**A Knowledge-Based Voting Algorithm for Automated Protein Functional Annotation**

Yu, GX[1] *, Glass, EM[1], Karonis, NT[1,2], and Maltsev, N[1]

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

[2]Department of Computer Science
Northern Illinois University, DeKalb, IL 60115, USA

To whom correspondence should be addressed:
Yu07_2000@yahoo.com
*current address: Virginia Bioinformatics Institute,
Washington Street,
Blacksburg, VA 24061, USA
Phone: 540 231 4226

Key words: protein function prediction, knowledge system, protein function groups, rules, voting procedure, alternative functional assignments

**ABSTRACT**
Automated annotation of high-throughput genome sequences is one of the earliest steps toward a comprehensive understanding of the dynamic behavior of living organisms. However, the step is often error-prone due to its underlying algorithms, which rely mainly on a simple similarity analysis and lack guidance from biological rules. We present here a knowledge-based protein annotation algorithm. Our objectives are to reduce errors and to improve annotation confidences. This algorithm consists of two major components: a knowledge system, called "RuleMiner," and a voting procedure. The knowledge system, which includes biological rules and functional profiles for each function, provides a platform for seamless integration of multiple sequence analysis tools and guidance for function annotation. The voting procedure, which relies on the knowledge system, is designed to make (possibly) unbiased judgments in functional assignments among complicated, sometimes conflicting, information. We have applied this algorithm to ten prokaryotic bacterial genomes and observed a significant improvement in annotation confidences. We also discuss the current limitations of the algorithm and the potential for future improvement.

## INTRODUCTION

The number of completely sequenced genomes has dramatically increased during the past several years, and this momentum is likely to continue. At the time of this writing, Genomes OnLine Database (GOLD) lists over 211 completely sequenced genomes [1]. Additional 522 prokaryotic genomes and 436 eukaryotic genomes are listed as ongoing sequence projects. Knowledge about protein components, functional capacities, and overall metabolic potentials of these genomes will advance progress toward a comprehensive understanding of the genetic mechanisms of diverse biochemical processes [2]. The challenge is that the experiments needed to determine biological functionalities for the composed gene components in the sequenced genomes are labor-intensive and very expensive.

In an effort to complement such experiments, several computational approaches have been developed to automate the annotation processes [3-6]. These approaches, however, are often error-prone because their underlying algorithms rely mainly on a BLAST or FASTA-based sequence similarity analysis[7]. In contrast, the diversity of cellular functions has created complicated and unpredictable sequence-function relationships [8]. Evolutionary processes add further complexities [9, 10]. Consequently, the similarity analysis cannot always provide relevant relationships between functions and sequences [7, 11]. The resulted annotations are difficult to interpret and error-prone, and annotation confidences are hard to evaluate [12].

Current bioinformatics research offer a variety of sequence analysis tools and each of them addresses different problems and has its unique features and capability[13-16]. It is thus essential to integrate these tools to achieve an enhanced computational capacity for recognizing and differentiating cellular functions. All these tools, however, have been independently developed and have resulted in incompatible nomenclatures [16]. As a result, the integration can be enormously difficult and could compromise the efficacy of these tools for the annotating protein function. The lack of clear principles or rules present another challenge [7, 11, 17], especially where multiple sequence analysis algorithms and heterogeneous biological datasets have to be integrated [9, 10].

Our previous efforts [11] in this direction have focused on developing a knowledge system, called "RuleMiner," for high-throughput genome sequence analysis. The knowledge system consists of three components: protein function groups (PFGs), PFG profiles, and rules. Established from an integrated analysis of the current knowledge in Swiss-Prot database [18] and family-based protein classifications, the PFGs cover all possible cellular functions available in the database. Characterized by sequence conservations (BLAST and BLOCKS), the occurrences of sequence-based motifs (BLOCKS), domains (Pfam), and species distributions, the PFG profiles illustrate detailed protein features for each PFG. The rules, mined from the PFG profiles, describe the clear relationships between the PFGs and all possible features. As a result, the knowledge system can provide an enhanced capability for protein function analysis. For example, the results from sequence analysis tools for given proteins can be comparatively analyzed and much-needed guidance is readily available for such an analysis. If the rules describe unique relationships between the protein features and the PFGs—for instance, one to one and many to one (one or many features to one unique PFG)—then these features can be used as unique functional identifiers and

cellular functions of unknown proteins can be reliably determined. Otherwise, additional information must be provided.

In this paper, we present a high-throughput protein annotation algorithm extended from the knowledge system development. Our goal is to develop an analysis system with a seamless integration of multiple sequence analysis tools, biological rules, and PFG profiles in order to reduce annotation errors, improve confidences. An additional goal is to categorize the annotation confidences and associate them explicitly with specific protein annotations. For these goals, a voting procedure, which relies on rules and the functional profiles in the knowledge system, is developed to make (possibly) unbiased judgments in functional assignments among complicated and sometimes conflicting information from the sequence analysis tools.

The judgments are based on the answers to the following questions: Does the knowledge system have any PFGs corresponding to the target proteins? Are the domains or motifs identified for the proteins unique to these PFGs (rules)? Are the features of the target proteins consistent with the profiles of the PFG candidates? Depending on the answers, we categorize the annotations into different confidence categories, in which annotations that satisfy all these questions would have the highest confidence.

We have applied this algorithm to ten prokaryotic bacterial genomes and observed significantly improved annotation confidences. We believe this algorithm will be a great asset to those interested in using the annotation data. For example, researchers will be able to decide to what degree the annotation data can be trusted and design their experiments accordingly with the genome annotation data and the annotation programs available on request.

## MATERIALS

In this section, we first describe genome sequence data and the procedure for primary genome analysis. We then define three new terms—digit scoring system, annotation confidence category, and protein version. Finally, we illustrate the rule-based algorithm with a detailed description of the analysis procedure and examples.

### Data Preparation

We downloaded ten completely sequenced genomes (Table I) from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). These genomes cover a variety of organisms (2 prokaryote superkindoms, seven phylums, nine classes and ten orders), ranging from genomes that are well studied (e.g., *Escherichia coli K12* and *Bacillus subtilis)* to those that are barely known (e.g., *Halobacterium sp.* and *Aeropyrum pernix*. With such variety, this genome data can be an excellent data set to evaluate the performance of our algorithm.

Table I

### Genome Sequence Data Processing

For the purpose of this research, a parallel process for BLAST, BLOCKS, and Pfam to run on a 512-node Linux cluster was developed at Argonne National Laboratory. A parallel process of this capacity is necessary to provide computational power because of the exceptionally large sequence data and computational time needed for each of the tools. The output of the tools is then processed and stored in an Oracle database. The database design is an important issue in the management of biological data because of its complexity and the exponential growth of related data. However, we will not describe the details of database design and managements here, which are beyond the scope of this paper.

