ProtEST: protein multiple sequence alignments from expressed sequence tags

James A. Cuff[1,2], Ewan Birney[3], Michele E. Clamp[2,†] and Geoffrey J. Barton[2,*]

[1]Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, [2]European Molecular Biology Laboratory Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and [3]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on June 4, 1999; revised on August 16, 1999; accepted on August 25, 1999

Abstract

Motivation: *An automatic sequence searching method (ProtEST) is described which constructs multiple protein sequence alignments from protein sequences and translated expressed sequence tags (ESTs). ProtEST is more effective than a simple TBLASTN search of the query against the EST database, as the sequences are automatically clustered, assembled, made non-redundant, checked for sequence errors, translated into protein and then aligned and displayed.*

Results: *A ProtEST search found a non-redundant, translated, error- and length-corrected EST sequence for >58% of sequences when single sequences from 1407 Pfam-A seed alignments were used as the probe. The average family size of the resulting alignments of translated EST sequences contained >10 sequences. In a cross-validated test of protein secondary structure prediction, alignments from the new procedure led to an improvement of 3.4% average Q_3 prediction accuracy over single sequences.*

Availability: *The ProtEST method is available as an Internet World Wide Web service at http://barton.ebi.ac.uk/servers/protest.html The Wise2 package for protein and genomic comparisons and the ProtESTWise script can be found at: http://www.sanger.ac.uk/Software/Wise2*

Contact: *geoff@ebi.ac.uk*

Introduction

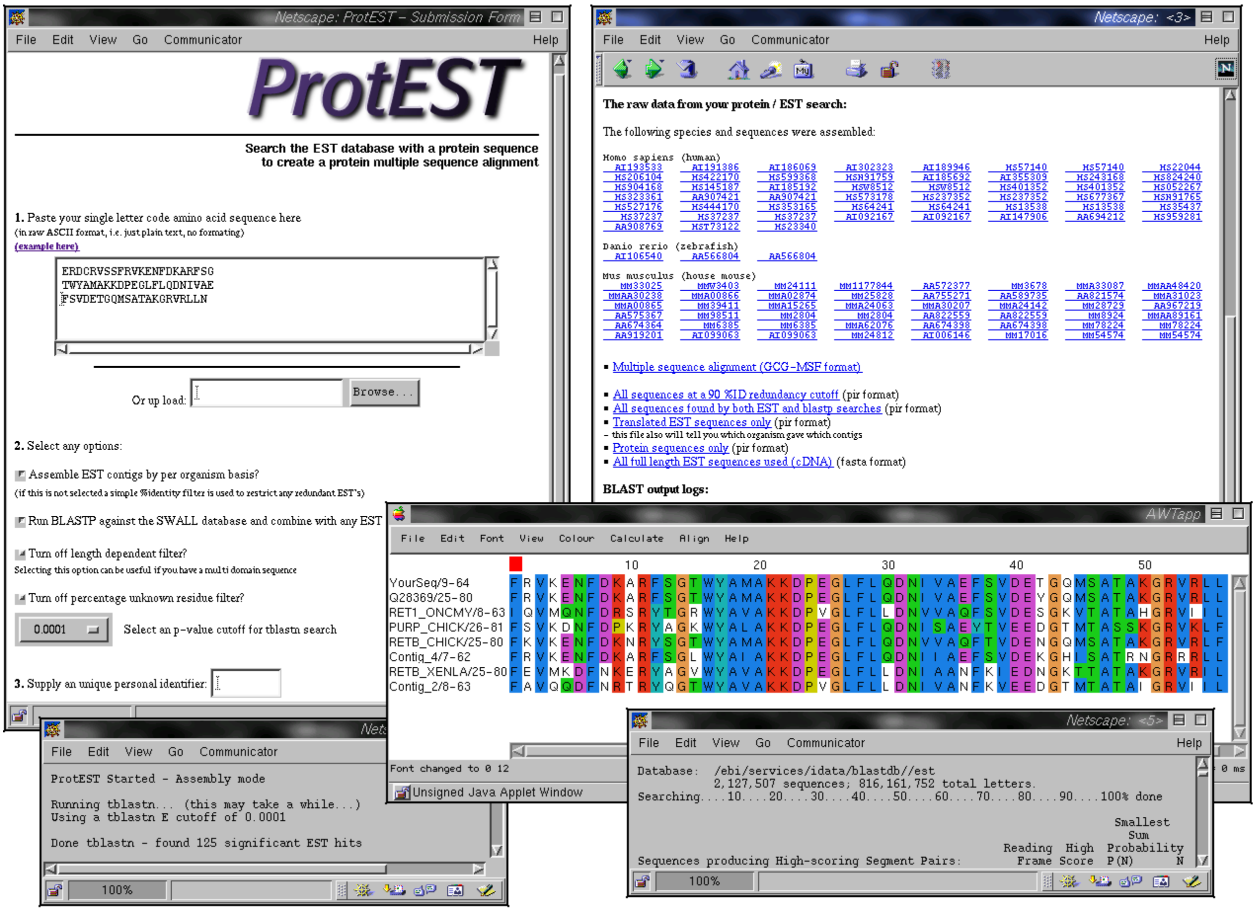
The prediction of functional residues (Casari *et al.*, 1995; Livingstone and Barton, 1996), secondary structure (Barton, 1995; Cuff and Barton, 1999) and the detection of weak sequence similarity by profile methods (Barton, 1990; Gribskov *et al.*, 1990; Eddy, 1996) rely on the analysis of multiple protein sequence alignments for optimal results. In general, greater reliability in function and structure prediction is obtained by increasing the number of sequences in the multiple sequence alignment (Rost and Sander, 1993). However, while the protein sequence databases such as SWISS-PROT (Bairoch and Apweiler, 1998) will often contain homologues to the sequence of interest, >65% of all sequences deposited to the EMBL/GenBank/DDBJ Nucleotide Sequence Database (version 58) (Stoesser *et al.*, 1999) are expressed sequence tags (ESTs). ESTs present a valuable resource to aid protein sequence analysis, but a major drawback is that they are determined by only single or double gel reads and so are error prone. Errors in the DNA sequence, particularly frame-shift errors, make it difficult to align ESTs reliably with full-length protein sequences. In addition, ESTs tend to code for fragments of full-length proteins and so may provide inconsistent information along the sequence. The nature of the sequencing process also makes EST sequence databases very redundant.

