

Protein and nucleic acid sequence database searching: a suitable case for parallel processing

A. F. W. COULSON*, J. F. COLLINS AND A. LYALL

Department of Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR

Sequence analysis of protein and nucleic acid databases by exhaustive string-matching algorithms is effectively implemented on large processor-array machines, such as the I.C.L. DAP. An improved method of assessing the significance of the best alignments for proteins is described. Examples involving the cystic fibrosis antigen and Drosophila vitellogenins illustrate the usefulness of this approach.

Received June 1987

1. INTRODUCTION

Molecular biology has been revolutionised by the development of fast sequencing techniques for nucleic acids. The rate of acquisition of protein sequence data has correspondingly accelerated, and molecular biological research now depends heavily on gene cloning, sequencing and the translation of open reading frames (which code for possible proteins using the triplet genetic code). This has led to the urgent need for adequate comparative sequence analysis, to promote the efficient use of other research resources.

Proteins and nucleic acids (the genetic material) are linear polymers whose sequences may be represented by character strings, with a 20-letter alphabet for proteins (denoting the individual amino-acid residues), and a 4-letter alphabet for nucleic acids (denoting the individual bases in the DNA or RNA polymers). The international database collections of sequences are prime resources for molecular biological research. These databases are currently small; the protein database has c. 1 000 000 characters of sequence, and the genetic database has c. 10 000 000 bases of sequence information, but already the task of searching them has led to the development of a number of approximate methods for making comparisons. However, the application of the exhaustive inexact string-matching algorithms, reviewed by Sellers,¹ has been beyond the capacity of many workstations and mainframe computers. The situation will deteriorate further as the databases are growing exponentially, doubling in size every two years or less.

We report here our experience using the I.C.L. 64 × 64 Distributed Array Processor (DAP)² for exhaustive database searching. DAP programs for inexact string-matching have been developed by Lyall *et al.*³ of these the most valuable, especially for the case of novel proteins, has implemented the 'Best Local Similarity' algorithm of Smith and Waterman.⁴

It is common for the sequence of part or the whole of a protein to be determined before its function is known. Prediction of function from the analysis of secondary (i.e. local folding along the chain) and tertiary (i.e. assembly of folded regions into a stable structure) structures cannot yet be achieved, and the most profitable approach has been to find analogies with or within the sequences of known proteins.⁵

As the databases grow larger, the number and the scores of alignments with unrelated protein subsequences increase. It is therefore an important issue to determine the significance of the best alignments found, and we describe here a method which is applicable precisely because the whole database has been searched.

2. ALGORITHM AND DAP IMPLEMENTATION

The 'Best Local Similarity' algorithm⁴ is related to other algorithms for sequence comparison and alignment (see Ref. 1), and uses dynamic programming techniques to track the best paths through a match matrix. Each path represents an alignment of the whole or part of the two sequences being compared. At each point in the matrix, the best path is determined by the best cumulative score of paths already running, such that

$$Score(i,j) = MAX(Score(i-1,j-1) + Sim(aa_i,aa_j), \\ Score(i-1,j) + gap\ penalty, \\ Score(i,j-1) + gap\ penalty)$$

where $Sim(aa_i,aa_j)$ is the similarity score for amino-acids aa_i in one sequence and aa_j in the other.

The cumulative scoring is justified in the Dayhoff⁶ analysis by the use of a log(odds) table, the odds being those that a particular pair of residues is found in a significant alignment rather than in a random selection of two residues from the whole population of residues. The figures were derived by Dayhoff from the 71 families of aligned proteins then available. She also described how to produce a series of log(odds) tables corresponding to different evolutionary spans, referred to as the PAM tables (1 PAM corresponds to the appearance of 1 substituted amino-acid residue in a pair of related proteins, per 100 residues aligned). The gap penalty is set to limit the proportion of gaps in the alignments reported to a (subjectively) appropriate level, and to maintain the triangle inequality. Paths are allowed to start at any location inside the match matrix from zero, and are tracked until the score declines to 0 or less, or competing paths block further path extension. Cells scoring less than 0 are reset to 0 before the computation is extended. The best local alignments are found from the maximum path scores, and tracked back through the matrix to their origin.

