PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information

Jiang Qian, Brad Stenger, Cyrus A. Wilson, Jimmy Lin, Ronald Jansen, Sarah A. Teichmann[1], Jong Park[2], Werner Krebs, Haiyuan Yu, Vadim Alexandrov, Nathaniel Echols, Mark Gerstein*

Department of Molecular Biophysics and Biochemistry
Yale University
PO Box 208114, New Haven, CT 06520, USA

[1]Department Biochemistry & Molecular Biology, University College London, Darwin Bldg, Gower St, London WC1E 6BT, UK and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

*To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: Mark.Gerstein@yale.edu

Revised version sent to Nuc. Acids. Res. 23 Feb. 2001 .

Abstract

As the number of protein folds is quite limited, a mode of analysis that will be increasingly common in the future, especially with the advent of structural genomics, is to survey and re-survey the finite parts list of folds from an expanding number of perspectives. We have developed a new resource, called PartsList, that lets one dynamically perform these comparative fold surveys. It is available on the web at bioinfo.mbb.yale.edu/partslist and www.partslist.org. The system is based on the existing fold classifications and functions as a form of companion annotation for them, providing "global views" of many already completed fold surveys. The central idea in the system is that of comparison through ranking; PartsList will rank the ~420 folds based on more than 180 attributes. These include: (i) occurrence in a number of completely sequenced genomes (e.g. it will show the most common folds in the worm vs. yeast); (ii) occurrence in the structure databank (e.g. most common folds in the PDB); (iii) both absolute and relative gene expression information (e.g. most changing folds in expression over the cell cycle); (iv) protein-protein interactions, based on experimental data in yeast and comprehensive PDB surveys (e.g. most interacting fold); (v) sensitivity to inserted transposons; (vi) the number of functions associated with the fold (e.g. most multi-functional folds); (vii) amino acid composition (e.g. most Cys-rich folds); (viii) protein motions (e.g. most mobile folds); and (ix) the level of similarity based on a comprehensive set of structural alignments (e.g. most structurally variable folds). The integration of whole-genome expression and protein-protein interaction data with structural information is a particularly novel feature of our system. We provide three ways of visualizing the rankings: a profiler emphasizing the progression of high and low ranks across many pre-selected attributes, a dynamic comparer for custom comparisons, and a numerical rankings correlator. These allow one to directly compare very different attributes of a fold (e.g. expression level, genome occurrence, and maximum motion) in the uniform numerical format of ranks. This uniform framework, in turn, highlights the way that the frequency of many of the attributes falls off with approximate power-law behavior (i.e. according to V^{-b}, for attribute value V and constant exponent b), with a few folds having large values and most having small values.

Introduction

Protein folds can be considered the most basic molecular parts. There are a very limited number of them in biology. Currently, about 500 are known, and it is believed that there may be no more than a few thousand in total (1-3). This number is considerably less than the number of genes in complex, multicellular organisms (>10,000 for multicellular organisms (4)). Consequently, folds provide a valuable way of simplifying and making manageable complex genomic information. In addition, folds are useful for studying the relationships between evolutionarily distant organisms since, in making comparisons, structure is more conserved than sequence or function.

In a general sense, how should one approach the analysis of molecular parts? A simple analogy to mechanical parts may be useful in this regard. Given the "parts" from a number of devices (e.g. a car, a bicycle, and a plane) one might like to know which ones are shared by all and which are unique (say, wings for a plane). Furthermore, one might want to know which are common, generic parts and which are more specialized. Finally, one might like to organize the parts by a number of standardized attributes (e.g. the most flexible parts, the parts with the most functions, and the biggest parts). PartsList aims to provide answers to simple questions such as these for the domain of protein folds.

Properties related to protein folds can be divided into those that are "intrinsic" versus "extrinsic". Intrinsic information concerns an individual fold itself -- e.g. its sequence, 3D structure, and function -- while "extrinsic" information relates to a fold in the context of all other folds -- e.g. its occurrence in many genomes and expression level *in relation* to that for other folds. Web-based search tools already provide intrinsic information about protein structures in the form of reports about individual structures. Valuable examples include the PDB Structure Explorer (5), PDBsum (6), and the MMDB (7). However, current resources lack the ability to fully present extrinsic information.

Likewise, while there are many databases storing information related to individual organisms (e.g. SGD, MIPS and FlyBase (8-10)), comparative genomics (PEDANT and COGs (9,11)), gene expression (GEO, the Gene Expression Omnibus at the NCBI, and ExpressDB (12)), and protein-protein interactions (DIP and BIND (13,14)), none of these integrates gene sequences, protein interactions, expression levels and other attributes with structure. (However, it should be mentioned that the Sacc3D module of SGD and PEDANT do tabulate the occurrence of folds in genomes.)

PartsList is arranged somewhat differently from most other biological resources. In a usual database (e.g. GenBank(15)) the number of entries increases as the database develops, while each entry has a fairly fixed number of attributes to describe it. In contrast, PartsList is envisioned to have a relatively stable number of entries, i.e. the finite list of protein folds, while the attributes that describe each entry are expected to increase considerably. In the current version of PartsList the properties for a protein fold include: amino acid composition, alignment information, fold occurrences in various genomes, statistics related to motions, absolute expression levels of yeast in different

experiments, relative expression ratios for yeast, worm, and *E. coli* in various conditions, information on protein-protein interactions (based on whole genome yeast interaction data and databank surveys), and sensitivity of the genes associated with the fold to inserted transposons.

