).

Analysis of Specific Phylogenetic Patterns of Fold Occurrence

Further analysis of the overall occurrence matrix involves detailed inspection of specific patterns of fold occurrence. Some notable patterns are shown in the schematic in Figure 4. Many of these are indicative of particular evolutionary processes -- e.g. gene loss or horizontal transfer. Other patterns may indicate convergent evolution -- i.e. two folds may occur in different families of proteins that carry out the same role in different organisms but have evolved independently. Others are obvious: folds in all organisms or folds in only one. The last pattern, unique folds in certain organisms, may be useful for identifying potential drug targets. A fold present in a pathogen but not in the human genome (or in any other organism) would naturally serve as an ideal target of a highly specific drug (antibiotic or vaccine). (A detailed list of unique folds is available from the website.)

Insert Fig. 5 -- “Schematic”

The analysis that follows shows that most of these interesting fold occurrence patterns were present in the overall occurrence matrix. The only exception is a pattern of *totally* complementary folds throughout the 20 genomes. Such a pattern is less likely to be found, as folds can be transferred between related organisms. However, we found several incomplete complementary patterns and a number of examples for horizontal fold transfer.

Gene Loss

There are a number of instances where folds (or structural superfamilies) are missing only from a single organism or clade. The most notable of these are 5 superfamilies that are missing from *Rickettsia* and present in all the other genomes.

Complementary Patterns of Fold Usage: Possible Convergent Evolution

Parts A-C of Table 2 show examples of superfamilies occurring in the different superkingdoms, performing similar or identical functions. Part A shows two superfamilies, both engaged in the control of cell division. One of them, a bacterial tubulin, is present only in archaeal and bacterial genomes (also in plants), while the other one, CKS1, a cyclin-dependent kinase, occurs only in eukaryotes.

Insert Table 2 -- “Complementary Transfer Sfams”

Horizontal Transfer

It is widely recognized now that importing and reutilizing genes from foreign organisms is quite common among microbes [48, 49]. Parts D and E of Table 2 list a number of possible cases of horizontal gene transfer among the three different clades. We carefully analyzed each potential candidate by collecting all proteins in Swissprot that contain domains with the same superfamily classification, and also by running reverse BLAST searches against the nonredundant (NR) protein database with the microbial ORFs as queries. Part D of the table shows 3 possible examples of such transfer from Archaea to Bacteria, while Part E lists 6 instances from Eukaryotes to Bacteria.

DISCUSSION AND CONCLUSION

We presented here an analysis of 20 completely sequenced genomes in terms of their usage of protein folds. This occurrence analysis has been done very carefully, choosing the searching and iterating parameters in a way that provided a good balance between sensitivity and robustness. All our results are built upon a large table, which we call a fold occurrence matrix. Thus, we were able to rank folds in terms of their overall commonness and to broadly compare organisms in terms of sharing folds. We have also focused on specific patterns of fold usage: complementary patterns between two or more folds, unique folds in certain organisms (which are potential antibiotic targets), and horizontal transfer.

The comparison of 20 genomes in structural terms from all three kingdoms of life also provided a glimpse into the emergence and spread of new folds and superfamilies. As we noted previously [50], the worm has many specific superfamilies not present in yeast or bacteria. They are basically concerned with multicellular life, evident from the high proportion (~ 70 %) of worm-specific superfamilies that are secreted or partially extracellular. On the other hand, the eukaryote-specific superfamilies present only in the worm and yeast are typically engaged in signaling and eukaryotic-type replication, appearing mostly in multidomain proteins or protein complexes (see website for details).

The specific phylogenetic patterns reveal several interesting features of the evolution of folds and superfamilies. As it is apparent from Figure 3 and as has also been discovered by others, there is a conserved set of proteins and superfamilies that invariably are present in every genome studied so far. These completely conserved superfamilies are involved mostly in replication, and usually appear in large multidomain proteins. Furthermore, in spite of the small number of these 'essential' superfamilies, they amount to less than 10 percent of the total of 471 superfamilies represented in this study. However, the corresponding matches involve about one third of the total number of matching ORFs in the 20 genomes (numbers listed in Table 1). This shows that the conserved superfamilies and folds are largely over-represented in the genomes.

Another interesting point, apparent from Figure 3, is that there are so many superfamilies that appear in one particular or only a few organisms. Besides the 25 worm-specific superfamilies we explored previously [50], the unique superfamilies are available from the website at http://bioinfo.mbb.yale.edu/genome/20. Many of these are related to their specific life-style, e.g. the ones in *Synecocystis* are mostly related to photosynthesis, whereas pathogen bacteria often carry pathogenicity-related genes, such as the virally coded KP4 toxin in *C.pneumoniae* or the tetracycline repressor and a pollen allergen in *M.tuberculosis*.

Future directions

Obviously, our analysis is obviously done with an incomplete list of domains, as we do not know all the protein folds. However, our analysis foreshadows the large-scale views we will have in the future after the completion of large-scale structural genomics projects. It is worthwhile to conclude here with an enumeration of the broad types of analysis structural genomics will make possible in the future and how our work here is related to them.

(a) The complete set of protein folds will enable us to take an overall view of the occurrence of structure in nature. We will be able to see which folds occur in which organisms and which functions they are associated with. To construct the complete list of folds we will need to consider a wide variety of organisms, as it has been demonstrated that there are a number of folds specific to various phylogenetic groups.

(b) Structural genomics will much better define the actual "modules" or regions of annotation for the genome. Modules are defined by 3D structure much more precisely than by sequence patterns or motifs, and the eventual, "final" annotation of the various regions in the human genome will undoubtedly be in reference to structural modules [8].

(c) Structural genomics will let us map the whole of protein structure space and take a global, unbiased viewpoint on the physical properties of proteins. Our view of protein structure and the conditions needed for structural stability (i.e. the size of a typical fold, the degree to which salt bridges confer thermostability, etc.) is currently strongly colored by the entries in the databanks, and this in turn is determined by the collective biases of many individual investigators following various hypothesis driven trajectories (i.e. the proteins we look at are always under the "lamppost"). It has, in fact, been shown that the proteins in the databank are NOT at all representative of those in a complete genome [51, 52].

(d) Structural genomics will improve our understanding of distant evolution. Protein folds are among the most conserved elements in biology. In terms of folds, a great amount of redundancy and reuse occurs (as is evident in the duplication section above). Consequently, folds are ideal for probing distant evolutionary relationships, across which there is no sequence conservation. If one had a complete set of protein folds, one could see the degree to which distantly related organisms share the same underlying biochemical parts, even if the underlying genes no longer have any sequence identity.

(e) Structural genomics will enable us to see which proteins are truly generic scaffolds that occur over and over again in nature and can be used for many functions, and which are more specialised parts. In combination with gene expression and protein abundance studies [53, 54] we will also be able to see which protein folds are more highly expressed and make up the bulk of actual physical mass in a cell. Our analysis here in conjunction with other preliminary analyses suggests that the TIM barrel fold may be a most common and versatile protein part [46, 55].

FIGURE AND TABLE CAPTIONS

Table 1A – The 20 genomes, Coverage and Duplication

Part A gives an overview of the coverage and duplication in the 20 genomes. The first column shows the 4-letter abbreviation used throughout the paper, the second column contains the full Latin names of the organisms. The literature references for the genomes are the following: Aaeo [56], Aful [57], Bbur [58], Bsub [59], Cpne [60], Ctra [61], Cele [62], Ecol [63], Hinf [64], Hpyl [65] Mgen [66], Mja, [67], Mpne [68], Mthe, [69]; Mtub, [70]; Phor [71], Rpro [72] Scer [73], Syne [74], Tpal [75]. The third column contains the total number of ORFs in the genomes, and the fourth shows the number of ORFs that have at least one match with one of the SCOP 1.39 domains. The sixth and seventh columns show the total number of amino acids in each proteome and the number of amino acids matched by a structural domain, respectively. The fifth and eighth columns contain the percentage values of the matched ORFs and matched amino acids, respectively. (For C. elegans, we used the ORF file associated with it in the original publication, which contained 19099 ORFs [62]. Subsequently, new versions of WormPep have come out, revising this number slightly.) The ninth and tenth columns show the number of folds and the number of superfamilies, respectively, found in the 20 genomes. The eleventh column lists the total number of matches (having eliminated the overlapping matches earlier) for each genome. The twelth column shows the domain length for each organism. In the last two columns we calculated the fold and superfamily duplication levels, by dividing the total number of matches by the number of folds and superfamilies, respectively, present in that particular genome.

