Using Protein Motif Combinations to Update KEGG Pathway Maps and Orthologue Tables

Frédéric Nikitin[1,4] frederic.nikitin@genebio.com
Bastien Rance[2] rance@lri.fr
Masumi Itoh[3] itoh@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[3] kanehisa@kuicr.kyoto-u.ac.jp
Frédérique Lisacek[1,4,5] frederique.lisacek@genebio.com

[1] Geneva Bioinformatics, 25 avenue de Champel, 1206 Geneva, Switzerland
[2] LRI, Bat. 490, University Paris XI, 91405 Orsay cedex, France
[3] Bioinformatics Center, Institute for Chemical Research, University of Kyoto, Uji 611-0011, Japan
[4] Swiss Instiute of Bioinformatics, 1 rue Michel Servet, 1211 Geneva, Switzerland
[5] Génome & Informatique, Tour Evry 2, 91034 Evry Cedex, France

Abstract

We have studied the projection of protein family data onto single bacterial translated genome as a solution to visualise relationships between families restricted to bacterial sequences. Any member of any type of family as defined in the Pfam database (domains, signatures, etc.) is considered as a protein *module*. Our first goal is to discover rules correlating the occurrence of modules with biochemical properties. To achieve this goal we have developed a platform to quantify information found in protein databases and to support the analysis of the nature of modules, their position and corresponding frequencies of occurrence (in isolation or in combination) in association with pathway knowledge as found in KEGG.This paper focuses on two pathways: the two-component system and the aminophosphonate metabolism, that are partially but not completely documented. Proteins involved in those pathways were listed separately in each organism to analyse module composition and rules constraining pathway interactions were identified. It is shown how these results can be used to update KEGG pathways and orthologue tables.

Keywords: protein, proteome, classification, bacteria, pathways, orthologue, knowledge representation

1 Introduction

As a result of efforts spread over time, current protein families are a mosaic of functional, structural and sequence features, as visible, for instance, in the federated InterPro web resource [10]. In fact, protein families reflect changes in protein data availability, that is, a sequence production ranging from independent and unorganised initiatives to systematic genome sequencing programmes.

Such mixed sources have given rise on the one hand, to classifications of proteins into families based on local common features, e.g., binding sites or other short conserved regions, both potentially rationalised by functional or structural studies. On the other hand, whole translated genome comparison has favoured an evolutionary interpretation of protein similarities and collections of translated orthologous genes have complemented existing protein families. KEGG (Kyoto Encyclopaedia of Genes and Genomes) and COG (Clusters of Orthologous Genes) are the most commonly used resources [7, 14]. The specificity of proteins originating from complete bacterial genomes is accounted for in TIGRfam [6], HOBACGEN [12] and HAMAPfam [3] sets. These collections are mainly based on multiple genome comparisons and express global sequence similarities. Indeed, when no specific feature of the protein whether functional and/or structural is known, protein families are equated to clusters of orthologues. Poorly characterised proteins commonly belong to a family only determined by evolutionary relationships. Such a family can only grow with the adjunction of other orthologues, which minimises the possibilities of overlap with other families. In contrast, when a domain is structurally and/or functionally defined the corresponding family can grow with the adjunction of a large variety of proteins minimally sharing one functional/structural feature and potentially different otherwise. As a result, the various and ever growing web resources for protein families provide a heterogeneous picture, due to mixing criteria for collecting proteins with distinct motivations (structure, function, Evolution, etc.).

Furthermore, protein family databases were originally conceptualised while assuming families were independent of each other. In recent years, considerable effort has been invested into highlighting potential and actual relationships between families. Emphasis is put on such protein *modularity* in InterPro [10], Pfam [1], CDART [4], etc., where instances of proteins belonging to several families are displayed. But in all cases, most statistics on protein properties are performed using large datasets of proteins from all possible origins, which makes the appreciation of the potential subtlety of protein characteristics difficult. Large variations in family size and species coverage generate very uneven entries in protein family databases where information piles up with no preset priority. Assessing the reliability of shown dependencies remains a challenge.

We have studied the projection of protein family data onto single bacterial translated genome as a solution to visualise relationships between families restricted to bacterial sequences. Our work follows a tradition of exploring the consistency of different representations of a same problem for determining the correlations and/or the equivalence as well as how complementary various approaches can be [15, 8]. The pattern recognition approach emphasises the key importance of knowledge representation. For instance, a sentence can be represented as a succession of phonemes or as a succession of words. Phonological rules are distinct from syntactic rules but both govern the occurrence of letters in sentences. In much the same way, constraints on amino acid and on peptides (peptides as domains, active or binding sites, evolutionary units, etc.) apply in proteins in a compatible way that is not clearly established. We suggest changes of representations as a means of elucidating the relative roles of these constraints.

