

# RuleMiner: a Knowledge System for Protein Function Annotations

Gong-Xin Yu\*

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

[yug@ornl.gov](mailto:yug@ornl.gov)

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

In this paper, a knowledge system for protein function annotation, called “RuleMiner”, is represented. This system consists of three essential components, **Protein Function Groups (PFGs)**, **PFG profiles** and **rules**. The *PFGs*, established from an integrated analysis of current knowledge of protein functions from Swiss-Prot database and protein family-based sequence classifications, cover all possible cellular functions available in the database. The *PFG profiles* illustrate detailed protein features in the *PFGs* as in sequence conservations, the occurrences of sequence-based motifs, domains 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 RuleMiner is able to provide an enhanced capability for protein function analysis. Such as, results from the integrated sequence analysis tools for given proteins can be comparatively analyzed due to the clear feature-*PFG* relationships. Also, much needed guidance is readily available for such analysis. If the rules describe one-to-one (unique) relationships between the protein features and the *PFGs*, then these features can be utilized as unique functional identifiers and cellular functions of unknown proteins can be reliably determined. Otherwise, additional information has to be provided. Also in this paper, a special section is dedicated to illustrate two real-world applications of this knowledge system in protein function analysis.

Key words: protein function groups, protein features, rules, knowledge system, and protein function annotations.

## 1. Introduction

The massive accumulation of completely sequenced genomes has generated tremendous demands for algorithm developments for systematically analyzing these data and giving them biochemical, physiological, and ecological meanings<sup>1</sup>. One of the earliest and critical steps in this analysis is to elucidate putative functions of open reading frames (ORFs) on these genomes (<ftp://ftp.ncbi.nih.gov/genbank/>). For this purpose, several large, high-throughput sequence analysis systems have been developed including, naming a few, WIT, GenQuiz, MAGPIE, and PEDANT<sup>2-5</sup>. However, there are several problems in those and other annotation systems. If not properly addressed, these problems can result in uncertainties in the annotations.

The function annotations in these systems primarily depend on pair-wise sequence similarity, a feature usually defined by Blast<sup>6</sup> or FastA<sup>7</sup> although a variety of other approaches are often included in the systems. However, protein function is a very complex concept, which includes given protein’s interactions with other proteins, its participation in various cellular processes, regulation, transportation, signal transduction, enzymatic activity, and many other features<sup>1,8</sup>. Complicated evolutionary and structure/function relationships among different proteins may add additional complexity in the annotation process<sup>9</sup>. Thus, it is necessary to integrate all possible features of protein sequences to achieve an enhanced computational capacity in recognizing and differentiating cellular functions.

The said integration is critically important because each of these tools addresses different sequence analysis problems. For instance, the Blast and FastA can reveal sequence similarity at the level of individual amino acids so that they can distinguish, to certain level, homologous proteins<sup>6,7</sup>. However, they could not identify evolutionarily divergent yet functionally related proteins. Information noises, accumulated over millions of year’s evolution, are often overwhelming in the protein sequences<sup>10,11</sup>. As a consequence, biologically meaningful relationships between these protein sequences are made unrecognizable by these tools.

Signature database-based sequence analysis tools represent a different but critical group of approaches for sequence analysis<sup>12-16</sup>. These tools can capture evolutionarily divergent and biologically important protein signatures. Thus, they can identify distant and clear-cut relationships in novel sequences. These tools, however, often fail to detect the differences among some homologous proteins<sup>17, 18</sup>. We can, by that, conclude that the signature defining tools and Blast or FastA can be well complemented with each others and that their integration would considerably increase the overall differentiation capability.

Furthermore, differences also exist among the signature extracting tools. Diagnostically, they have various areas of optimum applications because of different strengths and weaknesses of their underlying analysis methods<sup>19</sup>. Pfam, for example, focuses on the divergent domains<sup>14</sup>, Prosite focuses on the functional sites<sup>20</sup>, and Prints focuses on the families, specializing in hierarchical definitions from super-family down to sub-family levels in order to describe specific functions<sup>13</sup>. Blocks tool provides uncapped multiple alignments for protein families<sup>21, 22</sup>. So then the integrations of these tools would provide a better picture of overall protein feature structures.

All the aforesaid sequence analysis tools, however, are being independently developed. Consequently, these databases are independent and their nomenclature systems are incompatible<sup>19</sup>. Their integration, for this reason, can be enormously difficult and may severely compromise the efficacy in their applications in the protein function annotations. This is possibly one of the main reasons why information collected from signature-defined databases like Pfam and Blocks plays only a minor role in the current sequence analysis systems, instead of a critical one if judged by their exceptional computational capability in recognizing and differentiating protein cellular functions. They are often provided as support evidence for protein function annotations or most often as a supplementary data for user's references. That is why a global reference system is necessitated so that the results from multiple sequence analysis tools can be compared and cross-validated under this system.

Another problem in the annotation systems is that there are no rule-systems developed specifically for the protein annotation as of yet<sup>23, 24</sup>. Ideally, the rules must provide an ability to illustrate the precise associations between these cellular functions and protein features in current protein space. The rules can be, by that, utilized as guidance for protein function annotations. The guidance is especially important when alternative assignments are to be judged on some complicated cellular functions due to multifaceted evolutionary processes such as convergent and divergent evolutions<sup>25, 26</sup>.

Current computational sequence analysis tools, as described above, describe *localized* sequence features no matter whether they represent degrees of pair-wise sequence similarities (e.g., Blast), evolutionarily conserved sequence motifs (e.g., Blocks, Prints)<sup>21, 22, 13</sup>, or biologically important function domains (e.g., Pfam, Prosite)<sup>14, 20</sup>. There are no apparent associations established between the localized features and the specific cellular functions. It is sometimes difficult to infer functions based on occurrence of the localized features because they may exist in hundreds of proteins and are involved in diverse cellular functions<sup>27</sup>. The identification of these features in unknown protein sequences, by itself, will not provide enough evidence to make definite functional assignments if no other information can be acquired. Family-oriented function classification databases such as SCOP, Interpro, and PIR could not provide such guidance either although they have enormous applications in genomics studies<sup>19, 24, 28</sup>. Many protein families defined in these databases often cover many different cellular functions with un-differentiable protein sequences<sup>17</sup>.

