TIBS 23 – *NOVEMBER 1998*

Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases

Protein sequence database searches have become essential to much research in molecular biology and are central to genomic analysis. The need for fast search methods arose from large-scale sequencing and was met by the development of the popular FASTA[1] and BLAST programs[2–4]. Both are based on the Smith–Waterman algorithm[5], and identify and align regions of similarity between two sequences. Rapid exact-match methods locate regions that are most likely to be related, and detailed analysis is applied only to these regions. This heuristic strategy allows FASTA and BLAST to run 10–100 times faster than Smith–Waterman, at the cost of occasionally overlooking a significant similarity.

The original BLAST programs[2] could find only local alignments without gaps, but were able to apply a rigorous test of their statistical significance[6–8]. Specifically, the *E* value reported for a given result represents the number of alignments, with an equivalent or greater score, that would be expected to occur purely by chance. Computer simulations suggest that the same statistical theory applies to alignments that are allowed to have gaps[3,9,10], although there is as yet no mathematical proof for this. By estimating the relevant statistical parameters, FASTA and recent versions of BLAST that permit gapped alignments[3,4] can therefore accurately assess significance[10]. Extensive benchmarking[4,11] demonstrates that the sensitivities of these programs approach that of the rigorous Smith–Waterman algorithm.

Sequence profiles

Many functionally and evolutionarily important protein similarities are recognizable only through comparison of three-dimensional structures[12,13]. When such structures are not available, patterns of conservation identified from the alignment of related sequences can aid the recognition of distant similarities. There is a large literature on the definition and construction of these patterns, which have been variously called motifs, profiles, position-specific score matrices, and Hidden Markov Models[4,14–21]. In essence, for each position in the derived pattern, every amino acid is assigned a score. If a residue is highly conserved at a particular position, that residue is assigned a high positive score, and others are assigned high negative scores. At weakly conserved positions, all residues receive scores near zero. Position-specific scores can also be assigned to potential insertions and deletions[15,20,21].

The power of profile methods can be further enhanced through iteration of the search procedure[4,17–19]. After a profile is run against a database, new similar sequences can be detected. A new multiple alignment, which includes these sequences, can be constructed, a new profile abstracted, and a new database search performed. The procedure can be iterated as often as desired or until convergence, when no new statistically significant sequences are detected.

The design of PSI-BLAST

Iterated profile search methods have led to biologically important observations but, for many years, were quite slow and generally did not provide precise means for evaluating the significance of their results. This limited their utility for systematic mining of the protein databases. The principal design goals in developing the Position-Specific Iterated BLAST (PSI-BLAST) program[4] were speed, simplicity and automatic operation. The procedure PSI-BLAST uses can be summarized in five steps.

(1) PSI-BLAST takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST program.

(2) The program constructs a multiple alignment, and then a profile, from any significant local alignments found. The original query sequence serves as a template for the multiple alignment and profile, whose lengths are identical to that of the query. Different numbers of sequences can be aligned in different template positions.

(3) The program compares the profile to the protein database, again seeking local alignments. After a few minor modifications, the BLAST algorithm can be used for this directly.

(4) PSI-BLAST estimates the statistical significance of the local alignments found. Because profile substitution scores are constructed to a fixed scale[6], and gap scores remain independent of position, the statistical theory and parameters for gapped BLAST alignments[3] remain applicable to profile alignments[4].

(5) Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence.

Profile-alignment statistics allow PSI-BLAST to proceed as a natural extension of BLAST; the results produced in iterative search steps are comparable to those produced from the first pass. Unlike most profile-based search methods, PSI-BLAST runs as one program, starting with a single protein sequence, and the intermediate steps of multiple alignment and profile construction are invisible to the user.

A PSI-BLAST example

PSI-BLAST uncovers many protein relationships missed by single-pass database-search methods and has identified relationships that were previously detectable only from information about the three-dimensional structure of the proteins[4,22,23]. Here, we illustrate how to operate PSI-BLAST by using a comparison of proteins from thermophilic archaea and bacteria as an example[24]. We employed the WWW version of PSI-BLAST, which can be accessed at http://www.ncbi.nlm.nih.gov/BLAST.

