BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 8 2005, pages 1451–1456
doi:10.1093/bioinformatics/bti233

Sequence analysis

A structure-based method for protein sequence alignment

Maricel G. Kann, Paul A. Thiessen, Anna R. Panchenko, Alejandro A. Schäffer, Stephen F. Altschul and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20894, USA

Received on October 5, 2004; revised on December 2, 2004; accepted on December 15, 2004
Advance Access publication December 21, 2004

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022

ABSTRACT

Motivation: With the continuing rapid growth of protein sequence data, protein sequence comparison methods have become the most widely used tools of bioinformatics. Among these methods are those that use position-specific scoring matrices (PSSMs) to describe protein families. PSSMs can capture information about conserved patterns within families, which can be used to increase the sensitivity of searches for related sequences. Certain types of structural information, however, are not generally captured by PSSM search methods. Here we introduce a program, Structure-based ALignment TOol (SALTO), that aligns protein query sequences to PSSMs using rules for placing and scoring gaps that are consistent with the conserved regions of domain alignments from NCBI's Conserved Domain Database.

Results: In most cases, the alignment scores obtained using the local alignment version follow an extreme value distribution. SALTO's performance in finding related sequences and producing accurate alignments is similar to or better than that of IMPALA; one advantage of SALTO is that it imposes an explicit gapping model on each protein family.

Availability: A stand-alone version of the program that can generate global or local alignments is available by ftp distribution (ftp://ftp.ncbi.nih.gov/pub/SALTO/), and has been incorporated to Cn3D structure/alignment viewer.

Contact: bryant@ncbi.nlm.nih.gov

INTRODUCTION

The ongoing completion of genome projects presents scientists with the regular and increasing challenge of processing and understanding proteins using only their amino acid sequences. This has spurred the creation of protein family databases, as well as the design of programs for protein classification and annotation and the recognition of distant protein relationships. Software tools such as SSEARCH (Smith and Waterman, 1981; Pearson, 1991), FASTA (Pearson and Lipman, 1988), BLAST (Altschul *et al*., 1990), PSI-BLAST (Altschul *et al*., 1997), IMPALA (Schäffer *et al*., 1999), SAM (Karplus *et al*., 1998) and HMMER (Eddy, 1998), among others, have become popular due to their considerable speed and sensitivity in retrieving homologous sequences. Specifically, hidden Markov models (HMMs) and the related position-specific scoring matrices (PSSMs), derived from multiple alignments of related sequences, have been shown generally to increase the sensitivity of such searches (Gribskov *et al*., 1987; Baldi *et al*., 1994; Krogh *et al*., 1994; Altschul *et al*., 1997; Henikoff and Henikoff, 1997; Durbin *et al*., 1998; Eddy, 1998; Karplus *et al*., 1998; Park *et al*., 1998; Schäffer *et al*., 1999).

Explicitly, including structural information in these models presents a challenge. This can be addressed to some extent by modeling the existence and lengths of gaps which correspond to the loop regions between secondary structure elements in various positions and by requiring the alignment of certain key residues or motifs. In some sequence comparison programs, however, gap penalties are chosen *ad hoc* and key residues are aligned only if this will increase the alignment score. Here, we introduce a method, Structure-based ALignment TOol (SALTO), that can be used to generate global or local alignments between a query sequence and a PSSM derived from multiple alignments from the Conserved Domain Database (CDD) (Marchler-Bauer *et al*., 2003). SALTO forces the alignments it produces to incorporate structural features of the CDD alignments. Because the statistics of global alignment scores are not well understood, this paper will focus only on SALTO in its local alignment mode (SALTO_LOCAL).

CDD multiple alignments are curated manually to be consistent with structure–structure alignments. Each CDD alignment consists of a set of conserved core structures, or blocks. It has been shown that alignments based upon only these blocks can predict experimentally determined functional sites (Panchenko *et al*., 2004). SALTO constructs position-specific scoring matrices (PSSMs) from multiple alignments in a manner similar to IMPALA and its heuristic cousin RPS-BLAST, but it makes use of CDD's block structure by aligning only PSSM positions that derive from these blocks; PSSM positions outside the blocks, possibly corresponding to structural loops, are not aligned. The alignments SALTO produces are thus expected to include those residues that are key to protein structure and function. The program takes advantage of CDDs, explicit gap model by allowing gaps (at no alignment cost) only between blocks and therefore presumably within structural loops.