So to alleviate the aforementioned problems, we have developed a knowledge system, called RuleMiner. The uniqueness of this system is a simple algorithm developed for high-resolution protein function classifications. Compared to other biological database systems (e.g. Pfam, Prosite, Prints and Blocks), in which all of protein classifications are based mostly on the sequence information, the classification approach is unique in that it combines both knowledge-based and sequence-based protein analysis. In addition, the classification is based solely on information from Swiss-Prot

protein database, a highly reliable function resource<sup>29</sup>. This integration increase resolution in protein function classifications due to the fact that it goes beyond limitations, suffered by the sequence-only-based protein classification algorithms, especially when proteins with different functions share an extensive homology<sup>19,30</sup>. As a result, a series of high resolution of protein function groups (*PFGs*) are developed, which acquire an ability to possibly describe the smallest functional or evolutionary units, encoded possibly by individual protein encoding genes.

Another unique component is the establishment of *PFG profiles*. The *PFG profiles* illustrate diversity of proteins in the *PFGs*, including the variability of protein features such as sequence conservations, the occurrences of sequence-based motifs and domains, and their species distributions (Archaea, bacteria, Eukaryote, plasmid, Mitochondrion, chloroplast and virus as defined in Swiss-Prot database). This information is important for the rule-inferring purpose. Only individual *PFGs* are built to represent unique cellular functions, and the *PFG profiles* be precisely established to characterize these functions, then rules can be clearly defined to illustrate the unambiguous relationships between the functions and the protein features. Consequently, the RuleMiner is able to provide an enhanced capability for protein function annotations as followings; 1) Results from the integrated tools for given proteins can be integrated due to the clear feature-*PFG* associations; 2) much needed guidance or rules are readily available for the complex function annotations. If the rules describe one-to-one (unique) relationships between the protein features and the functions, then, these features can be utilized as unique functional identifiers so that cellular functions of unknown proteins can be reliably determined. Otherwise, additional information has to be provided to pinpoint the detailed cellular functions.

In this paper, we first introduce some definitions and notations that will be used in this paper. Then, we will describe in detail procedures for protein classification, protein feature characterization, and rule establishments. Lastly, we will illustrate the applications of this information system in protein function analysis, possible problems, and potential solutions to these problems.

## **2. Definitions and notations**

### **2.1. Protein features**

Main protein features are defined as localized sequence features, detected by some highly dedicated computational sequence analysis tools. They can represent degrees of pair-wise sequence similarities or evolutionarily and functionally conserved motifs and domains. Addition protein features include conservation of protein sequences within given cellular functions and occurrences of organism species in these functions.

Currently, three sequence analysis tools are utilized in our research. They are Blocks, Blast and Pfam. Each of these tools is able to identify different protein sequence features. To start, the Blocks are ungapped, multiple alignments corresponding to the most conserved regions of given protein families<sup>21,22</sup>. Blocks analysis pinpoints the conserved blocks corresponding to certain protein families. The results of the analysis can assign these proteins to one or possibly more homologous Blocks protein families.

While the Blocks pinpoints the localized motif of conserved regions and relationships between individual proteins to the sequence profiles of protein families, Blast algorithm emphasizes more on pair-wise amino acid sequence similarities<sup>6</sup>. And compared with Blast and Blocks, Pfam database is unique in that it contains curated multiple sequence alignments for each family, as well as profile hidden Markov models for finding these domains in new sequences<sup>14</sup>. Thus, Pfam search results can determine sequence features that illustrate the occurrence of certain biologically important domains.

### **2.2. Protein function groups (*PFGs*)**

The establishment of *PFGs* is intended to describe the smallest biochemical or evolutionary units encoded possibly by individual protein-encoding genes.

Let

$$K_X = (k_{x1}, k_{x2}, \dots, k_{xi}, \dots, k_{xn}) \quad k_{xn} \in \text{Swissprot} - \text{protein} - \text{database},$$

where  $k_{xn}$  is a **Knowledge-based Protein Function Category (KPFC)**, derived from a text mining procedure on protein function descriptions in Swiss-Prot protein database.

Let

$$F_X = (f_{x1}, f_{x2}, \dots, f_{xj}, \dots, f_{xm}) \quad f_{xm} \in \text{Blocks} - \text{database},$$

where

$f_{xm}$  is a **Sequence-based Protein Function Category (SPFC)**, established based on the best hit of Blocks analysis.

Then

$$\forall_{\text{protein}} S \in \text{pfg}(k_{xi}, f_{xj}) \text{ if } S \text{ has } k_{xi} \text{ in the Swiss-Prot database and the } f_{xj}$$

in the Blocks database,

where

$$\text{pfg}(k_{xi}, f_{xj}) \text{ is defined as a } \text{Protein Function Group (PFGs)}.$$

The definition of protein function groups is more conceptual in which both *KPFCs* and *SPFCs* are assumed to be determined in parallel. Then, their intercepts are defined as *PFGs*. In practice, the protein function classifications are hierarchical in which *KPFCs* are classified based on the analysis of function descriptions in the Swiss-Prot database; then, proteins in each of the category are further classified into smaller function categories, *SPFCs*, based on the Blocks analysis. It is the *KPFCs* and *SPFCs* pairs that define protein function groups (*PFGs*).

### 2.3. PFG profiles

Let us define some statistic parameters to characterize *PFGs* before we introduce the concept of *PFG profiles*.

**2.3.1. Confident value vectors:** Corresponding to each protein function group *PFGs*, a confident value vector is  $V_X = (v_{x1}, v_{x2}, \dots, v_{xn})$ , where  $n$  is the number of proteins in the *PFGs*. For Blocks-based similarity indicators,  $v_{xi}$  is logarithm of *E*-value between the  $i^{\text{th}}$  protein  $x$  and the protein family  $\text{spf}c_j$ ; for Blast-based similarity indicators  $v_{xi}$  is the average of logarithm of *E*-value between the  $i^{\text{th}}$  protein  $x$  and every other protein members in the *PFGs*. *E*-value is a parameter to evaluate the sequence similarity in the computational sequence analysis tools, e.g. such Blast, Block and Pfam.

