Here we present PSI-Search, an iterated profile search application for identifying distantly related protein sequences. PSI-Search is similar to PSI-BLAST, but substitutes a rigorous Smith–Waterman local alignment (Smith and Waterman, 1981) search strategy (SSEARCH, Pearson, 1991) to produce optimal local alignment scores from the profile PSSM. PSI-Search includes an optional alignment boundary-masking procedure that reduces HOE errors in the PSSM profile. SCANPS (Walsh et al., 2008) implements a similar iterative search strategy using Smith–Waterman alignments; however, it does not currently scale to large protein databases and does not include boundary masking.

2 METHODS

In PSI-Search, library searches are performed with search, selected hit sequences from the result are processed with an automated sequence boundary-masking procedure, and PSSM profiles are built using blastpgp. The PSI-Search iteration workflow (Fig. 1a) iterates through search and alignment/PSSM construction steps:

(1) The initial iteration is a normal search run with a sequence input.
(2) During the second iteration, aligned sequences with statistically significant scores from the previous search are retrieved using fastacmd. Details of the alignment boundaries are stored; sequence regions outside the boundaries are masked with 'X's to remove potential HOE regions; masked sequences are formatted into BLAST indexes using formatdb with an additional 10 000 random protein sequences created by makeprotseq (Race et al., 2000); and a PSSM checkpoint constructed with a blastpgp search; finally search is run with the input sequence, using the generated PSSM, to complete the second iteration and output alignments.
(3) Further iterations repeat Step (2). To avoid HOE, PSI-Search always uses the alignment boundary information from the first significant alignment in which a library sequence appears. Thus, if the first significant alignment with a library sequence aligned residues 25–125 at iteration i, later alignment boundaries at iteration i + 1 and beyond are ignored; only the initially aligned region (25–125) is used to form the PSSM.

3 RESULTS

Five iterative search strategies—PSI-BLAST (standard and HOE-reduced), PSI-Search (standard and HOE-reduced) and JackHMMER (Eddy, 2011)—were evaluated on the ReProtDom benchmark database using an E-value threshold of 0.001. JackHMMER is another iterative search tool that uses Hidden Markov Models.
Weighted true-positives and false-positives are calculated as strengths. PSI-BLAST, HOE-reduced PSI-BLAST, HOE-reduced PSI-Search are more specific than PSI-BLAST, HOE-reduced PSI-Search is more selective, but slightly less sensitive, than JackHMMER. JackHMMER is the most sensitive tool, but HOE-reduced PSI-Search is the most selective iterative tool.

ACKNOWLEDGEMENTS

Funding: This research was supported by the National Library of Medicine (NIH grant LM04969 to W.R.P.), European Molecular Biology Laboratory, and European Commission Research Infrastructures of the FP7 [grant agreement number 226073 SLING (Integrating Activity)].

Conflict of Interest: none declared.

REFERENCES

Agrawal,A. and Huang,X. (2009) PSI-BLAST_PairwiseStatSig: reordering PSI-BLAST hits using pairwise statistical significance. Bioinformatics, 25, 1082–1083.

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.

Altschul,S.F. et al. (2005) Protein database searches using compositionally adjusted substitution matrices. FEBS J., 272, 5101–5109.

Altschul,S.F. et al. (2009) PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res., 37, 815–824.

Bhadbhade,R. et al. (2006) Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains. Nucleic Acids Res., 34, W143–W146.

Eddy,S.R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol., 7, e1002195.

Gonzalez,M.W. and Pearson,W.R. (2010a) Homologous over-extension: a challenge for iterative similarity searches. Nucleic Acids Res., 38, 2177–2189.

Gonzalez,M.W. and Pearson,W.R. (2010b) RefProtDom: a protein database with improved domain boundaries and homology relationships. Bioinformatics, 26, 2361–2362.

Johnson,L.S. et al. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics, 11, 431.

Kim,B.H. et al. (2010) HangOut: generating clean PSI-BLAST profiles for domains with long insertions. Bioinformatics, 26, 1564–1565.

Li,Y. et al. (2011) A performance enhanced PSI-BLAST based on hybrid alignment. Bioinformatics, 27, 31–37.

Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. J. Mol. Biol., 226, 133–148.

Przybylski,D. and Rost,B. (2008) Powerful fusion: PSI-BLAST and consensus sequences. Bioinformatics, 24, 1987–1993.

Rice,P. et al. (2000) EMBOSS: the European molecular biology open software suite. Trends Genet., 16, 276–277.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol., 147, 195–197.

Stepourenov,S. et al. (2008) The effectiveness of position- and composition-specific gap costs for protein similarity searches. Bioinformatics, 24, i15–i23.

Walsh,T.P. et al. (2008) SCANPS: a web server for iterative protein sequence database searching by dynamic programming, with display in a hierarchical SCOP browser. Nucleic Acids Res., 36, W25–W29.

Fig. 1. (a) HOE-reduced PSI-Search iteration workflow. (b) Fraction of true-positives versus false-positives found by PSI-BLAST, PSI-BLAST HOE-reduced, PSI-Search, PSI-Search HOE-reduced, and JackHMMER. Weighted true-positives and false-positives are calculated as strengths. At iteration 5 and totalf, the total number of homologs for query f in the RefProtDom benchmark database. Alignments containing HOEs with >50% of the alignment outside the homologous boundary are counted as both true positive and false positives (TPs) and false positives (FPs, Fig. 1b). At 50% coverage, PSI-Search reduces the weighted fraction of errors from 4.5% (PSI-BLAST) to 2.9% (PSI-Search). Reducing HOE improves sensitivity even more, to 1.7% for HOE-reduced PSI-BLAST and 0.5% for HOE-reduced PSI-Search. At 50% coverage, JackHMMER performs very well using its statistical alignment envelope, producing only 1% weighted FPs, but its selectivity is worse than PSI-Search or HOE-reduced PSI-Search at 60% and 75% coverage. Overall, HOE-reduced PSI-Search is 9-fold more selective than PSI-BLAST. At the end of iteration 5, 78.3, 79.5, 77.3, 81.8 and 82.5% of weighted homologs are found by PSI-BLAST, PSI-Search, HOE-reduced PSI-BLAST, HOE-reduced PSI-Search and JackHMMER respectively. Thus, (i) HOE-reduction greatly improves search selectivity with a small cost in sensitivity in both PSI-BLAST and PSI-Search; (ii) Both PSI-Search and JackHMMER are more sensitive and selective than PSI-BLAST; (iii) HOE-reduced PSI-Search is more selective, but slightly less sensitive, than JackHMMER. JackHMMER is the most sensitive tool, but HOE-reduced PSI-Search is the most selective iterative tool.