Protein Sequence Analysis Program

—Multiple Sequence Alignment by Parallel Iterative Aligner—

ABSTRACT

Every protein is a chain of twenty kinds of amino acids. Its structure and function are determined by the sequence of these amino acids. Protein sequence analysis is vital to the prediction of protein structure and function, and to the analysis of the evolutional relationship among proteins. We built the "Parallel Iterative Aligner" for this analysis using a PIM machine. It solves a typical protein sequence analysis problem, that of multiple sequence alignment. This kind of combinatorial problem generally requires a large amount of computation. Our Parallel Iterative Aligner, however, can quickly solve practical-sized problems, producing high-quality answers.

KEY FEATURES

Iterative Improvement

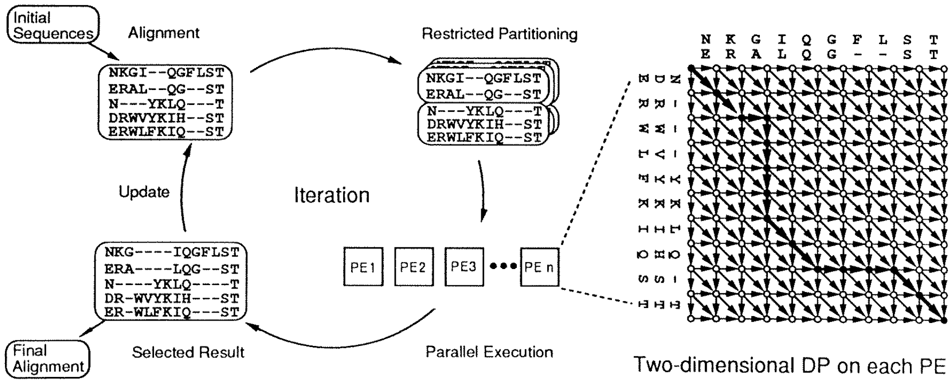
The system partially improves temporary alignment in an iterative way, and can effectively reach high-quality alignment as a result.

Speed-up by Parallel Execution

Multiple branches of a search tree in this combinatorial problem are evaluated in parallel by using many processing elements in each iteration.

Pruning of Search Tree

The heuristic method, "Restricted Partitioning Technique," prunes a large number of branches in the search tree and makes it possible to solve the combinatorial problem in a practical amount of time.



Operation of Parallel Iterative Aligner

What is protein sequence analysis?

Proteins are made up of twenty kinds of amino acids, which we distinguish by twenty different code letters. A protein has about two hundred amino acids on average and is represented by a sequence of code letters. Because every amino acid has its own properties of volume, hydrophobicity, polarity and so on, the order of the amino acids in the protein sequence gives the structure and function of the protein.

The protein sequence determination technique has been established for so long that more than twenty thousand sequences have been specified by the letters; this number is growing day by day. The structures of proteins are also being solved. Methods such as X-ray crystallography reveal how a chain of amino acids folds together. But these methods take many months to complete, so only three hundred protein structures have been determined so far.

An important way of discovering new biological information is by inferring the unknown structure of a protein from its sequence. We do this by analyzing the sequence of amino acids, because, fortunately, proteins that have similar sequences have similar structures. Multiple sequence alignment is one of the most typical methods of sequence similarity analysis. The alignment of several protein sequences can provide valuable information for researching the function or structure of proteins, especially if one of the aligned proteins has been well characterized.

Let us show an example of multiple sequence alignment. The following set of sequences represents an alignment of six different protein sequences. `HEKL` stands for a row of Histidine, Glutamic acid, Lysine and Leucine.

```
---------------HEKLLHPGIQKTTKLF-GET---YYFPNSQLLIQNIINECSICNLAKTEHRNTDM--P-TKTT
--------------LHQ-LTHLSFSKMKALLERSHSPYYMLNRDRTL-KNITETCKAC--AQVNASKSAVKQG-TR--
-PVLQ---LSPA-ELHS-FTHCG---QTAL--TLQ----GATTTEA--SNILRSCHAC---RGGNPQHQMPRGHI---
QATFQAYPLREAKDLHT-ALHIG---PRAL--SKA---CNISMQQA--REVVQTCPHC------NSAPALEAG-VN--
--ISD--PIHEATQAHT-LHHLN---AHTL--RLL---YKITREQA--RDIVKACKQC---VVATPVPHL--G-VN--
--ILT--ALESAQESHA-LHHQN---AAAL--RFQ---FHITREQA--REIVKLCPNC---PDWGSAPQL--G-VN--
               ^    ^                                 ^  ^
```