**2.3.2. Statistic parameters:** To evaluate the sequence conservation of proteins in the protein function groups *PFGs*, series of statistic parameters are calculated. These parameters are  $v_{\text{max}}$ : the highest confidence values;  $v_{\text{min}}$ : the lowest confidence values;

$\mu_X$ : the average defined as  $\mu_X = \sum v_{xi} / n$ ;  $m_X$ : median, and  $C.V._X$ : the co-efficiency of variance, defined as  $C.V._X = \delta_X / \mu_X * 100$ , where

$\delta_X = \sqrt{\sum (v_{xi} - \mu_X)^2 / (n - 1)}$ . Generally the C.V. from 0% to 10% indicates that data is consistent while those close to or greater than 50% indicates that there is a tremendous variability.

**2.3.3. PFG profiles (pfgPF):** To describe the diversity of the proteins in the *PFGs*, including the variability of protein features such as sequence conservations, the occurrences of sequence-based motifs and domains, and species associations. The species categories are defined 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>).

Let

$pf\text{g}PF = (S, Blk, Pm, BlkSp, PmSp)$ , where

1.  $S$  illustrates protein sequence conservations characterized by a statistics parameter vector  $P = (v_{\max}, v_{\min}, \mu_X, m_X, C.V._X)$ .
2.  $Blk$  illustrates the diversity of Blocks motif patterns describing the occurrence of conserved blocks in protein families  $spf\text{c}_j$ .
3.  $Pm$  illustrates the diversity of domain patterns describing the occurrence of biologically important biological domains (Pfam).
4.  $BlkSp$  describes the unambiguous relationships between  $Blk$  and the Swiss-Prot species categories.
5.  $PmSp$  describes the unambiguous relationships between  $Pm$  and the Swiss-Prot species categories.

#### 2.4. Rule determination

The rules are defined as instantly recognizable relationships between individual  $PF\text{G}$ s and protein features defined in the  $PF\text{G}$  profiles.

Let

$Rule = Relationship(F, PF\text{G}_n)$ .

where  $F$  is the protein features defined in the profiles,  $PF\text{G}$ s are protein function groups, and

$n \in (1, 2, \dots, k)$ , where,  $K$  represents any integer.

If  $n$  is equal to one, then the relationships between the features and the  $PF\text{G}$ s are unique, in which one feature corresponds to one particular  $PF\text{G}$ . Otherwise, the relationships are non-unique, in which one feature may correspond to multiple protein function groups or vice versa.

### 3. Method and algorithm

To build RuleMiner, we followed a three-step procedure to process functional and sequence data in the Swiss-Prot protein database<sup>29</sup>. We first classify Swiss-Prot proteins into  $PF\text{G}$ s. Then, we generate  $PF\text{G}$  profiles. Lastly, we extract rules to illustrate unequivocal relationships between the protein features defined in the  $PF\text{G}$  profiles and individual  $PF\text{G}$ s. The detailed procedure is described as following.

#### Step 1: Classify Swiss-Prot proteins into protein functional groups

The central idea of the algorithm is the integration of the current knowledge of well-defined protein functions and highly confident protein family information to establish highly differentiable  $PF\text{G}$ s in protein functions. To achieve this goal, we downloaded a version (6/25/2002) of the Swiss-Prot protein database, and extracted the function descriptions in the “DE” fields. Based on these descriptions, the Swiss-Prot proteins were assigned to either an “enzymatic” category if their EC numbers were known, otherwise, a “non-enzymatic” category.

The “enzymatic” proteins were then classified based on their EC numbers. For “non-enzymatic” proteins, text mining algorithms and controlled vocabularies were devised to extract phrases with functional meanings from the protein function descriptions. This resulted in the formation of **Knowledge-based Protein Function Categories (KPFs)**. The resultant **KPFs** were verified manually. Such a check is essential for quality control because of the annotation inconsistencies across various database entries.

The protein sequences in the *KPFCs* were further processed with the Blocks tool (Henikoff et al. 2000). For the classification purpose, the algorithm assigned the best Blocks hit to each protein. Based on these assignments, proteins within the *KPFCs* were further divided into smaller groups, named *Sequence-based Protein Function Categories (SPFCs)*. As a consequence,  $m=21,656$  *PFGs* were established. Each of the *PFGs* was defined by a pair of (*KPFC*, *SPFC*).

### **Step 2: Generate *PFG profiles***

To generate *PFG profiles*, we analyzed protein sequences within every *PFG* with Blast and Pfam. First, the protein patterns (evolutionary conserved motifs in Blocks and biologically important domains in Pfam) were evaluated to describe the pattern diversities. Then, associations were established between these patterns and species categories defined in Swiss-Prot database (e.g. Archaea, Eubacteria, Eukaryote, Plasmid, Mitochondrion and Chloroplast). Last, sequence conservations were evaluated (Please see Definitions and Notations). As a result, detailed *PFG profiles* were established for all of the *PFGs*.

### **Step 3: Define rules**

The aim of this step is to build instantly recognizable relationships between the protein features in the *PFG profiles* and individual cellular functions (*PFGs*) within the protein space covered by Swiss-Prot protein database. For this purpose, we first extracted lists of protein features either Blocks motif patterns or the patterns of biologically important Pfam domains from the *PFG profiles*. Then we determined, for each particular protein feature, what *PFG(s)* it may associate with. The number of the *PFGs* associated with the protein feature can be one or more. If it is one, then a one to one unique relationship will be constructed. Otherwise the relationships are non-unique in which one protein feature corresponds to multiple cellular functions or vice versa. The relationships, unique or not, define the rules, which will provide guidance in protein function determinations (see the Definitions and Notations for details).

## **4. Results**

### **4.1. *PFGs have a capacity to differentiate components in protein complexes***

One of the most significant components of the knowledge system is *PFG* in that the detailed information of given protein complexes can be illustrated. The information includes, but not limited to, subunits, species-specific protein units, and highly homologous but substrate-specific functional units (Table 1).