A search is initiated by pasting one's query sequence (here, the uncharacterized protein MJ0414 from *Methanococcus jannaschii*[25]) into the PSI-BLAST Web page (Fig. 1a). At this point, one can immediately press the 'Submit Query' button or, instead, first tailor the search. Specifically, substitution and gap costs can be selected, and the cutoff *E* value that PSI-BLAST uses when constructing a profile for the next iteration can be altered. The default *E* value is rather conservative (0.001); in this example, we changed it to 0.01.

Figure 1b shows the results of the program's initial gapped BLAST search; the only significant hits are very strong ones to the query sequence itself, and to uncharacterized proteins from three other archaea and the thermophilic bacteria *Aquifex aeolicus*. However, iterating the search by using the derived profile

0968–0004/98/$ – Published by Elsevier Science. PII: S0968-0004(98)01298-5

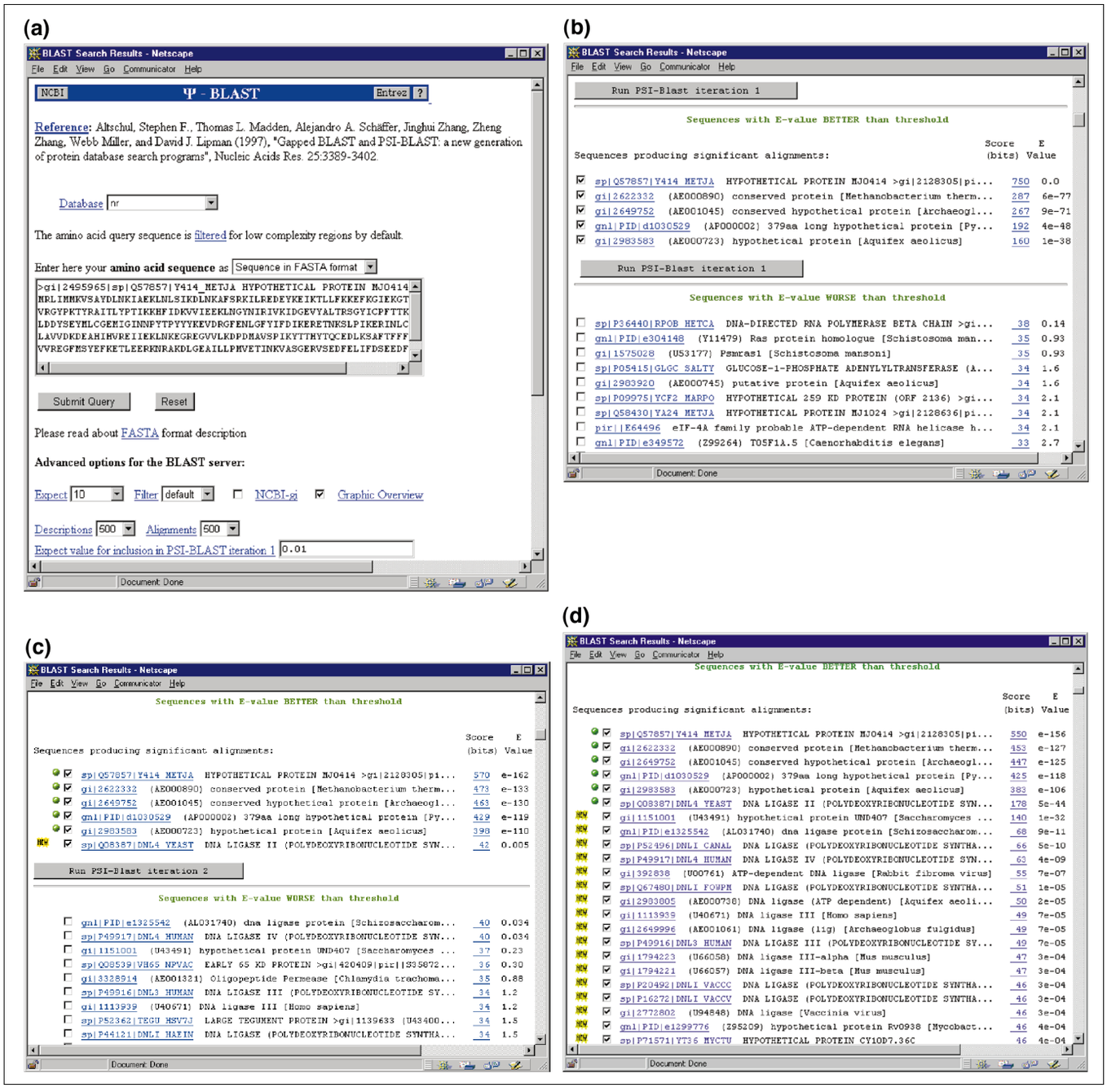

Figure 1

Progress of a PSI-BLAST search. **(a)** The WWW page for launching a PSI-BLAST search. **(b)** The initial gapped BLAST search results. **(c)** The first profile-based iteration of PSI-BLAST. **(d)** The second profile-based iteration of PSI-BLAST. All images are taken directly from WWW PSI-BLAST output; only a section of each page is shown. After the initial gapped BLAST search, sequences that for the first time have *E* value better than the cutoff are marked 'NEW'.

uncovers yeast DNA ligase II (Ref. 26); the *E* value is 0.005, which is moderately significant (Fig. 1c). Because we have used 0.01 as the cutoff *E* value for recruitment of alignments into successive profiles, the ligase sequence is included at this stage; consequently, the next iteration produces many highly significant alignments that involve other DNA ligases (Fig. 1d).

How do we interpret these results? Once a single sequence from a highly conserved family (here, the DNA ligases) is used in constructing a profile, the rest of the family will almost certainly be retrieved (and have *E* values that are highly significant) in subsequent iterations. Impressive *E* values for sequences retrieved in later iterations depend upon the validity of earlier inferences and therefore should not be taken as automatic proof of homology. In the example presented here, the best evidence for a possible relationship between the thermophile protein family and DNA ligases is the alignment produced in the first PSI-BLAST