ARC: Automated Resource Classifier for agglomerative functional classification of prokaryotic proteins using annotation texts†

Muthiah Gnanamani, Naveen Kumar and Srinivasan Ramachandran*

G N Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India

**Corresponding author (Fax, 91-11-2766-7471; Email, ramuigib@gmail.com)*

Functional classification of proteins is central to comparative genomics. The need for algorithms tuned to enable integrative interpretation of analytical data is felt globally. The availability of a general, automated software with built-in flexibility will significantly aid this activity. We have prepared ARC (Automated Resource Classifier), which is an open source software meeting the user requirements of flexibility. The default classification scheme based on keyword match is agglomerative and directs entries into any of the 7 basic non-overlapping functional classes: Cell wall, Cell membrane and Transporters ($\mathcal{C}$), Cell division ($\mathcal{D}$), Information ($\mathcal{I}$), Translocation ($\mathcal{L}$), Metabolism ($\mathcal{M}$), Stress($\mathcal{R}$), Signal and communication($\mathcal{S}$) and 2 ancillary classes: Others ($\mathcal{O}$) and Hypothetical ($\mathcal{H}$). The keyword library of ARC was built serially by first drawing keywords from *Bacillus subtilis* and *Escherichia coli* K12. In subsequent steps, this library was further enriched by collecting terms from archaeal representative *Archaeoglobus fulgidus*, Gene Ontology, and Gene Symbols. ARC is 94.04% successful on 6,75,663 annotated proteins from 348 prokaryotes. Three examples are provided to illuminate the current perspectives on mycobacterial physiology and costs of proteins in 333 prokaryotes. ARC is available at *http://arc.igib.res.in.*

[Gnanamani M, Kumar N and Ramachandran S 2007 ARC: Automated Resource Classifier for agglomerative functional classification of prokaryotic proteins using annotation texts; *J. Biosci.* **32** 937–945]

1. Introduction

The complete sequence determination of more than 300 micro-organisms has opened new opportunities for comparative analyses. An essential prerequisite step in this exercise is to annotate the functional roles of newly identified proteins and classify them into biological groups of common overall activity. The traditionally used and perhaps widely known system of classification was originally proposed by Riley, more than a decade ago (Riley 1993). Recently, some microbial genes have been annotated using the controlled vocabulary system of Gene Ontology (GO) consortium (Harris *et al* 2004). Compared with the Riley's scheme, the GO could be considered as an agglomerative approach enabling comparative analysis with respect to the ontologies. The prime mover towards this trend is the realization by many biologists that we need to transit from the reductionist schematic to the integrative schematic as elucidated by Ren´e Descartes in his second and third precepts, 365 years ago (Auffray *et al* 2003; Van Regenmortel 2004).

"The second, to divide each of the difficulties which I would examine into as many parts as it would be possible, and as might be required to resolve them best."

"The third, to conduct my thoughts in order, beginning with the simplest and easiest objects to know, to rise little by

Keywords. Automated resource; functional classification; integrative biology

†Additional material pertaining to this article is available with authors.

little, as it were by steps, up to the knowledge of the most complex; and assuming even order between those which do not precede each other naturally."

While the GO approach is systematic, the appearance of a given protein either in more than one node within an ontology or in more than one ontology can confound comparative analyses. These effects are particularly noticeable when the breadth of the functional class is narrow. Although narrowly sectioned functional classes offer specialized views of biological phenomena suited for specific investigations, they tend to limit straightforward interpretations from holistic perspectives (Van Regenmortel 2004). Early origins of agglomerative approaches can be traced back to Adams *et al* (1995) and Andrade *et al* (1999). Adams *et al* (1995) classified proteins and their encoding genes into functional classes such as energy metabolism, cell structure, homeostasis and cell division, RNA and protein synthesis and processing, cell signaling and communication. Andrade *et al* (1999) classified proteins into superclasses ENERGY, METABOLISM and INFORMATION. Our classification scheme is based on these early foundations laid for comparative analyses. Our goal is to use a functional classification scheme such that a given protein can be classified singularly into one functional class. Although this goal is ambitious considering the multiple functional roles exhibited by some proteins, our broad definition of functional class offers sufficient space to accommodate such candidates as well. In this work we develop a software using this approach involving classification of proteins into 7 basic and 2 ancillary agglomerative non-overlapping functional classes to enable integrative interpretation of analytical data.