For reasons detailed in [11], we chose to map protein families of the Pfam database with complete bacterial translated genomes. In particular, we justified our choice in full knowledge of underlying biases in databases of protein families. In the following, we will refer to the term of *module* to designate any identified family membership in a protein. Any member of any type of family as defined in the Pfam database (domains, signatures, etc.) is considered therefore as a module.

Preliminary results presented in [11] showed biases towards set module combinations that distinguish gram-positive from gram-negative bacteria or combinations that allow further investigation of unknown functions. Such exploratory studies justify the search of contextual rules governing module combinations in an attempt to produce explanatory protein classification (as opposed to blind clusterisation). Another underlying motivation is to express family membership in terms of explicit necessary and sufficient conditions. Whenever a family is associated with unknown or poorly characterised functions, membership conditions are usually only necessary.

This paper focuses on two pathways: the two-component system and the aminophosphonate metabolism. Proteins involved in those pathways were listed separately in each organism to analyse module composition and rules constraining pathway interactions were identified. It is shown how these results can be used to update KEGG pathways and orthologue tables.

2 Method and Results

2.1 Initial Data

Protein sequences of *B. subtilis* and *E. coli* were downloaded from the GenoList bacterial databases [18] with corresponding Swiss-Prot/UniProt accession numbers. Cross-references to Pfam (release 14.0) in the corresponding UniProt (release 2.1) entries were extracted.

The two bacterial translated genomes were scanned with HMMER 2.3 [17] to complement the set of domains selected in the previous step. Pfam-A (release 14.0) was used for profile construction.

All other bacterial translated genomes were extracted from UniProt (release 2.1).

2.2 Update of the Two-Component System Map and Related Orthologue Tables

2.2.1 Predefined Rules and Preliminary Observations

E. coli K12 and *B. subtilis*, the two most documented bacterial genomes, were first chosen to determine the variability of module combinations in the sensor proteins as well as the response regulator proteins of the two-component system. Given initial observations made in *B. subtilis* and *E. coli*, we investigate domain combinations in other bacterial protein sequences. Our first goal is to discover rules correlating the occurrence of modules with biochemical properties. To achieve this goal we have developed a platform to quantify information found in databases, which is at present carried out through the analysis of the nature of modules, their position and corresponding frequencies of occurrence (in isolation or in combination) in association with pathway knowledge[1] as found in KEGG and biochemistry handbooks.

Several criteria were exploited to finalise the selection of authentic pairs of the two-component system, namely the histidine kinase sensor protein and the response regulator protein:

- A sensor kinase contains minimally a "Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase" domain (PF02518).

- A response regulator contains minimally a "response regulator" domain (PF00072).

- All sequences are related to EC.2.7.3.- whether explicitly or implicitly, which excludes the chemotaxis proteins (classically referred to as CHEA, B, ..., W, Y, Z).

- A sensor kinase and a response regulator are often encoded in adjacent genes.

Subsequently, sensor proteins were classified upon their content in modules. All Pfam entries found in one or more protein of the two-component system are listed in Table 1(a,b,c). Both the curated (Pfam-A) and the automatic complement (Pfam-B) of Pfam were used. In the case of Pfam-B, only those families spanning over more than ten species were considered.

2.2.2 Preliminary Observations

The primary goal of the platform achieving the mapping between complete translated genomes and protein families is to offer a user the possibility of exploring module combinations, visualising and focusing on combinations of interest. In this framework, the platform is considered as assisting the formulation of hypotheses as opposed to providing fully automated analysis. It is therefore considered as a preliminary step in a data mining workflow.

Given the preset constraints listed in 2.2.1, the careful examination of data has led to set the basis of the general architecture described in Figure 1.

[1] Potentially the platform also allows association with subcellular location, bacterial lifestyle, etc.

Table 1a: List of Pfam-A families mapped in the regulator protein of the bacterial two-component system.

