ECCB 2002

The SYSTERS Protein Family Webserver

A. Krause, T. Meinel, J. Stoye, H. A. Schmidt, H. Luz, and M. Vingron

Max Planck Institute for Molecular Genetics
Dept. Computational Molecular Biology
Ihnestr. 73, D-14195 Berlin, Germany

`systers.molgen.mpg.de`

With the overwhelming growth of biological sequence databases comes the question of how to effectively handle these amounts of data. Protein sequences constitute one such data type for which the databases have grown to an impressive size.

A protein family contains evolutionarily related sequences. Generally, this is reflected by **sequence similarity**. Therefore, one aims at organizing the set of all protein sequences into family clusters based on their sequence similarity.

Clustering a large set of sequences as opposed to dealing only with the individual sequences offers several advantages. A frequent problem is the identification of sequences that are similar to a new query sequence. This task can be executed much faster when **only one comparison to an entire cluster** has to be performed rather than one comparison per database sequence.

Another important application lies in the possibility of analyzing **evolutionary relationships** among the sequences in a cluster and the species they come from. Additionally, a clustered protein sequence database can be used for selecting candidates for protein structure analysis.

SYSTERS [1] is a method for grouping protein sequences hierarchically into **superfamily** and **family clusters**. The classification is based on an **all-against-all database search** using gapped BLAST [4]. The **graph-based algorithms** take into account the topology of the sequence space induced by the data itself.

We have applied our algorithms to a set of 395,089 non-redundant sequences from the Swiss-Prot [6], TrEMBL [6], and PIR [8] databases. The data splits into 64,282 superfamilies, which are further divided into 82,450 family clusters with an overall number of 55,182 single sequence clusters.

So far our hierarchy consists of two layers representing protein superfamilies and families. For the third layer located at the domain level we currently rely on the **Pfam** domain database [5].

In the SYSTERS web server, information of the original data set are recorded as well as cross-references to the databases concerning protein structure (**PDB**, **IMB**), nucleotide sequence (**EMBL**), protein function (**ENZYME**), and protein domains (**PROSITE**).

The sequences in every family cluster have been **multiply aligned** using ClustalW [7], and for each cluster an unrooted phylogenetic **tree** is available. For each family cluster a MView [3] output is generated, and a majority **consensus sequence** is calculated from the resulting partial multiple alignment.

The SYSTERS consensus sequences and/or the original sequences build a searchable database. The result of a **BLAST search** is visualised as a sequence alignment.

SYSTERS protein families can be selected by the **sequence accession number** of the **original** as well as the **cross-linked** databases, any **keyword**, a **Pfam domain**, or a **taxon** (based on the NCBI taxonomy [9]). The taxonomic selection cannot only be entered on the species level but also at any other taxonomic rank.