2. Implementation

2.1 *Rationale*

Three attributes are associated with the function of a protein. These are the process it performs, the sub-cellular location where it performs the process, and the purpose for which the process is used. For example, the process of a DNA polymerase is to synthesize DNA. The cellular location in prokaryotes for this activity is intracellular. The purpose is to enable cell multiplication and division or DNA repair. Focusing on process will direct the entry of this protein into the class of replication and extending this further agglomeratively, will place the protein in the class of Information. On the other hand, if a protein is annotated only as GTP binding or ATP binding, although the process is clear, it is required in many processes, purposes and locations. Therefore, we would have to render it as unclassified. If a protein is annotated as 'membrane protein', then, only the location is known. This protein will find its entry into the class of Cell Wall, Cell Membrane and Transporters. Among all three attributes, process and location are clear whereas purpose is highly subjective and poorly defined. Therefore, in this system of classification, we focus on process (molecular function) and location (cellular component). The GO equivalent of purpose is Biological process.

2.2 *Functional class description*

ARC classifies proteins into seven basic functional classes: *C*: Cell wall and Cell membrane and Transporters, *D*: Cell division and binary fission, *I*: Information (Replication, Transcription, Translation), *L*: Translocation and Secretion, *M*: Metabolism, *R*: Stress, *S*: Signalling and communication and two ancillary functional classes: *O*: Others, *H*: Hypothetical proteins.

2.3 *Keyword collection*

Keywords qualifying the entry patterns for proteins with annotated functions were first collected from two bacteria *Bacillus subtilis* (Kunst *et al* 1997) and *Escherichia coli* K-12 (Blattner *et al* 1997), which are well studied compared to others. This approach is similar to that followed in constructing the Gene and Protein Synonyms Database (GPSDB) (Pillet *et al* 2004). Subsequently, ARC was operated on an archaeal representative, *Archaeoglobus fulgidus* (Klenk *et al* 1997). Additional keywords were collected from proteins with known function that were not classified using the keyword library prepared from *B. subtilis* and *E. coli* K-12. Subsequently, keywords were also collected from GO terms belonging to Cellular Component (for cellular locations) and to Molecular Function (processes). The terms belonging to Biological Process serve to describe the biological goals, which are more subjective and overlapping and therefore were not considered in our scheme. The keyword library was organized into 26 files encompassing all the functional classes arranged alphabetically.

2.4 *KEYWORD_FunctionalClassSymbol*

Functional Class Symbol is a single character corresponding to the name of a functional class. This nomenclature can be modified by the user to extend the symbol character length up to 10 characters without any space. ARC does not support those keyword entries which are without the FunctionalClassSymbol. ARC is case insensitive; user can enter keywords and associated class symbols in either case or a mixture of both. The keywords along with their associated

FunctionalClassSymbol are separated by an underscore and stored as ASCII files. This simple format supports the user to edit the keyword library by either deleting or modifying existing keyword entries or including keywords of their own choice with corresponding FunctionalClassSymbols. Users can also update the keyword library as and when needed. Organization of keywords need to follow a simple directional rule of decreasing complexity from the top. For example, the keyword 'histone lysine_M' is placed before 'histone_I' so that ARC first searches for the complex keyword followed by searching for simpler keywords, which are substring of the complex keyword. Similarly, Signal peptide_L keyword is placed before Signal_S. This organization is founded on the time honoured principle that complex keywords attribute higher clarity and specificity than simpler keywords.

2.5 *Gene symbol and synonym collection*

Gene symbols along with their synonyms (associated annotation information) were retrieved from the websites of SubtiList (*Bacillus subtilis ListiList8*) [SubtiList], Colibri (*Escherichia coli* K-12) [Colibri], TubercuList (*Mycobacterium tuberculosis* H37Rv [TubercuList], Leproma (*Mycobacterium leprae* TN) [Leproma], BoviList (*Mycobacterium bovis* AF2122/97) [BoviList], ListiList (*Listeria monocytogenes* EGD-e, *L. innocua* CLIP 11262) [ListiList], LegioList (*Legionella pneumophila* Paris, *L. pneumophila* Lens) [LegioList], SagaList (*Streptococcus agalactiae* nem 316) [SagaList], PhotoList (*Photorhabdus luminescens* TT01) [PhotoList], and UniProt [UniProt]. The gene symbols with their synonyms were alphabetically sorted into 26 accessory files. These files contain 47,456 gene symbols. The gene symbols in these files are not nonredundant. For example, both gene symbols *adh1* and *adh-1* (alcohol dehydrogenase) are included in these files.