Pfam-A ID	Family size*	Family name
PF00072	4082	Response regulator receiver domain
PF00158	723	Sigma-54 interaction domain
PF00165	1448	Bacterial regulatory helix-turn-helix proteins, AraC family
PF00196	1163	Bacterial regulatory proteins, luxR family
PF02954	609	Bacterial regulatory protein, Fis family
PF03861	57	ANTAR domain
PF04397	196	LytTr DNA-binding domain

*exclusively in bacteria

Table 1b: List of Pfam-A families mapped in the sensor kinase of the bacterial two-component system.

Pfam-A ID	Family size*	Family name
PF00497	563	Bacterial extracellular solute-binding proteins, family 3
PF00512	2332	His Kinase A (phospho-acceptor) domain
PF00672	2016	HAMP domain
PF00785	531	PAC motif
PF00989	901	PAS domain
PF01590	651	GAF domain
PF01627	324	Hpt domain
PF02518	4825	Histidine kinase-, DNA gyrase B-, HSP90-like ATPase
PF06580	117	Histidine kinase
PF07536	56	HWE kinase
PF07568	49	Histidine kinase

Table 1c: List of Pfam-B families with large coverage of bacterial species, mapped in the bacterial two-component system.

Pfam-B ID	Family size*	Family coverage in bacterial species
PfamB-000120	148	Actinobacteria (43), Deinococcus-th.(4), Proteobacteria (54), Firmicutes (47)
PfamB-000398	75	Firmicutes (75)
PfamB-001176	39	Proteobacteria (39)
PfamB-001840	33	Actinobacteria (6), Proteobacteria (15), Fusobacteria (1), Firmicutes (11)

2.2.3 Discovered Rules

A number of rules could be derived from studying the presence and absence of modules in proteins of the two-component system, as well as the relative position of modules. First, a direct observation of our dataset reveals that modules are more or less dispensable. Consequently, we have attempted to identify key modules in the various combinations. A sketch of our strategy is given in Figure 2.

Comparative studies led to emphasise the decisive presence of the “phospho-acceptor” domain along with a series of potential substitutes in all sensor proteins. This module is systematically directly upstream PF002518. As shown in Figure 1, this position is highly variable. Interestingly, this variability could be mapped with specific module combinations in corresponding regulator proteins.

The largest Pfam family defining the “phospho-acceptor” domain is PF00512 (approx. 2300 bacterial proteins). We therefore studied the variations in module combinations in corresponding response regulators depending on the presence or absence of PF00512.

Given the necessary presence of PF02518, the conditional presence/absence of PF00512 can be summarised as follows:

If PF00512 in sensor then, PF00486 or (PF00158 and PF02954) in regulator
If PF06580 and PF01590 in sensor then, PF00072 and PF04397 in regulator
If PF03861 (ANTAR) in regulator then, PF07536 in sensor

