Establishment of a Knowledge Base for Function Annotations in High-Throughput Sequence Analysis

G. X. Yu* and E. Marland

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

*To whom correspondence should addressed.

ABSTRACT

Motivation: The rapid accumulation of sequence data and information describing regulatory and metabolic networks has triggered the development of integrated systems for genome sequence analysis. However, a great deal of uncertainty exists in the annotations found in these systems because of the heterogeneities in the public databases and limitations in current computational approaches. Conflicts in assignments based on different computational tools add additional uncertainty to the annotations, and the situation is compounded by a lack of tools for cross-verification. These uncertainties have greatly affected the performance of genome analysis systems, specifically with regard to the accuracy of functional assignments to the genes. In order to minimize the effect of these uncertainties, a biological knowledge base is needed to provide rules for guiding function annotations and a global reference system for cross-verification of the results obtained by analysis using different computational tools.

Results: In this study, we have developed a rule-based knowledge system specifically for automated high-throughput genetic sequence analysis. It includes 22,612 protein function groups and their evolutionary spaces (distributions), which are characterized by protein sequence conservations, the phylogenetic distribution of protein motifs and domains, and their relationships to biological functions. Our knowledge base demonstrates that tremendous variations exist among protein functional groups. Over half of the protein functional groups are highly diversified in sequence similarities (53.6%, and 51.4% in Blast and Blocks measurements, respectively). With regard to protein relationships, we found that Pfam patterns have much higher resolution and broader coverage than Blocks families. Out of 10,604 protein functional groups that Blocks covered, 811 (7.6%) can be uniquely identified. In contrast, Pfam patterns cover 13,803 significant protein functional groups, and 1,899 (almost 14%) of them have unique identifiers. However, most of the relationships between protein functions and protein families or Pfam patterns are complex. Each of the protein families or Pfam patterns can correspond to multiple functions or vice versa. Hence, these families or patterns need to be further defined or additional tools introduced so that each function can be identified through its own unique set of features. One of the important applications of our knowledge base is cross-verification of protein function annotations obtained by different computational tools. Additional applications of this knowledge base are discussed in the paper.

Contact **gxyu@mcs.anl.gov**

INTRODUCTION

The past several decades has witnessed an unprecedented accumulation of genome sequence data. Sequence data from over 800 organisms can be found in NCBI Entrez Genomes page. These genomes cover all three main domains of life – Eubacteria, Archaea, and Eukarya (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). The rapid accumulation of functionally uncharacterized open reading frames (ORFs) in the genomes has triggered the development of integrated systems for genome sequence analysis, for example, GenQuiz (Andrade et al. 1999), MAGPIE (Gaasterland and Sensen 1996), and PEDANT (Frishman and Mewes 1997). These systems enable users to analyze many sequences in a consistent and efficient manner (Scharf et al. 1994; Casari et al. 1996). One of the main principles of sequence analysis is the correlation of functions with the similarities of sequences and structures, by which functional information is transferred from known proteins to the unknown proteins. Sometimes, however, these transfers can be extremely uncertain because of the following factors.

First, the sequence similarity is not always strongly associated with function, although it is one of the main principles on which most automated annotation programs have been developed (Frishman and Mewes 1997; Gaasterland Sensen 1996; Andrade et al. 1999). In most cases, the general rule is that the higher the similarity in sequences and structures between two sequences, the more certain that they will have similar functions. Exceptions exist, however. On the one hand, among some orthologous genes in phylogenetically distant organisms, sequence similarities are no longer recognizable (Shimamoto and Kyozuka 2002; Xiong and Bauer 2002). Furthermore, only marginal or no sequence similarity can be detected for those that have resulted from convergent evolution (Mardulyn et al. 1997; Csete and Doyle 2002; Gregory et al. 2002). On the other hand, the general rules that resulted from subfunctionalization (Lynch and Force 2000; Massingham et al. 2001; Van de Peer et al. 2001) and enzyme recruitment (Naumann et al. 2002; Collins and Mitchell 2002) cannot be used for differentiating the genes. It is essential to systematically categorize all these genes based on their evolutionary status. Therefore, a set of computational tools is needed for each gene category so that its function can be uniquely identified.

Second, the annotations and formats currently found in public databases are highly heterogeneous, thereby presenting a real challenge in their application in genome analysis systems. For example, some databases such as GenBank and EMBL may provide only minimal information for gene functions, which is primarily from gene function prediction programs (Benson et al. 2002; Stoesser et al. 2002). In contrast, other databases, such as WormPep, TREMBL, PIR, and Swiss-Prot, offer function descriptions that are extensive and of relatively high quality (Bairoch and Apweiler 2000; Barker et al. 2001; Stein et al. 2001; Stoesser et al. 2002). Furthermore, the annotations provided are sometimes inconsistent in the use of nomenclature and may not be computer compatible. These factors significantly complicate comparative analysis, a key method for functional characterization. Thus, the choice of reliable data resources and development of a knowledge base that is independent of individual database formats will play a critical role in highly efficient protein annotation.

Third, results from computational tools used in genetic sequence analysis and gene functional annotations are often incompatible. Current computational tools can be classified into three major groups based on the features of protein sequence they capture. The first group includes Blast (Altschul et al. 1990) and FastA (Lipman and Pearson 1985), which evaluate global sequence similarities. The second group consists of Blocks (Henikoff 2000), Pfam (Bateman et al. 2002), Prosite (Falquet et al. 2002), Smart (Schultz et al. 1998) and Prints (Attwood et al. 2000); these tools identify unique protein sequence-motifs, which may have some important biological functions. The third group of tools includes SWISS-MODEL (Peitsch 1996) and VAST (Madej et al. 1995), which provide information about protein structures. The integration of these three groups of tools is critical for reliable determination of a function of a gene. Moreover, a global reference system is necessary for these tools so that results can be evaluated in a systematic way and conflicts can be reliably resolved.

To minimize the aforementioned uncertainties and improve the accuracy of functional assignments, we have developed a rule-based knowledge system. The biological rules are defined based on the Swiss-Prot database (Bairoch and Apweiler 2000). Our choice of Swiss-Prot is driven by the fact that this curated protein sequence database provides a high level of annotation, a minimal level of redundancy, and a high level of integration with other databases (Bairoch and Apweiler 2000). Our knowledge base contains 22,612 protein functional groups based on a combination of Blocks analysis and lexical processing of function descriptions in the Swiss-Prot database. The establishment of these protein functional groups is intended to describe the smallest biochemical or evolutionary units encoded by single genes in protein complexes like subunits. For each functional group, the sequence conservations (Blast), protein signatures (Blocks and Pfam), and their evolutionary distributions are calculated. These features define the evolutionary spaces of these functional groups, which represent their current evolutionary status. The unique relationships between functions with Blocks protein families and Pfam domains are also included in our knowledge system.

In summary, our knowledge base can function in two distinct ways. As an evolutionary guide, it can check the evolutionary status of protein functions and build unique relationships between the sequences and functions. As a global reference system, it can cross-verify results from different computational tools, resolving conflicts in annotations. The use of our knowledge base therefore will enhance user confidence in annotations and will prevent overinterpretation. Further applications of this knowledge base, problems related to its use in protein functional annotations, possible solutions to these problems, and future directions of development are discussed.