Table 1B – Represented Superfamilies and Their Average Distribution

Total number and average occurrence of the represented superfamilies in the six soluble fold classes for the genomes *A.fulgidus*, *E.coli*, yeast and worm. The last row contains the number of represented superfamilies in the 20 genomes for each class, the last column shows the total number of superfamilies in the four organisms and the total of 20 genomes.

Table 2 - Examples of interesting fold usage patterns: complementary clades and horizontal transfer.

The occurrence of dots indicates whether a particular superfamily was found in a particular genome. The table also lists the SCOP descriptions for the superfamilies, a Swissprot protein and its function containing the superfamily. **A**/ Complementary clades, i.e. similar or identical functions performed by different superfamilies in the different superkingdoms between bacterial/archaeal and eukaryotic genomes. **B**/ Complementary clades between bacterial and eukaryotic/archaeal genomes. **C**/ Other complementary patterns, not restricted to a particular superkingdom. **D/** Examples of horizontal gene transfer between Archaea and Bacteria. **E**/ Examples of horizontal gene transfer between Eukaryotes and Bacteria.

Figure 1 – Overall fold occurrence matrix and most frequents folds and superfamilies.

The figures show two views of the "occurrence matrix" that tabulates the number of folds and superfamilies in the six soluble fold classes for each of the 20 genomes. Each row represents a fold; each column, one of the 20 genomes; and each cell represents the occurrence of a particular fold in a genome. The order of the organisms in part A is arranged according to the ribosomal tree in Figure 2A; the order is the same in part B.

In both parts, the occurrence of dots indicates the presence or absence of superfamilies and folds. However, if the particular superfamily or fold is among the top ten occurrences within the genome, the cell shows a statistic relating to the matches of that fold in the genome. (Precisely, it shows 10*f(i,j)*, see below.) The top occurrence in each genome is shaded in black, the next four in gray, and sixth to tenth in light gray.

The ranking scheme for folds and superfamilies is as follows: For each fold *i* in genome *j*, we first calculate the fraction of domains in the genome that have this fold: *f*(*i*,*j*)= *N(i,j)* / *D(j)*, where *N(i,j)* is number of times fold *i* occurs in genome *j* and *D(j)* is the estimated total number of domains in the genome. For the latter quantity we use *A(j)*/170, where *A(j)* is the number of amino acids in the proteome of genome *j* (from Table 1), and 170 is an estimate of the average size of a structural domain in the PDB [8]. Notice how the calculation of *f(i,j)* compensates for the fact that some genomes are dramatically larger than others and that the average size of a gene (in terms of amino acids and hence possible structural domains) also differs between genomes. Next, we determine an average value of *f(i)*, the fraction matched for fold *i*, over all genomes as follows: $f(i) = ?_j\ w(j)f(i,j)$, where the weighting factor *w(j)* is 1/6 for the two eukaryote genome, 1/12 for the four archaeal genomes, and 1/42 for the 14 bacterial genomes. The weighting factor is set so that each of the three kingdoms contributes equally to the average, and the large number of bacterial genomes does not overly skew the average. Finally, the folds or superfamilies are ranked in terms of *f(i)*.

Part A of the figure shows a schematic of the whole occurrence matrix, where the folds are first broken into major classes and then ranked in terms of *f(i)*. Part B shows a close-up of the top-ranking folds and superfamilies, including all the classes. The lines connecting the folds to the corresponding superfamilies indicate how the common folds are associated with common superfamilies. The dotted horizontal lines indicate missing lines (cuts) in the big table so that top folds in specific genomes that are not within the top total ranking can be shown. Along with each fold, the fold description and a domain identifier from SCOP 1.39 [4] are given. The entire listing is available on the website (http://bioinfo.mbb.yale.edu/genome/20). Part C shows a view similar to the previous parts, now focusing on the patterns of fold-pair sharing, where a fold pair is defined as two distinct folds occurring in tandem in a protein. The numbers indicate the number of times the fold pairs occur; if it is greater than 6, it is shaded black, between 3 and 6, gray, and below 3, white. The blank spaces show instances in which one of the pairs do not occur in the organism at all. For the complete listing of all the fold-pairs, please visit http://bioinfo.mbb.yale.edu/genome/20.

Figure 2 - Trees of the 20 organisms.

A/ The traditional tree based on pairwise sequence similarities among the ribosomal RNA small subunits of the 20 organisms. B/ Fold occurrence tree: the pairwise distances were based on the fraction of shared folds between the pairs of genomes of the 20 organisms.

Figure 3 - Distribution of the occurrence of the superfamilies among the 20 genomes.

This figure with an associated data table shows the number of SCOP superfamilies that occur in a given number of genomes. The SCOP superfamilies are divided into the usual six structural classes. For instance, the value 19 in the upper left corner of the data table denotes the 19 different all-alpha superfamilies that were found to be present in exactly one genome.

Figure 4 - Schematic.

This figure illustrates a number of interesting patterns of fold usage: (i) Present/Absent. The first pair of profiles shows two patterns in which the fold is only present in one genome, while the second pair shows patterns where the fold is absent from a single organism. The graph of the abundance of folds in each organism can be used to derive more information from the two aforementioned pairs of profiles. (ii) Complementary. The top right shows complementary patterns, in which some organisms have apparently one fold/superfamily, while other organisms have another fold/superfamily in a complementary manner. This could suggest that the two different folds/superfamilies have similar functions. However, this (complete) pattern is less likely to be found, as folds are often transferred between closely (or sometimes even remotely) related organisms. Complementary patterns in which one clade of organisms has one fold, while another one has a different fold, are more likely (middle right in the schematic). (iii) Loss/Transfer. The last two schematics show possible evidence for horizontal transfer (top of the pair) and gene loss (bottom of the pair). Horizontal transfer can be observed when one clade of organisms and just one member of the other clade have the same fold. An evolutionarily most parsimonious explanation for such a pattern is that the fold has been transferred from the dominant clade to a single organism. Gene loss can be observed when most members of the clade have the fold, whereas a few organisms do not.

References

1. Chothia, C. Proteins — 1000 families for the molecular biologist. Nature 357: 543-544, 1992.
2. Govindarajan, S, Recabarren, R & Goldstein, R A. Estimating the total number of protein folds. Proteins 35: 408-414, 1999.

3. Wolf, Y I, Grishin, N V & Koonin, E V. Estimating the number of protein folds and families from complete genome data. J Mol Biol 299: 897-905, 2000.

4. Murzin, A, Brenner, S E, Hubbard, T & Chothia, C. SCOP: A Structural Classification of Proteins for the Investigation of Sequences and Structures. J. Mol. Biol. 247: 536-540, 1995.

5. Orengo, C A, Michie, A D, Jones, S, Jones, D T, Swindells, M B & Thornton, J M. CATH--a hierarchic classification of protein domain structures. Structure 5: 1093-1108, 1997.

6. Holm, L & Sander, C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 26: 316-319, 1998.

7. Gerstein, M. A Structural Census of Genomes: Comparing Eukaryotic, Bacterial and Archaeal Genomes in terms of Protein Structure. J. Mol. Biol. 274: 562-576, 1997.

8. Gerstein, M & Hegyi, H. Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 22: 277-304, 1998.

9. Frishman, D & Mewes, H-W. Protein structural classes in five complete genomes. Nature Struct. Biol. 4: 626-628, 1997.

10. Wolf, Y I, Brenner, S E, Bash, P A & Koonin, E V. Distribution of protein folds in the three superkingdoms of life. Genome Res 9: 17-26, 1999.

