Comparative genomics using data mining tools

TANNISTHA NANDI, CHANDRIKA B-RAO and SRINIVASAN RAMACHANDRAN*

Functional Genomics Unit, Centre for Biochemical Technology, Mall Road, Delhi 110 007, India

Corresponding author (Fax, 91-11-725-7471; Email, ramu@cbt.res.in).

We have analysed the genomes of representatives of three kingdoms of life, namely, archaea, eubacteria and eukaryota using data mining tools based on compositional analyses of the protein sequences. The representatives chosen in this analysis were *Methanococcus jannaschii*, *Haemophilus influenzae* and *Saccharomyces cerevisiae*. We have identified the common and different features between the three genomes in the protein evolution patterns. *M. jannaschii* has been seen to have a greater number of proteins with more charged amino acids whereas *S. cerevisiae* has been observed to have a greater number of hydrophilic proteins. Despite the differences in intrinsic compositional characteristics between the proteins from the different genomes we have also identified certain common characteristics. We have carried out exploratory Principal Component Analysis of the multivariate data on the proteins of each organism in an effort to classify the proteins into clusters. Interestingly, we found that most of the proteins in each organism cluster closely together, but there are a few 'outliers'. We focus on the outliers for the functional investigations, which may aid in revealing any unique features of the biology of the respective organisms

1. Introduction

The number of complete genome sequences available currently is about 33 (http://www.ncbi.nlm.nih.gov). The lists of organisms whose genomes have been sequenced or are being sequenced include the archaea, the bacteria and the eukaryotes. The set of microbes whose genomes have been sequenced so far is a diverse one, ranging from organisms living under extreme conditions of environment to the model organisms of biology, and to some important human pathogens. The genome sizes of these organisms also range from the smallest known so far (*Mycoplasma genitalium* 580 kb; Fraser *et al* 1995), to much larger ones (human genome 3×10^6 kb). The success in completing microbial genome sequencing projects at a rapid pace has sparked off the investigations of interesting questions in biology. For example, the minimal gene set required to sustain the existence of a modern-type cell can now be assessed (Mushegian and Koonin 1996; Hutchinson *et al* 1999).

The availability of complete genome sequences from numerous microorganisms and the human genome sequence holds the promise to unravel the mysteries of the complicated biology of these organisms. It is now possible to address the most fundamental questions in biology through functional genomics. These include the number of genes encoded in an organism, their functional roles and the analysis of biochemical pathways at different states of the organism that manifest themselves into the present living form as we see them.

The post genome sequencing period is expected to be dominated by the functional characterization of the genome sequences and this activity is referred to as functional genomics. Some of the new experimental approaches in this area include Microarray and Gene Chip technology. Other activities include the analysis of proteins in an organism and it is referred to as proteomics. The combined experimental approaches are expected to enhance our understanding of the biology of these organisms.

The present literature consists of reports describing the comparative analysis of genomic sequences. These analyses have revealed certain interesting features such as

the generally high degree of conservation of microbial proteins – approximately 70% of them contain ancient conserved regions. This allowed some researchers to delineate families of orthologs across a wide phylogenetic range and, in many cases, predict protein functions with considerable precision. Numerous orthologous and paralogous relationships have also been brought to light and a database of Cluster of Orthologous Groups (COG) has been developed (Tatusov *et al* 1997; Koonin *et al* 1998; Tatusov *et al* 2000).

Another set of activity is the classification of the genes into different functional categories. In this case a single representative organism was selected for each of the 3 kingdoms of life. Comparative analysis is subsequently used to study the different features between the different genomes (Andrade *et al* 1999). These analyses reveal that proteins related with ENERGY processes are generally represented in all three domains, while those related with COMMUNICATION represent the most distinctive functional feature of each single domain and functions related with INFORMATION processing (translation, transcription, and replication) show a complex behaviour. Proteins in the superclass archaea are related with proteins of either eukarya or bacteria, as recognized previously. The distribution of functional classes in the three domains accurately reflects the principal characteristics of cellular life forms.

Comparative genomics can be carried out in many ways. For example, some authors have been able to predict transcription regulatory sites in Archaea by a comparative genomics approach. The results demonstrated the feasibility of prediction of at least some transcription regulatory sites by comparing poorly characterized prokaryotic genomes, particularly when several closely related genome sequences are available (Gelfand *et al* 2000). Using comparative genomics approach it has been possible to identify novel *cis*-acting regulatory elements in genomes (Raghavan *et al* 2000).

We are interested in developing data mining tools using various approaches to decipher important patterns from the genome sequence databases. In one such work we have developed an approach to analyse sequence complexities of protein sequences. This approach has enabled us to partition the encoded proteins in different genomes into different categories of sequence complexities. After analysing the functional composition of these categories we observed that we could segregate the encoded proteins (and their corresponding genes) into groups that would help us in designing suitable experiments for functional genomics (S Ramachandran, T Nandi, Chandrika B-Rao, S K Brahmachari and D A Dash, unpublished results).

In the present work, we describe our data mining analysis of the genomic sequences from three representatives

of the 3 kingdoms of life as chosen previously (Andrade *et al* 1999). We have developed a software that carries out the analysis of different attributes of proteins. Principal Component Analysis is subsequently used to examine clustering patterns. The functional composition analyses of these data are presented.

2. Methods

2.1 *Protein attributes*

Software programs were written in PERL (Practical Extraction and Reporting Language).

Variate 1 is the length (L) of the protein in number of amino acids.