METHOD AND ALGORITHM

We followed a three-step procedure to process data and build our rule-based knowledge base. At step 1, we classified the Swiss-Prot database proteins by using a dual-procedure algorithm that we had developed. The first procedure grouped proteins by lexical processing of function descriptions. Specifically, we extracted the annotations from “DE” field in the Swiss-Prot database and categorized the protein sequences into “enzymatic” if EC numbers were assigned, “non-enzymatic” otherwise. To further classify the enzymatic proteins, we simply grouped them based on their EC numbers. To further classify the non-enzymatic proteins, we devised a filtering procedure to eliminate words with no functional meanings (e.g., intergenic, similar, potential, or possible) and a dictionary that associated keyword characteristics with function descriptions. The results of the classification were verified manually; such a check is essential for quality control because of the inconsistencies in the annotations of some database entries. In the second procedure, the protein sequences in these groups were processed with the Blocks search tool and further separated into subgroupsThe Blocks database consists of blocks constructed from documented families of proteins using the automated PROTOMAT system (Henikoff and Henikoff 1991). The blocks are ungapped multiple alignments corresponding to the most conserved regions of given protein families (Hofmann et al. 1999; Henikoff et al. 2000). Blocks analysis therefore can pinpoint conserved regions with quantification by E-value and can result in the assignment of particular protein families to analyzed sequences.

At step 2, we calculated the evolutionary spaces for the functional groups. These spaces comprise all possible features that describe the overall picture of protein evolution in the functional groups. The features include the current status of sequence conservations, protein signatures, and their distributions among the different life domains. In addition to Blocks, we included Blast and Pfam in the analysis. Each of these computational tools is able to identify unique features of analyzed proteins. The results of Blocks analysis describe the occurrence of conserved blocks within the protein groups, the association of these patterns with the life domains, and the variation in the degree of conservation. While the Blocks tool pinpoints the localized module of conserved regions and relationships between individual proteins to the consensus of protein families, the Blast algorithm emphasizes more global sequence similarities. Furthermore, pairwise relationships among protein members within the functional groups can be accurately defined. Thus, the evolutionary status of the protein functional groups can be characterized at a global sequence level. Pfam is a database of protein domain families; it contains curated multiple sequence alignments for each family, as well as profile hidden Markov models for finding these domain in new sequences (Bateman et al. 2000). Pfam search results can determine the occurrence of certain biological functional domains with a reliability measured by E-value. As a result, the domain occurrences and their relationships with protein functions can be determined.

At step 3, we processed the data further to extract "rules" related to the biological functions. These rules include the unique relationships between protein signatures and cellular functions, the minimal requirements of Pfam domains. Finally, information including protein functional classification, the evolutionary space for protein functional groups, and the biological rules was formatted and stored in the database for use as a knowledge base.

RESULTS

Classification of protein functional groups provides considerable power to differentiate subgroups or versions in protein machines.

The Swiss-Prot database, version of 6/25/2002 (ftp://ftp.expasy.org/databases/swiss-prot/), which includes 111,046 protein entries, was downloaded and analyzed. The dual procedure of protein classification resulted in 5,164 enzymatic functional groups and 17,448 non-enzymatic functional groups. The enzymatic groups consist of 54,321 genes, which cover 1,944 enzymes. The non-enzymatic functional groups include 36,621 genes. As a result of the classification processes, detailed compositions of protein machines such as subunits, function specificities, or evolutionary origins can be illustrated. **Table 1** displays the number of subgroups in 10 enzymatic and 6 non-enzymatic functional groups. As indicated in the table, as many as 18 subgroups (H(+)-transporting two-sector ATPase) and 62 subgroups (50S ribosomal protein) can be differentiated in the enzymatic and non-enzymatic function set, respectively. These subgroups, in some cases, represent unique versions of the protein machines. Alcohol dehydrogenase (EC 1.1.1.1) is one of the examples (**Table 2**). It has been categorized into three distinct groups, which represent different versions of the enzyme. The first version of the enzyme is short-chain alcohol dehydrogenase; all of its 50 proteins come from the Eukarya life domain. The second version is zinc-containing alcohol dehydrogenase; its 81 proteins occur in all three life-domains (Archaea, Eubacteria, and Eukarya). The last version is iron-containing alcohol dehydrogenase; 3 of its proteins occur in Eubacteria and 2 in Eukarya. In some other cases, the subgroups represent subunits of functional protein complex. For example, H(+)-transporting two-sector ATPase (3.6.3.14) is a multi-subunit and non-phosphorylated enzyme involved in ion transport. Eighteen subgroups were determined for the enzyme (**Table 2**). This enzyme occurs in mitochondria, chloroplasts, and Eubacteria. The subgroups correspond to proteins in different sectors in the cells ,such as a membrane sector (F(o), V(o), A(o)) and a cytoplasmic-compartment sector (F(1), V(1), A(1)). Subgroups also include the enzymes that operate in a rotational mode, and the extra-membrane sector (containing 3 alpha- and 3 beta- subunits) is connected via the delta-subunit to the membrane sector by several smaller subunits (Friedl et al. 1983;

Capaldi and Aggeler 2002). Some of these subgroups are universal in species distribution. Genes in subgroups defined by protein families IPB003255, IPB000194, IPB002843, IPB002842, IPB002699, and IPB000454 code beta subunit, alpha and beta subunit, C/AC39 subunit, E/31 kDa subunit, subunit D, and subunit C of ATP synthase, respectively. Other subgroups are species-specific, such as those defined by vacuolar ATP synthase 16kD subunit signature (PR00122); all thirty-two genes in this subgroup, which code vacuolar ATP synthase 16kD subunits, are eukaryotic. Furthermore, subgroups can represent subunits and also the types of protein complex. The protein translation elongation factor, as its name suggests, involves a very important step of protein synthesis. Eight subgroups were differentiated. Genes in the protein family IPB001662 and IPB001326 code for gamma chain and beta/beta/delta chain for elongation factor 1 in Archaea and Eukarya, respectively. Genes in the other six subgroups code different types of elongation factors**.** Among them, GTP-binding elongation factor is universal; it occurs in all three life domains plus chloroplasts, cynalles, and mitochondria. In summary, protein functional groups represent detailed, but universal, functional descriptions that can be applied easily in different computational tools and to different genomes. Therefore, this information can be used as a global reference system to coordinate function annotations and genomewide comparative analysis.

Evolutionary spaces of functional groups can accurately describe the overall picture of gene evolution.