We have found that when SALTO is used in its local alignment mode, for most PSSMs the alignment scores it produces for sets of random sequences fit an extreme value distribution. Even though SALTO constrains its alignments to agree with structural evidence, its performance in a search for distant biological relationships is very similar to that of IMPALA. SALTO is now included in the Cn3D (Wang *et al*., 2000) protein structure viewer (Cn3D is available at http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) as an

*To whom correspondence should be addressed.

© The Author 2004. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

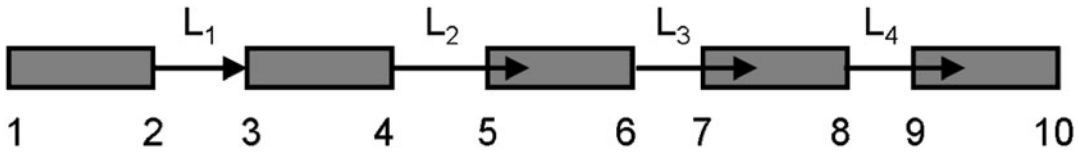

Fig. 1. Diagram showing a typical PSSM derived from CDD, where conserved regions are represented as blocks and non-conserved regions or loops of maximum length L_n are depicted as arrows. In SALTO_LOCAL a query sequence can be aligned to a PSSM starting at the beginning of any block and terminating at the end of the same or later block. For example, an alignment beginning at position 3 can end at 4, 6, 8 or 10. The maximum length of the corresponding loops L_n is determined by user-set parameters (see text for details).

option for aligning new sequences and is also available by ftp distribution (ftp://ftp.ncbi.nih.gov/pub/SALTO/).

MATERIALS AND METHODS

Database of Conserved Domains

SALTO uses the PSSMs and the definition of blocks or conserved regions from CDD (Marchler-Bauer *et al*., 2003). CDD contains curated domain alignments, many originally imported from Smart (Letunic *et al*., 2004) and Pfam (Bateman *et al*., 2002). CDD alignments are refined using known three-dimensional structural information and structure–structure alignments and conserved regions are identified that are likely to be present in all family members.

For this study, we began with a set of 271 alignments from CDD version 1.60, the current version of which is available at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml (Marchler-Bauer *et al*., 2003). We excluded from this test set 73 alignments containing repetitive regions. A PSSM was calculated for each of the remaining 198 alignments, using the procedure employed by PSI-BLAST, IMPALA and RPS_BLAST (Altschul *et al*., 1997). To compare alignments involving different PSSMs with one another, reasonably accurate *E*-values are required, so it is important to be able to approximate the random score distribution of any PSSMs included in a search database. As described below, for the parameters chosen for this study, 170 of the 198 PSSMs we constructed are well described by an extreme value distribution, hence we confine database searching using SALTO to this subset of PSSMs.

Algorithm

We outline the program SALTO, which uses dynamic programming to find an alignment with maximum score between a query protein sequence and a PSSM, subject to certain constraints. Controlled by a command line option, SALTO can be used in either global or local mode.

In its local mode, SALTO is similar to NCBI's IMPALA and RPS_BLAST, but uses a different approach to aligning a sequence with a PSSM and to introducing and scoring gaps. IMPALA and RPS_BLAST allow gaps anywhere in the alignments they produce and charge each gap a penalty that is an affine function of its length. In contrast, SALTO constrains the location and length of gaps using structure-based information from the CDD database, but does not penalize for their existence.

CDD partitions its multiple alignments into conserved blocks, which do not contain gaps, separated by regions of variable length within the sequences represented. SALTO considers this block structure to be inherited by the PSSMs constructed from the CDD alignments and uses it when aligning a sequence to a PSSM. SALTO aligns a new sequence only to the conserved blocks of a PSSM and no gaps are allowed within each block-alignment. An alignment must involve complete blocks and if a pair of blocks is contained in an alignment, then all intermediate blocks must be included as well. In SALTO's global mode, all blocks must be aligned, but in its local mode, only a contiguous subset (Fig.1).

