BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 18 2008, pages 1987–1993
doi:10.1093/bioinformatics/btn384

Sequence analysis

Powerful fusion: PSI-BLAST and consensus sequences

Dariusz Przybylski[1,2,*] and Burkhard Rost[1,3]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, [2]Broad Institute of MIT and Harvard University, 320 Charles St., Cambridge, MA 02141 and [3]Columbia University Center for Computational Biology and Bioinformatics (C2B2), NorthEast Structural Genomics Consortium (NESG), New York Consortium on Membrane Protein Structure (NYCOMPS), 1130 St Nicholas Ave. Rm. 802, New York, NY 10032, USA

Received on January 30, 2008; revised on July 6, 2008; accepted on July 22, 2008

Advance Access publication August 4, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: A typical PSI-BLAST search consists of iterative scanning and alignment of a large sequence database during which a scoring profile is progressively built and refined. Such a profile can also be stored and used to search against a different database of sequences. Using it to search against a database of consensus rather than native sequences is a simple add-on that boosts performance surprisingly well. The improvement comes at a price: we hypothesized that random alignment score statistics would differ between native and consensus sequences. Thus PSI-BLAST-based profile searches against consensus sequences might incorrectly estimate statistical significance of alignment scores. In addition, iterative searches against consensus databases may fail. Here, we addressed these challenges in an attempt to harness the full power of the combination of PSI-BLAST and consensus sequences.

Results: We studied alignment score statistics for various types of consensus sequences. In general, the score distribution parameters of profile-based consensus sequence alignments differed significantly from those derived for the native sequences. PSI-BLAST partially compensated for the parameter variation. We have identified a protocol for building specialized consensus sequences that significantly improved search sensitivity and preserved score distribution parameters. As a result, PSI-BLAST profiles can be used to search specialized consensus sequences without sacrificing estimates of statistical significance. We also provided results indicating that iterative PSI-BLAST searches against consensus sequences could work very well. Overall, we showed how a very popular and effective method could be used to identify significantly more relevant similarities among protein sequences.

Availability: http://www.rostlab.org/services/consensus/

Contact: dariusz@mit.edu

1 INTRODUCTION

PSI-BLAST achieves a remarkable compromise between speed and quality. Ideally, an alignment method should accurately identify related sequences in today's rapidly growing databases within the shortest possible time. While we want to simultaneously optimize speed and reliability, in practice there is a tradeoff: very accurate alignment methods are relatively slow (e.g. profile–profile alignment algorithms), while very fast methods are far less sensitive than we might wish (e.g. BLAST; Altschul *et al*., 1990). PSI-BLAST (Altschul *et al.*, 1997) strikes an excellent compromise between speed and sensitivity.

Consensus sequences improve PSI-BLAST performance. Consensus sequences were used early on to improve alignments (Patthy, 1987). The initial approaches mimicked profile-sequence alignments (Henikoff and Henikoff, 1997; Sonnhammer and Kahn, 1994). Many improvements followed (Finn *et al*., 2006; Kahsay *et al*., 2005; Letunic *et al*., 2006; Marchler-Bauer *et al*., 2002; Merkeev and Mironov, 2006; Schaffer *et al*., 1999; Schultz *et al*., 1998; Servant *et al*., 2002; Thelen *et al*., 1999). However, none of those methods approached the success of PSI-BLAST. We have recently proposed a simple add-on to PSI-BLAST that substantially improves its performance (Przybylski and Rost, 2007). The add-on did not require any code change in PSI-BLAST. It consisted of adding a final step of 'freezing' the profile after the standard, iterative search against native sequences and then using it to search a database with the native sequences replaced by their consensus counterparts. This simple add-on improves the performance throughout the entire sensitivity curve. However, it is not clear how the underlying residue composition of database sequences affects the statistics of alignment scores. This is an important issue because users rely on the estimates of statistical significance to judge retrieved alignments. In addition, incorrect scoring might invalidate iterative searches against consensus sequences; a single false alignment in one of the intermediate searches might pollute a scoring profile and thereby all subsequent searches.