Different protein families have diverged to a different extent in the course of evolution Therefore, it is essential to accurately define unique evolutionary spaces for each specific protein functional group. For this purpose, we analyzed the protein functional groups with Blocks, Blast, and Pfam. For Blast data, we first built scoring matrices, then calculated the variation of conservation within protein functional groups. For Blocks data, in addition to this calculation, we also determined the patterns of Blocks motifs and their distribution among different organisms and organelles as defined by the Swiss-Prot database (Archaea, Eubacteria, Eukarya, viruses and phages, chloroplasts, cyanelles, mitochondria, and plasmids) (http://www.expasy.ch/sprot/sprot-top.html). For Pfam data, we used Pfam to analyze the occurrences of biologically significant functional domains. **Figure 1** illustrates the conservation distribution of enzymatic protein functional groups in Blocks patterns and Blast global sequence similarity. The conservation is measured by the coefficient of variation (C.V.). C.V values from 0% to 10% indicate that sequences are consistent in Blocks pattern and highly conserved in sequences. C.V. values close to or greater than 50% indicate that the proteins in these protein groups are under great pressure to evolve. A total of 10,604 protein functional groups have been analyzed, which represent protein groups with a median E-value less than 1e-04. Panel A represents protein functional groups with median E-values from 1e-04 to 1e-20, Panel B corresponds to those with median E-values from 1e-20 to 1e-70, and Panel C, those with median E-values less than 1e-70. As indicated in the figure, some protein functional groups are highly conserved while others are quite variable. Compared with Blocks data (the triangle curves in **Fig. 1**), the C.V. curves (the oval curves in **Fig. 1**) in Blast are much flatter, especially in protein functional groups with low and medium ranges of conservation (Panel A and Panel B in **Fig. 1**). The results indicate much greater variation within protein functional groups in their global sequence similarity. This variation is not a surprise because the Blocks search focuses only on the conserved fragments of protein families so that phylogenetically distant homologous relationships can be detected with higher confidence. **Table 3** lists two extreme categories of variations that occurred in protein functional groups based on Blocks analysis. The first category includes genes encoding methylmalonate-semialdehyde dehydrogenase (1.2.1.27), Photosystem 44 kDa reaction center protein, photosystem D2 protein, photosystem P680 chlorophyll A apoprotein, and preprotein translocase. These functional groups are highly consistent in Block patterns and have extremely conserved domains. They exist in all varieties of life domains and may represent some of the core components in living organisms with stringent system requirements (Fraser et al. 2002). The second category includes viral genes that encode RNA-directed RNA polymerase, apoptosis inhibitors, and nucleocapsid proteins. These protein groups are extremely variable in both sequence conservations and Blocks patterns. They probably represent those required for adaptation and evolution (Reischl et al. 2001; Liu et al. 2002). This property is critical for pathogens such as infectious bacteria and viruses, in which constant change and adaptation are required for surviving various defense systems in their host.

Figure 1.

Rules in the knowledge base determine unique relationships between sequence signatures and cellular functions.

Blocks protein families either can uniquely determine cellular functions or can be shared by multiple cellular functions. **Figure 2** presents overall distribution of relationships between Blocks protein families and cellular functions. As indicated in the figure, the relationships show tremendous variation. For some of the protein families. the function determinations are ambiguous. For example, IPB001395, which codes for aldo-keto reductase family, can define over 24 different functionalities including a number of related monomeric NADPH-dependent oxidoreductases, such as aldehyde reductase, aldose reductase, prostaglandin F synthase, xylose reductase, and rho crystalline. All share a similar structure, with a beta-alpha-beta fold characteristic of nucleotide binding proteins; the same Pfam patterns, with up to three copies of aldo_ket_red domains for catalytic function; and undistinguishable Blocks patterns. In contrast, other Blocks protein families are function-specific: their occurrences are always associated with given functions. For example, IPB001006, the Lysyl hydrolase proton family, is unique to Procollagen lysine 5-dioxygenase (EC 1.14.11.4), and the protein family_IPB000682 is protein family that unique to protein-L-isoaspartate(D-aspartate) O-methyltransferase (EC 2.1.1.77).

Figure 2.

Four categories of relationships are identified between cellular functions and Blocks protein families (**Table 4**). The first category is “one to one,” in which the protein family and only the protein family determine some given cellular functions. Approximately 3% (321) of 10,604 protein functional groups are in this category. Among them, 126 are catalytic, in which proteins have EC number assignments. Some members of enzymatic proteins in this category are monomers, or enzymes with multiple homogeneous subunits. Other protein members are enzymes with heterogeneous subunits, but only one of these has been defined in the Blocks database. The remaining 490 protein functional groups are proteins such as 33 kDa chaperonin, acyl carrier protein, and melatonin receptor, which have non-catalytic cellular functions, or enzymatic proteins such as intron maturase, colipase, and fumarate reductase, which do not yet have EC number assignments in the Swiss-Prot database.

The second category defines relationships between cellular functions and Blocks protein families as "one to many". In this category, Blocks protein families can uniquely determines cellular functions despite of the fact that other protein families can also determine these functions. Members in this category are proteins with heterogeneous subunits such as DNA-directed DNA polymerase, DNA-directed RNA polymerase, photosystem (**Table 1**) and ATP synthase or with different cofactors such as protein kinases (**Table 2**). Members may also include proteins that resulted from convergent evolution such as DNA-directed DNA polymerase, which catalyzes DNA-template-directed extension of the 3-end of an RNA strand one nucleotide at a time**.**

The third category is "many to one". In this category, single protein families determine multiple functions. Over half (5,555) of the protein functional groups belong to this category, which involves 24,857 Swiss-Prot genes. The possible members are proteins that have evolved from functional recruitment or subfuctionalization and enzymes with non-enzymatic homologues or undefined enzymatic homologues or vice versa.

The last category is "many to many," which includes 3,729 protein functional groups. In this category, protein functions are determined by multiple protein families, which may be due to the combined factors as indicated in categories 2 and 3. For example, again, DNA-directed DNA polymerase and DNA-directed RNA polymerase are proteins that require heterogeneous subunits for their functions and originated from convergent evolution.

Compared with Blocks data, Pfam domain patterns have broader coverage of cellular functions and much higher resolution in the identification of protein functional groups. Pfam patterns detect 13,803 significant protein functional groups where Blocks data covers only 10,604. **Figure 3** illustrates the distribution of relationship between Pfam pattern and cellular functions. Of the Pfam patterns, 14% (1,899) can be used as unique identifiers for protein functional groups in which Pfam patterns are unique to functions. The majority of the Pfam patterns, however, are ambiguous in functional identification. For example, 7tm_1 is required for over 100 protein functional groups.

Figure 3.

In summary, some relationships between protein signatures (defined by Blocks and Pfam) and cellular functions are very specific; in this case, these signatures can be used as unique functional identifiers. On the other hand, the majority of the relationships are complex. In these cases, Pfam patterns and Blocks protein families can be ambiguous in functional identifications, demonstrating the limitation of these computational tools.

In addition to the relationship, we also determined the minimal Pfam domain requirements for each functional group. **Table 5** lists examples of Pfam domains that are required for cellular functions. Two domains, IGPS and PRAI, are required for a multifunctional enzyme, tryptophan biosynthesis protein TRP1, that includes indole-3-glycerol phosphate synthase (EC 4.1.3.27 EC 4.1.1.48). An additional GATase domain is required for anthranilate synthase component II (EC 4.1.3.27 EC 4.1.1.48 EC 5.3.1.24). While these Pfam patterns are able to determine unique cellular function for the proteins, others are required for multiple functional groups. Mtap_PNP is Pfam pattern with a single domain. It is required for catalytic functions of purine-nucleoside phosphorylase in Eubacteria and Eukarya and 5-methyl-thio-adenosine phosphorylase in Eukarya. It is also the minimal domain requirement for the multicopy enhancer of UAS2, a non-enzymatic protein.