The species-specific functional units represent one of unique classes of *PFGs* in the Knowledge-based protein analysis system (Table 1, Section A). These groups can identify different evolutionary versions of functional proteins, which are possibly reinvented at separate times by so-called convergent evolution<sup>31-33</sup>. This can best be illustrated by RNA and DNA replication machines. Archaea and Eukaryotes have extensive similarities in the principal protein components of these machines but are dramatically different from that of Eubacteria<sup>34,35</sup>. Another unique class includes *PFGs*, which are highly homologous but perform different cellular functionalities. In Zinc-containing alcohol dehydrogenase protein family, 14 enzymatic *PFGs* have been classified, representing functionalities with distinct substrate specificities despite the fact that proteins in this family are essentially un-differentiable by current sequence analysis tools (Table 1, Section B). Many more examples can be found in [http://www-wit.mcs.anl.gov/Gongxin/cgi-bin/access\\_data.cgi](http://www-wit.mcs.anl.gov/Gongxin/cgi-bin/access_data.cgi) or <http://www-wit.mcs.anl.gov/svmmmer/>.

Table 1. Two example categories of function units described by the *PFGs*

<i>PFG</i>		Function Description	Species Distribution <sup>a</sup>
<i>KPFC</i>	<i>KPFC</i>		
A. Subunits and species-specific version of protein function units			
EC 2.7.7.6		Bacterial Version of DNA directed RNA polymerase	
	IPB001700	Bacterial RNA polymerase, alpha chain	B <sup>b</sup> :41 E <sup>c</sup> :2 chl <sup>d</sup> :31 cynal <sup>e</sup> :1 mit <sup>f</sup> :1
	IPB000722	RNA polymerase, alpha subunit	B:27 E:5 V:4 chl:13 cynal:1
	IPB003716	RNA polymerase omega subunit	B:20
	IPB001572	RNA polymerases beta subunit	A <sup>g</sup> :14 B:29 E:22 V <sup>h</sup> :5 chl:15 cynal:1 plasm <sup>i</sup> :1
		Archaea and Eukaryote version of DNA directed RNA polymerase	
	IPB001514	RNA polymerases D/30 to 40 Kd subunits	A:11 E:10
	IPB002092	Bacteriophage-type RNA polymerase family	E:6 V:4 mit:6
	IPB001572	RNA polymerases beta subunit	A:14 B:29 E:22 V:5 chl:15 cynal:1 plasm:1
	IPB003221	DNA directed RNA polymerase, 7 kDa subunit	E:4
	IPB000268	RNA polymerases N/8 Kd subunits	A:10 E:6 V:7
	IPB001529	RNA polymerases M/15 Kd subunits	A:5 E:7
	IPB001725	RNA polymerases K/14 to 18 Kd subunits	A:9 B:1 E:6 V:3
	IPB000783	RNA polymerase H/23 kD subunit	A:11 E:4
	IPB001306	RNA polymerases L/13 to 16 Kd subunits	A:7 E:12
B. Highly homologous but substrate-specific functional units			
EC 1.1.1.1			
EC 1.2.1.1			
EC 1.1.1.73	IPB002328	Alcohol dehydrogenase class III	E:1 B:4 plasm:1 E:16
EC 1.1.1.1			
EC 1.2.1.1	IPB002328	Alcohol dehydrogenase class III	
EC 1.1.1.1	IPB002328	Alcohol dehydrogenase	A:1 B:7 E:73 plasm:1
EC 1.2.1.1	IPB002328	Formaldehyde dehydrogenase (glutathione)	B:2 E:4
EC 1.1.1.195	IPB002328	Cinnamyl-alcohol dehydrogenase	E:18
EC			E:10
1.1.1.255	IPB002328	Mannitol dehydrogenase Mycothiol-dependent formaldehyde dehydrogenase	B:1 plasm:1
EC 1.2.1.66	IPB002328	Aryl-alcohol dehydrogenase	B:2 E:7
EC 1.1.1.14	IPB002328	L-iditol 2-dehydrogenase	B:4
EC			
1.1.1.103	IPB002328	L-threonine 3-dehydrogenase	B:1
EC 1.2.1.46	IPB002328	Formaldehyde dehydrogenase	E:1
EC 1.1.1.9	IPB002328	D-xylulose reductase	B:1
EC			
1.1.1.251	IPB002328	Galactitol-1-phosphate 5-dehydrogenase	B:4 E:2
EC 1.1.1.2	IPB002328	Alcohol dehydrogenase (NADP+)	

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

#### 4.2. *PFG profiles can illustrate diversity in the evolution of given protein functions*

The sequence conservation is one of the most important features in the *PFG profiles*. A total of 8834 *PFGs* were analyzed. They represent protein function groups with an average of at least two proteins so that the profiles could be calculated (See Definitions and Notations for details). The mean *E*-values and co-efficiency of variations (*C.V.*) are two main parameters to measure the degree of protein sequence conservation. Ten categories of conservations are defined based on the *C.V.*, each of which has 10 point ranges, e.g. a *C.V.* of 0 to 10 is grouped as category 1; a *C.V.* of 10 to 20 as category 2; until a *C.V.* of 90 and greater, which is assigned as category 10. Those with a mean *E*-value of less than  $1e-20$  and a *C.V.* of less than 10 are considered as highly conserved *PFGs*. Blast and Blocks analysis both reveal that approximately 16.34% of the *PFGs* belong to this category. However, *PFGs* distribution among the 10 conservation categories is more biased to highly conserved category (e.g. category 1 or 2) in Blocks than that in Blast (Data not shown). The latter is much flatter and closer to a normal distribution, indicating that their variations are much greater. Such a phenomenon is not surprising because the Blocks analysis focuses on the conserved motifs of given protein families while Blast analysis on an overall sequence similarity.

In addition to sequence conservation, other protein features in the profiles are also helpful to characterize the *PFGs*. Among these features, the occurrences of Pfam domain patterns, Blocks motifs patterns, and their relationships with Swiss-Prot species categories are especially important. Both Pfam domains and Blocks motifs are clear indicators of biologically important functional units and they represent some basic function properties for the *PFGs*. Two extreme categories of variations are illustrated in Table 2 to demonstrate the characterization capability of the *PFG profiles*. The first category includes genes encoding methylmalonate-semialdehyde dehydrogenase (EC 1.2.1.27), Photosystem 44 kDa reaction center protein, photosystem D2 protein, photosystem P680 chlorophyll A, apoprotein and preprotein translocase. These *PFGs* are highly consistent in Block motifs and have extremely conserved Pfam domains. They exist in all varieties of life domains and may represent some of the core components in living organisms with stringent system requirements<sup>36</sup>. The second category includes viral-gene encoding proteins for RNA-directed RNA polymerase, apoptosis inhibitors, and nucleocapsid proteins. These *PFGs* are extremely variable in both sequence conservations and the occurrences of the Blocks motifs and Pfam domains. They may represent those required for frequent adaptation and evolution, which is critical for pathogens such as infectious bacteria and viruses, where constant change and adaptation are required for surviving various defense systems in their host<sup>37,38</sup>.

