2238–2246 *Nucleic Acids Research, 2007, Vol. 35, No. 7* *Published online 16 March 2007*
doi:10.1093/nar/gkm107

Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments

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

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave. Rm. 801, New York, NY 10032, USA and [3]NorthEast Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

Received December 20, 2006; Revised February 5, 2007; Accepted February 6, 2007

ABSTRACT

Sequence alignments may be the most fundamental computational resource for molecular biology. The best methods that identify sequence relatedness through profile–profile comparisons are much slower and more complex than sequence–sequence and sequence–profile comparisons such as, respectively, BLAST and PSI-BLAST. Families of related genes and gene products (proteins) can be represented by consensus sequences that list the nucleic/amino acid most frequent at each sequence position in that family. Here, we propose a novel approach for consensus-sequence-based comparisons. This approach improved searches and alignments as a standard add-on to PSI-BLAST without any changes of code. Improvements were particularly significant for more difficult tasks such as the identification of distant structural relations between proteins and their corresponding alignments. Despite the fact that the improvements were higher for more divergent relations, they were consistent even at high accuracy/low error rates for non-trivially related proteins. The improvements were very easy to achieve; no parameter used by PSI-BLAST was altered and no single line of code changed. Furthermore, the consensus sequence add-on required relatively little additional CPU time. We discuss how advanced users of PSI-BLAST can immediately benefit from using consensus sequences on their local computers. We have also made the method available through the Internet (http://www.rostlab.org/services/consensus/).

INTRODUCTION

Improved database search and alignment methods boost biology

Sequence alignments are fundamental to modern molecular biology. They are used to detect evolutionary relationships among proteins and genes; they also provide the basis for most advanced predictions of structure and function for biomolecules. The more organisms are sequenced, the more the need for sensitive and accurate database search and alignment methods increases. In conjunction with an appropriate scoring (decision) function, sequence alignment methods can often distinguish homologous from non-homologous genes/proteins. Alignments are also used to establish residues that are conserved between related sequences. This helps to identify residues that are most important for function and to transfer three-dimensional (3D) coordinates in comparative modeling of protein structures. Since most relations between genes or proteins are observed at large evolutionary distances, small improvements in the sensitivity and accuracy of database searches and alignments may translate to thousands of novel annotations that could guide and accelerate experimental biology.

PSI-BLAST strikes a very good compromise between speed and sensitivity

Ideally, an alignment method should accurately identify and align related sequences in today's rapidly expanding databases within the shortest possible time. While we want to simultaneously optimize speed and reliability, in practice, there is a tradeoff; more 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 (1)]. Generally, the most

*To whom correspondence should be addressed. Tel: +1 212 851 4669; Fax: +1 212 851 5176; Email: dsp23@columbia.edu

sensitive and accurate methods use profile–profile comparisons (2–5). In those algorithms, nucleic/amino acid substitution patterns are used for both sequences being aligned. One downside of profile–profile alignments is that they are relatively slow. When aligning two sequences of lengths m and n they require on the order of S^*m^*n operations (where S is the size of sequence alphabet—20 for proteins). Moreover, the algorithm is not easily amenable to acceleration. In contrast, the less powerful sequence–profile alignment methods can be easily accelerated. This is most impressively visible in the case of PSI-BLAST (6) that combines techniques for acceleration [FASTA (7), BLAST (1)] with accurate profile-based dynamic programming (8), and with an automated iterative refinement of the search. As a result, the PSI-BLAST search and alignment could be even two orders of magnitude faster (9) than the corresponding Smith–Waterman (8) alignment algorithm and almost as sensitive. This is an impressive solution that clearly is one reason for the enormous popularity of PSI-BLAST. Often, everyday sequence analysis applies a two-tier approach: first a search with a reliable and fast PSI-BLAST followed by a search with programs that generate more accurate alignments but are neither fast enough nor set up for database searches such as ClustalW (10), T-Coffee (11), MAFFT (12), MUSCLE (13). Note that in the following, we use a slight deviation from the usual connotation, namely the term profile–sequence instead of sequence–profile alignment to differentiate between the query (profile) and the template/database (sequence); PSI-BLAST by this notation is a profile–sequence method.