DISCUSION

The accumulation of amounts of sequence data and information regarding regulatory and metabolic networks, at unprecedented speed and massive scale, presents researchers, as never before, new opportunities and challenges as well. The principal opportunity is that, if proper computational tools for data mining and interpretation are developed, we can gain a much better understanding of the systematic behavior of living organisms. The principal challenge is that all this data is extremely heterogeneous in both data format and reliability. The enormous amount of data and the variability of that data make manual data curation nearly impossible. Therefore, computational tools have been independently developed for sequence data analysis. Each of the tools is able to capture particular features of proteins. The problem is how to integrate these tools to get consistent and highly confident function annotations. To this end, we have developed a knowledge base.

The classification of protein functional groups is one of the important features of our system, which, as indicated above, can differentiate components or versions of proteins. However, not all proteins can be classified into protein families. **Table 6** gives a partial list of protein functions that have no significant Blocks family assignments. H(+)-transporting two-sector ATPase is one of the examples. This enzyme has 75 genes that are not included in any of its 18 subgroups; they distribute over all the life domains. DNA-directed DNA polymerase (EC 2.7.7.6) is another example. Fourteen different subgroups have been determined for this function, with 30 genes that do not have the assignments of protein families. The possible reason is that Blocks database does not have enough coverage for protein cellular functions. The database has not been updated since its last release in August 2001. In addition, current Blocks families cannot provide enough differentiation power for some closely related protein functions. For example, IPB002328, a protein family coding for zinc-containing alcohol dehydrogenase, corresponds to over 20 different cellular enzymatic functions. All these functions share the same Blocks patterns, indicating that this computational tool lacks the ability to capture unique features for each of these functions.

Another unique feature in our knowledge base is the calculation of evolutionary spaces for the protein functional groups. Determination of evolutionary spaces for each protein functional group is an essential, but often ignored, step in large-scale functional annotation systems. The reason is that the quality of the predictions largely depends on the degree of sequence homology in distantly related protein families and overlapping of closely related protein families. In the course of evolution, different protein families have diverged to a different extent. Therefore, flat cutoff scores, commonly used to separate clusters belonging to different protein families, cannot provide reliable separations. In contrast, the evolutionary spaces, which describe the current evolutionary status of protein functional groups, can provide dynamic separation lines and excellent insight for deciding functions for individual proteins. This unique feature has been integrated into our high-throughput genetic sequence analysis system WIT3 at Argonne National Laboratory (http://www-wit.mcs.anl.gov/wit3). The knowledge base provides all necessary measurements for determining protein functions. The measurements used are as follows. Is the target gene within the evolutionary spaces of the functions? Is the protein family identified for the gene unique to the function? Are the Pfam domain(s) satisfied the minimal domain requirements defined for this function? Is the gene function definition consistent with species distribution for this group? Answering these questions will lead to enhanced annotation confidence and the prevention of over-interpretation. If conflicts occur, a voting strategy (Yu et al. unpublished data) will be applied to determine the most likely functions according to the rules in the knowledge base, thereby resolving conflicts among different computational tools.

In addition, with our knowledge base we can determine the version for the target genes, based on the species distribution of Blocks patterns and Pfam domains in defined functional groups. Pectate lyase (EC 4.2.2.2/PR00807) is an example. This group of genes encodes proteins that function in eliminative cleavage of pectate. The reaction gives oligosaccharides with 4-deoxy-alpha-D-gluc-4-enuronosyl groups at their nonreducing ends (Tamaru and Doi 2001). As indicated in **Table 3**, this protein functional group consists of 16 eukaryotic genes and 6 eubacterial genes and is highly variable both in Blocks patterns and in the degree of conservation. The Blocks patterns and Pfam domain requirements, on the other hand, are well separated between different species. There are four Blocks patterns (DEG, CDE, CD, DE) for Eubacteria and three patterns (BCDE, ABCDEFGH, ADE) for Eukarya. At least one Pfam domain (pec_lyase) is required. In our automatic annotation system, two genes**,** gi|2633080 of *Bacillus subtilis* and gi|4980940 of *Thermotoga maritime,* are assigned as pectate lyase (EC 4.2.2.2). In both genes, EC 4.2.2.2 comprises the best hits by Blast searching with Blocks family of PR00807 (DE) and Pfam domain of pec_lyase. The results indicate that both genes fit well into the evolutionary space of the protein functional group in the knowledge base. Furthermore, both genes can be categorized as version of Eubacteria, since they possess the Blocks DE pattern found only in Eubacteria.

Growing evidence indicates that an increasing number of proteins with apparently different structures may share common ancestors. It is also clear that similar local structures have been reinvented multiple times by so-called convergent evolution (Dodson and Wlodawer 1987; Makarova and Grishin 1999; Ponting and Russell 2002). As indicated above, our knowledge base covers the information about sequences, domains, motifs, and their species distributions for all available cellular functions. This information leads to the differentiation of the detailed composition of protein machines. By analyzing this information for particular functions, we can determine whether they are candidate genes that are involved in convergent evolution. For example, DNA-directed RNA polymerase (EC 2.7.7.6) has fifteen subgroups **(Table 2)**. Each of these subgroups has unique species distribution and corresponds to different subunits, as well as to the different versions of this multiple heterogeneous enzyme. It appears that two separate systems have evolved, one in Archaea and another in Eubacteria, although some components are shared extensively. For example, DNA-directed RNA polymerase genes in the family IPB000684 are universal in all domains of life, which defined subunit A for Archaea, largest subunit for Eukarya and delta chain for Eubacteria. Genes in the family IPB001529 are specific to Archaea (subunit M) and Eukarya (14.5 kda polypeptide), while IPB003716 defines genes that occurred only in Eubacteria (omega chain). Blast analysis discovered no significant sequence similarities between these genes. This result illustrates that they may have separate evolutionary origins. EC 1.1.1.1, alcohol dehydrogenase, is another good example. This enzyme is a homodimer or homotrimers. Protein classification resulted in three significant different protein families (**Table 2**). No significant sequence homologies were detected among these protein families, indicating that they are potential candidates for convergent evolution. Systematically analyzing data in our knowledge base will extract all possible candidates for convergent evolution for current protein functional groups. This information will be invaluable for studying the evolution of metabolic pathways for alternative pathways, which may associate the reinventions of the proteins, their structures and functions.

Acknowledgments

We are grateful to Dr. Natalia Maltsev for the initial idea of creation of a knowledge base and the development of a voting algorithm for automated prediction of gene functions. We also thank Dr. Nagiza Samatova for substantial help in preparation of this manuscript. This work was supported in part by U.S. Department of Energy, under Contract W-31-109-Eng-38.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol.,* **215,** 403-10.

Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., Daruvar, A. de, Reich, C., Franchini. A., Tamames, J., Valencia. A., Ouzounis, C., and Sander, C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15,** 391–412. **http://jura.ebi.ac.uk:8765/ext-genequiz/.**

Attwood, T. K., Croning, M. D. R., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., Selley, J. N., and Wright, W. (2000) PRINTS-S: The database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bairoch, A., and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28,** 45–48.

