 Nucleic Acids Research, 2001, Vol. 29, No. 1 33–36

CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins

Evgenia V. Kriventseva*, Wolfgang Fleischmann, Evgeni M. Zdobnov and Rolf Apweiler

EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received August 28, 2000; Revised and Accepted October 17, 2000

ABSTRACT

The CluSTr (Clusters of SWISS-PROT and TrEMBL proteins) database offers an automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences. Analysis has been carried out for different levels of protein similarity, yielding a hierarchical organisation of clusters. The database provides links to InterPro, which integrates information on protein families, domains and functional sites from PROSITE, PRINTS, Pfam and ProDom. Links to the InterPro graphical interface allow users to see at a glance whether proteins from the cluster share particular functional sites. CluSTr also provides cross-references to HSSP and PDB. The database is available for querying and browsing at http://www.ebi.ac.uk/clustr.

INTRODUCTION

With the rapid growth of protein sequence databases, there is an increasing need for automatic sequence analysis procedures. One approach is to pre-process a protein database into sets of homologous proteins (i.e. proteins that have evolved from the same ancestor) and use derived information for further analysis.

The CluSTr database, the database of Clusters of SWISS-PROT and TrEMBL (1) proteins, is built on the basis of sequence similarity. CluSTr can be used for: prediction of functions of individual proteins or protein sets; automatic annotation of newly sequenced proteins (2); removal of redundancy from protein databases (3); searching for new protein families; proteome analysis (4); and provision of data for phylogenetic analysis.

METHODS AND ALGORITHMS

The clustering approach is based on two steps. First, a similarity matrix of 'all-against-all' protein sequences is built. The similarity matrix is computed using the Smith–Waterman algorithm (5). A Monte-Carlo simulation, resulting in a *Z*-score (6) is used to estimate the statistical significance of similarity between potentially related proteins. That is, we calculate a raw Smith–Waterman score between sequences A and B and if this score is higher than a certain threshold we compare the sequence A with N shuffled sequences of B (B*). Sequences B* have the same length and amino acid composition as the initial sequence B.

$$Z(A,B) = (SW(A,B)–M)/\sigma$$

Where: $SW(A,B)$ is the raw Smith–Waterman score, M is the average Smith–Waterman score between sequence A and sequences B* and σ is the standard deviation.

Next sequence B is compared with N shuffled sequences A* and $Z(B,A)$ is calculated. The final Z-score is, Z-score = $\min(Z(A,B),Z(B,A))$. The Z-score obtained depends only on the sequences compared, not on the size and composition of the sequence database. This allows us to update the CluSTr database incrementally by keeping all scores of unchanged sequences and only calculating 'new-against-new' and 'new-against-unchanged' which avoids time-consuming recalculations.

Secondly, clusters are built using a single linkage algorithm for different levels of protein similarity. There are two main complications in the automatic clustering procedures: different protein families have different levels of sequence similarity and the clusters of proteins with different domains get pulled together by multidomain proteins. One of the approaches to tackle these problems is hierarchical clustering that allows us to work with clusters at different levels of sequence similarity. The LASSAP package (7) is used to calculate similarities and to build clusters.

Clusters for mammalian proteins, plant proteins and the three complete eukaryote genomes (*Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*) have been built. All the data is stored in a relational database and a web interface, via Java servlets, is provided.

STORAGE AND UPDATE PROCEDURE

The CluSTr data is stored in a relational database (Oracle). This allows us to handle large amounts of data and to facilitate comprehensive data updates. Multiple users have direct access to the database via Java servlets.

The main building blocks of the schema are Proteins, Groups, Similarities and Clusters. The Proteins table describes SWISS-PROT+TrEMBL entries, Groups describes protein sets for which clusters were built and the history of comparison runs, Similarities contains the pairwise scores between proteins and the Clusters table represents the information about and relationships between different clusters (Fig. 1).

*To whom correspondence should be addressed. Tel: +44 1223 494 430; Fax: +44 1223 494 468; Email: evgenia.kriventseva@ebi.ac.uk

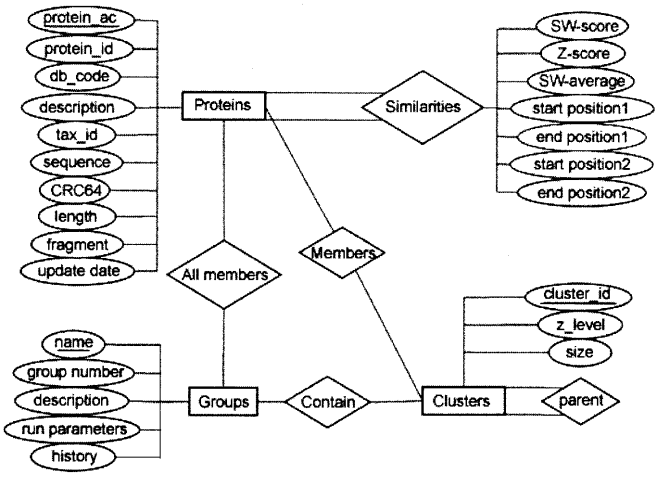

Figure 1. Entity-Relationship diagram for the CluSTr database.