One reason to build the database is to compare protein folds in a rich context and in a unified way. This was achieved through ranking. This allows users to directly compare very different attributes of a fold in a uniform numerical format. The rankings can be visualized in three ways: a profiler emphasizing the progression of high and low ranks across many pre-selected attributes, a rankings comparer for custom comparisons, and a numerical rankings correlator. This can help users gain insight into the functions of protein folds in the context of the whole genome. Our system makes it very easy to answer questions like: “What is the most common fold in the worm as compared to *E. coli*?” “What is the most highly expressed fold in yeast and how does this compare to the fold that changes most in expression level during the cell-cycle?” And "which fold has the most protein-protein interactions in the PDB and is it highly ranked in terms of protein motions?"

One of the strengths of the uniform numerical system of ranks in PartsList is that it puts everything into a common framework so that one can see hidden similarities in the occurrence of parts ordered according to many different attributes. In particular, as we describe below, we found that the frequency of many of the attributes falls off according to a power-law distribution (i.e. according to V^{-b}, for attribute value V and a constant b), with a few folds having large attribute values and most having small values. For instance, there are only a few folds that occur many times in the yeast genome, and most only occur once or twice. Likewise, most folds only have a few functions associated with them, but there are a few "Swiss-army-knife" folds that are associated with many distinct functions. Similar power-law-like expressions have been found to apply in a variety of other situations relating to proteins -- for instance, in the occurrence of oligo-peptide words (16-18), in the frequency of transmembrane helices (19) and sequence families with given size (20), and in the structure of biological networks, with a few nodes having many connections and most have only a few (21,22).

PartsList is built on top of the Structural Classification of Proteins (SCOP) (23) fold classification and acts as an accompanying annotation to this system. SCOP is divided into a hierarchy of five levels: class, fold, superfamily, family and protein. The "parts" in our system can be either SCOP folds or superfamilies. However, sometimes for ease of expression we will just refer to “folds” when we really mean “folds and/or superfamilies.” We currently use 420 folds and 610 superfamilies in PartsList. Each is represented by a representative domain, which is also the key for each entry of protein fold.

While we chose to use the SCOP classification, we could equally well have based the system on the other existing fold classifications, e.g. CATH (24), FSSP (25), or VAST (26,27). Moreover, for most attributes, we could also have developed our system around non-structural classifications of protein parts -- e.g. Pfam (28), Blocks (29), or SMART

(30). However, basing it around actual structural folds has the advantage that each part is more precisely and physically defined.

Attributes that can be ranked: Information in the system

Currently the attributes for each entry (i.e. protein fold) can be separated into several main categories: statistical information from a comprehensive set of structural alignments, amino-acid composition information, fold occurrences in various genomes, expression levels in different experiments, protein interactions, macromolecular motion, transposon sensitivity and miscellaneous.

We have developed a formalism for expressing each of the attributes, which is described in Table 1. In the table the term *PART* refers to either fold or superfamily, depending on which of these is being ranked. Essentially, we have a database of attributes where each attribute is given a standardized description and associated with a precise reference. In the following, we describe some main categories of attributes.

Genome Occurrence

The data in this category reveal fold occurrences in 20 different genomes, including 4 archaea, 2 eukaryotes, and 16 bacteria; (additional details online).

The data were obtained in the following fashion: Once a library of folds has been constructed, representative sequences can be extracted (50). Then one can use these to search genomes by comparing each representative sequence against the genomes using the standard pairwise comparison programs, FASTA (55) and BLAST (56) and well-established thresholds (57).

Alternatively, one can build up profiles by running each representative sequence against PDB with PSI-Blast and then comparing these profiles against each of the genomes. This later procedure is more sensitive than pairwise comparison and relatively efficient once the profiles are made up. However, in doing large-scale surveys one has to be conscious of the potential biases introduced due to the profiles being more sensitive for larger families, which often results in the big families getting even bigger.

After the structure assignment, it becomes easy to enumerate how often a fold or structure feature occurs in a given genome or organism. Detailed information can be found in (19,31,32,58). This pools assignments from previous work (59,60).

Alignment

Number of Structures. We did a comprehensive set of structural alignments of structures in the PDB structure databank (35,61,62). The number of structures and aligned pairs used in these comparisons, which are based around Astral (50), give approximate measures of the occurrence of folds in the PDB. Comparison of these values to those for

genome occurrence provides a measure of how biased the composition of the PDB is (63).

Sequence Diversity. The scores from the alignments indicate the sequence diversity between the related structures within folds or superfamilies, in terms of percent sequence identity and a sequence-based P-value. P-values are useful measures of statistical significance of the similarity calculation. A P-value is the probability that one can obtain the same or better alignment score from a randomly composed alignment. A smaller P-value is less likely to have been obtained by chance than a larger P-value. Large P-values close to 1.0 indicate that the similarity is characteristically random and thus insignificant.