**Table 2.** Characterization of protein functional groups based on Blocks and Pfam searching results

<i>PFG</i> <sup>a</sup>		Characterization of protein population within Protein Functional Groups				
<i>KPFC</i> <sup>b</sup>	<i>SPFC</i> <sup>c</sup>	<i>C.V.</i> <sup>d</sup>	Blocks Pattern	Species	Pfam patterns	Species
<b>Highly conserved protein functional groups</b>						
EC 1.2.1.27 (Methylmalonate-semialdehyde dehydrogenase)	IPB002086	3.54	ABCDE <sup>e</sup> :6	B E	Aldedh <sup>f</sup>	B:2 E:4
EC 1.2.1.9 (Glyceraldehyde-3-phosphate dehydrogenase)	IPB002086	4.19	ABCDEF:4	B E	Aldedh	B:1 E:3
EC 2.7.1.48 (Uridine kinase)	PR00988	3.10	ABCDEF:26	B E	PRK	B:19 E:7
EC 1.18.96.1 (Superoxide reductase)	IPB002742	10.05	ABC:4	A B	Desulfoferrodox	A:3 B:1
Photosystem 44 kDa reaction center protein	IPB000932	4.18	ABCDEF:24	B E chl cynal B E chl	PSII	cynal:1 B:3 chl:18 E:2 cynal:1 B:4
Photosystem D2 protein	IPB000484	2.82	ABCD:26	cynal	photoRC	chl:17 E:4
<b>Highly diverged protein functional groups</b>						
EC 1.2.99.5 (Formylmethanofuran dehydrogenase)	IPB002489	146.0	ABCDEF:2 BC:5 ABC:1	A	DUF14 Molydop_bindi ng	A:2
					DUF14	A:6
EC 4.2.2.2 (Pectate lyase)	PR00807	145.4	DEG:1 BCDE:1 ABCDEFGH: 4 CD:1 DE:12 ADE:1 CDE:2	B E	pec_lyase pec_lyase pec_lyase	B:1 B:15 E:6
EC 4.1.1.23 (Orotidine-5-phosphate decarboxylase)	IPB001754	71.97	ABE:15 ABCDEF:45 E:2 BCE:4 AB:1 AE:7 ABCE:3 BE:6 CE:2	A B E	OMPdecase OMPdecase PcrB NO Pfam pattern	B:1 A:1 A:1 B:1
					OMPdecase	A:6 B:25 E:43
Apoptosis inhibitor	IPB001370	75.10	C:2 AC:1 ABC:6	E V	BIR BIR BIR BIR BIR zf- C3HC4	V:1 E:1
					BIR BIR BIR zf- C3HC4	V:2 V:3 E:1
Glycoprotein	PR00668	64.69	ABCDEFGF:3 ACDF:5 ABCDEF:1 ABCDFG:2	V	Marek_A Marek_A Marek_A	V:3 V:8
E.C 2.7.7.48 (RNA-directed RNA polymerase)	IPB000224	73.49	ABCDE:4 ABE:3	V	Phosphoprotein	V:7

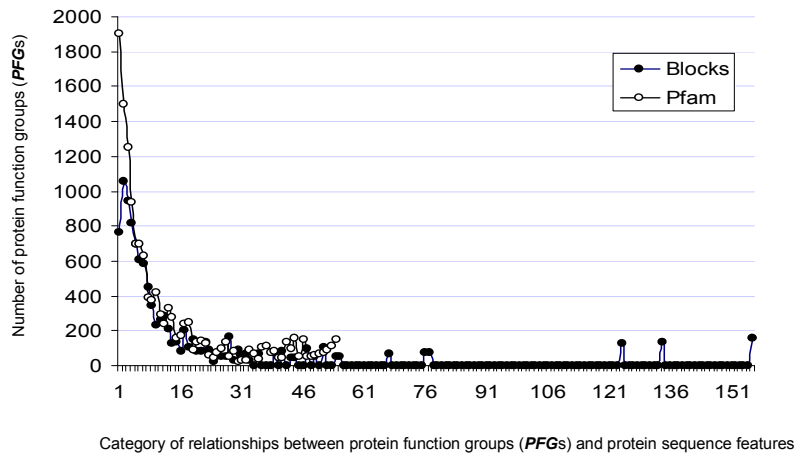
<sup>a</sup>*PFG* stands for protein function group; <sup>b</sup>*KPFC* for knowledge-based protein function category; <sup>c</sup>*SPFC* for sequence-based protein function category; <sup>d</sup>*C.V.* for Co-efficiency of variance; <sup>e</sup>ABCDE for a Blocks Pattern, each letter represents a Blocks motifs; <sup>f</sup>Aldedh for Pfam domain.

### 4.3. Rules can define clear relationships between the PFGs and protein features

Comparative analysis of the protein function groups (*PFGs*) and protein features in their profiles has generated clear cut relationships between protein features and individual cellular functions (Fig. 1). The relationships, in some cases, can be highly specific, in which one given feature corresponds to one given *PFG*. However, a majority of the relationships are especially complex, in which one feature may correspond to multiple functions. These relationships, to that extent, will have a profound affect on the sequence feature-based function annotations and their confidences.

If it was one to one unique relationships, then the protein features can be utilized as unique function identifiers so that the cellular functions for unknown proteins can be reliably determined. For instance, both RuBisCO\_large and RuBisCO\_small are unique to their cellular functions. If these domains are detected within the unknown proteins, it is highly likely that these proteins are large chain or small chain for Ribulose biphosphate carboxylase. Therefore, we can be positive that these proteins are involved in carboxylation/oxygenation of ribulose-1,5-biphosphate (RuBP)<sup>39</sup>, a critic step for carbon fixation in Calvin cycle.