Variate 2 is the percent of charged amino acids in a given protein. The charged amino acids were aspartic acid (D), glutamic acid (E), lysine (K) and arginine (R). Percent charge is given by

$$\frac{\text{Number of charged amino acids}}{\text{Total number of amino acids}} \times 100 \qquad (1)$$

Variate 3 is the percent hydrophobicity of the protein. We have used the Fauchere and Pliska scale (Fauchere and Pliska 1983) to classify the amino acids into hydrophobic group. Percent hydrophobicity is given by

$$\frac{\text{Number of hydrophobic amino acids}}{\text{Total number of amino acids}} \times 100 \qquad (2)$$

Variate 4 is a measure of distance of a protein sequence from a fixed reference point. The distance is measured according to the formula:

$$\text{Distance } (D)_{\text{fixed}} = \sqrt{\sum_{i=1}^{20} (O_i - E_i)^2} \qquad (3)$$

Where O_i is the observed number of amino acid of type 'i' in the concerned protein and E_i, the expected number of amino acid of type 'i' in the same protein. E_i is $L/20$ considering all amino acid to be uniformly distributed in the protein. We refer to this point as the fixed reference point. D_{fixed}/L is a normalized measure of distance for the concerned protein.

Variate 5 is the distance of a protein sequence from a variable reference point. The distance $D_{\text{var, globular}}$ has the same formula as that in Variate 4 but the E_i is calculated according to the formula:

$$E_i = f_i \times L \qquad (4)$$

where L is the length of the concerned protein in amino acids and f_i is the average frequency of occurrence of the ith amino acid in the set of proteins that are of high sequence complexity and are predicted to have globular fold within the same genome. For this purpose, we first run the protein sequences encoded in the genome through our sequence complexity analysis computer program and classify these proteins into 3 sets, namely, high complexity, medium complexity, and low complexity according to the fraction of the low complexity sequences present in each of the proteins.

The average frequency of each of the 20 amino acids from the high complexity set of proteins was computed by calculating the number of occurrences of the ith ($i = 1$ to 20) amino acid in the proteins set divided by the total number of amino acids in the same set. The sequence complexity of the proteins were computed in a related work (S Ramachandran, T Nandi, Chandrika B-Rao, S K Brahmachari and D A Dash, unpublished results) wherein we have developed a formula and have shown that our formula computes the sequence complexity with matching accuracy (high complexity – 94% correct; medium complexity – 66% correct; low complexity – 67% correct) as other computer programs and the results can be used to correlate with the protein folding type in space such as globular vs non-globular. We refer to this point as a variable reference point because the frequency of the different amino acids appearing in the globular set of proteins are unequal and variable in different genomes. As in Variate 4, $D_{var, globular}/L$ is a normalized measure of distance with respect to the variable reference point.

2.2 *Statistical*

All statistical procedures were carried out using the SAS package (SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, USA). Principal Component Analysis using correlation coefficients between the variates was carried out using this package.

2.3 *Sequences*

Sequences were retrieved from the National Centre for Biotechnology Information by anonymous ftp transfer from the subdirectory/genbank/genomes/.

2.4 *Sequence analysis*

Sequence analysis was carried out using the Wisconsin Package Version 10·0, Genetics Computer Group (GCG), Madison, Wisconsin, USA.

3. Results and discussion

We selected one representative organism whose complete genome has been sequenced from each of the 3 kingdoms of life namely, archaea, bacteria and eukaryota. These are *Methanococcus jannaschii*, *Haemophilus influenzae* and *Saccharomyces cerevisiae*. The normalized class distributions of the different attributes (variates) of the different proteins within the different genomes are shown in figures 1–5. The entire range of data points was divided
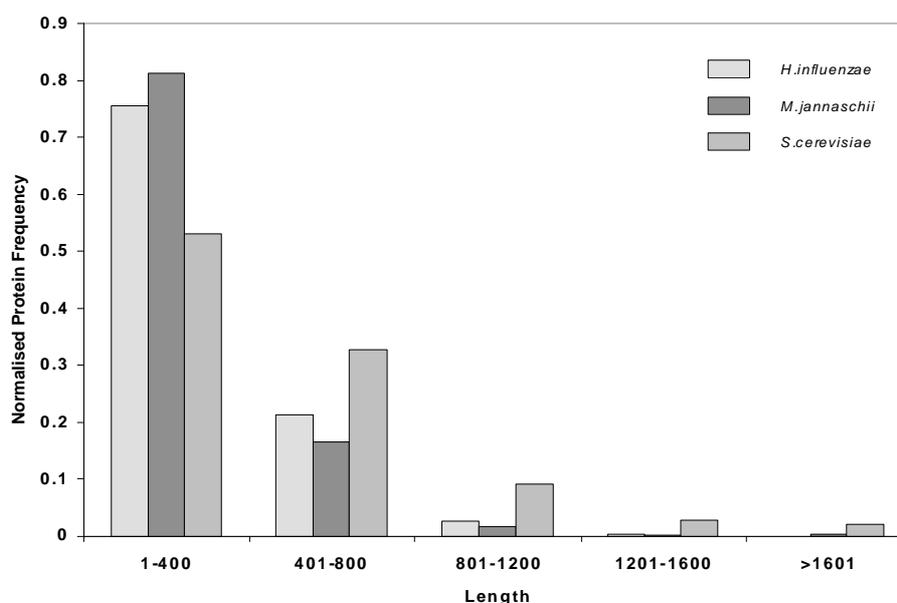

Figure 1. The length variate distribution pattern the 3 different genomes.

into equal sized class intervals. The number of proteins present at each interval, in each genome, was computed and the resulting distribution was plotted. The distribution pattern of the variates, namely, the normalized protein frequency was plotted against the length, percent charge, percent hydrophobicity, D_{fixed}/L, $D_{var, globular}/L$ are shown in figures 1–5.