Barker, W. C., Garavelli, J. S., Hou, Z., Huang, H., Ledley, R. S., McGarvey, P. B., Mewes, H., Orcutt, B. C., Pfeiffer, F., Tsugita, A., Vinayaka, C. R., Xiao, C., Yeh, L. L., and Wu, C. (2001) Protein Information Resource: A community resource for expert annotation of protein data. *Nucleic Acids Res.,* **29,** 29–32.

Bateman, A., Birney, E., Cerruti, L. Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E.L. L. (2002) The Pfam protein families database. *Nucleic Acids Res.,* **30,** 276–280.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2002) GenBank. *Nucleic Acids Res.*, **30,** 17–20.

Capaldi, R. A., and Aggeler, R. (2002) Mechanism of the F(1)F(0)-type ATP synthase, a biological rotary motor. *Trends Biochem Sci.,* **27,** 154–60.

Casari, G., de Daruvar, A., Sander, C., and Schneider, C. R. (1996) Bioinformatics and the discovery of gene function. *Trends in Genetics,* **12,** 244–245.

Collins, K, and Mitchell, J. R. (2002) Telomerase in the human organism. *Oncogene,* **21,** 564–579.

Csete, M E., and Doyle, J. C. (2002) Reverse engineering of biological complexity. *Science*, **295,** 1664-1669.

Dodson, G., and Wlodawer, A. (1998) Catalytic triads and their relatives. *Trends Biochem. Sci.*, **23,** 347–352

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30,** 235–358.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002) Evolutionary rate in the protein interaction network. *Science,* **296,** 750–752.

Friedl, P., Hoppe, J., Gunsalus, R. P., Michelsen, O., von Meyenburg, K., and Schairer, H.U. (1983) Membrane integration and function of the three F0 subunits of the ATP synthase of Escherichia coli K12. *EMBO J.*, **2,** 99–103.

Frishman, D., and Mewes, H. (1997) Pedantic genome analysis. *Trends Genet.,* **13,** 415–416. http://pedant.gsf.de/.

Gaasterland, T., and Sensen, C. (1996) Fully automated genome analysis that reflects user needs and preferences—a detailed introduction to the MAGPIE system architechure. *Biochimie,* **78,** 302–310. http://genomes.rockefeller.edu/magpie/.

Gregory, V., Kryukov, R., Kumar, A., Koc, A., Sun, Z., and Gladyshev**,** V. N. (2002**)** Selenoprotein R is a zinc-containing stereo-specific methionine sulfoxide reductase *Proc. Natl. Acad. Sci. USA*, **99,** 4245**–**4250.

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

Henikoff, S., and Henikoff, J. G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.,* **19,** 6565–6572.

Hofmann, K., Bucher, P., Falquet,L., and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

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

Liu, G. R., Rahn, A., Liu, W. Q., Sanderson, K .E., Johnston, R. N., and Liu, S. L. (2002) The evolving genome of *Salmonella enterica Serovar Pullorum*. *J Bacteriol.,* **184,** 2626–2633.

Lynch, M., and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics,* **154,** 459–73

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

Makarova, K. S., and Grishin, N. V. (1999) Thermolysin and mitochondrial processing peptidase: how far structure-functional convergence goes. *Protein Sci.* **8,** 2537–2540.

Mardulyn, P., Milinkovitch, M. C., and Pasteels, J. M. (1997) Phylogenetic analyses of DNA and allozyme data suggest that Gonioctena leaf beetles (Coleoptera; Chrysomelidae) experienced convergent evolution in their history of host-plant family shifts. *Syst Biol*., **46,** 722–747.

Massingham, T., Davies, L. J., and Lio, P. (2001) Analysing gene function after duplication. *Bioessays*, **23,** 873–876.

Naumann, C., Hartmann, T., and Ober, D. (2002) Evolutionary recruitment of a flavin-dependent monooxygenase for the detoxification of host plant-acquired pyrrolizidine alkaloids in the alkaloid-defended arctiid moth Tyriajacobaeae. *Proc. Natl. Acad. Sci. USA,* **99,** 6085–6090.

Peitsch, M. C. (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans*., **24**:274-279.

Ponting,C.P., and Russell,R.R. (2002) The natural history of protein domains, *Annu. Rev. Biophys. Biomol. Struct.,* **31,** 45-71.

Reischl, A., Reithmayer, M., Winsauer, G., Moser, R., Gosler, I., and Blaas, D. (2001) Viral evolution toward change in receptor usage: adaptation of a major group human rhinovirus to grow in ICAM-1-negative cells. *J. Virol*., **75,** 9312-9319.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., and Sander, C.. (1994) GeneQuiz: A workbench for sequence analysis. In: *Proc. Second International Conference on Intelligent Systems for Molecular Biology*, edited by R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls. AAAI Press, Menlo Park, California, 348–353.

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

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J (2001) WormBase: Network access to the genome and biology of *Caenorhabditis elegans. Nucleic Acids Res.,* **29,** 82–86.

Shimamoto,K., Kyozuka,J. (2002) Rice as a model for comparative genomics of plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol*., **53,** 399–419.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M. A., Tzouvara, K., and Vaughan, R. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.,* **30,** 21–26.

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.,* **30,** 13–16.

Tamaru, Y, and Doi, R.H. (2001) Pectate lyase A, an enzymatic subunit of the Clostridium cellulovorans cellulosome. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 4125-9.

Van de Peer, Y., Taylor, J. S., Braasch, I., and Meyer, A. (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.,* **53,** 436–446.

Xiong, J., and Bauer, C. E. (2002) Complex evolution of photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol*., **53,** 503–521.