Recent developments in dynamic programming techniques offset the problem of errors in ESTs by allowing a protein sequence to be compared directly with DNA, while also considering frame-shifts and in-frame stop codons (Birney and Durbin, 1997; Pearson *et al.*, 1997; Zhang *et al.*, 1997). ESTWISE (Birney, 1998) takes this a stage further by comparing a protein sequence with a DNA sequence by a probabilistic model of both protein evolution *and* potential sequencing error. A maximum likelihood path through the model then provides an alignment that can account both for protein evolution and sequencing error when matching the two sequences.

In this paper, we describe a procedure for building protein multiple sequence alignments that exploits the additional information available from EST sequences.

*To whom correspondence should be addressed.

† Present address: The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK



Netscape: ProtEST – Submission Form
File Edit View Go Communicator Help
ProtEST

Search the EST database with a protein sequence
to create a protein multiple sequence alignment

1. Paste your single letter code amino acid sequence here
(in raw ASCII format, i.e. just plain text, no formating)
(example here)

```
ERDCRVSSFRVKENFDKARFSG
TWYAMAKKDPEGLFLQDNIVAE
FSVDETGQMSATAKGRVRLLN
```

Or up load: Browse...

2. Select any options:

☑ Assemble EST contigs by per organism basis?
(if this is not selected a simple %identity filter is used to restrict any redundant EST's)

☑ Run BLASTP against the SWALL database and combine with any EST

□ Turn off length dependent filter?
Selecting this option can be useful if you have a multi domain sequence

□ Turn off percentage unknown residue filter?

0.0001 Select an p-value cutoff for tblastn search

3. Supply an unique personal identifier:

The raw data from your protein / EST search:

The following species and sequences were assembled:

```
Homo sapiens (human)
  AI193533   AI191386   AI186069   AI302323   AI189946    HS57140    HS57140    HS22044
  HS206104   HS422170   HS599368   HSN91759   AI185692   AI355309   HS243168   HS824240
  HS904168   HS145187   AI185192    HSW8512    HSW8512   HS401352   HS401352   HS052267
  HS323361   AA907421   AA907421   HS573178   HS237352   HS237352   HS677367   HSN91765
  HS527176   HS444170   HS353165    HS64241    HS64241    HS13538    HS13538    HS35437
   HS37237    HS37237    HS37237   AI092167   AI092167   AI147906   AA694212   HS959281
  AA908769   HST73122    HS23340

Danio rerio (zebrafish)
  AI106540   AA566804   AA566804

Mus musculus (house mouse)
   MM33025    MMW3403    MM24111  MM1177844   AA572377     MM3678   MMA33087  MMAA48420
 MMAA30238   MMA00866   MMA02874    MM25828   AA755271   AA589735   AA821574   MMA31023
  MMA00865    MM39411   MMA15265   MMA24063   MMA30207   MMA24142    MM28729   AA967219
  AA575367    MM98511     MM2804     MM2804   AA822559   AA822559     MM8924  MMAA89161
  AA674364     MM6385     MM6385   MMA62076   AA674398   AA674398    MM78224    MM78224
  AA919201   AI099063   AI099063    MM24812   AI006146    MM17016    MM54574    MM54574
```

- Multiple sequence alignment (GCG-MSF format)

- All sequences at a 90 %ID redundancy cutoff (pir format)
- All sequences found by both EST and blastp searches (pir format)
- Translated EST sequences only (pir format)
- this file also will tell you which organism gave which contigs
- Protein sequences only (pir format)
- All full length EST sequences used (cDNA) (fasta format)

BLAST output logs:

```
                          10        20        30        40        50
YourSeq/9-64      FRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLL
Q28369/25-80      FRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDEYGQMSATAKGRVRLL
RET1_ONCMY/8-63   IQVMQNFDRSRYTGRWYAVAKKDPVGLFLLDNVVAQFSVDESGKVTATAHGRVIIL
PURP_CHICK/26-81  FSVKDNFDPKRYAGKWYALAKKDPEGLFLQDNISAEYTVEEDGTMTASSKGRVKLF
RETB_CHICK/25-80  FKVKENFDKNRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATAKGRVRLF
Contig_4/7-62     FRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGHISATRNGRRRLL
RETB_XENLA/25-80  FEVMKDFNKERYAGVWYAVAKKDPEGLFLLDNIAANFKIEDNGKTTATAKGRVRIL
Contig_2/8-63     FAVQQDFNRTRYQGTWYAVAKKDPVGLFLLDNIVANFKVEEDGTMTATAIGRVIIL
```

```
Font changed to 0 12
```

```
ProtEST Started - Assembly mode

Running tblastn... (this may take a while...)
Using a tblastn E cutoff of 0.0001

Done tblastn - found 125 significant EST hits
```

```
Database:  /ebi/services/idata/blastdb//est
           2,127,507 sequences; 816,161,752 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                        Smallest
                                                                          Sum
                                                          Reading  High  Probability
Sequences producing High-scoring Segment Pairs:            Frame  Score  P(N)      N
```

Fig. 1. An outline of the ProtEST method as implemented in the Internet World Wide Web server. The submission form allows a single sequence to be inserted, and selection of any options. Options include a choice of assembling the sequences, or filtering on the basis of percentage identity. Both length-dependent filters and unknown residue filters can also be bypassed, along with the p-value cutoff to use for the TBLASTN step. (The slower, more accurate TFASTX algorithm may also be optionally selected at this stage.) The protein-searching stage may be bypassed if required. Automatic searching of UniGene EST clusters may also be bypassed if required. Output takes the form of both HTML pages, and a Java viewer to render the resulting multiple sequence alignment. If assembly is selected, the original EST fragments that make up the contigs can be retrieved, as can the raw data from the initial BLASTP and TBLASTN/TFASTX searches. The server generates a complete report of each step, which can be examined.

EST searching and clustering by ProtEST is carried out interactively for each query sequence, and so the effectiveness of the ProtEST method will scale with the increasing EST database. The software has been constructed to be as flexible as possible while still maintaining efficiency. The procedure combines the new techniques for DNA to protein sequence comparison (Birney, 1998) together with conventional techniques for DNA sequence database searching (Altschul *et al.*, 1990) and assembly (Green, 1996; Gordon *et al.*, 1998). A Web server has also been constructed (Figure 1).

We also show that on a benchmark for protein secondary structure prediction (Cuff and Barton, 1999), the inclusion of EST sequences consistently improves prediction accuracy.

Method

The stages of the ProtEST method are shown in Figure 2. Firstly, the TBLASTN (version 2) sequence comparison algorithm (Altschul *et al.*, 1990) is used to search the query sequence against the EMBL-EST (Stoesser *et al.*, 1999) database (step 1, Figure 2).