This study was motivated by the following three assumptions: (1) For a given residue substitution scoring matrix, the statistical significance of alignment scores depends on the residue compositions of aligned sequences. Assume that a particular scoring matrix highly rewards the alignment of tryptophan. This implies that sequences rich in tryptophan will likely generate higher alignment scores than those with average tryptophan content. (2) In general, the composition of consensus sequences differs from that of native sequences. Therefore, the distribution of alignment scores is likely different for consensus and native sequences, at least when using the same scoring matrix for both [such as BLOSUM62 (Henikoff and Henikoff, 1992) or the corresponding

*To whom correspondence should be addressed.

position-specific scoring matrices]. (3) PSI-BLAST is very popular, well-maintained, and has a great impact on the community of scientists that use sequence alignments. Therefore, it is desirable to improve PSI-BLAST performance without changing its alignment parameters (including scoring matrices and gap scores) with which the community is already familiar. In order to accomplish this, we have asked the following questions: how much do the parameters of alignment score distribution change for various types of consensus sequences? Can PSI-BLAST compensate for compositional variations through its internal composition-based adjustments (Schaffer *et al*., 2001)? Or, can we build consensus sequences in a way that renders statistical significance reported by PSI-BLAST as valid? Finally, can we apply PSI-BLAST to iteratively search consensus sequence databases?

2 METHODS

2.1 Generation of consensus sequences

We derived the consensus sequences from position-specific scoring matrices (PSSM, also known as scoring profiles) generated by iterative PSI-BLAST ('blastpgp') (Altschul *et al.*, 1997) searches of the redundancy-reduced UniProt (Apweiler *et al*., 2004) database containing about 1.5 million sequences. The sequence redundancy was reduced with CD-HIT (Li *et al*., 2001) such that pairs of sequences had <80% identical residues (globally). We allowed up to five PSI-BLAST iterations, i.e. the *frozen* profile was computed based on the fourth iteration or the next to the last one for early converging queries. The *E*-value threshold for inclusion in PSSMs was set to 0.001 and we increased the maximum number of aligned sequences to 2000 [blastpgp options '-j 5 -h 0.001 –v 2000 –b 2000 -Q PSSM(ASCII)']. Other options were left unchanged, including the default compositional adjustment of alignment score statistics and gap scores of $-(11+k)$ for gaps of length k. The determination of consensus sequences was based on ASCII PSSMs. For a given sequence and a residue position, we looked at the corresponding column of its PSSM and/or the frequency profile also present in the PSI-BLAST output. We explored three alternative ways for computing consensus residues at a given position i of a sequence: (1) *MF*: *maximal frequency*—the consensus residue j had the highest occurrence frequency f_{ij} in the profile column, (2) *MET: maximal relative entropy term*—we chose the residue j with the highest relative entropy term $f_{ij}ln(f_{ij}/b_j)$ with respect to the background frequency b_j, (3) *MR: maximal ratio of frequencies*—we chose the residue with the highest frequency ratio f_{ij}/b_j. In addition, we studied full (*MF-full*, *MET-full*, *MR-full*) and partial (*MF-partial*, *MET-partial*, *MR-partial*) versions of consensus sequences. For the (1) full consensus sequences, we computed the consensus residue at each sequence position, and for the (2) partial consensus, we computed the consensus in a constrained way, e.g. only for the more *informative* positions. The more informative positions were those having profile frequency columns with the relative entropy equal or above 0.6 (as reported in the PSI-BLAST output).

2.2 Alignments

All of the alignments (except those used to estimate the asymptotic values of the alignment score distribution parameters) were generated using PSI-BLAST version 2.2.15. The *frozen* scoring profiles (PSSMs) for the non-iterative profile-sequence alignments were generated in the same way as those used for generation of consensus sequences, except that a file containing the binary version of a PSSM was also stored [blastpgp option '-C PSSM(binary)']. Those binary PSSMs were used for a final (non-iterative) PSI-BLAST search against the appropriate consensus or native sequence databases [blastpgp options: '-j 1 -R PSSM(binary)']. For the non-profile-based sequence–sequence alignments the default BLOSUM62 (Henikoff and Henikoff, 1992) scoring matrix was used (blastpgp options: '-j 1'). When studying iterative searches against consensus sequence databases, we compared the performance for various number of iterations. The consensus version of the redundancy-reduced UniProt database used in iterative consensus searches was computed over a period of a few months using spare CPUs of a large computing cluster.