Figures

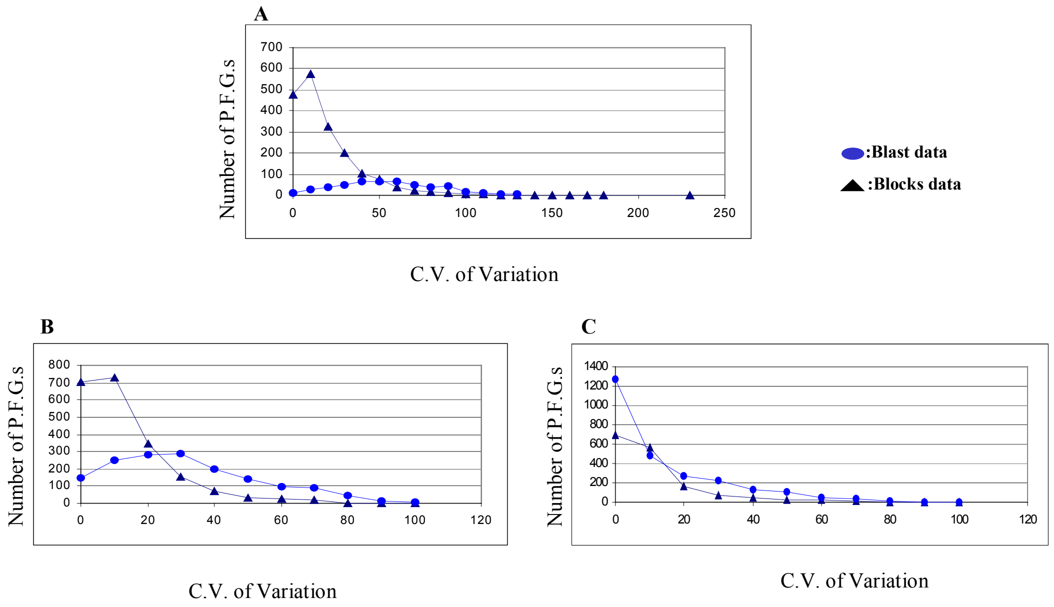

Fig.1. Distribution of conservation of protein sequences in three classes of all protein functional groups (P.F.G.s). The conservation is measured by Co-efficiency of Variation (C.V.) in conserved Blocks motifs (the triangle curves) and global sequence similarities (the oval curves). The C.V.s are represented in the x-coordinate. The y-coordinate indicates the number of protein function groups. Panel A represents protein functional groups with E-value greater than 1e-20, Panel B with E-values from 1e-20 to 1e-70, and Panel C with E-value less than 1e-70.

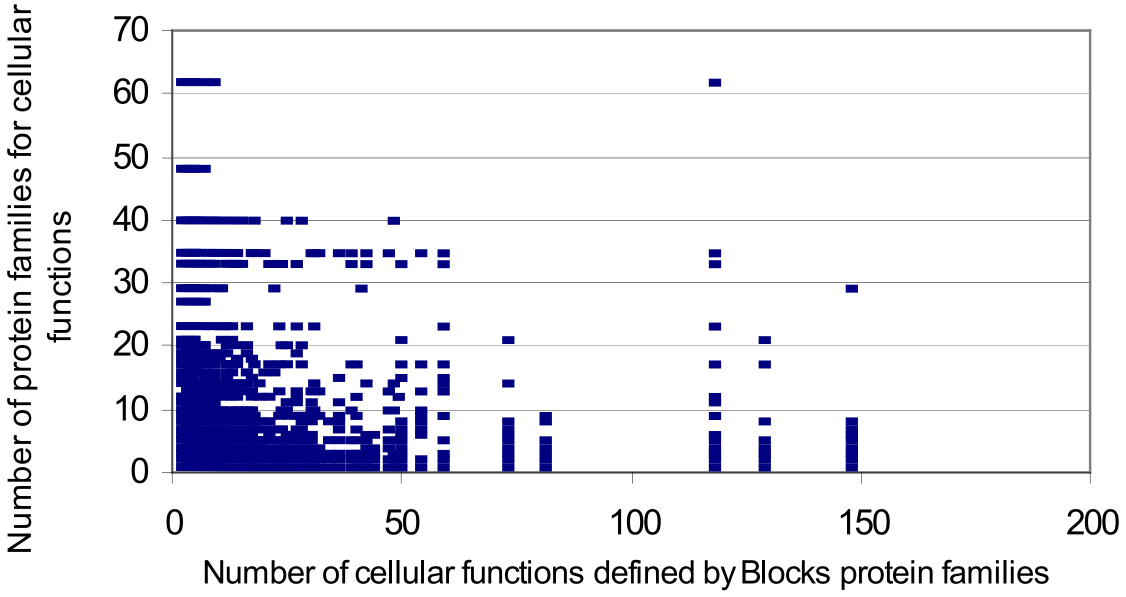

Fig. 2. The two-way relationships between protein families and cellular functions. The relations of the protein families and functions Include one to one and many to one, in which the protein family uniquely determines the functions. The relations also includes many to one and many to many, in which the relationships are very complex and additional evidence is needed for functional determination if these Information is to used for protein function annotations.

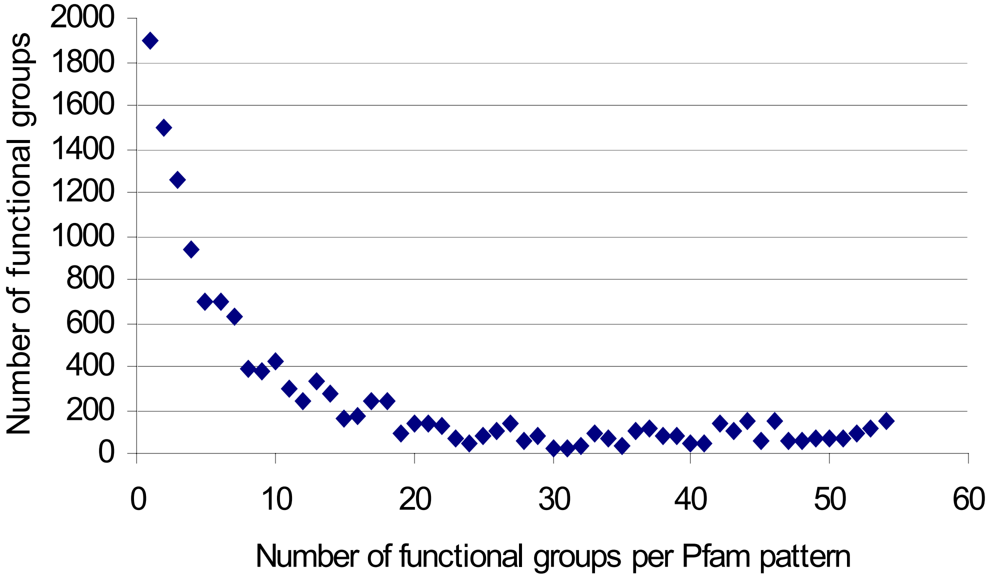

Fig. 3. The relationship between Pfam domain patterns and cellular functions. The relations of the domain patterns and the functions include one to one, in which Pfam patterns uniquely determine the cellular functions. The relations also includes one to many, in which single Pfam domain patterns are corresponding to many functions.

Tables

Table 1. Partial list of enzymatic and non-enzymatic functions in which subgroups can be clearly differentiated by the protein classification procedure

Function Description	Function	Number of Subgroups
H(+)-transporting two-sector ATPase	3.6.3.14 (heteromultimer)	18
Protein kinase	2.7.1.37 (homodimer)	17
DNA-directed RNA polymerase	2.7.7.6 (heteromultimer)	14
Cytochrome-c oxidase	1.9.3.1 (heteromultimer)	11
DNA-directed DNA polymerase	2.7.7.7 (heteromultimer)	9
1-phosphatidylinositol 3-kinase	2.7.1.137(heterodimer)	4
NAD(P)(+) transhydrogenase (AB-specific)	1.6.1.2 (heterodimer)	3
tRNA (5-methylaminomethyl-2-thiouridylate) methyltransferase	2.1.1.61	3
Alcohol dehydrogenase	1.1.1.1 (homodimers or tetramer)	3
Mercury (II) reductase	1.16.1.1(homodimer)	3
50S_ribosomal_protein	Non-enzymatic	62
60S_ribosomal_protein	Non-enzymatic	49
Transcription_factor	Non-enzymatic	34
Transcriptional_activator	Non-enzymatic	9
Replication_protein	Non-enzymatic	9
Elongation_factor	Non-enzymatic	8

