Computer Corner

Getting the most from PSI–BLAST

David T. Jones and Mark B. Swindells

Most biologists now conduct sequence searches as a matter of course. But how do we know that a relationship predicted by a homology search is a true, rather than false, hit with the same score? Many biologists design their own experiments with exquisite care yet still assume that results from programs with more than 20 adjustable parameters are 100% reliable. This article explains some of the key steps in getting the most from PSI–BLAST, one of the most popular and powerful homology search programs currently available.

Since its release less than five years ago, the new family of BLAST programs (which encompasses Gapped-BLAST and PSI–BLAST [1]) have gained an almost unprecedented acceptance with users and developers alike. There is no single reason for this, rather a combination of factors. First, the overall quality of the search results, particularly those produced by PSI–BLAST, put powerful methods into the hands of researchers who previously had to ask for expert help. Indeed, not only are distant evolutionary relationships predicted but also the confidence of each prediction (which is assessed using an *E*-value) becomes fairly easy to interpret after a few trial runs (although there are problems for the unwary). Second, the programs are freely available in both source code and binary versions, and attractive interfaces have been developed at the NCBI to encourage their use. It is no surprise that the paper describing these algorithms is already one of the world's most cited papers and looks likely to eclipse the success of earlier BLAST papers [2].

Given the popularity and ease of use of PSI–BLAST, along with recent updates to the original program, it seems timely to reiterate that, no matter how easy a program is to run, all bioinformatics tools are predictive and should be used with appropriate care and attention. After reading this and an earlier *Trends in Biochemical Sciences* article by Altschul and Koonin [3], you should have a good understanding of the practical issues with which you should be concerned.

Gapped-BLAST and PSI–BLAST

This article concentrates on protein–protein comparison through Gapped-BLAST and PSI–BLAST [1], although other flavours of the algorithm are also available from the NCBI, to which similar messages apply. Before going into detail, it is best to start with a simple description of each program and the associated tools. Despite the similarity in their names and the format of the results they return, Gapped-BLAST and PSI–BLAST should be considered separately by those unfamiliar with the field.

Gapped-BLAST is simply a logical development of the original BLAST algorithm which, similar to many algorithms, compared sequences using an amino acid substitution matrix such as BLOSUM62. The most visible improvement in Gapped-BLAST is, as the name suggests, that gaps are placed in the sequence alignments, resulting in added practicality. However, the sensitivity and speed of BLAST have also been dramatically improved. As an indication of the improvements made, recent comparisons have shown that the sensitivity of Gapped-BLAST is much closer to that achieved by a comprehensive search method (such as the well-known Smith–Waterman search [4]), even though it achieves the result in a fraction of the time [5]. The sensitivity of PSI–BLAST was also considerably higher [5–7].

PSI–BLAST (PSI stands for position-specific iterated) sits on top of the Gapped-BLAST program and will only work if Gapped-BLAST has already identified homologues of the query sequence. Without giving too much detail here, the key to the success of PSI–BLAST is its ability to assess the probable substitutions at each sequence position using the results of a previous Gapped-BLAST (or, by logical extension, PSI–BLAST) search. In practice, this is achieved by generating a profile (also known as a position-specific scoring matrix) from the results of a previous search and then applying it to the subsequent search. For example, with trypsin (a serine proteinase) as the query, Gapped-BLAST will identify related serine proteinase sequences and generate a profile in which position Ser195 of the active site prefers this residue type much more than other serine-containing positions. By knowing the requirements for each position, rather than treating all locations with the same amino acid type as equal, profiles are able, in principle, to find relationships that fall beyond traditional search methods while restricting the number of false positives predicted. Further refinement of the profile should also be possible until no new relationships are identified.

Example of success

With PSI–BLAST, it becomes possible to identify previous 'difficult' cases such as exfoliative toxin A from *Staphylococcus aureus* as a member of the trypsin-like serine proteinase superfamily, even though the sequence identity is only 16%. This protein was, in fact, a target for the *2nd Critical Assessment of Structure Prediction* experiment (CASP2), for which proteins likely to have their three-dimensional structures determined by the time of the meeting (held at the end of the experiment) had their structures predicted ahead of time by various methods. The results of these methods were then analysed to evaluate their strengths and weaknesses [8] (Fig. 1).