11. Lin, J & Gerstein, M. Whole-Genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels. Genome Research (in press), 2000.

12. Fischer, D & Eisenberg, D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc Natl Acad Sci U S A 94: 11929-11934, 1997.

13. Fischer, D & Eisenberg, D. Predicting structures for genome proteins. Curr Opin Struct Biol 9: 208-211, 1999.

14. Huynen, M, Doerks, T, Eisenhaber, F, Orengo, C, Sunyaev, S, Yuan, Y & Bork, P. Homology-based fold predictions for Mycoplasma genitalium proteins. J Mol Biol 280: 323-326, 1998.

15. Teichmann, S A, Park, J & Chothia, C. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. Proc Natl Acad Sci U S A 95: 14658-14663, 1998.

16. Sanchez, R & Sali, A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A 95: 13597-13602, 1998.

17. Dubchak, I, Muchnik, I & Kim, S H. Assignment of folds for proteins of unknown function in three microbial genomes. Microb Comp Genomics 3: 171-175, 1998.

18. Muller, A, MacCallum, R M & Sternberg, M J. Benchmarking PSI-BLAST in genome annotation. J Mol Biol 293: 1257-1271, 1999.

19. Salamov, A A, Suwa, M, Orengo, C A & Swindells, M B. Genome analysis: Assigning protein coding regions to three-dimensional structures. Protein Sci 8: 771-777, 1999.

20. Makarova, K S, Aravind, L, Galperin, M Y, Grishin, N V, Tatusov, R L, Wolf, Y I & Koonin, E V. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9: 608-628, 1999.

21. Wolf, Y I, Aravind, L & Koonin, E V. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet 15: 173-175, 1999.

22. Doolittle, R F, Feng, D F, Anderson, K L & Alberro, M R. A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol 31: 383-388, 1990.

23. Andersson, J O & Andersson, S G. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev 9: 664-671, 1999.

24. Stawiski, E W, Baucom, A E, Lohr, S C & Gregoret, L M. Predicting protein function from structure: unique structural features of proteases. Proc Natl Acad Sci U S A 97: 3954-3958, 2000.

25. Eisenstein, E, *et al.* Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. Curr Opin Biotechnol 11: 25-30, 2000.

26. Pellegrini, M, Marcotte, E M, Thompson, M J, Eisenberg, D & Yeates, T O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288, 1999.

27. Marcotte, E M, Pellegrini, M, Thompson, M J, Yeates, T O & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function [see comments]. Nature 402: 83-86, 1999.

28. Enright, A J, Iliopoulos, I, Kyrpides, N C & Ouzounis, C A. Protein interaction maps for complete genomes based on gene fusion events [see comments]. Nature 402: 86-90, 1999.

29. Zhang, L, Godzik, A, Skolnick, J & Fetrow, J S. Functional analysis of the Escherichia coli genome for members of the alpha/beta hydrolase family. Fold Des 3: 535-548, 1998.

30. Rychlewski, L, Zhang, B & Godzik, A. Functional insights from structural predictions: analysis of the Escherichia coli genome. Protein Sci 8: 614-624, 1999.

31. Fetrow, J S, Godzik, A & Skolnick, J. Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 282: 703-711, 1998.

32. Tatusov, R L, *et al.* Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Curr Biol 6: 279-291, 1996.

33. Galperin, M Y & Koonin, E V. Searching for drug targets in microbial genomes. Curr Opin Biotechnol 10: 571-578, 1999.

34. Hacker, J, Blum-Oehler, G, Muhldorfer, I & Tschape, H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23: 1089-1097, 1997.

35. Kallberg, Y & Persson, B. KIND-a non-redundant protein database. Bioinformatics 15: 260-261, 1999.

36. Wootton, J C & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. Computers and Chemistry 17: 149-163, 1993.

37. Wootton, J C & Federhen, S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266: 554-571, 1996.

38. Altschul, S F, Madden, T L, Schaffer, A A, Zhang, J, Zhang, Z, Miller, W & Lipman, D J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997.

39. Teichmann, S A, Chothia, C & Gerstein, M. Advances in structural genomics. Curr Opin Struct Biol 9: 390-399, 1999.

40. Altschul, S F & Koonin, E V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23: 444-447., 1998.

41. Park, J, Karplus, K, Barrett, C, Hughey, R, Haussler, D, Hubbard, T & Chothia, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284: 1201-1210., 1998.

42. Pearson, W R. Effective Protein Sequence Comparison. Meth. Enz. 266: 227-259, 1996.

43. Pearson, W R & Lipman, D J. Improved Tools for Biological Sequence Analysis. Proc. Natl. Acad. Sci. USA 85: 2444-2448, 1988.

44. Pearson, W R. Empirical statistical estimates for sequence similarity searches. J Mol Biol 276: 71-84, 1998.

45. Jones, D T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287: 797-815, 1999.

46. Hegyi, H & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 288: 147-164, 1999.

47. Orengo, C A, Jones, D T & Thornton, J M. Protein superfamilies and domain superfolds. Nature 372: 631-634, 1994.

48. Pennisi, E. Versatile gene uptake system found in cholera bacterium [news]. Science 280: 521-522, 1998.

49. Lake, J A, Jain, R & Rivera, M C. Mix and match in the tree of life. Science 283: 2027-2028, 1999.

50. Gerstein, M, Lin, J & Hegyi, H. Protein Folds in the Worm Genome. Pac. Symp. Biocomp. 5: 30-42, 2000.

51. Das, R & Gerstein, M. The Stability of Thermophilic Proteins: A Study Based on Comprehensive Genome Comparison. Functional & Integrative Genomics 1: 33-45, 2000.

52. Gerstein, M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des 3: 497-512, 1998.

53. DeRisi, J L, Iyer, V R & Brown, P O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686, 1997.

54. Gygi, S P, Rochon, Y, Franza, B R & Aebersold, R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720-1730, 1999.

55. Jansen, R & Gerstein, M. Analysis of the Yeast Transcriptome with Broad Structural and Functional Categories: Characterizing Highly Expressed Proteins. Nuc. Acids Res. 28: 1481-1488, 2000.

56. Deckert, G*, et al.* The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature 392: 353-358, 1998.

57. Klenk, H P*, et al.* The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon Archaeoglobus fulgidus [published erratum appears in Nature 1998 Jul 2;394(6688):101]. Nature 390: 364-370, 1997.

58. Fraser, C M*, et al.* Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi [see comments]. Nature 390: 580-586, 1997.

59. Kunst, F*, et al.* The complete genome sequence of the gram-positive bacterium Bacillus subtilis [see comments]. Nature 390: 249-256, 1997.

60. Kalman, S*, et al.* Comparative genomes of Chlamydia pneumoniae and C. trachomatis. Nat Genet 21: 385-389, 1999.

61. Read, T D*, et al.* Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. Nucleic Acids Res 28: 1397-1406, 2000.

62. Consortium, T C e S. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012-2018, 1998.

63. Blattner, F R*, et al.* The complete genome sequence of Escherichia coli K-12 [comment] [see comments]. Science 277: 1453-1474, 1997.

64. Fleischmann, R D*, et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd [see comments]. Science 269: 496-512, 1995.

65. Tomb, J F*, et al.* The complete genome sequence of the gastric pathogen Helicobacter pylori [see comments] [published erratum appears in Nature 1997 Sep 25;389(6649):412]. Nature 388: 539-547, 1997.

66. Fraser, C M*, et al.* The minimal gene complement of Mycoplasma genitalium [see comments]. Science 270: 397-403, 1995.

67. Bult, C J*, et al.* Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii [see comments]. Science 273: 1058-1073, 1996.

68. Himmelreich, R, Hilbert, H, Plagens, H, Pirkl, E, Li, B C & Herrmann, R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 24: 4420-4449, 1996.

69. Smith, D R*, et al.* Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J Bacteriol 179: 7135-7155, 1997.

70. Cole, S T*, et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence [see comments] [published erratum appears in Nature 1998 Nov 12;396(6707):190]. Nature 393: 537-544, 1998.