iteration ($E = 0.005$). This should be taken as a hint that requires corroboration. Fortunately, the PSI-BLAST alignment of our uncharacterized protein and yeast DNA ligase here provides such corroboration (Fig. 2). The best-conserved portions of the alignment correspond perfectly to the set of conserved motifs identified in ATP-dependent DNA ligases[27], including the catalytic lysine residue that forms a covalent adduct with AMP (Fig. 2). Although the *E* values

```
sp|Q08387|DNL4_YEAST DNA LIGASE II (POLYDEOXYRIBONUCLEOTIDE SYNTHASE (ATP)) (DNA LIGASE
          IV HOMOLOG) >gi|2131241|pir||S66870 DNL4 protein - yeast
          (Saccharomyces cerevisiae) >gi|1420096|gnl|PID|e252317
          (Z74913) ORF YOR005c [Saccharomyces cerevisiae]
          Length = 944

 Score =  178 bits (447), Expect = 5e-44
 Identities = 42/209 (20%), Positives = 76/209 (36%), Gaps = 28/209 (13%)


                                                      *
Query: 61  VIFLNDNLDVVRGY---PKTYRAITL-YPTIKKHFIDKVVIEEKLNGYNIRI--VKIDGE 114
           V   +D+L +  G+   P+  + + L Y  I +   D  ++EEK++G  I++  +
Sbjct: 239 VRLKDDDLSIKVGFAFAPQLAKKVNLSYEKICRTLHDDFLVEEKMDGERIQVHYMNYGES 298


Query: 115 VYALTRSG--YICPFTTKKVKKFLN--LEILDDYSEYMLCGEMI--GINNPYTPYYYKEV 168
           +   +R G  Y   +        ++  L   D   E +L GEM+           +
Sbjct: 299 IKFFSRRGIDYTYLYGASLSSGTISQHLRFTDSVKECVLDGEMVTFDAKRRVILPFGLVK 358


Query: 169 DRGFENLGF------------YIFDIKERETNK--SLPIKERINLCEKYNLPYVKPLAVV 214
             E L F             +FD+          LP+ +R          P    + +V
Sbjct: 359 GSAKEALSFNSINNVDFHPLYMVFDLLYLNGTSLTPLPLHQRKQYLNSILSPLKNIVEIV 418


Query: 215 DKDEAH--IHVREIIEKLNKEGREGVVLK 241
                +    +++ +E     G EGVVLK
Sbjct: 419 RSSRCYGVESIKKSLEVAISLGSEGVVLK 447
```

Figure 2

Alignment of a predicted novel archaeal DNA ligase with yeast DNA ligase II. The alignment is taken directly from the second iteration shown in Fig. 1d. The motifs conserved in DNA ligases[27] are shown against a colored background, and the catalytic lysine is indicated by an asterisk. Header information generated by PSI-BLAST describes the database sequence involved in the alignment and provides summary statistics for the alignment. Query lines report residues from the archaeal sequence provided as input to PSI-BLAST; subject (Sbjct) lines report residues from yeast DNA ligase II. Conserved residues are echoed on lines between the two sequences; + indicates Sbjct residues that have received a positive profile score.