Sections of a sequence between two regions aligned to blocks are considered 'loops' and a two-parameter constraint is placed on the maximum allowed length of each loop. Specifically, the loop between the matching sequence pieces aligning to blocks i and $i+1$ must be no larger than a constant (a) plus a chosen percentile (p) of the loop lengths between blocks i and $i+1$ in the CDD multiple alignment. This captures the observed variability of loop lengths in the CDD database, but generally allows for further variation. SALTO uses dynamic programming to find an alignment with maximum score, subject to these structure-based constraints.

Score statistics

To evaluate the sequence similarities returned by a database search method, it is important to be able to estimate the level of similarity one may expect to arise by chance. Analytic results (Karlin and Altschul, 1990; Dembo *et al*., 1994) show that, using standard substitution matrices, the scores of optimal ungapped alignments of random sequence pairs approach a Gumbel or extreme value distribution (EVD) (Gumbel, 1958). Under certain conditions, the same holds true empirically when gaps are allowed (Smith *et al*., 1985; Pearson, 1998; Mott, 2000; Altschul *et al*., 2001), or when sequences are aligned with PSSMs (Altschul *et al*., 1997). Briefly, the probability that a random score $S \geq x$ is given by

$$P(S \geq x) = 1 - \exp(-Kmn\mathrm{e}^{-\lambda x}) \quad (1)$$

where m and n are the lengths of the sequences compared, λ is the 'scale' parameter of the EVD and K is related to the EVD's characteristic value u by

$$K = \frac{\mathrm{e}^{\lambda u}}{mn}. \quad (2)$$

Although one may expect that random scores for the type of alignment constructed by SALTO_LOCAL will also be well approximated by an extreme value distribution, random simulations are needed to determine whether this is the case and, if so, to estimate the statistical parameters K and λ for any particular PSSM.

To generate a distribution of random scores, we modeled a protein as a sequence of independently selected amino acids, chosen using standard 'background' frequencies (Robinson and Robinson, 1991). Each PSSM was aligned, for several gap-length parameters p and a (described in the previous section), to 6000 random sequences of length 1000, and an EVD was fitted to the resulting collection of optimal scores using the maximum-likelihood method (Lawless, 1982). In each instance, a χ^2 goodness-of-fit test was performed between the EVD and the data from which it was estimated. For the great majority of PSSMs, there was a range of gap-length parameters for which the EVD could not be rejected, i.e. it had χ^2 P-value $\geq$0.005 (see Results section). Because reasonably accurate statistics are needed to compare alignments involving different PSSMs, these parameter ranges are to be preferred. P-values can then be reported using Equation (1), or E-values using the equation

$$E = Kmn\mathrm{e}^{-\lambda x}. \quad (3)$$

When the χ^2 P-value is <0.005, such EVD P-value or E-value estimates for SALTO_LOCAL alignment scores, based upon a questionable model of the random distribution, may be inaccurate. Note that bias in the amino acid composition of protein families is another potential source of inaccuracy in estimating E-values (Schäffer *et al*., 1999, 2001). This can be mitigated by estimating the EVD parameters for a given PSSM using, in place of standard amino acid frequencies, those that characterize the protein family used to generate the PSSM.

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022

A benchmark for comparing search methods

In this paper, we study the ability of a search of a PSSM database to find domains related to a query protein sequence. For this purpose, we will not record a high-scoring alignment between a sequence and a related PSSM as a true positive unless the alignment involves the correct region of the query sequence. Therefore, to evaluate the sensitivity of the search method, we require not only a test set of sequences and PSSMs and a list of 'true positive' relationships between them, but also a record of the sequence regions in which the relationships are valid.

We have described above the collection of 170 PSSMs used, each corresponding to a domain from CDD, with an associated multiple alignment and structures. A set of 6480 'non-redundant' query sequences was employed. This set was constructed by single-linkage clustering, based on BLAST E-values of 10^{-80} or less, of the entries in the MMDB structure database (Chen *et al*., 2003), which is available at http://www.ncbi.nlm.nih.gov/Entrez/. We recorded a sequence–PSSM pair as a true positive if it satisfied one of two criteria described below. For each criterion, we describe as well the region of the sequence in which the relationship is considered 'valid'. The test sets, annotated list of true positives and valid sequence regions are available from the authors.