71. Kawarabayasi, Y*, et al.* Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (supplement). DNA Res 5: 147-155, 1998.

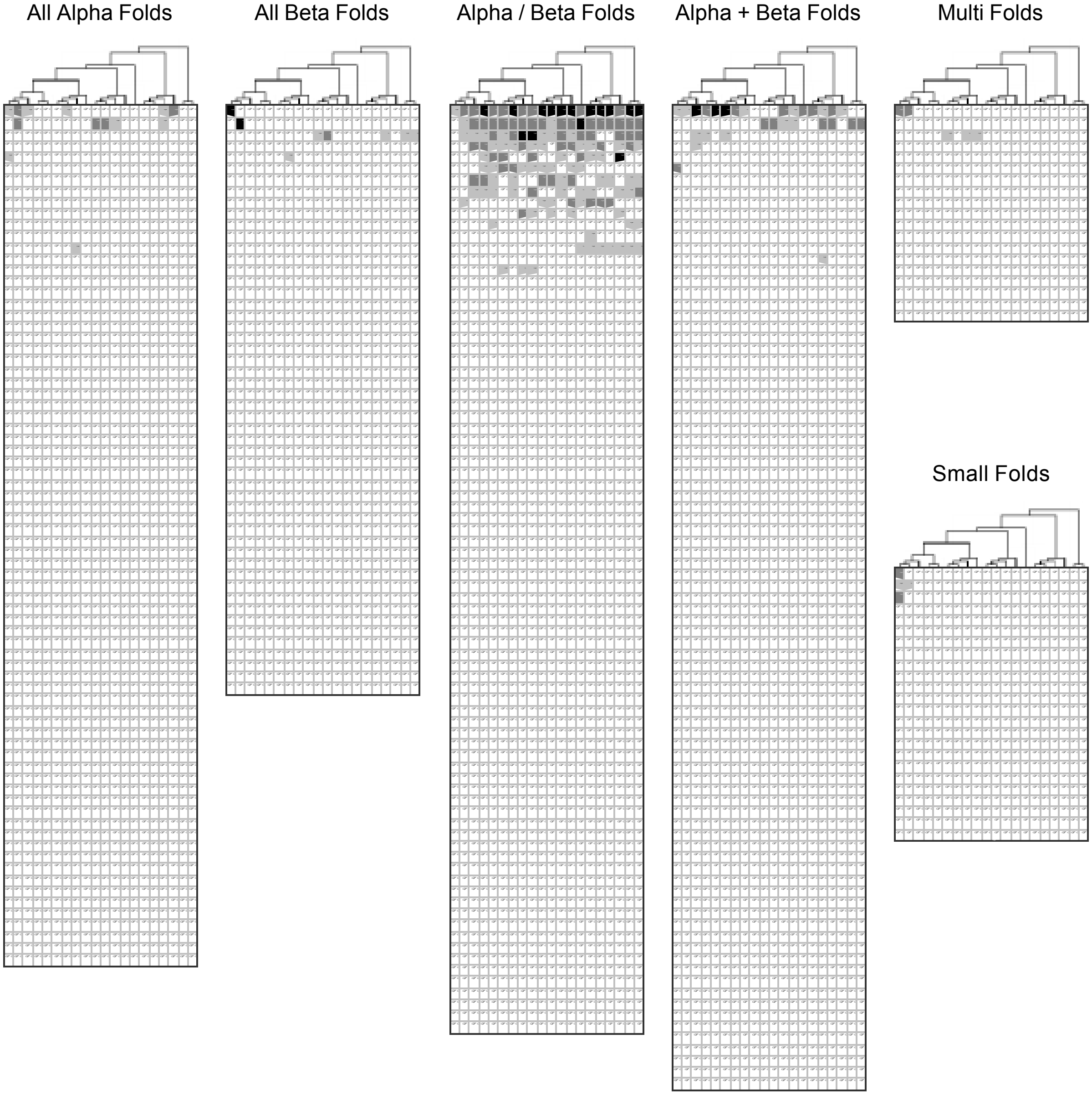
72. Andersson, S G*, et al.* The genome sequence of Rickettsia prowazekii and the origin of mitochondria [see comments]. Nature 396: 133-140, 1998.

73. Goffeau, A*, et al.* Life with 6000 genes [see comments]. Science 274: 546, 563-547, 1996.

74. Kaneko, T*, et al.* Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3: 109-136, 1996.

75. Fraser, C M*, et al.* Complete genome sequence of Treponema pallidum, the syphilis spirochete [see comments]. Science 281: 375-388, 1998.

