BIOINFORMATICS

Vol. 16 no. 5 2000
Pages 451–457

GeneRAGE: a robust algorithm for sequence clustering and domain detection

Anton J. Enright and Christos A. Ouzounis*

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received on January 27, 2000; revised and accepted on March 13, 2000

Abstract

Motivation: *Efficient, accurate and automatic clustering of large protein sequence datasets, such as complete proteomes, into families, according to sequence similarity. Detection and correction of false positive and negative relationships with subsequent detection and resolution of multi-domain proteins.*

Results: *A new algorithm for the automatic clustering of protein sequence datasets has been developed. This algorithm represents all similarity relationships within the dataset in a binary matrix. Removal of false positives is achieved through subsequent symmetrification of the matrix using a Smith–Waterman dynamic programming alignment algorithm. Detection of multi-domain protein families and further false positive relationships within the symmetrical matrix is achieved through iterative processing of matrix elements with successive rounds of Smith–Waterman dynamic programming alignments. Recursive single-linkage clustering of the corrected matrix allows efficient and accurate family representation for each protein in the dataset. Initial clusters containing multi-domain families, are split into their constituent clusters using the information obtained by the multi-domain detection step. This algorithm can hence quickly and accurately cluster large protein datasets into families. Problems due to the presence of multi-domain proteins are minimized, allowing more precise clustering information to be obtained automatically.*

Availability: *GeneRAGE (version 1.0) executable binaries for most platforms may be obtained from the authors on request. The system is available to academic users free of charge under license.*

Contact: *ouzounis@ebi.ac.uk*

Introduction

The enormous growth of public sequence databases and continuing addition of fully sequenced genomes has created many challenging problems in the field of bioinformatics. Large scale protein sequence comparison is increasingly becoming an effective way to extract useful biological information from genome sequences. Due to the increasing sizes of protein databases, methods of this kind need to be as accurate, efficient and automatic as possible. Accurate prediction of protein function necessitates the identification of ‘paralogous’ proteins within genomes and ‘orthologous’ proteins between genomes. Clustering algorithms are designed to take sequence databases and assign each protein to a family. Some measure of similarity is needed to assign homologous proteins to families.

Many methods exist that can cluster sequences according to similarity (or distance) information. However, most of these methods are either too computationally intensive or require too much manual input to be usable for large databases, such as genomic protein databases. Many such techniques do not take into account the problems associated with clustering databases containing multi-domain proteins. We believe that multi-domain proteins are sufficiently common in genome databases to merit explicit algorithmic treatment.

One of the most widely used methods to detect similarities between proteins for clustering purposes is the detection of homologues using single-sequence similarity search algorithms such as BLAST (Altschul *et al.*, 1997), FASTA (Pearson and Lipman, 1988) and Smith–Waterman (Smith and Waterman, 1981). These search algorithms can detect homologous groups of proteins in a protein database, by comparing every sequence in the database with every other. This type of all-against-all analysis forms the basis for many of the available clustering applications.

Other methods to determine similarity relationships between proteins involve multiple alignment procedures. These methods can also produce excellent information for clustering purposes, but tend to be significantly more CPU intensive than single-sequence alignment procedures. Newer breeds of search tools such as PSI-BLAST (Altschul *et al.*, 1997) and HMM-based methods (e.g. SAM-T98; Karplus *et al.*, 1998) have improved the detection of remote homologues. These types of search

*To whom correspondence should be addressed.

tools generally involve the detection of homologous intermediate proteins that provide a common link between remote homologues. However, they are generally slower, due to the iterative implementation of the algorithm. This makes them less desirable to use for the clustering of large datasets. For this type of genomic sequence clustering, we believe that local-alignment search tools are sufficient for the production of accurate clustering information.

Results from sequence comparison programs can be used to create a list of all hits with their associated similarity scores and statistical values. Clustering approaches can then be employed to create clusters from these data.