2.3 Evaluation of similarity search capability

We evaluated the ability to identify remotely related proteins using SCOP (Murzin *et al*., 1995) (release 1.69). We used the usual, descending hierarchy levels of 'fold', 'superfamily' and 'family' to define true and false relationships. Our positives consisted of pairs of protein domains from the same SCOP superfamily, but different SCOP families (i.e. the relatively easy pairs from the same family were not counted). However, for the more sensitive iterative searches against consensus sequences, we also counted pairs from the same SCOP-fold as positives. The negatives belonged to different SCOP-folds. We removed domains with: discontinuous sequences, missing coordinates in their three-dimensional structures, NMR and low-resolution structures (>2.5 Å), and the short ones (<50 residues). Next, we reduced the sequence redundancy of the set so that no pair of sequences could be aligned by BLAST with *E*-values better than 10^{-3} (when computed on UniProt database of ∼2 000 000 sequences), or at levels of sequence identity and alignment length that corresponded to homology-derived structures of proteins (HSSP)-values above 0 (Rost, 1999; Sander and Schneider, 1991) (whichever of the two criteria applied). This yielded a dataset of 2476 sequences for which we applied the all-against-all test.

2.4 Score statistics

PSI-BLAST provides statistical significance of alignment scores in terms of expectation values (*E*-values) that are given by:

$$E \approx Kmne^{-\lambda * score} \quad (1)$$

where m and n are the effective lengths (Altschul and Gish, 1996) of aligned sequences (query and database), *score* is a raw alignment score (as given by the values in scoring matrix and gap penalties), and K and λ are the parameters of the score distribution that depend on a scoring system and the residue composition of aligned sequences. Note that the computation of the *E*-value primarily depends upon a proper estimate of λ and much less so on that for K.

2.5 Determining parameters of alignment score distributions

The problem of estimating the statistical significance of alignment scores has been studied extensively (Altschul and Gish, 1996; Karlin and Altschul, 1990; Mott, 1992; Waterman and Vingron, 1994). We computed λ and K parameters [Equation (1)] with our implementation of the 'island' approach (Altschul *et al*., 2001; Olsen *et al*., 1999) for a case of scoring profiles. This approach is appropriate as the primary methods studied in this article rely on searching databases of consensus sequences with precomputed PSSMs. We have also estimated the score distribution parameters for profile-based searches against native sequences to relate our results to the earlier studies. First, we obtained the initial PSSMs for hundreds of thousands of randomly selected UniProt sequences. Most of them were too short to study the score distribution in the asymptotic limit of very long sequences. Therefore, we concatenated them in random order and then cut them into final long PSSMs, each composed of 7000 columns. We ended up with 75 000 of such long PSSMs. To generate corresponding long random consensus sequences, we first computed consensus sequences from each of the long PSSMs and used them to compute consensus residue background frequencies. Those background frequencies were used to generate random sequences used for studying asymptotic alignment score distribution parameters. For partial consensus sequences, we computed two separate sets of backgrounds—inside and outside of consensus regions and used them accordingly for

generation of random partial consensus sequences (with the informative positions indicated by the original PSSM relative entropy values at each sequence position).

2.6 Studying the compositional adjustment of alignment score statistic in PSI-BLAST

The newer versions of PSI-BLAST can adjust alignment score statistics based on varying residue compositions of query and database sequences ('-t' option in PSI-BLAST). In particular, we looked at the performance of the default adjustment implemented in the 2.2.15 version of the software. We have generated random sequence databases based on the native and consensus background residue frequencies. The numbers of random sequences and their sizes were the same as those found in the non-redundant UniProt database. We queried those databases with about 20 000 randomly chosen native sequences and the corresponding PSI-BLAST profiles (PSSMs). We recorded the average cumulative numbers of alignments per query that had *E*-values better than a given threshold value.

