# PairsDB protein alignment database

**Kimmo Mattila**

CSC - Scientific Computing Ltd., Finland

## Introduction

Sequence similarity searching with the BLAST algorithm is a cornerstone of molecular biology.

The new PairsDB service provides access to pre-calculated BLAST and PSI-BLAST based alignments for a comprehensive set of protein sequences. The service allows you to explore protein sequences and their similarity relationships quickly and easily. The web interface for the PairsDB-service can be found in:

`http://pairsdb.csc.fi`

## Structure of the PairsDB database

PairsDB is based on a non-redundant set of protein sequences and their hierarchical clustering. The sequences of PairsDB are collected from UniProt, PDB, RefSeq and ENSEMBL databases. Identical sequences are merged into a single entry (in PairsDB sequences are considered identical only if they have the same length and 100 % sequence identity). This first pruning of the source data produces a sequence set non-identical protein sequences called NRDB100 (Non Redundant sequence DataBase).

Next a sequence set containing less than 90% identical sequences, the NRDB90 set, is created. This pruning step is done with CD-HIT program. CD-HIT sorts all sequences by their lengths in decreasing order. Starting from the longest sequence, the procedure removes all sequences from the set that align over their full length and are more than 90% identical to the selected sequence. The procedure then takes the second longest sequence and does the same. The procedure continues, until all sequences have been processed. Because of the high similarity threshold most alignments need not be calculated explicitly, but instead a fast tuple lookup algorithm is sufficient. As a result the NRDB100 sequences are clustered into sequence families that contain a long representative sequence and group of shorter family members that are more than 90 % identical compared to the representative sequence. These representative sequences form the NRDB90 sequence set

For the NRDB90 set a BLASTP analysis is run in an all-against-all fashion. The results from this massive BLAST analysis step are stored into a relational database. Using these BLAST results non-redundant databases are created also for 80%, 70%, 60%, 50%, 40%, and 30% sequence identity. As a final step an all-against-all PSI-BLAST analysis is run using the NRDB40 sequence set.

When data is retrieved from the PairsDB database, this hierarchical sequence classification and pre-calculated alignments are used to construct a set of similar sequences and their alignments. For single query sequence the NRDB90 family and its representative sequence is first checked from the database. Also the alignment between the query and the representative sequences is retrieved. Using the pre-calculated BLAST results, other NRDB90 level sequences and their family members can then be collected.

## PairsDB WWW interface

### Finding name for your sequence

PairsDB interface is operated using the UniProt, PDB or ENSEMBL sequence names like CYC_HUMAN or 1J3S-A ( this refers to the A-chain of PDB entry 1J3S). If you do not know the name of your sequence you can use the "*Sequence Space Filter*" to check it. Sequence space filter is found in the top bar of the PairsDB interface. With this search tool you can try to find the sequence name by searching the sequence descriptions finding sequences that match 100% to your query sequence or a fragment of it. Often already a fragment of 10-20 amino acids is enough to identify your sequence. If the sequence is not found, the reason may be that it was not yet in


Figure 1. The BLAST query interface of PairsDB service.

the public databases when the last PairsDB data set was collected.

Sequence Space Filter can also be used to collect sequence data sets using combination of several search criteria. For example you could easily collect all sequences that are from an organism and contain a given InterPro domain.

## BLAST and PSI-BLAST based searches

PairsDB provides two ways to look for similar sequences for your query sequence. BLAST in nrdb90 level and PSI-BLAST in NRDB40 level. Both of them use the same logic to construct the sequence relationships from the database. Here we discuss only about the BLAST search interface but the same features exist also in the PSI-BLAST interface.

The BLAST search interface can be opened from the BLAST link in the top bar of the interface. There are two ways to do the search. The more simple "*BLAST results in NRDB90*" collects the NRDB90 level BLAST results for the query sequence or its NRDB90 level representative. Remember that you should feed the name of the sequence to the "*Query sequence*" filed, not the actual query sequence.

The second search option, "*BLAST results expanded to NRDB100*" retrieves also the family members of the NRDB90 level hit sequences. This search checks first the NRDB90 representative sequence for the given query sequence. BLAST hits for the representative sequence are then collected at the NRDB90 level. After this the hit list is expanded to NRDB100 level so that also those sequence neighborhood members that have overlapping match region with the query sequence are selected. The list of hit sequences can be filtered using following features:

- e-value (can vary between 1 - 0)
- fragments, hypothetical or transmembrane proteins
- source database (UniProt, PDB, RefSeq or ENSEMBL) or certain NRDB hierarchy level
- domains from InterPro, SCOP, CATH or ADDA domain databases. For InterPro and ADDA standard database identifiers are used. For SCOP and CATH domains PairsDB uses coding system, that can be checked from help pages of PairsDB
- any taxonomy ID number

The data retrieval can be started with the "Search" button. Typically retrieving and filtering the data takes 5 -15s.

