3772–3778 *Nucleic Acids Research, 2005, Vol. 33, No. 12*
doi:10.1093/nar/gki687

A sequence sub-sampling algorithm increases the power to detect distant homologues

Catrióna R. Johnston* and Denis C. Shields

Department of Clinical Pharmacology, Bioinformatics Group, Royal College of Surgeons in Ireland, 123 St Stephens Green, Dublin 2, Ireland

Received January 12, 2005; Revised and Accepted June 14, 2005

ABSTRACT

Searching databases for distant homologues using alignments instead of individual sequences increases the power of detection. However, most methods assume that protein evolution proceeds in a regular fashion, with the inferred tree of sequences providing a good estimation of the evolutionary process. We investigated the combined HMMER search results from random alignment subsets (with three sequences each) drawn from the parent alignment (Rand-shuffle algorithm), using the SCOP structural classification to determine true similarities. At false-positive rates of 5%, the Rand-shuffle algorithm improved HMMER's sensitivity, with a 37.5% greater sensitivity compared with HMMER alone, when easily identified similarities (identifiable by BLAST) were excluded from consideration. An extension of the Rand-shuffle algorithm (Ali-shuffle) weighted towards more informative sequence subsets. This approach improved the performance over HMMER alone and PSI-BLAST, particularly at higher false-positive rates. The improvements in performance of these sequence sub-sampling methods may reflect lower sensitivity to alignment error and irregular evolutionary patterns. The Ali-shuffle and Rand-shuffle sequence homology search programs are available by request from the authors.

INTRODUCTION

Protein homology has long been used as a means for identifying similarity in protein function or structure, based on the observation that most sequences with extensive similarity usually share an evolutionary ancestor. Sequence homology searches of novel sequences against known protein databases are a key element of sequence and genome annotation. While strong similarities may be readily identified by searching a single sequence against a database [e.g. BLAST searching (1)], weak similarities may be difficult to distinguish from background noise. In this case, the power to detect similarity is increased by searching an alignment against the database, where weighting amino acid probabilities at each residue position (2) provides a more sensitive comparison. One approach is to implement a profile hidden Markov model (HMM) (3). Each profile HMM of a protein alignment incorporates information from all proteins present in the alignment. Alignments containing accurately aligned diverse proteins are the most efficient for identifying distantly related proteins (4).

In such alignment-based searching, it is advantageous to weight the component sequences to favour more strongly those sequences that are more distantly related to other sequences in the sequence set (5–8). In HMMER, the sequences are weighted by default using the Gerstein–Sonnhammer–Chothia (GSC) method of weighting (9), additional methods are also available (10).

An alternative remote searching algorithm, PSI-BLAST (1), initiates the search process with a single sequence, and iteratively adds similar sequences found in the database to a scoring matrix of the aligned residues. PSI-BLAST's sequence weighting scheme (1) is a modified version of the one proposed by Henikoff and Henikoff 1994 (5). Essentially, PSI-BLAST takes the mean number of different residue types observed in columns of the multiple alignment in order to determine the weight assigned to each sequence within the alignment in generating the position-specific score matrix (PSSM) (1). Gap characters are treated as a 21st distinct character and any columns consisting of identical residues are ignored in calculating weights. The PSSM construction at each iteration has to make a decision from a number of different path possibilities. These decisions are controlled by the requirements for automation, speed and general simplicity (1). These decisions can cause a directionality of the PSI-BLAST algorithm that may increase noise in the results, e.g. if a non-family member with a domain in common with the protein family is identified and incorporated into the PSSM, it may mislead successive iterations to potentially overlook alternative routes to genuine distant homologues.

*To whom correspondence should be addressed. Tel: +353 1 4022790; Fax: +353 1 4022453; Email: kjohnston@rcsi.ie

Both the weighting schemes above assume that some kind of average of all the distantly related proteins is desirable. This would indeed be true if evolution proceeded in a regular fashion, since such an average would then be the best guess of what a distantly related protein might be like. However, proteins do not evolve according to a completely uniform process of independent amino acid change along all branches of an evolutionary tree (11–13). Single functionally important amino acids, short motifs, longer domains or residues clustered within the structure may be preserved or lost as units (14–16). Some critical regions of proteins may be shared between more distantly related proteins, but not shared between more closely related proteins, showing an independence from phylogeny.