Structural Diversity. We also give analogous measures of the diversity of the structures with a given fold, allowing one to rank folds by their degree of variability. We tabulate untrimmed and trimmed RMS, along with the structural P-value. RMS, root-mean-squared deviation in alpha carbon positions, has been the traditional statistic that gauges the divergence between two related structures. Smaller RMS scores indicate more closely related structures. However, sometimes a few ill-fitting atoms may significantly increase the RMS of structures known to be similar. To compensate for this we also report a "trimmed" RMS for a conserved core structure, which is based on the better fitting half of the aligned alpha-carbons, and structural P-value, which compensates for other effects such as structure size. For details, see Wilson et al. (35).

Composition

This allows us to see which folds are most biased in composition of particular amino acids. We use various levels of the Astral clustering of the SCOP sequences to arrive at the composition (50).

Expression

Three techniques are frequently used to obtain genome-wide gene expression data. They are Affymetrix oligonucleotide gene chips, SAGE (Serial Analysis of Gene Expression), and cDNA microarrays (43,64,65). SAGE and, to some degree, gene chips measure the absolute expression levels (in units of mRNA transcripts per cell), while microarrays are used to obtain the expression level changes of a given ORF as the ratio to a reference state.

A main motivation for expression experiments is often to study protein function and to characterize the functions of unannotated genes. However, this does not preclude relating other attributes of proteins, such as their structure, to expression data. For instance, it may be that highly expressed protein folds share a number of characteristics, such as a particularly stable architecture or a composition biased in a certain way. Relating expression and structure involved matching the PDB structure database against the genome and then summing the expression levels of all ORFs containing the same fold. However, if one is trying to find genes expressed in a particular metabolic state, PartsList is not the right place to look.

Absolute. The absolute expression level data gives a good representation of highly expressed genes. All the experiments currently indexed by PartsList are for yeast. For each experiment, in addition to ranking based on the average expression level for a fold, we also consider the composition in the transcriptome and the enrichment of this value relative to its composition in the genome. Transcriptome composition is the fractional composition of a fold (relative to that for other folds) in the mRNA population. In other words, it is the composition of a fold in the genome weighted by the expression levels of each of the genes. The enrichment is the relative change between the composition of a fold in the genome and the transcriptome. For more details, see (33,66). We report values for experiments from a number of different labs (41-44) and a single reference set that merges and scales all the expression sets together.

Ratio. The expression ratio data shows the most actively changing genes over a period of time (e.g. cell cycle) or based on a change in states (e.g. healthy vs. diseased). Source data for expression ratios are the fluctuations in expression of a certain fold over a period of time (e.g. the cell cycle). These are measured in terms of standard deviations for a particular fold, which is calculated from the average of the expression ratio standard deviations for each gene that matches the fold structure.

Interactions

Information on protein-protein interactions is derived from surveys of the contacts in the PDB and the experiments in yeast.

PDB. To determine which domains interact with one another in the PDB entries indexed by SCOP (9,580 at the time of the analysis), the coordinates of each domain were parsed to check whether there are five or more contacts within 5 Å to another domain, as described in (67). The distance of 5 Å was chosen, as this is a conservative threshold for interaction between two atoms, where the atoms are either Cα's or atoms in side-chains. The 5-contact threshold was chosen to make sure the contact between the domains was reasonably extensive. (In fact, the number of domains identified as contacting each other hardly changed for thresholds between 1 and 10 contacts and 3 to 6 Å distances).

Yeast. The interactions between structural domains in the yeast genome were obtained by assigning protein structures to the yeast proteins using PSI-BLAST and PDB-ISL as described in Teichmann et al (39,68). Assigned structural domains contained within the same ORF that were adjacent within 30 amino acids were assumed to interact. (This is generally true of the domains in the PDB, with a few exceptions, such as domains in transcription factors like adjacent zinc fingers, or variable and constant immunoglobulin domains.) To derive intermolecular interactions in the yeast genome we combined three sets of protein-protein interactions: (i) the MIPS web pages on complexes and pairwise interactions (February 2000)(9), (ii) the global yeast-two-hybrid experiments by Uetz et al. (51) and (iii) large-scale yeast two-hybrid experiments by Ito et al. (52). Out of all these pairwise interactions known for yeast ORFs, there is a limited set in which both partners are completely covered by one structural domain (to within 100 residues). This set of protein pairs was used to derive a further set of domain contacts in the yeast genome as described in (67).

Motions

Information on motions is from the Macromolecular Motions Database (36,37). We consider a set of approximately 4400 motions automatically identified by examining the PDB and a smaller, manually curated set of motions. For each fold we determine the number of entries in the motions database that are associated with it. Then over this set of motions we either average or take the maximum value of a number of relevant statistics describing the motion, i.e. the maximum Cα displacement in the motion, the overall rotation of the motion, and the energy difference between the start and endpoints of structures involved in the motion.

Transposon Sensitivity

Ross-MacDonald et al. (40) developed a procedure for randomly inserting transposons throughout the yeast genome. They investigated the phenotypes resulting from each insertion in 20 different growth conditions in comparison to wild-type growth. The experiment for each insertion in each condition was repeated several times. If the observed phenotype of the mutant deviates from the average wild-type phenotype, this could be either because of a real effect of the mutation on the cell or it could just a be typical variation of the phenotype of wild-type cells. We developed a P-value score that measures the degree of confidence that the observed phenotype results from randomly changing wild-type cells. The negative logarithm of this P-value rises with the significance of the phenotype measurements and can be understood as the sensitivity of the cell to mutations in a particular gene. We calculated a value for the transposon sensitivity for protein folds by geometrically averaging the P-values of the associated genes.