First, each CDD family contains a segment or segments from at least one sequence with known structure. We defined the region from the first sequence residue represented by the CDD family to the last as the CDD family footprint. A CDD family may contain several sequences with structures and each then has a family footprint. Using the VAST structure alignment algorithm (Gibrat *et al*., 1996), all structures represented by a CDD family are compared to all protein structures in db6480; complete protein chains or MMDB domains are employed for this comparison, not just those regions aligned within the CDD family. The list of structurally similar VAST neighbors is available at http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml. The region between the first and last residue aligned by VAST defines a VAST footprint in the CDD family sequences. If the overlap of the CDD family footprint and the VAST footprint comprises at least 80% of each of their lengths, then the relationship is taken to be a true one. The footprint of the VAST alignment within the db6480 sequence is recorded as the valid region. If more than one structure from a CDD family implies a true relationship to a db6480 sequence S, then the entire region between the first and last residues in S contained in some valid region is considered valid. Note that VAST has been compared to other methods for structural alignment and classification and has been shown to have comparable sensitivity (Marchler-Bauer *et al*., 1997; Matsuo and Bryant, 1999; Shapiro and Brutlag, 2004; Sierk and Pearson, 2004).

Second, using BLAST for each sequence segment from one of the CDD families represented and PSI-BLAST for each corresponding multiple alignment, a search of db6480 was performed. Whenever a hit with E-value 10^{-4} or less was returned, a true relationship between the sequence involved and the corresponding PSSM was recorded. The valid region within the db6480 sequence was determined as above, using BLAST and PSI-BLAST footprints in place of VAST footprints.

Our list of true relationships is a challenging one for database search programs. Each of the methods described above for inferring a relationship is triggered by a VAST, BLAST or PSI-BLAST alignment. A histogram of the percent identity implied by these alignments is shown in Figure 2. For PSI-BLAST alignments, percent identity is calculated using a consensus sequence for the multiple alignment used to construct the PSSM. As can be seen, most relationships are inferred from alignments with $<25\%$ sequence identity. The great majority of these alignments derive from VAST structure relationships, with only $\sim$10% of the relationships inferred using only the second criterion above.

Search performance

There are various ways to quantify the performance of a search method on a given test set, but recently receiver operating characteristic (ROC) analysis has been gaining widespread acceptance (Bamber, 1975; Swets, 1988). Assume that the results of a search are returned in a specific order and that they may

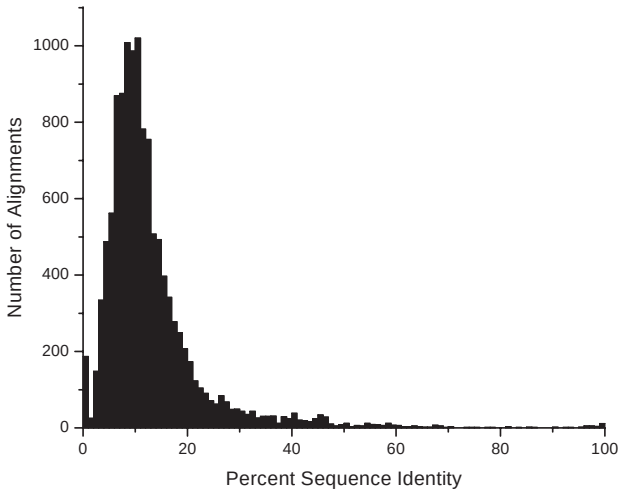

Fig. 2. Distribution of the percent sequence identity among related pairs of proteins. This test set was used for measuring the effectiveness of SALTO and IMPALA as search tools. The majority of the sequence pairs were identified by VAST as structurally similar proteins.