It has been shown that families have different outliers, so multiple profiles are needed to model these outliers (17) and also that distant homologues are better detected when using profiles that incorporate diverse sequences (4). We developed two search methods, Rand-shuffle and Ali-shuffle that attempt to directly address these points and are less sensitive to evolutionary assumptions by performing multiple searches of sequence trios. We compare these methods with other non-structural methods [PSI-BLAST and Rand-shuffle (control)] for searching sequence datasets that may have no further added value, such as profiles, alignments or protein family structures. Specifically, we sought to develop an alternative approach with increased power, relying on automated detection of subsets of sequences to free the method from the limiting assumptions of regular evolution.

MATERIALS AND METHODS

Algorithms

We set out to determine whether sub-sampling of sequences for alignment profile searches, and combination of the sub-sample search results, was superior to a single search using the entire alignment. For these investigations, we chose the HMMER (10) implementation of HMM searching. Ali-shuffle and Rand-shuffle could be set up to use other profile methods.

The Rand-shuffle (Random shuffle) implementation of the sub-sampling process (described below) selects a randomly chosen set of sequence subset alignments. This approach is likely to get around some errors in the full profile, since some subsets will omit incorrectly aligned sequences. Second, it may consider certain sequence subsets that may be closer than the all-sequence profile to distantly related sequences. However, the Rand-shuffle approach does not consider what particular trios might be optimal, i.e. Rand-shuffle does not perform any particular weighting towards certain subsets of trios. Therefore, in addition to Rand-shuffle we examined a second implementation of the sub-sampling approach called Ali-shuffle (Alignment shuffle), which is a heuristic method that additionally weights towards the more informative sequence subsets.

Subsets of three sequences (trios) were selected from the alignment in order to keep computational complexity to a minimum. Searching all possible combinations would be currently computationally too complex, even for subsets of size three.

The 'Rand-shuffle' algorithm. In the Rand-shuffle algorithm, a random selection of sequence trios were searched. The number selected was chosen to be the same as the number of combinations defined using the Ali-shuffle algorithm (discussed below) for each protein family, making the two algorithms directly comparable in terms of computational complexity. A sequence was allowed to occur in more than one trio; however, no two searched sequence trios contained the same combination of sequences. HMMER was used to perform the searches.

The 'Ali-shuffle' algorithm. Rather than implementing a standard weighting from the literature, we chose a weighting that would potentially enrich for sequence trios that share minor frequency residues, regardless of whether those sequences were overall more or less similar. Columns that are completely conserved were ignored in the choice of sequence trios, since they are represented in the overall profile. In addition, regions that were very un-conserved were also ignored, as being less likely to be of functional importance. The algorithm examined minor amino acids in columns of intermediate conservation. Columns classified as intermediately conserved were used to define sequence trios. For a given set of minor residues at such a column, sequence trios were defined that tended to contain biochemically similar amino acids. At first glance, this may appear counterintuitive, since most sequence weighting algorithms endeavour to combine distantly related sequences, rather than similar ones. However, the chosen algorithm has two likely effects. First, it enriches for sampling combinations of minor sequences while ignoring the majority effect, which is well represented by the original whole alignment HMMER search. Second, it admits the possibility that sequences preserving a conserved region that is scattered across a few evolutionary branches may be considered within a trio that shares this region. The process is repeated until all columns of the alignment have been examined producing a list of sequence trios to be searched. The net effect of this algorithm is that any sequence that has a minor amino acid at an intermediately conserved column is guaranteed to be included within a sequence trio search. Its likely partners in such a trio are sequences that share, at a moderately conserved column, similarities in the properties of the minor amino acids at that column. The precise details are given in the Supplementary Material. This algorithm has the effect of greatly reducing the total number of potential trios searched: with the PFAM dataset used, there was a typical 100-fold reduction from the potential complete set of trios for alignments containing $\geqslant$30 sequences.