The success of PSI–BLAST rests on the ability to combine search results with robust statistics to build and apply profiles that avoid the sea of unrelated sequences. This idea is not new but it appears to work in a much more reliable and automated way in PSI–BLAST than any previous profile-based search tool. Of course, similar to most things in life, using these programs to investigate your particular protein sequence might not go smoothly. Assuming that researchers are interested in identifying distant relatives of a query sequence (if not, they should not be using PSI–BLAST!), they are dependent on the statistics successfully identifying these distant relationships while simultaneously avoiding false hits – yet, if care is not exercised, it is under these very conditions that statistics can be thrown off course. It is unlikely (although not impossible) that a

http://tibs.trends.com 0968-0004/02/$ – see front matter PII: S0968-0004(01)02039-4

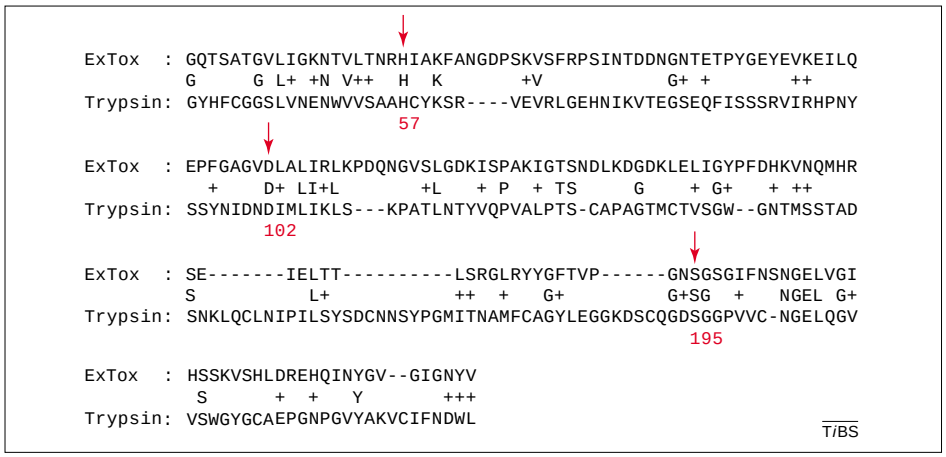

Fig. 1. Third-iteration PSI–BLAST result from querying exfoliative toxin (ExTox) A from *Staphylococcus aureus* (BAA97652.1) against the non-redundant database. His, Asp and Ser residues are indicated with arrows and numbered as for trypsin [anionic; complexed with the inhibitor benzamidine (1bit)]. The alignment has 15% identity (32/206) and the *E*-value = 6×10^{-21}. The threshold *E*-value for inclusion in the profile was 0.005 and the effective search space was 22 926 875 677.

human is going to be better at distinguishing distant homologues from false positives by eye than PSI–BLAST, so care should be taken when using web interfaces manually to influence hit lists for subsequent profile searches.

The Devil is in the detail

As with many recent sequence comparison methods, BLAST-based programs estimate the statistical significance of each alignment score through an *E*-value. This can be thought of as the number of times one would expect to get a false relationship with a similar score. The limit for safe searching is commonly taken as $E = 0.001$ (although, at the time of writing, the current default threshold for the NCBI BLAST Web server has been loosened, first to 0.002 and most recently to 0.005). So, for a particular result set, the worst case is a 1 in 1000 chance of hitting a false positive (1 in 200 for the web server). However, as we shall see, this is only half the story.

Database size

One of the less commonly known facts of database searching is that the *E*-value depends on database size. For example, if you perform the same database search each day at the NCBI website, and even if you pull back the same set of hits, their *E*-values will be less significant on the second day because the database will have increased in size. In practice, you are unlikely to see this tiny difference on a daily basis because it would only materially affect hits close to the threshold and, furthermore, *E*-values are only reported to a couple of decimal places. But the change is happening nevertheless. As long as the database being searched is comprehensive (such as the NCBI non-redundant database, with >700 000 sequences), an *E*-value of 0.001 is reasonable. However, if a small dataset is used, one must be careful not to over-interpret the results. For example, a search through all proteins of known structure (e.g. Protein Data Bank, Table 1) with $E = 0.001$ is akin to searching the NCBI NRDB with $E = \sim 0.1$ because the NRDB is nearly 100 times larger. As such, the two results are not comparable unless one corrects for database size.