3 RESULTS AND DISCUSSION

3.1 Alignment score parameters depended on consensus type

The variation of λ with the alignment score [Equation (1)] for gapped alignments has been described before (Altschul *et al*., 2001). Low-scoring alignments usually have fewer gaps and their score distribution differs from those obtained for high-scoring alignments with gaps. Here, we have focused mostly on asymptotic values of λ for high scores because they correspond to statistically significant alignments originating from searches of large sequence databases. In particular, we looked at λ for PSSMs generated with five iterations of PSI-BLAST. We observed that λ depended on the sequence types (Fig. 1). Computing consensus residues for the full sequence produced largest changes in λ (open symbols in Fig. 1, i.e. *MR-full, MET-full* and *MF-full*). For each one of them, the asymptotic value of λ was less than 0.2 (more data points would be needed to establish a precise limit). The value of λ for the profile-sequence alignments of the native sequences was about 0.255 (Fig. 1; green squares). This is rather close to a value of 0.267 previously established for the sequence-sequence alignments with the BLOSUM62 scoring matrix (Altschul *et al*., 2001). For the partial consensus sequences, λ appeared to follow the value obtained for the native sequences (filled symbols in Fig. 1, i.e. *MF-partial* and *MET-partial*). To some extent this result is not surprising because partial consensus substitutions are more restricted than the full ones, i.e. change fewer residues (Table 1). As a result, we established that one could use PSI-BLAST without any modifications to perform profile-based search against partial consensus sequence databases and maintain proper estimates of *E*-values.

We have also estimated the location parameter *K* used for computing *E*-values [Equation (1)]. For example, we found it to be ∼0.015 for the full consensus sequences (*MF-full*), 0.030 for the partial consensus sequences *(MF-partial*) and 0.032 for the native sequences.

3.2 Search performance similar for all consensus types

Do some types of consensus sequences retrieve related sequences from a database better than others? For each type of consensus, we ordered all query alignments by PSI-BLAST *E*-values. Next, we computed the cumulative numbers of true positive relations (same

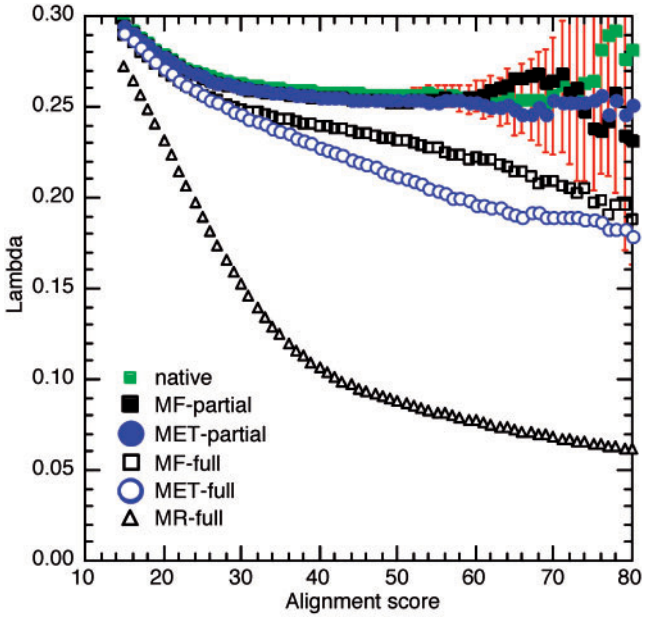

Fig. 1. Estimating λ. Score distribution parameter λ [Equation (1), *y*-axis] varies with alignment scores (*x*-axis). In practice, we are interested in the asymptotic value of λ for higher scores. Full consensus sequences affected λ significantly (open symbols) when compared to native sequences (green squares). In contrast, partial consensus did not significantly affect λ (filled black and blue symbols). Red error bars estimate the SD (for clarity only shown for native sequences). Note that high alignment scores were attained by few alignments.

Table 1. Pairwise residue identities of native and consensus sequences

		Native	Full consensus			Partial consensus		
		native	MR	MF	MET	MR	MF	MET
Native	native	100						
Full consensus	MR	65	100					
	MF	54	76	100				
	MET	51	80	90	100			
Partial consensus	MR	86	79	64	63	100		
	MF	83	72	71	67	93	100	
	MET	82	73	69	69	94	98	100

Shown are average percentages of pairwise residue identities between different types of sequences of a test set.

SCOP superfamily but different family) for increasing cumulative numbers of false positive pairs (different SCOP-folds). For any number of false positives (i.e. at any error rate), the profile-sequence searches against the databases of full consensus sequences yielded most true positives (Fig 2; top three curves: *MET-full*, *MF-full, MR-full*). Interestingly, it did not matter much how we compiled the full consensus (three top lines with open symbols in Fig. 2 are almost indistinguishable). The profile-based searches against partial consensus sequences (only most informative positions replaced by consensus) were somewhat less efficient, especially when more false hits were allowed (Fig 2; *MET-partial*). Nevertheless, they were significantly better than standard profile-sequence searches

