conditions that give rise to differing expression levels for different transcripts are elucidated. This implies that future genefinders will also need to take explicitly into account experimental data relating to differential expression, as well as the other types of data discussed here. It is anticipated that this task will occupy genefinding researchers for some years to come.

Acknowledgements

I gratefully acknowledge the support of DOE grant DE-FG03-95ER6211 and thank R. Guigó, D. Kulp and M. Reese for helpful suggestions.

References

1 Guigó, R. (1997) *Comput. Chem.* 21, 215–222
2 Claverie, J-M. (1997) *Hum. Mol. Genet.* 6, 1735–1744
3 Krogh, A. (1998) in *Guide to Human Genome Computing* (2nd edn) (Bishop, M.J., ed.), pp. 261–274, Academic Press
4 Gelfand, M.S. (1995) *J. Comput. Biol.* 2, 87–115
5 Staden, R. (1984) *Nucleic Acids Res.* 12, 505–519
6 Stormo, G.D. (1990) *Methods Enzymol.* 183, 211–220
7 Fickett, J.W. (1996) *Comput. Chem.* 20, 103–118
8 Borodovsky, M. and McIninch, J. (1993) *Comput. Chem.* 17, 123–133
9 Xu, Y. *et al.* (1994) in *Proceedings of the Conference on Intelligent Systems in Biology (ISMB)* (Altman, R. *et al.*, eds), pp. 376–383, AAAI/MIT Press
10 Claverie, J-M. (1992) *Comput. Chem.* 16, 89–91
11 Zhang, M.Q. (1998) *Hum. Mol. Genet.* 7, 919–932
12 Searls, D.B. (1992) *Am. Sci.* 80, 579–591
13 Durbin, R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press
14 Krogh, A., Mian, I.S. and Haussler D. (1994) *Nucleic Acids Res.* 22, 4768–4778
15 Cole, S. *et al.* (1998) *Nature* 393, 537–544
16 Thomas, A. and Skolnick, M. (1994) *IMA J. Math. Appl. Med. Biol.* 11, 149–160
17 Lukashin, A.V. and Borodovsky, M. (1998) *Nucleic Acids Res.* 26, 1107–1115
18 Henderson, J., Salzberg, S. and Fasman, K. (1997) *J. Comput. Biol.* 4, 119–126
19 Snyder, E. and Stormo, G. (1995) *J. Mol. Biol.* 248, 1–18
20 Kulp, D. *et al.* (1996) in *Proceedings of the Conference on Intelligent Systems in Biology (ISMB)* (States, D.J. *et al.*, eds), pp. 134–142, AAAI/MIT Press
21 Reese, M.G. *et al.* (1997) *J. Comput. Biol.* 4, 311–323
22 Burge, C. and Karlin, S. (1997) *J. Mol. Biol.* 268, 78–94
23 Kulp, D. *et al.* (1997) in *Proceedings of the Pacific Symposium on Biocomputing* (Altman, R.B. *et al.*, eds), pp. 232–244, World Scientific
24 Xu, Y. and Uberbacher, E.C. (1997) *J. Comput. Biol.* 4, 325–338
25 Gelfand, M.S. *et al.* (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 9061–9066
26 Smith, T.F. and Waterman, M.S. (1981) *Adv. Appl. Math.* 2, 482–489
27 Kozak, M. (1996) *Mamm. Genome* 7, 563–574
28 Nagel, R. *et al.* (1998) *RNA* 4, 11–23

Multiple-alignment & -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.

Sean R. Eddy
Dept of Genetics, Washington
University School of Medicine,
4566 Scott Ave, St Louis,
MO 63110, USA.
eddy@genetics.wustl.edu
http://www.genetics.wustl.edu/eddy/

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 slightly improved. New sequences are often found to be similar to several uncharacterized sequences, defining whole families of novel genes with no informative **BLAST** or **FASTA** similarities.

However, given a sequence family, 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 that is derived from the frequency with which 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 carried out.

Most new profile software is based on statistical models called hidden Markov models (**HMM**s). Here, a practical demonstration is given of a multiple-alignment-based

0167-7799/98/$ – see front matter © 1998 Elsevier Science. All rights reserved.

similarity search. Much more comprehensive reviews of the literature on profile HMM methods are available elsewhere[1–5]. 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 (**paralogs**), which 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; 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 *sra-4* (**Wormpep** AH6.8; **SWISS-PROT** SRA4_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)[8] 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 hits were 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 on a PC.

Sequence gathering

The first step for further analysis is to define more carefully 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) of Wormpep 13 using AH6.8 as the query pulled out 36 hits with highly significant ***P* values** of $<10^{-6}$. Most were ~350 amino acids long. As a crude protection against erroneous computational gene predictions, four sequences that were either longer than 500 amino acids or shorter than 200 amino acids were discarded, leaving 32 sequences. Sometimes, more must be done to define the sequences to align. For example, sequences might be related by a shared domain instead of over their entire length; thus, it might be necessary to isolate alignable subsequences (based on the bounds of BLAST alignments, for example) before making the multiple alignment. This step can involve quite a lot of manual work.

Multiple sequence alignment

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[10]. 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, SSR3_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 SSR3_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 (DIP5_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[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, 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 ($E = 1 \times 10^{-6}$) in the GPCR superfamily. However, the top-scoring non-nematode hit was actually a putative mitochondrial 60S ribosomal L11 protein (RM11_ACACA), which had an extremely significant E value of 5×10^{-49}. 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 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, it is implicitly assumed that all the aligned sequences belong to the same family. If the alignment errantly includes unrelated sequences, profile scores will be misleading. 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 unhelpful 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 homologs) with highly significant E values. In the above example, PSI-BLAST assigned a score to RM11_ACACA that was barely above the inclusion threshold in the first search. Once it was in the training set, RM11_ACACA then received an E value of 5×10^{-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; 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

URLs...

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://hmmer.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/

SWISS-PROT 35 database
http://expasy.hcuge.ch/sprot/

Other profile and profile HMM software packages

SAM
http://www.cse.ucsc.edu/research/compbio/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/searchlauncher/launcher.html

WashU IBC
http://www.ibc.wustl.edu/service/clustal.html

Other lists of pointers

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

One source for the Linux operating system

Red Hat Linux
http://www.redhat.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. 18–21.

References

1 Eddy, S.R. (1996) *Curr. Opin. Struct. Biol.* 6, 361–365
2 Eddy, S.R. *Bioinformatics* (in press)
3 Krogh, A. (1998) in *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) *Nucleic Acids Res.* 25, 3389–3402
9 Sonnhammer, E.L.L. and Durbin, R. (1997) *Genomics* 46, 200–216
10 Thompson, J.D. *et al.* (1994) *Nucleic Acids Res.* 22, 4673–4680
11 Thompson, J.D. *et al.* (1997) *Nucleic Acids Res.* 25, 4876–4882

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.

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

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 scenario, 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 proteins 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

0167-7799/98/$ – see front matter © 1998 Elsevier Science. All rights reserved.