Finally, a number of similarity detection approaches are based not on primary sequence, but on comparison of protein structures in three-dimensional (3D) fold-space (Holm and Sander, 1993). The detection of 3D structure similarities may imply some remote functional relationships. It has been shown that certain proteins sharing as little as 15% sequence identity can have almost identical 3D structures (Hubbard *et al.*, 1999). These approaches can effectively detect remote homologues that lie in the 'twilight-zone' of sequence homology. Sophisticated methods for pair-wise comparisons of protein structures have been developed and can be used in the generation of protein fold clusters (Holm and Sander, 1996).

This approach is limited by the relatively sparse amount of 3D structural data for known proteins. Accurate structure prediction could be of enormous value for the clustering of gene-products in a newly sequenced genome (for which there will probably be no 3D information; Jones, 1999).

Clustering techniques

Clusters are built by associating each protein sequence with a list of detected neighbouring sequences, where the distance measure is a score, or probability value, obtained from the initial sequence comparisons. Sequence space in this way may be represented as a graph whose vertices represent the sequences. Vertices may be linked by weighted directional edges. The edges are weighted to represent the degree of similarity between two vertices as determined from the initial similarity searches. Strongly connected sets of vertices should then represent clusters of related proteins in sequence space (Tatusov *et al.*, 1997).

The thresholds for defining an edge in this graph (e.g. a BLAST E-value cut-off) control how conservative the clusters will be. A lower expectation value means less relationships will be detected, but noise will be reduced to a minimum (high precision, low recall). A higher (more permissive) expectation value allows more relationships to be detected; however, this increases the risk of detecting false relationships (high recall, low precision). It is essential that the correct balance between conservative and permissive association be reached.

Remote homologues

The clustering of proteins into families involves two fundamental issues. The first of these issues is how to generate clusters given that not all hits will be detected by the similarity search stage. Remote homologues may form a separate cluster that cannot be linked to a sister cluster of similar proteins, due to the inability of the search algorithms to find a remote homology relationship. It is sometimes possible to link these two clusters together if conserved residue information is taken into account. This is a similar approach to sequence comparison tools such as PSI-BLAST and SAM-T98 (Altschul *et al.*, 1997; Karplus *et al.*, 1998, respectively).

This procedure can enhance the detection of remote homologues by as much as 70% in the case of the PDB40D database (Park *et al.*, 1997), with an error rate of approximately 1%. However, because this database decomposes proteins into domains, the multi-domain problem is not addressed.

Multi-domain proteins

The second issue in sequence clustering involves multi-domain proteins. These proteins contain two or more separate domains that may be similar to different sets of unrelated clusters (Figure 1). If such a protein cannot be determined to be a multi-domain protein then the two clusters will be artefactually linked. In this case the multi-domain protein is technically a member of both sets of clusters; however, the two clusters may not be related.

In order for clustering to work effectively, the presence of multi-domain proteins must be detected and explicitly recorded. Once detected, clusters containing these proteins can be broken down into separate clusters.

System and methods

The GeneRAGE algorithm is written in ANSI C, and developed on a Sun Ultra 10 Workstation. The code has been ported to the following operating systems: Solaris, Compaq Tru64 UNIX, SGI IRIX, AIX and Linux. An SMP parallel implementation of the code is also available (based on the POSIX Pthread standard), and has been tested in Linux, SGI IRIX and Compaq Tru64 multiprocessor environments. The minimum hardware requirements for the clustering of small genomes ($<$ 6000 proteins) is 32MB RAM and sufficient disk space to store the sequence database and search results.

Algorithm

Initial steps

An 'all-against-all' sequence similarity search is undertaken to determine significant similarity relationships within a query database of size n proteins. In this analysis the BLASTp package was used to determine similarity relationships between proteins, below a specified E-value cut-off (1×10^{-10}). All query sequences were filtered using the CAST algorithm (Promponas *et al.* submitted) prior to searching, to mask compositionally biased regions in these proteins. The filtering of sequences using the CAST algorithm reduces noise in the sequence similarity search, and makes E-values more reliable for sequence clustering.

BLAST v2.0 was chosen for this analysis, as it is relatively fast and sufficiently accurate to provide a solid basis for genome sequence clustering. It is possible, however, to use other search tools for this step, if more distant homology relationships are required. A bit-wise matrix (T) of size $n \times n$ elements is constructed from these pair-wise similarity relationships. Each bit in the matrix represents either the presence ('1') or absence ('0') of significant similarity between any two proteins in the database.