Miscellaneous

The miscellaneous section includes any information that does not fit into a major category. It includes: number of pseudogenes in worm associated with a fold (53), total number of functions and number of enzymatic functions associated with a fold (54), the average length of the sequence, and the year the domain structure was originally determined.

Errors

The above data, of course, have systematic and statistical errors. For some attributes we expect considerably smaller errors than others. For instance, we expect the numbers related to the sequence composition of different folds (e.g. the Ala composition) to be particularly accurate, since the only factors affecting these are errors in the underlying sequence of the protein and in the scop fold classification itself. In contrast, there is a considerable known rate of false positives associated with the global protein interaction experiments using the two-hybrid method (51,69), and this suggests statistics based on yeast interactions may be somewhat less accurate. Furthermore, the precise values for the rankings in PartsList are also contingent on the evolving contents of various databanks. Thus, over time as more structures are determined, one should expect statistics such as the most common folds in a particular genome to change somewhat. A very detailed

discussion of the expected errors in the various quantities in PartsList is available on the web from the help section.

Ranking all the folds based on extrinsic information

The PartsList resource facilitates exploring extrinsic information by dynamically ranking protein folds in different contexts, such as genome and expression levels. We provide three tools for visualizing the rankings: Comparer, Correlator, and Profiler. The overall structure of PartsList is schematically shown in Fig. 1.

Comparer

The motivation behind Comparer is to allow one to rank folds according to a given attribute and then see the ranks associated with other attributes. The ranking attribute and the additional attributes are selected by the user. Figure 2(a) shows an example. The most common folds in *E. coli* are shown alongside three other attributes: fold occurrence in yeast, fluctuation in expression level during the yeast cell cycle, and fluctuation in expression level in *E. coli* during heat shock. Which displayed attribute is used to rank the folds can be easily changed; in the example in Figure 2(a) the report can be re-sorted based on the other three attributes by clicking on arrows.

Profiler

In principle, Profiler presents the same information as Comparer. However, it shows the progressing pattern for several pre-selected categories and is intended to give people an easy-to use interface that gives some simple views of the data. Figure 2(b) shows an example that highlights the phylogenetic pattern of fold occurrence in 20 genomes.

Correlator

Correlator uses linear and rank correlation coefficients to measure the association between two selected attributes. The difference between these two types of correlation coefficients is that the former relates to the actual values while the latter relates to the ranks among the samples. The interpretation of the linear correlation coefficient can be completely meaningless if the joint probability distribution of the variables is too different from a binormal distribution. This is the reason for introducing the rank correlation coefficient. Correlator provides both coefficients for the selected quantities. In most cases, they are close. For example, the linear correlation coefficient and rank correlation coefficient for fold occurrence in genomes *A. fulgidus* and *M. jannaschii* (Aful and Mjan) are 0.88 and 0.77, respectively, while the corresponding coefficients for fold occurrence in *A. fulgidus* and *S. cerevisiae* (Scer) are 0.52 and 0.48, respectively. This is not surprising, as the first two genomes are both Archaeal, while in the second comparison one genome belongs to Archaea (Aful) and another to Eucarya (Scer). As one would expect, the fold occurrences for the more closely related genomes have a higher correlation.

In addition to the coefficients, Correlator displays a scatter plot to aid in visualizing the correlation between the selected fold attributes. Figure 2(c) shows the scatter plot for the second example above: the correlation between occurrences in the *A. fulgidus* and *S.*

cerevisiae genomes. One can easily observe that some folds appear frequently in Scer but seldom or never in *A. fulgidus*. By clicking on a point on the plot, one obtains detailed information about the corresponding fold. This kind of plot can reveal interesting folds with certain relationships between attributes even though in some cases the overall correlation coefficients between the two attributes are almost zero (i.e. no correlation).

Power-Law Behavior of Many Disparate Attributes

Going back and forth between Correlator and Comparer allows one to see interesting relationships between disparate attributes of proteins. Figure 3 illustrates a comparison of two attributes, functions and interactions. It shows a ranking of the folds that have the most interactions in the PDB in comparison to those that have the most functions. It is immediately apparent that there are only a few folds with large values of either attribute, i.e. many functions or interactions. Moreover, the most multi-functional folds also have the most distinct interactions with other folds, suggesting that a few a folds may function as general-purpose parts.

In fact, the uniform system of ranks in PartsList shows that "only a few folds having large values for an attribute" is a generally true statement for many of the disparate attributes catalogued by the system. Moreover, the falloff from high to low values for a given attribute often follows a power-law distribution. That is, the normalized frequency F that a number of distinct folds have a particular attribute value V follows a functional form like:

$$F(V) = a\ V^{-b}$$

where a and b are constants. Note that $F(V)$ is just the number of folds with an attribute value V divided by the total number of folds and that on a log-log plot this function becomes a straight line with slope $-b$. Often the attribute value V itself reflects the *occurrence* of a fold in a particular context -- e.g. V could be the number of times a given fold occurs in a particular genome. Quantities that follow a power-law-like behavior are often said to have a form like that of Zipf's law, which often occurs in the analysis of word frequency in documents (70).