In the case of length distribution, the maximum protein frequency was observed within the range of 400 residues

in all the three genomes indicating that it is in this category that most of the proteins group. The frequency sharply decreases as the length increases. The distribution pattern of *S. cerevisiae* was different from that of *H. influenzae* and *M. jannaschii.*

The percent hydrophobicity distribution of the proteins in the three genomes is shown in figure 2. We have used the Fauchere and Pliska scale as it has been reported that this method is the most direct method (Fauchere and
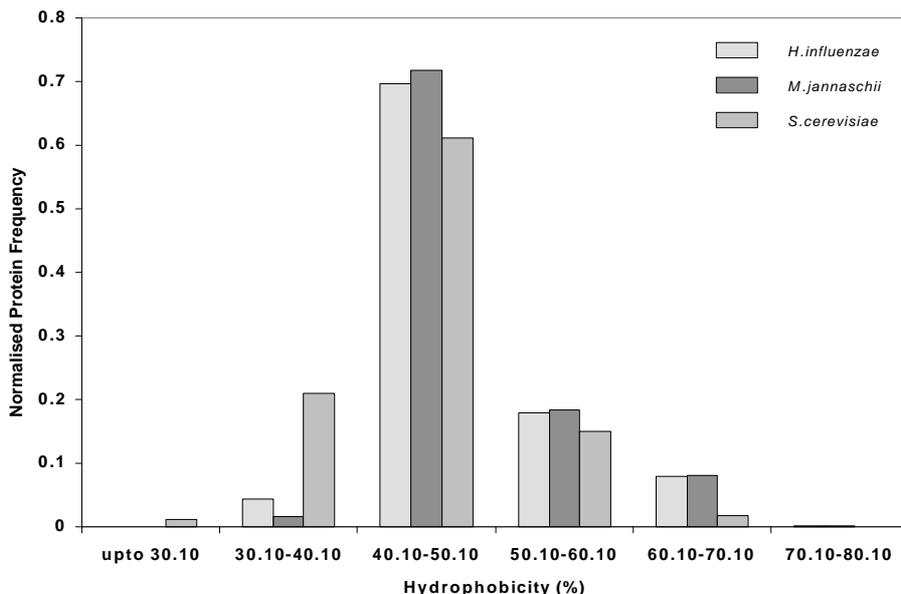

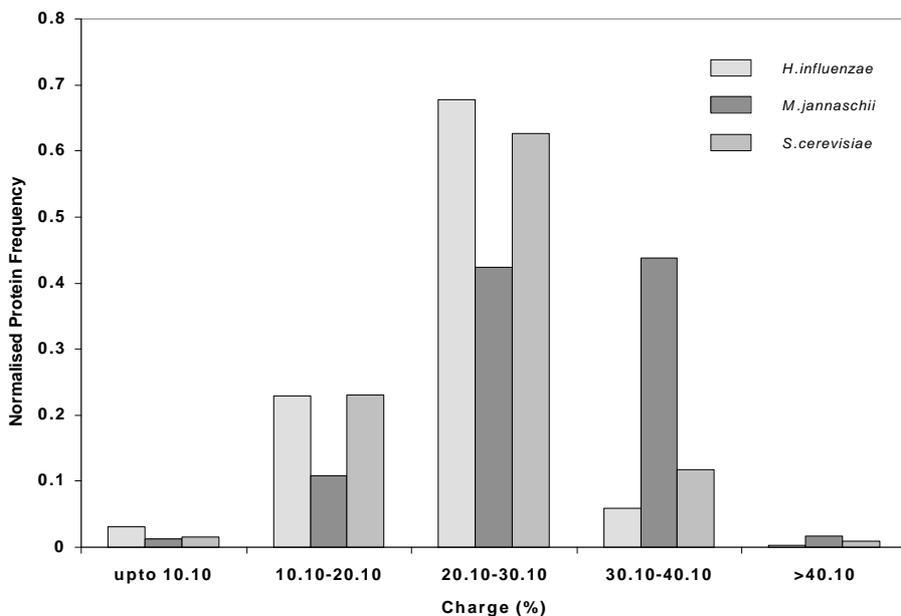**Figure 2.** The percent hydrophobicity distribution patterns in the three genomes

Figure 3. The percent charge distribution patterns in the three genomes.

Pliska 1983; Gribskov and Devereux 1992). In all the three genomes, the maximum number of proteins lies within the hydrophobicity range of 40%–50%. Also, the protein frequency decreases sharply on either side of this range in all the three cases. The distribution pattern was similar in the three cases except that in *S. cerevisiae* there is a larger fraction of proteins in the 30%–40% hydrophobicity range and a lower fraction in the 40%–50%

range. Among the proteins in the category 40%–50%, the majority of them were predicted to be globular using sequence complexity analysis in all the 3 genomes.