For both Rand-shuffle and Ali-Shuffle methods, each non-redundant combination selected was used to build a HMM profile with the HMMER package (10). The profile was calibrated and used in a search, by HMMER, against a given database. In addition to the selected trios, the full alignment search was also searched. The results of the separate alignment profile searches of all sequence subsets and the complete alignment's profile search are merged and ranked in order of score.

In practice, Ali-shuffle linearly expands the computational complexity of the search algorithm used (in this case HMMER) in relation to the length of the alignment, and more loosely in relation to the number of sequences in the alignment.

Other search methods. All PSI-BLAST and BLAST searches were conducted by using each sequence in the alignment separately and combining the search results. PSI-BLAST

searches were conducted to convergence, since it performs better at high rates of false positives when run until convergence. HMMER searches were performed using the whole alignment. We used the default parameter settings of the HMMER program in all assessments of both the straightforward HMMER search and the modifications to the search strategy using sub-samples of sequence sets. 'hmmcalibrate' was used for all profiles. PSI-BLAST and BLAST searches used the default parameter settings.

Benchmark dataset—688 protein families

Our objective is to develop a better algorithm for the detection of distantly related protein sequences based on alignments of primary sequences. In order to evaluate the performance of the different homologue detection methods, we used the SCOP classifications of Murzin *et al*. (18,19) to create a benchmark dataset. SCOP has been used before as a benchmark dataset in both method comparison studies (4,8,20,21) and in new method assessment studies (22,23). SCOP is a manually curated, hierarchical database, where each protein domain of known tertiary structure is classified into a family that in turn belongs to a superfamily. Each superfamily belongs to a fold that in turn belongs to a class. Proteins belonging to a family have a clear evolutionary relationship. Proteins in the same superfamily are of probable common evolutionary origin, and those in the same fold have major structural similarity. The SCOP classification's reliance on structural similarity to define superfamily relationships makes it an effective benchmark validation dataset for comparisons among algorithms that compare primary sequence identity. The protein alignments for comparing these algorithms were obtained from the PFAM database (24). A subset of protein families common to both databases made up our benchmark dataset (see below).

The 'Full' alignments of PFAM version 6.6 protein domain families (24) were downloaded from http://www.sanger.ac.uk/Software/Pfam/. The file astral-scopdom-seqres-gd-sel-gs-bib-100-1.57.fa from the SCOP protein database (18,19) version 1.57 was downloaded from http://astral.stanford.edu/scopseq-1.57/. Sequences from each PFAM family were searched against the SCOP database using BLAST (1). PFAM families that consistently hit a single family in the SCOP database and no other superfamily members outside of that family with an *E*-value less than a cut-off threshold ($<10^{-4}$) were denoted as having superfamily members that were 'difficult' to detect. The query PFAM families from the 'difficult' to detect dataset consist of 688 PFAM families that have an acceptable computational intensity (i.e. limited to 50 sequences where alignments are longer than 500 amino acids). This subset of 688 PFAM families along with their BLAST identified SCOP assignments (identical to the SCOP protein database version 1.65 PFAM-SCOP assignments) constituted our benchmarking dataset.

Comparison and assessment

SCOP is a hierarchical database. Both the superfamily and family levels of the hierarchy contain a degree of sequence similarity, and therefore both levels were applicable to this study. The true positive hits were defined as hits to the superfamily. Results could be categorized into three categories: true positives (tp), false-positives (fp) and false-negatives (fn). A true positive was defined as a search from PFAM that hit the correct superfamily. All other hits were defined as false positives. A false negative is a superfamily member that has not been hit. At a given search similarity threshold, we calculated the sensitivity [tp/(tp + fn)], and the proportion of false positives in all hits or 'false-positive rate' [fp/(fp + tp)] (25). Specificity was also investigated, but we present only the false-positive rate, since it has a simpler interpretation in assessing the biological significance of a database match.