Thus far, this general conclusion is described in language sufficiently abstract to accommodate the many different types of attributes in PartsList. A few concrete examples will make the conclusion clearer. For instance, we find that in genomes most folds occur only once while there are only a very few folds that occur many times. An illustration is shown in the upper panel of Fig. 5 for *E. coli*. The x-axis is the number of times a particular fold occurs in the *E. coli* genome and the y-axis shows the number of distinct folds that have same occurrence. (This is normalized by dividing by the total number of folds so that the maximum value on y-axis is 100%.) From the log-log format of the plot, one can immediately see that the falloff obeys a power-law, with a few folds occurring many times and most only once or twice. The middle panel shows other attributes that display similar power-law-like behavior, including expression level in

yeast, number of functions associated with a fold, and number of protein-protein interactions found in the PDB. Of course, not all attributes follow a power-law. The lower panel shows two of these less typical attributes: Asp composition in a fold and average number of residues involved in a motion.

One of the strengths of the uniform numerical system of ranks in PartsList is that it puts everything into a common framework so that one can see similarities across disparate attributes. We believe it would be difficult to see a common power-law behaviour for many aspects of protein structure without PartsList.

Traditional Single-Structure reports

In addition to the tools that compare and relate the extrinsic properties of protein folds, we provide traditional reports that are more focused on an individual structure.

Occurrence report. This allows users to see the number of times that a fold corresponding to the queried protein structure occurs in various genomes. This gives a phylogenetic profile of the occurrence of a particular fold in 20 genomes, similar in spirit to the fold patterns discussed earlier (19).

Function report. This summarizes the functional classification of the queried PDB structure. It merges a number of functional classifications, including FlyBase(10), ENZYME(71), GenProtEC(72) and MIPS(9). Our approach to functional classification is described in a number of previous publications (35,54). In short, we used pairwise comparison to cross-reference the PDB domains against Swissprot. Depending on whether they had an Enzyme Commission number, we were able to divide all entries into enzymes and nonenzymes, a division that represents the highest level in our classification. (For the enzyme category, we only transferred Enzyme Commission numbers to those SCOP domains with a one-to-one match to a Swissprot enzyme.) In the absence of an EC-type classification for nonenzymes, we assigned functions to nonenzymatic SCOP domains according to Ashburner's original classification of Drosophila protein functions. This classification is derived from a controlled vocabulary of fly terms, is available on the web, and is loosely connected with the FLYBASE database (10). It has recently been superceded by the GO functional classification (73). MIPS and GenProtEC classifications to SCOP domains were assigned based on sequence comparisons to classified yeast and *E. coli* ORFs, respectively. The SCOP domain most closely matching each ORF classified in MIPS or GenProtEC was assigned the corresponding MIPS or GenProtEC function number. Only matches of 80% sequence identity or greater were considered.

Alignment report. This gives detailed information on structural alignments available between pairs of protein domains associated with a fold. A pair viewer is provided, which gives many key statistics about the alignment (e.g. RMS, sequence identity, number of fit atoms, etc.), in addition to a listing of the actual aligned residues. Both HTML and parseable text views are available.

Interaction report. This shows all the pairs of protein-protein interactions associated with a fold based on either the PDB survey or yeast genome data.

Rank report. This highlights the top-five and bottom-five ranked attributes associated with a fold. It also shows all attributes ordered by the rank they are given in that fold. It, thus, highlights for a particular fold the attributes with respect to which it most stands out. That is, it highlights the "outlier attributes" of each fold, the way each fold is most unique. The rank report could be used, for example, by a protein engineer interested in determining the unique properties of a structure he is working on.

PDB report. This summarizes all the information concerning a domain or a representative PDB structure. It includes: (i) a summary of the occurrence report; (ii) a summary of the alignments available for structures in the same superfamily and fold; (iii) a description of motions and motion-movies associated with the structure in the Macromolecular Motions database (36,37); (iv) a summary of the merged functional classification; (v) a core structure, if available (74); (vi) ranking tables of the queried structure in various datasets; and (vii) a summary of the interactions report. Figure 4 shows a sample PDB report for structure 1AMA.

Fold report. This lists all the SCOP domains associated with the queried fold and provides information (similar to that in the PDB report) that is common to all -- i.e. genome occurrence, alignment report, and rankings.

Summary and Discussion

We developed a web-based system for dynamically ranking protein folds based on disparate attributes, including fold occurrence in various genomes, expression level, alignment statistics, protein-protein interactions, motion statistics, and transposon sensitivity. Three ranking tools are provided -- Comparer, Profiler, and Correlator -- which can help users to place one fold in context of all other ones. The uniform system of ranks employed by PartsList provides a good framework for comparing different experiments and gaining a broad perspective on the complexity of genomes.

We anticipate that PartsList will have a relatively stable number of entries (i.e. folds), while for each entry the attributes that describe it will increase over time. In the future as experiments yield new information, PartsList will include more and more attributes. In particular, we anticipate that much new expression information will be incorporated. We also plan to develop a form to allow automatic submission of new ranking attributes and to encourage people to submit any ranking information.

Acknowledgments:

We thank NIH Structural Genomics Program and the Keck Foundation for support.