Figure IA



All Alpha Folds

All Beta Folds

Alpha / Beta Folds

Alpha + Beta Folds

Multi Folds

Small Folds

FOLDS

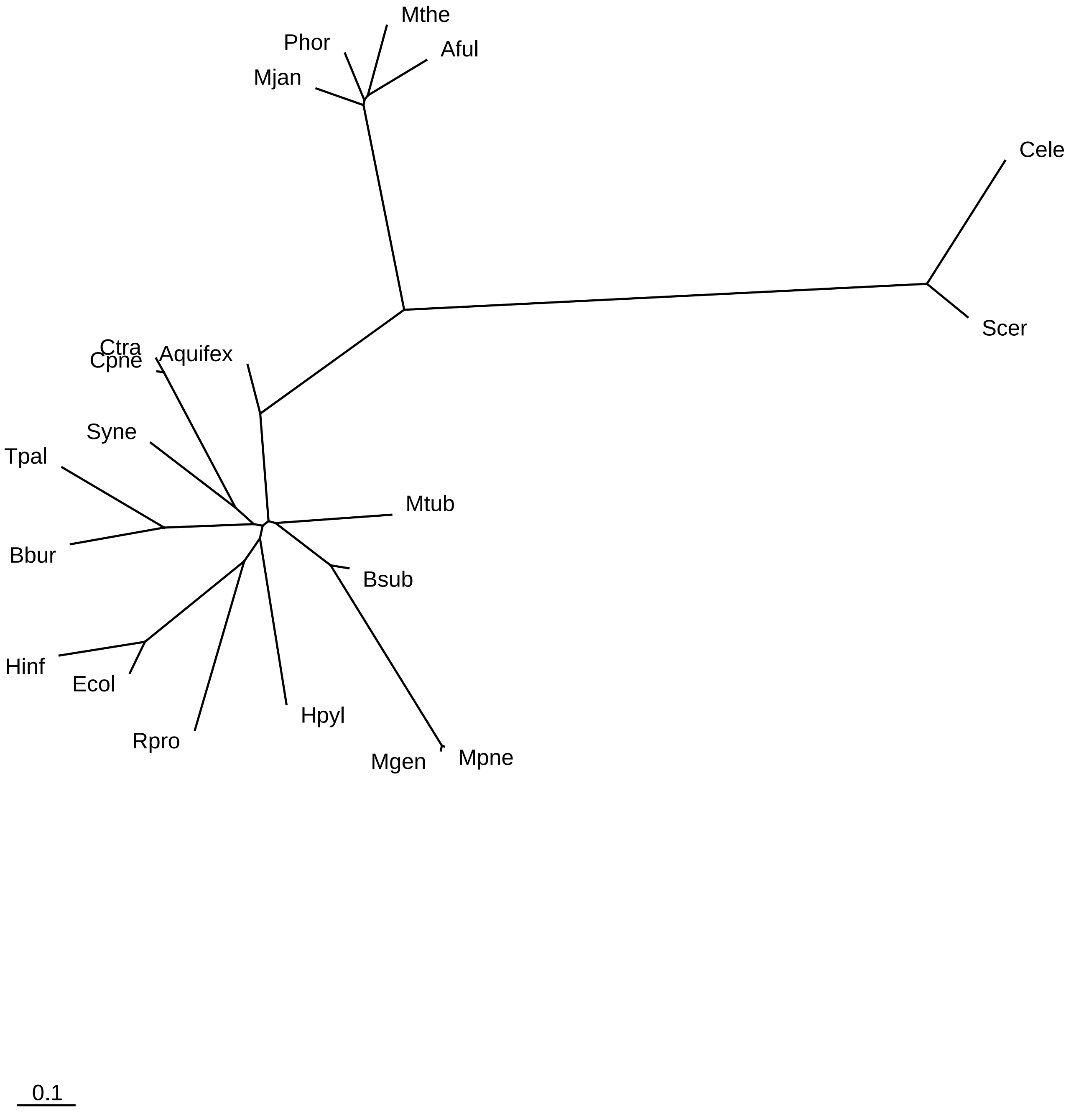
	cele	scer	mjan	phor	mthe	aful	aaeo	mtub	bsub	mpne	mgen	hpyl	rpro	ecol	hinf	bbur	tpal	syne	ctra	cpne	
P-loop cont. NTP hydrolase	5	8	21	18	17	14	20	7	9	24	33	16	23	9	17	26	21	10	20	18	3.29
Ferredoxin-like	4	7	29	6	35	24	11	6	·	·	·	7	5	10	10	2	5	6	·	·	4.34
beta/alpha (TIM)-barrel	·	5	13	9	14	12	10	10	12	6	7	7	5	12	11	8	7	8	10	9	3.1
Rossmann-fold	·	4	5	6	8	8	12	16	14	4	4	7	5	10	8	·	·	9	5	5	3.22
SAM-dep. met. transferases	·	·	12	10	5	7	8	7	4	4	·	10	5	4	6	4	·	5	5	4	3.53
Flavodoxin-like	·	·	·	4	8	8	·	·	8	·	·	5	·	8	6	3	·	11	·	·	3.14
alpha-alpha superhelix	5	7	7	·	·		9		·				·	·	·	·	5	6	·	·	1.91
FAD/NAD(P)-bndng domain	·	·	·	5	4	9	9	5	4	6	7	·	5	5	·	·	·	·	·	·	3.4
Adenine alpha hydrolase	·	·	7	6	5	·	5	·	·	6	9	5	6	·	·	4	3	·	4	4	3.17
PLP-dependent transferases	·	·	5	5	5	·	7	·	6	·	·	3	·	4	5	·	·	3	4	5	3.54
Protein kinases (PK)	10	8	·	·	·	·	·	·	·	·	·	·	·	·				·	·	·	5.1
Immunoglobulin-like	17	·	·		·	·	·	·	·			·		·		·	·	·	·		2.1
Ribonuclease H-like motif	·	5	·	·	·	·	·	·	·	4	5	·	7	6	8	5	6	·	·	·	3.47
Cl. II aaRS and biotin syn.	·	·	·	·	·	·	·	·	·	8	10	4	5	·	·	4	6	·	6	6	4.61
Acyl-CoA binding protein	·	7	·	·	·	·	·	·	·	11	13	3	·	·	·	·	3	·	·	·	1.105
alpha/beta-Hydrolases	·	·	·	·	·	·	·	9	5	·	4	·	5	·	·	·	·	4	·	·	3.56
Zincin-like	12	·		·				·	·	·	·		·		·	·		·			7.3
7-bladed beta-propeller	·	8						·					·		·			·	·	·	2.51
OB-fold	·	·	·	·	·	·	·	·	·	6	8	·	·	·	·	·	3	·	4	4	2.29
beta-Grasp	7	·																			4.105
Glucocorticoid rcptr DNA-bnd	6	·		·						·				·							7.33

SUPERFAMILIES

	cele	scer	mjan	phor	mthe	aful	aaeo	mtub	bsub	mpne	mgen	hpyl	rpro	ecol	hinf	bbur	tpal	syne	ctra	cpne		
3.29.1	5	8	21	18	17	14	20	7	9	24	33	16	23	9	17	26	21	10	20	18	d1gky__	P-loop containing NTP hydrolases
4.34.1	·	·	25	4	31	20	6	·	·			3	·	8	6		·	·			d1fxd__	4Fe-4S ferredoxins
3.22.1	·	4	5	6	8	8	12	16	14	4	4	7	5	10	8	2	2	9	5	5	d1xel__	NAD(P)-binding Rossmann-fold
3.53.1	·	·	12	10	5	7	8	7	4	4	·	10	5	4	6	4	2	5	5	4	d1vid__	SAM-dependent methyltransferases
3.4.1	·	·	·	5	4	9	9	5	4	6	7	·	5	5	·	·	3	3	3	·	d1grh__	FAD/NAD(P)-binding domain
3.54.1	·	·	5	5	5	5	7	·	6	·	·	3	·	4	5	·	·	3	4	5	d1map__	PLP-dependent transferases
5.1.1	10	8	·	·	·	·	·	·	·	·	·	·	·	·				·	·	·	d1hcl__	Protein kinases (PK), catalytic core
4.61.1	·	·	·	·	·	·	5	·	·	8	10	4	5	·	5	4	6	·	6	6	ds051__	Class II aaRS and biotin synthetases
3.1.5	·	·	5	3	7	·	·	·	4	·	·	·	·	·	·	2	2	·	·	·	d1ads__	NAD(P)-linked oxidoreductase
3.56.1	4	·	·	·	·	·	·	9	5	3	4	·	5	·	·	·	·	4	3	·	d1ax9__	alpha/beta-Hydrolases
1.105.4	·	6	·	·	·	·	·		·	11	12	3		·	·	·	·	·	·	3	d2tmaa_	Tropomyosin
3.47.1	·	4			·	·	·	·	·	3	4	·	4	3	5	4	5	·	·	·	d1ap8__	Translation initiation factor eIF4e
3.17.2	·	·	6	4	4	·	·	·	·	·	4	·	·	·	·	·	·	·	·	·	ds035__	adenine nucleotide alpha hydrolases
2.51.3	·	8						·					·		·			·	·	·	ds029__	Trp-Asp repeat (WD-repeat)
4.89.1	·	·	4	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	d1gsa_2	Glutathione synthetase ATP-binding
7.3.9	11								·	·						·					d1apo__	EGF/Laminin
3.82.1	·	·	·	4	·	·	·	·	·	·	·	·	·	4	5	4	3	4	4	4	d1rkm__	Periplasmic binding protein-like II
2.1.1	9	·	·		·	·		·								·		·	·		d1cd8__	Immunoglobulin
1.91.8	·	·	7		·		9									·		5			d1a17__	Tetratricopeptide repeat
2.29.4	·	·	·	·	·	·	·	·	·	5	6	2	4	·	·	·	3	·	4	3	ds025__	Nucleic acid-binding proteins
5.19.1	·	·			·	7	·	6	5			·	·	·	·	·	·	·	·	·	d1lci__	Firefly luciferase-like
3.14.2		·		·	4	·	·	·	5			3	·	5	·	3	·	9	·	·	d2che__	CheY-like
3.83.1	·	·	·	·	·	8	·	6	4			·	·	·	·	·		·	·	·	d1afwa1	Thiolase
4.34.7	·	5										·					·	·		·	d1fht__	RNA-binding domain
1.91.3	4	·											·	·		·	5	·			d1awcb_	Ankyrin repeat
4.105.1	7	·																			d1lit__	C-type lectin-like
7.33.1	6	·		·						·				·							d1gdc__	Glucocorticoid receptor DNA-binding