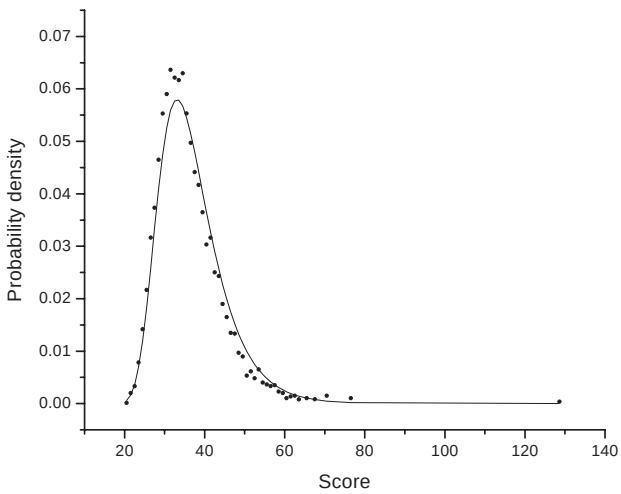

Fig. 3. SALTO_LOCAL alignment scores (dots) and the best fitting EVD with parameters $\lambda = 0.153$ and $K = 0.0026$ (solid curve). The scores are from local alignments of the PSSM for the glycoprotein hormone β-chain homologue CDD family (GHB) and 6000 random sequences generated using a standard amino acid distribution (Robinson and Robinson, 1991).

be classified as either true (related) or false (unrelated). Descending this list, one may plot the ROC curve—the aggregate numbers of true versus false positives returned. A ROC_n curve is produced if one truncates the plot after the first n false results (McClish, 1989; Gribskov and Robinson, 1996). An ideal retrieval method will return all the true results before any false ones. The area under a ROC curve is normalized so that ideal retrieval receives a score of 1.0 and this is called a ROC score.

Here, we are interested not only in whether a search program identifies a related PSSM for a given query, but also in whether the returned alignment involves the correct region of the PSSM. A variation on ROC and ROC_n curves appropriate to this situation is the 'localization-response operating characteristic' (LROC) curve (Chakraborty, 1989; Swensson, 1996; Edwards *et al*., 2002). Applied to our case, the number of related PSSMs returned with an alignment in the correct region is plotted against the number of unrelated

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022

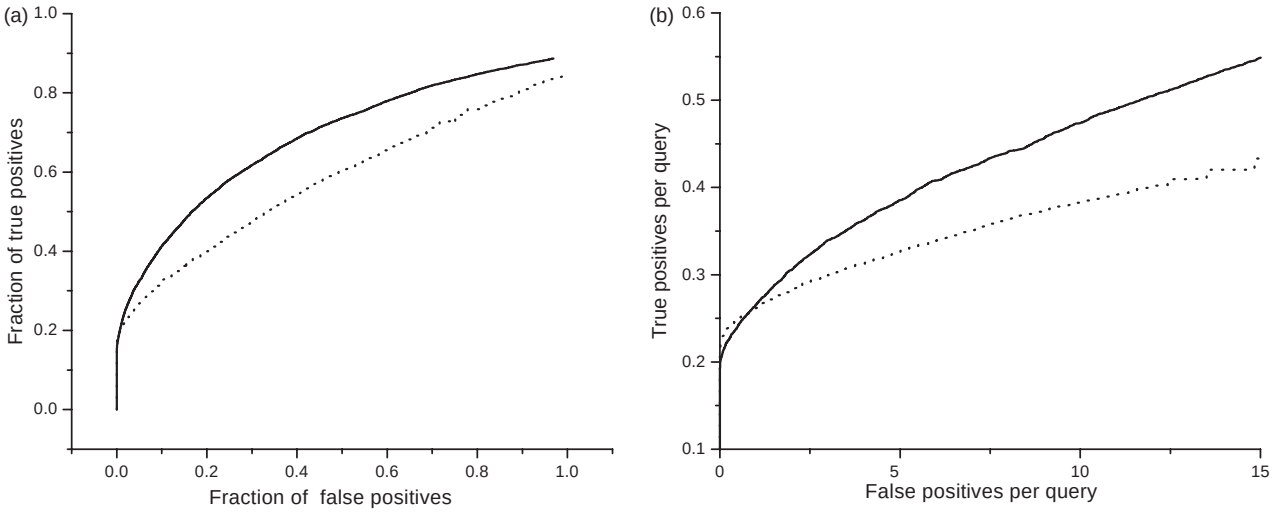

Fig. 4. Localization-response operating characteristic (LROC) curves. SALTO_LOCAL (solid) is compared to IMPALA (dotted). (**a**) Full LROC curves. (**b**) First portion of the LROC curve; values are normalized for the total number of queries.