2.6 *Algorithm*

The algorithm is shown in figure 1. ARC follows 'first hit and assign' approach. This strategy is based on the premise that the main process to which a given annotation refers to is written first, in the subject line, followed by other related information, if available. In a majority of annotations, this principle is evidently adhered to and leads to a singular decision, but in a small minority of cases (1.44% to 3.6% of annotated proteins), the annotation information leads to pluralistic decision. The algorithm however, directs the entry of these proteins to the corresponding functional class based on the first keyword it spots, but these entries are potential confounders. Annotations similar to these cases in any species will likely confound analysis. Fortunately in a majority of species, these cases are very low in number.

At the start of the algorithm, a positive hit for the presence of word 'nonribosomal' in the annotation text will assign 'M' as functional class because these proteins (nonribosomal peptide synthetases and associated proteins) are responsible for the synthesis of secondary metabolites (Grunewald and Marahiel 2006). For a negative hit, ARC will start searching for the presence of words or substrings – 'not' or 'non' – in the annotation text. A positive hit will place the protein into 'Unclassified' category. A negative hit will cause ARC to search for the common words or substrings 'synthesis' and 'synthetic'. If any of these general substrings is present in the annotation information, and a positive hit for a keyword for a functional class other than metabolism class, then, ARC directs the protein to the respective functional class. A negative hit for all keywords for other functional classes will cause ARC to direct the protein in the metabolism class. It is to be noted that the substrings 'non', 'not', 'synthesis', 'synthetic' and 'nonribosomal' are not part of the keyword library.

In subsequent steps, ARC will assign the destination functional class to the protein for a positive hit to keywords corresponding to that class. After exhausting all the possibilities ARC checks whether a protein should be placed in the Hypothetical class. If no keyword of its library is found in the annotation text, ARC starts searching for the presence of Gene Symbols. ARC is equipped to classify proteins having official gene symbol as their annotation text. For a positive hit, ARC searches for the keywords of its library in the corresponding synonym of the gene symbol. For a negative hit, ARC places the respective protein in the 'Unclassified' category.

2.7 *ARC Web server*

The ARC algorithm was coded in C and compiled using the GNU gcc compiler 3.4.3 in the Itanium 2 64 bit dual processor server running on RedHat Linux Enterprise version 4. The Web server was prepared in Apache version 2.0, Server side scripting in PHP version 5.0 using the graphics library JPgraph version 2.2. The client side scripting was done in HTML and AJAX. The C code of the algorithm can be compiled by other (UNIX based) C compilers as well.

File Formats: Three types of file formats can be accepted. An input file with (i) annotations only in FASTA format with or without sequence, (ii) annotations and expression fold change for microarray data, (iii) annotations and numeric value representing user specified data. Users also can upload MS Excel files. Graphical displays are