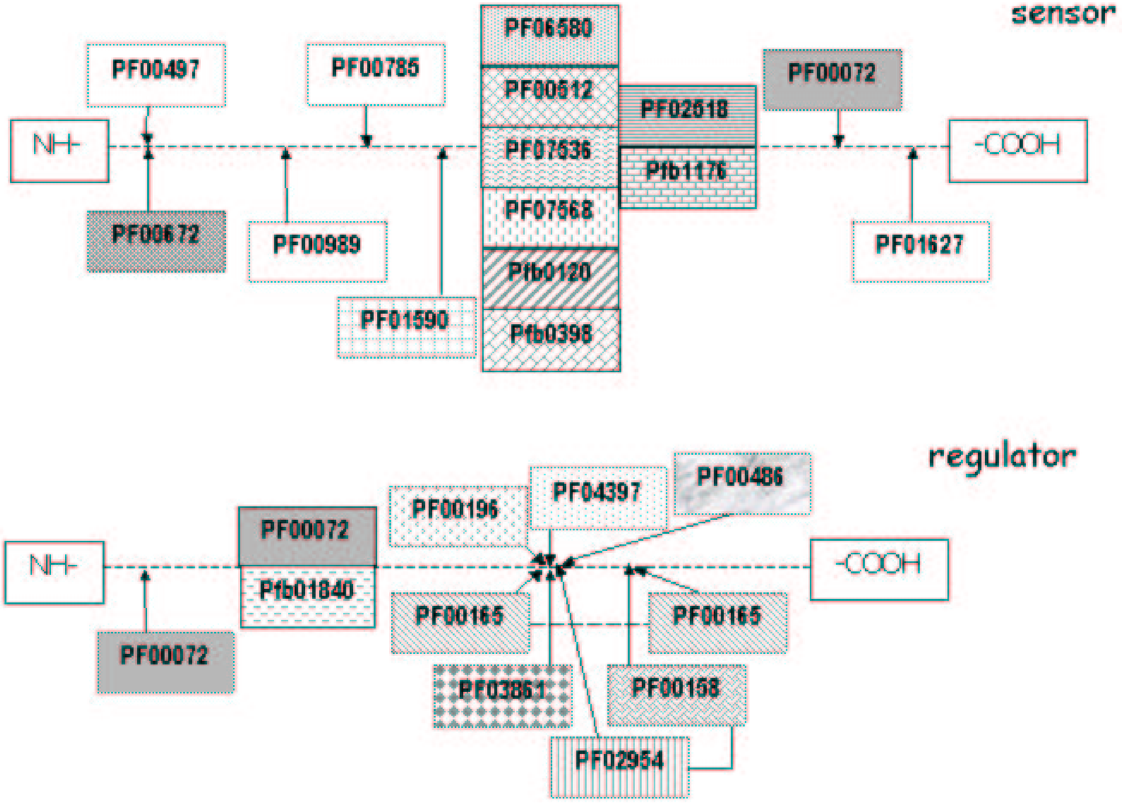

Figure 1: Common architecture of sensor and regulator proteins of the bacterial two-component system.

We also have identified mutually exclusive situations, such as:

> If PF06580 and PF00672 in sensor then, PF00072 and PF00165 in regulator
> If PF06580 and PF01590 in sensor then, PF00072 and PF04397 in regulator

Such an exploration of joint combinations between regulator and sensor proteins led to put forward some assumptions on the correlated changes. This information is summarised in Figure 3. Such correlated changes allow the straightforward identification of the LytR/LytS sub-class, for instance. In that particular case, a set module combination seems to provide the necessary and sufficient conditions we are seeking. Further predictions can be made as shown in the following (see Section 2.2.3).

Finally, we also tested releasing the imperative presence of PF02518 in sensors while requiring any of the modules identified as "phospho-acceptor" domains. This led to locate a new family in Pfam-B (PF001176) as an alternative to PF02518.

Figure 3 shows that at least seven consistent sub-classes of sensors can be defined upon correlated module combinations. This classification could be used to serve an objective common within our collaborative work which involves resolving the ambiguity of dashes in degenerated EC numbers such as EC.2.7.3.-.

2.2.4 Ethanolamine Sensors in Some Bacteria Species

A regular mapping was identified in a sub-category of (regulator, sensor) pairs, which is summarised in the following statement:

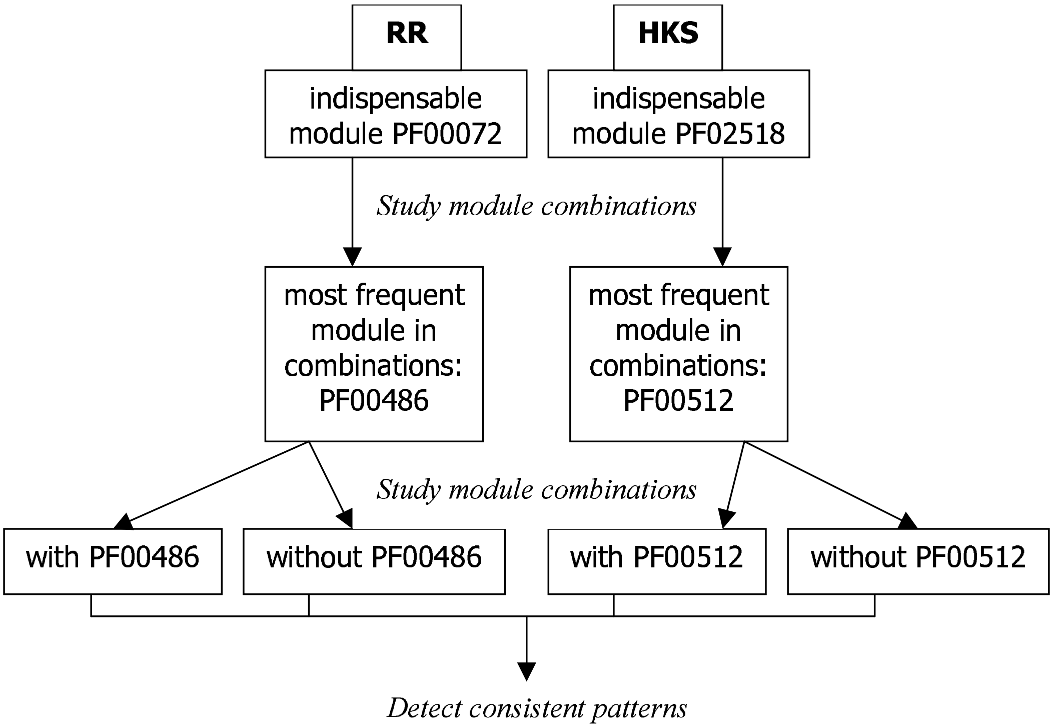
RR
HKS
indispensable module PF00072
indispensable module PF02518
Study module combinations
most frequent module in combinations: PF00486
most frequent module in combinations: PF00512
Study module combinations
with PF00486
without PF00486
with PF00512
without PF00512
Detect consistent patterns