A standard Receiver Operating Characteristic (ROC) curve (25) that consists of a diagram with the sensitivity plotted on the *x*-axis and the false-positive rate α, on the *y*-axis, was plotted. ROC curves and variations on them have been used in other studies in the past (20,26–30). In our analysis, the ROC curve was constructed by finding the sensitivity and the false-positive rate when varying *E*-value (1,10) threshold cut-offs in the results. Error bars for the ROC curves were calculated by the bootstrap method (31): the super-families in the dataset were sampled randomly, with replacement, 1000 times and 2.5% tails of the distribution were used to produce the 95% confidence intervals shown in the plots. The bootstraps were calculated to determine whether the results were sensitive to the particular families used in the assessment, and to determine significance of differences among methods. This was not carried out in previous method comparison studies.

Application—cytokine family search

To test the application of Ali-shuffle, the PFAM alignment of interleukin 8-like cytokines was searched against the complete human genome translated into all six possible reading frames. The resulting search results were combined, and a list of non-redundant hits made. Where there were two or more hits to the same location on the genome we took the hit with the lowest *E*-value (best hit). Hits to chromosome 4 were segregated into the following categories: known cytokines as annotated by Ensembl (32), other sequences, and finally, potential novel matches, defined as other sequences with a score >20.

RESULTS

The ROC curve permits the comparison of alternative search methods at a range of sensitivity and false-positive rate points (20,25–30). It plots sensitivity (*x*-axis) against the false-positive rate (*y*-axis), using a whole range of arbitrary cut-off points of sequence similarity. Thus, as the threshold (*E*-value) for considering a hit of interest is raised, there is both an increasing sensitivity and an increasing number of false positives. Figure 1 compares the search methods. PSI-BLAST is only comparable at higher false-positive rates, since the method always returns some false positives. At very low rates of false-positives (<1%) HMMER, Ali-shuffle and Rand-Shuffle are comparable with a sensitivity of ~7%.

The Rand-shuffle algorithm enhances HMMER's performance when a greater degree of false positives is considered. Thus, a simple algorithm combining sequences at random gives an immediate benefit over the existing method, at moderate false-positive rates of 5%. Use of the Ali-shuffle algorithm, to select which trio alignment subsets to use, results in a further improvement in performance, at false-positive rates >5%. PSI-BLAST is only directly comparable at the