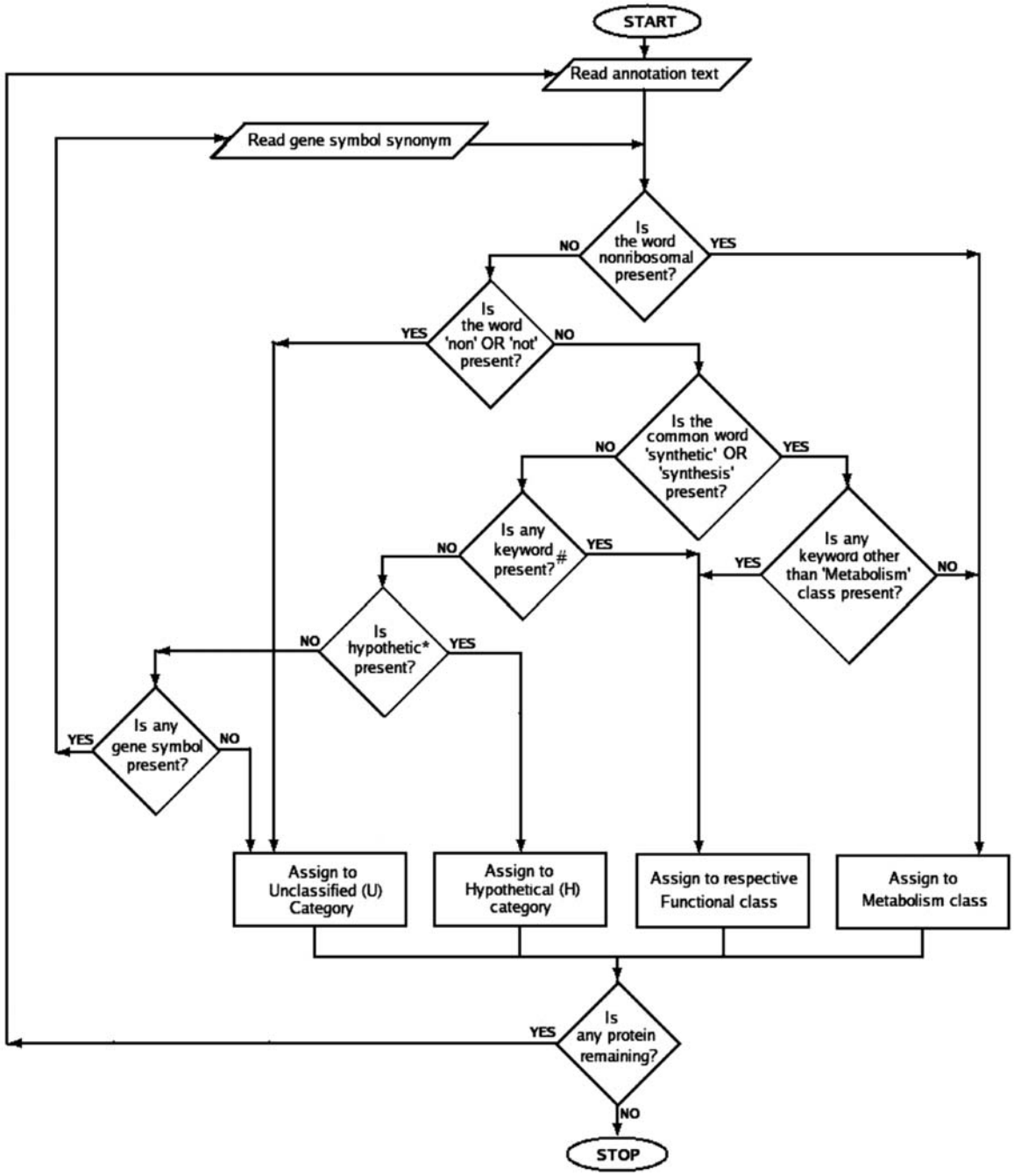

Figure 1. The ARC algorithm. Rules for classification of proteins performing distinct functions but have common keywords in their annotation texts: Signalling and communication: Presence of 'kinase' prefixed by 'serine' OR 'threonine' OR 'tyrosine' OR 'polo' OR 'pros' OR 'tele' OR 'SAP' OR 'protein'; Presence of 'phosphatase' prefixed by 'serine' OR 'threonine' OR 'aspartate' OR 'histidine' OR 'threonine' OR 'protein'; Presence of 'PAS', which is not a substring of any other word; Metabolism: Presence of kinase NOT prefixed by 'serine' OR 'threonine' OR 'tyrosine' OR 'polo' OR 'pros' OR 'tele' OR 'SAP' OR 'protein'; Presence of 'phosphatase' NOT prefixed by 'serine' OR 'threonine', OR 'aspartate' OR 'histidine' OR 'protein'; Presence of any of the words 'PII', 'PTS', 'P700', 'CoA', which is not a substring of any other word; Information: Presence of any of the words 'RNA', 'DNA', 'SOS', which is not a substring of any other word; Cell wall and Cell membrane: Presence of the word 'porin' which is not a substring of any other word. Proteins with the following key words or substrings are directed to the hypothetical class: 'hypothetic', 'unknown cds', 'conserved archael protein', 'conserved protein', 'conserved crenarchael protein', 'putative enzyme', 'uncharacterized protein', 'uncharacterized conserved protein', 'putative conserved protein', 'unknown protein', 'function unknown', 'unknown function'.

generated dynamically. The output files from ARC are tab delimited ASCII files containing the annotation text and the FunctionalClassSymbol.

2.8 *Sequence files and annotations*

The July 2006 release of genomic data of NCBI (NCBI) had 348 prokaryotes with a total number of proteins amounting to 1,056,660, containing 675, 663 annotated proteins.

2.9 *Orthologs and paralogs*

Orthologs were examined by mixing the proteins a given class from 3 species, *Mycobacterium tuberculosis*, *M. bovis* and *M. leprae*. Orthologs were noted if clusters of proteins had singular entries from all three species meeting the criteria $S = 1.5$, $L = 0.95$ in the program BLASTCLUST (Kondrashov *et al* 2002, NCBI). Paralogs in each species were identified by running BLASTCLUST on proteins of a given functional class within a species meeting the criteria $S = 0.8$, $L = 0.95$ (Subramanyam *et al* 2006). All other parameters were used at their default settings.