reported for the other ligase alignments do nothing to confirm the relationship, the alignments themselves conform to the conservation pattern shown in Fig. 2. Thus, we can conclude that the uncharacterized archaeal and *A. aeolicus* proteins probably make up a new family of ATP-dependent DNA ligases. This finding is interesting both in itself and in the context of the apparently massive horizontal gene exchange between thermophilic archaea and bacteria[24].

Notes on using PSI-BLAST

The WWW version of PSI-BLAST requires the user to decide after each iteration whether to continue. In some respects, this is a limitation, but it has the advantage that the user can hand-pick the sequences used for each profile construction, regardless of *E* value, by checking boxes next to the sequences' descriptions (Fig. 1). A stand-alone version of PSI-BLAST (obtainable from NCBI at ftp://ncbi.nlm.nih.gov/blast/executables/) allows the user to run the program for a chosen number of iterations or until convergence; it also allows the user to save the profile produced and use it subsequently to search another database.

PSI-BLAST is a powerful tool, but its use requires caution. The sources of error are the same as for standard BLAST but are easily amplified by iteration. The major source of deceptive alignments is the presence within proteins of regions that have highly biased amino acid compositions[28]. If such a region is included during production of a profile, otherwise unrelated sequences containing similarly biased regions will probably creep in during subsequent iterations, rendering the search nearly worthless. PSI-BLAST filters out biased regions of query sequences by default, using the SEG program[28]. Because the SEG parameters have been set to avoid masking potentially important regions, some bias can persist; PSI-BLAST can thus still generate compositionally rooted artefacts. These cases usually can be identified by inspection – especially when sequences that have a known bias, such as myosins or collagens, are retrieved. SEG (obtainable at ftp://ncbi.nlm.nih.gov/pub/seg/seg/) can be used with parameters that eliminate nearly all biased regions[28], and the user can apply locally other filtering procedures, such as COILS[29] (which detects coiled-coil regions), before submitting the appropriately masked sequence to PSI-BLAST.

Conclusion

PSI-BLAST alters the database-searching landscape. Many important but subtle relationships that previously were detectable only by the fairly laborious application of several methods, or by structural comparison, can now be identified more easily. Although a PSI-BLAST search is easy to launch, the program increases rather than removes the need for expertise, because there is more to interpret.

PSI-BLAST has room for improvement. The profile-construction process is fairly naive, and increased sophistication might increase the power of the profiles produced. The current simple procedure for generating multiple alignments could also be improved. PSI-BLAST can be modified to accept appropriately structured multiple alignments as input from other programs[30].

In addition, one can construct a curated library of PSI-BLAST compatible profiles and modify the program in order to compare a query directly to this library. We hope that future refinements that perhaps incorporate some of these ideas will further enhance our ability to make sense of protein sequences.

Acknowledgements

We thank the developers of PSI-BLAST, who include D. J. Lipman, T. L. Madden, W. Miller, A. A. Schäffer, J. Zhang and Z. Zhang. We also thank L. Aravind for his collaboration on the application of PSI-BLAST to the detection of subtle relationships among proteins.

