PairsDB protein alignment atlas - interface and database tables

Kimmo Mattila

CSC – IT Center for Science, Espoo, Finland

On EMBnet news vol. 13 (4) [1] we presented the PairsDB protein sequence alignment database [2] that contains directly computed or hierarchically inferred pairwise alignments for all known protein sequences. Since 2007 several small modifications have been made to the database and the WWW interface (http://pairsdb.csc.fi/). The database itself has been updated twice. The latest release (1/2009) is based on the protein sequences collected from protein databases in February 2009.

In this article we provide an updated description about the interface and also discuss the general features of the most essential SQL database tables of PairsDB. The data collected to the PairsDB database is freely available and provides a unique resource for studying the currently known protein universe with the methods of bioinformatics and data mining.

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 and RefSeq 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).

In the second pruning step the NRDB100 sequences are clustered with CD-HIT program. 90% identity is used as the threshold level for the clustering. As a result the NRDB100 sequences are sorted 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. The representative sequences form the NRDB90 sequence set.

For the NRDB90 set a BLAST analysis is run in an all-against-all fashion. Using the BLAST results non-redundant sequence sets are created for 80%, 70%, 60%, 50%, 40%, and 30% sequence identity levels. 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 BLAST or PSI-BLAST 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 promptly collected.

Below are some key figures from the 1/2009 version of PairsDB. This data gives an overview of both the size of PairsDB and also of the currently known protein universe.

- The total number of protein sequences collected from source databases (UniProt, PDB and RefSeq) **13,4 million**.
- Number of unique sequences (NRDB100) was **7,3 million**. Of these, 36% were found only once in the source databases. About 24% of the unique sequences were found to exist in more than one organism.
- Number of sequence families that are less than 90% identical to other sequences (NRDB90) is **4,4 million**. 80% of these families contain only one sequence.
- Number of sequence families that are less than 40% identical to other sequences (NRDB40) is **2,3 million**. 69% of these families contain only one sequence.
- Number of BLAST matches within the NRDB90 sequence set: **9428 million**
- Number of PSI-BLAST matches within the NRDB40 sequence set: **5003 million**

Finding name for your sequence

PairsDB interface is operated using the UniProt, RefSeq or PDB sequence names like CYC_HUMAN, NP_061820.1 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 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 a certain taxonomic group 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. To start the search, define the "Query sequence" and press "Search" button Remember that you should feed the name of