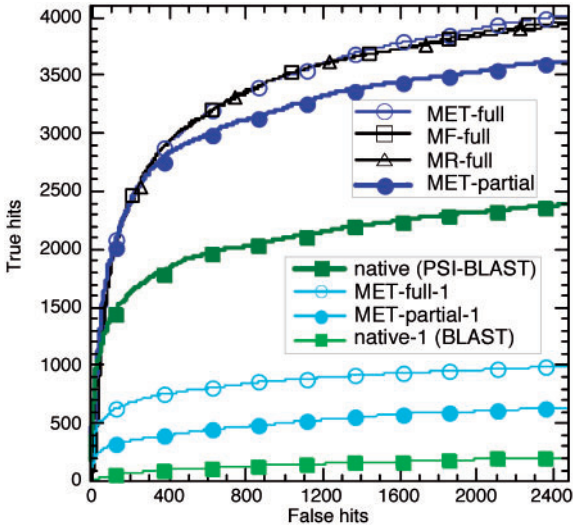

Fig. 2. Comparison of search performance. All-against-all alignments of the test set sequences were ordered by their PSI-BLAST *E*-values. The cumulative numbers of non-trivial true relations (same SCOP superfamily but different SCOP family) were plotted against the cumulative numbers of false positives (different SCOP-folds). The profile-sequence searches against the full consensus sequences performed best (top three curves: *MET-full*, *MF-full, MR-full*). Profile-sequence searches against partial consensus sequences were slightly less efficient (*MET-partial*), but they were still significantly better than standard profile-sequence (*native*). Sequence–sequence searches (one cycle of PSI-BLAST with BLOSUM62 matrix) were clearly inferior (*MET-full-1, MET-partial-1, native-1*).

of PSI-BLAST (Fig 2; *native*). For comparison, we also included the performance of sequence–sequence searches with pairwise BLAST against the native and consensus sequences (Fig. 2; *MET-full-1, MET-partial-1, native-1*). As expected, pairwise searches fared much worse than profile-sequence searches. The relative performance difference between the full and partial consensus sequences appeared larger for the sequence–sequence (Fig. 2; *MET-full-1, MET-partial-1*) than for profile-sequence searches.

3.3 Composition of consensus sequences varied

The search performance appeared not to differ between various types of full consensus sequences, although their average residue compositions were quite different (Fig. 3A). The consensus based on the maximum ratio of target and background frequencies (*MR-full*) weighed more heavily rare residues such as tryptophane (W). The consensus based on the most frequent residue (*MF-full*) weighed more heavily the more ubiquitous ones such as leucine (L). Finally, the consensus based on relative entropy (*MET-full*) produced the composition that appeared to be more balanced (Fig. 3A, blue bars). The average percent differences in residue identity (and SDs) between native and full consensus sequences were: 65 (±14) for *MR-full*, 54 (±16) for *MF-full* and 51 (±17) for *MET-full* consensus sequences. The partial consensus calculations resulted in average compositions that were much closer to the native ones (Fig. 3B). The corresponding residue identities with respect to native sequences were: 86 (±7)%, 83 (±8)% and 82 (±8)%. Thus, the consensus calculation (MR) that changed sequences the least in terms of the average residue identity has changed the score distribution parameters the most. Other pairwise residue identities are given in Table 1. All calculations were performed on our non-redundant SCOP test set.

3.4 PSI-BLAST compositional adjustments were partially successful

When compositions of aligned sequences differ from a standard one, PSI-BLAST can attempt to correct the estimates of statistical significance accordingly (Schaffer *et al*., 2001; Yu and Altschul, 2005). We studied how well the default adjustments perform on consensus sequences (non-default adjustments are not available for profile-based searches). Using PSI-BLAST profiles we searched against the consensus and native sequence databases (Section 2). For the comparison, we also searched with the BLOSUM62 substitution matrix (standard, non-profile BLAST search). In the latter case, the estimates of statistical significance were not very sensitive to compositional differences and the statistic adjustments worked well (Table 2, observed and expected counts similar; adjustments were conservative). However, for the profile-based searches the compositional differences played a significant role, particularly for the full consensus sequences (especially pronounced for *MR-full,* Table 3). The compositional adjustment of scores attempted by PSI-BLAST (-t option set to 1) failed to satisfactorily correct for the differences. In contrast, the *E*-value estimates were good for partial consensus sequences. For both native and partial consensus sequences, the compositional score adjustment sometimes resulted in slightly increased numbers of random alignments with significant *E*-values.