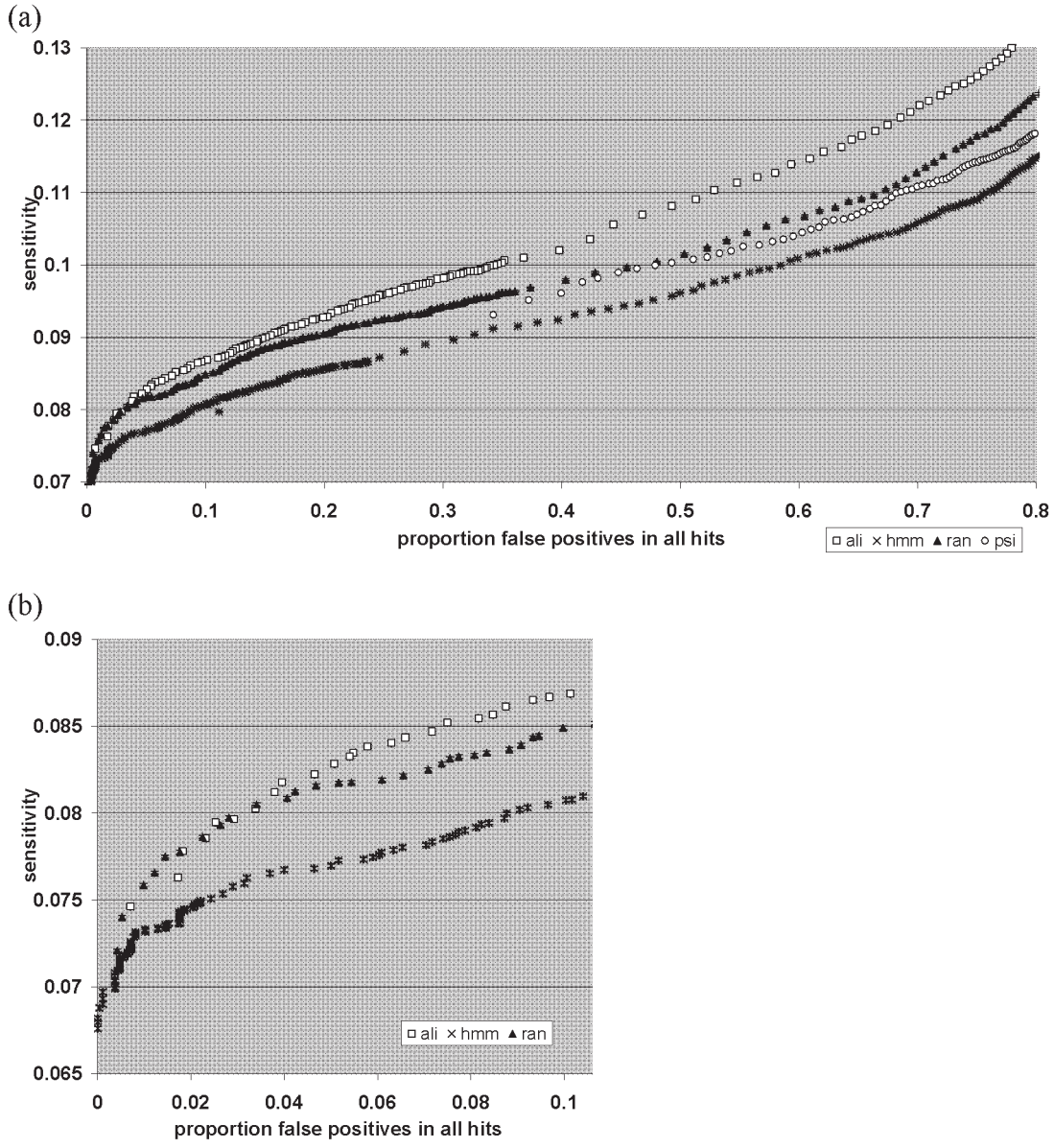

Figure 1. (**a**) Sensitivity (*y*-axis) against the false-positive rate (*x*-axis) for the search results of 688 protein families. Ali-shuffle (ali), HMMER (hmm), Rand-shuffle (ran) and PSI-BLAST (psi) are compared. A total of 95% confidence intervals are also included in the plot. (**b**) Detail.

higher false-positive rates. As noted before, in Lindahal and Elofsson's method comparison study, PSI-BLAST outperforms a HMM implementation [in their comparison ssHMM (26)] at higher specificities (20). Figure 1a indicates that where PSI-BLAST results are comparable (false-positive rate of >35%) it always outperforms HMMER. At these high false-positive rates, Rand-Shuffle marginally outperforms PSI-BLAST, while Ali-Shuffle demonstrates a substantial improvement.

In a typical search for novel sequences, investigators will start using BLAST, then proceed to a more complex method, such as PSI-BLAST or HMMER, and only then consider a more complex search, such as Ali-shuffle. The results in Figure 1 do not clearly distinguish how much better the Rand-shuffle and Ali-shuffle algorithms are at detecting sequences not detectable by BLAST. Therefore, we investigated the performance of the search methods among the true positives that are not detectable by BLAST. BLAST-detected hits (detected at *E*-value $\leqslant 10^{-4}$) were removed from consideration, and results presented for the remaining sequences (Table 1). At low false-positive rates (5%), both Rand-Shuffle and Ali-Shuffle demonstrate a marked improvement in HMMER's sensitivity (37.5 and 62.5%, respectively). Allowing increasing numbers of false positives, the pattern of performance fluctuates somewhat, with the advantage of Rand-shuffle tending overall to decline, while that of Ali-shuffle stays more stable (Table 1).