### A Digit Scoring System for BLAST Hits

In the voting algorithm, the E-value is one of the most important criteria for evaluating the sequence similarity in sequence analysis tools. Biological function domains, motifs, and BLAST hits with a lower E-value are more likely the proper function assignments; an E-value of zero represents the highest level of confidence in functional relevance [13]. Comparing results from Pfam [14] and BLOCKS [15] is straightforward because each can give unique assignments of the functional domains or motifs with certain E-values. Comparing the BLAST hits, on the other hand, is difficult because BLAST analysis can give multiple hits. All of them may share the same functional annotations but each of them may associate with a different E-value. To represent protein functions of the BLAST hits, we developed a novel scoring system

(Table II). In this system, eight confidence levels of scores are defined by extending the scoring scheme of GeneQuiz [3]. Two digits represent each confidence level so that maximum number of BLAST hits can accumulate up to 99 without mixing adjacent levels of confidences. We call this a "digit score" in order to differentiate it from the score that is built into the BLAST search. The scheme can be easily extended as needed to increase the capacity of the scoring system.

Table II

**Annotation Confidence Categories**

The confidence category indicates to what extent we can trust the annotation for given target proteins. We established three groups of annotation confidences based on possible combinations between tool-derived protein features and their potential entries in the knowledge system (Table III). Annotations in Groups I and III have a strong support from the knowledge system. A combinatory analysis of the protein features, rules, and PFG profiles can lead to highly confident functional assignments. The difference between the two groups of annotations is that proteins in Group I have unique functional assignments, whereas those in Group III have alternative or multiple functional assignments. Group II annotations, on the other hand, do not have such support, resulting in low confidence. We further classified the annotations of each group into four categories, depending on their E-values [3]. Annotations with an E-value of 1e-70 or less are considered to be highly confident (especially in Groups I and III), whereas those with an E-value of 1e-4 or greater are considered tentative or hypothetic (especially in Group II).

Table III

**Protein Versions**

The protein versions represent unique positions that the target proteins occupy in the evolutionary process—in this paper, the species categories. These categories are defined as is in the Swiss-Prot database: A: Archaea, B: Bacteria, E: Eukaryote, Plasm: Plasmid, Chl: Chloroplast, Mit Mitochondrion, V: Virus, and Cyan: Cyanelle (http://www.expasy.ch/sprot/sprot-top.html). The protein versions can be determined based on a comparative analysis of the BLOCKS patterns of the target proteins and the BLOCKS pattern-species associations in their corresponding PFG profiles of the knowledge system [11]**.**

The BLOCKS pattern is expressed in strings of uppercase letters (e.g., ABCDEF), each of the letter representing a conserved sequence motif for given BLOCKS families. The knowledge system established a specific association between the BLOCKS patterns and the species categories. It can be very complex: some patterns are universal to all species categories, whereas others are unique to certain species categories (Table IV). Nonetheless, most of the associations are well defined. Consequently, the protein version of the target proteins can be clearly determined by comparing the BLOCKS patterns of the proteins with their corresponding PFG profiles.

Table IV

Figure 1

**Procedure of Knowledge-Based Annotation Algorithm**

Figure 1 illustrates the procedure of our annotation algorithm. The procedure comprises three steps—data analysis, data processing, and voting—described as follows.

1. **Data analysis.** Analyze the genome sequence data (predicted proteins) with BLAST, BLOCKS, and Pfam in a high-throughput manner (note that they are the same sets of sequence analysis tools used in the knowledge system development).
2. **Data processing.** Process the tool-derived outputs from step I for every predicted protein in the genomes. For BLAST, the results include all homologous proteins and their corresponding E-values. Additional information included in the BLAST results is the detailed functional descriptions and their derived knowledge-based protein function categories (KPFCs) of these homologous proteins (the same procedure developed in the RuleMiner is used to extract KPFCs). For BLOCKS, the results include the best-hit BLOCKS families: the sequence-based protein function categories (SPFCs), BLOCKS motifs, and E-values. For Pfam, results include Pfam domains, their locations on the proteins and E-values. The Pfam results are further processed to form unique Pfam domain patterns in which domains are arranged on the proteins in the way that there are no overlaps.
3. **Voting.** Use the results from Step 2 (protein features, e.g., KPFCs, SPFCs, and Pfam domains) to query the knowledge system. PFGs have two components: KPFCs and SPFCs, which, together with other features in the PFG profiles, are stored in separate columns in the knowledge system. Therefore, querying the knowledge system with any of these features will result in the assignments

of possible PFG(s) and the identification of their related PFG profiles. Then, apply a voting procedure to determine the proper function annotations for target proteins and associate each of the annotations with confidence categories.

The voting procedure is complicated because there are many possible combinations of the sequence analysis tool-derived features and their potential entries (PFGs) in the knowledge system (Table III). For simplicity, three cases are established, which correspond to three annotation groups. In Case I, protein features such as BLOCKS motifs, Pfam domains, or their combinations are function-specific (e.g., one/many-to-one relationships between these features), and thus the corresponding PFGs in the knowledge system or PFG profiles can be used to recognize unique functions. In this case, the voting procedure would lead to specific functions and proteins would be annotated with high confidences (Group I). In Case II, protein features have no corresponding entries in the knowledge system. This will result in low confidences (Group II), especially when the E-value is large (function relevance with an E-value of zero is considered significant and that with an E-value of 0.1 or greater is considered unrelated). In Case III, the rule points to one/many to many, a non-unique feature-PFG relationship and as a consequence, voting procedure leads to multiple PFGs. In this case, there will be no decisions in choosing a specific function among these PFGs (Group III).

***Example of the Voting Procedure,*** The following example demonstrates how the knowledge system facilitates the voting procedure when multiple sequence analysis tools and knowledge system are incorporated. gi|1788071 is one of over 4,200 open reading frames (ORFs) in the genome of *Escherichia coli* K-12 MG1655. Because of the analytic process (Figure 2), two different functional assignments are given. One of the annotations is ribokinase with a protein function group of *PFG* (EC 2.7.1.15, IPB002173), and the other is 2-dehydro-3-deoxygluconokinase (3-deoxy-2-OXO-D-gluconate kinase) (KDG kinase), which belongs to *PFG* (EC 2.7.1.45, IPB002173). In this example, no unique function-specific protein features (rules and PFG profiles) can be identified in the knowledge system.

Figure 2

## RESULTS

One of the key features of our annotation algorithm is that we can obtain unique and highly confident functional annotations. Furthermore, each of the annotations is associated with confidence categories (e.g., category I.3 and I.4). In the *Escherichia coli* genome*,* over 51% of the proteins have such functional annotations (Figure 3A). About 24% of the protein annotations in *Archaeoglobus fulgidus* genome belong to these categories (Figure 3B). The principal reason that the knowledge-based annotation algorithm can achieve such a high confidence is that rules in the knowledge system can define a unique relationship between protein features and their corresponding cellular functions (PFGs). Among a total of 3,832 feature-PFG relationships examined [11], 1,821 are defined as unique by the BLOCKS analysis alone. Our analysis, which incorporates information from BLAST, BLOCKS and Pfam, would certainly strengthen the capability of the differentiation and the recognition of relevant function relationships and so increase the accuracy in computation-oriented function annotation.