FOLD PAIRS

fold 1	fold 2	aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro
3.29	2.32	4	3	4	3	12	14	6	7	8	4	6	7	5	3	3	4	5	3	4	4
2.29	4.61	1	1	1	2	6	3	2	4	5	4	4	3	3	3	4	1	2	3	3	2
4.1	4.34	1	1	1	1	5	3	1	3	1	1	2	1	1	1	1	2	2	1	1	1
1.28	3.29	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
3.4	4.48	4	1	1	2	3	4	1	2	4	3	5	2		2	2	1	1	1	1	2
3.22	4.42	1	1	1	0	4	5	3	4	5	4	4	3	4	1	1	2	2	3	3	1
2.32	4.1	1	1	1	1	4	2	1	3	1	1	2	1	1	1	1	2	2		1	1
2.32	2.33	2	1	1	1	1	2	2	1	2	1		2	1	1	1	1	1	1	1	1
4.32	3.1	1	1	1	2	5	1	1	1	4	5	1	1	1	1	1	1	1	1	1	0
3.23	4.89	3	3	3	0	9	10	6	5	6	8	7	2	4	0	0	1	1	2	2	2
3.47	5.17	0	0	1	0	12	10	1	3	3	1	1	2	1	1	1	2	1	1	1	2
4.72	5.13	1	0	0	0	1	3	1	1	2	2	1	2	1	2	2	2	1	2	2	2
3.22	4.1	1	1	1	0	3	3	2	1	1	2	1	1	0	1	1	1	0	1	1	1
3.5	3.1	1	1	2	1			1	1	5	1		1	1	1	1	1	1	1	1	1
4.61	3.42	2	2	2	2	2	1	1	1	1	1	1	1	0	2	2	1	1	1	0	0
1.76	3.3	1	0	1	1	2	1	1	1	1	2	1	1	1	0	0	0	1	1	1	1
4.29	4.1	1		1	1	2	1	1	1	1	1		1		0	1	1	1	1	1	1
2.32	4.34					2	1	1	2	1	2	2	1	2	1	1	1	1	1	1	1
3.22	1.79	1	1	1	0	3	1	2	2	2	4	3	2	1	0	0	1	1	1	1	0
3.52	2.34	0	0	0	0	0	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1


Mthe
Phor
Aful
Mjan
Cele
Scer
Ctra
Cpne
Aquifex
Syne
Tpal
Mtub
Bbur
Bsub
Hinf
Ecol
Rpro
Hpyl
Mgen
Mpne
0.1

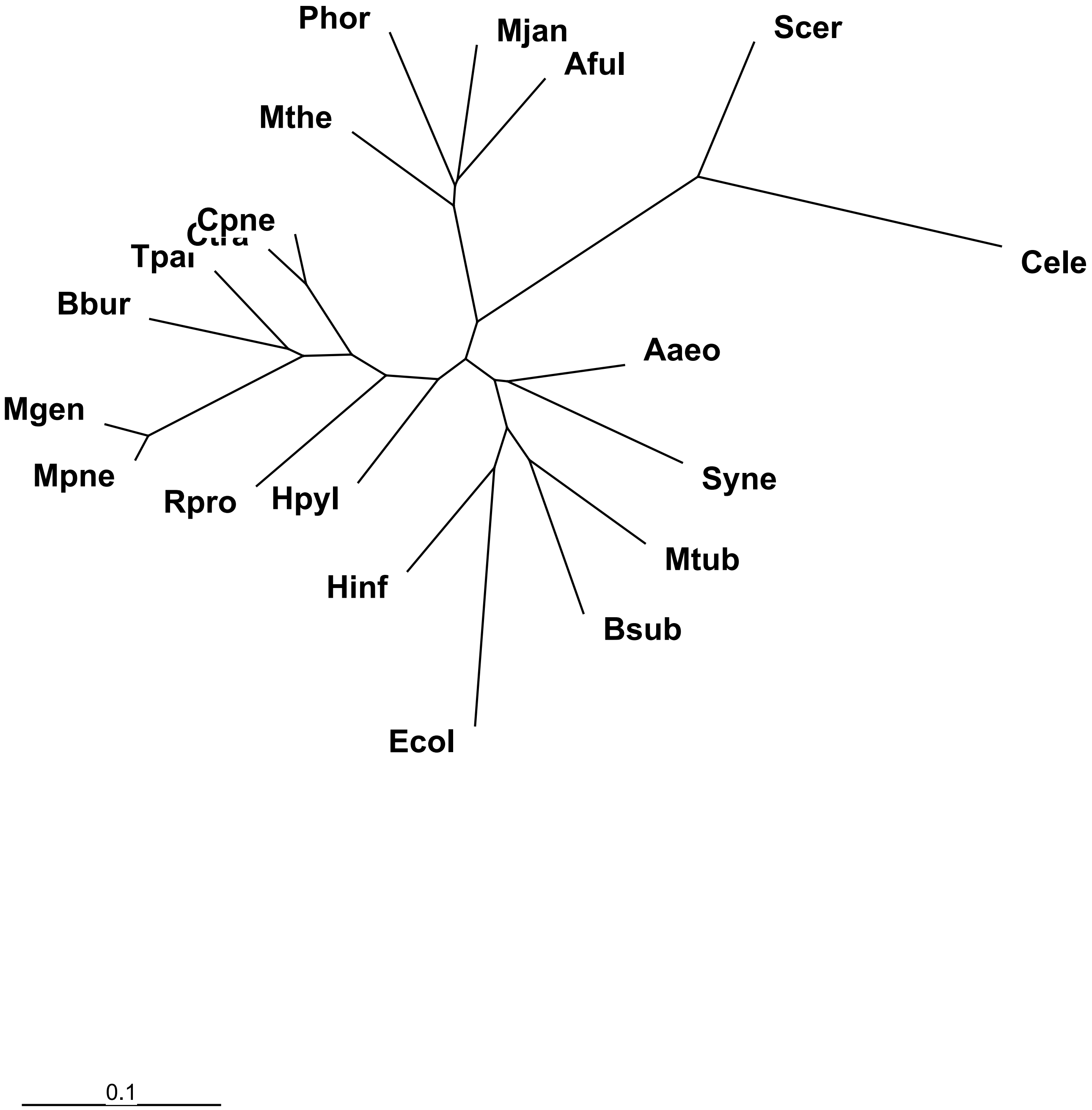

Phor
Mjan
Aful
Scer
Mthe
Cpne
Ctra
Tpal
Bbur
Cele
Aaeo
Mgen
Mpne
Syne
Rpro
Hpyl
Mtub
Hinf
Bsub
Ecol
0.1

Figure 3 -- Conservation of the superfamilies in the 6 structural classes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	cele	scer	mjan	phor	mthe	aful	aaeo	mtub	bsub	mpne	mgen	hpyl	rpro	ecol	hinf	bbur	tpal	syne	ctra	cpne
A	19	16	8	6	3	6	3	3	1	3	3	3	3	4	0	5	2	3	1	5
B	22	13	8	2	4	3	1	7	1	5	2	1	2	1	3	1	3	0	0	4
A/B	4	6	7	4	2	4	3	2	1	5	5	15	5	8	7	6	4	5	8	16
A+B	14	14	10	6	11	3	0	1	7	6	2	7	2	7	4	6	5	2	3	10
M	1	2	0	1	1	0	1	0	0	2	0	0	1	1	0	1	4	1	0	3
S	15	10	1	3	2	1	0	1	0	0	2	0	0	0	0	0	0	0	0	0