Investigation of the results from the different methods indicated that there were a proportion of families where Ali-shuffle considerably improved HMMER's sensitivity and a smaller number of families where HMMER outperformed Ali-shuffle. The properties of these families were investigated to determine

Table 1. Effect of excluding BLAST-detected hits on the relative percentage increase in sensitivity of Ali-shuffle and Rand-shuffle compared with HMMER alone

% False positives	Ali-shuffle Including blast	Excluding blast[a]	Rand-shuffle Including blast	Excluding blast[a]
5	7.79	62.50	5.19	37.50
10	7.41	50.00	2.47	33.33
15	8.43	57.89	6.02	42.11
20	8.14	60.00	5.23	50.00
25	10.34	63.64	5.75	45.45
30	8.89	83.33	4.44	50.00
35	9.89	64.29	5.49	35.71
40	10.87	51.61	6.52	25.81
45	12.77	57.58	6.38	21.21
50	13.54	61.11	5.21	16.67

[a]Excluding from consideration BLAST hits detectable with *E*-value of 10^{-4}.

whether alignments of a certain type may benefit particularly from a specific search algorithm. A comparison of phylogenetic distances and pairwise distances between all possible sequences in a given family was performed, to test whether those families with a greater sequence diversity and phylogenetic spread perform better with Ali-shuffle. The alignment length, the number of sequences in the alignment, the average tree distance and the average pairwise distance were also looked at.

The results of the different methods were investigated at the appropriate search cut-off that generated a false-positive proportion of 50%. Alignment length was weakly but significantly positively correlated with A − H (where A − H = Ali-shuffle sensitivity minus HMMER sensitivity, for families showing different sensitivities for the two methods; $r = 0.280$, $P = 0.014$). This implies that Ali-shuffle tends to work better with longer sequences compared with HMMER, while HMMER tends to work better with shorter sequence alignments.

The correlation between tree distances and pairwise distances for these families was not found to be significant ($r = -0.125$, $P = 0.281$), suggesting that greater sequence diversity and phylogenetic spread is not a predictor of Ali-shuffle's success. No significant correlation was found for the other attributes tested for the different family sets (number of sequences: $r = -0.036$, $P = 0.76$; average tree distance: $r = 0.013$, $P = 0.914$; average pairwise distance: $r = 0.0418$, $P = 0.72$).

An interleukin 8-like cytokine alignment was searched against the translated human genomic DNA with Ali-shuffle. As expected, there was a concentration of known hits within a region of chromosome 4. The seven apparent novel cytokine sequences identified by Ali-shuffle on chromosome 4 clustered strongly at the two sub-regions containing the known cytokine genes.

DISCUSSION

A validation of the Ali-shuffle and Rand-shuffle methods using the SCOP protein classification database demonstrated that consideration of such similarities does indeed increase the power to detect distantly related homologues. Rand-shuffle simply selects trios at random, and at 5% false-positive rates it gives a 37.5% increase in HMMER's sensitivity, in searches for related sequences not detectable by BLAST. Its success may result from by-passing alignment errors in the entire alignment, and secondly from allowing chance combinations of sequence to be favoured that would otherwise be obscured. Ali-shuffle further improves HMMER's performance, by enriching for trios that share properties at intermediately conserved columns not shared by the alignment as a whole. Since these sequence trios need not necessarily be closely related over all residue positions, this method provides a more general consideration of subsequence similarities, independent of the phylogeny of the aligned proteins. This contrasts with previously developed methods (4,6,8,9,33) that weight towards more distantly related proteins within the alignment-based on the overall sequence similarity. Thus, these search methods complement existing search approaches.

