BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 7 2000
Pages 656–657

NAIL—Network Analysis Interface for Linking HMMER results

Luis Sánchez-Pulido[1,*], Yan P. Yuan[2,3], Miguel A. Andrade[2,3] and Peer Bork[2,3]

[1]Protein Design Group, Centro Nacional de Biotecnología (CNB-CSIC), Campus Univ. Autónoma, Cantoblanco 28049, Madrid, Spain, [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany and [3]Max Delbrück Center for Molecular Medicine, Department of Bioinformatics, PO Box 740238, 13092 Berlin-Buch, Germany

Received on November 15, 1999; revised and accepted on February 11, 2000

Abstract

Summary: *Network Analysis Interface for Linking HMMER results (NAIL) is a web-based tool for the analysis of results from a HMMER protein database-search. NAIL facilitates the selection of protein hits and the creation of an alignment, which can be used for a new sequence similarity search.*
Availability: *From http://www.bork.embl-heidelberg.de/NAIL/*
Contact: *sanchez@cnb.uam.es*

Sequence comparison is a powerful tool for the computational analysis of protein sequences. Even remote similarities may allow the transfer of structural and functional features between homologous sequences. Iterative searches, using profiles of aligned sequences, are much more sensitive than pairwise comparison strategies (Bork and Gibson, 1996; Holm and Sander, 1996a; Neuwald *et al.*, 1997; Park *et al.*, 1997; Rost, 1999). HMMER (Eddy, 1998) provides a method to derive a profile (a Hidden Markov Model) from an alignment. This can be used for searching protein databases for regions in proteins similar to the aligned sequences. HMMER is widely used and the software is freely available (e.g. http://hmmer.wustl.edu/).

The output from a HMMER search is a text file showing:

1. a list of global hits with scores and protein identifiers,
2. a detailed list where multiple hits of the profile to the same sequence are displayed with indications of the position of the hit,
3. the alignments of each hit to the profile.

After extensive use of HMMER we realized that this output could be greatly improved in two respects:

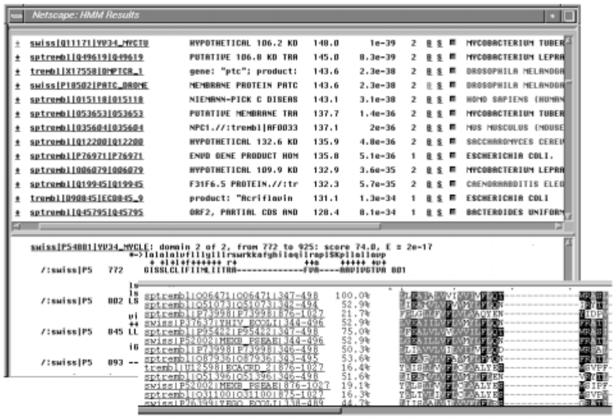
- Data visualization. It is difficult to see the connections between the different data, although they are obviously linked to sequences (e.g. a protein sequence is linked to its hits, which in turn are linked to their alignments to the profile).
- Subsequent use of the results. A typical use of the HMMER output is the construction of an alignment of the hits considered true positives, which can be re-used by HMMER or by other programs. This can be a quite laborious work if the number of hits is large.

In order to alleviate these two bottlenecks, we have set up a web server called Network Analysis Interface for Linking HMMER results (NAIL) which converts HMMER 2.1.1 text output into HTML, including (i) links between the hits, (ii) external links from the hits to databases, (iii) view of aligned hits, and (iv) retrieval of selected hits.

The user provides both the text output of a HMMER search and the HMM model used for the search. The server processes the text output into a HTML output which is split into two frames: the upper one containing global hits and domain hits, and the lower one initially displaying the alignment of each fragment to the profile. In the upper frame, each global hit has a link (indicated with a '+', see Figure 1) to the next domain hit in the same frame. Each domain hit has a '+' link to the next domain hit in the same sequence (or to the global hit if there is none left) and a '?' link which triggers the display of the corresponding alignment in the lower frame. At the right of each hit line there is a 'B' link to the BLAST2 server (Yuan *et al.*, 1997) [for a search of the fragment against several databases using BLAST2, see Altschul *et al.* (1997)] and a 'S' link to the SMART server [for identification of protein domains in the sequence, Schultz *et al.* (2000)], both at EMBL-Heidelberg.

All database identifiers are linked to the corresponding database entry through SRS (Etzold *et al.*, 1996). If the database identifiers correspond to PDB entries, they are

*To whom correspondence should be addressed.