Symmetrification of the matrix

To facilitate clustering, the first step used was the symmetrification of the similarity matrix T. This condition has been previously used in sequence comparison (Rivera *et al.*, 1998). We implemented this condition as follows: for every element of the matrix $T_{i,j}$ check:

$$\textbf{If :} \ \forall T_{i,j} \ \textbf{:} \qquad T_{i,j} = T_{j,i}$$

$\Rightarrow$ **Then:** Skip; **else**

$$\textbf{If :} \ \forall T_{i,j} \ \textbf{:} \qquad T_{i,j} \neq T_{j,i}$$

$\Rightarrow$ **Then:** A Smith–Waterman dynamic programming alignment is used to determine if this is a false positive/false negative assignment (Pearson, 1996). If a significant Z-score (e.g. $Z > 10$), obtained with a further 100 rounds of randomized alignments, is detected between proteins i and j, then the matrix is corrected by setting:

$$T_{i,j} = T_{j,i} = 1.$$

This situation represents a false negative case at the search step that is rectified at the symmetrification step.
Otherwise, if no significant similarity is detected (e.g. $Z \leq 10$) the matrix is corrected by setting:

$$T_{i,j} = T_{j,i} = 0.$$

This situation represents a false positive case at the search step that is eliminated at this step.
When this procedure is complete, matrix T satisfies the symmetrical properties of a sequence similarity matrix:

$$\textbf{Now :} \qquad \forall T_{i,j} \ \textbf{:} \ T_{i,j} = T_{j,i}.$$

Detection of multi-domain proteins

The detection of multi-domain proteins from the query database is important to allow accurate clustering of the matrix. Multi-domain proteins are detected by the following simple, yet effective protocol. If two proteins a and b hit a common protein c, does protein a hit protein b? In other words, does the transitivity criterion hold?

$$\textbf{If :} \qquad T_{a,c} = T_{c,a} = 1$$

and

$$\textbf{If :} \qquad T_{b,c} = T_{c,b} = 1$$

$\Rightarrow$ **Then:** check

$$\textbf{If :} \qquad T_{a,b} = T_{b,a} = 1$$

If this is not the case then c may be a multi-domain protein (Figure 2). The algorithm works as follows.

- For each protein c in matrix T, collect the set S_c of all proteins that exhibit significant similarity to protein c, from matrix T.

- For every pair of proteins (a, b) in set S_c, look up matrix T to check if any similarity exists between them.

  $$\textbf{If :} \qquad T_{a,b} = T_{b,a} = 1$$

  $\Rightarrow$ **Then:** Skip; **else**

  $$\textbf{If :} \qquad T_{a,b} = T_{b,a} = 0$$

  Confirm that no significant similarity exists by performing an additional Smith–Waterman dynamic programming alignment between a and b. If significant sequence similarity is detected (e.g. $Z > 10$), this is a false negative case that is corrected by setting $T_{a,b} = T_{b,a} = 1$. In this case $T_{a,c} = T_{c,a} = 1$ and $T_{b,c} = T_{c,b} = 1$ already holds, therefore a, b and c belong to the same family. If no significant similarity is detected (e.g. $Z \leq 10$), mark protein c as a candidate multi-domain protein, composed of two domains a', b' with similarity to a and b respectively.

Another important application for this technique is the detection of fusion proteins across genomes (Enright *et al.*, 1999). In these cases, proteins a and b represent component proteins in one genome and protein c represents a multi-domain composite protein in another genome with two domains a' and b' similar to a and b respectively. The fusion detection algorithm (called DifFuse) is a variant of GeneRAGE, where the second dynamic programming test is performed between entries of two databases (Enright *et al.*, 1999).