2.10 *Statistical methods*

Statistical significance of differences in proportions were carried out using the Binomial proportions test (Uitenbroek 1997). For data in table 1, expected proportions were computed assuming the same pattern of the reference organism to which a given species is compared. For data in table 2 observed proportions in each functional class were compared to global proportions of genes for the same functional class in each expression category assuming no preference for each category with any of the functional classes.

2.11 *Availability and requirements*

Project name: ARC; Project home page : *http://arc.igib.res.in*; Operating system(s): Linux, UNIX;

Programming language : C; Other requirements : gcc, cc or other equivalent compilers;

License : GNU GPL; Any restriction to use by non-academics : No

3. Results and discussion

3.1 *Efficiency*

ARC was tested on three model organisms *B. subtilis* (Kunst *et al* 1997), *E. coli* K-12 (Blattner *et al* 1997) and *A. fulgidus* (Klenk *et al* 1997), and found to classify 97.3%, 96.8% and 96.2% of the annotated proteins of these organisms. The proportions of confounding entries were 1.6%, 3.6% and 1.4% respectively. A run of ARC on 348 prokaryotic proteomes classified 635,390 proteins out of 675,663 annotated proteins representing an overall success rate of 94.04%. Given that the keyword library of ARC was initially built from 2 representative model organisms with subsequent serial enrichment, the high success rates of classification by ARC shows that most annotation groups have been coherent and in phase with the earliest genome sequencing groups. It is now possible to formulate standardized annotation scheme. These results also attest that a great majority of proteins could be classified with singular decisions directing entry into a single class. However, a minority of annotation cases pose persisting problems by presenting plural decisions. One probable cause for such confounders

Table 1. Comparative proteomics of mycobacteria[a]

Species	Total number of proteins	Number of proteins classified into seven functional classes							Number of proteins in the ancillary classes		
		C	*D*[d]	*I*	*L*	*M*	*R*[d]	*S*	*O*	*U*	*H*
Mycobacterium tuberculosis H37Rv	3989	625	13	471	43	1137	12	34	224	87	1343
Mycobacterium bovis AF2122/97	3920	617	13	438	46	1128	12	34	205	74	1353
Mycobacterium leprae TN	1605	157[b,1,2]	10	211	50[c, 1,2]	512[c, 1,2]	8	15	38[b,1,2]	31	573

[a]We observed that the genomes of *M. tuberculosis* CDC1551 and *M. avium* subsp. *paratuberculosis* had an alarmingly high number of hypothetical proteins and therefore, these species were dropped from this analysis.

[b] Significantly different. Lower than expected proportion ($P < 0.01$) compared with *M. tuberculosis* H37Rv[1] or *M. bovis*[2].

[c] Significantly different. Higher than expected proportion ($P < 0.01$) compared with *M. tuberculosis* H37Rv[1] or *M. bovis*[2].

[d] The test for statistical significance was not carried out for these classes because of poor reliability due to small occurrences.

Table 2. Classification of genes in 4 categories of expression of *M. tuberculosis* clinical isolates[a]

	Consistently expressed	Low expression	Unexpressed	Variable expression	Total
C	50	127	151	44	372
D[b]	2	3	4	1	10
I	85**(+)	86*(–)	138	78	387
L	9**(+)	3	2	3	17
M	142	314	392	156	1004
R[b]	4	1	3	1	9
S	5	9	16	10	40
O	17	44	83	38	182
Total	314	587	789	331	2021

[a]The observed proportion of genes belonging to a given category of expression in each class was tested against the global expected proportion. In this test, we determined the preferential association (over-representation or under-representation) of the 4 categories of expression with the functional classes.

[b]Statistical tests in these classes were not carried out because low reliability arising out of small sample numbers and class size.

** $P < 0.001$, * $P < 0.01$; + over-representation (positive association), – under-representation (negative association).

is the non-uniformity in the emphasis on the process. For example, two annotations "Transcriptional regulator of Serine/Threonine kinase" and "Serine/Threonine kinase transcriptional regulator" result in a duplex decision. The former annotation places transcriptional regulation as the first description and therefore the given protein should enter class I for Information processing. The latter annotation places Serine/Threonine kinase as the first description and therefore the given protein should enter class S for signalling and communication. Since transcriptional regulation is the primary process carried out by the given protein, the former annotation should be followed. Thus, annotations following the thumb rule of describing the primary process first followed by biological context are classified by ARC unambiguously. Confounders arise when this order is usually reversed due to deviation from uniformity. Confounders could also arise in case of proteins with multiple functional domains falling into different functional classes.