Figure 2: Basis of a strategy for the detection of consistent patterns of occurrence of combinations.

> Whenever the combination {PF00072 + PF03861} is detected in a response regulator, the corresponding sensor protein contains the combination {PF07536 + PF02518}.

Twenty-six {PF00072 + PF03861} combinations were found in UniProt bacterial entries. All but one of these proteins are found in fully sequenced organisms and there seems to be only one such combination per organism (that is, 25 sequences originate from 25 distinct complete translated genomes and the last sequence is isolated). Furthermore, a majority of these proteins are annotated as encoded by a gene named *nasT*, originally studied in *Azotobacter vinelandii* (not completely sequenced) and meant to be required for the expression of the nitrite-nitrate reductase operon [5]. In the same paper, *nasS* is described as the gene encoding for the corresponding sensor protein which contains the {PF07536 + PF02518} combination. Moreover, six out of 25 regulators are located on their respective chromosome in a position adjacent to the corresponding sensor kinase. In all six cases the sensor kinase contains the {PF07536 + PF02518} combination.

It appeared reasonable to attempt further investigation of {PF07536 + PF02518}. In fact, fifty-five such combinations were found in bacterial entries. In this case, several of such proteins originate from the same organism. However, it was possible to unambiguously match 16 pairs of (regulator, sensor) exclusively in F*irmicutes* and *Actinobacteria*. Interestingly, in *Clostridium tetani*, and *Fusobacterium nucleatum*, the corresponding annotation of both the regulator and the sensor entries mentions ethanolamine utilisation. Such specialisation could be extended to L*isteria* species (hypothetical proteins: lm01173/ lin1137) as acknowledged by Philippe Glaser (personal communication). We have yet to confirm this property in the remaining species.

2.2.5 Generalisation to Staphylococcus Aureus N315

Four of the seven sub-classes defined above and shown in Figure 3 were mapped in *Staphylococcus aureus* N315. Seven pairs of (sensor, regulator) not yet listed in the corresponding orthologue tables were identified as containing the relevant module combinations. Figure 4 illustrates the distribution of these pairs across the four detected sub-classes and how the KEGG tables were updated.