Consensus sequences can represent families of related proteins

Protein sequences are subject to continuous evolution. Random mutations and insertions/deletions of nucleic acids within genes are source of variability of protein sequences. The pressure to maintain biological function (and/or 3D structure) constrains the range of mutations. In general, proteins can have quite dissimilar sequences and still perform the same biological function and/or have very similar 3D structure. At each sequence position, i.e. for each residue, the mutational variability can be characterized by a vector of amino acid substitution frequencies. The resulting matrix is often referred to as a *sequence profile*. The substitution frequencies are typically computed from alignments of functionally and/or structurally related proteins. In subsequent steps (iterations), such profiles are then used as the basis for aligning protein sequences (in profile-sequence and profile–profile algorithms). A consensus sequence can be thought of as a one-dimensional simplification of such a profile that, e.g. substitutes the 20-dimensional vector (for 20 amino acids) in each column (residue position) by the most frequent or most informative amino acid observed at that position. The consensus can be applied globally (to all profile columns) or locally (only to some columns) (14,15). There also exist other, more specialized techniques for generating consensus sequences (16).

Consensus sequences empower alignment methods

Consensus sequences were used early on to improve alignments (17). Initially the substitution of profiles by consensus mimicked profile–sequence alignments (14,18) (more accurately leading to consensus–sequence or sequence–consensus comparisons). Those methods tapped into fast alignment algorithms such as FASTA or BLAST. This approach is used successfully with ProDom (19) and COBBLER (14) consensus sequences. The development of fast profile–sequence alignment methods such as PSI-BLAST halted the development of sequence–consensus methods. Although BLAST-based sequence–consensus searches may be considerably faster than PSI-BLAST searches, they are thought to also be considerably less accurate. A symmetric approach of aligning a query sequence with a database of profiles (sequence–profile alignments) is used, for example, in Blocks Searcher (20) and in RPS-BLAST (6,21) to search the Blocks (22,23), PRINTS (24) and CDD (25,26) databases. Another approach is to align a query sequence with profile-derived Hidden Markov Models (HHMs) as applied by, e.g. Pfam (27) and Smart (28,29). An interesting idea suggested for PSI-BLAST searches with consensus sequences was never tested nor implemented on a larger scale (30).

Profile–profile algorithms tend to be both most sensitive and most accurate (31,32). Unfortunately, profile–profile comparisons are also much slower and more complex than heuristically accelerated sequence–sequence and profile–sequence algorithms. For this reason their application to everyday searches of large sequence databases on a typical computer workstation is not practical. Recently, an algorithm that approximates profile–profile algorithms by performing consensus–consensus alignments (16) has been published. In this article, we propose a different approximation to profile–profile comparisons in which only one profile is substituted by a consensus sequence (profile–consensus alignment). A somewhat similar approach (without heuristic speed-up) was proposed for aligning quasi-consensus sequences with HMMs (33). Consensus sequences can be derived in various ways. In one approach the raw sequences are only replaced by consensus residues 'locally', i.e. for some of the residues, e.g. the evolutionarily conserved regions (as done by the COBBLER method based on Blocks). Alternatively, one could replace the complete sequence with a consensus sequence. Here, we tested both alternatives.

Which consensus alignment is best?

Given all possible variants of using consensus sequences: which one is best? A direct comparison of existing methods may not provide the most informative answer to this question because different methods generate profiles and consensus sequences in different ways (see Supplementary Data for such a comparison). Here, we set up an experiment where we could control all the parameters to study differences between various algorithmic approaches. The same sets of multiple alignments and the same algorithms for computing consensus residues were used. Also the same alignment algorithm

(PSI-BLAST) was used to make all alignments. We compared three possible ways of using consensus sequences in alignments—aligning raw with consensus sequences (sequence-consensus), aligning only consensus sequences (consensus–consensus) and (proposed here) aligning profiles with consensus sequences (profile–consensus). In addition, we studied whether protein sequences locally enriched with consensus information performed better than simple global consensus sequences. Since the alignment of consensus sequences is as widely applicable and potentially as fast as alignment of raw sequences we have also compared it with the standard raw sequence alignment methods—PSI-BLAST and BLAST. Finally, we have provided the first comprehensive analysis for the quality of consensus sequence alignments.

We found that profile–consensus alignments outperformed other consensus sequence alignments. Notably, the profile–consensus approach most closely resembled profile–profile algorithms. The profile–consensus searches with PSI-BLAST were significantly more sensitive and specific than the original PSI-BLAST searches with raw sequences. Improvements were particularly significant for more difficult tasks such as the identification and alignment of distant structural relations between proteins. Despite the fact that the improvements were higher for more divergent relations, they were consistent even at high accuracy/low error rates for non-trivially related proteins. The improvements were very easy to achieve; no parameter used by PSI-BLAST was altered and no single line of code changed. Moreover, the consensus sequence add-on required relatively little additional CPU time. This new way of search and alignment added onto the existing PSI-BLAST program is almost as fast and easily applicable as PSI-BLAST itself.