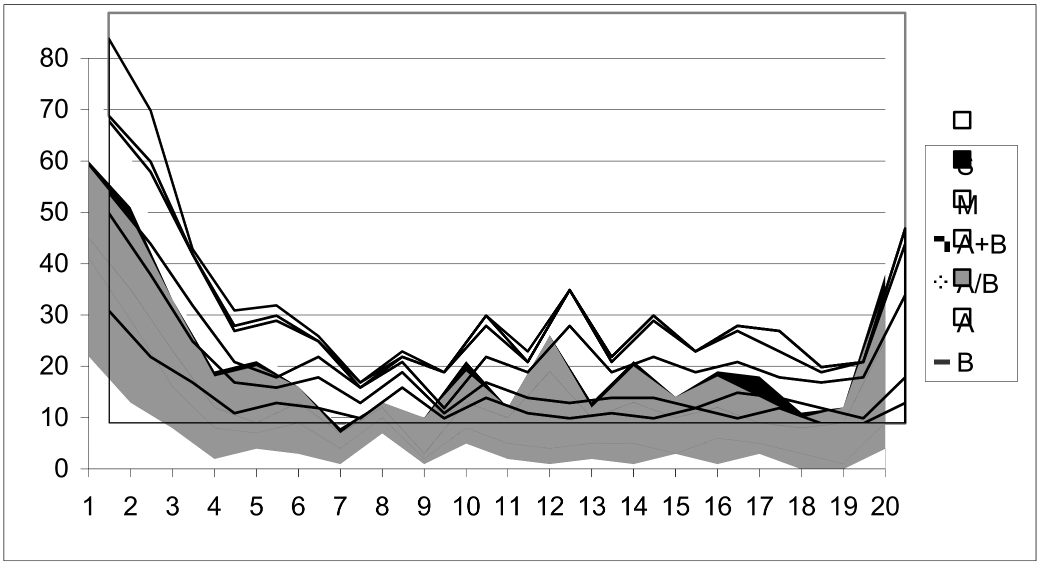


Figure 4 - Schematic of the different fold patterns

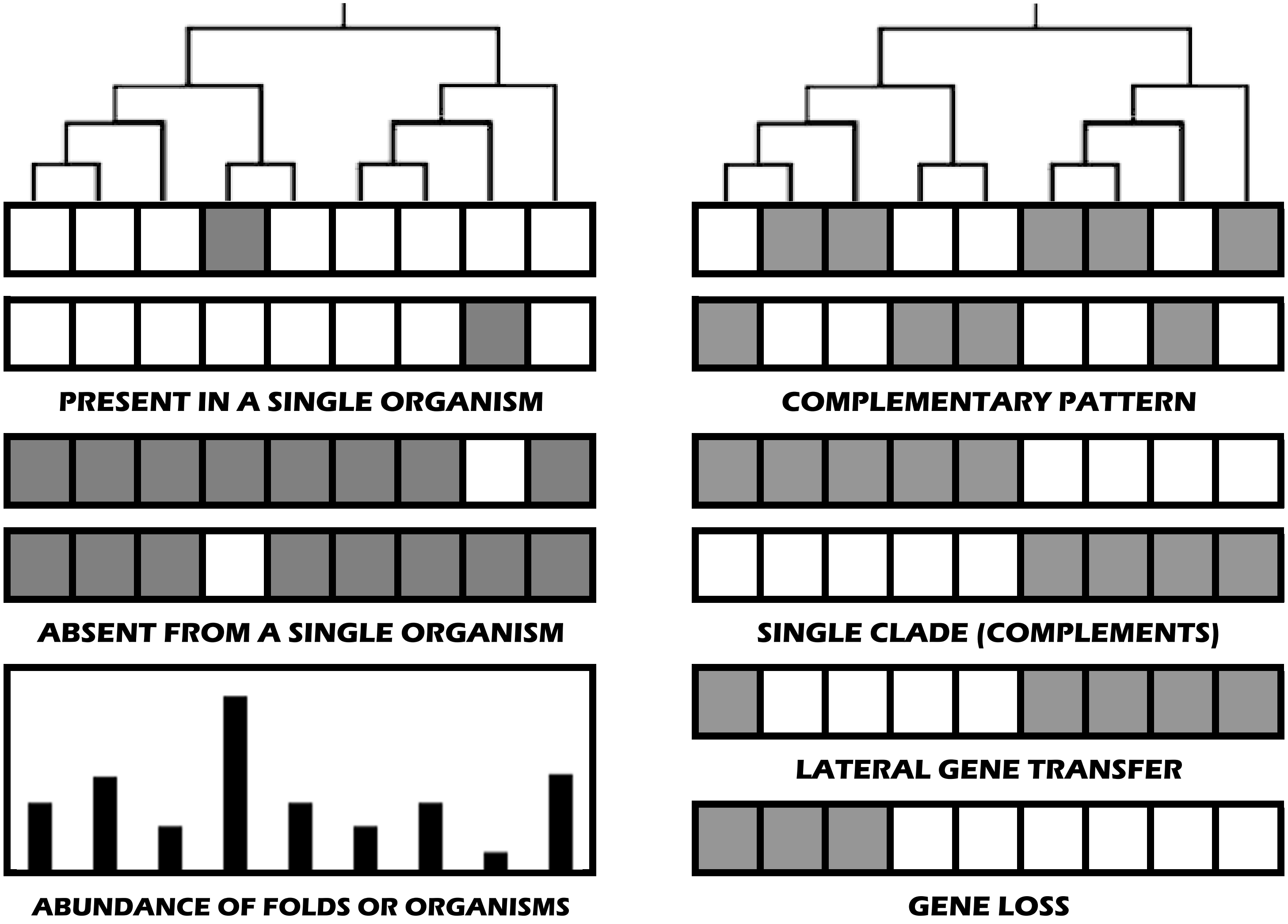

PRESENT IN A SINGLE ORGANISM
COMPLEMENTARY PATTERN
ABSENT FROM A SINGLE ORGANISM
SINGLE CLADE (COMPLEMENTS)
LATERAL GENE TRANSFER
ABUNDANCE OF FOLDS OR ORGANISMS
GENE LOSS

Table I / A

abbrev.	Species Name	ORF Coverage			Amino Acid Coverage			Domain Matches			Domain	Duplication	
		Total	Matching	m/t	Total	Matching	m/t	Folds	Sfam	Dom	Length	Fold	Sfam
Aaeo	*Aquifex aeolicus*	1522	527	34.6%	482512	116664	24.2%	162	205	690	169.1	4.26	3.37
Aful	*Archaeoglobus fulgidus*	2409	650	27.0%	663320	146655	22.1%	147	186	849	172.7	5.78	4.56
Bbur	*Borrelia burgdorferi*	1638	289	17.6%	432219	65816	15.2%	126	151	369	178.4	2.93	2.44
Bsub	*Bacillus subtilis*	4100	1121	27.3%	1217000	276596	22.7%	208	276	1460	189.4	7.02	5.29
Cele	*Caenorhabditis elegans*	19099	4586	24.0%	8096713	1136801	14.0%	247	304	7803	145.7	31.59	25.67
Cpne	*Chlamydia pneumoniae*	1052	274	26.0%	361694	66160	18.3%	136	165	367	180.3	2.70	2.22
Ctra	*Chlamydia trachomatis*	894	259	29.0%	312553	60295	19.3%	134	163	348	173.3	2.60	2.13
Ecol	*Echerischia coli*	4290	1191	27.8%	1363501	296762	21.8%	229	303	1611	184.2	7.03	5.32
Hinf	*Haemophilus influenzae Rd*	1707	528	30.9%	520930	125776	24.1%	190	243	710	177.1	3.74	2.92
Hpyl	*Helicobacter pylori*	1577	381	24.2%	500616	89025	17.8%	152	193	495	179.8	3.26	2.56
Mthe	*Methanobacterium thermoautotrophicum*	479	164	34.2%	174566	39680	22.7%	95	111	228	174.0	2.40	2.05
Mjan	*Methanococcus jannaschii*	1771	470	26.5%	501793	93299	18.6%	128	164	613	152.2	4.79	3.74
Mtub	*Mycobacterium tuberculosis*	677	178	26.3%	237651	43222	18.2%	101	118	251	172.2	2.49	2.13
Mgen	*Mycoplasma genitalium*	1871	522	27.9%	526205	105553	20.1%	135	179	675	156.4	5.00	3.77
Mpne	*Mycoplasma pneumoniae*	3924	1198	30.5%	1335687	291496	21.8%	199	253	1587	183.7	7.97	6.27
Phor	*Pyrococcus horikoshii*	2064	461	22.3%	568544	97276	17.1%	121	155	555	175.3	4.59	3.58
Rpro	*Rickettsia prowazekii*	837	264	31.5%	280233	60285	21.5%	135	160	350	172.2	2.59	2.19
Scer	*Saccharomyces cerevisiae*	6218	1699	27.3%	2906890	434481	14.9%	215	273	2346	185.2	10.91	8.59
Syne	*Synechocystis sp.*	3168	882	27.8%	1119717	196041	17.5%	199	255	1131	173.3	5.68	4.44
Tpal	*Treponema pallidum*	1031	252	24.4%	350676	58542	16.7%	123	150	346	169.2	2.81	2.31