Otherwise, it is going to be very difficult to pinpoint the detailed cellular functions without additional information. For example, 7tm\_1 is a ubiquitous pfam domain common to G-protein-coupled receptors GPCRs<sup>40-42</sup>. However, the GPCRs constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). Thus, the detection of this domain in a given protein could not lead to the detailed function annotation if no additional information was present. In summary, the rule-based system can give useful information not only to decide how confident given annotations can be but also to judge when additional evidence are needed to determine unique functions for unknown proteins.



**Fig. 1.** The relationships between protein function groups (*PFGs*) and protein sequence features. X-coordinate represents different categories of such relationships, e.g. “1” represents a one-to-one unique relationship, in which a given protein feature (Pfam domain or Blocks motif pattern) occur in one single *PFG* while others represent non-unique relationships, in which one given protein feature can occur in multiple *PFGs*. Y-coordinate represents the number of *PFGs* that are involved in each of the categories.

## 5. Application of the RuleMiner system

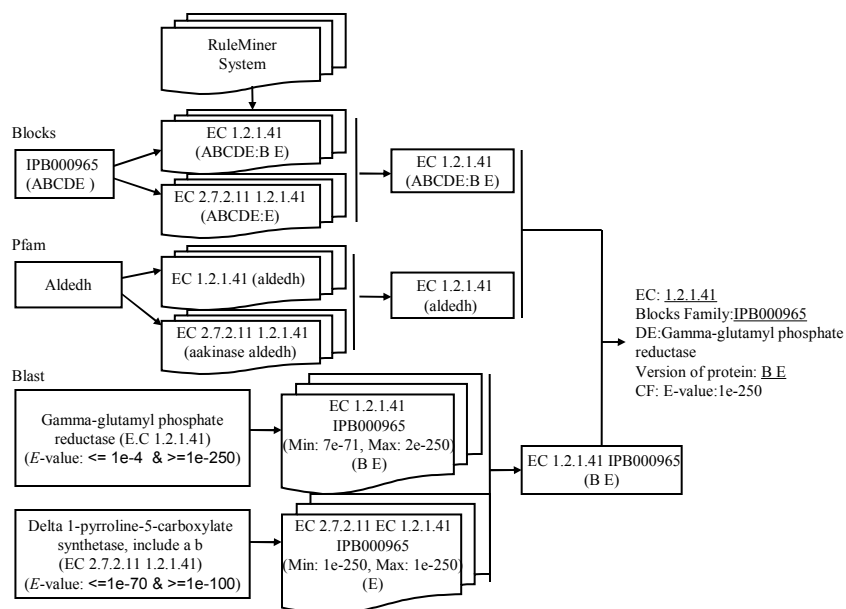
The section describes two real-world applications of the RuleMiner system in the protein function annotations.

### 5.1. WIT3 for protein function annotations

The RuleMiner, including three critical components: the *PFGs*, the *PFG profiles*, and the derived rules, has been the foundation for the development of WIT3 system for high-through protein sequence annotations in Argonne National Laboratory (<http://www-wit.mcs.anl.gov/Gongxin>). Each of these parts plays an important role in the development as demonstrated in the following annotation procedure. For an unknown protein,

1. Analyze with Blast Blocks and Pfam (Remember that they are the same sets of tools involved in the RuleMiner development!). Results include all homologous proteins and their corresponding *E*-values from Blast data. Additional information included in the Blast results is detailed functional descriptions and **Knowledge-based Protein Function Categories (KPFs)** of these homologous proteins (Use the same procedure to extract *KPFs* based on the detailed functional descriptions). For Blocks, results include Blocks families: the **Sequence-based Protein Function Categories (SPFs)**, 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 patterns in which domains are arranged on the proteins in the way that there are no overlaps.
2. Query the RuleMiner with the results e.g. *KPFs*, *SPFs* and Pfam domains from step 1. Outputs are assignments of possible *PFG(s)* and organized by the sequence analysis tools. Additional evaluations of these assignments include a procedure to determine if they are consistent with the *PFG profiles* of these candidate *PFGs* (e.g. the domain or motif-function relationships, the ranges of conservations, and the species-*PFG* associations).
3. Compare the *PFGs* among the three different tools to determine if they are consistent. If not, a voting procedure will be applied to vote for a winner or winners (Unpublished data). The principle is that if any assignments of the *PFG(s)* can be unambiguously determined by a single or combination of protein features, then it is most likely that the *PFG(s)* is/are the right function assignments. The results are that either single (unique) or a group of (undifferentiable) *PFG(s)* are assigned to the annotated proteins.

An example of the application is illustrated in Fig. 2. gi|1786438 is one of over 4200 open reading frames (ORF) in the genome of *Escherichia coli* K-12 MG1655. As a result of the analytic process, The ORF is annotated as gamma-glutamylphosphate reductase with protein function group of *pfg*(EC 1.2.1.41, IPB000965). In this example, Pfam result and its associated rules (unique relationships between Pfam domains and the *PFGs*) in the RuleMiner system provides a strong differentiation capacity, thus, giving a highly reliable function annotation to this ORF.



**Fig. 2.** Application of the RuleMiner system in the annotation of gi|1786438 of *Escherichia coli* K-12 MG1655 (Bacterium). The annotation is based on the sequence features identified by three sequence analysis tools and information in the RuleMiner system. As a result of comparative analysis, gi|1786438 is annotated as gamma-glutamylphosphate reductase with the protein function group of *PFG*(EC 1.2.1.41, IPB000965). The Pfam result and its associated rules (unique relationships between Pfam features and the protein function groups) provide the strongest differentiation ability in this example. “Aldedh” is unique to *PFG*(EC 1.2.1.41, IPB000965) while an additional domain “aakinase” is needed for *PFG*(EC 2.7.2.11 1.2.1.41, IPB000965). Feature-species associations provide an additional resolution. *pfg*(EC 1.2.1.41, IPB000965) occurs in both Eukaryote and Eubacteria, which is consistent with the source of our targeted gene. In contrast, *PFG*(EC 2.7.2.11 1.2.1.41, IPB000965) occurs only in Eukaryote.

## 5.2. SVMMER for functional differentiation of highly homologous PFGs

Evolutionary processes create complicated relationships between protein sequence/functions<sup>19, 24, 28</sup>. One of common phenomenon for such complexity is that proteins with highly homologous sequences perform clearly different functions, e.g. substrate specificities<sup>26</sup>. This phenomenon presented an immense problem in protein annotations. None of the current sequence analysis algorithms can differentiate these proteins. Development of algorithms to tackle this problem is one of the highest priorities in protein function analysis.