MATERIALS AND METHODS

Generation of consensus sequences

For each test sequence used in this study, we generated the position-specific scoring matrix (PSSM) using PSI-BLAST. We used a maximum of five iterations, an e-value threshold for inclusion in PSSM of 0.001 and no query filtering [blastpgp options '−j 5 −h 0.001 −F F −Q PSSM(ASCII)']. All profiles were generated by aligning against a redundancy-reduced version of the UniProt (34) database [80% sequence identity reduction using CD-HIT (35)]. The determination of consensus amino acids was based on the ASCII PSSMs. Each original residue was replaced with the amino acid that had the highest corresponding PSSM score (highest 'target' to background frequency ratio). Three types of consensus sequences were generated: In the 'global consensus' mode, we replaced all residues by the consensus; in the 'consensus$^{\text{top50\%}}$' mode we replaced the 50% of the residues associated with most informative profile columns (highest relative entropy) by the consensus; in the 'consensus$^{\text{low50\%}}$' mode we replaced the 50% of residues associated with least informative columns with consensus residues.

Alignments

All alignments were generated using the 'blastpgp' executable in the PSI-BLAST suite of programs. All profiles (PSSMs) used for alignments were generated in the same way as profiles used for generation of consensus sequences except that a file containing the binary version of a PSSM was also stored [blastpgp options: '−j 5 −h 0.001 −F F −C PSSM(binary)']. The binary PSSM was used for a final PSI-BLAST search and alignment of the database of consensus sequences using just one iteration [blastpgp options: '−j 1 −F F −R PSSM(binary)']. For non-profile-based alignments of sequences 'blastpgp' program with default BLOSUM62 (36) scoring matrix was also used (options: '−j 1 −F F'). For comparison of performance PSI-BLAST (the same options) was used to search the corresponding database of raw sequences. For convenience of analysis the alignments of consensus sequences were translated back to 'real' sequences using a simple Perl script (Figure 1).

Evaluation of performance

There is no commonly accepted means of evaluating the performance of database search and alignment methods. One way of generating test sets of sufficient size is to compare proteins with known 3D structures because for such comparisons standards-of-truth can relatively easily be generated automatically. We assessed both the ability to identify related proteins and the ability to correctly align them based on structural alignments (below). Evolution has conserved the principle components of protein 3D structures (often misleadingly referred to as 'the fold') at higher divergence than the principle aspects of protein function. Therefore, evaluations based on structural alignments tend to put emphasis on more diverged relationships than would comparisons that are based on functional features.

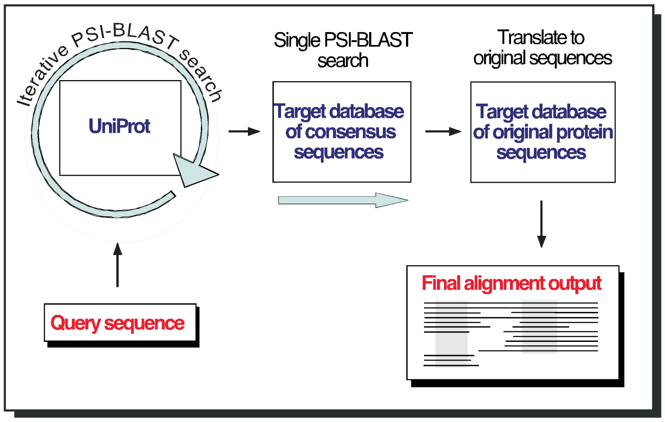
Iterative PSI-BLAST search
UniProt
Single PSI-BLAST search
Target database of consensus sequences
Translate to original sequences
Target database of original protein sequences
Query sequence
Final alignment output

Figure 1. Sketch of consensus search. First, the PSSM for a query protein sequence is built by an iterative PSI-BLAST search over a large database of proteins sequences (such as UniProt). The resulting PSSM is then used to search and align sequences contained in a target database of consensus sequences. Finally, consensus sequence alignments are translated to alignments of the native raw protein sequences.

PSI-BLAST as the point of reference

All our evaluations used PSI-BLAST and BLAST as points of reference. The rationale was manifold. First, PSI-BLAST alone is not sufficient because there are many different ways of running PSI-BLAST, i.e. we need a point of reference in order to track our way of running PSI-BLAST. For this we explored BLAST. Second, most recent assessments of new alignment methods are compared to PSI-BLAST and/or BLAST. Since it is rather unreasonable to compare results obtained on different data sets, we cannot directly compare our results to other publications. However, the two reference points allowed for the triangulation of a comparison. Third, our major purpose was to illustrate the advantage of adding our protocol onto existing PSI-BLAST searches, i.e. PSI-BLAST is the most important point of reference for our protocol. This is because PSI-BLAST is one of the few tools that can be used for fast and accurate searching of largest sequence databases and consensus sequence alignments can be used for the same purpose.