Table 1. Localization-response operating characteristic ($LROC_n$) at various false positive thresholds

	$ROC_{10\,000}$	$ROC_{25\,000}$	$ROC_{100\,000}$
SALTO_LOCAL	0.185 ± 0.005	0.213 ± 0.005	0.300 ± 0.005
IMPALA	0.187 ± 0.005	0.205 ± 0.005	0.258 ± 0.005

All search results of the benchmark query test sets, ranked by E-value, are pooled.

PSSMs returned. Correct PSSMs returned with an alignment in the wrong position are ignored; this is equivalent to a ROC curve in which such correct PSSMs with incorrect alignments are deemed to have been returned after all other PSSMs. With certain reasonable assumptions, it can be shown that, stochastically, LROC scores are linearly related to ROC scores (Swensson, 1996).

In this paper, when using SALTO or IMPALA to compare a sequence to a CDD-derived PSSM, we record a result as a true positive only if it involves a CDD determined to be related to the query sequence in question and the alignment produced overlaps at least 1% of the 'valid' sequence region.

We compare the performance of SALTO_LOCAL and IMPALA on the benchmark test set described above using $LROC_n$ scores. E-values can be used to combine the results of multiple searches into a single retrieval list, and a single $LROC_n$ score may be calculated. Alternatively, $LROC_n$ scores can be calculated for individual queries and an average score over a query set may then be used to compare the performance of different search methods. $LROC_n$ errors were estimated by a bootstrap procedure (Schäffer *et al*., 2001).

RESULTS

Loop length parameters

Just as each PSSM is tailored to a particular CDD family, the loop length parameters a and p used by SALTO are similarly tailored. In general, if these parameters are too loose or too restrictive, i.e. if they allow the introduction of long or only short loops in the alignment, respectively, the resulting random score distribution will not follow an EVD and it will not be appropriate to use the PSSM in a search database.

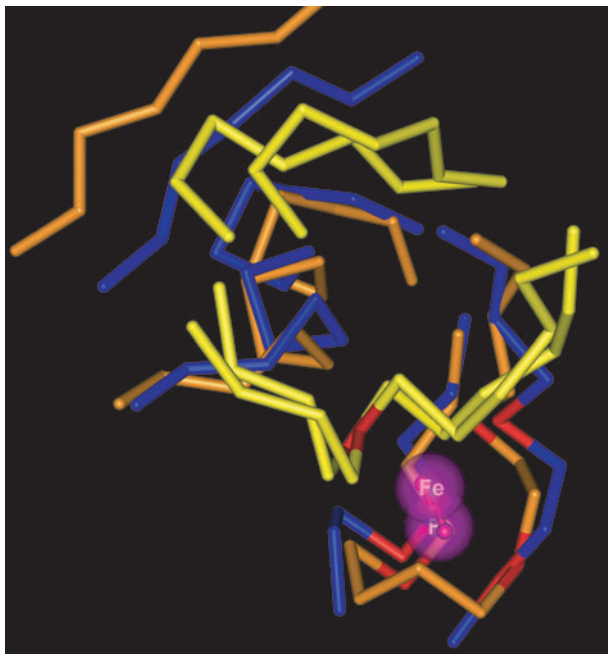

Fig. 5. 3D representation (using Cn3D) of the regions of two proteins with the 2Fe–2S iron–sulfur domain aligned by SALTO_LOCAL: PDB sequence1B9R_A (blue tubes), is shown aligned to1JRO_A (orange tubes). 1B9R_A is the representative sequence of the 2Fe-2S–iron-sulfur-cluster-binding domain (fer2) CDD family, and 1JRO_A is a structurally similar protein found by VAST. The four cysteins of the iron-binding site are highlighted in red. Yellow highlights a region aligned by SALTO_LOCAL but not by IMPALA (as shown in Fig. 6).

