Ready for a motif submission? A proposed checklist

The Protein Sequence Motifs column is now entering its fourth year, and its popularity is underlined not only by its longevity, but also by the steadily increasing number of submissions to the column. In order to improve reproducibility of the results and the widespread acceptance of new motifs, we would like to give some guidelines that reflect common practices in the field of sequence analysis. Some criteria for submission of Protein Sequence Motifs have already been discussed[1], and some additional guidelines are given in Box 1.

Box 1. Suggestions for figures and significance estimates

Figures

- An alignment of representative sequence segments that comprise the motif provides strong support for its existence.
- The selected set should contain distantly related members of the respective family rather than a subset that fits best with the proposed motif.
- The alignments should include information about (1) Protein names/abbreviations that correspond to the text (figure caption); (2) positions of the displayed sequence regions within the respective proteins; and (3) accession numbers from widely used sequence databases for unique identification.

Significance estimates

- As the currently used algorithms for evaluating the significance of a similarity are far from optimal [for example, the gap problem is not completely solved; the statistics depend strongly on elements such as empirically derived similarity matrices or the redundant (and partially erroneous) data of public databases], novel methodologies can be used, but they should be described explicitly and briefly (or extensively elsewhere). In such cases, however, there should be a justified reason why standard methods cannot produce similar results.
- The scoring schemes in the BLAST or FASTA series of programs already give hints for pairwise comparison but require the presence of globular domains, i.e. compositionally biased (low complexity) regions need to be filtered[8].
- The amino acid substitution matrix and gap parameters used should be mentioned if they differ from standard implementations (for example, BLOSUM62 for BLAST and PAM250 for FASTA). Furthermore, different matrices should be applied in comparisons, and should, ideally, result in similar alignments.
- The probability of homologies within a multiple alignment increases if certain amino acids are conserved in all sequences. Programs such as MACAW[10] and MoST[11], algorithms associated with the Blocks database[12] and motif finders (such as those base on Gibbs-samplers or hidden Markov models) can at least provide support for ungapped motifs by assigning probabilities of matching by chance.
- Again, these tests should be complemented by searches for compositionally biased regions such as coiled coils, transmembrane sequences, short repeats or segments rich in certain amino acids[13].

Sequence motifs can be classified into two basic types: most comprise patterns that imply common descent of the proteins or domains under consideration (divergent evolution), while a few are analogous (convergent) functional motifs. Before either type of motif is submitted, it is necessary to check whether a genuine motif is being described for the first time, or whether extension of a previously identified motif is of sufficiently novel functional significance to warrant publication. Second, and most important, the demonstration of statistical significance is, and should remain, essential.

Assessing statistical significance is particularly difficult for convergent motifs (for example, short functional motifs such as glycosylation sites, RGD motifs or PEST sequences), since it is only based on accuracy estimates (ratio of putative true positives over false positives) within the current set of database proteins. Thus, even significant deviations from a random distribution are not sufficient to predict the existence of such a motif in a particular protein. Although convergent functional motifs might indeed guide further experimentation, reports of these should not focus on particular examples, and adequate statistical tests on the databases should be applied.

However, the majority of motifs appearing in this column reflect common ancestry, and several significance estimates can therefore be applied. These estimates should be used before fitting a subtle similarity to a tempting functional hypothesis. Pairwise identities as low as 3% might still reflect a homology, while in other cases amino acid identities above 30% may not be significant, particularly for short sequences (see, for example, Refs 2, 3).

Since similarity, and not just identity, of amino acids has to be taken into account, the exact and empirical theories available for this kind of problem, such as Poisson scores[4] or extreme value statistics[5], should be made use of for this kind of problem. Computer programs are available for such methods, and are widely used.

Another approach is the incorporation of the random background of sequences, as measured in several database search programs such as FASTA[6] and rigorously applied in BLASTP[7,8]. There are several powerful methods that take into account the conservation of amino acids in an alignment. Widely used pattern and profile search methods can verify the significance of a newly discovered motif (see, for example, Ref. 9 and subsequent improvements such as better sequence weights).

Only a rigorous analysis of a putative novel motif hints at its significance and functional relevance; any conclusion drawn regarding function requires an even more careful interpretation of the background of the respective proteins. For the transfer of functional information on the basis of sequence similarity, no recipe can be given here, but experience shows that care is required at this stage to prevent erroneous functional predictions and to save efforts in experimental analysis.

References

1 Boguski, M. S. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71

2 Doolittle, R. F. (1986) *On ORFs and URFs*, University Science Books

3 Sander, C. and Schneider, R. (1991) *Proteins* 9, 56–68

4 Karlin, S. and Altschul, S. F. (1990) *Proc. Natl Acad. Sci. USA* 87, 2264–2268

5 Arratia, R., Gordon, L. and Waterman, M. S. (1990) *Ann. Stat.* 18, 539–570

6 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl Acad. Sci. USA* 85, 2444–2448

7 Altschul, S. F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410

8 Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994) *Nature Genetics* 6, 119–163

9 Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987) *Proc. Natl Acad. Sci. USA* 84, 4355–4358

10 Schuler, G. D., Altschul, S. F. and Lipman, D. (1990) *Proteins* 9, 180–190

11 Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994) *Proc. Natl Acad. Sci. USA* 91, 12091–12095

12 Henikoff, S. and Henikoff, J. G. (1992) *Proc. Natl Acad. Sci. USA* 89, 10915–10919

13 Bork, P., Ouzounis, C. and Sander, C. (1994) *Curr. Opin. Struct. Biol.* 4, 393–403

PEER BORK AND CHRISTOS OUZOUNIS

EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

JO MCENTYRE

Trends in Biochemical Sciences.