An alternative to using TBLASTN for the initial search would be to use TFASTX (Pearson *et al.*, 1997). TFASTX produces an optimal Smith–Waterman alignment of the query and translated-library sequence while also calculating similarity scores that allow for frame-shifts. However when comparing TFASTX against TBLASTN on 4 SGI R10K processors, run time was 04:35 min for TBLASTN as opposed to 17:10 min for TFASTX.

The ProtEST Web server has an option to run the

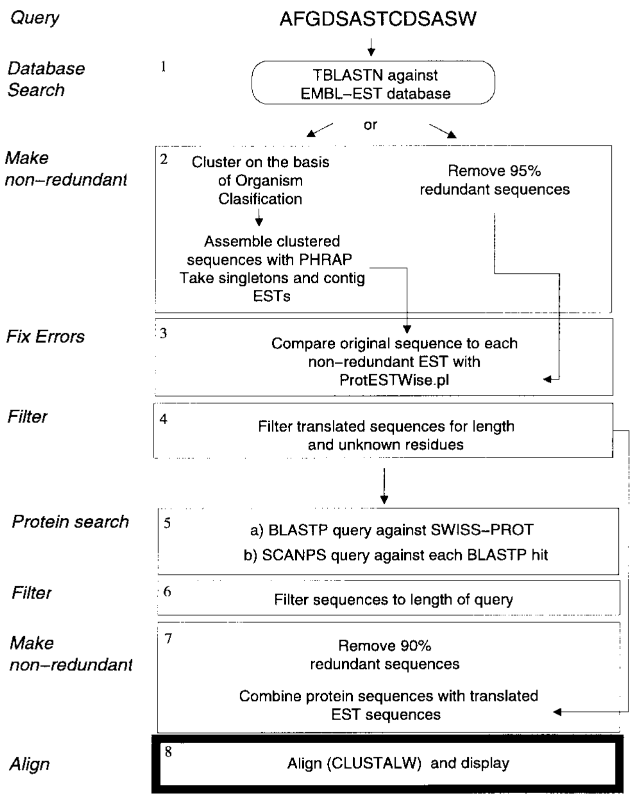

Query
AFGDSASTCDSASW
Database Search
1
TBLASTN against EMBL–EST database
or
Make non–redundant
2
Cluster on the basis of Organism Clasification
Assemble clustered sequences with PHRAP Take singletons and contig ESTs
Remove 95% redundant sequences
Fix Errors
3
Compare original sequence to each non–redundant EST with ProtESTWise.pl
Filter
4
Filter translated sequences for length and unknown residues
Protein search
5
a) BLASTP query against SWISS–PROT
b) SCANPS query against each BLASTP hit
Filter
6
Filter sequences to length of query
Make non–redundant
7
Remove 90% redundant sequences
Combine protein sequences with translated EST sequences
Align
8
Align (CLUSTALW) and display

Fig. 2. The ProtEST method. See 'Method' for a full description of each step.

slower TFASTX method, but in this work we applied the TBLASTN method.

Sequence identifiers from matches down to a TBLASTN p-value of 10^{-4} are taken from this search and used to retrieve the corresponding full-length sequence from the EMBL-EST database. The sequences found by TBLASTN are first clustered on the basis of their organism classification, then each of the sequence clusters are assembled by the PHRAP (Green, 1996; Gordon *et al.*, 1998) sequence assembly algorithm (step 2, Figure 2). Assembling sequence based on organism classification is likely not to assemble all the available sequence. Any sequences within the cluster that cannot be assembled will be ejected as singletons. If the singletons turn out to be too short, they are removed by the length filters applied later in the process.

The contigs and any singletons from the assembly process replace the sequences found by the initial TBLASTN search as organism-specific and non-redundant EST sequences. Application of the assembly program means that if the query protein sequence matches two separate EST fragments that overlap, it is possible to assemble the two overlapping EST fragments into a single longer contig.

A simple alternative to assembling the sequences is to remove EST redundancy by excluding sequences that are $>95\%$ identical to the query sequence. However, with this simple approach there is no opportunity to combine the short EST fragments into a single, longer contiguous sequence.