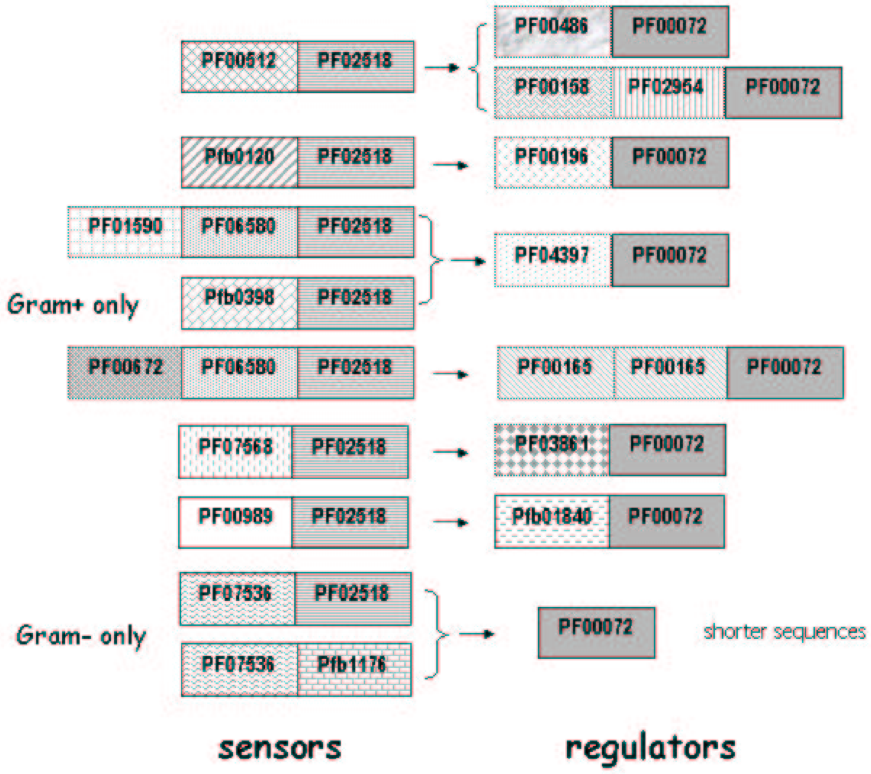

Figure 3: Correspondence between sensor and regulator respective architecture.

2.3 Update of the Glycerolipid and Aminophosphonate Metabolism Maps and Related Orthologue Tables

2.3.1 Predefined Rules and Preliminary Observations

As stated above, the prediction of values for replacing dashes in degenerated E.C. numbers is a shared goal within our collaboration. We have considered the update of glycerolipid and aminophosphonate metabolism due to the common presence of EC.2.7.8.-. The ambiguous annotation of sequences bearing two copies of the phospholipase D active site motif (PF00614) in *E. coli* K12, led to focus on this initial architecture. Two hundred and sixty bacterial sequences containing two copies of the phospholipase D active site motif (PF00614) were extracted from Uniprot (release 2.1). Redundant data were discarded which reduced the original set to 246 proteins. No other prerequisite was stated. Given that the relative position of modules is a property of interest in our analysis, potential positional constraints have been searched.

2.3.2 Unexpected Predictive Regularity

Varying lengths between 150 and 1100 amino acids (on average 450) did not affect the architecture in the 246 proteins but the dual motif was apparently shifted towards the C-terminal end in longer sequences. This simple observation led to calculate the distance between the last amino acid of the second occurrence of PF00614 and the last C-terminal amino acid. This distance was surprisingly constant depending on the functional annotation associated with the proteins considered (see Figure 5). Such regularity is summarised in Table 2.

We have not checked all proteins not falling into the obvious categories (i.e., wrong C-terminal prediction cannot be ruled out). However, the overwhelming case of a 60 amino acid distance between the end of PF00614 and the last C-terminal amino acid (∼ 40% of proteins) matching the “cardiolipin synthase” annotation deserves attention. Likewise, the “CDP-diacylglycerol-serine O-phosphatidyltransferase” annotation is very consistent with the 72 amino acid distance.