The DAP host sets up a 2Mbyte DAP core image, and the results are returned as data blocks to the host after

* Principal author, to whom correspondence should be addressed.

the DAP program has terminated. As the sequence alignments are of unknown size at the outset, the program was designed to store the essential details of all runs in a fixed format within the DAP. Implementation with the complete match matrix in main memory is impossible, and for the purposes of coding the algorithm in the DAP, store limitations require that all results be acquired in a single forward pass from data corresponding to a small part of the match matrix, rather than from a double pass through the whole match matrix.

Two rows of the match matrix are used, each 4096 elements long (the DAP long vector length), together with two sets of path data, representing the previous and the current row and path data. The current row is updated; assignments of the score are carried out in parallel by matrix operations for score extensions with diagonal or vertical steps in the paths. The scoring for paths best extended horizontally can then be determined, using recursive doubling combined with logical masks to detect whether further improvements have been made at each cycle. A maximum of 12 iterations completely exhausts all the horizontal path extensions in each segment of the comparison matrix. Paths with an improved maximum score which exceeds a threshold value are reported into the results registers. The results are processed:

- (i) by overwriting any existing inferior path details starting from the same coordinates;
- (ii) or by writing the result into a free location;
- (iii) or, if there are no free locations, by
 - (a) discarding all path details from the lowest class currently stored, marking these locations as available and incrementing the threshold to the score of the discarded class;
 - (b) if the path score to be stored exceeds the new threshold, returning to (ii).

The current paths and details are then written into the 'previous' row and details registers. As the database is considered in 4096-long segments, details of paths leaving at the end of each row are stored, to be made available at the beginning of each row in the next segment of the database. Paths can therefore be tracked wherever they may occur within the match matrix for the whole comparison process.

An advantage of this strategy is that the results accumulated are guaranteed to be the best available by this algorithm, and can be sorted within the DAP; a key is returned to allow host generation of the alignments in order of diminishing score, under user control. In essence, any run can be reconstructed if the coordinates of the start and stop positions are known. However, serial alignment programs calculate possible path states at many locations never included in any path; in the DAP, therefore, the maximum deviations above and below a diagonal path from the start of each path being traced have been added to the set of path details. This provides the host with additional information defining the narrowest band within which each alignment must lie. In the cases of highly related sequences with few gaps, there is a major saving in time in generating the alignments.

The DAP search of version 11 of the NBRF (National Biomedical Research Foundation) protein database, containing 1066790 amino-acids, takes *c.* 1.8 DAP second per residue in the query sequence.

3. SIGNIFICANCE

The assessment of the significance of total or partial alignments between genes or proteins has usually been approached by asking whether the query sequence produces significantly better alignments than sequences derived by randomly reordering the query sequence.⁷ Two points arise here; the time to search a database is significant even on the DAP, and the investment of more CPU time to discover the statistical behaviour of random sequences each time is not attractive. Secondly, the database is not a collection of sets of randomly ordered characters; in general, proteins share characteristic structure features in the natural folded state (for instance, the alpha-helix, beta sheet and various types of turn) and these are reflected in short-range ordering within the protein sequence. Therefore, real proteins are likely to contain regions of better local similarity with each other than with random rearrangements of the same residues.

The DAP program returns data for the 4096 best alignments, which form in most searches the upper end of a much larger distribution of scores. The database is highly diverse, and no single family of related proteins is represented much more than 100 times. Hence the majority of the best results will be of alignments between regions of the query sequence and proteins in the database which have no close connection; in other words, these are alignments representing the noise-level in comparisons with a large collection of unrelated proteins. If we can determine the underlying shape of this distribution, we can predict the frequency of occurrence of an alignment of any score arising from unrelated proteins, and so establish the likelihood that any particular alignment belongs to this class or not.

We can regard the alignments reported by the 'Best Local Homology' algorithm as a series of aligned pairs which may be scored positively or negatively, and unmatched residues in either strand, where gaps have been introduced (under penalty). Each reported alignment starts with a positive score, and terminates where the cumulative score reaches a maximum.

The analysis of significance does not require knowledge of the complete distribution of scores. All that is needed is a model for the expected value of the ratio of the number of alignments scoring ($n+1$), to the number of alignments scoring (n). The most important route by which an alignment could improve from a score of n to $n+1$ (or beyond) is from the position at which the current maximum score n was reached. The probability that, within a region of the match matrix through which the path can be extended with net loss of x in score, there is a matching region from which a net gain of $(x+1)$ can subsequently be obtained, must be independent or nearly independent of the current maximum score, once this has exceeded a low value. This implies that the distribution of path scores will decline exponentially, and this is indeed found experimentally to be the case.