The non-redundant sequences from PHRAP are then compared with the query, using ProtESTWise (step 3, Figure 2). ProtESTWise, developed with the PERL API (Birney, 1998) of the Wise2 sequence comparison package, provides an interface to the EST comparison algorithm ESTWISE (Birney, 1998). ESTWISE gives a maximum likelihood position for frame-shift errors due to sequencing error. A 'corrected' protein sequence is then generated with 'X' marking potential sequencing errors and in-frame stop codons. ESTWISE was used in the alignment phase as (in our tests) it places the insertion or deletion due to sequencing error more accurately than TFASTX. However, ESTWISE is not practical for the database searching phase as it does a complete dynamic programming pass of each EST, making it computationally expensive. The combination of either TFASTX or TBLASTN, which provide good sensitivity for finding the ESTs, and ESTWISE, which provides accurate alignment considering frame-shifts, is a good compromise.

In order to prune overhangs, the translated sequences are then filtered by applying a length cutoff of 3/2 (step 4, Figure 2). For example, if the query sequence is N residues long, the sequence length would have to range between $2N/3$ and $3N/2$ residues to be included. If sequences exceed the length criterion, they are truncated by removing residues from each end until the length of the sequence satisfies the cutoff value. Sequences falling short of the lower length limit are discarded. The value of 3/2 for the length cutoff was reached by visual inspection of a number of multiple sequence alignments, produced with different cutoff values. This filter removes short sequences but does allow sequences that are longer than the query, and are related, to be included after truncation. A filter was also applied to remove those sequences that had over 3% of residue marked as 'X'. This situation would occur, for example, if there were a large number of potential sequencing errors and in-frame stop codons located by ESTWISE in the assembled EST sequence.

Protein searching

A BLASTP (version 2) (Altschul *et al.*, 1990) search of the SWISS-PROT non-redundant SPTR database (Bairoch and Apweiler, 1998) is also performed for the query sequence (step 5, Figure 2). The BLAST (Altschul *et al.*, 1990) output is then screened by SCANPS (Barton, 1993), an implementation of the Smith–Waterman dynamic programming algorithm (Smith and Waterman, 1981), with

Table 1. Data for sequences found from the EST and protein-searching portions of ProtEST

	PDB ProtEST	PDB UniGene	PDB Prot.	Pfam ProtEST	Pfam Prot.
Size of dataset (sequences)	513	513	513	1407	1407
Coverage (%)	257 (50.0%)	225 (43.8%)	513 (100%)	827 (58.7%)	1407 (100%)
Total number of sequences	2442	2951	14 880	8479	88 988
Average number of sequences per family (median)	9.5 (4.0)	13.1 (4.0)	29.0 (16.0)	10.6 (6.2)	63.5 (35.0)
Total number of residues	286 099	401 844	2 294 040	2 233 212	27 026 091
Number of residues marked as 'X'	970	231	316	6828	5407
Average sequence length (residues)	117	136	154	127	303

'PDB' refers to the 513 protein test set for secondary structure prediction (Cuff and Barton, 1999). 'Pfam' refers to the 1407 set of single sequences from the Pfam-A seed alignments (Bateman *et al.*, 1999). 'ProtEST' refers to sequences that were derived from the assembled, non-redundant translated and error-checked ESTs obtained from the ProtEST method (Figure 2, steps 1–4). 'Prot.' refers to the protein searching portion of ProtEST (Figure 2, steps 5–7). 'UniGene' refers to the sequences found by searching the UniGene database with TBLASTN. The number of 'X' residues corresponds to positions where there are either unknown residues, or there are positions marked as unknown by the dynamic programming algorithm used to translate the DNA to protein.

length-dependent statistics. Sequences are rejected if their SCANPS probability score is $> 10^{-4}$. Sequences are also rejected if they do not fit the length cutoff of 3/2 (step 6, Figure 2). The sequences from the EST search and the protein search are then combined. All pairs of sequences are compared by the AMPS package (Barton, 1990). The sequences are clustered on the basis of percentage identity by following complete linkage clustering (step 7, Figure 2). Finally, a 90% sequence identity cutoff is used to select clusters to give a non-redundant set of sequences which are then aligned by CLUSTALW (Thompson *et al.*, 1994) with default parameters (step 8, Figure 2).

Evaluation of ProtEST

The quality of cross-validated protein secondary structure predictions derived from the ProtEST alignments was examined. This test exploited the non-redundant test set of 513 proteins recently developed by Cuff and Barton (1999). This test provides a direct measure of the usefulness of ProtEST alignments in analysis of protein secondary structure prediction. However, since the 513 proteins are all from proteins of known three-dimensional structure, they may not give a fair representation of how many additional sequences can be found by ProtEST for a typical protein query.