The data update is another big challenge in the design and implementation of the CluSTr database. Our aim is to update CluSTr data incrementally in a synchronised manner with weekly updates of SWISS-PROT+TrEMBL. There are additional Oracle tables to facilitate this. The PROTEIN_NEW table gets populated with new protein data. We check for new, changed and deleted proteins using SWISS-PROT+TrEMBL accession numbers and the circular redundancy checksum (crc64). A list of new and changed proteins is created followed by the calculation of similarities for this set against itself and against unchanged proteins.

WEB INTERFACE

The CluSTr database is available for querying and browsing at http://www.ebi.ac.uk/clustr.

It is possible to query the CluSTr database directly by one or several SWISS-PROT+TrEMBL accession numbers as well as cluster IDs using the so-called 'simple search'. The 'advanced search' allows to query SWISS-PROT+TrEMBL via the SRS (8) 'AllText' datafield, which includes entry accession numbers, entry names, sequence annotation, keywords, taxonomic information and references to other datasources, and retrieves the clusters for the returned proteins. The result of the query is a graphical presentation of corresponding clusters at different levels of protein similarity (Fig. 2). A cluster of interest can be further investigated by clicking on its ID number. For each cluster the list of proteins, their descriptions and domain composition are provided (Fig. 3). The domain composition is defined using InterPro (http://www.ebi.ac.uk/interpro/) (9), a new integrated and annotated resource of protein families, domains and functional sites from PROSITE (10), PRINTS (11), Pfam (12) and ProDom (13). Links to the InterPro graphical view allow users to see at a glance whether proteins from the cluster share particular functional sites.

For each cluster the list of secondary structure cross-references from the Homology derived Secondary Structure of Proteins (HSSP) database (14) is generated dynamically. The database also provides links to the Protein Data Bank (PDB) resource (15). The links to SRS allow users to download selected proteins from a cluster.

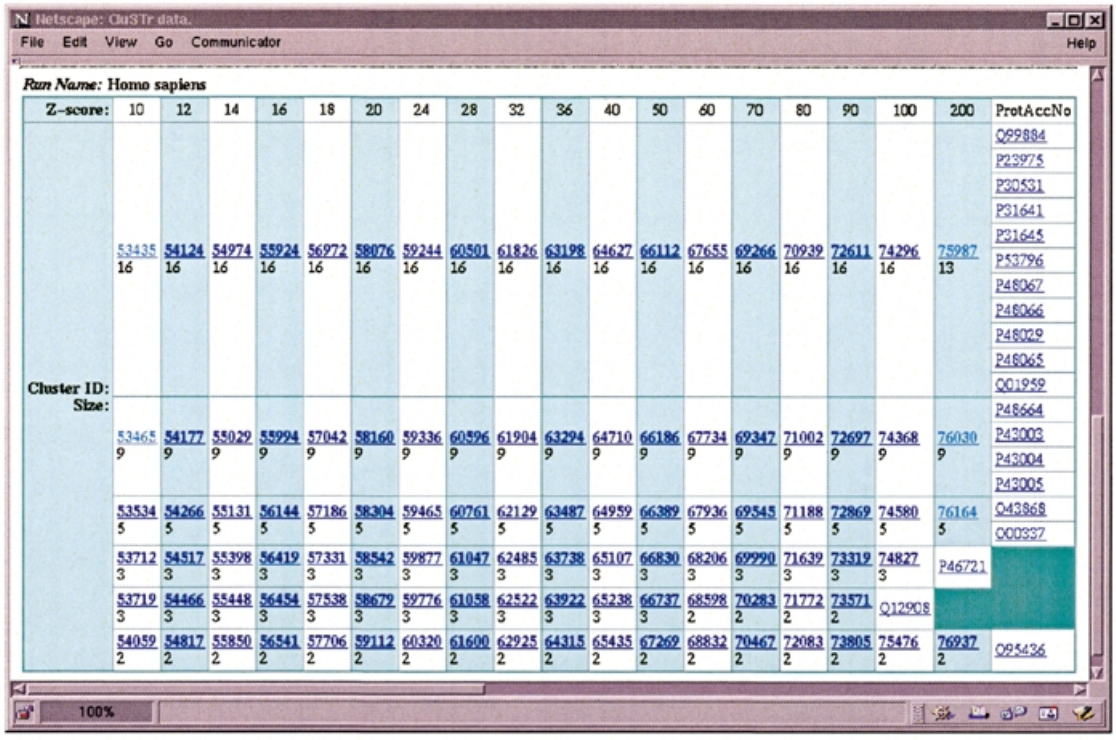

Figure 2. Searching the CluSTr database. Results for a query of 'human sodium transport' proteins. The table contains accesssion numbers of proteins with the words 'human' and 'sodium transport' in their annotation and the corresponding clusters at different *Z*-levels.