References

1 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
2 Altschul, S. F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
3 Altschul, S. F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
4 Altschul, S. F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
5 Smith, T. F. and Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197
6 Karlin, S. and Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
7 Karlin, S. and Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 5873–5877
8 Dembo, A., Karlin, S. and Zeitouni, O. (1994) *Ann. Prob.* 22, 2022–2039
9 Waterman, M. S. and Vingron, M. (1994) *Stat. Sci.* 9, 367–381
10 Pearson, W. R. (1998) *J. Mol. Biol.* 276, 71–84
11 Pearson, W. R. (1995) *Protein Sci.* 4, 1145–1160
12 Holm, L. and Sander, C. (1997) *Structure* 5, 165–171
13 Brenner, S. E., Chothia, C. and Hubbard, T. J. P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078
14 Schneider, T. S., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* 188, 415–431
15 Gribskov, M., McLachlin, A. D. and Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. U. S. A.* 84, 4355–4358
16 Staden, R. (1989) *Comput. Appl. Biosci.* 4, 53–60
17 Gribskov, M. (1992) *Gene* 119, 107–111
18 Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 12091–12095
19 Yi, T-M. and Lander, E. S. (1994) *Protein Sci.* 3, 1315–1328
20 Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996) *Comput. Chem.* 20, 3–23
21 Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press
22 Mushegian, A. R. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 5831–5836
23 Huynen, M. *et al.* (1998) *J. Mol. Biol.* 280, 323–326
24 Aravind, L. *et al. Trends Genet.* (in press)
25 Bult, C. J. *et al.* (1996) *Science* 273, 1058–1073
26 Sterky, F., Holmberg, A., Pettersson, B. and Uhlen, M. (1996) *Yeast* 12, 1091–1095
27 Shuman, S. and Schwer, B. (1995) *Mol. Microbiol.* 17, 405–410
28 Wootton, J. C. and Federhen, S. (1996) *Methods Enzymol.* 266, 554–571
29 Lupas, A. (1996) *Methods Enzymol.* 266, 513–525
30 Zhang, Z. *et al.* (1998) *Nucleic Acids Res.* 26, 3986–3990

STEPHEN F. ALTSCHUL AND EUGENE V. KOONIN

NCBI, National Library of Medicine, NIH, Bethesda, MD 20894, USA.
Email: altschul@ncbi.nlm.nih.gov

Eukaryotic rRNA methylation: the calm before the Sno storm

Much excitement has arisen from the discovery that small nucleolar RNA molecules (snoRNAs) function as guides for post-synthetic modifications of eukaryotic rRNA[1–6]. The title of one review, 'Sno storm in the nucleolus: new roles for myriad small RNPs'[6], abundantly conveys the excitement. A prerequisite for this work was the accurate mapping of the numerous modified nucleosides within eukaryotic rRNA. My colleagues and I had the good fortune to contribute to this earlier work, especially the mapping of the RNA methyl groups. Here, I look back on the mapping work, which, in retrospect, was a mini golden era of 'calm before the Sno storm'.

Albert Einstein College of Medicine, 1967–1969

In January 1967, I arrived as a postdoc in Jim Darnell's laboratory at the Albert Einstein College of Medicine, New York. I had completed my PhD, on ribosome-catalyzed peptidyl transfer, under Robin Monro at the Medical Research Council Laboratory for Molecular Biology, in Cambridge. Robin had been a postdoc with Fritz Lipmann and was an early leader in the characterization of the partial reactions of protein synthesis. The subject had interested me greatly, and my contribution made a mark[7], but I had also become interested in RNA biosynthesis in animal cells through reading Jim Darnell's work; when Jim offered me a postdoctoral position, I accepted enthusiastically.

Jon Warner – who was already well known for discovering polysomes during his PhD with Alex Rich at MIT – was closely associated with Jim's research group. Because of my background in ribosomes, we agreed that I should work jointly with Jon and Jim. This was an ideal arrangement and was further enhanced by the stimulating overall environment. Harry Eagle (of Eagle's medium) had been influential in attracting several leading cell biologists to 'Einstein', and there were excellent interactions between research groups. A wide range of topics was amenable to the (then) illuminating methods of radioactive labelling, cell fractionation, sucrose-gradient centrifugation and polyacrylamide-gel electrophoresis, and many features of cellular and viral composition and biosynthesis were being revealed.

Sheldon Penman had developed a cell-fractionation procedure that allowed purification of the nucleolar precursors of rRNA (Ref. 8). When I arrived, Jon Warner was developing this procedure further for isolation and characterization of nucleolar preribosomal RNP particles[9]. In a separate study, Jim's group discovered 5.8S rRNA (initially called 7S rRNA)[10]. The discovery exemplified Jim's perceptive eye; he had noticed that 28S rRNA sedimented slightly more slowly after extraction with hot phenol than after extraction with cold phenol. He guessed that a small piece of noncovalently attached RNA might be released by heat treatment, and he and colleagues sought and characterized this RNA[10]. Jim and Jon were both most interested in regulation, however. One approach to regulation was to observe the effects of depriving cells of a nutritionally essential amino acid – and thereby slowing down protein synthesis to turnover levels – on ribosome formation. I used valine deprivation as a model[11]; withholding this essential amino acid led to a reversible slowing down of pre-rRNA synthesis and processing, but not to complete cessation.

 PII: S0968-0004(98)01304-8