ARC was tested on a total of 9903 GO terms of the categories 'molecular function' and 'cellular component' (Version-1.0, dated: 23:08:2006) (Harris *et al* 2004). Out of 7973 molecular function terms, 6742 were classified and out of 1930 cellular component terms, 1095 were classified by ARC. The remaining 2066 terms do not occur in any prokaryote annotation text. The number of confounders is 10. The low number of confounding terms in the GO scheme is presumably due to the careful preparation of GO terms by the GO group (Harris *et al* 2004).

3.2 *Applications*

ARC can be used efficiently to analyse (i) proteomes using annotation files with or without sequence, (ii) microarray data, and (iii) other forms of user defined data. Here, we describe one example of each case.

3.2.1 *Annotation files with or without sequence:* We take the case of mycobacteria. The annotations of complete proteomes of *M. tuberculosis* H37Rv (Cole *et al* 1998), *M. bovis* AF2122/97 (Klenk *et al* 2003), *M. leprae* TN (Cole *et al* 2001) appear to be of reasonably acceptable quality as judged by the lesser number of hypothetical proteins compared with the rest. The results of classification of these proteomes by ARC are displayed in table 1. The close relationship between *M. tuberculosis* and *M. bovis* is readily apparent in that the numbers of proteins in each class are similar and there is no observable statistically significant difference between the proteome content between these two species. The comparison of the proteome of *M. leprae* with *M. tuberculosis* and *M. bovis* presents some interesting features. According to the proteome size of *M. leprae*, it has lower number of proteins than expected in the functional classes *C* (cell wall, cell membrane and transporters) and *O* (others including virulence factors). On the other hand, the number of proteins in the classes *L* (translocation and secretion) and *m* (metabolism) are higher than expected.

Essentially, the protein content in the class *L* of *M. leprae* is very similar to the other two virulent mycobacteria despite the loss of so many genes. It has been explained that the reduction in proteome size of *M. leprae* is due to the accumulation of pseudogenes, whose counterparts in *M. tuberculosis* are intact (Cole *et al* 2001). Furthermore, it has been proposed that these pseudogenes may have arisen due to a combination of the loss of proofreading activity by DNA polymerase and the loss of sigma factors controlling

the transcription of their cognate genes (Madan Babu 2003). These two proposals provide a physical mechanistic explanation for the accumulation of pseudogenes in *M. leprae*. Our observation strongly suggests that the functional role of the genes also had played a significant role in selection of gene loss towards selected functional classes during reductive evolution. It is clear that the reductive forces have not touched the protein content (and their encoding genes) in class L. It is also interesting to note that while several metabolic pathways have disappeared in *M. leprae*, the metabolic protein content is still higher than that expected based on its proteome size. This result suggests that reductive elimination was shared between classes C and M and, the class C took the larger share of reductive evolution in order to preserve some metabolic pathways for its survival.

In terms of pathogenesis, it appears that targeting the proteins belonging to the class L should be accorded priority because of their conservative nature in relation to the niche colonized by mycobacteria. This set was analysed for orthologs and paralogs. Strikingly, this set of proteins had very low number of paralogs. *M. leprae* had no paralogs suggesting that the reductive evolution had erased gene duplications. Even *M. tuberculosis* and *M. bovis* had only one paralog pair annotated as secreted antigen 85-C and 85-A. The list of orthologs sharing common ancestry to all three species includes translocases, secreted protease, signal recognition particle, translocase SecY, secreted proteins, secreted antigens and signal peptidases. One secreted protease shares a common ancestry only between *M. bovis* and *M. leprae*. There are 37 secreted proteins unique to *M. leprae*. It is probable that the differences between the three species in secreted proteins may contribute towards different specificities during colonization of vertebrate hosts. The difference between *M. bovis* and *M. tuberculosis* although appears to be minor, is still worth investigating. It is apparent that our analytical tool has served to illuminate the differences between the mycobacterial species with respect to their biological characteristics.