Figure 3

Ribulose bisphosphate carboxylase (EC 4.1.1.39) (RuBisCO) is an example of such an annotation. RuBisCO catalyzes the initial step in Calvin cycle, the photosynthetic dark reaction pathway in plants cyano-, purple, and green bacteria [19]. It consists of a large catalytic unit and a small subunit of an undetermined function. The information in the knowledge system indicates that BLOCKS protein families and Pfam domains for both subunits are unique to their functions. The properties enable our annotation algorithm to discover two subunits in the genome of *Synechocystis* sp. and to assign unique functions to these subunits. We also found one or two copies of RuBisCO large subunits in nonphotosynthetic bacteria such as *Bacillus subtilis*, as well as Archaea including *Archaeoglobus fulgidus* and *Methanococcus jannaschii*. As was shown by Finn and Tabita [20], recombinant forms of the Archaeal enzymes catalyze a bona fide RuBP-dependent $CO_2$ fixation reaction, and it was recently shown that *Methanococcus jannaschii* and other anaerobic Archaea synthesize catalytically active RubisCO in vivo. In our study, all the functional assignments of ribulose bisphosphate carboxylase for the proteins in these genomes are classified as Category I.4.

Another unique feature of our annotation algorithm is that alternative annotations are given to some proteins (Category III). For example, 5% of *Escherichia coli* proteins and 9% of *Archaeoglobus fulgidus*

proteins are annotated as such (Figure 3). The reason is that proteins with such assignments are often highly homologous but have different sub-functions (e.g., enzymes with different substrate/ligand binding specificity); furthermore, no function-unique features can be defined for these proteins. For example, the BLOCKS protein family zinc-dependent dehydrogenase covers 17 different sub-functions. All of these sub-functional enzymes share similar catalytic mechanisms [21, 22].

Examples of such alternative functional assignments are shown in Table V for six genes in the *Aquifex aeolicus* genome. They cover a variety of cellular functions, including phosphatase, ATP-binding transporter, cytochrome oxidase, and transcriptional repressor and regulatory functions. In these families, BLOCKS patterns are essentially un-differentiable among all sub-functions. In addition, they possess identical Pfam domains. In the knowledge system, the features and PFGs for these functions are represented as one/many-to-many relationships. Obviously, the lack of unique protein feature identifiers for those highly homologous functions prevents our annotation algorithm from making final decisions about their functions. This situation contrasts the existing annotation systems, in which a brute-force approach is often used: functions are assigned mostly by whatever appears as the top hit of BLAST search.

Table V

**Comparison of Multiple Genome Annotations**

The annotation distributions in multiple genomes are compared in Figure 3. The genomes are arranged in a doughnut figure (see Table I for the detailed description of the species). The first five genomes are Eubacteria, and the rest are Archaea. In general, Archaea genomes are far less informative than those of Eubacteria in regards to functional inferences. If the genomes are arranged by their ratios of hypothetical protein to the total number of ORFs in these genomes, then five Archaea genomes will be located in the top five places, with *Aeropyrum pernix* in the first. Almost 60% (1,584) of the 2,694 proteins in the genome end without any functional clues. *Pyrococcus horikoshii* ranks second; about 46% of the 2,064 ORFs in the genome are hypothetical. The Eubacterial genomes generally have much lower ratios of hypothetic proteins. *Escherichia coli* K12 has the lowest of all, in which only 6% of its 4,248 ORFs are hypothetical. The other four genomes have around 20% hypothetical annotations. If these genomes are arranged by the ratio of proteins with Category I.4 annotations over the total ORFs in these genomes, their ranks are approximately reversed, with *Escherichia coli* at the top and *Aeropyrum pernix* at the bottom. So far, *Archaeoglobus fulgidus* has been shown to be the best-studied genome (11%) among the five Archaea.

## DISCUSSION

In this paper, we present knowledge-based voting algorithm for high-throughput protein function annotations, in which multiple sequence analysis tools, biological rules, and functional (PFG) profiles are seamlessly integrated. The objective is to reduce annotation errors, improve confidences, and relate the annotations with confidence categories. For the first time, a knowledge system has been established and incorporated into the protein annotation process. The results from the integrated sequence analysis tools for given proteins can be comparatively analyzed. In addition, much-needed guidance is now available to enhance such analysis for an accurate function assignment.

The annotations are further categorized based on confidence levels in the algorithm. The annotations with a strong support from the knowledge system are categorized at the highest level of confidence because of PF*G*s, well-defined PFG profiles, and clear-cut feature-function relationships; annotations without such support are considered tentative. The confidence information will be critical to researchers in deciding to what extent the annotation data can be trusted and then enable them to design experiments that are more reliable.

Alternative functional assignments represent another unique feature in our annotation system. With our algorithm, no conclusion is forced if the evidence is not strong enough. Our analysis revealed that about 7% of the proteins in the analyzed genomes (from 5% to 9%) have such assignments (Figure 4). This figure strongly contrasts with the results from other current annotation systems. These results are often inconsistent because of their reliance on a brute-force approach [3, 4, 6] that selects the best-scoring proteins regardless of the sequence databases used in the analysis [11].

Figure 4

The comparison of different genome annotation systems is difficult because of the lack of a standard system for function representations. Although we have not attempted to compare our rule-based annotation system with any others, the alternative functional assignment presents one of the real improvements in the field. This feature helps accurately reflect the complexity of the biological functions in which the proteins are involved [8-10] and prevent the spread of mistaken annotations [7, 11].

Alternative function assignments also open an opportunity to fill gaps in the metabolic pathway for certain organisms, in which some enzymes are mysteriously missing in current annotation data. For example, EC 5.3.1.8 (mannose-6-phosphate isomerase) is listed as a missing function from *Synechocystis* PCC6803 and other cyan-bacteria genomes (http://www.genome.jp/kegg/). In our analysis, however, alternative functions are assigned for single proteins in these genomes, including a mono-functional enzyme of EC 2.7.7.22 (Mannose-1-phosphate guanylyltransferase) and a bi-functional enzyme of EC 2.7.7.22 5.3.1.8 (Figure 5). These alternatives provide scientific evidence to generate working hypotheses for researchers to design experiments to fill such metabolic and regulatory pathway gaps [23].

Figure 5

The analysis of the distribution of annotation confidences among multiple genomes indicates a strong discrepancy in the representation of current knowledge. *Escherichia coli* has the highest ratio of proteins (over 50% of 4,289) that have annotations of the highest confidence (Categories I.3 and I.4). In contrast, *Aeropyrum pernix*, a crenarchaeota genome, represents one of the most poorly studied genomes. Only 5% of its 2,694 predicted ORFs have the annotations classified as such. The majority (59%) of ORFs have no functional clues at all. On one hand, the poorly annotated genomes in general and Archaea genomes in particular reflect the current limitations of computational tools in function determinations. On the other hand, they present an opportunity to find new functions if efforts are made to systematically studying these genomes and their corresponding organisms.