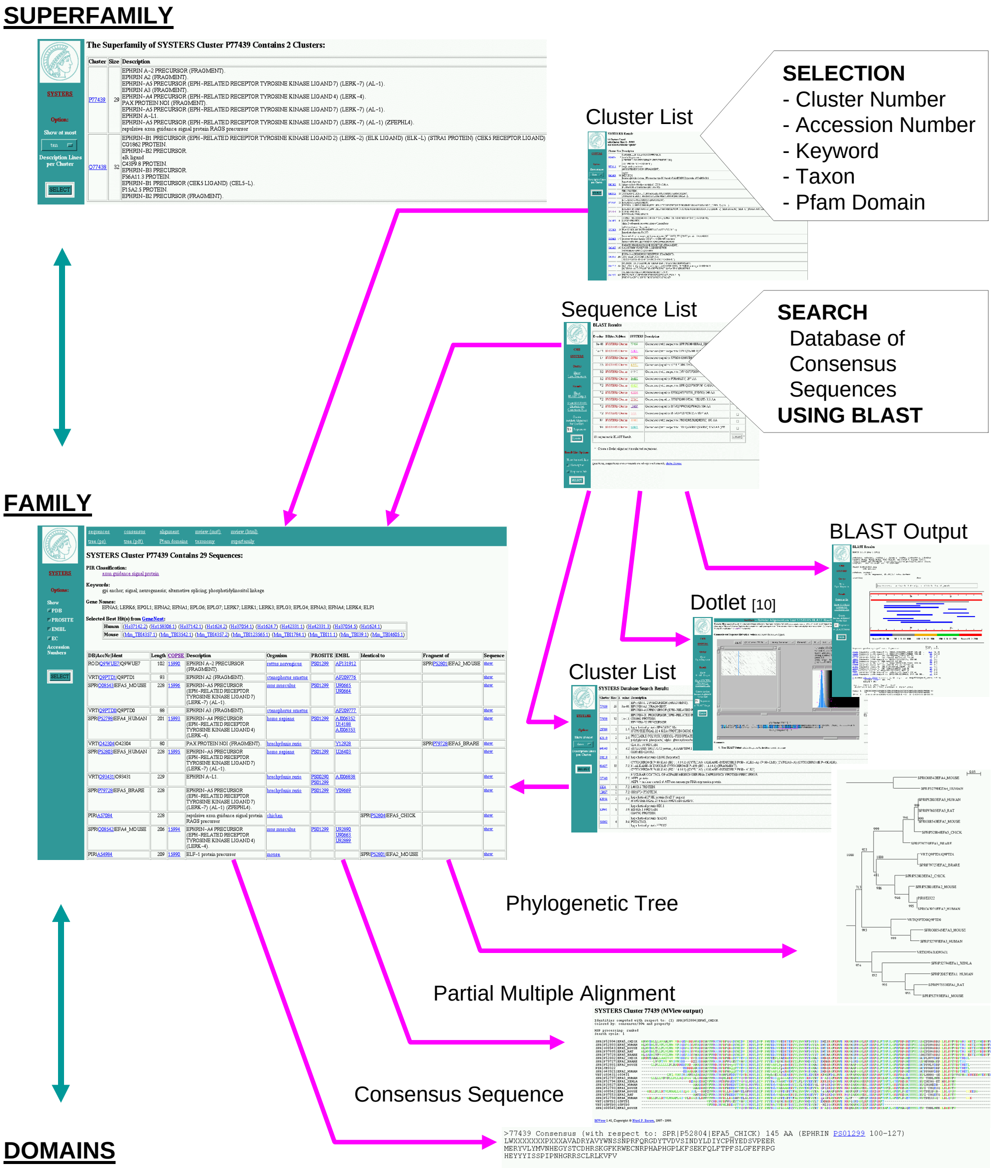
SYSTERS is integrated into a database framework of mRNA/EST consensus sequences, GeneNest [2],

`http://genenest.molgen.mpg.de`

and genomic DNA, SpliceNest,

`http://splicenest.molgen.mpg.de`

Links from SYSTERS to GeneNest and vice versa permit an **over-all exploration** of the whole sequence space.



SUPERFAMILY

The Superfamily of SYSTERS Cluster P77439 Contains 2 Clusters:

Cluster	Size	Description
P77439	29	EPHRIN A-2 PRECURSOR (FRAGMENT). EPHRIN A2 (FRAGMENT). EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1). EPHRIN A3 (FRAGMENT). EPHRIN-A4 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 4) (LERK-4). PAX PROTEIN NOI (FRAGMENT). EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1). EPHRIN A-L1. EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1) (ZFEPHL4). repulsive axon guidance signal protein RAGS precursor
O77438	32	EPHRIN-B1 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 2) (LERK-2) (ELK LIGAND) (ELK-L) (STRA1 PROTEIN) (CEK5 RECEPTOR LIGAND) CG1862 PROTEIN. EPHRIN-B2 PRECURSOR. elk ligand C43F9.8 PROTEIN. EPHRIN-B3 PRECURSOR. F56A11.3 PROTEIN. EPHRIN-B1 PRECURSOR (CEK5 LIGAND) (CEL5-L). F15A2.5 PROTEIN. EPHRIN-B2 PRECURSOR (FRAGMENT).

Cluster List

SELECTION
- Cluster Number
- Accession Number
- Keyword
- Taxon
- Pfam Domain

Sequence List

SEARCH
Database of Consensus Sequences
USING BLAST

BLAST Output

Dotlet [10]

Cluster List

FAMILY

SYSTERS Cluster P77439 Contains 29 Sequences:

PIR Classification:
axon guidance signal protein

Keywords:
gpi anchor; signal; neurogenesis; alternative splicing; phosphatidylinositol linkage

Gene Names:
EFNA5; LERK6; EPGL1; EFNA2; EFNA1; EPLG6; EPLG7; LERK7; LERK1; LERK3; EPLG3; EPLG4; EFNA3; EFNA4; LERK4; ELF1

Selected Best Hit(s) from GeneNest:

Human	(Hs37142.2) (Hs158306.1) (Hs37142.1) (Hs1624.2) (Hs37054.1) (Hs1624.7) (Hs42331.1) (Hs42331.3) (Hs37054.5) (Hs1624.1)
Mouse	(Mm_TBI4357.1) (Mm_TBI3542.1) (Mm_TBI4357.2) (Mm_TBI123565.1) (Mm_TBI1784.1) (Mm_TBI11.1) (Mm_TBI39.1) (Mm_TBI4605.1)