Rand-shuffle and Ali-shuffle both improve HMMER's performance at high false-positive rates ($\geqslant$5%, see Table 1), although Ali-shuffle out performs Rand-shuffle. We suggest that the Rand-shuffle algorithm improves HMMER's performance because individual trios of sequences are less contaminated by alignment errors seen elsewhere in the full alignment (4) in addition to selected subsets detecting distantly related proteins because their profile is a closer match than the full alignment profile. Alignment error may arise through sequencing (e.g. frameshifting) errors, presence of splice variants or simply the difficulty in reconstructing the true evolutionary relationships of sequences. We have not formally assessed the impact of alignment error on these methods, since it is difficult to quantify the extent and the nature of errors in real datasets, and simulation models may not reflect the true distribution of such errors. The further improvement in performance by the Ali-shuffle algorithm suggests that the pattern of evolution is an important factor. We postulate that Ali-shuffle's independence from phylogeny enables the detection of both non-phylogenetically linked and phylogenetically linked motifs shared by distantly related homologues that may be missed by a straightforward HMMER search. The extent that Ali-shuffle improves performance suggests that departures from regularity in evolution may be quite marked, particularly when comparing very distantly related proteins. This is not entirely surprising, since most of the original studies that identified a typically regular pattern of protein evolution (34) relied on studies of relatively conserved proteins.

Ali-shuffle improves HMMER's performance at higher permitted false-positive rates in our benchmark tests. Particularly striking is the sensitivity increase observed when the BLAST identified hits are removed from the results (see Table 1). This reflects the typical use of database searching by a biologist who will carry out a BLAST search and then subsequently investigate the slower, more powerful, search techniques.

While some researchers would not wish to consider high levels of false positives, other researchers increasingly use correlation with other evidence sources, such as expression patterns and gene structure to help refine candidate homologues. Clustering of the seven cytokine-like sequences identified by Ali-shuffle at the two sub-regions containing the known cytokine genes indicate that they may represent either functional genes or pseudogenes arising through local duplication of DNA. This illustrates the utility of Ali-shuffle in detecting novel related sequences. It indicates how information from

searches with lax parameters allowing a high false-positive rate may be effectively combined with additional sources of information, justifying the continued development of search methods with greatest power at higher false-positive rates.

Sequences of proteins where Ali-shuffle and Rand-shuffle improved HMMER were generally longer. These results indicate that alignment length is of some importance in the success of Ali-shuffle, as is the selection of the subsets searched. Increase of alignment length means that there are a larger number of columns, and within the current implementation of the algorithm this usually results in a larger number of trio searches being carried out; we postulate that it is this that leads to the added success of Ali-shuffle. While there is a computational burden associated with using Ali-shuffle, increasing availability of high-performance computing will help overcome this, at least for searches focused within single genomes.

While previous work has established the principle that considering subsets of alignments permits the identification of distantly related proteins that might otherwise not be detected (17), our algorithm takes a more directed approach in terms of the combination of these subsets without using structural information while limiting the number of searches performed. Clearly, there are many other possible variations of this principle that may also be possible. Subsets could be established representing sub-trees within an alignment, which would then have greater power to detect rapidly evolving proteins belonging to that particular clade of sequences. We investigated whether altering the number of sequences included in the sub-sampling (from 3 to 7 sequences) altered the performance, and found it to be quite similar (data not shown). However, the more generic approach of Ali-shuffle incorporates a wider variety of possible evolutionary scenarios and does not depend on the assumption of regular evolution at all regions in the protein.

The algorithm presented here provides a reasonable compromise between computational intensity, increased sensitivity, controlling the signal-to-noise ratio and practical implementation. Further improvements to the computational intensity of the algorithm may well be possible, cutting down on the number of sub-alignments searched according to a variety of alternative criteria. This method differs from profile–profile algorithms (35), which require that the target database of sequences is well-aligned, reducing the search space but also excluding the discovery of certain potential true positives. Two-track HMMs (36,37) require structural data, which again limits possible comparisons to sequences of known structure. A direct comparison of the utility of these three methods is not practical, since each is optimal for the searching of different databases. In the presence of structural data, two-track HMMs may well be a superior approach. We explored whether the sub-sampling approach we applied was also applicable to the improvement of profile–profile methods, using the PRC implementation of profile–profile searches (M. Madera, unpublished data) (http://supfam.mrc-lmb.cam.ac.uk/PRC). However, we did not find a significant improvement in PRC's performance by applying the Ali-shuffle sub-sampling to the construction of search profiles (data not shown). Partly due to the computational intensity of the Ali-shuffle algorithm, we do not suggest that it be used for routine genome annotation, as has been proposed for PSI-BLAST (22) and HMMs (17), but rather by researchers with a specific protein family of interest to search for distantly related proteins in datasets that do not have alignments or structural data available. The combination of the output of Ali-shuffle with other information sources, such as chromosomal location, provides a powerful way to quickly provide corroborative information regarding true homology.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Matt Sullivan, Rich Edwards and Stephen Park for discussions. This work was supported by Enterprise Ireland and by the Programme for Research in Third Level Institutions administered by the Higher Education Authority, Ireland. Funding to pay the Open Access publication charges for this article was provided by Science Foundation Ireland (SFI).