As indicated above, the sequence-based functional annotations, while useful in certain cases, are limited in their coverage of protein functional space. Function references based on protein networks present another layer of genome analysis methods complementary to sequence-based analysis. We believe that proteins often form structured interaction network modules to accomplish specific functions, such as transcriptional regulatory, metabolic synthesis, and signal transductions. Therefore, hypothetic proteins that have highly confident links with these network modules are likely to have similar functions [24]. To test this hypothesis, we plan to develop an integrated network construction system and incorporate network information into our annotation algorithm to expand functional coverage and increase annotation accuracy.

In the current knowledge system and knowledge-based voting algorithm, we applied three sequence analysis tools, BLAST, BLOCKS and Pfam for their distinguished capabilities, broad sequence and functional coverages, rich annotation information, unique yet complementary attributes [11]. In brief, Blast, revealing sequence similarity at the level of individual amino acids, could recognize and distinguish, to certain level, homologous proteins but could not identify evolutionarily divergent yet functionally related ones. Pfam and BLOCKS, two family-based signature databases, could well complement the BLAST tool since they can detect divergent domains and conserved motifs, therefore, having ability to identify distant and clear-cut relationships in novel sequences [14, 15].

However, we do not intend, by any means, to exclude other signature databases because those signature databases address different sequence analysis problems and have their own strength. Quite the opposite, building such knowledge-based annotation frame makes it easier to incorporate additional sequence analysis tools, thus, facilitating the development of more complex and smarter annotation systems in the future. As a final point of this paper, we have to emphasize that Gene Ontology (GO) will play an especially important role in such systems since GO can provide structured vocabularies to describe genes and gene products [25]. As a consequence, GO would facilitate function representation, which is an essential part of algorithm in the development of our knowledge system.

**ACKNOWLEDGMENT**

This work was supported in part by the U.S. Department of Energy under Contract W-31-109-Eng-38.

**REFERENCES**


1. Bernal, A., U. Ear, and N. Kyrpides, *Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.* Nucleic Acids Res., 2001. **29**: p. 126-127.
2. Heffelfinger, G.S., et al., *Carbon Sequestration in Synechococcus Sp.: from molecular machines to hierarchical modeling.* OMICS, 2002. **6**: p. 305-330.
3. Andrade, M.A., et al., *Automated genome sequence analysis and annotation.* Bioinformatics, 1999. **15**: p. 391-412.
4. Gaasterland, T. and C. Sensen, *Fully automated genome analysis that reflects user needs and preferences-a detailed introduction to the MAGPIE system architechure.* Biochimie, 1996. **78**: p. 302-310.
5. Frishman, D. and H. Mewes, *Pedantic genome analysis.* Trends Genet., 1997. **13**: p. 415-416.
6. Overbeek, R., et al., *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction.* Nucleic Acids Res., 2000. **28**: p. 123-125.
7. Wu, C.H., et al., *The Protein Information Resource: an integrated public resource of functional annotation of proteins.* Nucleic Acids Res., 2002. **30**: p. 35-37.
8. Strauss, E.J. and S. Falkow, *Microbial pathogenesis: genomics and beyond.* Science, 1997. **276**: p. 707-712.
9. Massingham, T., L.J. Davies, and P. Lio, *Analysing gene function after duplication.* Bioessays, 2001. **23**: p. 873-876.
10. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**: p. 459-473.
11. Yu, G.X., *RuleMiner: a knowledge system for supporting high-throughput protein function annotations.* JBCB, 2004: p. 595-617
12. Galperin, M.Y. and E.V. Koonin, *Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption.* In Silico Biol, 1998. **1**(1): p. 55-67.
13. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol., 1990. **215**: p. 403-410.
14. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res., 2002. **30**: p. 276-280.
15. Henikoff, J.G., et al., *Increased coverage of protein families with the Blocks database servers.* Nucleic Acids Res., 2000. **28**: p. 228-230.
16. Mulder, N.J., et al., *The InterPro Database, 2003 brings increased coverage and new features.* Nucleic Acids Res., 2003. **31**: p. 315-318.
17. Kretschmann, E., W. Fleischmann, and R. Apweiler, *Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT.* Bioinformatics, 2001. **17**: p. 920-926.
18. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Res., 2000. **28**: p. 45-48.
19. Miziorko, H.M. and G.H. Lorimer, *Ribulose-1,5-bisphosphate carboxylase-oxygenase.* Annu Rev Biochem., 1983. **52**: p. 507-535.
20. Finn, M.W. and F.R. Tabita, *Synthesis of catalytically active form III ribulose 1,5-bisphosphate carboxylase/oxygenase in archaea.* J Bacteriol., 2003. **185**(10): p. 3049-3059.
21. Sun, H.W. and B.V. Plapp, *Progressive sequence alignment and molecular evolution of the Zn-containing alcohol dehydrogenase family.* J Mol Evol., 1992. **34**: p. 522-535.
22. Joernvall, H., B. Persson, and J. Jeffery, *Characteristics of alcohol/polyol dehydrogenases. The zinc-containing long-chain alcohol dehydrogenases.* Eur J Biochem, 1987. **167**: p. 195 -201.
23. Osterman, A. and R. Overbeek, *Missing genes in metabolic pathways: a comparative genomics approach.* Curr Opin Chem Biol., 2003. **7**(2): p. 238-251.
24. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-453.
25 Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al.* The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, 2003. **13**: p. 662–672