The simplest way to compare searches of different databases is to keep a value known as the 'search space' constant by setting the '-Y' option. A suitable value can be observed at the end of any file that shows the results of searching NRDB (Fig. 1), although it is really the order of magnitude that is important rather than the exact number. With PSI–BLAST, it is important for users to realize that the *E*-value, and hence the database size, affects the number of hits qualifying for inclusion in subsequent profiles. As a result, the final predictions depend on the thresholds applied during all previous iterations, as well as on the final threshold.

Thus, as long as the variety of assumptions made to calculate *E*-values holds true, inclusion of false relationships should be a rare occurrence. However, if they do not hold true, trouble could await the user. For PSI–BLAST, the effect of holding false relationships in a result set can be particularly severe because those results are used to seed the following search. One of the basic assumptions behind *E*-value estimates is that each sequence has an average amino acid composition. In practice, this is rarely the case and the question becomes how different each is from the average and what effect that difference is likely to have on the search results. Another key assumption is that *E*-values for gapped alignments will have the same characteristics as ungapped alignments (for which the equations are most reliable). However, this discussion is beyond the scope of our article and the interested reader is referred to Refs [9,10].

Sequence complexity

'High complexity' proteins make full use of the 20 amino acids, whereas 'low complexity' proteins use more restricted sets at correspondingly higher frequencies. Predictably, searches are most powerful when only high complexity proteins are compared, because background scores are substantially lower than for real hits. In protein terms, globular domains generally have the highest complexity, whereas transmembrane regions and coiled coils have low to intermediate complexity. Beyond this, there are sequences that are so biased that even the untrained eye can detect them. For completeness, it is worth remembering that all rules exist to be broken and that there are even globular proteins whose sequences deviate considerably from the norm (such as hisactophilin, with nearly 25% histidine content).

The data used to compute the composition of an average protein by BLAST and PSI–BLAST are comparable to a typical globular domain. This means that high quality results should be obtained provided that only high complexity regions are compared. This is the reason for applying filtering (also known as masking) before searching, so that regions of sequence that are known to perform worse in database searches are avoided. A popular masking program (again from the NCBI) is SEG [11], which concentrates on identifying and removing regions of very low complexity. However, with this approach, other regions of low or intermediate complexity, such as transmembrane helices and coiled coils, will be left untouched.

What is the real error rate?

During work to annotate a bacterial genome, Huynen *et al.* [12] used a

Table 1. Internet sites with sequence tools

Web and FTP sites	Description
http://www.ncbi.nlm.nih.gov/BLAST	Simple access to the complete family of BLAST programs. These can compare any combination of protein and DNA through six-frame translation of DNA. Gapped-BLAST and PSI–BLAST can only compare a protein sequence against a protein database.
http://www.rcsb.org/pdb	The Protein Data Bank – the key source of experimental 3D structure data.
http://www.ensembl.org	Annotation of the human genome.
http://genome.ucsc.edu	Golden Path annotation of human genome.
ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools	Source code for compilation.
ftp://ftp.ncbi.nlm.nih.gov/blast/executables	Binary versions for a variety of machine types.
ftp://ftp.ncbi.nlm.nih.gov/pub/seg/seg	SEG masking software for stand-alone usage.
ftp://bioinf.ucl.ac.uk/pub/pfilt	*pFilt* masking software for stand-alone usage. Identifies transmembrane helices, coiled coils and 12-residue windows with low complexity. Also applies two composition filters.

[a]Low-complexity regions are identified as 12-residue windows in which the average residue occurrence is >4. Local and global composition filters masked out residues >4 and >5 standard deviations, respectively, from the mean values in Swissprot.

threshold *E*-value of 0.001 to assign protein folds to open reading frames, on the assumption that this was a safe threshold. As part of this work, they tested these assumptions using PSI–BLAST, a set of known relationships between proteins of known three-dimensional structure and the NRDB database from the NCBI. They found that the actual false positive rate for these sequences was ~1.8%. This is effectively 18 times higher than suggested by the theoretical *E*-value and they were concerned that even this might be a low estimate given the absence of complications (transmembrane, coiled coil and other non-globular regions) in typical PDB sequences.

