

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

Yu, GX<sup>1\*</sup>, Glass, EM<sup>1</sup>, Karonis, NT<sup>1,2</sup>, and Maltsev, N<sup>1</sup>

<sup>1</sup>Mathematics and Computer Science Division  
Argonne National Laboratory, Argonne, IL 60439, USA

<sup>2</sup>Department of Computer Science  
Northern Illinois University, DeKalb, IL 60115, USA

To whom correspondence should be addressed:

yug@ornl.gov

\*current address: Computer Science and Mathematics Division  
Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37830, USA

## **ABSTRACT**

**Motivation:** Automated annotation of genome sequences is one of the earliest and indispensable steps toward a comprehensive understanding of the dynamic behavior of living organisms. It is, however, often an error-prone procedure because of the underlying algorithms in current analysis systems, which rely mainly on simple similarity analysis and lack guidance from biological rules. We present here a knowledge-based protein annotation algorithm. Our objectives are to reduce annotation errors, improve and categorize confidences, and explicitly associate them with the annotations.

**Results:** 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 guidance for function annotation. The voting procedure relies on the knowledge-system and is designed to make unbiased judgments in functional assignments among complicated and sometimes conflicting information. We have applied this algorithm on ten prokaryotic genomes and observed a significant improvement in annotation confidences. Current limitations of the algorithm and the potential for future improvement are also discussed.

**Availability:** The results can be found by querying the WIT3 database at <http://compbio.mcs.anl.gov/wit3>.

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

## **INTRODUCTION**

The number of whole sequenced genomes has dramatically increased during the past several years, and this trend is likely to continue at an accelerated pace. Currently, the Genomes OnLine Database (GOLD) lists over 211 completely sequenced genomes [1]. An 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 dramatically accelerate progress toward a comprehensive understanding of the genetic mechanisms involved in diverse biochemical processes pertinent to medicine, biotechnology, environmental management, and agriculture [2]. The challenge is that the experiments needed to systematically determine functionalities for all predicted proteins in the sequenced genomes are extremely labor-intensive and prohibitively expensive.

In an effort to complement such experiments, several computational approaches have been developed to automate the annotation processes [3-6]. Automated annotation, however, is often an error-prone procedure because of underlying algorithms in the analysis systems [7]. These algorithms rely mainly on Blast or FastA-based sequence similarity analysis, although a variety of other approaches may be included in the systems. In contrast, the diversity of functions in which proteins are involved in various cellular processes has created complicated, sometimes unpredictable, sequence-function relationships [8]. Evolutionary processes may add complexity to the annotation process [9, 10]. Similarity analysis thus cannot always recognize and differentiate between relevant function relationships [7, 11]. Hence, results are often difficult to interpret and error-prone, and the annotation confidences are hard to evaluate [12]. Integration of multiple sequence analysis tools (e.g., Blast for similarity-analysis [13], Pfam for biologically important domains [14], and Blocks for highly conserved motifs [15]) and introduction of biological rules to provide relevant guidance will be essential to achieve an enhanced computational capacity for recognizing and differentiating cellular functions in the function annotations.

Integration is critical because each of the sequence analysis tools addresses different sequence analysis problems and has its unique features and capability [16]. All these tools, however, have been independently developed and have resulted in incompatible nomenclatures [16]. Hence, the integration can be enormously difficult and can severely compromise the efficacy of these tools for annotating protein function. A lack of clear principles or rules in protein functional analysis presents 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, named “RuleMiner,” for high-throughput genome sequence analysis. The knowledge system consists of three components: protein function groups (PFGs), PFG profiles, and rules. The PFGs, established from an integrated analysis of current knowledge of protein functions from the Swiss-Prot database [18] and protein family-based sequence classifications, cover all possible cellular functions available in the database. The PFG profiles illustrate detailed protein features for each PFG as in sequence conservations (Blast and Blocks), the occurrences of sequence-based motifs (Blocks), domains (Pfam), and species distributions.

The rules, extracted from the PFG profiles, describe the clear relationships between these PFGs and all possible features. As a result, the knowledge system provides an enhanced capability for protein function analysis. For example, results from sequence analysis tools for given proteins can be comparatively analyzed. Also, much-needed guidance is readily available for such analysis. If the rules describe unique relationships between the protein features and the PFGs—for example, 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 has to be provided.