The lower-scoring 98% of the recorded alignments were therefore analysed by fitting the best line to the $\log(\text{no. of alignments})$ *v.* score with excellent results, providing parameters to estimate the expected frequency of any scoring alignment, as well as standard deviation for the distribution about the line. The high-scoring outlying alignments can be tested for their significance by seeing how well they conform to this distribution; especially,

how many standard deviations they are above the expected frequency, thus expressing the likelihood of any alignment occurring with unrelated proteins.

However, it must be emphasised that any alignment can potentially provide the molecular biologist with useful information, and between 50 and 200 are normally collected for display.

4. PATTERN DETECTION AND SEARCHING

Additional processing of the results can provide further useful displays; for example, when a query sequence is related to a number of sequences in the database, the alignments can be accumulated, or 'learnt', so that it is possible to display a large number of alignments with respect to the query sequence. This is a sensitive method of finding conserved residues of short-sequence features, which are difficult to detect in individual alignments. The learning process can be guided by different criteria, to allow the disclosure of patterns relating different types of sequence within the same set of alignments. Each search for patterns can then be reinforced by re-searching the database with the pattern detected, to establish and refine its ability to discriminate sequences fitting the pattern from the bulk of unrelated proteins.

The ability to detect patterns can be extended by using a more general method of describing a pattern. In principle, a pattern search could be carried out with specific values for all matching possibilities at each position. However, we have provided simple general extensions which have been of considerable value, defining four types of character:

(i) normal characters, matched using the similarity table values, and attracting a gap penalty if unmatched in an alignment;

(ii) residues which attract scores from the similarity table used, but which must be matched with a positive score in any reported alignment;

(iii) residues which may be matched with any residue, with zero score, but which cannot be unpaired without attracting the gap penalty; and

(iv) residues which may match any character without preference, with zero score, and which may be omitted without penalty.

This has provided a flexible and versatile pattern-detection and searching tool. A specific advantage that distinguishes this mode of pattern detection from the 'regular expression' pattern-searching programs is that, in addition to exact fits to the specified pattern, other near-fits are scored and can be reported, including those with a wider range of character substitution or spacing than envisaged in the specification of the pattern. That is, it is possible to discover unexpected ways in which the pattern is variable, without explicit definition of these alternatives.

Such searches are providing interesting results with complex polypeptides deduced from viral gene sequences, where the ability to detect conserved features may help define regions of importance in the normal function of the polypeptide in the viral lifecycle, helping to define these features for further research.

For example, a pattern for a zinc-ion binding site has been proposed⁸ to be $CX_{2-4}CX_{2-15}(C \text{ or } H)X_{2-4}(C \text{ or } H)$, where C stands for the amino-acid cysteine, H for the

amino-acid histidine, and X stands for an unspecified amino-acid. The database can be searched with this pattern in c. 60 DAP seconds and the best matching regions readily listed to verify this hypothesis.

5. EXAMPLES

5.1 Cystic-fibrosis associated antigen

A gene cloned by Dorin *et al.*⁹ coded for a protein found at elevated levels in the serum of cystic fibrosis patients and carriers. The gene was translated into the protein sequence, and the database search found that there were significant homologies with calcium-binding proteins, using the 250 PAM similarity table. The search was repeated with a variety of PAM tables, and the maximum significance was found with the 80 PAM table. The set of results is shown in Fig. 1. The best alignment (Fig. 2) had an expected frequency of 1.7×10^{-10} (44 s.d.s above expectation), and clearly indicated that the alignment belonged to a different class from those forming the bulk of the reported results, and which arise from biologically unrelated proteins.

Twenty out of the next 21 alignments were with proteins all known to bind calcium ions, and this fact would have indicated the same property in the cystic fibrosis antigen, even in the absence of the proteins giving very high-scoring alignments.