Table 2. Subcategories of ATP synthase and their species distribution

Function	Subgroup	Protein Family	Function Description	Species Distribution[a]
ATP synthase	1	IPB003255	ATP synthase beta subunit, C-terminal	A[b]:14 B[c]:3 E[d]:23
	2	IPB000790	ATP synthase alpha subunit, C-terminal	B:37 E:10 chl[e]:17 cynal[f]:1 mit[g]:15
	3	IPB000194	ATP synthase alpha and beta subunit, N-terminal	A:14 B:78 E:42 chl:28 cynal:1 plasm[h]:4
	4	IPB000131	ATP synthase gamma subunit	B:31 E:17
	5	PR00122	Vacuolar ATP synthase 16kD subunit signature	B:1 E:32
	6	IPB002843	ATP synthase (C/AC39) subunit	A:9 B:3 E:11
	7	IPB002842	ATP synthase (E/31 kDa) subunit	A:9 B:2 E:16
	8	IPB002699	ATP synthase subunit D	A:9 B:8 E:13
	9	IPB000568	ATP synthase A subunit	B:30 E:5 chl:16 mit:88
	10	IPB002490	V-type ATPase 116kDa subunit family	A:1
	11	IPB003238	Mammalian mitochondrial ATPase subunit 8	E:7 mit:34
	12	IPB001421	Mitochondrial ATPase subunit 8	E:19 mit:41
	13	IPB000454	EuBacterial and plasma membrane ATP synthase subunit C	A:1 B:34 E:18 chl:15 cynal:1 mit:20
	14	IPB001469	ATP synthase, Delta/Epsilon chain	B:62 E:15 chl:21 cynal:1
	15	IPB002841	ATP synthase (F/14-kDa) subunit	A:9 B:2 E:11
	16	IPB002146	ATP synthase B/B CF(0)	B:39 E:4 chl:23 cynal:2
	17	IPB000711	ATP synthase, delta (OSCP) subunit	B:31 E:12 chl:6 cynal:1
	18	IPB003445	Cation transport protein	B:1
Alcohol dehydrogenase	1	IPB003030	Short-chain alcohol dehydrogenase family	E:50
	2	IPB002328	Zinc-containing alcohol dehydrogenase	A:1 B:7 E:72 plasm:1
	3	IPB001670	Iron-containing alcohol dehydrogenase	B:3 E:2
Elongation factor	1	IPB001662	Elongation factor 1 gamma chain	E:16
	2	IPB001326	Elongation factor 1 beta/beta/delta chain	A:1 E:26
	3	IPB000640	Elongation factor G, C-terminus	A:17 B:52 E:22
	4	IPB001816	Elongation factor Ts	B:49 E:3 chl:3
	5	IPB001059	Elongation factor P (EF-P)	B:30
	6	IPB000795	GTP-binding elongation factor	A:16 B:68 E:83 chl:11 cynal:1 mit:1
	7	IPB003163	Yeast DNA-binding domain	E:1
	8	IPB001140	ABC transporter transmembrane region	E:6
EC 2.7.7.6	1	IPB000684	Eukaryotic RNA polymerase II heptapeptide repeat	A:21 B:1 E:27 V:1 chl:13 cynal:1
	2	IPB001572	RNA polymerases beta subunit	A:14 B:29 E:22 V:5 chl:15 cynal:1 plasm:1
	3	IPB001700	Bacterial RNA polymerase, alpha chain	B:41 E:2 chl:31 cynal:1 mit:1
	4	IPB000722	RNA polymerase, alpha subunit	B:27 E:5 V:4 chl:13 cynal:1
	5	IPB001514	RNA polymerases D/30 to 40 Kd subunits	A:11 E:10
	6	IPB002092	Bacteriophage-type RNA polymerase family	E:6 V:4 mit:6
	7	IPB003716	RNA polymerase omega subunit	B:20

Table 2. Subcategories of ATP synthase and their species distribution (continued).

8	IPB003221	DNA directed RNA polymerase, 7 kDa subunit	E:4
9	IPB000268	RNA polymerases N/8 Kd subunits	A:10 E:6 V:7
10	IPB001529	RNA polymerases M/15 Kd subunits	A:5 E:7
11	IPB001725	RNA polymerases K/14 to 18 Kd subunits	A:9 B:1 E:6 V:3
12	IPB000783	RNA polymerase H/23 kD subunit	A:11 E:4
13	IPB001306	RNA polymerases L/13 to 16 Kd subunits	A:7 E:12
14	IPB001222	TFIIS zinc ribbon domain	E:2 V:4
15	IPB000135	High mobility group proteins HMG1 and HMG2	E:4 V:1

[a]The species distribution illustrates where the genes in a protein functional group are located among different organisms and organelles as defined by Swiss-Prot database; [b]A represents for Archaea; [c]B for EuBacteria; [d]E for Eukarya; [e]chl for chloroplast; [f]cynal for cyanelle; [g]mit for mitochondrion; and [h]plasm for plasmid.

Table 3. Examples of characterization of protein functional groups based on Blocks searching results