3.2.2 *Microarray data:* We take the example of an elegant work by Gao *et al* (2005). In this work, gene expression diversity among *M. tuberculosis* clinical isolates was examined using microarrays. The list of genes is available as supplementary MS Excel file and they are categorized into 4 groups: Consistently expressed, Low expression, Unexpressed, Variable expression. We classified the genes in these categories using ARC. The results are displayed in table 2. In order to examine the statistical significance of the proportions of genes of each class appearing in the 4 categories, we formulated the following null hypothesis from statistical standpoint: In absence of any preferential association (negative or positive) of a given category to any of the functional classes, we expect the proportion observed in each functional class equal or nearly equal to the global proportion of genes in that class for a given category. Statistically significant deviation from this pattern in a given category would suggest positive or negative association of that category with the corresponding functional class.

We observed that the genes belonging to the classes I and L have significant positive association with consistent expression among the clinical isolates. Another interesting feature is the negative preference of low expressed genes with class I. These observations show that Information class genes are underrepresented in low expressed category. The category of consistently expressed genes would consist of housekeeping genes many of which include genes belonging to classes I and M. Housekeeping genes are generally predicted to be consistently expressed. This prediction appears to be met satisfactorily by the genes belonging to class I but not to class M. In the previous section, we observed that the class of secreted proteins is important for the biology of mycobacteria. Given that these genes also exhibit consistent expression, taken together, these results strongly point to an important role for the secreted proteins in mycobacterial physiology.

3.2.3 *Other forms of user defined data:* We examined the average cost of proteins belonging to each functional class across 333 prokaryotic species. The results are displayed in figure 2. In this case, the input data has functional annotation information along with user defined data on protein costs. The cost of a protein was determined as the sum total cost of the individual amino acids in the protein. The biosynthetic cost of each amino acid was taken from Akashi and Gojobori (2002). Although these biosynthetic costs are based on *E. coli* and *B. subtilis*, since the basic biosynthetic pathways for amino acid biosynthesis is anciently evolved and conserved, the costs in terms of high energy phosphate bonds (~*P*) are likely to be same across other prokrayotes. It is evident that the classes I and M show a relatively narrow distribution of average costs across all 333 species compared with other classes. The narrow distribution of cost in these classes is instructive. Class I consists of proteins such as ribosomal proteins and aminoacyl tRNA-synthetases, which are known to be conserved in sequence. Similarly class M consists of enzymes belonging to metabolic pathways, several of which are known to be conserved. Therefore, conservation of protein cost could emerge from a high degree of sequence conservation displayed by these proteins. In other classes, the wider distribution is indicative of greater diversity.

It must be noted that these examples demonstrate one possible approach for obtaining biological insights using ARC. Other forms of analysis are highly probable attempting to address multiple hypotheses.

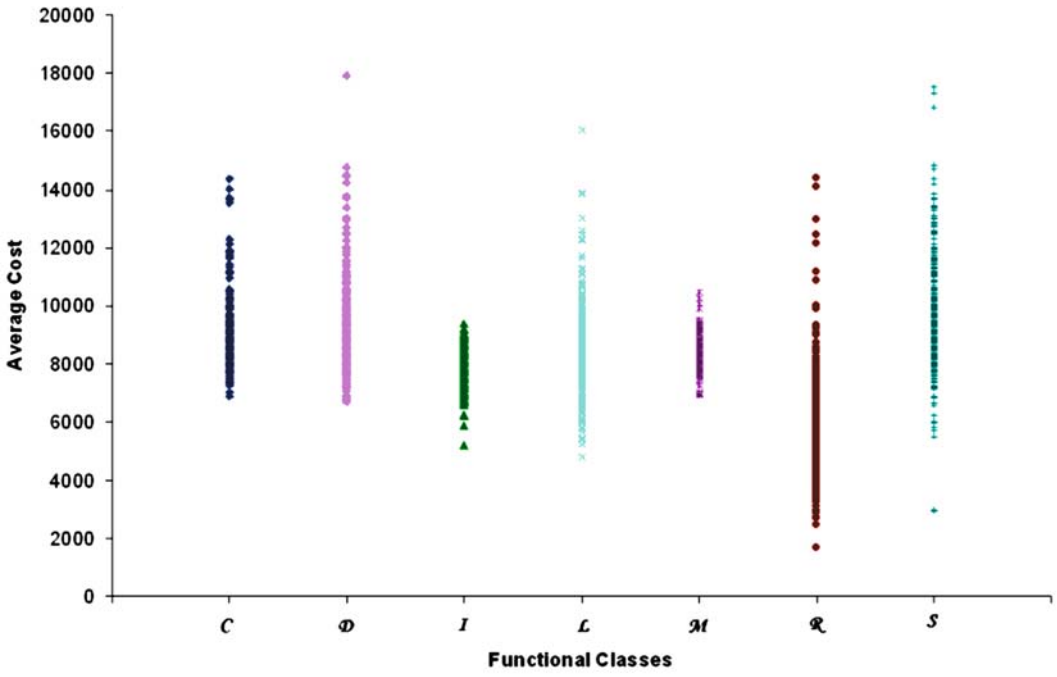

Figure 2. Classification of user defined protein costs data by ARC in 333 prokaryotes. Each point in a given functional class denotes the average cost of the proteins in one species in that class. Only 333 prokaryotes could be analysed because the annotations in some species appeared to be of low quality in that the numbers of proteins in some functional classes were zero.