## BLAST Results

The BLAST results page starts with information about the query sequence and the corresponding representative sequence in NRDB90 level. Detailed information about the query or representative sequence can be obtained using the links in Shortcuts column. Note that the actual BLAST results are computed using the NRDB90 representative sequence, not the original query. Except in those cases, where the query is also the NRDB90 representative.

## Match Overview

The Match Overview table lists the found BLAST/PSI-BLAST hits. The first column displays the location of the matching region between the hit and representative sequence. The original query sequence id is represented by a red bar and its NRDB90 representative sequence as a green bar. The matching sequences that originate from NRDB90 are shown as dark yellow bars while the corresponding NRDB100 level family members are presented as light yellow bars. Using the shortcuts (I,B,P) you can directly go to the sequence

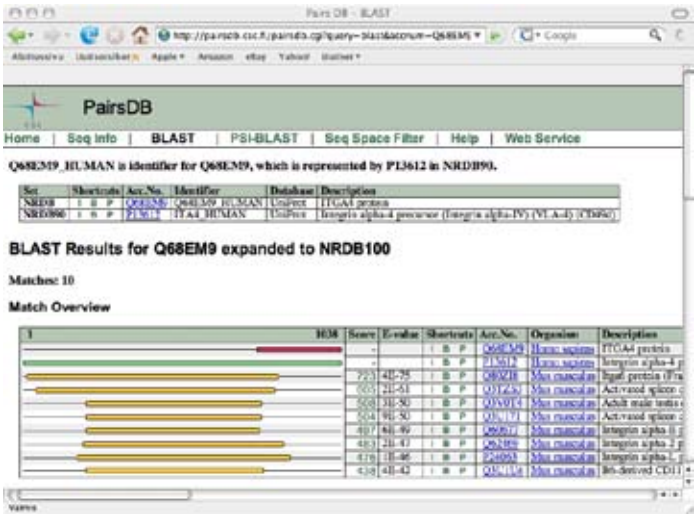


Figure 2. BLAST result page of PairsDB service

info, BLAST or PSI-BLAST page of any of these sequences.

Note also that one hit in NRDB100 level can represent several entries in the source databases. Thus if the result list seems to lack a UniProt entry or PDB structure that should be there, it may be presented by some other sequence name. E. g. UniProt entries CYC _ GORGO, CYC _ HUMAN and the A chain of PDB entry 1J3S have identical sequences so they are presented by only one hit, in this case named as 1J3S-A.

## Stacked Multiple Alignment

The stacked multiple alignment shows those regions of the hit sequences that align with query sequence. The density of the colour refers to how well conserved a specific amino acid is in the alignment. In the stacked alignment the hit sequence regions that do not align with the query sequence, are not shown. Thus the query-anchored stacked alignment is NOT a multiple sequence alignment.

## Pairwise Alignments

This section displays the pairwise alignments between the query and hit sequences. The score and E-values refer to the values of the NRDB90 level BLAST hits thus they are not exactly correct values.

Using the section options in the BLAST query page you can also choose to show the stacked multiple sequence alignment or hit sequence list in FASTA format.

## Benefits of PairsDB

PairsDB is very useful tool when you need to do several slightly modified BLAST searches. E. g., when you want to get familiar with a protein sequence with which you have not yet worked before you right want to quickly see if there is some structural data available for the query sequence or are there known homologues in certain taxonomic group? All this you could of course do with normal BLAST too, however this would be much slower. One normal BLAST search may take several minutes and more specific queries may require laborious filtering of BLAST results or construction of users own BLAST databases. PairsDB produces essentially the same results in few seconds.

If the WWW interface of PairsDB does not suite all your needs, you can utilize the Web Service interface of the service or install the PairsDB to your local MySQL server. The files needed for building a local installation of PairsDB can be found at `ftp://ftp.funet.fi/pub/sci/molbio/pairsdb`.

## Acknowledgment

PairsDB was developed by Prof. Liisa Holm and Dr. Andreas Heger, and it is maintained jointly with CSC.