Evaluation of search capability

We evaluated the ability to identify related proteins with SCOP (37) (release 1.69). For the assessment we omitted protein pairs from the same SCOP family (considered rather easy to recognize) and pairs that belonged to different SCOP superfamilies but to the same SCOP fold (considered too difficult for sequence alignment methods). Thus, our positives were pairs of proteins from the same SCOP superfamily while negatives were pairs of proteins from different SCOP folds.

Evaluation of alignment quality

Comparative modeling is a technique that allows the modeling of a 3D structure for a query protein Q based on a template T of experimentally known structure (38,39). In the simplest implementation comparative modeling first aligns Q and T and then copies the co-ordinates from T to model the structure of Q based on this alignment. Alignment mistakes significantly impair the quality of such models. We measured the quality of alignments implicitly, namely by assessing the quality of the comparative models originating from the alignments.

We superposed all models (represented by C_α atom coordinates) with experimentally determined 3D structures using one particular automatic method for structural superposition, namely LGA (40); this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction [CASP (41)]. First, we computed a Global Distance Test (GDT) (40) that corresponds to the largest, not necessarily continuous subset of residues superimposable within a specified distance threshold. Second, we also computed Longest Continuous Segments (LCS) (40) of residues (consecutively modeled residues) that can fit under a specified RMSD cutoff. The second measure provided us with a local alignment quality test. Note that we chose a subset of pairs (Q,T) such that for all pairs experimental structures were available; we built the model for Q using the known structure of T and assuming that Q had no known structure, but we evaluated the accuracy of the model using the experimentally known structure for Q. We reported results for two different thresholds. The first was rather stringent (2 Å); it focused on the essential core similarities between model and experiment. The second was rather relaxed (5 Å) thereby capturing more generic, coarse-grained similarities. Note that GDT computation uses the actual distance threshold while LCS uses average distance (RMSD).

Note that we assess a real-life situation in which we model structures for proteins Q that are not identical to the experimentally known structures T. This implies that the quality of a model also depends crucially on the divergence between Q and T: at high evolutionary distances, the two structures will differ so much in detail that even accurate alignments will not give as accurate models as inaccurate alignments between more closely related pairs. We accounted for this effect by structural alignments: we used the 3D alignment method MAMMOTH (42) to align the known structures of Q and T. This approximated an upper limit for what could be achieved by simplistic comparative modeling that only copied coordinates. The quality of models based on MAMMOTH alignments was also evaluated using LGA.

Data sets

We analyzed the ability to correctly identify and align related proteins on a subset of SCOP. We removed domains with discontinuous sequences, structures with missing coordinates, NMR structures, low-resolution structures (<2.5 Å) and short proteins (<50 residues). The resulting set of proteins was tailored differently for assessing search and alignment quality.

To assess search capability (homology/fold recognition), we reduced the redundancy of the sequence set so that no pair of sequences could be aligned by BLAST at 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 HSSP-values above 0 (43,44) (whichever of the two criteria applied). This yielded a data set of 2476 sequences for which we applied an all-against-all test.

The choice of datasets for studying alignment quality was motivated by the observation that the quality of sequence alignments deteriorates rapidly below levels of around 30% pairwise sequence identity (45). In order to assess the ability of our add-on consensus approach to correctly align more distant pairs, we did not consider alignments with >30% pairwise sequence identity. Within this set of distant relatives, we monitored two different levels of alignment difficulty correlated with standard everyday uses of sequence alignment algorithms. First, we chose only those protein pairs that could be aligned by PSI-BLAST with e-values ranging from 10^{-3} to 10 when searching large public sequence databases. Second, we looked at the more difficult task of aligning protein pairs belonging to the same SCOP superfamily but different SCOP families (with e-values of up to 100 when computed on sequence unique subset of SCOP).

Those sets were composed of 1647 (set 1: most related, non-trivial pairs), and 5551 (set 2: more difficult, most diverged) protein pairs respectively. The final data sets were 'pairs non-redundant' in the following sense: no protein in any pair could be aligned with any protein from any other pair at PSI-BLAST e-values better than 1000 (calculated on UniProt database).