The percent charge distribution of the proteins in the three genomes is shown in figure 3. In this case, the distribution pattern of *M. jannaschii* was different from those of *H. influenzae* and *S. cerevisiae*. The distribution has a maximum in the range of 20%–30% charge in
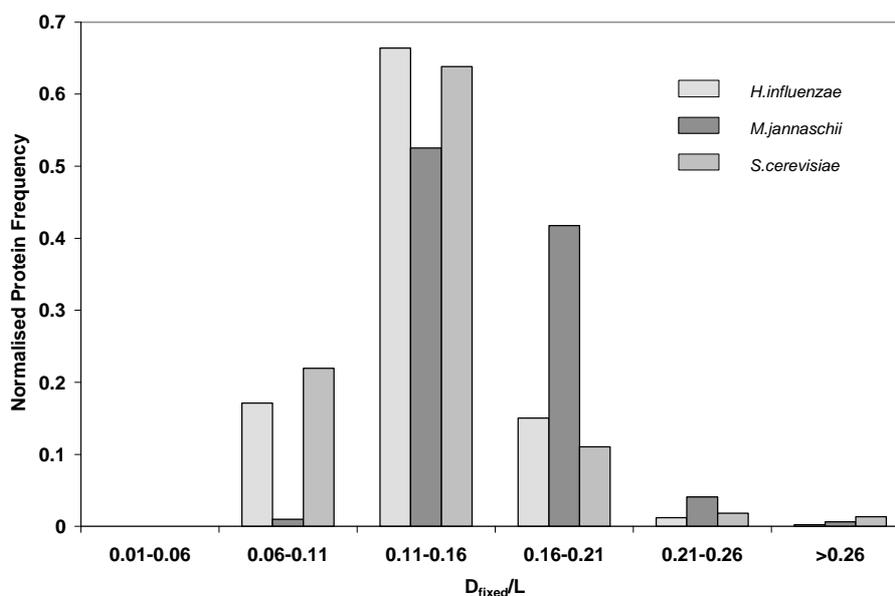
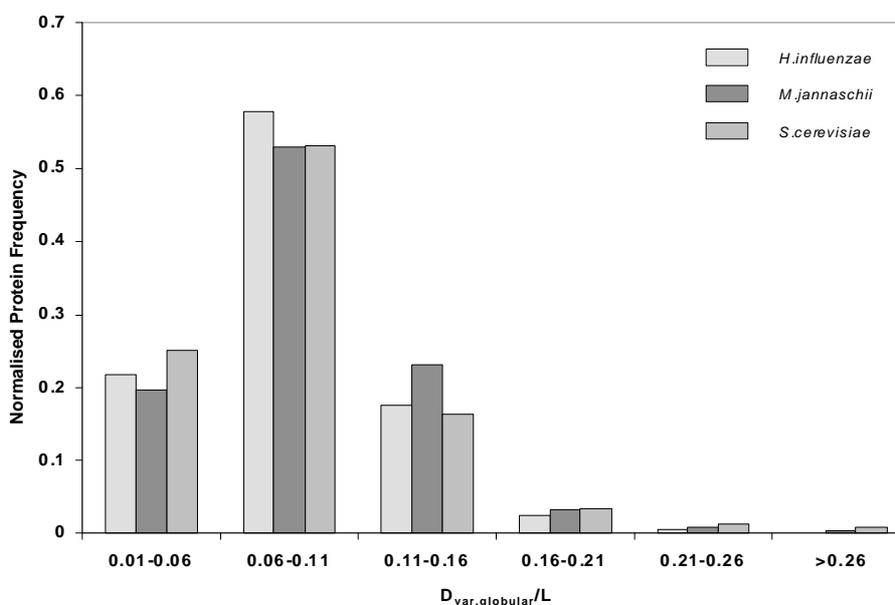**Figure 4.** The normalized fixed distance (D_{fixed}/L) distribution in the three genomes.

Figure 5. The normalized variable distance ($D_{var,\ globular}/L$) distribution in the three genomes.

case of *S. cerevisiae* and *H. influenzae*. In case of *M. jannaschii*, the number of proteins in this range was lower whereas the number of proteins in the range 30%–40% was higher compared to the other 2 organisms. This suggests that in *M. jannaschii* there is a greater fraction of proteins with higher amount of charged amino acids. The higher proportion of charged amino acids might aid in providing additional structural support for these proteins in the form of electrostatic interactions that would be required for this organism since *M. jannaschii* was isolated from hot temperature environment (Natesh *et al* 1999).

The normalized measure of fixed distance distribution of the proteins in the three genomes is shown in figure 4. In this case it was observed that the major fraction of the proteins in the bacterial and the yeast genomes group into a modest distance from the fixed reference point. The *M.*

jannaschii genome however shows a deviation from this pattern in that there is a greater number of proteins at a slightly larger distance interval indicating that the number of proteins deviating from the fixed reference point is greater as compared to the other two genomes, namely, *H. influenzae* and *S. cerevisiae*. The differences in charge composition in *M. jannaschii* contributes to the differences observed in the fixed distance distribution pattern of this genome compared with the others. The number of proteins at lower or larger distances is uniformly low in all the three organisms.