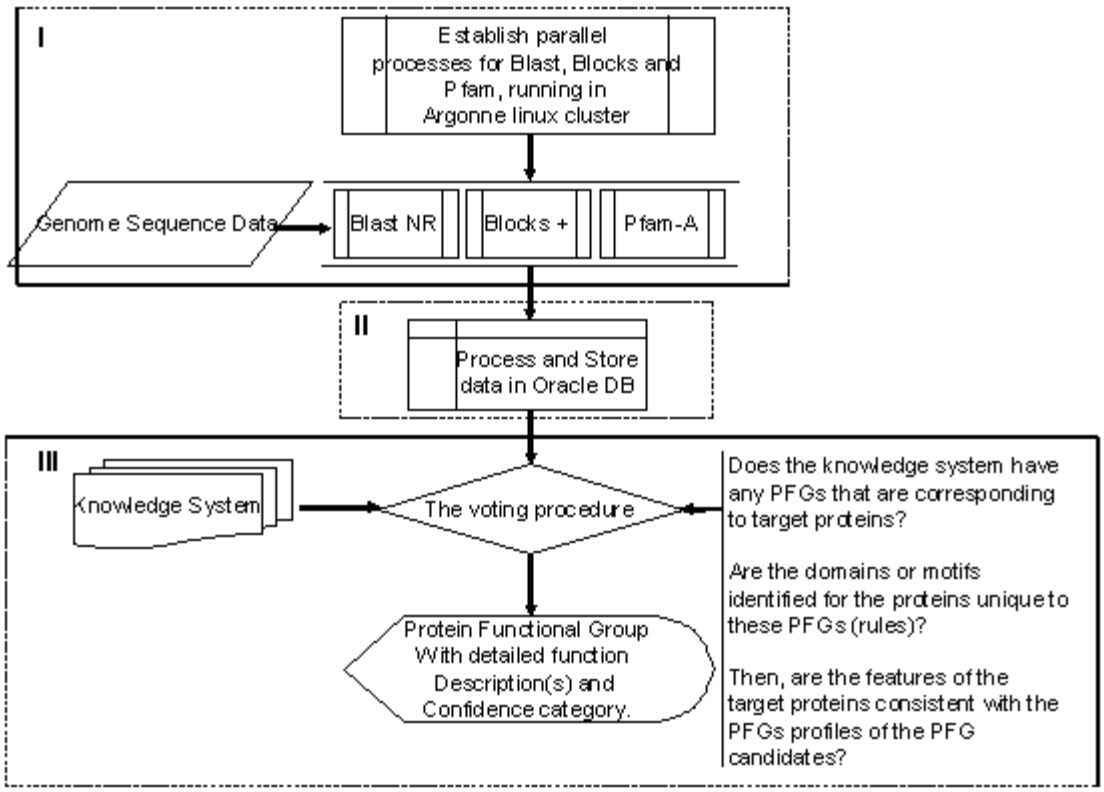


**Figure 1.** The knowledge-based annotation procedure includes three steps: I. data analysis, II. data processing and III. functional annotations. In step I, genome sequence data are analyzed with BLAST, Pfam and BLOCKS; in the step II, outputs from step I are further processed to extract sequence features that can be used to directly query RuleMiner, the knowledge system; In the last step, the voting procedure are applied to determine the function annotations and assign function confidence categories. All these processes proceed in a high-throughput and automatic manner.

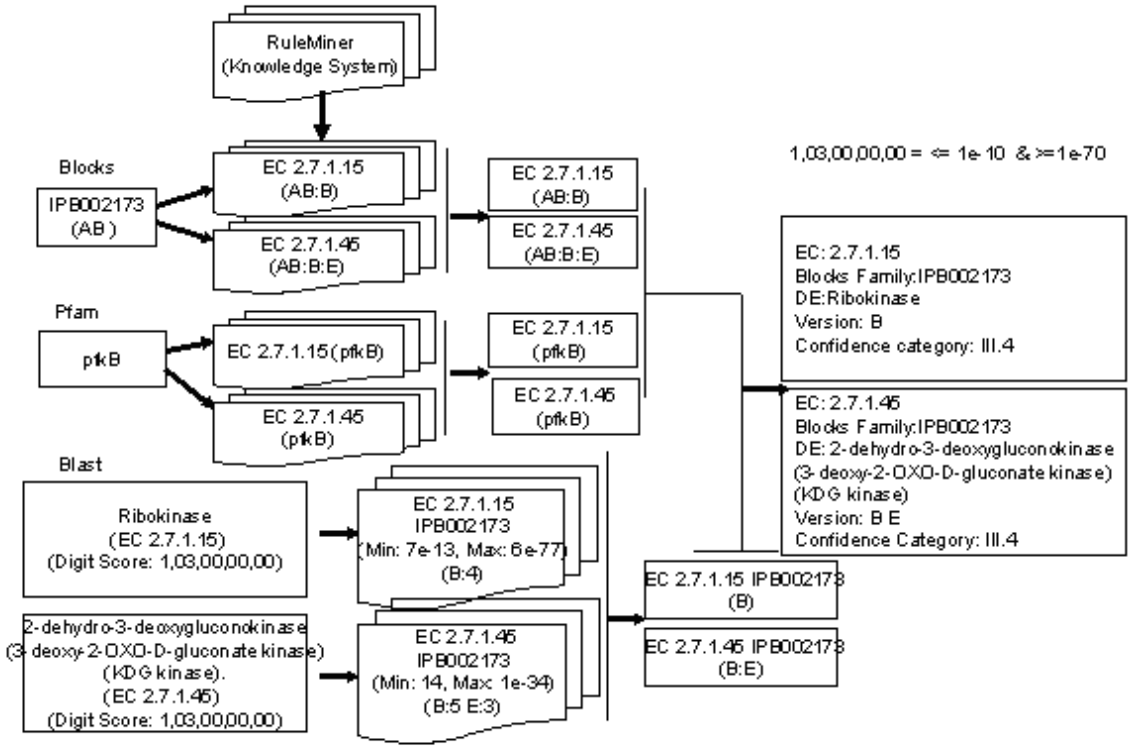


**Figure 2.** Application of the knowledge-based voting algorithm for annotation of gi|1788071, a protein from *Escherichia coli* K-12 MG1655 (Bacterium). The annotation is based on the sequence features identified by three sequence analysis tools and information in the knowledge system.

1. The Pfam domain-based rules (non-unique relationships between Pfam features and the protein functional groups) provide no differentiation ability in this example. "pfkB" is essential for both PFG (EC 2.7.1.15, IPB000965) and PFG (EC 2.7.1.45, IPB000965).
2. Feature-species associations could not offer any extra information for this annotation: PFG (EC 2.7.1.15, IPB000965) occurs in Eubacteria and PFG (EC 2.7.1.45, IPB000965) in both Eukaryote and Eubacteria.
3. BLOCKS and BLAST-based PFG profile could not supply any support either in distinguishing two protein function groups. Protein features from both BLOCKS and BLAST analysis fit very well with profiles from both protein functional groups.

Consequently, no decision will be made between the alternative functions and no unique functions will be given until new evidence is obtained.