Table I / B: Represented Superfamilies and Their Average Duplication

Levels in the Soluble Fold Classes in A.fulgidus, E.coli, Yeast, Worm and the Total of the 20 Genomes

	All-alpha			**All-beta**			**Alpha/Beta**			**Alpha+Beta**			**Multidomain**			**Small**			**Total**
	Sfams	Dup	Copy#	Sfams	Dup	Copy#	Sfams	Dup	Copy#	Sfams	Dup	Copy#	Sfams	Dup	Copy#	Sfams	Dup	Copy#	Sfams
aful	**29**	**2.5**	**73**	**18**	**2.1**	**37**	**74**	**6.1**	**453**	**49**	**4.2**	**207**	**12**	**5.7**	**68**	**4**	**2.8**	**11**	**186**
ecol	**55**	**2.9**	**159**	**44**	**4.1**	**181**	**105**	**8.3**	**872**	**78**	**3.9**	**313**	**16**	**5.0**	**81**	**5**	**1.0**	**5**	**303**
scer	**56**	**7.9**	**448**	**35**	**9.5**	**333**	**88**	**9.3**	**823**	**72**	**5.2**	**351**	**14**	**14.6**	**204**	**13**	**14.3**	**187**	**273**
cele	**62**	**20.3**	**1319**	**52**	**27.8**	**1633**	**81**	**18.1**	**1482**	**72**	**15.7**	**1140**	**14**	**42.6**	**598**	**23**	**63.7**	**1631**	**304**
20	**97**		**3197**	**83**		**3069**	**117**		**8976**	**120**		**4046**	**19**		**1598**	**35**		**1898**	**471**

Table II: Patterns of Complementary Superfamilies and Horizontal Transfer

A/ Complementary clades between bacterial/archaeal and eukaryotic genes

B/ Complementary clades between bacterial and eukaryotic/archaeal genomes

C/ Other complementary patterns

D/ Horizontal Transfer between Archaea and Bacteria

E/ Horizontal Transfer between Eukaryotes and Bacteria

	aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro	Sfam	domain	SCOP Function	Swissprot	Swissprot Function
A	•	•	•	•			•	•	•	•	•	•	•	•	•	•	•			•	3.25.1	**d1fsz_1**	Tubulin, GTPase domain	**FTSZ_ECOLI**	CELL DIVISION PROTEIN FTSZ
					•	•															4.57.1	**d1puc__**	Cell cycle regulatory proteins	**CKS1_YEAST**	CELL DIVISION CONTROL PROTEIN CKS1
B	•	•	•	•	•	•															1.22.1	**d1tafb_**	Histone-fold	**T2D5_YEAST**	TRANSCRIPTION INITIATION FACTOR TFIID
	•	•	•	•	•	•															1.63.1	**d1kxu_2**	Cyclin-like	**TF2B_RAT**	TRANSCRIPTION INITIATION FACTOR IIB
	•	•	•	•	•	•															7.35.3	**d1qyp__**	Rubredoxin-like transcriptional factor domain	**TFS2_YEAST**	TRANSCRIPTION ELONGATION FACTOR S-II
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	1.34.1	**d1coo__**	C' domain of RNA polymerase alpha subunit	**RPOA_TREPA**	DNA-DIRECTED RNA POLYMERASE
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	1.75.1	**d1gln_1**	Glu-tRNA synthetase AC-binding domain	**SYE_BACSU**	GLUTAMYL-TRNA SYNTHETASE (EC 6.1.1.17)
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	1.88.1	**d1sig__**	RNA polymerase, sigma70 subunit	**RPOS_ECOLI**	RNA POLYMERASE SIGMA FACTOR
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	4.104.1	**d2def__**	Peptide deformylase catalytic core	**DEF_HAEIN**	POLYPEPTIDE DEFORMYLASE (EC 3.5.1.31)
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	4.11.7	**d1tif__**	Translation initiation factor, N' domain	**IF3_BORBU**	TRANSLATION INITIATION FACTOR IF-3
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	4.28.1	**d2reb_2**	RecA protein, C-terminal domain	**RECA_HAEIN**	RECA PROTEIN (RECOMBINASE)
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	4.36.1	**d1ife__**	Translation initiation factor IF3	**IF3_BORBU**	TRANSLATION INITIATION FACTOR IF-3
							•	•	•	•	•	•	•	•	•	•	•	•	•	•	4.88.1	**d1div__**	Ribosomal protein L9	**RL9_BACSU**	50S RIBOSOMAL PROTEIN L9
C								•		•											4.40.1	**d2chsa_**	Chorismate mutase	**CHMU_BACSU**	CHORISMATE MUTASE (EC 5.4.99.5)
					•				•			•									1.101.1	**d5csma_**	Chorismate mutase II	**CHMU_YEAST**	CHORISMATE MUTASE (EC 5.4.99.5)
		•					•	•	•	•	•									•	1.81.1	**d1cem__**	Glycosyltransferases of the superhelical fold	**GUN6_DICDI**	ENDOGLUCANASE (EC 3.2.1.4)
		•		•		•		•		•	•	•									2.21.1	**d1yna__**	ConA-like lectins/glucanases	**GUN1_TRIRE**	ENDOGLUCANASE EG-1 (EC 3.2.1.4)
			•	•		•	•	•	•	•	•	•				•		•	•		3.1.1	**d1edt__**	Glycosyltransferases	**GUNB_NEOPA**	ENDOGLUCANASE B (EC 3.2.1.4)
						•	•	•	•	•		•	•			•	•			•	4.2.1	**d153l__**	Lysozyme-like	**CHIT_SOLTU**	ENDOCHITINASE PRECURSOR (EC 3.2.1.14)
	•	•	•		•	•	•	•	•	•	•	•	•					•	•	•	2.65.2	**d1hcz_2**	Rudiment single hybrid motif	**PYC_PICPA**	PYRUVATE CARBOXYLASE (EC 6.4.1.1)
									•	•		•		•	•	•					2.65.3	**d1f3z__**	Duplicated hybrid motif	**PTGA_BACSU**	PTS SYSTEM, GLUCOSE-SPECIFIC IIABC COMP

	aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro	Sfam	domain	SCOP Function	Swissprot	Swissprot Function
D	•	•		•					•												1.86.1	**d1aora1**	Aldehyde FerOR C' domain	**AOR_PYRFU**	ALDEHYDE:FERREDOXIN OXIDOREDUCTASE
	•	•		•					•												4.94.1	**d1aora2**	Aldehyde FerOR N' domain	**AOR_PYRFU**	ALDEHYDE:FERREDOXIN OXIDOREDUCTASE
	•	•		•				•		•											3.1.10	**d5ruba1**	RuBisCo, C' domain	**RBL_NITVU**	RUBISCO LARGE SUBUNIT
E					•				•			•									1.101.1	**d5csma_**	Chorismate mutase II	**CHMU_ARATH**	CHORISMATE MUTASE (EC 5.4.99.5)
					•	•		•					•								1.37.1	**d1rec__**	EF-hand	**TPC2_DROME**	TROPONIN C
						•		•													2.1.5	**d1suh__**	Cadherin	**CAD5_HUMAN**	VASCULAR ENDOTHELIAL-CADHERIN
						•			•					•							2.45.1	**d1eal__**	Lipocalins	**PGHD_HUMAN**	PROSTAGLANDIN-H2
					•	•			•												3.7.1	**d2bnh__**	Leucine-rich repeats	**RINI_PIG**	RIBONUCLEASE INHIBITOR
					•	•					•										4.70.1	**d1axx__**	Cytochrome b5	**NIA1_MAIZE**	NITRATE REDUCTASE (EC 1.6.6.1)
						•													•		4.112.1	**d1toh__**	Tyrosine hydroxylase	**TY3H_HUMAN**	TYROSINE 3-HYDROXYLASE (EC 1.14.16.2)