The normalized measure of variable distance distribution of the proteins in the three genomes is shown in figure 5. In this case, the distances of the proteins in amino acid composition is with respect to the frequencies of the different amino acids of the high complexity proteins within the same genome. The maximum number
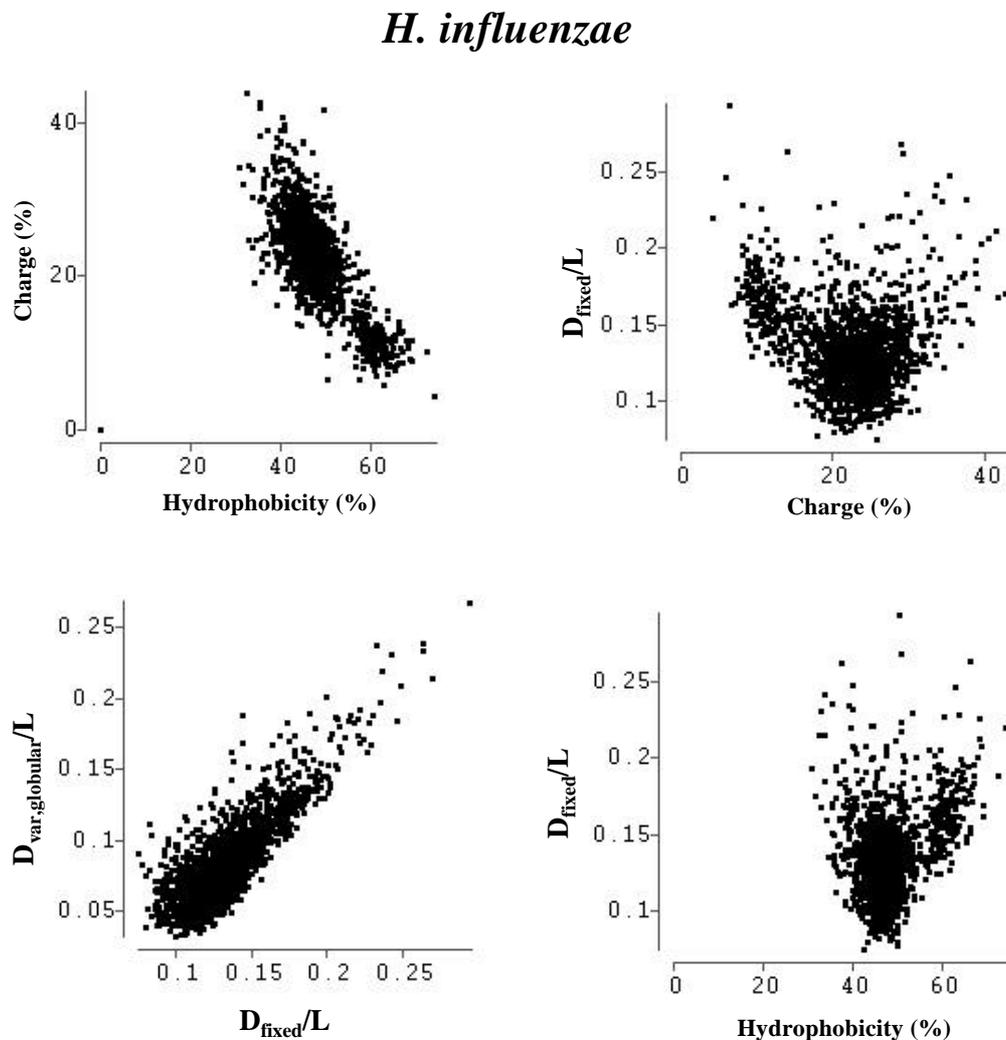
H. influenzae

Figure 6.   The bivariate plots for the *H. influenzae* genome.

of proteins in all the three-organism lie within a moderate range of distance and the number of proteins having short or long distances are few. The striking feature in this case of distribution compared with that of the fixed distance distribution is the near uniformity of the distribution pattern between all the three genomes despite the fact that in case of *M. jannaschii* we observed some deviation in other distribution patterns compared to the other two genomes. Because the distances in this case are with respect to the same genomes, it is likely that the basic compositional characteristics of the variable reference point in each genome represents its own characteristic. This indicates that the compositional variation from the variable reference points even though they are different in each of the genomes, is similar between the different genomes.

The bivariate plots for the three different genomes are shown in figures 6–8. It can be observed that the bacterial and the archaeal genomes have two clusters in the hydrophobicity plots whereas there is no such feature in the yeast genome. In all the bivariate plots it can be observed that most of the proteins encoded in the different genomes cluster into a dense group. A few proteins remain outside the cluster. These proteins are placed outside the main cluster in these bivariate plots because they have different compositional features from the rest of the proteins in the different genomes.

To analyse these genomes further and to identify the outliers in a composite approach using all the variates and to obtain useful patterns giving new leads to designing experiments on selected genes, we have used the principal component analysis (PCA). This statistical method enables the reduction of the large amount of multivariate data to a small number of dimensions that can be selected objectively, and hence facilitates the
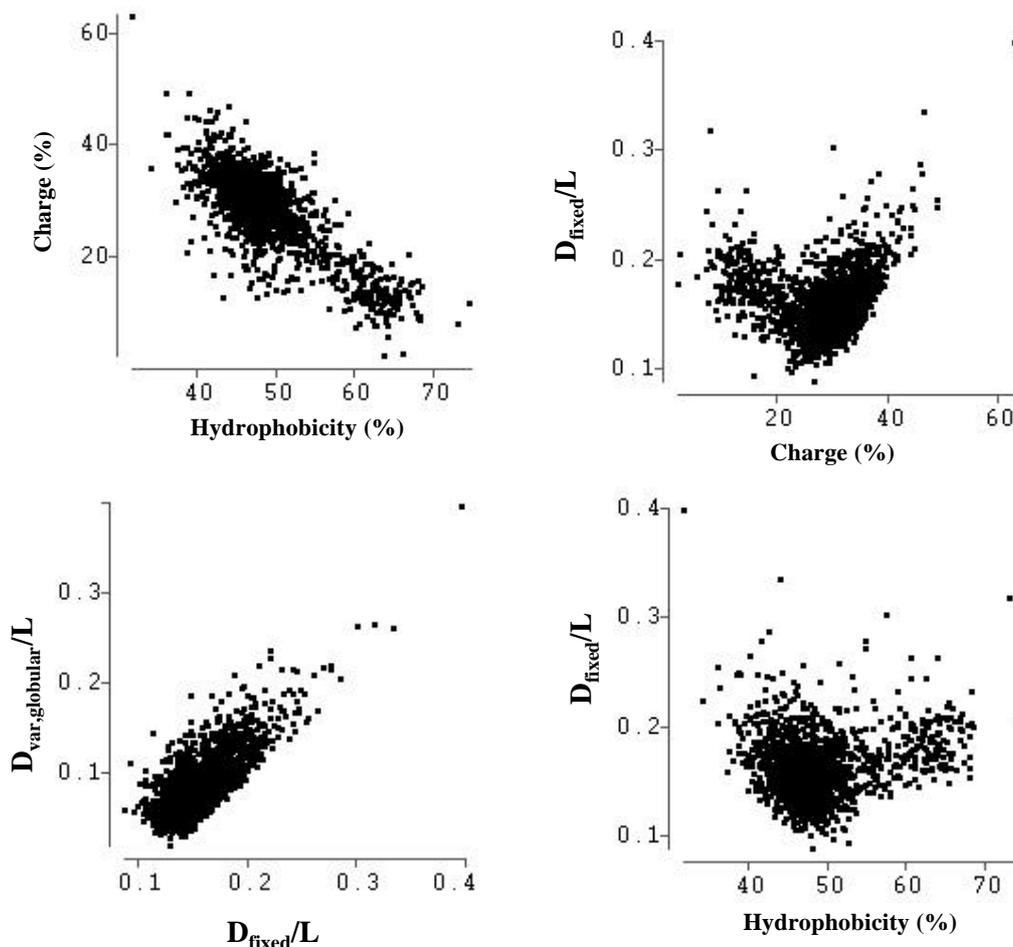
M. jannaschii