Acknowledegement

MG, NK and SR thank the Council of Scientific and Industrial Research, New Delhi for funding support in the form of a grant CMM0017 Task Force "In Silico Biology for Drug Target Identification". We also thank Prof. Vani Brahmachari and Dr Beena Pillai for useful discussions, and Dr Souvik Maiti for encouragement.

References

Adams M D, Kerlavage A R, Fleischmann R D, Fuldner R A, Bult C J, Lee N H, Kirkness E F, Weinstock K G *et al* 1995 Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence; *Nature (London) (Suppl.)* **377** 3–174

Akashi H and Gojobori T 2002 Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis; *Proc. Natl. Acad. Sci. USA* **99** 3695–3700

Andrade M A, Ouzounis C, Sander C, Tamames J and Valencia A 1999 Functional classes in the three domains of life; *J. Mol. Evol.* **49** 551–557

Auffray C, Imbeaud S, Roux-Rouquié M and Hood L 2003 Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space-time; *Philos. Trans. A Math. Phys. Eng. Sci.* **361** 1125–1139

Blattner F R, Plunkett III G, Bloch C A, Perna N T, Burland V, Riley M, Collado-Vides J, Glasner J D *et al* 1997 The complete genome sequence of *Escherichia coli* K-12; *Science* **277** 1453–1474

Cole S T, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V, Eiglmeier K *et al* 1998 Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence; *Nature (London)* **393** 537–544

Cole S T, Eiglmeier K, Parkhill J, James K D, Thomson N R, Wheeler P R, Honore N, Garnier T *et al* 2001 Massive gene decay in the leprosy bacillus; *Nature (London)* **409** 1007–1011

Gao Q, Kripke K E, Saldanha A J, Yan W, Holmes S and Small P M 2005 Gene expression diversity among Mycobacterium tuberculosis clinical isolates; *Microbiology* **151** 5–14

Grunewald J and Marahiel M A 2006 Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides; *Microbiol. Mol. Biol. Rev*. **70** 121–146

Harris M A, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S *et al* 2004 The Gene Ontology (GO) database and informatics resource; *Nucleic Acids Res.* **32** D258–D261

Klenk H P, Clayton R A, Tomb J F, White O, Nelson K E, Ketchum K A, Dodson R J, Gwinn M *et al* 2003 The complete genome sequence of *Mycobacterium bovis*; *Nature (London)* **100** 7877–7882

Klenk H P, Clayton R A, Tomb J F, White O, Nelson K E, Ketchum K A, Dodson R J, Gwinn M *et al* 1997 The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*; *Nature (London)* **390** 364–370

Kondrashov F A, Rogozin I B, Wolf Y I and Koonin E V 2002 Selection in the evolution of gene duplications; *Genome Biol.* **3** research 0008.1–0008.9

Kunst F, Ogasawara N, Moszer I, Albertini A M, Alloni G, Azevedo V, Bertero M G, Bessieres P *et al* 1997 The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*; *Nature (London)* **390** 249–256

Madan Babu M 2003 Did the loss of sigma factors initiate pseudogene accumulation in *M leprae*?; *Trends Microbiol.* **11** 59–61

Pillet V, Zehnder M, Seewald A K, Veuthey A L and Petrak J 2004 GPSDB: a new database for synonyms expansion of gene and protein names; *Bioinformatics* **21** 1743–1744

Riley M 1993 Functions of the Gene Products of *Escherichia coli*; *Microbiol. Rev*. **54** 862–952

Subramanyam M B, Gnanamani M and Ramachandran S 2006 Simple sequence proteins in prokaryotic proteomes; *B.M.C. Genomics* **11** 141

Uitenbroek D G 1997 SISA Binomial. Southampton *http://home.clara.net/sisa/binomial.htm*

Van Regenmortel M H 2004 Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism; *EMBO Rep.* **5** 1016–1020

*e*Publication: 16 June 2007