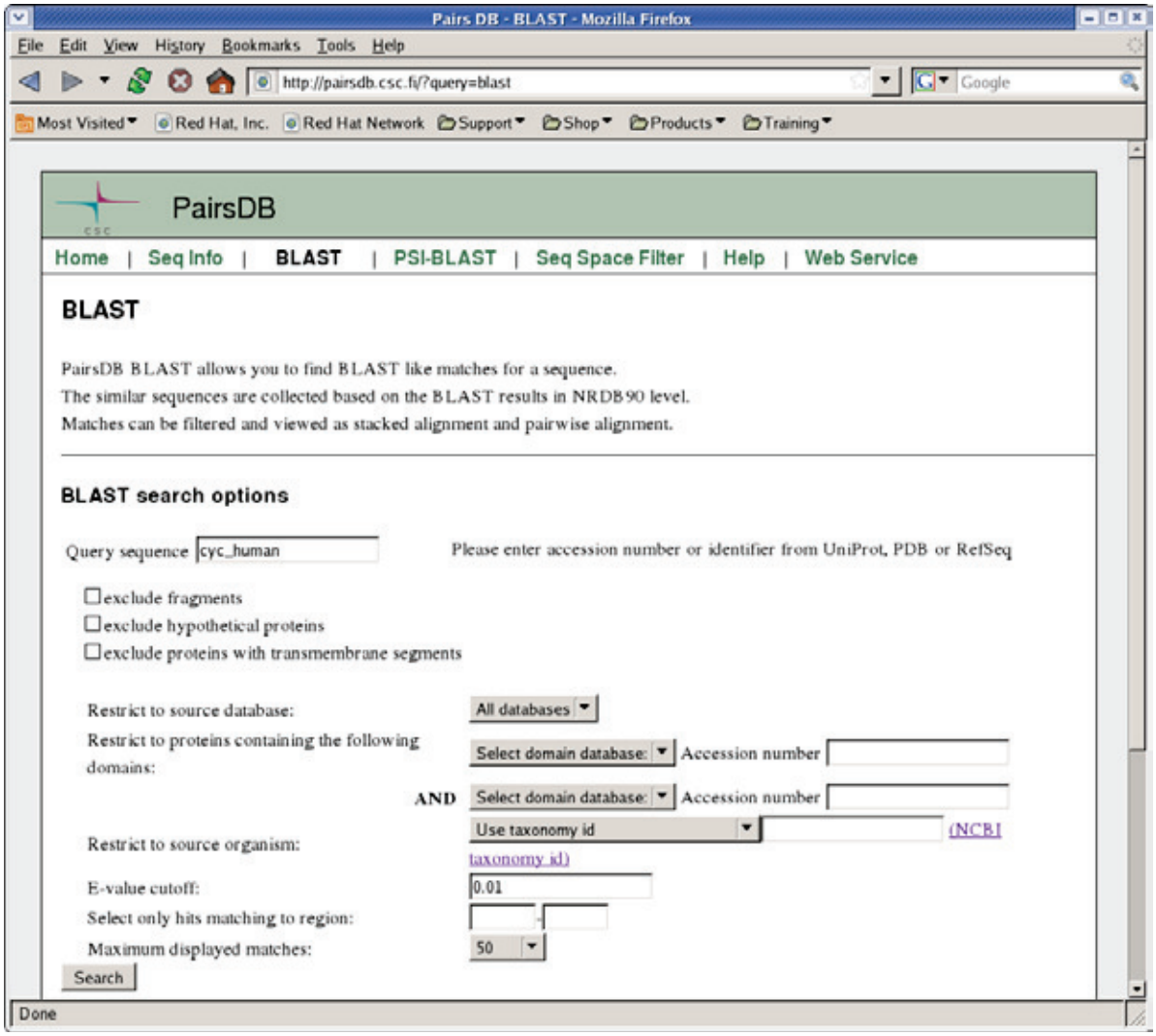

Figure 1. The BLAST query interface of PairsDB.