So, how realistic are these estimates of error? To shed light on this, we conducted some simple tests to obtain worst-case estimates using data for which all standard precautions (such as masking) had already been performed. We took 1349 human protein sequences from SWISS-PROT and searched them against a database that contained all masked entries from NRDB95, plus 'reverse-complemented' versions of each sequence to act as decoys. [Reverse-complemented sequences were generated by writing out each masked sequence back to front (ACDEF becomes FEDCA), and then mutating each residue to what it would most probably have been on the opposite strand of DNA and finally masking once again.] Reverse-complemented sequences do not exist in real life but retain physicochemical properties (including the periodicity) comparable to real sequences (as opposed to randomized versions of a sequence). So, if we identify a relationship to a reverse-complemented sequence during a search, we know that it must be an error rather than a distant homologue.

To investigate the errors that arise from simple masking methods leaving many regions of lower, but not low, complexity unmasked, two distinct databases were generated. One was a SEG-filtered version of NRDB95 and the other was a more stringently masked database in which all predicted transmembrane regions [13], coiled coils [14], lower complexity regions and residues with notably higher frequencies than the average, were masked out. For convenience, we shall refer to the approach that masks sequences in this more stringent manner as *pFilt* (Table 1) but many of these individual algorithms (such as COILS [14]) are available independently from other sites. Finally, in contrast to other approaches [15], we decided to restrict the opportunity for false positives further by masking both query and database sequences during the searches.

Before going any further, it is important to point out that masking is not the only way of tackling this problem. From version 2.2, PSI–BLAST has been able to down-weight the significance of hits between sequences that deviate from the expected composition. Initial reports suggest that this is successful for modest variations and all the following searches were run with this option switched on.

Every search was given the opportunity to run for five iterations, although not all of them made it that far. The results were surprising. With the SEG filtering (and composition-based statistics), 13.2% of the searches had picked up a false positive by the fifth iteration. For searches with *pFilt* filtering (and composition-based statistics), the observed error was slightly better at 4.8%, but both values are considerably higher than any of the statistical estimates suggest. On a per-iteration basis, the results show further reason for caution (Fig. 2), with many searches bringing back false hits even in the first search (which is essentially Gapped-BLAST). Clearly, deviations from the ideal sequence composition have a considerable effect and none of the available methods reliably handle all such cases.

Although the above might surprise those unfamiliar with the details of such tools, a more positive message can be provided overall. First, even with the fairly radical construction of false sequences for the dataset, 86.8% of searches with SEG and 95.2% of those with *pFilt* had no empirical evidence of false positives. Second, the choice of human target sequences and the large number of decoy sequences in our database might represent a worst-case view. In reality, the numbers for most searches will be between those presented here and the much more conservative results from earlier papers. Third, the errors will be at *E*-values close to the threshold (in our example, 0.001) rather than those for which there is considerable statistical confidence. Finally, there are many other important ways to add confidence to a prediction. For example, known active-site residues or nucleotide-binding sites would be expected to have strong conservation even when sequences have <10% identity when correctly aligned. A good example of this is the prediction that histidine kinases have an ATP-binding domain similar to DNA gyrase and HSP-90, in which a DxGxG motif responsible for binding phosphate in each protein is correctly aligned by PSI–BLAST [16,17]. For proteins lacking such clear signals, other information such as co-location on a pathway might help strengthen the case for homology.

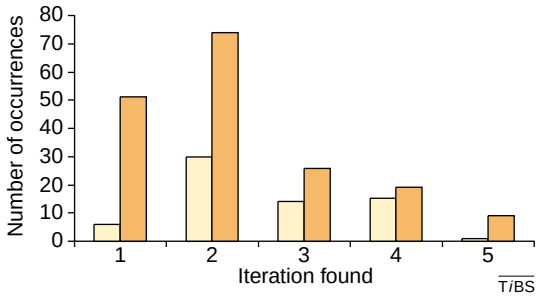

Fig. 2. Number of searches that pick up a false positive in the *pFilt* (yellow) and SEG (orange) masked non-redundant database (NRDB) 95 database. For these searches, a conservative approach was adopted in which both the query and all database sequences were masked. All searches used version 2.2 of blastpgp with the composition-based statistics option switched on. There were 737 064 sequences in the database and the effective database length was 221 466 393.