RESULTS

Approximation of profile–profile alignments performed best

For each alignment method tested here, we ordered all alignments of all queries by e-values. Next, we computed the cumulative number of true positive relations (same SCOP superfamily but different family; note that cases with the same superfamily and the same family were carefully filtered out from our data set to reduce redundancy) for increasing cumulative numbers of false positives (pairs of proteins with different SCOP folds). For any cumulative number of false positives (i.e. at any error rate) searching with profiles against a database of global consensus sequences yielded most true positives (profile-consensus, Figure 2A and Supplementary Data). Such a search was the closest approximation of profile–profile alignments since only one of the profiles was replaced by the corresponding consensus sequence. Replacing both profiles by consensus sequences and scoring alignments with a generic scoring matrix (BLOSUM62) did not perform as well (consensus–consensus, Figure 2A). Although this approach seemed to have some advantage over PSI-BLAST in a low error region (few false positives), the loss of some profile information for both profiles was largely detrimental. Finally, searching with a raw sequence and a generic scoring matrix against a database of consensus sequences performed worse than other consensus sequence methods but significantly better than BLAST (sequence–consensus, Figure 2A).

We also observed that global consensus sequences performed better than sequences with partial consensus information. For example, searching with profiles against consensus$^{\mathrm{top50\%}}$ sequences (50% of the residues in most informative positions replaced by consensus) performed somewhat worse than searching against global consensus sequences (profile–consensus$^{\mathrm{top50\%}}$, Figure 2A). Interestingly, the search with the least conserved/informative half of the residues replaced by consensus (profile–consensus$^{\mathrm{low50\%}}$, Figure 2A) still improved performance over raw (no consensus) sequences!

Few corrupted profiles can produce many false positives with very significant scores. Alternatively, few very good profiles with many related proteins present in the database can identify them preferentially. Thus, plots of the cumulative number of true versus false positives according to alignment scores may be locally dominated by such a bias. Counting the cumulative number of true positives according to the alignment score rank obtained in each individual query search (i.e. considering the first n alignment pairs from each query) tends to reduce the bias. This test demonstrated that few outliers did not skew the results. Instead, the search against a database of global consensus sequences produced the largest number of true positives at any rank considered (Figure 2B).

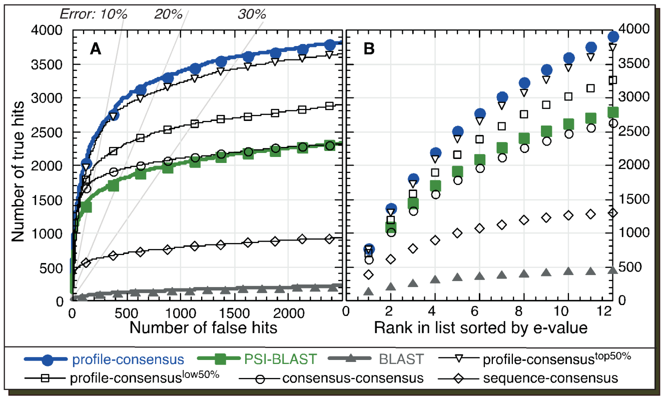

Figure 2. Consensus sequences performed better at any error rate. We compared the performance of BLAST and PSI-BLAST, with different strategies for consensus add-ons profile-consensus marked our standard approach of aligning a PSI-BLAST profile of the query against a database of consensus sequences (blue circles); profile-consensus$^{\mathrm{top50\%}}$ aligned query profiles against a database in which only the 50% most informative residues (Methods) were replaced by consensus sequence (black inversed triangles); profile-consensus$^{\mathrm{low50\%}}$ aligned query profiles against a database in which only the 50% least informative residues were replaced by consensus sequence (black rectangles); consensus-consensus marked BLAST-based comparisons between consensus sequences on both sides, i.e. for the database and the query (black circles); sequence-consensus were BLAST-based comparisons between native sequences on the query side and a database with consensus sequences (black diamonds). For reference, results of original sequence-based PSI-BLAST (green rectangles), and pairwise BLAST (gray triangles) are also shown. True pairs were sequences from the same SCOP superfamily (similar structure), while false ones belonged to different SCOP folds (different structure) (Methods). (**A**) Alignments (2476 sequences, all versus all) were sorted by e-values. True versus false computed over all matches found below a given e-value threshold. By construction, we excluded all pairs that were trivially related (Methods), which explained why the curves for the pairwise BLAST were so low. Profile alignments of global consensus sequences performed best. The transparent gray lines marked the levels of 10, 20 and 30% errors. For instance, at the 10% error (90% accuracy) level, the profile-based search of global consensus sequences revealed over 66% more correct relations than PSI-BLAST (global-consensus-based = 2483 true positives; PSI-BLAST = 1490). (**B**) To rule out that the improvements of consensus sequence-based searches (A) originated from few families, we counted the cumulative number of correctly classified pairs (structural similarity recognized) for the first best scoring n alignment pairs (rank n) from each query search (i.e for rank n equal 2 we looked at 4952 pairs (2 times 2476). The searches of global consensus sequences performed best at all ranks.

