

Multiple alignment and multiple sequence based searches

Sean R. Eddy

Dept. of Genetics, Washington University School of Medicine

4566 Scott Ave., St. Louis MO 63110 USA

eddy@genetics.wustl.edu

Phone: +1-314-362-7666; FAX: +1-314-362-7855

keywords: multiple sequence alignment, profiles, hidden Markov models

August 7, 1998

Abstract

Multiple sequence alignments reveal patterns of conservation that can be exploited in database searches using “profile” methods. Starting with a single nematode sequence that has no informative BLAST hits, I give a real example of the use of multiple alignment and profile search software to detect informative remote homologies.

It used to be that most new sequences were novel, with no informative similarity to anything in the sequence database. Thanks to genome sequencing projects, things are slightly better now. New sequences are often similar to several uncharacterized sequences, defining whole *families* of novel genes with no informative BLAST or FASTA similarities.

Given a sequence family, though, powerful alternative similarity search methods can be applied. Software packages are available that can take a multiple sequence alignment and build a “profile” of it. Profiles incorporate position-specific scoring information derived from the frequency that a given residue is seen in an aligned column. Because sequence families preferentially conserve certain critical residues and motifs, this information can sometimes allow more sensitive database searches to be done. Most new profile software is based on statistical models called “hidden Markov models” (HMMs).

Here, I show a practical demonstration of a multiple alignment based similarity search. Much more comprehensive reviews of the literature on profile hidden Markov model methods are available elsewhere [1, 2, 3], including two recent books [4, 5].

An example sequence

In the *C. elegans* genome, several large paralogous gene families that were first thought to be nematode specific have since been classified as putative G-protein coupled receptors (GPCRs) [6, 7]. Detecting similarity between these nematode sequences and known GPCRs in other organisms is a nontrivial sequence analysis task. I arbitrarily chose the putative GPCR gene *sra-4* (Wormpep AH6.8; SWISS-PROT SRA4_CAEEL; 329 aa) as an example. The task is to find a significant similarity between AH6.8 and a protein of known function in another organism.

A WWW BLAST search at NCBI [8] using AH6.8 as a query (BLASTP 2.0.4, default options, vs. 319,187 sequences in the NR database on 7/30/98) shows 46 hits with E-values less than 0.01, but all but one of them are to uncharacterized *C. elegans* sequences. The top scoring non-worm hits are a mitochondrial

L11 ribosomal protein (SWISS-PROT RM11_ACACA; E=0.002) and an ornithine decarboxylase (SWISS-PROT DCOR_YEAST; E=0.44). The E-value of the RM11_ACACA is very marginal; I wouldn't trust it by itself. Thus, at first look, AH6.8 is a member of a large but nematode specific gene family.

Much of the analysis that follows requires installing and running software on a local UNIX machine. Basic familiarity with UNIX is essential for bioinformatics. For many labs, the most convenient and inexpensive way to run UNIX is to install the free Linux operating system on a PC. URLs to the software packages I use are given in the accompanying box.

Sequence gathering

The first step for further analysis is to more carefully define a nonredundant set of sequences that belong to the novel family.

The Wormpep 13 database is the authoritative nonredundant source of nematode predicted protein sequences [9]. A WU-BLASTP 2.0a18 search (W. Gish, unpublished; WU-BLASTP is the Washington University version of gapped BLASTP) of Wormpep 13 using AH6.8 as the query pulls out 36 hits with highly significant P-values less than 10^{-6} . Most are about 350 aa long. As a crude protection against erroneous computational gene predictions, I discarded four sequences longer than 500 aa or shorter than 200 aa, leaving 32 sequences. Sometimes, it is necessary to be more careful. Sequences may be related by a shared domain instead of over their entire length, so it may be necessary to isolate alignable subsequences (based on the bounds of BLAST alignments, for instance) before making the multiple alignment. This step can take a fair amount of manual work.

Multiple sequence alignment

The next step is to produce a multiple alignment. I favor the program ClustalW, a very good program that also happens to be free, well supported, capable of dealing with large numbers of sequences, and available for Macintosh, Windows, and various UNIXes [10]. There is also a graphical user interface, ClustalX [11].

Obtaining an acceptable multiple sequence alignment is usually easy, once the family is defined. Starting from a file of 32 sequences in FASTA format called `worm.fa`, I typed at my UNIX workstation's command line:

```
% clustalw worm.fa
```

After a couple of minutes, ClustalW produces a multiple alignment in a file called `worm.aln`. I took a quick look at the alignment in the graphical display of ClustalX, just to be sure that it seemed sensible. In a careful analysis, I might also edit and trim the alignment; a molecular biologist's eye is almost always a better judge than a program's.