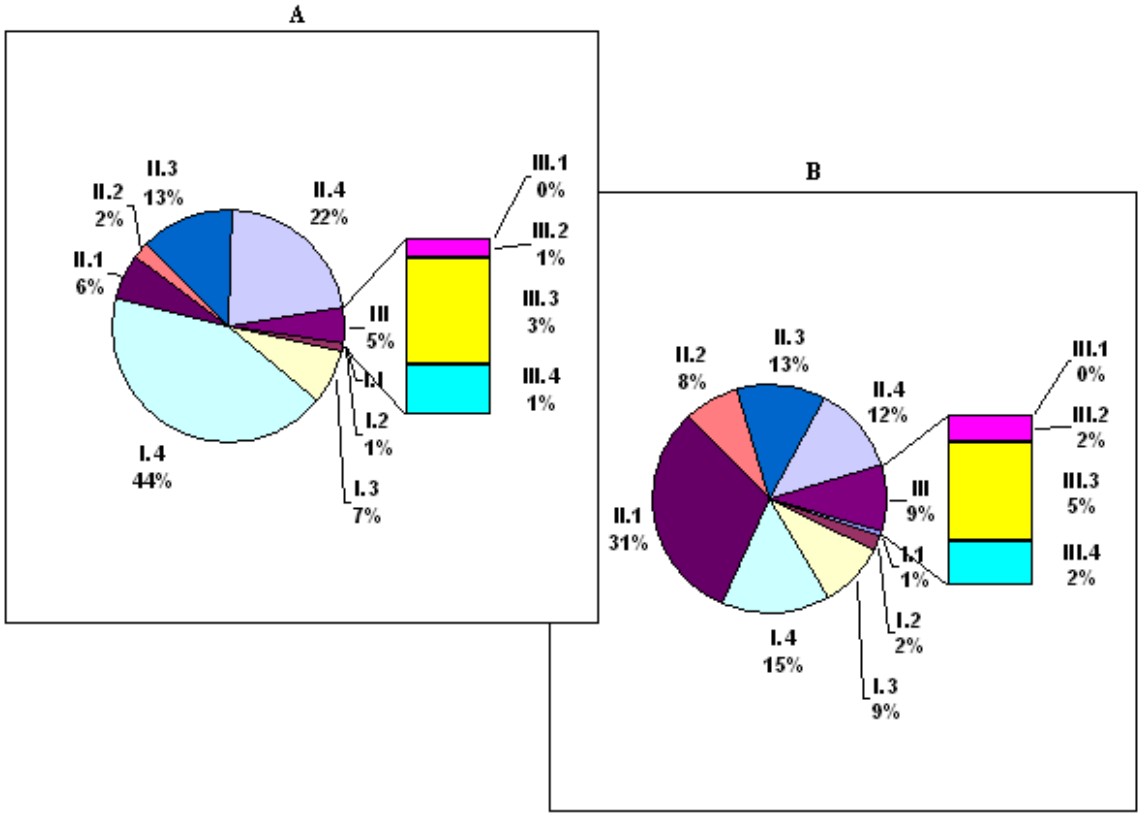


**Figure 3.** Percentage distribution of categories of annotation confidence of two genomes; one from bacterium: *Escherichia coli* (Panel **A**) and one from Archaea: *Archaeoglobus fulgidus* (Panel **B**). The labels indicate the categories of annotation confidences. Categories I.1 to I.4 present unique functional assignments with a strong knowledge system support. Categories II.1 to II.4 present annotations that do not have such support. Categories III.1 to III.4 present proteins with alternative functional assignments with knowledge base supports. From those confidence categories, X.1 is considered as hypothetical (or marginal) and tentative, and X.4 is considered as highly confident, where X presents any of the three confidence groups (Table II).

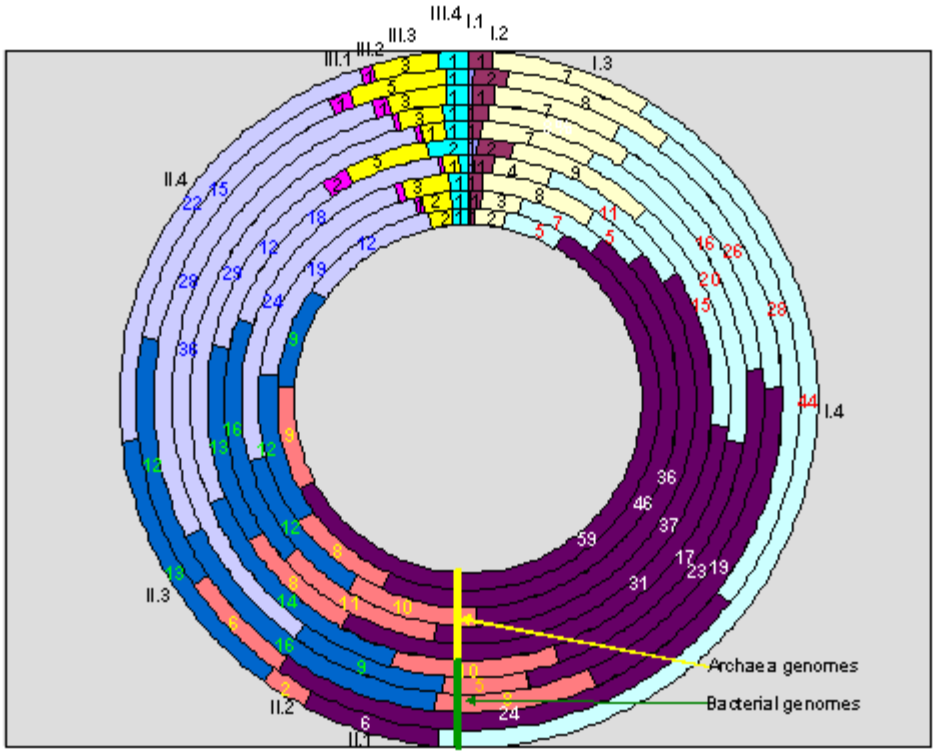


**Figure 4.** Comparisons of percentage distribution of categories of annotation confidence for ten genomes. The genomes are arranged in a doughnut as follows (from outer cycle to the inner): ECOLI, BACNA, AQUAE, CHLTR, RICPR, ARCFU, METJA, HALN1, PYRHO, AERPE (see Table I for detailed species descriptions). The labels indicate the categories of annotation confidences. Categories I.1 to I.4 present unique functional assignments with a strong knowledge base support. Categories II.1 to II.4 present annotations that do not have such support. Categories III.1 to III.4 present proteins with alternative functional assignments with knowledge system supports. From those confidence categories, X.1 is considered as hypothetical (or marginal) and tentative, and X.4 is considered highly confident, where X presents any of the three confidence groups (Table II).

A

| Gene | Assigned Function | Protein function group | | Version | Confidence |
|---|---|---|---|---|---|
| 1593 | Mannose-1-phosphate guanylyltransferase (GDP) (EC 2.7.7.22) (GDP- mannose pyrophosphorylase) (GMP) | E.C 2.7.7.22 | IPB001538 | B Plasmid | III.A |
| | Xanthan biosynthesis protein xanB Includes: Mannose-6-phosphate isomerase (EC 5.3.1.8) (Phosphomannose isomerase) (PMI) (Phosphohexomutase); Mannose-1-phosphate guanylyl transferase (GDP) (EC 2.7.7.22) (GDP-mannose pyrophosphorylase) (GMP) | E.C 5.3.1.8 E.C 2.7.7.22 | IPB001538 | B | III.A |