Figure 7. The bivariate plots for the *M. jannaschii* genome.

extraction of clustering patterns among the proteins. In addition we can identify proteins that are 'outliers' with respect to the main cluster. Such analyses have been used recently in developing procedures for identifying excreted or cytoplasmic proteins (Van Heel 1991; Casari *et al* 1995; Forster *et al* 1999; Schneider 1999). The results of these analyses are shown in table 1. It can be observed that, in different genomes, the identified proteins that are 'outliers' are different yet there are some common features.

For example in bacteria and in archaea it was observed that among several hypothetical proteins, there are also proteins whose putative functions have been identified through homology searches. The *S. cerevisiae* genome is strikingly different in this regard. Most of the proteins that are very different from the rest of the proteins in the whole genome are hypothetical at the present time. One protein whose function has been predicted to have a role in damage-response falls in this group.

The proteins that are far removed from the main cluster as identified through PCA could represent acquired pro-

teins either from horizontal transfer or they may have evolved for some special purpose that is required for some biological function for the particular niche which the organism occupies or they may have evolved for some biological purpose that is characteristic of the organism. On the other hand, the hypothetical proteins belonging to the main cluster may have a role in routine biochemical function such as, amino acid and nucleotide biosynthesis, cellular processes, central intermediary metabolism, energy metabolism, lipid metabolism, regulatory functions, replication, transcription, translation, transport and a few other categories or biochemical functions that are closely related to these.

Among the outliers, are a few ribosomal proteins (*M. jannaschii* and *H. influenzae*) or hydrogenase forming protein (*H. influenzae*). The significance of a few ribosomal proteins and hydrogenase formation protein falls as outliers is not clear at present. The rest are either hypothetical or conserved hypothetical proteins. BLAST search results show that the conserved hypothetical protein
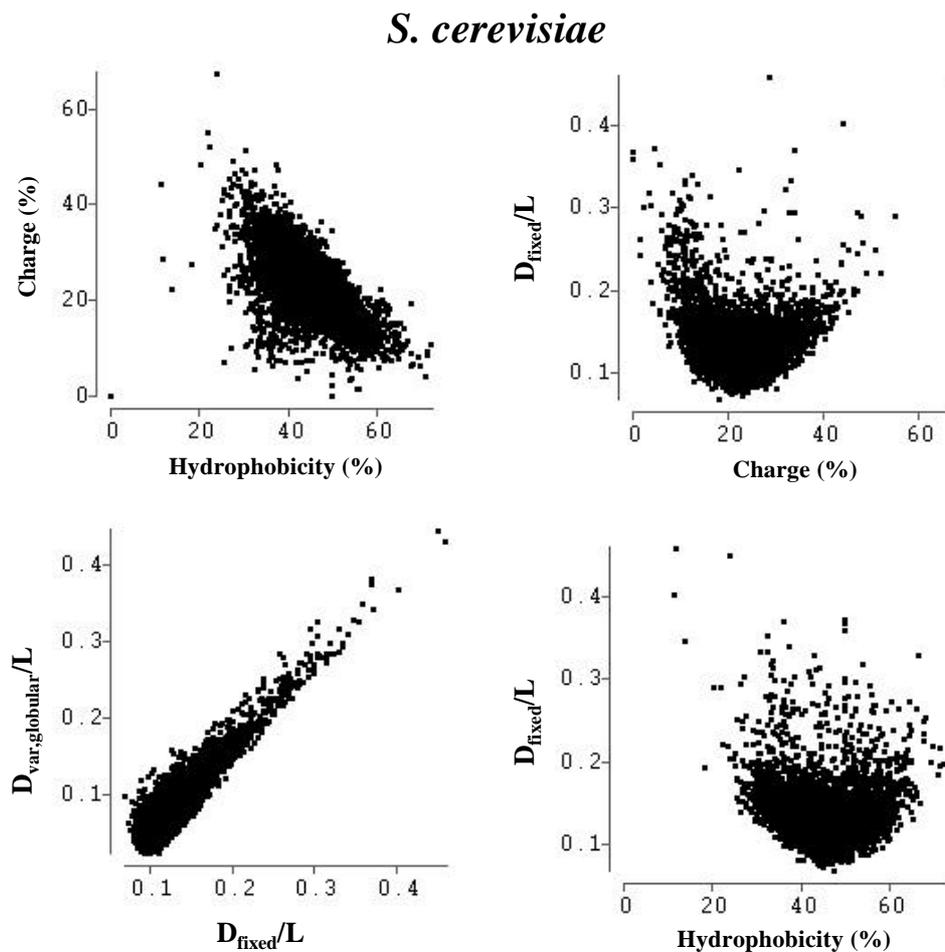

Figure 8.   The bivariate plots for the *S. cerevisiae* genome.