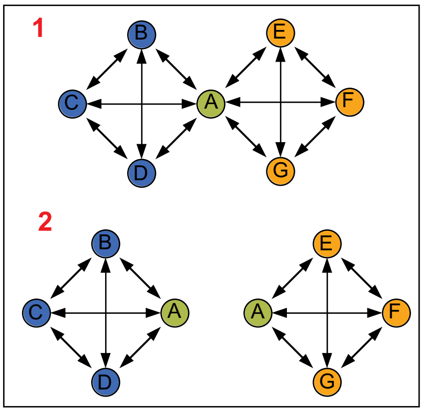

Fig. 1. The multi-domain problem in sequence clustering. Graphs representing sequence relationships in a hypothetical case for a multi-domain protein family. Vertices represent sequences and edges represent homology relationships between these sequences. Color indicates membership into a specific family. In case (1), if sequence A is not detected as a multi-domain protein, then all sequences A–G may be considered as members of a single family by virtue of their relationship to the intermediate sequence A. In case (2), sequence A has been succesfully detected as a multi-domain protein, and two distinct fully connected clusters have been detected and protein A has been re-assigned on the basis of its domain structure.

Single linkage clustering

The processed matrix is recursively clustered by beginning a clustering operation for each row of the matrix. If a protein corresponding to this row i of matrix T is not already clustered, then a new cluster is created containing sequence i. New sequences are added to this cluster by processing across row i of the matrix and recursively sub-clustering each protein that is hit by protein i. As the clustering procedure descends through each row of the matrix, more and more proteins are added to each cluster.

At this stage, multi-domain proteins are clustered separately from single-domain families. When the initial clustering operation is complete, multi-domain family information from the second step of the algorithm is used to split clusters. Clusters that are deemed to contain two separate families linked by one or more multi-domain proteins are split into their constituent families. Multi-domain proteins can hence be members of more than one cluster. Finally, all clustering information (including multi-domain information) is represented in a clusters table. Additionally, for genome comparison studies, a similarity table is created for further analysis (Figure 2).

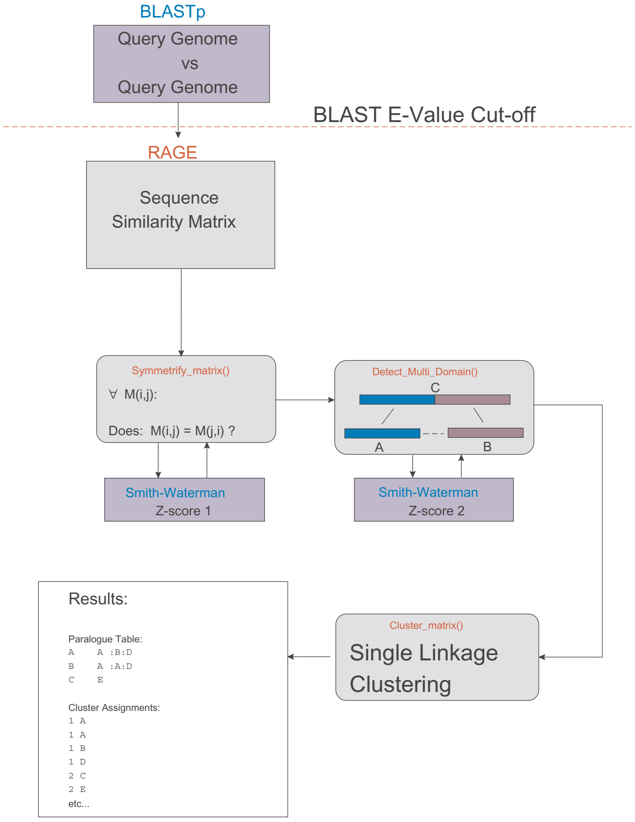

Fig. 2. Schematic representation (flowchart) of the algorithm. All $n \times n$ similarities within the query database detected using BLAST are stored in matrix T. For all non-symmetrical hits, a Smith–Waterman comparison is used to resolve false hits. The multi-domain detection algorithm identifies cases of the form depicted in the inset, where proteins a and b exhibit similarity to protein c but not to each other, by checking matrix T (which is further confirmed by an additional Smith–Waterman comparison). Both Smith–Waterman runs are executed an additional 100 times, with randomization of the sequences, and a Z-score is obtained: if the Z-score is higher than a threshold, the similarity is accepted as significant. A recursive single-linkage clustering operation is then performed on the corrected matrix T, to obtain a similarity table and a table of cluster assignments.