Precalculated data on the Web

When conducting your own searches, successfully navigating the potential pitfalls is down to you. However, many genome resources have recently become available that often provide precalculated data. Well-known academic examples of this include Ensembl and the Golden Path site at UCSC (Table 1), which are revolutionizing the ways that non-specialists gain access to genome data (there are also many other less well publicized sites). With external pressure to keep these sites up-to-date and relevant, it is frequently unclear, even to an expert, how to reproduce a search and confirm a set of hits. The end result for the user is that an unknown proportion of sequences will be incorrectly annotated and, more worryingly, that these might then be used to incorrectly annotate other proteins. It is possible that a great deal of wasted time and effort in some future project might one day be traced back to a single false positive match from a database search.

Take home messages

At the end of the day, the user of these algorithms needs to cast a critical eye over search results and to draw conclusions using their own expertise. Running PSI–BLAST and other computer programs might be gloriously easy but, at the end of the day, the results must be interpreted as with any other experiment (albeit, in this case, an *in silico* experiment). Experimentalists are all too aware of the need to treat experimental results with statistical caution, but are often willing to assume that the results of a database search are a certainty. PSI–BLAST has ~30 available options, of which most people will only ever use about four. Each option can be thought of as resembling the control parameters of a laboratory experiment. The database and its preparation might represent a cell line and its preparation, and the *E*-value might be analogous to the temperature at which an experiment is conducted. As with the design of any experiment, the specific choice of parameters is up to the user, even though default options that are believed to be safe are available. Simply stating that a match has been found using PSI–BLAST with an *E*-value of <0.001 is no better than stating the activity of an enzyme without describing the precise experimental conditions under which the activity was measured.

Naturally, the vast potential of discovery outweighs any of the concerns addressed here, so we encourage non-experts to use the methods more rather than less. Get familiar with the algorithms by using cases for which you already know the answer and see what else is picked out. You will probably find a great deal from such mining of genome data, but watch out for the odd nugget of fool's gold that might also come your way.

References

1 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
2 Russo, E. and Bunk, S. (1999) Hot papers in bioinformatics. *Scientist* 13, 15
3 Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI–BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447
4 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197
5 Salamov, A.A. *et al.* (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* 12, 95–100
6 Park, J. *et al.* (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284, 1201–1210
7 Muller, A. *et al.* (1999) Benchmarking PSI–BLAST in genome annotation. *J. Mol. Biol.* 293, 1257–1271
8 Vath, G.M. *et al.* (1997) The structure of the superantigen exfoliative toxin A suggests a novel regulation as a serine protease. *Biochemistry* 36, 1559–1566
9 Mott, R. (2000) Accurate formula for *P*-values of gapped local sequence and profile alignments. *J. Mol. Biol.* 300, 649–659
10 Mott, R. (1999) Local sequence alignments with monotonic gap penalties. *Bioinformatics* 15, 455–462
11 Wootton, J.C. and Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149–163
12 Huynen, M. *et al.* (1998) Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280, 323–326
13 Jones, D.T. *et al.* (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33, 3038–3049
14 Lupas, A. *et al.* (1991) Predicting coiled coils from protein sequences. *Science* 252, 1162–1164
15 Schaffer, A.A. *et al.* (2001) Improving the accuracy of PSI–BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 29, 2994–3005
16 Mushegian, A.R. *et al.* (1997) Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5831–5836
17 Tanaka, T. *et al.* (1998) NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ. *Nature* 396, 88–92

David T. Jones
Bioinformatics Unit, Dept Computer Science, University College London, Gower St, London, UK WC1E 6BT.
e-mail: d.jones@cs.ucl.ac.uk

Mark B. Swindells
Inpharmatica Ltd, 60 Charlotte St, London, UK W1T 2NU.
e-mail: swintech@inpharmatica.co.uk

Corrigendum

In the January issue, we published an article 'Structural genomics and signaling domains by James H. Hurley *et al.* (*TiBS* 27, 48–53). In Table 1, the entry for the PB1 domain under 'Function' should read: 'Protein–protein interaction with 'PC' motif in small G-protein guanine nucleotide exchange factors and others'. In addition, under the heading 'PB1', it is incorrectly stated that the PB1 motif interacts with the small G protein Cdc42p. This should read, instead, that the PB1 motif interacts with the small GTPase guanine nucleotide exchange factor Cdc24p.