References:

1. Chothia, C. *Proteins. One thousand families for the molecular biologist* (1992) *Nature*, **357,** 543-544.

2. Brenner, S. E., Hubbard, T., Murzin, A., Chothia, C. *Gene duplications in H. influenzae* (1995) *Nature*, **378,** 140.

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

4. Consortium, T. C. e. S. *Genome sequence of the nematode C. elegans: a platform for investigating biology.* (1998) *Science*, **282**(5396)**,** 2012-8.

5. Berman, H., M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. *The Protein Data Bank* (2000) *Nucleic Acids Res.*, **28,** 235-242.

6. Laskowski, R. A., Hutchinson, E.G., Michie. A.D., Wallace, A.C., Jones, M.L., Thornton, J.M. *PDBsum: a Web-based database of summaries and analyses of all PDB structures* (1997) *Trends Biochem. Sci.*, **22,** 488-490.

7. Wang, Y., Addess, K.J., Geer, L., Madej, T., Marchler-Bauer, A., Zimmernan, D., Bryant, S.H. *MMDB: 3D structure data in Entrez* (2000) *Nucleic Acids Res.*, **28,** 243-245.

8. Ball, C. A., Dolinski, K., Dwight, S.S., Harris, M.A., Issel-Tarver, L., Kasarskis, A., Scafe, C.R., Sherlock, G., Binkley, G., Jin, H., Kaloper, M., Orr, S.D., Schroeder, M., Weng, S., Zhu, Y., Botstein, D., Cherry, J.M. *Integrating functional genomic information into the Saccharomyces genome database* (2000) *Nucleic Acids Res.*, **28,** 77-80.

9. Frishman, D., Heumann, K., Lesk, A., Mewes, H. W. *Comprehensive, comprehensible, distributed and intelligent databases: current status* (1998) *Bioinformatics*, **14,** 551-561.

10. FlyBase. *The FlyBase database of the Drosophila Genome Projects and community literature* (1999) *Nucleic Acids Res.*, **27,** 85-88.

11. Tatusov, R. L., Galperin, M.Y., Natale, D.A., Koonin, E.V. *The COG database: a tool for genome-scale analysis of protein functions and evolution* (2000) *Nucleic Acids Res.*, **28,** 33-36.

12. Aach, J., Rindone, W., Church, G.M. *Systematic management and analysis of yeast gene expression data* (2000) *Genome Res.*, **10,** 431-445.

13. Bader, G. D., Hogue, C.W. *BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways* (2000) *Bioinformatics*, **16,** 465-477.

14. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D. *DIP: the database of interacting proteins* (2000) *Nucleic Acids Res.*, **28,** 289-291.

15. Benson, D. A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L. *GenBank* (2000) *Nucleic Acids Res.*, **28,** 15-18.

16. Konopka, A. K., Martindale, C. *Noncoding DNA, Zipf's law, and language [letter]* (1995) *Science*, **268,** 789.

17. Flam, F. *Hints of a language in junk DNA [news]* (1994) *Science*, **266,** 1320.

18. Bornberg-Bauer, E. *How are model protein structures distributed in sequence space?* (1997) *Biophys. J.*, **73,** 2393-2403.

19. Gerstein, M. *Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census* (1998) *Proteins*, **33,** 518-534.

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

21. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L. *The large-scale organization of metabolic networks* (2000) *Nature*, **407,** 651-654.

22. Amaral, L. A. N., Scala, A., Barthelemy, M., Stanley, H.E. *Classes of small-world networks* (2000) *Proc. Natl. Acad. Sci. USA*, **97,** 11149-11152.

23. Murzin, A. G., Brenner, S.E., Hubbard, T., Chothia, C. *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* (1995) *J. Mol. Biol.*, **247,** 536-540.

24. 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* (1997) *Structures*, **5,** 1093-1108.

25. Holm, L., Sander, C. *Mapping the protein universe* (1996) *Science*, **273,** 595-602.

26. Gibrat, J. F., Madej, T., Bryant, S.H. *Surprising similarities in structure comparison* (1996) *Curr. Opin. Struc. Biol.*, **6,** 337-385.

27. Madej, T., Gibrat, J-F., Bryant, S.H. *Threading a database of protein cores* (1995) *Proteins*, **23**(356-369).

28. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., Sonnhammer, E.L.L. *The Pfam protein families database* (1999) *Nucleic Acids Res.*, **27,** 260-262.

29. Henikoff, J. G., Greene, E.A., Pietrokovski, S., Henikoff, S. *Increased coverage of protein families with the blocks database servers* (2000) *Nucleic Acids Res.*, **28,** 228-230.

30. Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. *SMART, a simple modular architecture research tool: identification of signaling domains* (1998) *Proc. Natl. Acad. Sci. USA*, **95,** 5857-5864.

31. Hegyi, H., Lin, J., Gerstein, M. (2000) *submitted*.

32. Gerstein, M., Levitt, M. *A structural census of the current population of protein sequences* (1997) *Proc. Natl. Acad. Sci. USA*, **94,** 11911-11916.

33. Jansen, R., Gerstein, M. *Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins* (2000) *Nucleic Acids Res.*, **28,** 1481-1488.

34. Drawid, A., Jansen, R., Gerstein, M. *Genome-wide analysis relating expression level with protein subcellular localization* (2000) *Trends Genet.*, **16,** 426-429.

