New horizons in sequence analysis
Tim JP Hubbard

An ever increasing number of protein sequences are being compared, partly because of the availability of full sets of protein sequences from several completed genome-sequencing projects. The resulting problem of scale has shifted the emphasis of sequence analysis method development from sensitivity and flexibility, which relies on manual intervention and interpretation, to the automatic generation of results of known reliability.

Introduction
A major objective of the analysis of protein sequences is to discover which share a common ancestor. Demonstrating an evolutionary relationship is useful because it generally implies a common or similar function. In an era in which whole genomes have been sequenced, clustering together proteins of equivalent function is probably of more immediate use than predicting protein structure, as it allows the metabolic systems of different organisms to be compared and contrasted [1, 2••].

Sequence-comparison methods are used to detect evolutionary relationships. Probably the simplest scoring scheme for such methods is sequence identity, for example, significant similarity is indicated if two sequences can be aligned over a region of 100 residues with 40 residues being identical. Much better scoring schemes than this can identify much weaker similarities as being significant, however, the only real test for whether two sequences are related is to compare their 3D structures.

So far, an example of a pair of sequences that have clear sequence similarity but dissimilar 3D structures has not been found, so it is generally accepted that a common ancestor always implies a common 3D structure as well. Many structures have been found to be very similar even though their sequences appear unrelated by the most sensitive methods. It can be argued that such sequences do not need to be related, as the structures they adopt could have just converged to similar, physically favoured conformations. In a substantial number of cases, however, not only are the structures similar, but so are the functions, the active-site locations, etc., which makes it clear that the sequences actually are evolutionarily related. The conclusion is that many structural similarities are also evolutionary similarities but these cannot be detected using even the most sensitive sequence-comparison methods.

In this review, I discuss the effectiveness of sequence-comparison methods, their evaluation and what improvements to detection accuracy and coverage are probable in the near future.

Sequence-comparison methods and statistical scoring schemes
Sequence-comparison methods fall into two broad groups: pairwise methods in which pairs of sequences are directly compared; and methods in which models are built from pre-existing multiple sequence alignments and compared either with other sequences or other models. The major problem in using any sequence-comparison method is identifying which of the proposed alignments are biologically meaningful. Ideally, one would like a scoring scheme that has a known cut-off, above which all alignments represent evolutionary relationships, and below which all alignments are random with no biological meaning. In fact, the most that can be hoped for is a scheme that gives the probability that an alignment is significant, which allows a cut-off to be chosen that gives an acceptable balance between the detection rate for evolutionary relationships (the ‘coverage’) and the number of false positives. For example, when looking for distant members of a single protein family, a relatively large number of false positives may be acceptable, as each can be examined individually. Conversely, when automatically clustering a large database of sequences by a single linkage approach, a cut-off that gives a very low number of false positives must be selected to minimize the number of clusters containing groups of totally unrelated sequences. With the efforts to completely cluster all sequences into domain families [3••], particularly from the growing number of genomes, methods are required with a single cut-off that gives the highest possible coverage consistent with very few false positives.

A significant step towards such a method was the development of the pairwise program BLAST, which, as well as being very fast, applied analytical statistics to its fragmentary ungapped alignments to rank them by a P value—the probability of the alignment occurring by chance [4–6]. More recently, the same extreme value distribution (EVD) has been applied to gapped alignments [7, 8], and an empirical application has been incorporated...
into FASTA and SSEARCH implementations [9] of the Smith and Waterman algorithm [10].

Clearly, more information is used in methods in which a model has been constructed, regardless of whether is it a motif, pattern or profile (for terminology and a review, see [11•]). Such methods are accepted to be much more sensitive than pairwise ones, however, the large number of variables inherent in data derived from multiple sequence alignments make meaningful scoring schemes difficult to construct, and deficiencies in current ones are widely acknowledged [11•]. Recently, the application of Hidden Markov Models (HMM) to sequence searching has led to profile searching that has a statistical scoring scheme, although it is not based on EVD theory (for a review, see [12]).

A variation on the above methods is to incorporate structure information into the scoring matrix or model if the structure is known for one of the sequences to be compared. Algorithms that perform what is referred to as ‘threading’ extend this approach by scoring an alignment using a pseudoenergy function that contains 2D residue–residue pair information (for a review, see [13]). Threading is a relatively new technique and has special problems: first, the use of 2D information prevents the use of the fast dynamic programming algorithm to find the optimal alignment, which forces the use of various approximations. Second, these methods are designed and optimized for ‘fold recognition’, that is, they recognize similar structures, regardless of whether they have an evolutionary origin or not. This is especially hard in the nonevolutionary cases because of the plasticity of protein structures. Given that recognizing distant evolutionary relationships is hard enough, optimizing them to recognize only this subset of structural relationships might be profitable. At present, the use of threading methods is even more ad hoc than profiles [13], and I am only aware of one method that has a published statistical scoring scheme [14•], although with the requirements for statistically based predictions [15] for CASP2 [16•,17•] (see below) this may change.

Evaluation of comparison methods
How effective are statistical scoring schemes in practice? A standard way to test the efficiency of a given sequence-comparison method has been to apply it to one or a number of large protein families, for example, to take a globin sequence and see how many other globin sequences score above the first nonglobin sequence using that method. Unfortunately, this test is flawed as the list of ‘globins’ in any database is itself based mainly on sequence comparison, and, almost certainly, other sequences evolutionarily related to globins exist that are not labelled as such. As a result, assessing how many of the false positives are really false and how many as yet unrecognized ‘globins’ have been missed is impossible.