Among the 198 PSSMs studied, we found 170 for which the range of loop length parameters tested yielded random scores well approximated by an EVD. The remaining 28 families failed the test at χ^2 P-value ≥ 0.005; we further analyzed why they failed the

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022

```
SALTO
pdb|1B9R|A    3    VVFIdeqsgEYAVDAQd-GQSLMEVATQNGVPgIVAECGGScVCATCRIEIed   54
pdb|1JRO|A    3    IAFLln-geTRRVRIEdpTQSLLELLRAEGLTgTKEGCNEG-DCGACTVMIrd   53

pdb|1B9R|A   55    awveivgeanpdendlqstgepmtAGTRLSCQVFIdpsMDGLIVR   100
pdb|1JRO|A   54    a--------------------agSRAVNACLMMLp-qIAGKALR    76

IMPALA
pdb|1B9R|A    3    VVFIDEQSGEYAVDAQDGQSLMEVATQNGVPGIVAECGGSCvCATCRIEIED    54
pdb|1JRO|A    3    IAFLLNGETRRVRIEDPTQSLLELLRAEGLTGTKEGCNEGD-CGACTVMIRD    53
```

Fig. 6. Alignment of the representative sequence of the 2Fe–2S iron–sulfur-cluster-binding domain (fer2) CDD family, PDB 1B9R_A, with query sequence 1JRO_A using (**a**) SALTO_LOCAL and (**b**) IMPALA. Aligned residues are represented in bold uppercase, and unaligned residues in lower case. The iron-binding site, represented by four cysteins, is highlighted in red. The box highlights the binding site region aligned by SALTO_LOCAL and not by IMPALA.

test. We found that from those 28 entries, 10 CDD entries that were originally hand-curated to have three conserved regions or blocks, contain one block that dominates in size. Another nine CDD entries contain only one or two long blocks. For these entries, the random alignments produced by SALTO are virtually always to the same, single block. In these cases a good fit to the EVD cannot be expected (Gumbel, 1958). This particular problem is easy to avoid; newer versions of CDD may be changed so that long blocks are split and users may also edit these family models themselves for in-house searches.

Given loop length parameters from which to choose, we preferred the 'loosest' ones, allowing the introduction of large gaps between blocks. To illustrate the fit of an EVD to a random distribution of SALTO_LOCAL scores, we use the glycoprotein hormone β-chain homologue CDD family (GHB). Figure 3 shows a histogram of random alignment scores when the loop length parameters $p = 0.99$ and $a = 6$ are used; the χ^2 goodness-of-fit test, to an EVD with parameters $\lambda = 0.153$ and $K = 0.0026$, has P-value 0.88. For this CDD family, goodness-of-fit P-values >0.005 (corresponding to a good fit to an EVD) were obtained with a wide range of loop length parameters: values of p ranging between 0.20 and 0.99, each with several values of a. The parameters used for Figure 3 are the largest that yielded a good fit to the EVD and accordingly are those we associate with this CDD family in our database. Loop length parameters chosen for each of the 170 database CDD families are available by ftp (ftp://ftp.ncbi.nih.gov/pub/SALTO/).

Search accuracy

Using query sequences from db6480 and the database of 170 CDD families described above, the relative accuracy of the IMPALA and SALTO_LOCAL search methods, as measured by pooled LROC$_n$ scores, are compared in Table 1. LROC$_n$ scores are reported for $n = 10\,000$, 25 000 and 100 000 which correspond approximately to 1, 4 and 15 false positives per query. By these measures of retrieval accuracy, SALTO_LOCAL performs as well as IMPALA for low false positive values and outperforms IMPALA as one moves further into the 'twilight zone'; and the number of false positives increases (Figure 4a and b).

DISCUSSION AND CONCLUSION

Traditionally, to score insertions or deletions in sequence alignments, affine gap penalties have been used despite the fact that this simple model does not adequately describe the evolution of indels (Qian and Goldstein, 2001; Wrabl and Grishin, 2004). It is useful to have a PSSM-alignment search method that can incorporate models of loop location and length. In contrast to uniform affine gap costs, the model described here accommodates large loops in specified positions without extra cost, where curated protein-family multiple alignments suggest they may occur. This can improve the recognition of members of highly divergent protein families, with long insertions (e.g. SALTO outperforms IMPALA in detecting sequences related to the AAA-ATPase families that contain big domain insertions; data not shown). To illustrate the advantage of using SALTO_LOCAL, the alignment between 1B9R_A, the representative protein sequence of the 2Fe–2S iron–sulfur cluster binding domain (fer2) CDD family and 1JRO_A (a protein query that also contains the 2Fe–2S iron–sulfur domain) is shown in Figures 5 and 6. It can be seen that while SALTO_LOCAL aligns the four cysteins that constitute the iron-binding site (depicted in Figs 5 and 6), IMPALA aligns only three of them. In this example, SALTO_LOCAL has the advantage of assessing no penalty for the inclusion of a big loop before the penultion block, which allows it to incorporate the complete iron-binding site; the alignment score produced by IMPALA decreases considerably if that insertion is included, so IMPALA does not identify the functional site as a whole.