Table 1. The proteins that were markedly different from the main cluster identified using PCA in the different genomes.

Organism	Database identification number	Functional annotation
M. jannaschii	gi|1592277|||	LSU ribosomal protein L41E
	gi|1590961|	Predicted coding region MJ0223
	gi|1499306|	LSU ribosomal protein L12A(P1)
	gi|1590876|	LSU ribosomal protein L37E
	gi|2826390|	Predicted coding region MJ1282·2
	gi|1591221|	Predicted coding region MJ0519
	gi|1498959|	Predicted coding region MJ0185
H. influenzae	gi|1574457|	Predicted coding region HI1601
	gi|1573639|	Ribosomal protein L7/L12(rpl7/12)
	gi|1574794|	Predicted coding region HI1327
	gi|1574229|	Hydrogenase formation protein (hypG)
	gi|1574625|	Conserved hypothetical protein
	gi|3212183|	Predicted coding region HI0148·1
S. cerevisiae	gi|1431299|	ORF YDL 184c
	gi|1903290|	ORF YDL 133c-a
	gi|1903289|	ORF YDL 133c-a
	gi|1945332|	ORF YGR 159c
	gi|486581|	ORF YKR 092c
	gi|854443|	Ddr 48p (DNA damage-responsive protein 48)
	gi|1420189|	ORF YOR053w

from *H. influenzae* has homologues only in *Escherichia coli* and *Bacillus subtilis*. In the case of *S. cerevisiae*, except for a DNA damage-responsive protein, the rest are hypothetical. The damage-responsive protein has been shown to function in the production or recovery of mutations in *S. cerevisiae*. It is probable that the other proteins that are outliers may have some special function. Therefore, out of a large number of proteins, the proteins (and their corresponding genes) which fall, as 'outliers' could be important for experimental analysis through functional genomics approaches.

Our predictions finds support when we apply a statistical approach to the analyses of proteins encoded in genomic sequences. For example, in our previous work we have shown that some of the genes identified in the 21 different genomes using sequence complexity features belong to some special category of biological function. This function was found to be characteristic of the life style of the different organisms. In this work we have described the development and application of one data-mining tool that can be used to reveal the features of different genomes. Our tools can be used in variety of ways. For example the variable reference point used in this work could be framed differently to suit other applications.

Compositional differences in proteins have been used in various analyses and have been correlated with different aspects of protein evolution such as structure (Nakashima *et al* 1986; Wootton 1994; S Ramachandran, T Nandi, Chandrika B-Rao, S K Brahmachari and D A Dash, unpublished results), and overall cellular localization (Nakashima and Nishikawa 1992, 1994; Schneider and Wrede 1993; Schneider 1999). Although our work analyses compositional features, it uses different formulae. Through our analysis we were able to observe the differences between the 3 different genomes representing the three different kingdoms of life as well as the common features between these genomes. Using our approach we were able to correlate the compositional differences between different proteins and their overall function; whether they have a general biochemical function (the main cluster) or they have a role in certain specialized biological function (the proteins that are isolated from the main cluster).

It would certainly be most useful if we were able to predict the precise biological role of numerous hypothetical proteins that may aid in designing more specific experiments. We are currently exploring the neural network applications in this connection.

thanks Prof. Partha P Majumder and Prof. S K Brahma-chari for fruitful discussions.

Appendix

The method used to analyse sequence complexity of a protein sequence is described in the following text:

For computing complexity in protein sequences, the number of occurrences of each amino acid in the sequence, namely, the amino acid composition is first calculated. Subsequently, the maximum possible number of different dimers (dipeptides) N_{max} for the given amino acid composition is calculated using the dimer word count approach. The observed number of different dimers (dipeptides) in the sequence is calculated in 2 non-overlapping frames. An average of the number of different dimers in the 2 frames is taken. This average number is denoted by N_{obs}. The ration (N_{obs}/N_{max}) represents the fraction of the maximum possible different dimer words in the given sequence. For computing the complexity of a given sequence, we need to compute an additional "measure of skewness". This is because, natural sequences generally have compositional bias. Consequently, the N_{max} is lower than that would be in an ideally compositionally unbiased sequence in which the number of occurrences of each amino acid is equal. To compute the "measure of skewness" we have,

$$\Delta_{obs} = \sum_{I=1}^{20} (O_I - L/20)^2, \tag{1}$$

where L is the length of the sequence and it is maximized in the extreme case of a homopolymeric sequence where O_I for one amino acid is L and the rest 0 in equation (1). We denote this value Δ_{max}.

The ratio ($\Delta_{obs}/\Delta_{max}$) is a normalized measure of 'skewness' caused by the amino acid composition deviation from the equal frequency of occurrence of all the 20 amino acids in a protein of length L amino acids.

We define complexity of a sequence as

$$\text{Complexity} (\boldsymbol{g}) = \{(N_{obs}/N_{max})\} - (\Delta_{obs}/\Delta_{max}). \tag{2}$$

In the homopolymeric case, say $(Q)_n$ (a polyglutamine stretch), complexity is 0 as $N_{obs} = N_{max} = 1$ and $\Delta_{obs} = \Delta_{max}$. In the case of uniform distribution, where all amino acids occur with equal frequency, the second term in equation (2) becomes 0 and complexity is simply a measure of the fraction of the maximum possible dimers that can be theoretically formed. Word counts of trimers and other higher lengths were not used as the likelihood of their repetitive occurrences become lower and did not provide additional information in assessing the sequence complexity.