3.5 Little additional CPU needed for add-on

In this study, we used separate databases for the iterative derivation of PSSMs (non-redundant UniProt) and for the final search and alignment against consensus sequences. On average, the entire iterative PSI-BLAST search took about 10 min per query (about 2 min per iteration on a single 3.2 GHz CPU with 2 GB of RAM using query sequences with average length of 415 residues). The additional time consumed by the add-on to search against a consensus sequence database of the same size depended on the sequence types. It took about 7 min to search *MR-full* and 4.5 min for *MF-full* consensus sequence databases. In the case of partial consensus it took about 2.5 min to search the *MR-partial* and about 2.2 min for *MF-partial* (compared to about 2 min needed to search native one with PSI-BLAST profile).

3.6 Iterative searches against consensus sequences yielded further improvements

We made the first attempt at analyzing iterative PSI-BLAST searches against consensus sequence databases. For this analysis, we pushed the envelope by running up to 20 iterations. We counted hits belonging to the same SCOP-fold but to different families as positives to reach deeper into remote protein–domain relationships. The iterative PSI-BLAST searches against the native sequence database resulted in near saturation of performance at about 10 iterations. Only a small improvement was observed in the subsequent 10 iterations (Fig. 4; top two green lines). The iterative searches against consensus sequences (*MF-full*) produced significantly more true hits with just three iterations. Five consensus