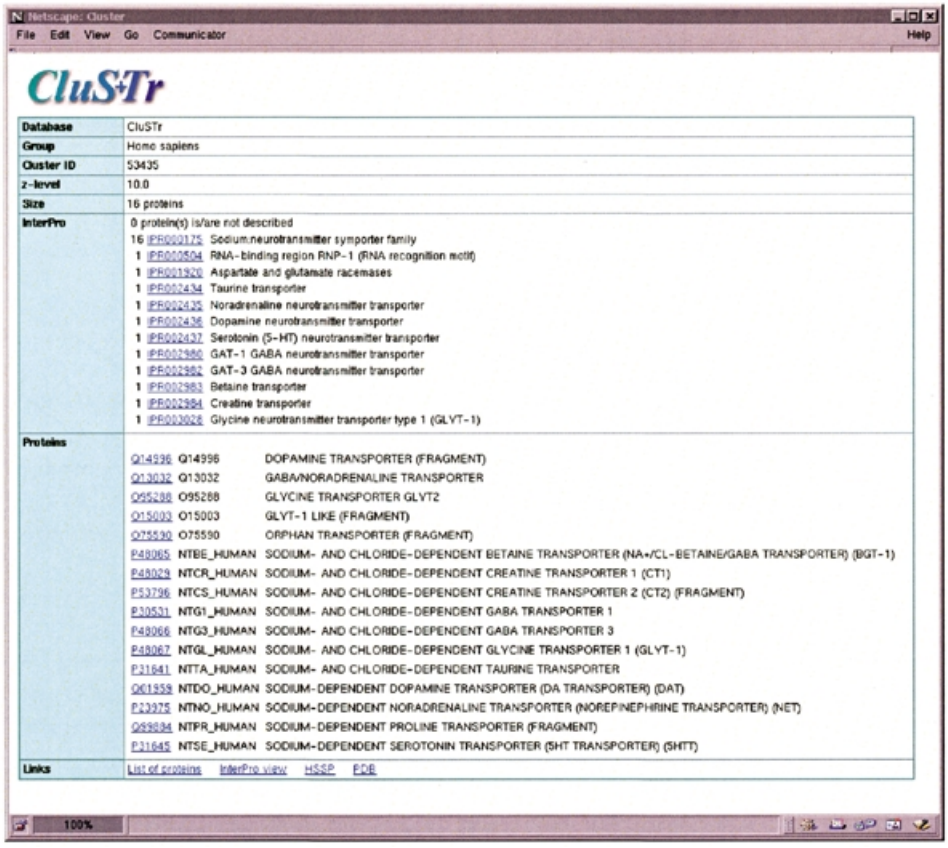

Figure 3. A cluster of the human sodium:neurotransmitter symporter proteins. The presentation contains general information, lists of proteins, their description and InterPro-based domain description of the cluster. At the bottom of the page are links to the InterPro graphical representation and the SRS-generated list of clustered proteins as well as links to the HSSP and PDB databases.

FUTURE PERSPECTIVES

We are going to use the CluSTr database for function prediction and automatic annotation of newly sequenced proteins. By analysing the annotation of related proteins we can also improve the consistency of information in SWISS-PROT+TrEMBL. Furthermore we will use CluSTr to make SWISS-PROT+TrEMBL an even less redundant protein sequence database. Proteins detected to have very close sequences are potential candidates for merging into a single entry. Clusters can also provide data for phylogenetic analysis. Finally, we can compare the domain and family composition of different organisms on the basis of clusters for different genomes.

ACKNOWLEDGEMENTS

We thank Gene-It for technical support. We are also grateful to Beate Marx for administration of the relational database and helpful comments. This work was supported in part by grant B104-CT97-2099 of the European Commission.

REFERENCES

1. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
2. Fleischmann,W., Moeller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
3. O'Donovan,C., Martin,M.J., Glemet,E., Codani,J.J. and Apweiler,R. (1999) Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics*, **15**, 258–259.
4. Apweiler,R., Biswas,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I. and Zdobnov,E. (2001) Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.*, **29**, 44–48.
5. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
6. Comet,J.P., Aude,J.C., Glemet,E., Risler,J.L., Henaut,A., Slonimski,P.P. and Codani,J.J. (1999) Significance of Z-value statistics of Smith–Waterman scores for protein alignments. *Comput. Chem.*, **23**, 317–331.
7. Glemet,E. and Codani,J.J. (1997) LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci.*, **13**, 137–143.
8. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
9. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
10. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
11. Attwood,T.K., Croning,M.D.R., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

12. 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.
13. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
14. Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 214–218.