Netscape: HMM Results

```
± swiss|Q11171|YV34_MYCTU       HYPOTHETICAL 106.2 KD   148.0     1e-39  2  B $ ■ MYCOBACTERIUM TUBER
± sptrembl|Q49619|Q49619        PUTATIVE 106.8 KD TRA   145.0   8.3e-39  2  B $ ■ MYCOBACTERIUM LEPRA
± trembl|X17558|DMPTCR_1        gene: "ptc"; product:   143.6   2.3e-38  2  B $ ■ DROSOPHILA MELANOGA
± swiss|P18502|PATC_DROME       MEMBRANE PROTEIN PATC   143.6   2.3e-38  2  B $ ■ DROSOPHILA MELANOGA
± sptrembl|O15118|O15118        NIEMANN-PICK C DISEAS   143.1   3.1e-38  2  B $ ■ HOMO SAPIENS (HUMAN
± sptrembl|O53653|O53653        PUTATIVE MEMBRANE TRA   137.7   1.4e-36  2  B $ ■ MYCOBACTERIUM TUBER
± sptrembl|O35604|O35604        NPC1.//:trembl|AF0033   137.1     2e-36  2  B $ ■ MUS MUSCULUS (MOUSE
± sptrembl|Q12200|Q12200        HYPOTHETICAL 132.6 KD   135.9   4.8e-36  2  B $ ■ SACCHAROMYCES CEREV
± sptrembl|P76971|P76971        ENVD GENE PRODUCT HOM   135.8   5.1e-36  1  B $ ■ ESCHERICHIA COLI.
± sptrembl|O06079|O06079        HYPOTHETICAL 109.9 KD   132.9   3.6e-35  2  B $ ■ MYCOBACTERIUM LEPRA
± sptrembl|Q19945|Q19945        F31F6.5 PROTEIN.//:tr   132.3   5.7e-35  2  B $ ■ CAENORHABDITIS ELEG
± trembl|D90845|ECD845_9        product: "Acriflavin    131.1   1.3e-34  1  B $ ■ ESCHERICHIA COLI
± sptrembl|Q45795|Q45795        ORF2, PARTIAL CDS AND   128.4   8.1e-34  1  B $ ■ BACTEROIDES UNIFORM
```

```
swiss|P54801|YV34_MYCLE: domain 2 of 2, from 772 to 925: score 74.0, E = 2e-17
                *->lalalalvflllylllrswrkkafyhilaqilrapiSKpllallavp
                   + +l+l+f++++++ r+            ++a    +++++ +v+
  /:swiss|P5   772    GISSLCLIFIIMLIITRA-------------FVA----AAVIPGTVA 801
```

```
sptrembl|O06471|O06471|347-498    100.0%
sptrembl|O51073|O51073|342-494     52.9%
sptrembl|P73998|P73998|876-1027    21.7%
swiss|P37637|YHIV_ECOLI|344-496    52.9%
sptrembl|P95422|P95422|347-498     75.0%
swiss|P52002|MEXB_PSEAE|344-496    52.9%
sptrembl|P73998|P73998|346-498     50.3%
sptrembl|O87936|O87936|343-495     53.6%
trembl|U12598|ECACRD_2|876-1027    16.4%
sptrembl|O51396|O51396|346-498     51.6%
swiss|P52002|MEXB_PSEAE|876-1027   19.1%
sptrembl|O31100|O31100|875-1027    16.3%
swiss|P76399|YEGO_ECOLI|338-489    44.7%
```

Fig. 1. Partial overview of the data accessible through NAIL. Background window: the list of hits (top) is linked to the list of alignments (bottom). Each hit has a tick box. A selection of checked hits can be retrieved in FASTA format for further use in other programs. Foreground window: HMMalign is used for making an alignment of the hits. This alignment is displayed using MView (Brown *et al.*, 1998).

linked to the corresponding FSSP file, offering possibilities for taking advantage of structural information (Holm and Sander, 1996b; Brenner *et al.*, 1998).

The species of each hit is displayed. A simple colour coding of the species name (yellow eukaryotes, blue prokaryotes, red archaea, white viruses) offers at a glance a view of the family span over the life kingdoms, which can be helpful to spot spurious hits.

Given a selection of hits, other options are readily available at the click of a button, such as getting sequences in FASTA format (with the possibility of manually changing the hit boundaries) or making an alignment, which is displayed with MView (Brown *et al.*, 1998), and can be obtained in CLUSTAL W (Thompson *et al.*, 1994) format.

In order to get the full functionality of NAIL we recommend the use of the following databases available from EMBL and EBI public ftp servers: SwissProt (Bairoch and Apweiler, 1999, ftp://ftp.EMBL-Heidelberg.de/EMBL-EBI/databases/swissprot/swiss), non-redundant database (Gish, 1992, ftp://ftp.EMBL-Heidelberg.DE/EMBL-EBI/databases/nrdb/nrdb) and a smaller version of the previous non-redundant database where sequences with more than 90% of similarity are joined under the same entry (Holm and Sander, 1998, ftp://ftp.ebi.ac.uk/pub/databases/nrdb90/nrdb90.gz).

The HTML data is deleted periodically but the generation of the web pages is instantaneous. Therefore, we recommend keeping the HMMER output file and using the server rather than linking to the corresponding HTML files.

The NAIL code was written in Perl making extensive use of Steven Brenners Perl cgi libraries (Brenner, 1998).

Acknowledgements

We would like to thank Enrique Morett and Jose Castresana for useful comments and suggestions during the development of NAIL.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Meth. Enzym.*, **266**, 162–184.

Brenner,S.E. (1998) The cgi-lib.pl Home Page. http://cgi-lib.stanford.edu/cgi-lib/.

Brenner,S.E, Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 6073–6078.

Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a Web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763. http://hmmer.wustl.edu/

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods in Enzymology*, **266**, 114–128.

Gish,W. (1992) nrdb program. NCBI/NLM. USA. ftp://ncbi.nlm.nih.gov/pub/nrdb.

Holm,L. and Sander,C. (1996a) Mapping the protein universe. *Science*, **273**, 595–602.

Holm,L. and Sander,C. (1996b) DALI/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res.*, **25**, 1665–1677.

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of distant sequence homologies. *J. Mol. Biol.*, **273**, 349–354.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Prot. Eng.*, **12**, 85–94.

Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: A Web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.

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

Yuan,Y.P., Lai,J. and Bork,P. (1997) The Advanced BLAST2 Search Server—EMBL-Heidelberg. Unpublished. http://www.bork.embl-heidelberg.de/Blast2/.