B

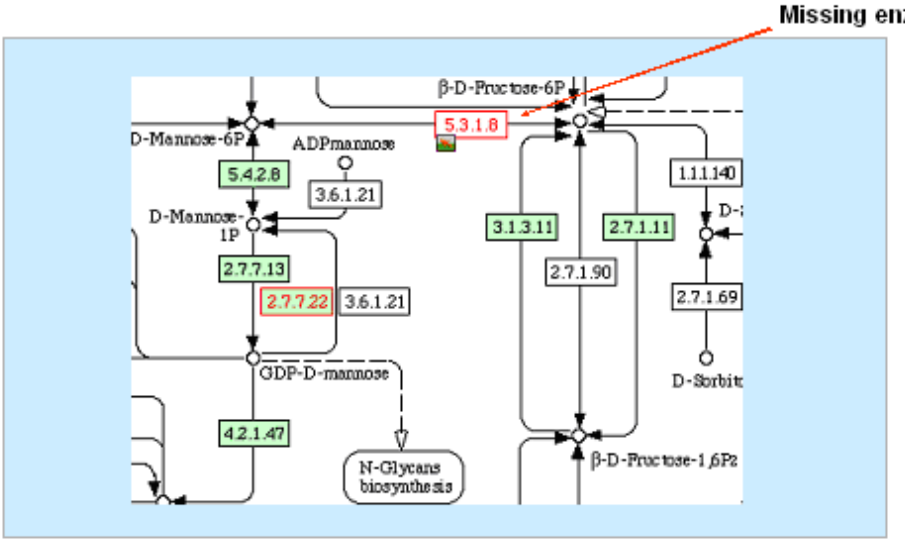


**Figure 5.** Annotation example by the RuleMiner and profile-based voting algorithm and its potential application. **A:** Alternative functions are assigned, each of which associates explicitly with annotation confidences for a *Prochlorococcus marinus* protein**. B:** Pathways (metabolic, regulatory) are refined through improved annotations: The function of EC 5.3.1.8 is missed in this and other several cyanobacteria genomes (*Anabaena*, *Synechocystis*, *Synechococcus,* and *Thermosynechococcus*) in the fructose and mannose metabolism pathway. Assignment of a bifunctional enzyme provides a mechanism to fill the gap [11]. Boxes with green background represent functions found in these organisms.

**Table I.** The complete microbial genomes tested and their basic information.

| Species Name | Species Symbol | Phylum | Class | Order | SuperKingdom | Genome Size (Kbps) | Number of ORFs |
|---|---|---|---|---|---|---|---|
| *Escherichia coli K12* | ECOLI | ProteobacteriaProteobacteria | Gammaproteobacteria | Enterobacteriales | Bacteria | 4639 | 4289 |
| *Bacillus subtilis* | BACNA | Firmicutes | Bacilli | Bacillales | Bacteria | 4215 | 4100 |
| *Aquifex aeolicus* | AQUAE | Aquificae | Aquificae | Aquificales | Bacteria | 1551 | 1522 |
| *Chlamydia trachomatis* | CHLTR | Chlamydiae | Chlamydiae | Chlamydiales | Bacteria | 1043 | 894 |
| *Rickettsia prowazekii* | RICPR | Proteobacteria | Alphaproteobacteria | Rickettsiales | Bacteria | 1111 | 834 |
| *Archaeoglobus fulgidus* | ARCFU | Euryarchaeota | Archaeoglobi | Archaeoglobales | Archaea | 2178 | 2420 |
| *Methanococcus jannaschii* | METJA | Euryarchaeota | Methanococci | Methanococcales | Archaea | 1665 | 1715 |
| *Halobacterium sp.* | HALN1 | Euryarchaeota | Halobacteria; | Halobacteriales | Archaea | 2014 | 2058 |
| *Pyrococcus horikoshii* | PYRHO | Euryarchaeota | Thermococci | Thermococcales | Archaea | 1739 | 2064 |
| *Aeropyrum pernix* | AERPE | Crenarchaeota | Thermoprotei | Desulfurococcales | Archaea | 1669 | 2694 |

**Table II.** Confidence levels, E-value, and score assignments.

| Confidence Levels | E-value | Digit scoring | Confidence in GenQuiz (Probability) |
|---|---|---|---|
| 1 | <1 | 10 | Unknown (0%) |
| 2 | <0.1 | 100 | Unknown (0%) |
| 3 | <1e-4 | 10000 | Marginal (30%) |
| 4 | <1e-10 | 1000000 | Tentative (70%) |
| 5 | <1e-20 | 100000000 | Clear (95%) |
| 6 | <1e-70 | 10000000000 | Clear (99%) |
| 7 | <1e-100 | 1000000000000 | Clear (>99%) |
| 8 | <1e-200 | 100000000000000 | Clear (>99%) |

**Table III.** Distribution of 12 confidence categories in the knowledge-based function annotation algorithm

| Group | Group Descriptions | Category | Confidences (E-value) |
|---|---|---|---|
| I | Unique Assignments: The integrated analysis of the tool-derived protein features, the rules and PFG profiles lead to a unique PFG (e.g. the BLOCKS patterns and Pfam domain patterns are function-specific in the knowledge system). | 1 | >1e-04 |
| | | 2 | < 1e-04 && >1e-20 |
| | | 3 | < 1e-20 && >1e-70 |
| | | 4 | < 1e-70 |
| II | Assignments without knowledge system supports: There are no functional descriptions related to the target proteins in the Swiss-Prot database. Therefore, there are no corresponding *PFG*s in the knowledge System | 1 | >1e-04 |
| | | 2 | < 1e-04 && >1e-20 |
| | | 3 | < 1e-20 && >1e-70 |
| | | 4 | < 1e-70 |
| III | Alternative Assignments: all tool-derived protein features, the rules and PFG profiles in the knowledge system point to several undistinguishable *PFG*s. In this circumstance, there will be no decisions about their functions unless we obtain additional evidence. | 1 | >1e-04 |
| | | 2 | < 1e-04 && >1e-20 |
| | | 3 | < 1e-20 && >1e-70 |
| | | 4 | < 1e-70 |

**Table IV.** Evolutionary distribution of BLOCKS patterns of PFGs in the knowledge system

| Protein Functional Group (PFG) Function Category | BLOCKS Family | Association of BLOCKS Pattern - Species Category in PFG profile |
|---|---|---|
| EC 2.7.7.6 | IPB001572 | ABCDEGH[a]:B[b], ABCDEFGH:A:B:E:chl:cynal, H:B, CDEGH:plasm, BCDEGH:E, EFGH:chl, CDEFGH:V, ABC:A, DEFGH:A:V, BCDEFGH:E |
| EC 4.2.99.9 | IPB000277 | ABCDEF:B:E, BCDF:E, BCF:E |

| | | |
|---|---|---|
| EC 1.6.99.3 | IPB000103 | AE:A, BCDE:B |
| EC 4.1.1.23 | IPB001754 | ABE:A:B, ABCDEF:E, BCE:B, E:B, AE:A, BE:B, ABCE:A:B, CE:B |
| High mobility group protein | IPB000910 | ABC:E, BC:E:V, C:E |
| ATP-dependent helicase | IPB000629 | A:B, AC:A:B, AD:B, AE:A, ABCDE:E, |
| E.C.4.2.2.2 | PR00807 | DEG:B BCDE:E ABCDEFGH:E CD:B DE:B ADE:E CDE:B |

Note: [a]. BLOCKS Patterns; [b]. Species categories defined in Swiss-Prot.