Little additional CPU needed for add-on. In this study, we used separate databases for iterative derivation of PSSMs (non-redundant UniProt) and for the final search and alignment ('sequence unique' SCOP; Figure 1). In this scenario, our un-optimized add-on consensus search and alignment nearly doubled the CPU time, in the following way. Five iterations of PSI-BLAST against SCOP would take about 5 s (on a single 2.8 GHz CPU with 1 GB RAM), one additional iteration of PSI-BLAST with the consensus sequence added another ~4 s. Most of the 4 s were spent on the search (3.2 s); very little additional time was needed to translate alignments of consensus sequences into 'raw' sequence alignments.

Note that we actually ran PSI-BLAST against UniProt, while we only applied the consensus addition to SCOP. The entire PSI-BLAST search against UniProt took about 5 min per query. If testing the consensus method on UniProt, we expect that this would lead to an additional 4 min of processing time.

Better alignments. We studied the alignment quality of the best performing consensus sequence search algorithm (profile–consensus) and compared it with the quality of raw sequence alignments, i.e. the original PSI-BLAST alignments. The quality was measured by assessing the 3D structure models that resulted from a simple comparative model-building strategy using these alignments. To provide a useful perspective on the results, we also evaluated 3D models obtained from structural superpositions carried out with MAMMOTH (42). We found that on average consensus sequence-based models had significantly more (not necessarily consecutive) residues in the vicinity of experimentally determined coordinates than did PSI-BLAST-based models (Table 1). This was true when measuring detailed structural similarities (stringent distance threshold of 2 Å) as well as when measuring coarse-grained structural similarities (relaxed distance threshold of 5 Å), and it was true for both levels of alignment difficulty (Table 1). However, the improvements from our add-on of consensus sequences were most significant for more difficult data sets and for more coarse-grained similarities. The comparisons with the models obtained from structural superpositions by MAMMOTH further underscored the relative significance of the gains from consensus-based searching. For example, at the threshold of 5 Å consensus-based searches increased the number of correctly modeled residues around half as much as MAMMOTH did. Surprisingly, at the stringent 2 Å threshold and the less difficult (Table 1; PSI-BLAST e-values 10^{-3}–10) consensus-aligned models had, on average, more superposed residues than the MAMMOTH models. This is rather surprising because it implies that the sequence-only alignment found a better superposition for two 3D objects than did the structure-only alignment method. The likely explanation for this puzzling finding is that MAMMOTH was optimized for the identification of structural similarities within 4 Å, i.e. a threshold more useful for more distantly related structures. In other words, structural alignments of MAMMOTH were likely not optimized for finding 'tight alignments' of closely related proteins. Nevertheless, the performance of consensus sequence-based models generated without structures was impressive in this case.

For local subsets of consecutive model residues we also found that the models resulting from the consensus alignments had longer segments of 'good quality' than did PSI-BLAST based models (Table 2). This was true for a more stringent RMSD threshold of 2 Å as well as for a more relaxed threshold of 5 Å. Again, the improvements from our add-on of consensus sequences were most

Table 1. Consensus sequences improve the global quality of structural models*

	<2 Å (C_α distance)			<5 Å (C_α distance)		
	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH
SCOP superfamily, only	15.8 (±0.2)	**18.1** (±0.2)	19.6 (±0.2)	22.6 (±0.2)	**27.2** (±0.3)	35.2 (±0.3)
PSI-BLAST e-values 10^{-3}–10	34.7 (±0.4)	**38.3** (±0.5)	36.5 (±0.4)	49.1 (±0.5)	**55.2** (±0.6)	58.0 (±0.5)

*For each protein in our data sets (query Q), we aligned a similar protein in the PDB (template T) and used the experimental structure of T to model the structure for Q by simply copying the C_α backbone of T onto Q according to the alignment provided. Since for all Qs in our experiment the correct answer was known (all Qs had known structure), we could then assess how accurate the model was by superposing the model and the known structure. For this superposition, we used the structural alignment method LGA. Here, the measure of accuracy was the percentage of C_αs that were closer to the real structure than some distant cutoff (<5 Å for the three rightmost columns, and <2 Å for columns 2–4). Note that the set of residues below a distance threshold was not necessarily consecutive in sequence. We compared the consensus sequence-based approach with that of the regular PSI-BLAST. The data for MAMMOTH was generated by optimally superposing the structures of Q and T without considering their sequences. In principle, this approximated an upper threshold for performance (Results). The two rows distinguished different data sets corresponding to different levels of alignment difficulty: '***SCOP superfamily only***' were pairs of proteins that fell into different SCOP families and into the same SCOP superfamily (coarse-grained structural relation), while '***PSI-BLAST e-values 10^{-3}–10***' were pairs of proteins with similar structure that fell into the corresponding interval of sequence similarity. Note that both rows reflected the performance for 'non-trivial' tasks. Standard errors are given in parentheses.