Ideally one would like to take each sequence in SWISS-PROT, generate ProtEST alignments, then assess how many additional sequences are found over a simple protein database search. However, this would require over 270 000 searches, and is currently computationally unrealistic. The Pfam 3.4 alignment database (Bateman *et al.*, 1999) gives >50% coverage of SWISS-PROT and so was taken to be representative of the database as a whole. The first sequence from each of the 1407 Pfam-A seed alignments was taken to measure the number non-redundant, error-checked EST sequences that the ProtEST method returned. The protein matches from the BLASTP section of ProtEST were also compared to the number of EST sequences found.

Results and discussion

The results are shown in Table 1. Of the 513 proteins of known three-dimensional structure that were used to test the method, 257 sequences (50%) matched at least one translated, non-redundant, error-checked EST contig sequence. Of the 257 sequences, 2442 contigs matched in all, giving an average of 9.5 extra translated EST sequences per sequence family. In contrast, the protein sequence database searching stage of ProtEST returned 14 880 sequences in total, which corresponds to an average of 29 sequences per family.

When the 1407 primary sequences from the seed alignments of Pfam version 3.4 were used to test the ProtEST method, there were 827 sequences that matched at least one non-redundant, error-checked, translated EST contig sequence. The success rate using the Pfam sequences was 58.7%, which is 8.7% higher than obtained using sequences of known three-dimensional structure. This

```
sequence      : SKGVITITDAEFESEVLKAEQPVLVYFWASWCGPCQLMSPLINLAANTYSDRLKVV
phd ProtEST   : ---EEEE----HHHHHE----EEEEEEE-----------HHHHHHHHHH---EEEE
phd single    : ---EEEEE-HHHHHHHHHH---EEEEEEE---------HHHHHHHH------EEEE
DSSP          : ___EEE_____HHHH_______EEEEEE______HHHHHHHHHHHHHH_____EEE

sequence      : KLEIDPNPTTVKKYKVEGVPALRLVKGEQILDSTEGVISKDKLLSFLDTHLN
phd ProtEST   : EEE--------EEEEEEE-EEEEEEE-------------HHHHHHHHHHH-- (71.3%)
phd single    : EEEE------EEEEEEEEEEEEEE---EEEEE----EEE--HHHHHHHH--- (60.2%)
DSSP          : EEE____HHHHHH_______EEEEEE__EEEEEEE____HHHHHHHHHHHH_
```

Fig. 3. Comparison of PHD (Rost and Sander, 1993), predictions for single sequences and EST sequences from ProtEST, for 1thx (Thioredoxin Electron Transport Protein) against the DSSP (Kabsch and Sander, 1983) secondary structure definition.

Table 2. Comparison of Q_3 prediction accuracy for cross-validated PHD predictions for 257 sequences

Data used for alignments	Q_3 accuracy (%)
Protein and ESTs alignments (ProtEST)	72.1
Protein alignments	72.0
ESTs from ProtEST	69.0
Single sequences	65.6

The table compares single sequences, and multiple sequence alignments created from ESTs alone (steps 1–4, Figure 2), protein alone (steps 5–7, Figure 2), and ProtEST sequences (all steps, Figure 2).

result is not surprising, given the redundancy levels in the EST database, and the sequence/organism bias within the known structure database [PDB (Bernstein *et al.*, 1977)]. For the Pfam sequences there was a total of 8749 EST sequences found, which is an average of 10.6 sequences per family. In comparison, the protein search yielded 88 988 sequences for each of the 1407, which corresponds to an average of 63.5 sequences per family.

The protein and the translated, error-checked EST sequences for the 257 successful EST searches for sequences of known structure were combined, and used to predict the corresponding secondary structure by the PHD (Rost and Sander, 1993) algorithm. PHD was chosen since it has been found to be the best single secondary prediction method available that uses multiple sequence alignments for prediction (Cuff and Barton, 1999). Table 2 shows that the resulting average Q_3 accuracy [prediction accuracy over three states (Schulz and Schirmer, 1979)] for the EST plus protein alignments (from ProtEST) was slightly higher, (0.1%) than the average Q_3 for the alignments generated only from the protein sequences. Although this is not a significant improvement, the addition of the EST sequences does not make the predictions any less accurate. Given that the neural networks that PHD has were not trained on alignments extended by EST sequences, without retraining, one would not expect a significant improvement in accuracy. We are currently investigating training a neural network prediction algorithm with EST-derived protein sequence alignments.