**Table V.** Alternative functional assignments in *Aquifex aeolicus*

| Genepid | BLAST-based Function Description | Digit Score | Knowledge-based Function Groups (KPFG) | Family-based Function Groups (SPFG) | BLOCKS Pattern | Pfam Domain | Spec Categ |
|---|---|---|---|---|---|---|---|
| 2984221 | | | | | | | |
| | Alpha-ribazole-5-phosphate phosphatase (EC 3.1.3.-). | 2000000 | 3.1.3.- | IPB001345 | ABCD | PGAM | B |
| | Phosphoglycerate mutase (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM) (MPGM) (BPG-dependent PGAM). | 2100000 | 5.4.2.1 | IPB001345 | ABCD | PGAM | BE |
| 2984332 | | | | | | | |
| | ATP-binding protein abc. | 200000000 | ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Choline transport ATP-binding protein opuBA. | 100000000 | Choline transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Dipeptide transport ATP-binding protein dppD. | 500000000 | Dipeptide transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | General L-amino acid transport ATP-binding protein aapP. | 100000000 | General L-amino acid transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Glutamine transport ATP-binding protein glnQ. | 200000000 | Glutamine transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Glycine betaine L-proline transport ATP-binding protein proV. | 200000000 | Glycine betaineL-proline transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Glycine betaine carnitine choline transport ATP-binding protein opuCA. | 100000000 | OxPP cycle protein | IPB001617 | AB | ABC_tran | B |
| | Glycine betaine transport ATP-binding protein opuAA (EC 3.6.3.32) (Quaternary-amine-transporting ATPase). | 100000000 | 3.6.3.32 | IPB001617 | AB | ABC_tran | B |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Histidine transport ATP-binding protein hisP. | 200000000 | Histidine transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Iron(III)-transport ATP-binding protein hitC. | 200000000 | Iron(III)-transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Maltose maltodextrin transport ATP-binding protein malK. | 300000000 | Maltosemaltodextrin transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Nitrate transport ATP-binding protein nrtC. | 400000000 | Nitrate transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Nopaline permease ATP-binding protein P. | 100000000 | Nopaline permease ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Oligopeptide transport ATP-binding protein oppD. | 1501000000 | Oligopeptide transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Possible ribonucleotide transport ATP-binding protein mkl. | 200000000 | Ribonucleotide transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| | Probable ABC transporter ATP-binding protein PEB1C. | 804000000 | ABC transporter ATP-binding protein | IPB001617 | AB | ABC_tran | B plasm |
| | Probable amino-acid ABC transporter ATP-binding protein yckI. | 500000000 | Amino-acid ABC transporter ATP-binding protein | IPB001617 | AB | ABC_tran | B plasm |
| | Putative ferric transport ATP-binding protein afuC. | 300000000 | Ferric transport ATP-binding protein | IPB001617 | AB | ABC_tran | B |
| 2984362 | | | | | | | |
| | Potential acrAB operon repressor. | 10000 | AcrAB operon repressor | IPB001647 | A | tetR | B |
| | Regulatory protein mtrR. | 10000 | Regulatory protein | IPB001647 | A | tetR | B |
| | Tetracenomycin C transcriptional repressor. | 10000 | Tetracenomycin C transcriptional repressor | IPB001647 | A | tetR | B |
| | Transcriptional repressor Bm3R1. | 10000 | Transcriptional repressor | IPB001647 | A | tetR | B |
| | Uid operon repressor (Gus operon repressor). | 10000 | Uid operon repressor | IPB001647 | A | tetR | B |
| 2984384 | | | | | | | |
| | Cytochrome O ubiquinol oxidase subunit III (EC 1.10.3.-). | 2000000 | 1.10.3.- | PF00510 | DE | COX3 | B |
| | Cytochrome c oxidase polypeptide III (EC 1.9.3.1). | 4582050000 | 1.9.3.1 | PF00510 | DE | COX3 | B E mi |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Quinol oxidase polypeptide III (EC 1.9.3.-) (Quinol oxidase aa3-600, subunit qoxC) (Oxidase aa(3)-600 subunit 3). | 2010000 | 1.9.3.- | PF00510 | DE | COX3 | B |
| 2983588 | | | | | | | |
| | Acetoacetate metabolism regulatory protein atoC (Ornithine arginine decarboxylase inhibitor) (Ornithine decarboxylase antizyme). | 10000000000 | Acetoacetate metabolism regulatory protein | IPB002078 | ABCD | sigma54 | B |
| | Alginate biosynthesis transcriptional regulatory protein algB. | 10000000000 | Glycosyltransferase | IPB002078 | ABCD | sigma54 | B |
| | Nitrogen regulation protein NR(I). | 70100000000 | Nitrogen regulation protein | IPB002078 | ABCD | sigma54 | B |
| | Repressor protein luxO. | 10000000000 | Repressor protein | IPB002078 | ABCD | sigma54 | B |
| | Transcriptional regulatory protein hydG. | 20814160000 | Transcriptional regulatory protein | IPB002078 | ABCD | sigma54 | B plasm |
| | Type 4 fimbriae expression regulatory protein pilR. | 10000000000 | 4 fimbriae expression regulatory protein | IPB002078 | ABCD | sigma54 | B |
| 2982837 | | | | | | | |
| | Acetoacetate metabolism regulatory protein atoC | 100000000 | Acetoacetate metabolism regulatory protein | IPB002078 | BCD | sigma54 | B |
| | Alginate biosynthesis transcriptional regulatory protein algB. | 100000000 | Glycosyltransferase-transcritional-regulatory | IPB002078 | BCD | sigma54 | B |
| | Formate hydrogenlyase transcriptional activator. | 200000000 | Formate hydrogenlyase | IPB002078 | BCD | sigma54 | B |
| | Hydrogenase-4 transcriptional activator. | 100000000 | Hydrogenase-transcriptional-activator | IPB002078 | BCD | sigma54 | B |
| | Nitrogen fixation protein anfA. | 200000000 | Nitrogen fixation protein | IPB002078 | BCD | sigma54 | B |
| | Signal-transduction and transcriptional-control protein. | 100000000 | Stanniocalcin | IPB002078 | BCD | sigma54 | B |
| | Transcriptional regulatory protein flbD. | 901000000 | Transcriptional regulatory | IPB002078 | BCD | sigma54 | B plasm |
| | Transcriptional regulatory protein xylR (67 kDa protein). | 100000000 | Xylose repressor | IPB002078 | BCD | sigma54 | plasm |

Note: [a]. Species categories defined in Swiss-Prot (A=Archaea; B=Eubacteria; E=Eukaryote; plasm=plasmid)