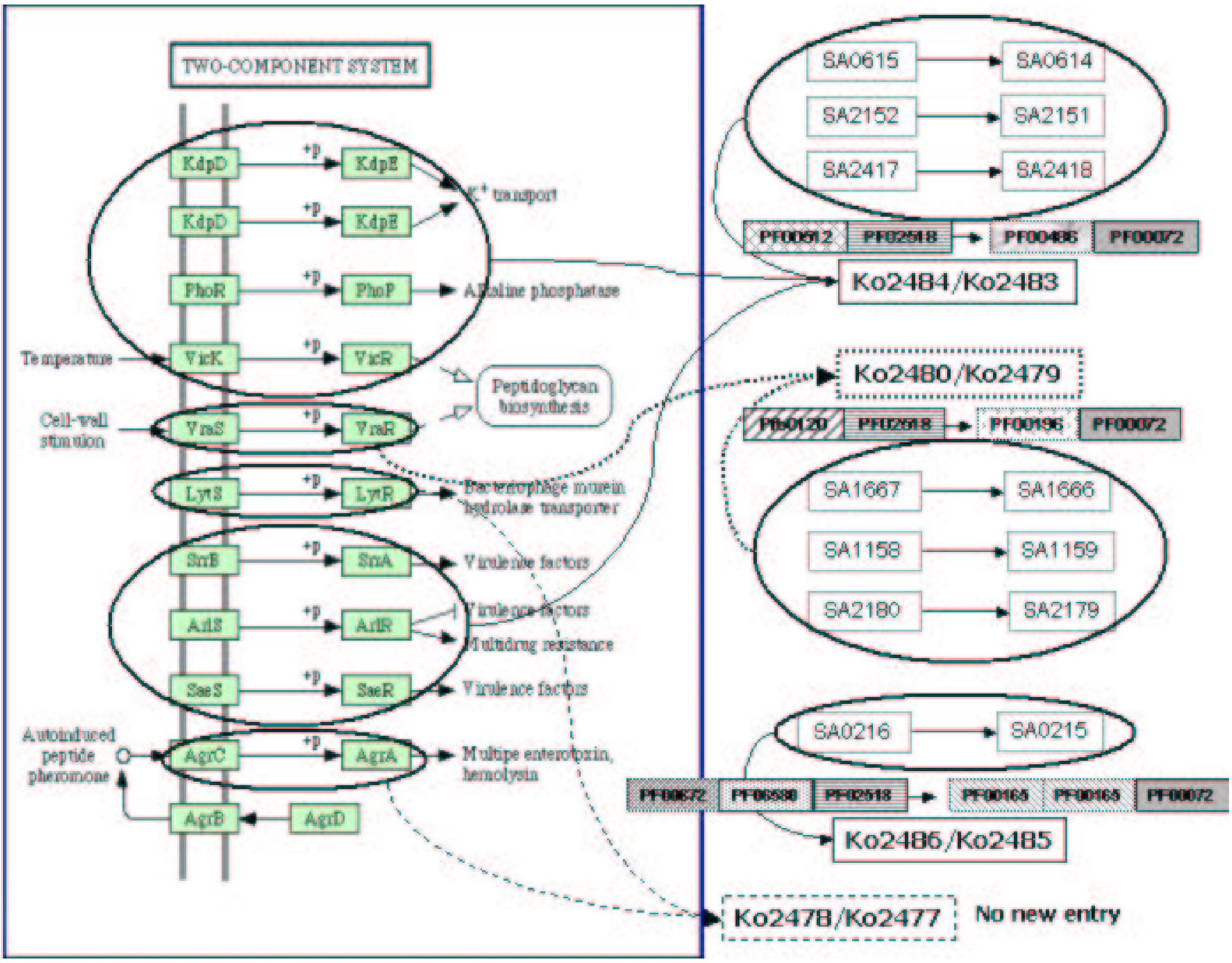

Figure 4: Example of update of KEGG orthologue table for *Staphylococcus aureus* N315.

3 Discussion

The study of module content in proteins taken in a complete translated genome has been used as a starting point for formulating hypothesis possibly valid in proteins across the bacterial kingdom. Results presented above show that regular features could be identified while describing proteins as succession of well or poorly known modules. An integrative platform was developed to ease the discovery of constraints or rules by allowing the quantification of module properties (e.g., presence/absence, relative frequency and position). So far rules were manually discovered but results are encouraging for future automation of some procedures (the work is ongoing). The outcome of this work can be directly input in KEGG orthologue tables.

As mentioned earlier, we have studied the projection of protein family data onto single bacterial translated genome as a solution to visualise relationships between families restricted to bacterial sequences. Examining each translated genome independently and comparing the occurrences of domain combinations was intended to (i) identify generalisable discriminating properties between bacteria, (ii) study various modular combinations of domains and formulate hypothesis on functional implications, (iii) set the basis of a similarity measure between translated genomes assuming that it is complementary and compatible with commonly used sequence-based measures.

The examples shown in this paper show that usual operations of insertion, deletion and substitution could be used to define a measure for comparing proteins within or across translated genomes based on the modular architecture of proteins. In the same vein, [16] have proposed a systematic global approach to evaluate the modular content of translated genomes. Our prospective work also aims towards that

Table 2: C-terminal length restricted variations in 246 proteins containing two occurrences of PF00614.

C-terminal	Functional annotation	Proteins	Species coverage
< 59	Mixed functions	28	Across bacterial species
60	Cardiolipin synthase	102	Across bacterial species
60±1	Cardiolipin synthase	17	Across bacterial species
< 61, 72<	Mixed functions	12	Pseudomonas + Chlamydia
72	CDP-diacylglycerol-serine O-phosphatidyltransferase	25	Proteobacteria only
< 72, 82<	Mixed functions	17	Proteobacteria + Chlamydia
82	Putative synthase	18	Proteobacteria only
83	Putative synthase	3	Proteobacteria only
< 85	Mixed functions	24	Proteo- / cyano- bacteria