35. Wilson, C. A., Kreychman, J., Gerstein, M. *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores* (2000) *J. Mol. Biol.*, **297,** 233-249.

36. Gerstein, M., Krebs, W. *A database of macromolecular motions* (1998) *Nucleic Acids Res.*, **26,** 4280-4290.

37. Krebs, W., Gerstein, M. *The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework* (2000) *Nucleic Acids Res.*, **28,** 1665-1675.

38. 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* (1998) *J. Mol. Biol.*, **284,** 1201-1210.

39. Teichmann, S., Chothia, C., Church, G., Park, J. *Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL* (2000) *Bioinformatics*, **16,** 117-124.

40. Ross-Macdonald, P. C., P.S.R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F.K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G.S., Snyder, M. *Large-scale analysis of the yeast genome by transposon tagging and gene disruption* (1999) *Nature*, **402,** 413-418.

41. Jelinsky, S. A., Samson, L.D. *Global response of Saccharomyces cerevisiae to an alkylating agent* (1999) *Proc. Natl. Acad. USA.*, **96,** 1486-1491.

42. Holstege, F. C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C. J., Green, M.R., Golub, T.R., Lander, E.S., Young, R.A. *Dissecting the regulatory circuitry of a eukaryotic genome* (1998) *Cell*, **95,** 717-728.

43. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr, Hieter, P., Vogelstein, B., Kinzler, K.W.,. *Characterization of the yeast transcriptome* (1997) *Cell*, **88,** 243-251.

44. Roth, F. P., Hughes, J. D., Estep, P.W., Church, G. M. *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation* (1998) *Nature Biotechnology*, **16,** 939-945.

45. Spellman, P. T., Sherlock, G., Zhang, M.Q., Iyer, V.R.,, Anders, K. Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B. *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization* (1998) *Mol Biol Cell*, **9,** 3273-3297.

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

47. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., Herskowitz, I. *The transcriptional program of sporulation in budding yeast* (1998) *Science*, **282,** 699-705.

48. Richmond, C. S., Glasner, J.D., Mau, R., Jin, H., Blattner, F.R. *Genome-wide expression profiling in Escherichia coli K-12* (1999) *Nucleic Acids Res.*, **27,** 3821-3835.

49. Wixon, J., Blaxter, M., Hope, I., Barstead, R., Kim, S. *Caenorhabditis elegans* (2000) *Yeast*, **17,** 37-42.

50. Brenner, S. E., Koehl, P., Levitt, M. *The ASTRAL compendium for protein structure and sequence analysis* (2000) *Nucleic Acids Res.*, **28,** 254-256.

51. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M. *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae* (2000) *Nature*, **403,** 623-627.

52. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., Sakaki, Y. *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins* (2000) *Proc. Natl. Acad. Sci. USA*, **97,** 1143-1147.

53. Harrison, P., Echols, N., Gerstein, M. *Digging for Dead Genes: An Analysis of the Characteristics of the Pseudogene Population in the C. elegans Genome* (2001) *Nucleic Acids Res.*, **29,** 818-830.

54. Hegyi, H., Gerstein, M. *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome* (1999) *J. Mol. Biol.*, **228,** 147-164.

55. Lipman, D. J., Pearson, W.R. *Rapid and sensitive protein similarity searches* (1985) *Science*, **227,** 1435-1441.

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

57. Brenner, S., Chothia, C. and Hubbard, T. *Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships* (1998) *Proc. Natl. Acad. Sci. USA*, **95,** 6073-6078.

58. Teichmann, S., Chothia, C., Gerstein, M. *Advances in structural genomics* (1999) *Curr. Opin. Struc. Biol.*, **9,** 390-399.

59. Gerstein, M., Lin, J., Hegyi, H. *Protein folds in the worm genome* (2000) *Pacific Symposium on Biocomputing*, **5,** 30-42.

60. Lin, J., Gerstein, M. *Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels* (2000) *Genome Res.*, **10,** 808-818.

61. Levitt, M. and Gerstein, M. *A Unified Statistical Framework for Sequence Comparison and Structure Comparison* (1998) *Proceedings of the National Academy of Sciences USA*, **95,** 5913-5920.

62. Gerstein, M. and Levitt, M. *Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the Scop Classification of Proteins* (1998) *Protein Science*, **7,** 445-456.

63. Gerstein, M. *How representative are the known structures of the proteins in a complete genome? A comprehensive structural census* (1998) *Folding & Design*, **3,** 497-512.

64. Brown, P. O. and Botstein, D. *Exploring the new world of the genome with DNA microarrays* (1999) *Nat Genet*, **21**(1 Suppl)**,** 33-7.

65. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. and Lockhart, D. J. *High density synthetic oligonucleotide arrays* (1999) *Nat Genet*, **21**(1 Suppl)**,** 20-4.

66. Gerstein, M., Jansen, R. *The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function* (2000) *Curr. Opin. Struc. Biol.*

67. Park, J., Lappe, M., Teichmann, S.A. *Mapping Protein Family Interactions: Intra- and Intermolecular Interactions Repertoires are Distinct* (2000) *J. Mol. Biol.***,** (in press).

68. Teichmann, S. A., Park, J., Chothia, C. *Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements* (1998) *Proc. Natl. Acad. Sci. USA*, **95,** 14658-14663.

