WWW access to the SYSTERS protein sequence cluster set

A. Krause¹, P. Nicodème¹, E. Bornberg-Bauer¹,², M. Rehmsmeier¹ and M. Vingron¹

¹Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, INF 280, D-69120 Heidelberg, Germany

Abstract
Summary: We present a Web server where the SYSTERS cluster set of the non-redundant protein database consisting of sequences from SWISS-PROT and PIR is being made available for querying and browsing. The cluster set can be searched with a new sequence using the SSMAL search tool. Additionally, a multiple alignment is generated for each cluster and annotated with domain information from the Pfam protein family database.

Availability: The server address is http://www.dkfz-heidelberg.de/tbi/services/cluster/systersform

Contact: {A.Krause|M.Vingron}@DKFZ-Heidelberg.de

The increasing size of protein sequence databases makes the grouping of the sequences into families useful for studying and understanding their functionality. Such a clustering can help in compressing the output produced by database search programs, can aid in the automatic derivation of multiple alignments and profiles, and can provide data for evolutionary analysis.

We present a Web server which allows access to the SYSTERS cluster set. The clustering is based on our searching algorithm called SYSTEmatic ReSearching (Krause and Vingron, 1998). SYSTERS iterates traditional search routines like BLAST (Altschul et al., 1990, 1997) or FASTA (Pearson and Lipman, 1988) to find for a given query sequence a set of related ones (called a cluster). Every newly found sequence in a search above a given cut-off (e.g. a P value of 10⁻³⁰) is added to the cluster and used for a re-searching step. The procedure is iterated as long as a new sequence is found.

The clustering procedure first generates a set of SYSTERS clusters by searching a database with each of its sequences. Then a set-theoretic clustering method uses this highly redundant cluster set to partition a large fraction of the sequence database into disjoint clusters. Only a small part of the database is assigned to clusters which overlap each other.

The cluster set underlying the SYSTERS Web server consists of 115 448 sequences from SWISS-PROT (Bairoch and Apweiler, 1998) and PIR (Barker et al., 1998) which are grouped into 39 963 single-sequence clusters (including 13 114 fragmental sequences) and 11 418 clusters with at least two sequences.

To guarantee fast, secure and flexible access to the cluster set, all data are stored in a PostgreSQL (http://www.postgresql.org) database. PostgreSQL is a freely available object-relational database management system, originally developed at the University of California at Berkeley (Stonebraker and Rowe, 1986). All searchable attributes are indexed using a B-tree index to improve access time.

Clusters can be selected by cluster number, size, organism, sequence accession number, identifier or by searching the complete sequence annotations for keywords. The keyword search is carried out with the indexing and query scheme GLIMPSE (GLobal IMPlicit SEarch; Manber and Wu, 1994).

The sequences in every cluster have been multiply aligned using ClustalW 1.7 (Thompson et al., 1994) and with each cluster an unrooted tree (computed using Neighbor-Joining; Saitou and Nei, 1987) is available.

All multiple alignments are annotated with known domains from the Pfam protein family database of alignments and HMMs (Sonnhammer et al., 1998). Each domain annotation in a multiple alignment is linked to the list of clusters containing this domain. Vice versa, clusters can be selected directly from a list of Pfam domains. Additionally, a new sequence can be used as a query against the cluster set. The search is performed by the similarity searching tool SSMAL (Shuffling Similarities with Multiple Alignments) (Nicodème, 1998) that uses features of the BLAST algorithm for scanning a database of multiple alignments. In contrast to profile searches, this approach is based on alignment graphs built on a distinction between well-conserved and weakly conserved regions in the multiple alignment. The run time is comparable.
Fig. 1. Overview of the SYSTERS Web server. Clusters of sequences are presented with, for example, the first three description lines of their annotation (top right). The annotation can be searched using GLIMPSE and PostgreSQL. SSMAL is used for sequence similarity searching versus the aligned clusters. The alignments are annotated with domain information from Pfam.

We thank Heiko Schmidt for running large amounts of database searches on a parallel machine. We acknowledge financial support from BMBF (Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie).

References