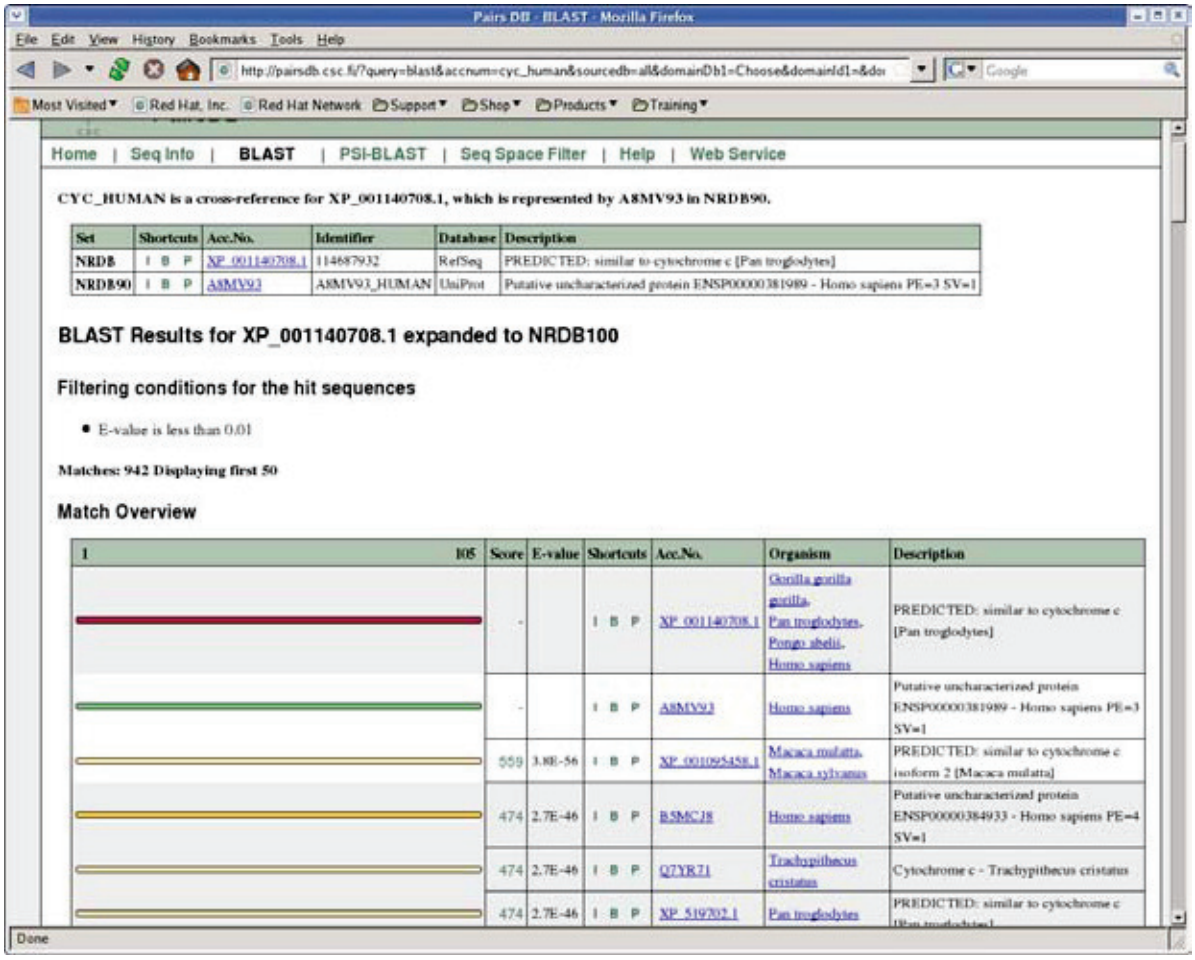

Figure 2. BLAST result page of PairsDB.

the sequence to the "Query sequence " filed, not the actual sequence data.

As a first search step the NRDB90 level representative sequence for the given query sequence is retrieved. Then BLAST hits for the representative sequence are collected at the NRDB90 level. After this the hit list is expanded to NRDB100 level so that only those sequence neighborhood members that have overlapping match region with the query sequence are selected to the result set.

The hit sequence list can also be filtered using following features:

- e-value (can vary between 1 – 0);
- exclude fragments, hypothetical or transmembrane proteins;
- select only hits that are from certain source database (UniProt, PDB or RefSeq) or that are included on 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;
- Taxonomy ID number;
- subregion of the query sequence.

Retrieving and filtering the data takes 10s to few minutes depending on the size or the result set and the selected filtering methods and output formats. The number of hits to be reported is by default limited to 50 but can be expanded up to 10,000.

BLAST Results

The BLAST results page starts with information about the query sequence and the corresponding representative sequence in NRDB90 level. After that, filtering conditions, used in the query,

are listed. By default the actual results are shown as match overview table, stacked multiple alignment and pairwise alignments.

Match Overview

The Match Overview table lists the found BLAST hits. The first column displays the location of the matching region between the hit and representative sequence. The original query sequence is represented as 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 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. For example 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 this query-anchored stacked alignment is NOT a multiple sequence alignment. Stacked multiple alignments are not shown for query sequences that are longer than 1000 amino acids.

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.