Protein Functional Group		**Characterization of protein population within Protein Functional Groups**					
Function Category	**Protein Family**	**Lowest CF**[a]	**Upper CF**	**C.V.**[b]	**SP_DIS**[d]	**PA_DIS**[d]	**PAT_SP**[e]
		Highly conserved protein functional groups					
E.C.1.2.1.27	IPB002086	5.2e-52	5.7e-58	3.54	B:2 E:4	ABCDE:6	ABCDE:B:E
E.C.1.2.1.9	IPB002086	3.7e-62	9.8e-70	4.19	B:1 E:3	ABCDEF:4	ABCDEF:B:E
E.C.2.7.1.48	PR00988	3e-51	7e-57	3.10	B:6 E:1	ABCDEF:7	ABCDEF:B:E
E.C.1.18.96.1	IPB002742	3.3e-33	3.4e-43	10.05	A:3 B:1	ABC:4	ABC:A:B
Hydrogenase expressionformation protein	PF01924	9.8e-199	2.6e-227	4.88	A:1 B:5 plasm:1	ABCDEFG:7	ABCDEFG:A:B:plasm
Photosystem 44 kDa reaction center protein	IPB000932	1.5e-206	3.2e-242	4.18	B:3 E:2 chl:18 cynal:1	ABCDEF:25	ABCDEF:B:E:chl:cynal
Photosystem D2 protein	IPB000484	5.9e-143	2.1e-160	2.82	B:4 E:4 chl:17 cynal:1	ABCD:28	ABCD:B:E:chl:cynal
Photosystem P680 chlorophyll A apoprotein	IPB000932	8.3e-216	1.1e-257	4.63	B:4 E:1 chl:19 cynal:1	ABCDEF:25	ABCDEF:B:E:chl:cynal
Preprotein translocase	IPB000185	1e-250	1.3e-281	3.10	B:30 E:2 chl:7	ABCDEFGHIJ:39	ABCDEFGHIJ:B:E:chl
		Highly diverged protein functional groups					
E.C.1.2.99.5	IPB002489	0.00025	8.8e-133	146.04	A:8	ABCDEF:2 BC:5 ABC:1	ABCDEF:A BC:A ABC:A
E.C.4.2.2.2	PR00807	1.5e-05	1.7e-131	145.42	B:16 E:6	DEG:1 BCDE:1 ABCDEFGH:4 CD:1 DE:12 ADE:1 CDE:2	DEG:B BCDE:E ABCDEFGH:E CD:B DE:B ADE:E CDE:B
E.C.4.2.99.9	IPB000277	2.3e-08	5e-115	56.28	B:6 E:4	ABCDEF:7 BCDF:1 BCF:2	ABCDEF:B:E BCDF:E BCF:E
E.C.1.6.99.3	IPB000103	6.5e-05	5.7e-80	65.19	A:1 B:2	AE:1 ABCDE:2	AE:A BCDE:B
E.C.4.1.1.23	IPB001754	0.28	8.4e-119	71.97	A:5 B:13 E:32	ABE:5 ABCDEF:32 BCE:1 E:2 AE:3 BE:4 ABCE:2 CE:1	ABE:A:B ABCDEF:E BCE:B E:B AE:A BE:B ABCE:A:B CE:B
E.C.2.7.7.48	IPB000224	4e-14	1.4e-135	73.49	V:7	ABCDE:4 ABE:3	ABCDE:V ABE:V
Apoptosis inhibitor	IPB001370	0.0022	2.4e-65	75.10	E:2 V:7	C:2 AC:1 ABC:6	C:V AC:V ABC:E:V
Metallothionein	PR00873	0.0061	4.9e-64	135.30	E:21	D:15 AD:1 ABCD:5	D:E AD:E ABCD:E
Glycoprotein	PR00668	1e-14	9.2e-100	64.69	V:11	ABCDEFG:3 ACDF:5 ABCDEF:1 ABCDFG:2	ABCDEFG:V ACDF:V ABCDEF:V ABCDFG:V
High mobility group protein	IPB000910	0.006	1.2e-148	58.23	E:14 V:1	ABC:11 BC:3 C:1	ABC:E BC:E:V C:E

Table 3. Examples of characterization of protein functional groups based on Blocks searching results (continued).

Methyl-accepting chemotaxis protein	IPB000122	1.2e-22	1.6e-194	61.70	B:9	ABCDEFG:5 BCEF:1 ABDEG:2 BCDEG:1	ABCDEFG:B BCEF:B ABDEG:B BCDEG:B
Nitrite extrusion protein	BP04821	1.4e-18	1.8e-202	61.69	B:3	ABCDEF:2 BDE:1	ABCDEF:B BDE:B
ATP-dependent helicase	IPB000629	0.0042	4.4e-73	111.39	A:3 B:4 E:3	A:1 AC:3 AD:1 AE:2 ABCDE:3	A:B AC:A:B AD:B AE:A ABCDE:E
Metallothionein	PR00872	0.0039	1.6e-34	114.63	E:28	A:9 B:15 AB:1 ABC:3	A:E B:E AB:E ABC:E
Nucleocapsid protein	BP03484	1.8e-13	1e-250	85.98	V:20	CFG:1 BCFG:5 CG:2 BCE:1 ABCDEFGHI:10 BCG:1 ABCDEF:1	CFG:V BCFG:V CG:V BCE:V ABCDEFGHI:V BCG:V ABCDEF:V
Outer membrane protein	IPB001702	0.00023	1.2e-198	71.93	B:22 V:1	ACFG:1 ABCDEFGH:14 BCEH:1 AF:5 ACF:1	ACFG:B ABCDEFGH:B BCEH:B AF:B ACF:B

[a.] C.F. represents the level of confidence measured by E-value; [b.] C.V. indicates the co-efficiency of variations; [c.] SP_DIS represents the species distribution of genes within protein functional groups; [d.] PA_DIS represents the number distribution of Blocks patterns; [e.] PAT_SP represents the species distribution of Blocks patterns; [f.] the ratios of Blocks: denominator indicates total number of Blocks for protein families and nominator the number of Blocks detected for individual genes.

Table 4. Categories of relationships between Blocks protein families and cellular functions

Category	Protein Group	
	Number of Groups	Number of Genes
1[a]	321	4045
2	490	8461
3	5555	24857
4	4238	29818

[a.] The number indicates the categories of unique relationships between Blocks protein families and cellular function: 1 for one to one (cellular functions are determined by one and only by one protein family); 2 for one to many (one protein family uniquely determines a cellular function, but this function is also defined by other protein families); 3 for many to one; and 4 for many to many.

Table 5. Unique relationships between Pfam patterns and protein functional groups

Pfam Pattern	Protein Functional Group		Species Distribution[a]
	Function Description	Protein Family	
IGPS PRAI	E.C.4.1.1.48, E.C.5.3.1.24	IPB001468	B E
GATase IGPS PRAI	E.C.4.1.3.27 E.C.4.1.1.48 E.C.5.3.1.24	IPB001468	E
GATase_2 Asn_synthase	E.C.6.3.5.4	IPB001962	A[b] B[c] E[d]
GATase_2 GATase_2	Glutamine amidotransferase	IPB000583	A B
GATase_2 GATase_2 Pribosyltran	E.C.2.4.2.14	IPB000583	B
GATase_2 Pribosyltran	E.C.2.4.2.14	IPB000583	A B E
GATase_2 SIS SIS	E.C.2.6.1.16	IPB000583	A B E plasm
Mtap_PNP	E.C.2.4.2.-	IPB001369	B
	E.C.2.4.2.1	IPB001369	B E
	E.C.2.4.2.28	IPB001369	E
	Multicopy_enhancer_of_UAS2	IPB001369	E

[a]The species distribution illustrates where the genes in a protein functional group are located among different organisms and organelles as defined by Swiss-Prot database; [b]A for Archaea; [c]B for Eubacteria; [d]E for Eukarya; and [e]plasm for plasmid.

Table 6. Partial list of enzymatic and non-enzymatic proteins in which no significant protein families can be assigned.

Function Description	**Number of Subgroups**	Number of Genes **without Blocks Families**	Species Distribution[a]
H(+)-transporting two-sector ATPase	18	**75**	A[b]: 3 B[c]:8 E[d]:56 mit:8
Transcriptional_activator	10	25	B:4 E:17 cynal[e]:1 chl[f]:3
DNA-directed RNA polymerase	14	30	A:3 B:1 E:13 V:13
Cytochrome-c oxidase	11	24	B:5 E:19
DNA-directed DNA polymerase	9	50	A:7 B:18 E:11 V:8 mit[g]:4 Plasm[h]:1
Transcription_factor	31	19	E:18 B:1
Replication_protein	7	9	B:2 E:6 V:1
Elongation_factor	8	5	A:5

[a] The species distribution illustrates where the genes in a protein functional group are located among different organisms and organelles as defined by Swiss-Prot database; [b] A for Archaea; [c] B for Eubacteria; [d] E for Eukarya; [e] cynal for cyanelle; [f] chl for chloroplast; [g] mit for mitochondrion; [h] V for viruses and phages; and [h] plasm for plasmid.