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schaffer,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.
2. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
3. 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.
4. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*., **30**, 4321–4328.
5. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol*., **243**, 574–578.
6. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci*., **10**, 19–29.
7. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (2001) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic acids, 1st edn*. Cambridge University Press, Cambridge, UK.
8. Karchin,R. and Hughey,R. (1998) Weighting hidden Markov models for maximum discrimination. *Bioinformatics*, **14**, 772–782.
9. Gerstein,M., Sonnhammer,E.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol*., **236**, 1067–1078.
10. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
11. Uzzell,T. and Corbin,K.W. (1971) Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096.
12. Langley,C.H. and Fitch,W.M. (1974) An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol*., **3**, 161–177.
13. Gillespie,J.H. (1986) Natural selection and the molecular clock. *Mol. Biol. Evol*., **3**, 138–155.
14. Pils,B. and Schultz,J. (2004) Evolution of the multifunctional protein tyrosine phosphatase family. *Mol. Biol. Evol*., **21**, 625–631.
15. Perutz,M.F., Bauer,C., Gros,G., Leclercq,F., Vandecasserie,C., Schnek,A.G., Braunitzer,G., Friday,A.E. and Joysey,K.A. (1981) Allosteric regulation of crocodilian haemoglobin. *Nature*, **291**, 682–684.
16. Braun,E.L. (2003) Innovation from reduction: gene loss, domain loss and sequence divergence in genome evolution. *Appl. Bioinformatics*, **2**, 13–34.

17. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
18. 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.
19. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
20. Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, **295**, 613–625.
21. 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.
22. Müller,A., MacCallum,R.M. and Sternberg,M.J.E. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.
23. Rehmsmeier,M. (2002) Phase4: automatic evaluation of database search methods. *Brief Bioinform.*, **3**, 342–352.
24. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
25. Armitage,P., Berry,G. and Matthews,J.N.S. (2002) *Statistical Methods in Medical Research, 4th edn*. Blackwell Publishing, Oxford, UK.
26. Hargbo,J. and Elofsson,A. (1999) A study of hidden Markov models that use predicted secondary structures for fold recognition. *Proteins*, **36**, 68–87.
27. Gribskov,M. and Robinson,N.L. (1996) The use of receiver operating characteristic (ROC) analysis to evaluate sequencing matching. *Comput. Chem.*, **20**, 25–34.
28. Spang,R., Rehmsmeier,M. and Stoye,J. (2002) A novel approach to remote homology detection: jumping alignments. *J. Comput. Biol.*, **9**, 747–760.
29. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
30. Rice,D.W. and Eisenberg,D. (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.*, **267**, 1026–1038.
31. Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.
32. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
33. Rehmsmeier,M. and Vingron,M. (2001) Phylogenetic information improves homology detection. *Proteins*, **45**, 360–371.
34. Ayala,F.J. (1986) On the virtues and pitfalls of the molecular evolutionary clock. *J. Hered.*, **77**, 226–235.
35. Jaroszewski,L., Rychlewski,L. and Godzik,A. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.*, **9**, 1487–1496.
36. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, **53** (Suppl. 6), 491–496.
37. Karchin,R., Cline,M., Mandel-Gutfreund,Y. and Karplus,K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*, **51**, 504–514.