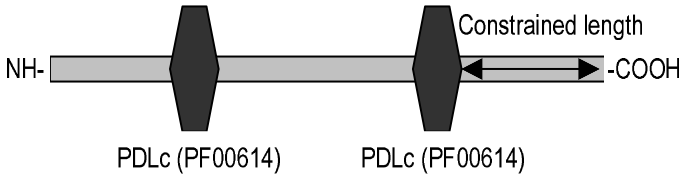

Figure 5: Basic architecture of members of the phospholipase D family. The arrow specifies the segment found to remain constant in length depending on protein function.

goal but we have chosen to limit ourselves to partially known pathways. Moreover, compensatory changes and mutually exclusive roles of modules are observed phenomena to be further explored. However, we are presently focused on contextual and local rules or constraints to be generalised across bacterial species in set contexts, e.g., that of a pathway. It is our deliberate choice not to be systematic at first. We tend to assume that differential processing of proteins might be more appropriate than bulk analyses for mining purposes.

Our work on the two-component system is very much related that described and reviewed in [2]. However, differing initial assumptions lead us to considering different constraints on proteins. In particular by examining the relative frequency and the positional variations of modules promoted the determination of hierarchical rules. The search for structured modularity is part of a general strategy for integrating proteomics data as described in [9]. This viewpoint is supported by increasing evidence that units of evolution are more likely to be protein components (see [13] for review).

Note added in proof: Figure 1 features architectures derived from information found in Pfam version 14. Version 15 was released shortly after this work was submitted. As it turns out, in version 15, PfamB_000120 was transferred to PfamA as PF07730, a new sub-family of histidine kinases, thereby confirming the appropriateness of our analysis.

References

[1] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R., The Pfam protein families database, *Nucleic Acids Res.*, 32:D138–D141, 2004.

[2] Galperin, M.Y., Bacterial signal transduction network in a genomic perspective, *Environ Microbiol.*, 6:552–567, 2004.

[3] Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J., Lachaize, C., Veuthey, A.L., Gasteiger, E., and Bairoch, A., Automatic annotation of microbial proteomes in Swiss-Prot, *Comput. Biol. Chem.*, 27:49–58, 2003.

[4] Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H., CDART: Protein homology by domain architecture, *Genome Res.*, 12:1619–1623, 2002.

[5] Gutierrez, J.C., Ramos, F., Ortnert, L., and Tortolero, M., NasST, two genes involved in the induction of the assimilatory nitrite-nitrate reductase operon (nasAB) of *Azotobacter vinelandii*, *Mol. Microbiol.*, 18:579–591, 1995.

[6] Haft, D.H., Selengut, J.D., and White, O., The TIGRFAMs database of protein families, *Nucleic Acids Res.*, 31:371–373, 2003.

[7] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32:D277–280, 2004.

[8] Lisacek, F., Methods of computational genomics: An overview, *Compact Handbook of Computational Biology*, Konopka, A.K. Crabbe, J.C., and Dekker, M., editors, New-York, 279–342, 2004.

[9] Lisacek, F., Chichester, C., Gonnet, P., Jaillet, O., Kappus, S., Nikitin, F., Roland, P., Rossier, G., Truong, L., and Appel, R.D., Shaping biological knowledge: Applications in proteomics, *Comp. Funct. Genom.*, 5:190–195, 2004.

[10] Mulder, N.J., Apweiler, R., 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., Vaughan, R., and Zdobnov, E.M., The InterPro database, 2003 brings increased coverage and new features, *Nucleic Acids Res.*, 3:315–318, 2003.

[11] Nikitin, F. and Lisacek, F., Investigating protein domain combinations in complete proteomes, *Comput. Biol. Chem.*, 27:481–495, 2003.

[12] Perriere, G., Duret, L., and Gouy, M., HOBACGEN: Database system for comparative genomics in bacteria, *Genome Res.*, 10(3):379–285, 2000.

[13] Ponting, C.P. and Russell, R.R., The natural history of protein domains, *Annu. Rev. Biophys. Biomol. Struct.*, 31:45–71, 2002.

[14] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V., The COG database: New developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 29:22–28, 2001.

[15] Watanabe, S., *Pattern Recognition: Human and Mechanical*, Wiley, 1985.

[16] Ye, Y. and Godzik, A., Comparative analysis of protein domain organisation, *Genome Res.*, 14:343–353, 2004.

[17] `http://hmmer.wustl.edu/`

[18] `http://www.pasteur.fr/GenoList/`