Results

To evaluate the performance of the algorithm, the complete genome of the archaeal methanogen *Methanococcus jannaschii* was analysed using GeneRAGE. This particular genome was chosen because of our extensive experience with this organism and the availability of updated manual annotations for all genes. The initial genome self-comparison using BLASTp (Altschul *et al.*, 1997) took approximately 20 min running in parallel on four workstations (using htBLAST, A. Enright, unpublished). Symmetrification, multi-domain detection and clustering for all 1771 proteins, took 8 min on a two-processor

R10000 SGI Octane workstation. The analysis was conducted using a BLASTp E-value of 1×10^{-6} using the CAST filter (Promponas *et al.*, submitted). Z-score cut-off values of 10 (symmetrification) and 7 (multi-domain detection) were used by GeneRAGE. These values are arbitrarily set by the user—we have set the above mentioned values based on experimentation and empirical observations.

Of the original 3391 hits obtained by BLASTp runs, 1026 were considered to be false positives and were removed, while 889 were considered as false negatives and were added, during the symmetrification step. The total number of hits after the processing of the similarity matrix was 3254.

Most proteins in *M.jannaschii* (69%) have no paralogues in the genome and hence clustered as individual sequences. Other clusters of varying sizes were formed from related sequences within the genome. The distribution of these clusters is illustrated in Figure 3. Multi-domain proteins detected within the genome were clustered by their individual domains. Clusters containing more than three members were analysed to check their validity (61 clusters). This analysis was performed by taking each protein in a cluster and examining its corresponding annotation from the *M.jannaschii* functions database (Kyrpides *et al.*, 1996). In addition, multiple alignments (Thompson *et al.*, 1994) were created to assist in the evaluation of generated clusters if needed. Of these 61 clusters, 95% (58/61) had manual annotations that were consistent with that cluster. The other three clusters had consistent, high quality alignments but conflicting annotations. These cases may represent incorrect annotations, functionally diverse families or false positive cases.

Multi-domain proteins detected within the context of this dataset proved to be consistent with the manual annotations (Kyrpides *et al.*, 1996) and further multiple alignment analysis (Thompson *et al.*, 1994). One example is the archaeal ATPase proteins (Koonin, 1997), which were successfully resolved into three distinct domains (not shown). Further examples of successful multi-domain detection include ABC transporter proteins, hydrogenases and dehydrogenases (not shown). Even domains as short as CBS (Bateman, 1997) or TPR (Kyrpides and Woese, 1998) were detected and assigned to consistent clusters within the *M.jannaschii* genome.

To further examine and validate the multi-domain detection algorithm, a complex test set containing multi-domain proteins was used. This set contains multi-domain relationships that are not present in the *M.jannaschii* genome described above. The test set consisted of 13 genes/proteins for aromatic amino acid biosynthesis (*aro* operon) from four different genomes. These proteins consist of single domains, that have fused together in some genomes, yet remain as separate proteins in other genomes (Duncan *et al.*, 1987). Figure 4 illustrates the arrangement of these proteins in four different genomes. Conventional clustering techniques may fail to cluster individual proteins from each genome into the correct clusters. We are not aware of any sequence clustering technique that can perform the above task automatically at this level of precision. GeneRAGE successfully detected the presence of multi-domain proteins in this test set and divided clusters accordingly (Figure 5). Each cluster generated hence represented a single functional unit. The same result can be reproducibly obtained from the corresponding complete genome sequences.

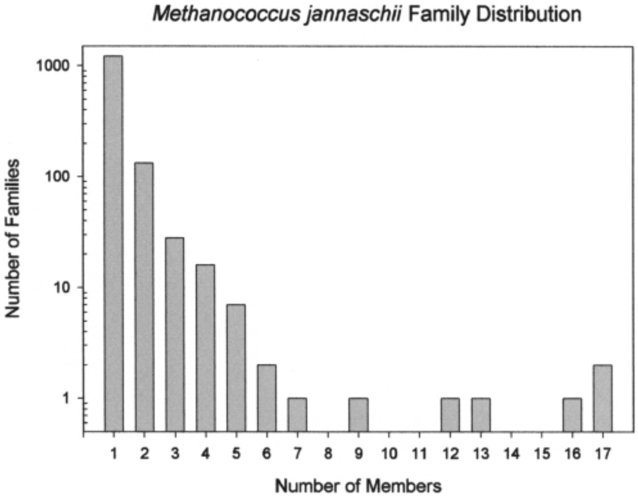

Fig. 3. The distribution of the family size in the genome of *Methanococcus jannaschii*. The family size is shown on the x-axis (linear) and the number of families on the y-axis (logarithmic). The majority of the genes are single-copy genes while the largest families contain up to 17 members.