DB|AccNr|Ident	Length	COPSE	Description	Organism	PROSITE	EMBL	Identical to	Fragment of	Sequence
ROD|Q9WUE7|Q9WUE7	102	15990	EPHRIN A-2 PRECURSOR (FRAGMENT).	rattus norvegicus	PS01299	AF131912		SPR|P52801|EFA2_MOUSE	show
VRT|Q9PTD1|Q9PTD1	93		EPHRIN A2 (FRAGMENT).	ctenophorus ornatus		AF209776			show
SPR|O08543|EFA5_MOUSE	228	15996	EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1).	mus musculus	PS01299	U90665 U90664			show
VRT|Q9PTD0|Q9PTD0	88		EPHRIN A3 (FRAGMENT).	ctenophorus ornatus		AF209777			show
SPR|P52798|EFA4_HUMAN	201	15993	EPHRIN-A4 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 4) (LERK-4).	homo sapiens	PS01299	AJ006352 U14188 AJ006353			show
VRT|O42304|O42304	80		PAX PROTEIN NOI (FRAGMENT).	brachydanio rerio		Y12928		SPR|P79728|EFA5_BRARE	show
SPR|P52803|EFA5_HUMAN	228	15995	EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1).	homo sapiens	PS01299	U26403			show
VRT|O93431|O93431	229		EPHRIN A-L1.	brachydanio rerio	PS00290 PS01299	AJ006838			show
SPR|P79728|EFA5_BRARE	228		EPHRIN-A5 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 7) (LERK-7) (AL-1) (ZFEPHL4).	brachydanio rerio	PS01299	Y09669			show
PIR|A57084	228		repulsive axon guidance signal protein RAGS precursor	chicken			SPR|P52804|EFA5_CHICK		show
SPR|O08542|EFA4_MOUSE	206	15994	EPHRIN-A4 PRECURSOR (EPH-RELATED RECEPTOR TYROSINE KINASE LIGAND 4) (LERK-4).	mus musculus	PS01299	U92890 U90663 U92889			show
PIR|A54984	209	15990	ELF-1 protein precursor	mouse			SPR|P52801|EFA2_MOUSE		show

Phylogenetic Tree

Partial Multiple Alignment

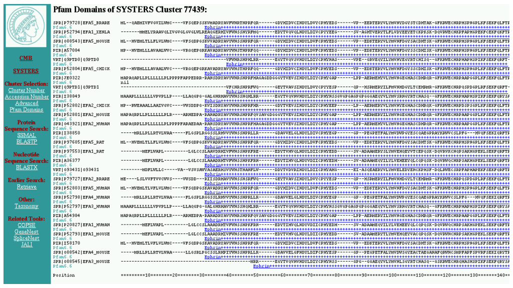
SYSTERS Cluster 77439 (MView output)

MView 1.41, Copyright © Nigel P. Brown, 1997-1999.

Consensus Sequence

```
>77439 Consensus (with respect to: SPR|P52804|EFA5_CHICK) 145 AA (EPHRIN PS01299 100-127)
LWXXXXXXXPXXXAVADRYAVYWNSSNPRFQRGDYTVDVSINDYLDIYCPHYEDSVPEER
MERYVLYMVNHEGYSTCDHRSKGFKRWECNRPHAPHGPLKFSEKFQLFTPFSLGFEFRPG
HEYYYISSPIPNHGRRSCLRLKVFV
```

DOMAINS



Pfam Domains of SYSTERS Cluster 77439:

[1] A. Krause, S.A. Haas, E. Coward, and M. Vingron. SYSTERS, GeneNest, SpliceNest: Exploring Sequence Space from Genome to Protein. Nucleic Acids Research, 30 (1):299-300, 2002

[2] S.A. Haas, T. Beißbarth, E. Rivals, A. Krause, and M. Vingron. GeneNest: automated generation and visualization of gene indices. Trends in Genetics, 16 (11):521-523, 2000

[3] N.P. Brown, C. Leroy, and C. Sander. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics, 14 (4): 380-381, 1998

[4] S.F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25 (17):3389-3402, 1997

[5] A. Bateman et al. The Pfam Protein Families Database. Nucleic Acids Research, 30 (1):276-280, 2002

[6] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research, 28 (1):45-48, 2000

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

[8] C.H. Wu et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Research, 30 (1):35-37, 2002

[9] David L. Wheeler et al. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Research, 30 (1):13-16, 2002

[10] T. Junier and M. Pagni. Dotlet: diagonal plots in a Web browser. Bioinformatics, 16 (2):178-179, 2000