SALTO implements one approach to capturing the structural information encoded in the pattern of ungapped blocks of CDD protein families. One problem any such specialized alignment model presents is the characterization of its statistics. Despite the requirement that specified conserved regions be aligned completely or not at all, for most protein families considered loop length parameters may be found that yield random score distributions well characterized by the EVD. Furthermore, as CDD increases in size, it is expected that models of conserved regions and the lengths of loops that separate them will become more accurate. This may improve the power of any tool, like SALTO, that employs such models in its search algorithm. Thus, in the future, it might prove possible to devise more sophisticated gap penalty functions and use them to improve SALTO's performance. Our observation here that good statistical power is obtained with structure-based alignment suggests that further investigation might be worthwhile. In the context of NCBI's CDD project, we have already found that the methods in SALTO, incorporated into Cn3D, lead to better semi-automatic curation of proteins alignments. The final alignments are easier to

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022

interpret in the context of the structure and function of the protein aligned.

In summary, our approach provides a reasonable and fast alternative method for using structural and functional knowledge from protein family models when aligning new sequences to those models. Our results show that, despite the constraints SALTO's model imposes on the alignments it generates, its retrieval accuracy, when used to search a database of protein models, is in general better than IMPALA's.

ACKNOWLEDGEMENTS

We thank Eva Czabarka and Sergey Sheetlin for calculating errors for the LROC measures, and John Spouge and Aron Marchler-Bauer for helpful discussions and comments on the manuscript. Support for this work was provided by the intramural research program of the National Institutes of Health.

REFERENCES

Altschul,S.F., Bundschuh,R., Olsen,R. and Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

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.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Bamber,D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.*, **12**, 387–415.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Chakraborty,D.P. (1989) Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med. Phys.*, **16**, 561–568.

Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.

Dembo,A., Karlin,S. and Zeitouni,O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, **22**, 2022–2039.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, England.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edwards,D.C., Kupinski,M.A., Metz,C.E. and Nishikawa,R.M. (2002) Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Med. Phys.*, **29**, 2861–2870.

Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput.Chem.*, **20**, 25–33.

Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, NY.

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

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.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Lawless,J.F. (1982) *Statistical Models and Methods for Lifetime Data*. Wiley, NY.

Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32** (Database issue), D142–D144.

Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.

Marchler-Bauer,A., Levitt,M. and Bryant,S.H. (1997) A retrospective analysis of CASP2 threading predictions. *Proteins*, **29** (Suppl), S83–S91.

Matsuo,Y. and Bryant,S.H. (1999) Identification of homologous core structures. *Proteins*, **35**, 70–79.

McClish,D.K. (1989) Analyzing a portion of the ROC curve. *Med. Decision Making*, **9**, 190–195.

Mott,R. (2000) Accurate formula for *P*-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, **300**, 649–659.

Panchenko,A.R., Kondrashov,F. and Bryant,S. (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T., and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Qian,B. and Goldstein,R.A. (2001) Distribution of Indel lengths. *Proteins*, **45**, 102–104.

Robinson,A.B. and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl Acad. Sci. USA*, **88**, 8880–8884.

Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

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.

Shapiro,J. and Brutlag,D. (2004) FoldMiner: structural motif discovery using an improved superposition algorithm. *Protein Sci.*, **13**, 278–294.

Sierk,M.L. and Pearson,W.R. (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci.*, **13**, 773–785.

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

Smith,T.F., Waterman,M.S. and Burks,C. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.*, **13**, 645–656.

Swensson,R.G. (1996) Unified measurement of observer performance in detecting and localizing target objects on images. *Med. Phys.*, **23**, 1709–1725.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.

Wrabl,J.O. and Grishin,N.V. (2004) Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. *Proteins*, **54**, 71–87.

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1451/251038 by guest on 20 August 2022