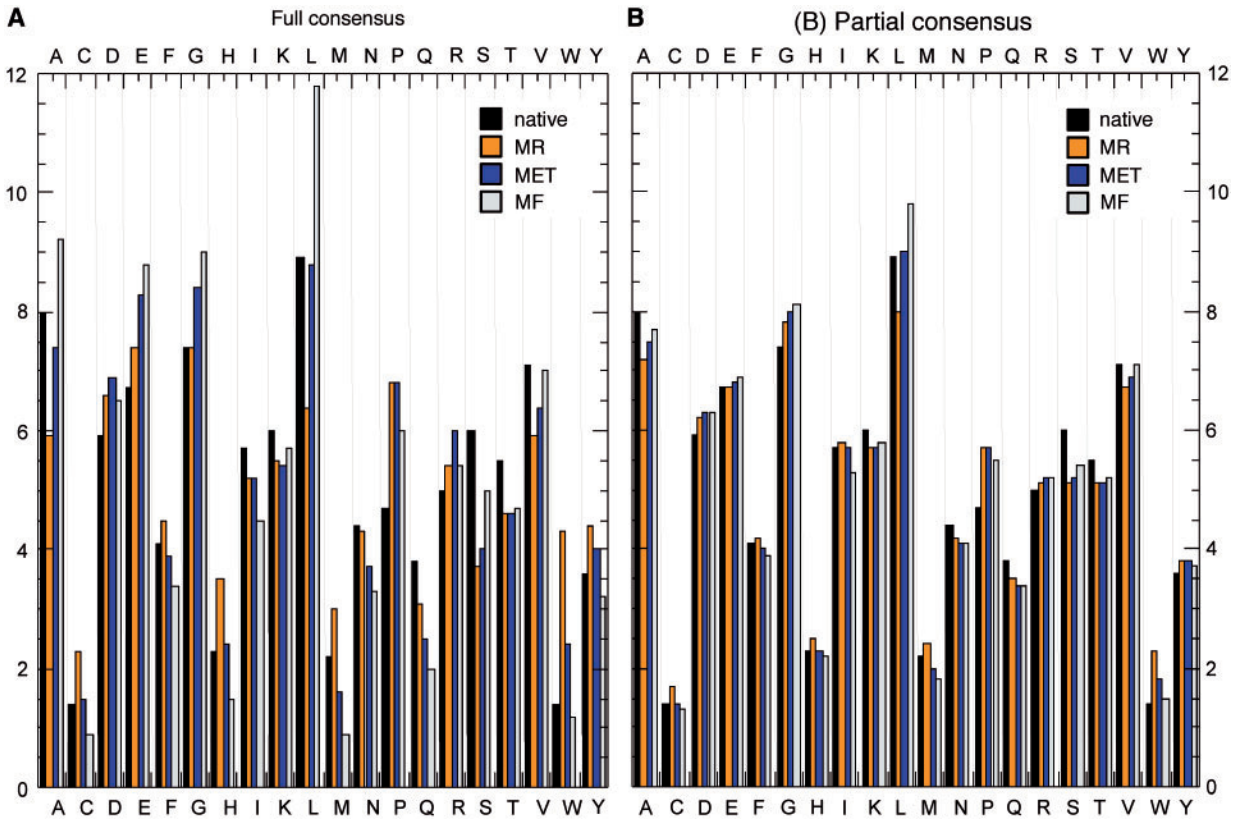

Fig. 3. Comparison of residue compositions. We computed the background residue compositions for consensus and native sequences in our test set. Full consensus sequences (**A**) differed more from native than partial consensus sequences (**B**). Choosing the consensus residue corresponding to the highest relative entropy term (blue bars) resulted, on average in smaller deviations from the native composition.

Table 2. Accuracy of BLAST *E*-values[a]

Observed	Native		Full consensus						Partial consensus					
Expected	native	native-adj.	MR	MR-adj.	MF	MF-adj.	MET	MET-adj.	MR	MR-adj.	MF	MF-adj.	MET	MET-adj.
0.001	0.0014	0.0010	0.0010	0.0002	0.0007	0.0006	0.0009	0.0003	0.0014	0.0008	0.0018	0.0008	0.0012	0.0009
0.01	0.010	0.007	0.006	0.004	0.006	0.005	0.008	0.004	0.011	0.006	0.012	0.005	0.011	0.006
0.1	0.09	0.07	0.07	0.04	0.03	0.06	0.08	0.06	0.09	0.07	0.11	0.07	0.10	0.07
1	0.9	0.7	0.7	0.5	0.2	0.7	0.9	0.6	0.9	0.7	1.1	0.7	1.0	0.7
10	9	7	7	6	18	7	9	7	9	8	11	8	10	8

[a]Shown are the expected and observed numbers of random alignment scores per query for ∼20 000 sequence queries on randomly generated databases (of UniProt size) of native and consensus sequences. Appendix '-adj.' indicates results obtained with the use of compositional adjustment of *E*-values with BLAST option '-t' set to 1.

iterations produced almost twice as many true hits as the native PSI-BLAST search produced with 20. For comparison, we showed the results of the profile-sequence search (profile obtained from 10 iterations of PSI-BLAST on a native database) against a final database of consensus sequences (Fig. 4; blue line, mixed). These results remain to be compared to the performance of profile–profile methods (Bujnicki *et al*., 2001; Fischer *et al*., 2003).

4 CONCLUSIONS

PSI-BLAST is an excellent, well-known, well-maintained and trusted resource for searching and aligning sequence databases. A simple add-on consisting of searching with a PSI-BLAST generated scoring profile against a database of consensus sequences significantly improved the performance in finding related sequences. Here, we specified in detail how different strategies of compiling consensus residues affected the estimates of statistical significance and performance. Profile-based PSI-BLAST searches against full consensus sequences improved the most over searches against native sequences. However, they sometimes suffered from problems in the estimates of statistical significance. The partial consensus sequences improved significantly over native sequences without sacrificing estimates of statistical significance. Our initial results for iterative searches against consensus sequences were very promising: a lower

Table 3. Accuracy of PSI-BLAST[a] *E*-values[b]

Observed	Native		Full consensus						Partial consensus					
Expected	native	native-adj.	MR	MR-adj.	MF	MF-adj.	MET	MET-adj.	MR	MR-adj.	MF	MF-adj.	MET	MET-adj.
0.001	0.0013	0.0033	66.4912	1.6637	0.0024	0.0008	0.0215	0.0093	0.0053	0.0040	0.0014	0.0028	0.0013	0.0038
0.01	0.008	0.020	98.880	4.3162	0.018	0.028	0.086	0.036	0.022	0.026	0.010	0.021	0.009	0.022
0.1	0.08	0.18	159.83	12.83	0.170	0.22	0.51	0.26	0.17	0.17	0.10	0.20	0.10	0.20
1	0.8	1.6	259.1	34.9	1.6	1.8	3.2	2.1	1.4	1.4	1.0	1.7	1.0	1.7
10	8	13	405	102	14	14	22	16	12	12	9	13	10	14

[a]Shown are the expected and observed numbers of random alignment scores per query for a set of about 20 000 profile (PSSM) queries on randomly generated databases (of UniProt size) of native and consensus sequences. Appendix '-adj.' indicates results obtained with a use of compositional adjustment of E-values with PSI-BLAST option '-t' set to 1.
[b]PSI-BLAST search was restarted from a stored profile.