In this paper, we present an algorithm for high-throughput protein annotations. 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, and relate the annotations with confidence categories. Our algorithm consists of two major components: the knowledge system “RuleMiner” and a voting procedure. The knowledge system provides guidance for function annotation; the voting procedure, which relies on rules and the functional profiles in the knowledge system, is designed 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 are categorized as having the highest confidence.

We have applied this algorithm on ten prokaryotic bacterial genomes and observed significantly improved annotation confidences. We believe this algorithm will be of a great help 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 can design their experiments accordingly. The genome annotation data and the annotation programs are 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—before we illustrate the annotation procedure of the rule-based algorithm. In addition, we give an example to illustrate the procedure.

### Data Preparation

We downloaded ten completely sequenced genomes (Table 1) 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 life domains, 6 classes and 10 orders), ranging from genomes that are well studied (e.g., *Escherichia coli* K12 and *Bacillus subtilis*) to those that are barely examined (e.g., *Halobacterium sp.* and *Aeropyrum pernix*). Thus, this genome data can be used for evaluating the performance of our algorithm.

### Genome Sequence Data Processing

To provide computational capability to process the high-throughput genomes sequence data, we developed a parallel process for Blast, Blocks, and Pfam to run on a 512-node Linux cluster at Argonne National Laboratory. Building such a parallel process is essential 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 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. Here, however, we do not describe the details of database design and managements, 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 the computational sequence analysis tools. Biological function domains, motifs, and Blast hits with a lower E-value are more likely the right 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 functional domains or motifs with certain E-values. Compare the Blast hits is difficult, however, because Blast analysis can give multiple hits with the same functional annotations, each of which associates with a particular E-value. To represent protein functions of the Blast hits, we developed a novel scoring system (Table 2). In this system, eight confidence levels of scores are defined by extending the scoring scheme of GeneQuiz [3]. Each confidence level is represented by two digits so that maximum number of Blast hits can be accumulated up to 99 without blending 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.

### **Annotation Confidence Categories**

The confidence category indicates how much 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 3). Annotations in Groups I and III have 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, in contrast, do not have such support, resulting in low confidence. We further classified the annotations of each group into four categories, depending on the E-values and their associated confidences [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 as tentative or hypothetical (especially in Group II).

### **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 in 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 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 as strings of uppercase letters (e.g., ABCDEF), each representing a conserved sequence motif for given Blocks families. One of the features defined in the knowledge system is the specific associations between these Blocks patterns and the species categories. The associations can be complex: some patterns are universal to all species categories, whereas others are unique to certain species categories (Table 4). Nonetheless, most of the associations are well defined for each protein family. 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.

### **Procedure of Knowledge-Based Annotation Algorithm**

Figure 1 illustrates the procedure of our annotation algorithm. Briefly, 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 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 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, 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 extraction of their related PFG profiles. Then, apply a voting procedure to determine the proper function annotations for target proteins and associates each of the annotations with confidence categories. The annotation confidence will depend on the answers to the following questions: First, does the knowledge system have any PFGs corresponding to the target proteins? Are the identified domains or motifs unique to these PFGs (rules)? Then, are the features of the target proteins consistent with the PFGs profiles of the candidate PFGs? Annotations that satisfy all these questions are considered as having the highest confidence.

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 3). For simplicity, roughly three cases can be 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 their corresponding PFGs in the knowledge system or PFG profiles can be used to recognize unique functions. The voting procedure thus leads to specific functions. In this case, proteins will be annotated as high confident annotations (Group I). In Case II protein features have no corresponding entries in the knowledge system. In this case, the annotations will have 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, all protein features and their related information in the knowledge system lead to multiple PFGs if the rule indicates one/many to many, a non-unique feature-PFG relationship. In this case, there will be no decisions in choosing a specific function among these (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. The example, 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 to the protein. One of the annotations is ribokinase with 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.

## 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 Blocks analysis alone. Our analysis, which incorporates information from Blast, Blocks and Pfam, would certainly add strength to the differentiation and recognition of relevant function relationships and hence increase the accuracy in computation-oriented function annotation.

Ribulose biphosphate carboxylase (EC 4.1.1.39) (RuBisCO) is an example of such an annotation. RuBisCO catalyzes the initial step in Calvin's reductive pentose phosphate cycle in plants as well as cyanobacteria, purple, and green bacteria [19]. It consists of a large catalytic unit and a small subunit of undetermined function. 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 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<sub>2</sub> 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 biphosphate 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 having such assignments are often highly homologous but have different subfunctions (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 subfunctions. All of these subfunctional enzymes share similar catalytic mechanisms [21, 22]. An additional 134 homologous protein groups (organized by protein superfamilies) can be visualized in <http://www-wit.mcs.anl.gov/svmmmer/>.