Profile searches

The next step is to construct a profile of the multiple alignment, and to search it against the sequence database. Using my HMMER 2.0 software (S.R.E., unpublished) for building profile hidden Markov models, starting with the ClustalW alignment of the 32 sequences in `worm.aln`, I typed the following series of commands:

```
% hmmbuild worm.hmm worm.aln
% hmmcalibrate worm.hmm
% hmmsearch worm.hmm swiss35
```

The `hmmbuild` command builds a profile `worm.hmm` from the alignment, taking a few seconds. The `hmmcalibrate` command automatically estimates some parameters needed for calculating accurate E-values in database searches, taking several minutes. The `hmmsearch` command searches Swissprot 35 (on local disk) with the profile, taking several hours. The output is a ranked list of hits, giving E-values.

The HMMER output shows a number of mammalian GPCRs with significant hits. The top scoring ones are somatostatin receptors in the GPCR superfamily (for instance, `SSR3_RAT`, $E=0.029$). An E-value of 0.029 is a marginal but significant hit. I typically use 0.05 as a trusted cutoff for HMMER. After 29 other GPCRs from other organisms comes the top scoring non-GPCR, a dicarboxylic amino acid permease (`DIP5_YEAST`, $E=0.45$).

Thus, from one multiple sequence based search, I can predict homology between AH6.8 and mammalian G-protein coupled receptors.

PSI-BLAST

One practical problem with this analysis is that I needed a number of software packages and databases installed on my workstation. Many biologists would prefer a Web server.

NCBI's PSI-BLAST ("position specific iterated BLAST") server provides such a Web service [8]. PSI-BLAST is an iterative profile search. A single sequence is first searched against the database using BLAST. The significant hits are aligned to the query, and a profile of the alignment is built. This profile is searched against the database to gather more hits and make a new alignment. This is iterated repeatedly, possibly until nothing new is found.

PSI-BLAST was designed to be an interactive tool. Compromises were made to favor speed over other considerations. In general, the profile HMM software packages are more sensitive and specific, but are far slower.

I submitted the AH6.8 sequence to the PSI-BLAST server (version 2.0.5, searching SWISS-PROT, default parameters). After just one iteration, the significant hits include a number of GPCRs from other organisms. The top scoring one is `P2YR_HUMAN`, a human endothelial purinergic receptor ($E=1e-6$) in the GPCR superfamily.

However, the top scoring non-nematode hit is actually a putative mitochondrial 60S ribosomal L11 protein (`RM11_ACACA`), that gets an extremely significant E-value of $5e-49$. Why such disparate results? Is AH6.8 a GPCR or a ribosomal L11 protein? A careful analyst would probably do more work and decide that AH6.8 is a probable GPCR, but a careless analyst might annotate AH6.8 as a ribosomal L11 protein based solely on its best PSI-BLAST hit. This is an example of the two biggest dangers in profile analysis.

Two pitfalls for the unwary

In building a profile, you implicitly assume that all the aligned sequences belong to the same family. If the alignment errantly includes unrelated sequences, profile scores will mislead you. A profile of a spurious multiple alignment of kinases and globins will recognize either a kinase or a globin with significant scores, but this would not mean that kinases and globins are homologous.

This is particularly nasty in iterative approaches like PSI-BLAST. Once PSI-BLAST mistakenly includes a nonhomologous sequence with a borderline score, the *next* iteration will almost certainly include that same sequence (and its homologues) with highly significant E-values. In the example, PSI-BLAST assigned `RM11_ACACA` a score that was barely over the inclusion threshold in the first search. Once it was in the training set, `RM11_ACACA` then received an E-value of $5e-49$ in the next iteration. All this score means is that `RM11_ACACA` is similar to itself, not that it is similar to AH6.8. A PSI-BLAST E-value reflects the significance of the match to the training set in the previous iteration, *not* the significance of the match to the original query sequence.

Therefore, after defining a family of sequences by any iterative search procedure, one should trace the chain of evidence linking each sequence to the rest of the family. One approach is cluster analysis of pairwise BLAST similarities, e.g. identifying weak links between groups of more clearly homologous sequences, so these links can be given more careful consideration. In a cluster analysis of the PSI-BLAST hits, RM11_ACACA stands out as an outlier.

The second pitfall is that database annotation is of variable quality, especially because dubious annotations are propagated to other sequence homologues. RM11_ACACA is annotated as a ribosomal L11 protein – but why? In fact, I cannot find experimental evidence for the classification of RM11_ACACA from the MEDLINE reference in the SWISS-PROT entry, and a cursory BLAST analysis shows only marginal similarity to other (putative!) L11 proteins. Without more evidence, I'm not confident of what RM11_ACACA really is. It's even possible that it is a GPCR. I have to call it an uninformative hit.

Conclusion

Other important applications of profile searches involve starting with multiple alignments of known sequence families, using characterized sequences in the public database. Pre-built multiple alignments and profiles are publicly available for hundreds of known sequence families. The uses of profile databases are discussed in Kay Hofmann's article elsewhere in this issue.

A web page with hyperlinks to the inputs and outputs of the example AH6.8 analysis in this paper is available at <http://www.genetics.wustl.edu/eddy/publications/tigs-9808/>.