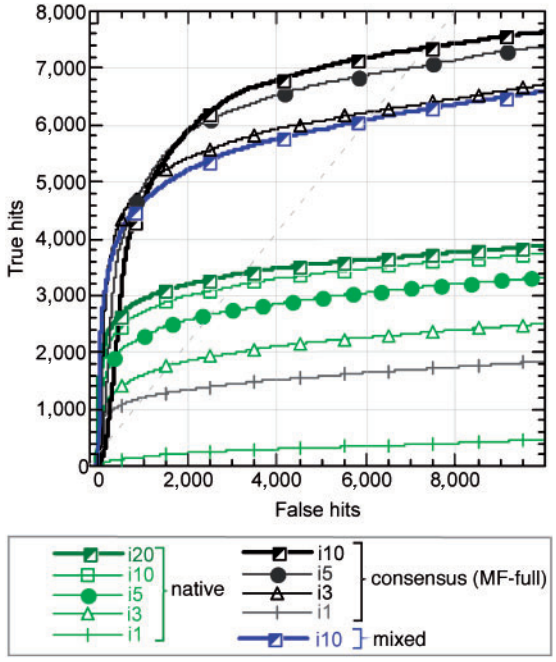

Fig. 4. Iterative PSI-BLAST searches against native and consensus sequences. Iterative PSI-BLAST searches and PSSM refinements on native sequence database (green lines) resulted in near saturation of performance at about 10 iterations (top two green lines). The corresponding searches on the database of consensus sequences (black lines) found significantly more true hits (same SCOP-fold but different family) with just three iterations (black triangles), while five iterations (black circles) retrieved almost twice as many true hits as the maximum for the native PSI-BLAST. For comparison, a result of the *frozen* profile-based search against a final database of consensus sequences (*MF-full*) is presented (blue line).

number of iterations used less CPU overall and yielded about twice as many correct hits at the same error rates as standard PSI-BLAST searches did. Hence, the fusion of PSI-BLAST and consensus sequences promises another leap in database searches.

ACKNOWLEDGEMENTS

Thanks to Johannes Söding (Max-Planck-Institute in Tübingen) for helpful comments and suggestions; to Beth Cockerham and Kiran Garimella (Broad Institute of MIT and Harvard) for reading and commenting on the article; to all who deposit their experimental data in public databases and to those who maintain these databases; and also to all who develop alignment tools and make them publicly available, in particular to those who develop and support PSI-BLAST!

Funding: This work was supported by the grants R01-LM07329-01 from the National Library of Medicine (NLM) and U54-GM074958-01 from the Protein Structure Initiative (PSI) of the National Institutes of Health (NIH).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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.* (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Apweiler,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Bujnicki,J.M. *et al.* (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Fischer,D. *et al.* (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins*, **53** (Suppl. 6), 503–516.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.

Kahsay,R.Y. *et al.* (2005) Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. *Bioinformatics*, **21**, 2287–2293.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Letunic,I. *et al.* (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Marchler-Bauer,A. *et al.* (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.

Merkeev,I.V. and Mironov,A.A. (2006) PHOG-BLAST – a new generation tool for fast similarity search of protein families. *BMC Evol. Biol.*, **6**, 51.

Mott,R. (1992) Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.*, **54**, 59–75.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Olsen,R. *et al*. (1999) Rapid assessment of extremal statistics for gapped local alignment. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 211–222.

Patthy,L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.*, **198**, 567–577.

Przybylski,D. and Rost,B. (2007) Consensus sequences improve PSI-BLAST through mimicking profile-profile alignments. *Nucleic Acids Res.*, **35**, 2238–2246.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schaffer,A.A. *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

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.

Schultz,J. *et al*. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Servant,F. *et al*. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.

Sonnhammer,E.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3**, 482–492.

Thelen,M.P. *et al*. (1999) A sliding clamp model for the Rad1 family of cell cycle checkpoint proteins. *Cell*, **96**, 769–770.

Waterman,M.S. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Yu,Y.K. and Altschul,S.F. (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, **21**, 902–911.