Examples of such alternative functional assignments are shown in Table 5 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 for all functions are essentially undifferentiable among the subfunctions. 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, a lack of unique protein feature identifiers for those highly homologous functions prevents our annotation algorithm from making final decisions about their functions. This situation is contrast to 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.

### Comparison of Multiple Genome Annotations

Annotation distributions in multiple genomes are compared in Figure 3. The genomes are arranged in a doughnut figure (see Table 1 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 as to functional inferences. If the genomes are ordered by their ratios of hypothetical protein to the total number of ORFs in these genomes, then five Archaea genomes will be located at the top 5, with *Aeropyrum pernix* in first place. 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 hypothetical proteins. *Escherichia coli* K12 has the lowest ratio of all, in which only 6% of its 4,248 ORFs are hypothetical. Four other genomes have around 20% hypothetical annotations. If these genomes are ordered by the ratio of proteins with Category I.4 annotations over the total ORFs in these genomes, their ranks are approximately opposite, 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 genomes.

## DISCUSSION

We present an algorithm for high-throughput protein annotations so that multiple sequence analysis tools, biological rules, and functional (PFG) profiles can be 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. Results from the integrated sequence analysis tools for given proteins can be comparatively analyzed. In addition, much-needed guidance is made available to enhance such analysis for an accurate function assignment.

One of the unique features of the algorithm for high-throughput sequence analysis is that protein annotations are clearly categorized based on confidence levels. Annotations with strong support from the knowledge system are categorized at the highest level of confidence because of PFGs, well-defined PFG profiles, and clear-cut feature-function relationships; annotations without such support are considered as tentative. The confidence information will be critical to researchers in deciding to what extent the annotation data can be trusted and will 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, which often have inconsistencies 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].

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 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 missed 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].

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 the 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 committed 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 [23]. 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.

#### 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., Ear, U., and Kyrpides, N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*, **29**, 126-127.
2. Heffelfinger, G.S., Martino, A., Gorin, A., Xu, Y., Rintoul III, M.D., Geist, A., Al-Hashimi, H.M., Davidson, G.S., Faulon, J.L., Frink, L.J., Haarland, D.M., Hart, W.E., Jakobsson, E., Lane, T., Li, M., Locascio, P., Olken, F., Olman, V., Palenik, B., Plimpton, S.J., Roe, D.C., Samatova, N.F., Shan, M., Shoshoni, A., Strauss, C.E.M., Thomas, E.V., Timlin, J.A., and Xu, D. (2002) Carbon Sequestration in *Synechococcus* Sp.: from molecular machines to hierarchical modeling. *OMICS, A Journal of Integrative Biology*, **6**, 305-330.
3. Andrade, M.A., Brown, N.P., Leroy, C., S. Hoersch, A.d. Daruvar, C. Reich, F. A., J. Tamames, V. A., C. Ouzounis, and C. Sander (1999) *Automated genome sequence analysis and annotation*. *Bioinformatics*, **15**, 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 architecture*. *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., N. Larsen, G.D. Pusch, M. D'Souza, J. Selkov E, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov, *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction*. *Nucleic Acids Res.*, 2000. **28**: p. 123-125.
7. Wu, C.H., H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z.Z. Hu, R.S. Ledley, K.C. Lewis, H.W. Mewes, B.C. Orcutt, B.E. Suzek, A. Tsugita, C.R. Vinayaka, L.S. Yeh, J. Zhang, and W.C.

- Barker, *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. 2: 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., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool*. J Mol Biol., 1990. **215**: p. 403-410.
  14. Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L.L. Sonnhammer, *The Pfam protein families database*. Nucleic Acids Res., 2002. **30**: p. 276-280.
  15. Henikoff, J.G., E.A. Greene, S. Pietrokovski, and S. Henikoff, *Increased coverage of protein families with the Blocks database servers*. Nucleic Acids Res., 2000. **28**: p. 228-230.
  16. Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R.R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S.E. Orchard, M. Pagni, D. Peyruc, C.P. Ponting, J.D. Selengut, F. Servant, C.J.A. Sigrist, R. Vaughan, and E.M. Zdobnov, *The InterPro Database, 2003 brings increased coverage and new features*. Nucl. 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., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein, *A Bayesian networks approach for predicting protein-protein interactions from genomic data*. Science, 2003. **302**(5644): p. 449-453.

The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory ("Argonne") under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.