5.2 Vitellogenins in *Drosophila melanogaster*

Garabedian *et al.*¹¹ have shown that the sequences of three storage proteins (YP 1, YP 2 and YP 3) found in the eggs of the fruit fly share a long region of highly conserved sequence. A database search revealed that this

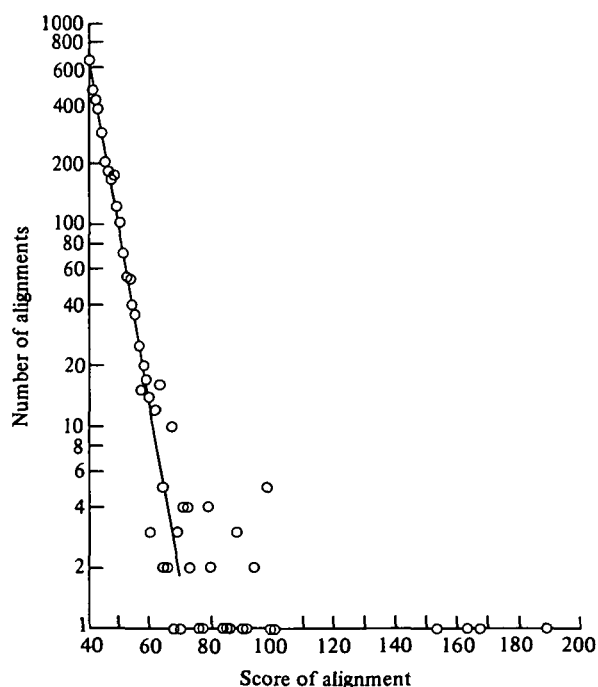


Figure 1. Distribution of the best alignment scores reported by the DAP protein database searching program, using the cystic fibrosis-associated antigen sequence, and the similarity table calculated for 80 PAMs.⁶ The line of best fit to the lower 98% of the results is shown. The highest-scoring alignment would be expected at a frequency of 1.7×10^{-10} , 44 s.d.s from the expected value.

2 SELETAMETLINVFHAHSGKEGDKYK LSKKELKELLQTELSGFLDAQKDADAVIDKVMKELDEDGDGEVDVFQEYVVLV 78
3 TELEKALNSIIDVYHKYSLIGN FHAVYRDDLKLLTECPOYI RKKGAD V W FKELDINTDGAVNFOEFLILV 75

Figure 2. Alignment of the amino-acid sequences of the cystic fibrosis antigen (lower sequence) and the bovine s-100a alpha protein chain (upper sequence), in the highest-scoring alignment reported by the 'Best Local Homology' algorithm. The one-letter notation for amino-acids is that recommended by the IUPAC-IUB Commission on Biochemical Nomenclature (the *Biochemical Journal*, 113, 1 (1969). Numbers at the ends of the sequence segments indicate the position within the entire protein chain of the aligned residues. Identical pairs of residues are starred; pairs scoring positively but non-identical are dotted; all others attract penalties.

```

146  .*.***.....* ** ** . * * .***** * . * * .** *** **..
239  VHVIGHSLGSHAAGEAGRR T NG TIERITGLDPAEPCFQGTPE LVRLDPSDAKFVDVIHTDAAP 208
    IHLIGQGISAHVAGAAGNKYTAQTGHKLRRITGLDPAKV LSKRPQILGGLSRGDADFVDIAHT ST 303

    . . * . . * . * .*** . . ** .. * ... .... * . * . * . * * . **
209  IIPNLGFGMSQTVGHLDFFPNG GKQMPGCQKNI LSQIVD I DGIWEGT RDF VACNHLRSYK 268
304  F A MGTP I R CGDVDLYPNGPSTGVPGSENVIEAVARATRYFAESVRPGSERNFPAVPANSLKOYK 367

```

Figure 3. The best scoring alignment found between the vitellogenin YP 3 from *Drosophila melanogaster* (lower sequence), and pig lipase (upper sequence). Expected frequency for an alignment with this score: 1.17×10^{-7} ; 53 S.D.s above expectation.

conserved region otherwise aligned best with part of a sequence from a pig digestive lipase, and that this similarity was highly significant (expected frequency 1.17×10^{-7} ; 53 S.D.s above the predicted expectation) (Fig. 3). However, one residue thought to play a role in the catalytic activity of the lipase was not matched in the alignment (Fig. 4). The vitellogenins, in fact, do not show lipase activity. It was postulated that the similarity found is related to the ability of these proteins to bind lipids or lipid-like materials. Subsequent tests have shown that the vitellogenins strongly bind a natural lipid-like derivative of the insect hormone ecdysone, which is involved in the control of embryonic development. The embryo breaks down vitellogenin at the stage when the hormone is known to be released, and it now appears that these proteins may have an important role in regulating embryonic development.

shows that the DAP can match this more powerful machine (and at a fraction of the cost).