One way to test methods is to compare them against a database of sequences in which the evolutionary relationships are known. Databases in which similar structures are clustered have been available for sometime [18,19], however, as some structures may be similar as a result of convergent evolution, such databases are not ideal for benchmarking. Recently, the SCOP database [20] has been created, which subdivides clusters of similar structures into ‘superfamilies’ whose members are all expected to share a common ancestor due to conserved structural and sequence features. The number of evolutionary relationships in SCOP will be an underestimate, as similar structures are only grouped into superfamilies if the evolutionary evidence is clear, however, SCOP should not contain any false positives. A database of sequences of domains extracted from SCOP, in which no sequence has pairwise sequence identity of >40% with any other, was used to calibrate the SSEARCH [9], an implementation of the Smith and Waterman algorithm [10], prior to using it to search for duplications within the Haemophilus influenzae genome [21]. For this database, only around 15% of pairwise evolutionary relationships could be detected at an error rate of 1% false positives [22]. One implication of this remarkably low figure is that the amount of duplication within, and similarity between, genomes that can be currently be detected is a gross underestimate and that statements such as “protein X does not exist in genome Y” should be treated with extreme caution.

Another way to test methods is by blind prediction, in which the structures of pairs of sequences are only known after they have been predicted to be either similar or different. Two large scale prediction experiments have been organized, CASP1 [23•] and CASP2 [16•,17•], in which targets (sequences for which the structure is about to be solved) were distributed to predictors who had to submit predictions before the structure was announced. Although seen as a structure-prediction experiment, the results can also be evaluated to assess the limits of sequence-searching methods. For some targets, which showed no obvious sequence similarity to any sequence of known structure, not only did their structure turn out to be similar to a known fold, but the relationship appears to be an evolutionary one [24•]. Although the disadvantage of such experiments is that sample sizes are small, the results will allow a rough measurement of how much more can be predicted as a result of the manual interpretation of results by ‘experts’, and with what accuracy. Studying the ‘rules’ used by the predictors who successfully detected these relationships and attempting to incorporate them into scoring schemes may be beneficial.

Improving the detection rate
As far as I am aware, there is currently no publicly available implementation of EVD statistics to any profile method, although there seems to be no a priori reason why it should not be possible. The application of EVD should allow higher coverage of evolutionary relationships than those of
pairwise methods or existing profiles’ scoring schemes for the same error rates. For the subset of sequences in which a structure is known, threading may lead to even higher coverage than for sequence-only profiles, if statistics can be incorporated to allow the definition of a high reliability cut-off. The weakness of threading is the requirement for atomic coordinates, although, as it is widely assumed that there are only a limited number of protein folds [25–27], and therefore a limited number of evolutionarily related superfamilies, it is likely that it will be possible to cluster the majority of sequences to a sequence of known structure in the foreseeable future as more and more structures are solved.

Threading requires that the structure of at least one of the sequences being compared is known; however, other structure-prediction methods are available that are not based on a template (ab initio methods), which may extract additional structural information from a sequence family that is orthogonal to the information contained in a normal profile. Secondary-structure prediction is now sufficiently reliable [28] to be used to enhance sequence-searching methods. Combined with HMMs, it has been used for fold recognition [24,29] and has been noted as providing useful additional information to determine if two families linked by a statistically insignificant alignment are in fact related [11+]. The approach is to predict the secondary structure of each family and then to examine if these predicted segments match in the existing alignment. In order to assemble predicted secondary structural segments into a 3D model, contacts between these segments (so called ‘long range interactions’) need to be predicted. Two methods have been developed to do this [30,31] and, although their predictive power is currently too low for structure prediction, the information from these methods might be used similarly to the way predicted secondary structure is already used to discriminate sequence alignments.

Conclusions

In order to classify protein sequences in the way that protein structures have been classified, sequence-comparison methods of higher sensitivity and better reliability are needed. Improved techniques for calibrating and benchmarking should allow the iterative refinement of methods and scoring schemes to increase coverage while maintaining acceptably low error rates. Similar techniques that use structural comparison data for benchmarking are being used to assess and refine the quality of alignments (e.g., see [32+]), and these may also be useful in improving methods to tackle the problem of clustering while taking into account domain boundaries within protein chains [2**,33], neither of which has been discussed in this review. Protein structures contain information that is not obvious from a sequence, or else the folding problem would already be solved, so, ultimately, sequence-structure threading methods will probably recognize more distant relationships than sequence-only methods. Finally, there is a tendency to make a distinction between sequence comparison in which a structure is known and sequence comparison in which it is not, however, sequences still imply structures even if neither of them is known. Recognizing this, it may be possible to use ab initio structure-prediction information to systematically discriminate weak sequence alignments.

Acknowledgements

I would like to thank all my colleagues, particularly Steven Brenner, Alexey Murzin and Cyrus Chothia (with whom the NCSP database was developed), and Sean Eddy, Graeme Mitchison and Richard Durbin for the many useful discussions concerning the problems of sequence analysis. I am grateful to the Medical Research Council/Department of Trade and Industry/ZENECA LINK programme for financial support.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest


This paper describes the transformation of raw threading scores into approximate P-values (cf., P-values in BLAST [4]) by correcting for the effects of length, composition and the number of alternative alignments considered.

These two addresses access the same Web page and, together with [17*], give full details of the CASP2 prediction experiment, including the submission formats, evaluation criteria and results of their application to the >900 predictions received. The full results will be published in a special issue of Proteins in July 1997.

17. Protein structure prediction center on World Wide Web URL: • http://predictioncentre.llnl.gov/ See annotation [16*].


This entire issue of proteins is devoted to the first structure prediction experiment, CASP1, with articles by both the assessors and the predictors.

An example of a CASP2 target that is considered to be evolutionarily related to existing structures in the PDB.


An example of the use of a structural alignment database used to compare a number of sequence-alignment methods, with references to previous tests.