The algorithm has also been used for the discovery of novel protein families in archaeal genomes. To this end, all complete archaeal genomes (four species) were clustered into families using GeneRAGE. For each family, all members were compared against the Pfam database (Bateman *et al.*, 2000), using BLASTp. Families whose members had no significant homology to any known family in the Pfam database were submitted, processed and curated by the Pfam project members. Of these, 294 families were subsequently made available in the Pfam database release 5 (Bateman *et al.*, 2000).

Discussion

We have described a fast and efficient method for clustering protein sequences according to similarity. The algorithm has been designed for the clustering of protein sequences within and between complete genomes. We believe, however, that the algorithm also has wide ranging

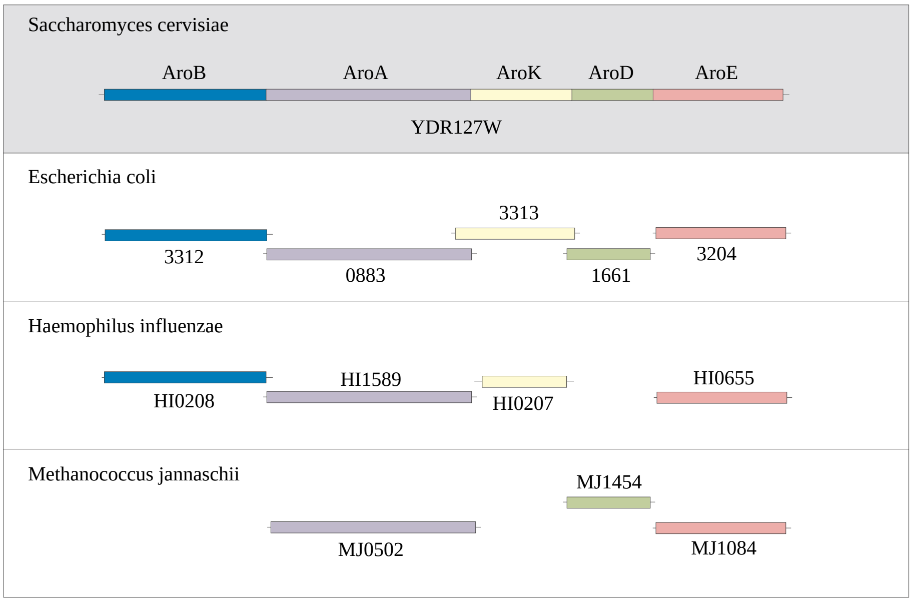

Fig. 4. Pictorial representation of the *aro* cluster in a number of species. In *Saccharomyces cerevisiae*, the gene YDR127W encodes a multi-functional protein (Duncan *et al.*, 1987), while in the three other species, the corresponding enzymes are encoded by different, not necessarily proximal, genes (gene identifiers shown). The equivalent genes are color-coded. Certain genes appear to be absent from *Haemophilus influenzae* or *M.jannaschii*.

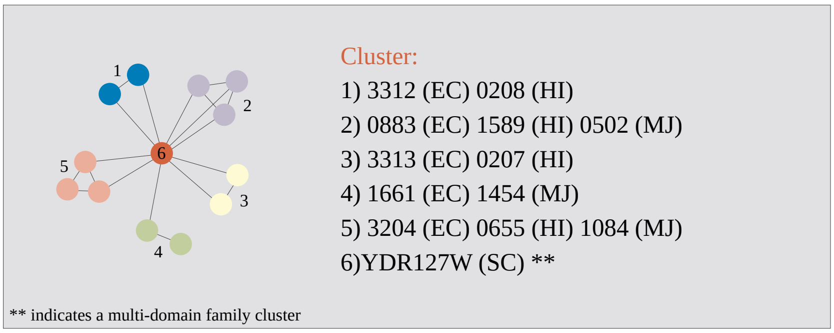
1
2
3
4
5
6
Cluster:
1) 3312 (EC) 0208 (HI)
2) 0883 (EC) 1589 (HI) 0502 (MJ)
3) 3313 (EC) 0207 (HI)
4) 1661 (EC) 1454 (MJ)
5) 3204 (EC) 0655 (HI) 1084 (MJ)
6)YDR127W (SC) **
** indicates a multi-domain family cluster