69. Schwikowski, B., Uetz, P. and Fields, S. *A network of protein-protein interactions in yeast* (2000) *Nat Biotechnol*, **18**(12)**,** 1257-61.

70. Knuth, D. (1973) The Art of Computer Programming:, 3., Addison-Wesley, Readin, MA.

71. Bairoch, A. *The ENZYME data bank* (1993) *Nucleic Acids Res.*, **21,** 3155-3156.

72. Riley, M., Labedan, B. (1996) In Neidhardt, F., Curtiss, III, R., Lin, E.C.C., Ingraham, J., Low, K.B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M., Umbarger, H.E. (ed.), Escherichia coli and Salmonella: Cellular and Molecular Biology. ASM Press, Washington D.C., pp. 2118-2202.

73. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium* (2000) *Nat Genet*, **25**(1)**,** 25-9.

74. Schmidt, R. B., Gerstein, M., Altman, R.B. *LPFC: an Internet library of protein family core structures* (1997) *Protein Science*, **6,** 246-248.

Figure and Table captions

Table 1

This table shows all the attributes ranked by PartsList. The formalism for specifying an attribute has two parts: an overall category, denoted by a single uppercase symbol, and some parameter choices, which are denoted by lower-case arguments to the first symbol. Some examples for folds will suffice to make this clear: G(aful) is genome occurrence of a particular fold in *A. fulgidus*; M(nhinges,goldstd) is the maximum value of the number of hinges statistic from surveying a set of motions in the gold-standard subset of the Macromolecular Motions Database, where this statistic is only calculated for the entries in the motions database that are associated with a particular fold; And I(pdball,inter) is the number of distinct types of protein-protein interactions found in a survey of the PDB, subject to the restriction that the interactions must be between folds on different chains.

Figure 1

The overall structure of PartsList. Three tools (Profiler, Comparer, and Correlator) provide an easy way to access and manipulate the display of the dataset. With these tools, users can isolate interesting folds and obtain fold reports about them. Further clicks take one to PDB report, which gives detailed information about an individual structural domain, including its genome occurrence, alignment information, molecular motions, functional annotation, interactions, and core structure.

Figure 2

Sample displays. **(A)** a sample Comparer display: the four selected attributes are the fold genome occurrence in yeast, the analogous quantity for *E. coli*, fluctuation of expression level for CDC28 synchronized yeast cell during the cell cycle , and the corresponding values for *E. coli* to heat shock. (Using the nomenclature in Table 1 these quantities are G(scer), G(ecol), F(cdc28), and F(heatec).) The folds are ranked in terms of fold occurrence in *E. coli* and the most common fold here is the TIM-barrel (represented by the SCOP domain d1aj2__). If one clicks the "Display ranks" button, the values in the cells will be replaced by the ranks in their respective columns. By clicking the "re-rank" arrows, one can also obtain other views by sorting on other attributes. **(B)** Shows the occurrences of folds in 20 genomes in Profiler. **(C)** Shows the correlation between the fold occurrences in the *A. fulgidus* and *S. cerevisiae* genomes (G(aful) and G(scer)) . Both linear and rank correlation coefficients are calculated. The linear correlation coefficient is defined as: $R = \frac{1}{N-1}\mathbf{X} \bullet \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are two vectors with N elements. Each element of the **X** vector is normalized thus: $X_i = \frac{X_i' - \overline{X}}{\sigma_x}$, where $\overline{X}$ and σ_x are the average and standard deviation of the values of the original data vector **X'**, respectively. **Y** is normalized in a similar fashion. For two perfectly correlated datasets, $R = 1$, while for two completely uncorrelated datasets, $R = 0$. If we replace X_i by its rank among all the other X_i in the sample (i.e., 1,2,3…,N), then we get the rank correlation coefficient. A scatter plot is also shown to help in visualizing this correlation.

Figure 3

The relation between the number of functions associated with a protein fold and the number of distinct protein-protein interactions it has (based on a survey of the PDB databank). These are X(func) and I(pdball,none) using the nomenclature in Table 1. This relationship can be displayed both in Comparer (left) and Correlator (right).

Figure 4

An sample PDB report for structure 1AMA. The report summarizes the relevant information for this domain, including genome occurrences, alignment, motions, function classification, core structure and rankings. By clicking on the headers, one can get the detailed reports for these quantities.

Figure 5

Some novel relationships that are highlighted by the PartsList system.

Upper panel shows the occurrence of folds in the *E. coli* genome plotted on a log-log scale -- i.e. G(ecol) using the nomenclature in Table 1. The x-axis is the fold occurrence in the genome, while the y-axis is the number of folds with a particular occurrence. The fit of the points to a straight line shows that the falloff obeys a power law with constants a=0.35 and b=1.3 (see text).

Middle panel shows other attributes that also follow power-law behavior: the average expression level according to our merged and scaled set (L(ref) with a=.3 and b=1.2), the number of protein-protein interactions (I(pdball,none) with a=.52 and b=1.6), and the number of functions (X(func) with a=.76 and b=2.5).

Lower panel shows some attributes that do not follow power-law behavior: the Asp composition of the fold (B(Ala,pdb100)) and the number of mobile residues during a motion (M(nresidue,auto)). The fold occurrence in *E. coli* is plotted as a reference.