Other output options

You can modify the BLAST results display in the BLAST query page: You can print the hit sequences or stacked multiple alignment in fasta format or switch of some part of the output. Often the most time consuming part of the PairsDB result processing is constructing and downloading the HTML presentation of the stacked and pairwise alignments. Using only Match overview presentation can make PairsDB to act much faster.

PairsDB SQL Tables

The PairsDB www-interface allows an easy way to use the PairsDB database as a handy substitute for BLAST. However the real power of PairsDB can be obtained by using the database directly through SQL queries. CSC does not provide tools that would allow any user to submit free MySQL queries to the database, but the database content is freely available at the FTP site of CSC:

ftp://ftp.funet.fi/pub/sci/molbio/pairsdb/

There are two limiting issues in using the data, however. Firstly, the size of the current PairsDB version is about 1,5 TB. Another drawback is that the database is not well documented.

All together the PairsDB consists of 50 different tables. We present here the most important tables of the system to help potential users to get started. Installing instructions for the PairsDB tables can be found from the README document at the FTP site.

nrdb

The NRDB table is the most central table of the database. It contains all the unique sequences that form the non-redundant data set. For each unique sequence a unique id number: **nid**, is assigned. This nid number is used in all PairsDB tables to identify the sequence. In addition to the nid number the nrdb table contains columns for the actual sequence string, description, sequence length, date and a filter column that describes the position of the sequence in the PairsDB hierarchy. Zero value in the filter column means that the sequence is obsolete and no more in use in the other tables.

Each nid has accession number and identifier values too, however you should note that the accession number, identifier and description, presented in the nrdb table are not necessary the only ones that in the source databases refer to this sequence. The possible other values can be found from the cross _ references table.

cross_references

This table contains information about the names and accession numbers that have been used for

a certain sequence (i.e. nid) in the source databases. So if you would like to know the nid of a sequence you are working with (say RIMM_ECOLI). You could check it with SQL query:

SELECT nid FROM cross_references WHERE identifier="RIMM_ECOLI";

cross _ references table also has column to identify the source database where the accession number was used (1= UniProt, 3 = PDB and 12 = RefSeq), and the sequence description present in the source database.

pairsdb_90x90 and psiblast_40x40

The all-against-all BLAST results for the NRDB90 sequence set and the all-against-all PSI-BLAST results for the NRDB40 sequence set are stored into tables pairsdb_90x90 and psiblast_40x40. These very large tables have identical structures. The two first columns hold the nid numbers of query (query_nid) and hit sequences (sbjct_nid). The e-value is stored to the third column as the logarithm of the actual value (log(e)). The following six columns contain information about the alignment between the two sequences. In addition to the starting, and end residues of the actual structure of alignment is stored too. The alignment structure is stored to query_ali and sbjct_ali columns in a format where +X means X aligning residues and -X X gaps in the alignment. So for example BLAST alignment:

```
Query:     ALES-SAS
           |   | |::
Hit:       A--SESVA
```

Would be stored in following format:

query_ali: +4-1+3
sbjct_ali: +1-2+5

The last two columns of this table contain the score and identity percent of the alignment. As pairsdb_90x90 and psiblast_40x40 tables contain billions of rows, indexing of the columns that will be used in the queries is essential.

pairsdb_100x90 and pairsdb_100x40

If the query sequence does not belong to NRDB90 or NRDB40 sequence sets, one has to be able to check what is the nrdb90/40 level representative sequence for the query sequence and how the sequence aligns with the representative sequence. This information is stored to the pairsdb_100x90 and pairsdb_100x40 tables. The alignment between the reference sequence (rep_nid) and the member sequence (mem_nid) of the sequence family is coded in the same way as in the pairsdb_90x90 and psiblast_40x40 tables.

Acknowledgment

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

References

1. Mattila K (2007) PairsDB protein alignment database. EMBnet news 13 (4):22-24.
2. Heger A, Korpelainen E, Hupponen T, Mattila K, Ollikainen V, Holm L (2008) PairsDB atlas of protein sequence space. Nucleic Acids Res. 36(Database issue):D276-D280.