Complexity analysis: Sequence analyses were carried out by scanning each sequence with a window of size 45 in steps of one amino acid. From the test cases we inferred that sequences in which a complexity value of less than 0·8 was registered, corresponded to 'low complexity segments' as defined earlier (Wootton 1994). Conversely, sequences that had a complexity value of greater than 0·8, corresponded to high complexity segments and they had globular structures as evidenced from their crystal structures. The inferred protein sequences in different genomes were classified on the basis of the fraction ($F_{complexity}$, expressed as percentage) of low complexity segments of the total number of segments in the protein. This enables direct comparisons and partitioning of the proteins encoded in different genomes into different categories of complexity.

First we tested our algorithm on a few proteins whose results using the SEG algorithm was published earlier (Woottom 1994). The results of 4 cases such as myosin, collagen, human cartilage specific proteoglycan core protein, and the human CAN protein are shown in figure 1. When compared with the results from SEG we observed that both the plots were very similar. The segments of the protein sequences differing significantly in sequence complexity can be clearly inferred. Thus, the dimer word count approach gives very similar results, if not identical, to the SEG program

Correlation with PDB structures: A total of 382 proteins were analysed. An inverse relation was observed between the overall 3-D structure and the complexity results. The results were compiled into 3 categories of complexity. These 3 categories are 'high complexity' ($F_{complexity} = 0$), 'medium complexity' ($F_{complexity} = 0-15$), and 'low complexity' ($F_{complexity} > 15$). Proteins that were of 'high complexity' had globular or near globular structures (94% of the cases). Proteins with 'medium complexity' also had globular or near globular structures (66%) whereas the proteins that were of low 'complexity' had mostly rod-like structure and were far from being globular (67% of the cases). In a few exceptions, we also analysed those sequences using SEG. The SEG program with default parameters also gave very similar results (data not shown) indicating that, in these sequences the predictions cannot be made correctly on the basis of sequence information alone.

References

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

Casari G, Sander C and Valencia A 1995 A method to predict functional residues in proteins; *Struct Biol.* **2** 171–178

Fauchere J L and Pliska V 1983 Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides; *Eur. J. Med. Chem.-Chim. Ther.* **18** 369–375

Forster M J, Heath A B and Afzal M A 1999 Application of distance geometry to 3D visualization of sequence relationships; *Bioinformatic* **15** 89–90

Fraser C M, Gocayne J D, White O, Adams M D, Clayton R A, Fleischmann R D, Bult C J, Kerlavage A R, Sutton G, Kelley J M, Fritchmann J L, Weidman J F, Small K V, Sandusky M, Fuhrman J, Nguyen D, Utterback T R, Saudek D M, Phillips C A, Merrick J M, Tomb J F, Dougherty B A, Bott K F, Hu P C, Lucier T S, Peterson S N, Smith H O, Hutchison III C A and Ventor J C 1995 The minimal gene complement of *Mycoplasma genitalium*; *Science* **270** 397–403

Gelfand M S, Koonin E V, Mironov A A 2000 Prediction of transcription regulatory sites in Archaea by a comparative genomic approach; *Nucleic Acids Res.* **28** 695–705

Gribskov M and Devereux J (eds) 1992 *Sequence Analysis Primer* (Oxford: Oxford University Press) pp 67–71

Hutchison C A, Peterson S N, Gill S R, Cline R T, White O, Fraser C M, Smith H O and Venter J C 1999 Global transposon mutagenesis and a minimal Mycoplasma genome; *Science* **286** 2165–2169

Koonin E V, Tatusov R L and Galperin M Y 1998 Beyond complete genomes: from sequence to structure and function; *Curr. Opin. Struct. Biol.* **8** 355–363

Mushegian A R and Koonin E V 1996 A minimal gene set for cellular life derived by comparison of complete bacterial genomes; *Proc. Natl. Acad. Sci. USA* **93** 10268–10273

Nakashima H and Nishikawa K 1992 The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins; *FEBS Lett.* **303** 141–146

Nakashima H and Nishikawa K 1994 Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies; *J. Mol. Biol.* **238** 54–61

Nakashima H, Nishikawa K and Ooi T 1986 The folding type of a protein is relevant to the amino acid composition; *J. Biochem.* **99** 153–162

Natesh R, Bhanumoorthy P, Vithayathil P J, Sekar K, Ramakumar S and Viswamitra M A 1999 Crystal structure at 1·8 Å resolution and proposed amino acid sequence of a thermostable xylanase from *Thermoascus aurantiacus*; *J. Mol. Biol.* **288** 999–1012

Raghavan S, Hariharan R and Brahmachari S K 2000 Polypurine polypyrimidine sequences in complete bacterial genomes: preference for polypurines in protein-coding regions; *Gene* **242** 275–283

Schneider G 1999 How many potentially secreted proteins are contained in a bacterial genome?; *Gene* **237** 113–121

Schneider G and Wrede P 1993 Development of artificial neural filters for pattern recognition in protein sequences; *J. Mol. Evol.* **36** 586–595

Tatusov R L, Galperin M Y, Natale D A and Koonin E V 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution; *Nucleic Acids Res.* **28** 33–36

Tatusov R L, Koonin E V and Lipman D J 1997 A genomic perspective on protein families; *Science* **278** 631–637

Van Heel M 1991 A new family of powerful multivariate statistical sequence analysis techniques; *J. Mol. Biol.* **220** 877–887

Wootton J C 1994 Non globular domains in protein sequences: Automated segmentation using complexity measures; *Comput. Chem.* **18** 269–285