Table 2. Consensus sequences improve the local quality of structural models*

	<2 Å (C_α RMSD)			<5 Å (C_α RMSD)		
	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH
SCOP superfamily, only	14.4 (±0.1)	**15.7** (±0.2)	16.5 (±0.2)	23.8 (±0.3)	**27.9** (±0.3)	35.3 (±0.3)
PSI-BLAST e-values 10^{-3}–10	26.8 (±0.4)	**28.0** (±0.4)	26.1 (±0.4)	51.6 (±0.6)	**58.4** (±0.7)	62.8 (±0.7)

*Data sets identical to those as in Table 1; the difference is that accuracy is now measured by considering a single sequence-consecutive segment in the model that falls below a certain distance threshold. The longest consecutive segments were identified by the program LGA. Note that thresholds reflect cutoffs in terms of C_α RMSD, i.e. the distance averaged over the entire segment. In contrast, the values in Table 1 reflect actual C_α thresholds for spatial distances.

significant for more difficult data sets and for more coarse-grained similarities.

DISCUSSION

Here we demonstrated that both the search and alignment quality of PSI-BLAST can easily be improved without having to alter the code. Performance improved substantially with simply replacing the last iteration of the standard PSI-BLAST search against a database of raw sequences with a search against a database of consensus sequences. The improvements were most significant for non-trivial tasks such as the identification (Figure 2) and alignment of distant structural similarities. All improvements translated directly into better initial models for comparative modeling (Tables 1 and 2).

The analysis provided a worst-case scenario for the performance of consensus sequences resulting from simply piggybacking a new idea (usage of consensus sequences directly for the alignment) onto an old method (PSI-BLAST). We neither altered gap penalties (11 for opening and 1 for extension), nor substitution matrices, nor any other parameter optimized for raw rather than consensus sequences. Preliminary tests (data not shown) indicated that consensus sequence-based searches did not change the robustness/sensitivity with respect to such parameters. We also found that using the most frequent amino acid type at each position instead of the amino acid with maximal PSSM score did not reduce the gain significantly. On the other hand, the adverse consequence of not optimizing any of the PSI-BLAST parameters was that searching a database of consensus sequences took almost four times as long as searching a comparable database of raw sequences (~3.2 versus ~0.8 s on a non-redundant SCOP). Lately, we have realized that it was largely due to using parameters such as thresholds for extending hits (high-scoring residue words), triggering gapped alignments and gap penalty values themselves that were not optimal for consensus sequences (our preliminary results indicate that raising the threshold for extending hits by about 20% almost doubles the speed and affects the sensitivity negligibly). Those details, as well as the scoring matrix, remain to be optimized for the particular concept of consensus sequences.

To generate global consensus sequences, we replaced each amino acid in the template by the amino acid that scored highest in the associated column of a profile PSSM produced by a standard PSI-BLAST search. Thereby, we maximized the self-score of the resulting consensus sequence with respect to its PSSM. As a consequence, any two proteins having similar profiles are also likely to have a higher alignment score when consensus sequences are aligned. Our results suggest that the corresponding change of the alignment score for unrelated proteins was considerably smaller. Surprisingly, replacing only the least informative half of all residues by consensus also improved performance (profile-consensus$^{\mathrm{low50\%}}$, Figure 2). This may suggest that even weakly or non-conserved positions are associated with specific constraints on random amino mutations that can be utilized to detect similarities.

The best performance of profile–consensus search was achieved when the profile that was used to generate the consensus sequence was obtained in the same way as the profile used for the alignment scoring. For example, when the profile used to compute the consensus was obtained after fewer PSI-BLAST iterations, performance deteriorated. Improving the searches through consensus databases that apply more involved ways of using consensus sequences such as ProDom and COBBLER may therefore require one to search with the same type of scoring profiles that was used to generate the database in the first place. Unfortunately, the algorithms used for their creation are considerably more involved and more time consuming. In contrast, our add-on protocol is very simple. The global consensus sequences can be generated easily from PSI-BLAST ASCII matrices. The optimal search of such database requires similarly easily obtainable PSI-BLAST binary profiles. Any PSI-BLAST user could easily accomplish this. However, the generation of a large consensus database is computationally costly. Therefore, we decided to provide an up-to-date consensus sequence version of Swiss-Prot (46) and PDB (47) through our website (http://www.rostlab.org/services/consensus/). We plan to provide consensus sequences for the entire UniProt in the near future. We have also provided a simple Perl program for translating PSI-BLAST ASCII matrices into consensus sequences. In addition, for the convenience of users we have provided a script for the conversion of aligned consensus sequences into the corresponding alignments of real sequences. We have also made profile–consensus searches available through the PredictProtein server (48) (http://predictprotein.org).