The EST contig sequences found by the first stage of the ProtEST search (Figure 2, step 1–4) were then aligned separately. PHD predictions were carried out for each of these EST alignments as before. The final Q_3 accuracy was 3.4% better than predicting from just single sequences, and 3% less accurate than predicting from protein sequence alignments (Table 2). Figure 3 shows an example PHD prediction for 108 residues of the electron transport protein, Thioredoxin (1thx). The average family size for the protein sequences (29) is on average over three times larger than the corresponding EST sequence family (9.5). It is this feature that most likely leads to the difference in secondary structure prediction accuracy.

UniGene (Schuler, 1997) gene clusters of human, mouse and rat sequences were combined to search for contigs. The total number of contig sequences found in the UniGene search was 2951 as opposed to 2442 for ProtEST. UniGene contains longer contigs than can be generated with the organism cluster/PHRAP assembly method of ProtEST. During the length-checking stage, where the translated contigs are compared with the original target protein, 1761 sequences were removed from UniGene contigs; however, 2049 were removed for the ProtEST-generated contigs. This result reflects the very real problem of short contigs, which is apparent in both the UniGene sequences and the automatically generated sequences from ProtEST. For example, the average sequence length of the protein sequences applied in this test was 154 residues, but the ProtEST and UniGene average sequence lengths were 117 and 136, respectively.

When compared with the automatic method of ProtEST, the coverage of UniGene sequences for the PDB test was 43.8%. If the sequences found by both ProtEST and UniGene were combined, the coverage was boosted to 54%. UniGene includes GenBank mRNA and GenBank Genomic sequence as well as the EST database to build its contigs. Although these sections form only a small proportion of the EST sequence (1.5%), including them still adds to the database size and scope. The ProtEST

World Wide Web server (Figure 1) provides an option also to search the UniGene database and combine any results with contigs generated from the ProtEST organism cluster/PHRAP assembly method.

UniGene is updated monthly, while ProtEST dynamically translates the sequences to form protein multiple alignments. The advantages of both approaches are available through the ProtEST server.

This study shows that the quality of the EST sequences and their resulting alignments can significantly improve secondary structure prediction accuracy. This work shows the benefits of using EST sequences in one application of protein multiple sequence alignments. However, it is likely that the addition of EST sequences generated by ProtEST will be beneficial in the future prediction of functional residues and the development of more sensitive profile searching methods.

Acknowledgements

We would like to thank Dr B.Rost and Prof P.Green for supplying PHD and PHRAP, respectively. We also thank Dr J.Parsons for informative discussions. Finally, we would like to thank our reviewers for providing suggestions for improvement to both the manuscript and the ProtEST World Wide Web server. E.B. is a Wellcome Trust student. J.C. is an Oxford Centre for Molecular Sciences/MRC student.

References

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.

Bairoch,A. and Apweiler,R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.*, **183**, 403–428.

Barton,G.J. (1993) An efficient algorithm to locate all locally optimal alignments. *Comput. Applic. Biosci.*, **9**, 729–734.

Barton,G.J. (1995) Protein secondary structure prediction. *Curr. Opin. Struct. Biol.*, **5**, 372–376.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L.L. (1999) Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer basedarchival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Birney,E. (1998) Wise2. http://www.sanger.ac.uk/Software/Wise2/

Birney,E. and Durbin,R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *ISMB*, **5**, 56–64.

Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Cuff,J.A. and Barton,G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins Struct. Funct. Genet.*, **34**, 508–519.

Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

Green,P. (1996) The PHRAP documentation. http://bozeman.genome.washington.edu/phrap.docs/phrap.html

Gribskov,M., Luthy,R. and Eisenberg,D. (1990) Profile analysis. *Methods Enzymol.*, **183**, 146–159.

Kabsch,W. and Sander,C. (1983) A dictionary of protein secondary structure. *Biopolymers*, **22**, 2577–2637.

Livingstone,C.D. and Barton,G.J. (1996) Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.*, **266**, 497–512.

Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **1**, 24–26.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **10**, 694–698.

Schulz,G.E. and Schirmer,R.H. (1979) *Principles of Proteins Structure*. Springer, New York.

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

Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **27**, 18–24.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weigh matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Zhang,Z., Pearson,W.R. and Miller,W. (1997) Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.*, **3**, 339–349.