Fig. 5. The result of the automatic clustering and domain detection for the thirteen *aro* genes shown in Figure 4. GeneRAGE correctly detects homologous genes, assigns them into clusters and defines the relationships within the set. Sequence identifiers as in Figure 4.

uses in the clustering of protein sequences in general. The key abstractions made by the algorithm are the representation and symmetrification of sequence similarity information in a matrix, and the subsequent detection of multi-domain proteins (Figure 2). The symmetrification step is an important abstraction, as it not only detects false positive/false negative relationships, but also provides a consistent similarity construct for further processing and analysis. The storage of similarity information as binary relationships in the matrix makes the algorithm more efficient and less memory intensive. This allows the analysis of much larger sequence datasets (>50 000 sequences). Given the fully automatic implementation of this code and its precision, we believe that the algorithm represents a significant improvement over currently available clustering techniques.

The multi-domain detection step, although based on a simple abstraction (Figure 2) is an important advance. Because multi-domain proteins are detected due to inconsistencies in the similarity matrix, this step not only detects multi-domain proteins, but also detects and corrects further false negative relationships. The precision of this step has been demonstrated in many cases. The detection of multi-domain proteins is, however, highly dependent on the cut-off scores specified. Multi-domain proteins containing two domains that are non-similar are relatively easy to detect. The difficulty lies in the detection of multi-domain proteins that contain two or more very similar domains. We are currently investigating ways of dynamically modifying the cut-off values used in the multi-domain detection step, allowing more efficient detection of these cases.

The algorithm is sufficiently general that various methods can be used at each step. For instance, any similarity search algorithm can be used for the initial step of homology detection. For the clustering step, different methods can also be used, such as complete and average linkage clustering. Further investigation will involve comparison with unsupervised machine learning algorithms for the clustering of the corrected similarity matrix.

Acknowledgements

We would like to thank our colleagues Alex Bateman (Sanger Centre), Mark Carroll and Steven Searle (EMBL-EBI), Nikos Kyrpides (Integrated Genomics, Inc., Chicago IL, USA) and Eamonn O'Toole (Compaq Computer Corporation, HPTC Group, Galway, Ireland) for technical advice and various discussions.

This work was fully supported by the European Molecular Biology Laboratory. C.O. acknowledges further support from the European Commission and IBM Research. Patent application filed on behalf of EMBL.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bateman,A. (1997) The structure of a domain common to archaebacteria and the homocystinuria disease protein. *Trends Biochem. Sci.*, **22**, 12–13.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Duncan,K., Edwards,R.M. and Coggins,J.R. (1987) The pentafunctional arom enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains. *Biochem. J.*, **246**, 375–386.

Enright,A.J., Iliopoulos,I., Kyrpides,N. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Hubbard,T.J., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **27**, 254–256.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Koonin,E.V. (1997) Evidence for a family of archaeal ATPases [letter; comment]. *Science*, **275**, 1489–1490.

Kyrpides,N.C., Olsen,G.J., Klenk,H.P., White,O. and Woese,C.R. (1996) *Methanococcus jannaschii* genome: revisited. *Microb. Comp. Genomics*, **1**, 329–338.

Kyrpides,N.C. and Woese,C.R. (1998) Tetratrico-peptide-repeat proteins in the archaeon *Methanococcus jannaschii*. *Trends Biochem. Sci.*, **23**, 245–247.

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol*, **266**, 227–258.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rivera,M.C., Jain,R., Moore,J.E. and Lake,J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA*, **95**, 6239–6244.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

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

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.