The RuleMiner is able to classify these highly homologous proteins into different *PFGs* based on their individual functions, e.g. substrate specificities. For example, Blocks’ Zinc-dependent dehydrogenase protein family (IPB002328) covers multiple *PFGs* (Table 1). All of these enzymes share similar catalytic mechanism; of which zinc atoms play an essential role<sup>17</sup>. Additional 134 homologous protein groups (organized by protein superfamilies) can be visualized in <http://www-wit.mcs.anl.gov/svmmer/>. This property of the rule system gives us a clear advantage. Instead of struggling in noise, heterogeneous databases of proteins in number of thousands, we can concentrate ourselves on a limited number of highly homologous *PFGs*. Supervised data mining

algorithms can be efficiently applied to find differentiable features among these otherwise un-differentiable homologous function groups.

SVMMER, a discriminative-tool-based prediction approach, is developed, based on the RuleMiner system to differentiate these highly homologous *PF*Gs and predict functions for unknown proteins (<http://www-wit.mcs.anl.gov/svmmmer/>). The main purpose for this algorithm is, given two well-defined, homologous *PF*Gs and their related protein sequences, to find discriminating features to separate these *PF*Gs. Support vector machine (SVM), a supervised margin classifier, was used as discriminator<sup>43</sup>. The basic procedure is as following:

1. Construct multiple sequence alignment (MSA) from these highly homologous protein sequences;
2. Build scoring vectors for member sequence based on the MSA and its Hmm model. The length of the scoring vectors corresponds to the number of positions in the MSA and their values to that of the Hmm scores.
3. Establish training classes of the scoring vectors for the SVM classifier. The class labels (+/-) for the scoring vectors correspond to two targeted *PF*Gs.
4. Train and test SVM data models.
5. Predict functional classes for unknown proteins based on the model.

Our underlying assumption is that information encoded in protein sequences dictates their functions in cellular environments, including their selectivity for substrates, cofactor, and other binding ligands. If that is the case, the SVM classifier would be able to recognize the differences between two highly homologous *PF*Gs and build a hyperplane in high dimensional spaces to separate them. SVMMER has tested 10 groups of *PF*Gs in the RuleMiner system and results demonstrated that this method can achieve an average of 95% accuracy in functional differentiations and protein function predictions, illustrating its effectiveness in tackling with the bioinformatics complex problem.

## 6. Discussion and Conclusion

In this paper, we tried to address two problems that are closely associated with the uncertainties that are likely to occur in the protein function annotations. To begin, computational sequence analysis tool integration represents an enormous problem. The 1999's emerge of Interpro, an integrated documentation resource of protein families, domains, and functional sites, has greatly accelerated such research. It amalgamates the major protein signature databases into one comprehensive resource<sup>19</sup>. However, the functional classifications in Interpro mainly focus on protein families, not aimed to provide a resolution as required in this study. To advance this research and achieve our goal in building the high-resolution knowledge-based protein analysis system, we present a unique form of cellular function categorizations, based on a simple protein function classification algorithm. The main purpose of the protein function classification is to establish high-quality representatives of protein cellular functions. Ideally, these representatives must be universal so that they can describe individual functions across different species and, yet, accurate enough to represent unique functions, encoded possibly by individual protein-encoding genes. This is the primary reason why we include the knowledge-based function analysis into our classification algorithm because functional knowledge in Swiss-Prot adds additional information, so additional differentiation capability.

The second problem is that there is no commonly recognized guidance in the protein function annotations. Accordingly, we are equally enthusiastic about building a system that has capabilities to define principles or rules for such guidance. Only when the *PF*Gs are clearly established and their protein features accurately characterized, can unequivocal relationships (rules) between individual functions and the protein features be established. We have shown that the relationships, in some cases, are highly specific in which one feature corresponds to one function. We also demonstrated that the relationships are very complex. One feature may correspond to multiple protein function

groups or vice versa. These relationships, thus, will have a useful affect on the sequence feature-based function annotations. If it is one to one unique relationships, the protein features can be utilized as unique functional identifiers so that cellular functions of unknown proteins can be reliably determined. Otherwise, additional information has to be provided to pinpoint the detailed cellular functions or additional algorithms e.g. SVMMER, have to be integrated.

The *PFG profiles* can provide additional information for the protein function annotation. The reason is that the quality of the predictions largely depends on, in most cases, the recognition and differentiation capability of the analysis tools on homologous proteins. In the course of evolution, different protein families have diverged to a different extent<sup>19,24,28,44</sup>. Therefore, flat cutoff scores, commonly used in the function determinations, cannot provide reliable separations<sup>2-5</sup>. In contrast, the *PFG profiles*, which describe the sequence conservations, the diversities of protein features (motifs and domains), and the occurrences among organism species, can provide necessary information to determine dynamic cutoffs in protein function annotations.

Despite of the aforementioned properties of the RuleMiner system, there are number of problems that have to be addressed. First, number of protein function groups (*PFGs*) is un-proportionally large considered a limited number of Swiss-Prot proteins involved in the process. Over 21,000 *PFGs* are established among 74,000 Swiss-Prot proteins (a text mining procedure filtered out proteins, which function annotations deem to be unreliable, e.g. proteins with function descriptions containing hypothetical or putative), an average of less than 4 proteins per *PFG*. Format of function descriptions in Swiss-Prot database may be partially responsible for this problem. Ideally, protein with the same functions would have same functional descriptions such as EC numbers for enzymatic proteins so that they can be classified as the same *KPFCs* by the knowledge-based function classification procedure. However, some abnormalities in the function descriptions in Swiss-Prot database, especially for proteins, which are less studied and non-enzymatic, cause significant problems. In some cases, proteins with the same functions can not be clustered together due to the abnormalities.

Second, a biased representation of protein functions in the database may cause an additional problem. Some functions are extremely over-represented e.g. alcohol dehydrogenase (132 proteins) and Rubisco (526 proteins) while majority of them have only one protein entry. As a consequence, less than half (8834) of the *PFGs* are covered in the *PFG profiles*. The rules based on these profiles, therefore, have some limited coverage of protein functions. The third problem is that some proteins and their functions can not be classified into protein function groups (*PFGs*) at all. Table 3 gives a list of proteins that have no *PFG* assignments, which further limits the coverage of protein functions by the RuleMiner system.

