genefinder program), that displayed significant sequence similarity to *pros*. Closer inspection showed that the major block of sequence conservation between K12H4.1 and pros is confined to a block of about 160 amino acids at the extreme carboxyl terminus of both proteins (Fig. 1). A few smaller blocks of sequence similarity can be found upstream of this region (two are shown in Fig. 1), but both proteins contain regions rich in amino acids such as serine, glutamine and proline, so that these similarities might just be chance occurrences. Translation of the whole cosmid in all three reading frames, and comparison with the first 1240 amino acids of the prospero protein, did not reveal any further exons with sequence similarity to K12H4.1 that had been missed by genefinder.

The highest sequence similarity (79% identity) occurs in the region predicted to be a homeodomain. This supports the notion that this region is indeed an atypical homeodomain despite the weak similarity. The block of 100 amino acids between the homeodomain and the carboxyl terminus shows 60% identity between pros and K12H4.1. This motif, termed the *prospero* domain, is novel, and databank searches did not reveal any significant similarities to other motifs or proteins.

In conclusion, *pros* and K12H4.1 define a novel class of atypical homeobox genes that encode evolutionarily highly conserved homeodomains. In addition, a novel motif extends from th homeodomain to the carboxyl terminus in both proteins. Given that some of the conserved residues in the *prospero* domain are basic residues, the possibility arises that this domain might be DNA binding.

THOMAS R. BÜRGLIN

Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School, Wellman 8, Boston, MA 02114, USA.
email: burglin@frodo.mgh.harvard.edu

References

1 Chu-Lagraff, Q., Wright, D. M., McNeil, L. K. and Doe, C. Q. (1991) *Development* (Suppl.) 2, 79–85
2 Matsuzaki, F. *et al.* (1992) *Biochem. Biophys. Res. Commun.* 182, 1326–1332
3 Vaessin, H. *et al.* (1991) *Cell* 67, 941–953
4 Bürglin, T. R. (1993) in *A Guidebook for Homeobox Genes* (Duboule, D., ed.), pp. 25–71, Oxford University Press
5 Ceska, T. A. *et al.* (1993) *EMBO J.* 12, 1805–1810
6 Leiting, B. *et al.* (1993) *EMBO J.* 12, 1797–1803
7 Sulston, J. *et al.* (1992) *Nature* 356, 37–41
8 Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.* 12, 387–395

I think therefore I publish

The *Protein Sequence Motifs* column of *TIBS* was initiated as a forum for discussion of novel sequence similarities and their possible biological significance. The main criteria for publication in this column were, and still are, that (1) the sequence(s) be either published or freely available from an electronic database, and (2) that the observation of homology be significant and original. However, sophisticated software technology for the automated detection and annotation of sequence similarities has raised some new issues and caused us to ponder the meaning of originality and, hence, the 'publishability' of sequence motifs.

The accompanying letter by Thomas Bürglin is a case in point. That a *Drosophila prospero* homolog exists in *Caenorhabditis elegans* had already been documented in the GenBank record (accession no. L14331) for the cosmid sequence that contains it. This homology information was generated by a semi-automated system developed for the nematode genome project, verified by a human analyst and submitted to the database (L. Hillier, pers. commun.). This discovery was made a second time and independently, however, by a software robot (without any human intervention at all) and distributed to thousands of subscribers to the *Entrez* integrated database and retrieval system[1] (see article on p.94 of this issue for a review of *Entrez* software). In this latter case, the *prospero* discovery was implicit in the sense that, initially, only a computer was 'aware' of it.

Automated data analysis and annotation have arisen out of necessity. GenBank[2] currently contains 164 million nucleotides in 151000 sequences and is doubling in size every 21 months. In addition, the impact of high-throughput genomic sequencing of 50 megabases or more per year[3] has yet to be felt. The effective management of these data, described by the term 'bioinformatics', involves not only the analysis and annotation of DNA and protein sequences, but also the establishment of all the relevant links to other databases, such as those of three-dimensional structures (Protein Data Bank), genetic map locations (Genome Data Base) and the biomedical literature (MEDLINE). A good deal of this analysis and annotation occurs in the laboratories producing the data, but the public databases have an important role and responsibility to provide up-to-date annotation, to establish and maintain links to other databases, and to provide easy access for all.

Powerful software has been developed to link diverse databases and automatically search for homologs of new sequences. For each release of *Entrez*[1], for example, all DNA and protein sequences in the databases are searched against all others, and significant relationships automatically become part of the database. (A type of 'homology' searching, called 'neighboring', is also carried out for publications.) Many of these relationships represent answers to questions that have yet to be asked, as in the case of the *prospero* homolog described above. Researchers can access these pre-computed results via the Internet or on CD-ROM (contact info@ncbi.nlm.nih.gov) and thereby check whether a homology has been documented in an electronic database or in the conventional published literature.

Because such automated systems are keeping track of new homologies for us, what then constitutes an original, publishable observation of sequence similarity? For the present, it would seem reasonable to suggest that any homology noted by a human, whether or not it has been noted already by a software robot, should be judged in terms of the added value of human interpretation and the desirability of communicating this information to a specific audience. But as information technologies are implemented for the automatic and targeted dissemination of new findings to all potentially interested parties, many aspects of traditional publication may have to be reassessed.

References

1 NCBI News (1993) 2, 1–4
2 Benson, D., Lipman, D. J. and Ostell, J. (1993) *Nucleic Acids Res.* 21, 2963–2965
3 Collins, F. and Galas, D. (1993) *Science* 262, 43–46

MARK BOGUSKI

National Center for Biotechnology Information, National Library of Medicine, NIH, 8600 Rockville Pike, Bethesda, MD 20894, USA.

JO MCENTYRE
Editor

TIBS
welcomes
letters to
the Editor.