References

- [1] Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.*, 6:361–365
- [2] Eddy, S. R. (1998) *Bioinformatics*, in press
- [3] Krogh, A. (1998) *Computational Methods in Molecular Biology* (Salzberg, S., Searls, D., and Kasif, S., eds), pp. 45–63, Elsevier Science
- [4] Durbin, R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press
- [5] Baldi, P. and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press
- [6] Troemel, E. R. *et al.* (1995) *Cell*, 83:207–218
- [7] Robertson, H. M. (1998) *Genome Res.*, 8:449–463
- [8] Altschul, S. F. *et al.* (1997) *Nucl. Acids Res.*, 25:3389–3402
- [9] Sonnhammer, E. L. L. and Durbin, R. (1997) *Genomics*, 46:200–216
- [10] Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucl. Acids Res.*, 22:4673–4680
- [11] Thompson, J. D. *et al.* (1997) *Nucl. Acids Res.*, 25:4876–4882

Glossary

Cluster analysis A process of assigning data points (sequences) into groups (clusters), starting from pairwise distances. Useful for identifying outliers and weak links between groups. Fairly easy to do by hand for small data sets.

E-value For a given score, the number of hits in a database search that we expect to see by chance with this score or better. The E-value takes into account the size of the database that was searched. The lower the E-value, the more significant the score is. See P-value.

Gene family Two or more genes that are related by divergent evolution from a common ancestor, either by speciation or gene duplication.

Graphical user interface Software that allows a user to interact via “user-friendly” menu and mouse-driven commands, as is typical of Macintosh and Windows applications, and less common for UNIX applications; as opposed to a “command line interface” of typed or scripted commands.

Hidden Markov model A kind of formal probabilistic model that is well suited to providing a mathematical framework for profile analysis.

Iterative search A search procedure that is repeated, usually with increasing sensitivity in each round. For instance, taking all the significant hits from an initial BLAST search and using each of them as a query for a new round of BLAST searches would be one form of iterative search.

Linux A freely available but commercial-strength clone of the UNIX operating system. A godsend for starting bioinformaticians on a budget. Easily installed alongside Windows on a PC, so the same machine can be booted into either Linux or Windows.

Multiple alignment An alignment of three or more sequences, with gaps (spaces) inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column of the multiple alignment.

Paralogs Two homologous sequences (e.g. sequences that share a common evolutionary ancestor) that diverged by gene duplication, as opposed to *orthologs*, which diverged by speciation. A gene family within a single organism is necessarily composed only of paralogs (barring horizontal transmission of genes from another species).

Profile A linear model of the consensus of a multiple alignment. For each column of a protein alignment, a profile assigns 20 residue scores (one per amino acid), and one or more gap penalties for insertions of extra residues adjacent to this column or a deletion of the consensus residue at this column. Profiles are also called “position specific scoring matrices” (PSSMs). Profiles that don’t allow insertions and deletions are also called “weight matrices”.

P-value Like an E-value, but a P-value is the probability of a hit occurring by chance with this score or better, as opposed to the expected number of hits. A P-value has a maximum of 1.0, while an E-value has a maximum of the number of sequences in the database that was searched. For small (significant) P-values, P and E are approximately equal, so the choice of one or the other in a software package is arbitrary. NCBI BLAST 2.0, FASTA, and HMMER report E values. WU-BLAST 2.0 reports P values.

Software and databases used in example analysis:

NCBI BLAST2.0 server	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast?Jform=1
WUBLAST software	http://blast.wustl.edu/
CLUSTALW software	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/
CLUSTALX software	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
HMMER software	http://hmmmer.wustl.edu/
NCBI PSI-BLAST server	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast
Wormpep 13 database	http://www.sanger.ac.uk/Projects/C_elegans/wormpep/
Swissprot 35 database	http://expasy.hcuge.ch/sprot/

Some other profile and profile HMM software packages:

SAM	http://www.cse.ucsc.edu/research/complibio/sam.html
PFTOOLS	http://ulrec3.unil.ch:80/profile/
HMMpro	http://www.netid.com/
GENEWISE	http://www.sanger.ac.uk/Software/Wise2/
PROBE	ftp://ncbi.nlm.nih.gov/pub/neuwald/probe1.0/
META-MEME	http://www.cse.ucsd.edu/users/bgrundy/metameme.1.0.html
BLOCKS	http://www.blocks.fhcrc.org/

Web servers for multiple alignment:

BCM Search Launcher	http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html
WashU IBC	http://www.ibc.wustl.edu/service/clustal.html

Other lists of pointers:

EBI's BioCatalog	http://www.ebi.ac.uk/biocat/biocat.html
------------------	---

One source for the Linux operating system:

Red Hat Linux	http://www.redhat.com
---------------	---

Table 1: A sample of Internet URLs for software, databases, and servers relevant to multiple alignment and multiple alignment based searches.