As the database grows, the biologist is as interested in increasing the variety of known sequences as in providing new examples of known protein types. Increasingly, more and more of the database for proteins is likely to be hypothetical proteins, inferred from gene sequences, whose physical and chemical properties and biological role have not been observed.

For this reason, results which have strong statistical significance are only part of the value the biologist can draw from these searches. If the query sequence is related, distantly, to a database sequence which is unique, that fact may be enough to generate a biological hypothesis which can then be tested further. The value to the biologist of having a complete and exhaustive search carried out is much more, therefore, than finding the single best alignment; valuable information may be gained from the presence of groups of related alignments and even from a single alignment of low statistical significance. This fact differentiates the biological database search problem from more conventional database searching problems.

It will be important to maintain some facility which can fulfil this search role in the near future, while the sizes and rate of increase in the databases can still be handled. When the promised improvements in sequencing technology are implemented, and gene sequences can be accumulated at 1000000 bases per day, a new crisis will have to be faced if this information is going to be of a significant use to the biological community.

Acknowledgements

We should like to thank Professor S. Michaelson (Department of Computer Science) for his help and advice; Dr S. Reddaway and Dr D. Hunt (I.C.L.) for their insights into the DAP; Carolyn Bucholtz (C.S.I.R.O., Sydney) for assistance. A.L. gratefully acknowledges an SERC CASE award with I.C.L.

.*. *** ** ** . *

VHVIGHSLGSHAAGEAGRR T

IHLIGOGISAHVAGAAGNKYT

Figure 4. Alignment of vitellogenin YP 3 (lower sequence) and pig lipase (upper sequence), in the region of the active serine (S) residue in lipase. The serine is aligned against a glycine (G) residue, which would not be expected to substitute functionally for the serine.

6. DISCUSSION

The nature of protein and nucleic acid sequences makes them immediately suitable for processing by network-connected arrays of processors. The efficiency of the DAP 1-bit processors is particularly high, since much of the arithmetic can use 1- to 2-byte variables, and there is a large component of logical operations. Data movements are predominantly long vector shifts, which are slightly more complex than simple vector shifts. The published account of sequence comparison on a Cray-1 machine¹⁰

REFERENCES

1. P. H. Sellers. The theory and computation of evolutionary distances: pattern recognition. *Journal of Algorithms* **1**, 359–373 (1980).
2. P. Flanders, D. J. Hunt, S. Reddaway and D. Parkinson. Efficient high speed computing with the Distributed Array Processor. In *High Speed Computer and Algorithm Organisation*, edited D. J. Kuck, D. H. Lawrie and A. H. Sameh, pp. 113–120. Academic Press, London (1978).
3. A. Lyall, C. Hill, J. F. Collins, and A. F. W. Coulson. Implementation of Inexact String Matching Algorithms on the I.C.L. DAP. In *Parallel Computing '85*, edited M. Feilmeier, G. Joubert and U. Schendel, pp. 235–240. North-Holland, Amsterdam (1986).
4. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).
5. J. F. Collins and A. F. W. Coulson. Molecular sequence comparison and alignment. In *Nucleic Acid and Protein Sequence Analysis: A Practical Approach* edited M. Bishop & C. Rawlings, pp. 323–358. I.R.L. Press, Oxford (1987).
6. M. O. Dayoff, R. M. Schwartz and B. C. Orcutt. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3, edited M. O. Dayhoff, pp. 345–352. N.B.R.F., Washington (1978).
7. D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searching. *Science* **227**, 1435–1441 (1985).
8. J. M. Berg. Potential metal-binding domains in nucleic acid binding proteins. *Science* **232**, 485–487 (1986).
9. J. R. Dorin, M. Novak, R. E. Hill, D. J. H. Brock, D. S. Secher and V. van Heyningen. A clue to the basic defect in cystic fibrosis from cloning the CF antigen gene. *Nature* **326**, 614–617 (1987).
10. T. F. Smith, M. S. Waterman and C. Burks. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13**, 645–665 (1985).
11. M. J. Garabedian, A. D. Shirras, M. Bownes and P. Wensink. The nucleotide sequence of the gene coding for *Drosophila melanogaster* yolk protein 3. *Gene*, in press.