The first two problems, however, are highly associated with data quality of Swiss-Prot database. With an improvement of data quality and an expansion of protein function coverage in the database, the importance of these problems are likely to lessen. The third problem is due to a lack of protein families in the Blocks database. One of the possible reasons 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 (<http://www.blocks.fhcrc.org/>). The current Blocks database covers approximate 2000 protein families vs. over 8400 in InterPro release 7.0 (<http://www.ebi.ac.uk/interpro/>). An obvious solution for this problem is to update Blocks database based on Interpro protein families, which would significantly increase coverage of cellular functions by the RuleMiner system.

Table 3. Proteins with no *PFG* assignments for eight types of cellular functions

<i>KPFC</i>	Number of proteins with no <i>PFGs</i>	Species Distribution
H(+)-transporting two-sector ATPase	75	A: 3 B:8 E:56 mit:8
Transcriptional_activator	25	B:4 E:17 cynal:1 chl:3
DNA-directed RNA polymerase	30	A:3 B:1 E:13 V:13
Cytochrome-c oxidase	24	B:5 E:19
DNA-directed DNA polymerase	50	A:7 B:18 E:11 V:8 mit:4 Plasm:1
Transcription_factor	19	E:18 B:1
Replication_protein	9	B:2 E:6 V:1
Elongation_factor	5	A:5

### Acknowledgements

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

### References

- Zhou, J.Z., Miller, J.M. (2002) Microbial genomics – challenges and opportunities: the 9<sup>th</sup> international conference on microbial genomes. *J. of Bacteriology*, 184: 4327-4333.
- Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, Kyrpides N, Fonstein M, Maltsev N, Selkov E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28:123-5.
- 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/>.
- Gaasterland T. and Sensen C. (1996) Fully automated genome analysis that reflects user needs and preferences—a detailed introduction to the MAGPIE system architecture. *Biochimie* 78 302–310. <http://genomes.rockefeller.edu/magpie/>.
- Frishman D. and Mewes H. (1997) Pedantic genome analysis. *Trends Genet.* 13 415–416. <http://pedant.gsf.de/>.
- 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.
- Lipman D. J. and Pearson W. R. (1985) Rapid and sensitive protein similarity searches. *Science* 227 1435-1441.
- Riley, M. (1998) Systems for categorizing functions of gene products, *Curr. Opin. Struct. Biol.*, 8, 388-92
- Strauss EJ, Falkow S. (1997) Microbial pathogenesis: genomics and beyond. *Science*, 276:707-712.
- Huynen MA, Bork P. (1998) Measuring genome evolution. *Proc Natl Acad Sci U. S. A.*, 95:5849-5856.
- Bansal, A.K., Bork, P. and Stuckey, P.J. (1998) Automated pair-wise comparison of microbial genomes, *Math. Model. Sci. Comput.*, 9, 1-23.
- 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.

13. 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.
14. Bateman A. Birney E. Cerruti L. Durbin R. 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.
15. 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.
16. 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.
17. Plapp B.V., Sun H.-W. (1992) Progressive sequence alignment and molecular evolution of the Zn-containing alcohol dehydrogenase family. *J. Mol. Evol.* **34**: 522-535.
18. Jeffery J., Persson B., Joernvall H. (1987) Characteristics of alcohol/polyol dehydrogenases. The zinc-containing long-chain alcohol dehydrogenases. *Eur. J. Biochem.* **167**: 195- 201.
19. Mulder N.J., Apweiler ., Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R.R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S.E., Pagni M., Peyruc D., Ponting C.P., Selengut J.D., Servant F., Sigrist C.J.A., Vaughan R, Zdobnov E.M. (2003). *The InterPro Database, 2003 brings increased coverage and new features.* [Nucl. Acids. Res. 31: 315-318.](#)
20. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* **3**:265-374.
21. Hofmann K. Bucher P. FalquetL. and Bairoch A. (1999) The PROSITE database its status in 1999. *Nucleic Acids Res.* **27** 215–219.
22. Henikoff S. and Henikoff J. G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19** 6565–6572.
23. Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics.* 2001 **17**:920-926.
24. Wu C.H, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LS, Zhang J, Barker WC. (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.* **30**:35-7.
25. Massingham T. Davies L. J. and Lio P. (2001) Analysing gene function after duplication. *Bioessays* **23** 873–876.
26. 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.
27. Duguet M., Confalonieri F. (1995) A 200-amino acid ATPase module in search of a basic function *Bioessays*, **17**: 639- 650.
28. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**:264-267.
29. Bairoch A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28** 45–48.
30. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**:22-28

31. Dodson G. and Wlodawer A. (1998) Catalytic triads and their relatives. *Trends Biochem. Sci.* **23** 347–352.
32. Makarova K. S. and Grishin N. V. (1999) Thermolysin and mitochondrial processing peptidase: how far structure-functional convergence goes. *Protein Sci.* **8** 2537–2540.
33. Ponting, C.P. and Russell, R.R. (2002) The natural history of protein domains *Annu. Rev. Biophys. Biomol. Struct.* **31** 45-71.
34. Mushegian AR, Koonin EV. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A.*, 93:10268-10273.
35. Brown JR, Doolittle WF. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev.*, 61:456-502.
36. 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.
37. 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.
38. 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.
39. Broglie R., Coruzzi G., Lamppa G., Keith B., Chua N.H. (1983) Structural analysis of nuclear genes coding for the precursor to the small subunit of wheat ribulose-1,5 bisphosphate carboxylase. *Biotechnology* **1**: 55-61.
40. Birnbaumer L. (1990) G-proteins in signal transduction. *Annu. Rev. Pharmacol. Toxicol.* **30**: 675-705
41. Casey P.J., Gilman A.G. (1988) G-protein involvement in receptor-effector coupling. *J. Biol. Chem.* **263**: 2577-2580.
42. Attwood T.K., Findlay J.B.C. (1993) Design of a discriminating fingerprint for G-protein-coupled receptors. *Protein Eng.* **6**: 167-176
43. Vapnik, V. (1998) *Statistical Learning Theory*. Wiley.
44. 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.