Packages are available that can take a multiple sequence similarity search methods can be applied. Software with no informative

It used to be that most new sequences were novel, with no informative similarity to anything in the sequence database. As a result of genome sequencing projects, the situation is now

Comparisons of multiple sequences can reveal gene functions that are not clear from simple sequence homologies. The important parameters in multiple alignment and multiple-sequence-based searches, using an example from Caenorhabditis elegans are described.

It used to be that most new sequences were novel, with no informative similarity to anything in the sequence database. As a result of genome sequencing projects, the situation is now

**Multiple-alignment and -sequence searches**

Comparisons of multiple sequences can reveal gene functions that are not clear from simple sequence homologies. The important parameters in multiple alignment and multiple-sequence-based searches, using an example from Caenorhabditis elegans are described.

It used to be that most new sequences were novel, with no informative similarity to anything in the sequence database. As a result of genome sequencing projects, the situation is now

Comparisons of multiple sequences can reveal gene functions that are not clear from simple sequence homologies. The important parameters in multiple alignment and multiple-sequence-based searches, using an example from Caenorhabditis elegans are described.

It used to be that most new sequences were novel, with no informative similarity to anything in the sequence database. As a result of genome sequencing projects, the situation is now
similarity search. Much more comprehensive reviews of the literature on profile HMM methods are available elsewhere. A Web page with hyperlinks to the inputs and outputs of the example AH6.8 analysis discussed in this article is available at http://www.genetics.wustl.
edu/eddy/publications/tigs-9808/.

An example sequence

In the Caenorhabditis elegans genome, several large paralogous gene families (paralogos), which were first thought to be nematode specific, have since been classified as putative G-protein-coupled receptors (GPCRs). Detecting similarity between these nematode sequences and known GPCRs in other organisms is a nontrivial sequence analysis task as a simple BLAST search is not sufficient to detect the remote similarity. Here, I will describe the steps taken to find a significant similarity between the putative GPCR gene A8X4-4 (Wormpep AH6.8; SWISS-PROT SRM4_CAEEL; 329 amino acids) and a protein of known function in another organism.

A World Wide Web (WWW) BLAST search at the National Center for Biotechnology Information (NCBI) using AH6.8 as a query (BLASTP 2.0.4, default options, versus 319 187 sequences in the nr database on 30 July 1998) showed 46 hits with E values of <0.01, with all but one of the hits corresponding to uncharacterized C. elegans sequences. The top-scoring non-worm hit was 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 a borderline significance score, but it is not low enough to be trusted without further information. Thus, at first glance, AH6.8 appears to be 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. The next step is to produce a multiple alignment. ClustalW is 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 UNIX systems. There is also a graphical user interface – ClustalX (Ref. 11). Obtaining an acceptable multiple sequence alignment is usually straightforward, once the family is defined. Starting from a file of 32 sequences in FASTA format, called worm.fa, the following command was typed at the command line of the UNIX workstation:

`% clustalw worm.fa`  
ClustalW then produces a multiple alignment in a file called worm.aln. It is important, at this point, to inspect the alignment in the graphical display of ClustalX, to make sure that it seems sensible. In a careful analysis, the alignment can also be edited and trimmed.

Profile searches

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

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

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

The HMMER output showed several mammalian GPCRs with significant hits. The top-scoring hits were somatostatin receptors in the GPCR superfamily (for example, SSRC3_RAT; E = 0.029). An E value of 0.029 represents a marginal but significant hit. A trusted cutoff for HMMER is typically 0.05. The SSRC3_RAT hit was followed, with increasing E values, by 29 other GPCRs from other organisms and then the top-scoring non-GPCR, a dicarboxylic amino acid permease (DHPL_YEAST; E = 0.45). Thus, from one multiple-sequence-based search, homology between AH6.8 and mammalian GPCRs can be predicted.

PSI-BLAST

One practical problem with this analysis is that several software packages and databases need to be installed on the local workstation. Many biologists prefer a Web server.
The NCBI PSI-BLAST server provides such a Web service. 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, and 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.

When the AH6.8 sequence was submitted to the PSI-BLAST server (version 2.0.5, searching SWISS-PROT; default parameters), the significant hits after just one iteration included a number of GPCRs from other organisms. The top-scoring hit was P2YR_HUMAN, a human endothelial purinergic receptor (E \textsuperscript{-}value of 5 \times 10\textsuperscript{-5}). So why did this search produce such disparate results, and is AH6.8 a GPCR, or a ribosomal L11 protein? A careful analyst would probably do more work and decide that AH6.8 is probably a GPCR, but a careless analyst might annotate AH6.8 as a ribosomal L11 protein. The second pitfall is that database annotation is propagated to other sequence homologs. 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; for example, identifying weak links between groups of more clearly homologous sequences so that 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 homologs. RM11\_ACACA is annotated as a ribosomal L11 protein – but why? In fact, there is no experimental evidence for the classification of RM11\_ACACA from protein – but why? In fact, there is no experimental evidence for the classification of RM11\_ACACA from protein – but why? In fact, there is no experimental evidence for the classification of RM11\_ACACA from
Protein classification & functional assignment

There are several collections of amino acid sequence motifs that indicate particular structural or functional elements. Web-based searches of these collections with a newly identified sequence allow reasonably confident functional predictions to be made.

A variety of genome and cDNA sequencing projects is producing raw sequence data at a breathtaking speed, creating the need for a large-scale functional classification effort. On a smaller scale, the average molecular biologist can also be faced by a new sequence without any a priori functional knowledge. Any hint as to whether the newly identified gene encodes a transcription factor, a cytoskeletal protein or a metabolic enzyme would certainly help to interpret the experimental results and would suggest a direction for subsequent investigations.

Protein versus domain classification

The first step in addressing this question is usually a database search with BLAST or a similar program (see the article by Steven Brenner on pp. 9–12). In the best case scenario, the BLAST output would show a clear similarity to a single, well-characterized protein spanning the complete length of the query protein. In the worst case scenarios, the output list would fail to show any significant hit. In reality, the most frequent result is a list of partial matches to assorted proteins, most of them uncharacterized, with the remainder having dubious or even contradictory functional assignments.

Much of this confusion is caused by the modular architecture of the proteins involved. An analysis of known 3-D protein structures reveals that, rather than being monolithic, many of them contain multiple folding units. Each unit, termed a domain, has its own hydrophobic core and satisfies most of its residue–residue contacts internally. In order to fulfill these conditions, independent domains must have a minimum size of ~50 residues unless they are stabilized by metal ions or disulfide bridges. Analysis of protein sequences corroborates this structural notion; sequence pairs frequently exhibit localized regions of similarity; the remainder of the protein being totally dissimilar. A folding independence for all of these so-called homology domains has not been demonstrated experimentally. However, they do show context independence; that is, they can occur in various domain arrangements and are occasionally

References

2 Eddy, S.R. Bioinformatics (in press)

Kay Hofmann
MEMOREC Stoffel GmbH,
D-50829 K鰊, Germany
kay.hofmann@memorec.com

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, it is not clear what RM11_ACACA really is. It is even possible that it is a GPCR. It should therefore be called 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. Prebuilt multiple alignments and profiles are publicly available for hundreds of known sequence families. The uses of profile databases are discussed by Kay Hofmann on pp. 19–21.