Our results suggested that sequence–profile method (i.e. methods that search database of profiles with a sequence) such as IMPALA and the methods used to search CDD (25,26) might also benefit from mimicking profile–profile alignments through searching database of profiles with a consensus sequence (consensus–profile alignment). Similarly, methods that use sequences to search HMM-derived profile databases such as in Pfam and SMART might also improve performance by replacing a raw query sequence with a consensus sequence as proposed in this manuscript, although the HMM-derived consensus sequences may be more appropriate (33). Finally, it is also likely that methods using bidirectional profile–sequence/sequence–profile scoring (49,50) will benefit from using profile–consensus/consensus–profile approach.

One advantage other than improved performance is that consensus sequence-based alignments are likely less sensitive to sequencing errors. This may be particularly appealing in the age of massive sequencing efforts that grind up indiscriminately what is found in oceans, soils and polluted environments. Finally, it remains to be shown that the advantage of using consensus sequence-based searches for the identification and alignment of remote structural similarities between proteins will hold more generally, e.g. for the nucleotide sequences, and for the usage of with other alignment algorithms, such as ClustalW or T-Coffee.

One consequence of our improvements was that the consensus sequence-based alignment profiles were both more diverse and more accurate than those generated by the ordinary PSI-BLAST. Prediction methods that use alignment profiles, such as those predicting aspects of protein structure, tend to improve proportionally with better profiles (51–54). It is therefore reasonable to assume that our consensus sequence add-on to PSI-BLAST will clearly boost the performance of downstream methods for the prediction of protein structure and function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Thanks to Henry Bigelow, Eileen Guilfoyle, and Kazimierz Wrzeszczynski for reading and commenting on the manuscript and to Guy Yachdav for help in setting up the server. Thanks to all those who develop alignment tools and make them available to the public. In particular thanks to people developing and providing support for the PSI-BLAST program. Thanks to Adam Zemla for providing LGA and to the creators and developers of MAMMOTH. The support for the work of DP and BR and funding to pay the Open Access publication charge was provided by the grant U54-GM074958-01 to the Northeast Structural Genomics consortium (NESG) from the Protein Structure Initiative (PSI) of the National Institutes of Health (NIH) and by the grant RO1-LM07329-01 from the National Library of Medicine (NLM). Last but not least, thanks to all of those who maintain excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
3. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
4. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
5. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
6. Altschul,S., Madden,T., Schäffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
8. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
9. 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.
10. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
11. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
12. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
13. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
14. Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.
15. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.
16. 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.
17. Patthy,L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.*, **198**, 567–577.
18. Sonnhammer,E.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3**, 482–492.
19. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinformatics*, **3**, 246–251.
20. Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.
21. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–196.
22. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
23. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
24. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
25. Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
26. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
27. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–251.
28. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–260.
29. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool:

identification of signaling domains. *Proc. Natl. Acad. Sci. USA*, **95**, 5857–5864.
30. Thelen,M.P., Venclovas,C. and Fidelis,K. (1999) A sliding clamp model for the Rad1 family of cell cycle checkpoint proteins. *Cell*, **96**, 769–770.
31. Kryshtafovych,A., Venclovas,C., Fidelis,K. and Moult,J. (2005) Progress over the first decade of CASP experiments. *Proteins*, **61** Suppl 7, 225–236.
32. Rychlewski,L. and Fischer,D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.*, **14**, 240–245.
33. Kahsay,R.Y., Wang,G., Gao,G., Liao,L. and Dunbrack,R. (2005) Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. *Bioinformatics*, **21**, 2287–2293.
34. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–119.
35. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
36. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
37. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
38. Fiser,A., Feig,M., Brooks,C.L. 3rd and Sali,A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.*, **35**, 413–421.
39. Ginalski,K. (2006) Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 172–177.
40. Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
41. Moult,J., Fidelis,K., Rost,B., Hubbard,T. and Tramontano,A. (2005) Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins*, **61** Suppl 7, 3–7.
42. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
43. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
44. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
45. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
46. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
47. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
48. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–326.
49. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
50. Przybylski,D. and Rost,B. (2004) Improving fold recognition without folds. *J. Mol. Biol.*, **341**, 255–269.
51. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
52. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.
53. Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
54. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.