388–389 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg035


# Improvements to CluSTr: the database of SWISS-PROT + TrEMBL protein clusters

**E. V. Kriventseva, F. Servant* and R. Apweiler**

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

Received September 14, 2002; Accepted September 24, 2002

## ABSTRACT

**The CluSTr database (http://www.ebi.ac.uk/clustr/) offers an automatic classification of SWISS-PROT + TrEMBL proteins into groups of related proteins. The clustering is based on analysis of all pair-wise sequence comparisons between proteins using the Smith–Waterman algorithm. The analysis, carried out on different levels of protein similarity, yields a hierarchical organization of clusters. Information about domain content of the clustered proteins is provided via the InterPro resource. The introduced InterPro 'condensed graphical view' simplifies the visual analysis of represented domain architectures. Integrated applications allow users to visualize and edit multiple alignments and build sequence divergence trees. Links to the relevant structural data in Protein Data Bank (PDB) and Homology derived Secondary Structure of Proteins (HSSP) are also provided.**

## INTRODUCTION

With sequencing projects producing large amounts of data lacking functional characterization, there is an increasing need for automated sequence annotation procedures. We created the CluSTr (Clusters of SWISS-PROT + TrEMBL proteins) database (1), a resource for an automatic classification of SWISS-PROT + TrEMBL (2) proteins into groups of related sequences. The clustering is based on analysis of all pair-wise sequence comparisons between proteins using the Smith–Waterman algorithm (3). A Monte-Carlo simulation, resulting in a *Z*-score (4), is used to estimate the statistical significance of raw Smith–Waterman scores between potentially related proteins. Clustering is carried out at different levels of protein similarity, yielding a hierarchical organization of the protein groups.

## DATA CONTENT AND ACCESS

Currently data for more than 70 completely sequenced proteomes, including the eukaryotic proteomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* and more than 65 prokaryotic ones, is represented in the database. The complete list of available proteomes is provided on the documentation page (http://www.ebi.ac.uk/clustr/documentation.html).

Web access to the data stored in a relational database (ORACLE) is provided using Java servlet technology. The CluSTr database is available for querying and browsing from http://www.ebi.ac.uk/clustr.

## IMPROVED WEB INTERFACE

Two types of search forms are provided: a 'simple search' that queries directly the CluSTr data in Oracle and an 'advanced search' that allows free-text querying via the EBI SRS (Sequence Retrieval System) server (5). The result page contains the description of the requested cluster, a list of the grouped proteins with SWISS-PROT + TrEMBL description and the InterPro (6) based information on domain/family signatures represented in the cluster. A number of links are provided for further data analysis. These links allow to download entries in the clusters of interest, to look at the graphical representation of known functional signatures, to inspect/edit corresponding multiple alignments, to see the resolved structural domains, and to create dynamic SRS links to other biological databases.

SRS allows retrieving information for clustered proteins from other databases using indexed links. The SWISS-PROT + TrEMBL accession numbers stored in the CluSTr database are used to access the corresponding proteins, which SRS allows to download in various formats. The retrieved SWISS-PROT + TrEMBL records could be linked further inside SRS to other databases. For example, on the basis of information from OMIM (7) (a catalogue of human genes and genetic disorders), it is possible to see whether proteins from a cluster are associated with a disease.

The CluSTr interface is enriched by information on the underlining domain architecture via InterPro resource. For each cluster an InterPro section provides the summary of the domain content, showing the homogeneity of a group in terms of represented domains. Visual representation of this data is provided through the InterPro graphical interface. The analysis

*To whom correspondence should be addressed. Tel: +44 1223 49494686; Fax: +44 1223 494468; Email: flo@ebi.ac.uk

of a cluster domain composition is even more apparent with the condensed graphical view, which shows a single representative for clustered proteins with exactly the same domain architecture.

The DisplayFam multiple alignment browser has been recently included to visualize summarized multiple alignments based on sequence divergence trees and consensus sequences. For the users who want to edit multiple alignments of the clustered proteins, we integrated Jalview (8). This tool has many useful colouring schemes, which highlight different features of multiple alignments. For example, it is possible to colour amino acids according to their biochemical properties, to identify most conservative residue columns and to see secondary structure predictions.

Structural information provides an important insight into the understanding of protein functions. For each cluster, the list of secondary structure cross-references to the Homology derived Secondary Structure of Proteins (HSSP) database (9) is generated dynamically. The database also provides links to the Protein Data Bank (PDB) resource (10), the archive of structural data of biological macromolecules.

## UPDATE PROCEDURE

An automated procedure has been developed to update CluSTr data incrementally in a synchronized manner with weekly releases of SWISS-PROT + TrEMBL. Additional ORACLE tables are used to facilitate the update procedure, which identifies new, updated, unchanged and deleted proteins using SWISS-PROT + TrEMBL accession numbers and the circular redundancy checksum (CRC64) of sequences. The list of new and changed proteins is used to calculate the similarity between proteins in this set and against proteins in the unchanged set. Clusters are built for the updated similarity matrix on different levels of *Z*-scores using a single linkage algorithm.

An automatic procedure for tracing cluster identifiers between releases was developed based on the MatchDom algorithm (11). The algorithm transfers the cluster identifiers from a reference cluster set to a target set, looking for cluster overlaps between two successive releases of CluSTr. The matching clusters are sorted and the best overlapping cluster inherits the identifier of a reference cluster. The other clusters get new identifiers. The procedure allows maintaining stable cluster identifiers corresponding to particular protein families.

## APPLICATIONS

Apart from its use as an interactive web resourse, CluSTr has been applied to the analysis of complete proteomes http://www.ebi.ac.uk/proteome (12). The CluSTr section at the Proteome Analysis database provides a general description of homologous protein groups in completely sequenced genomes as well as a list of candidates with novel sequence domains. The developed methodology has also been used for the in depth study of protein families (13).

## CONCLUSIONS

CluSTr is an evolving resource. New developments include an improved web interface, an automated update procedure and better coverage of SWISS-PROT + TrEMBL data, which will allow us to use the resource in the TrEMBL automatic annotation routine, known as a RuleBase (14).

## ACKNOWLEDGEMENTS

We thank the EBI external service group headed by Rodrigo Lopez for help with the SRS server and web pages. We express gratitude to Gene-It for technical support. We are also thankful to Wolfgang Fleischmann for helpful comments. This work was supported in part by grant B104-CT97-2099 of the European Commission.

## REFERENCES

1. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) CluSTr: a database of clusters of SWISS-PROT + TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.
2. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
3. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
4. 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.
5. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.
6. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
7. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
8. Clamp,M. (1998) Jalview. http://www2.ebi.ac.uk/~michele/jalview/.
9. Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.
10. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
11. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, 246–251.
12. Apweiler,R., Biswas,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E.V., Mittard,V., Mulder,N., Phan,I. *et al.* (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.
13. Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
14. Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.