Each sequence is shifted by inserting gaps (dash characters). Each column of the resultant alignment has the same or similar amino acids. An identical pattern such as `H....H` and `C..C` is considered to be an important site called a *sequence motif,* or simply a *motif,* because an important protein sequence site has been conservative along with evolutional cycles between mutation and natural selection. Multiple sequence alignment is useful not only for inferring the structure and function of proteins but also for drawing a phylogenetic tree along the evolutional histories of the creatures.

Dynamic programming on sequence alignment

Dynamic programming (DP) is a basic method to find an optimal alignment. The method is regarded as the best path search in the N-dimensional network. In the method, if two groups of sequences are given, a two-dimensional network that has a number of nodes connected by arrows is formed (Figure

1). A score is assigned to each arrow. We search a path from the top left node to the bottom right node, maximizing the total score of the arrows. In this case, the set of arrows that connect black circle nodes is the best path. This best path corresponds to the optimal alignment.

Scores on arrows should reflect similarity between compared characters. In the case of protein sequence alignment, Dayhoff's odds matrix is the most popular way of obtaining the scores. The matrix was obtained by statistical analysis of the mutation probability of amino acids.

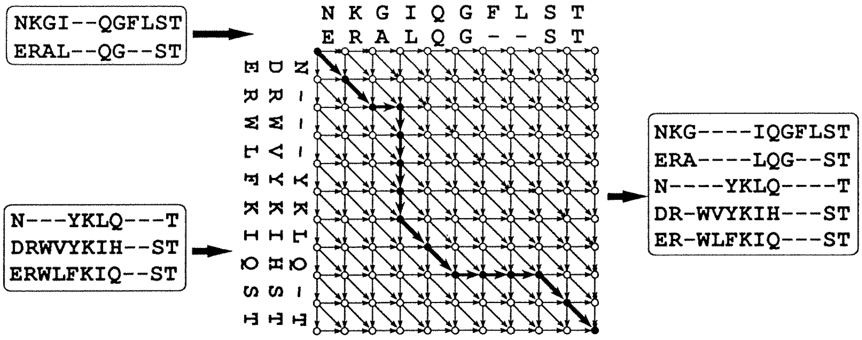


Figure 1 Two-dimensional DP

Theoretically, N-dimensional DP provides an optimal alignment of N groups of sequences. However, N-dimensional DP operates in exponential time as N grows. When N is more than three, it does not complete in a realistic time frame. Conventionally, researchers in the biological field make a multiple sequence alignment by merging groups of aligned sequences. A conventional algorithm, called the *tree-based algorithm*, merges them in tree-like order using two-dimensional DP.

Though the execution time of the conventional algorithm is manageable, the quality of its resultant alignment is not high enough yet. Thus, researchers fairly often have to do multiple sequence alignments by hand. The large number of sets of sequences to be aligned have become a burden on those researchers.

Parallel iterative aligner

We developed a *parallel iterative aligner*, whose execution will be demonstrated, in order to improve the quality of automatic multiple sequence alignment. The algorithm of this parallel iterative aligner is based on the Berger-Munson algorithm. Firstly, we introduce the B-M algorithm, and then we explain the parallel iterative aligner.

The B-M algorithm features a novel randomized iterative strategy so as to generate a high-score multiple sequence alignment. Figure 2 illustrates the iterative strategy, whose procedure is as follows: the initially aligned sequences are randomly divided into two groups. By fixing the alignment of sequence members within each group, we can optimize the alignment between the groups, using two-dimensional DP. The resultant alignment, in turn, is the starting point for the next alignment of a different pair of groups. Each iteration that improves the alignment between two sequence groups will also improve the global alignment.

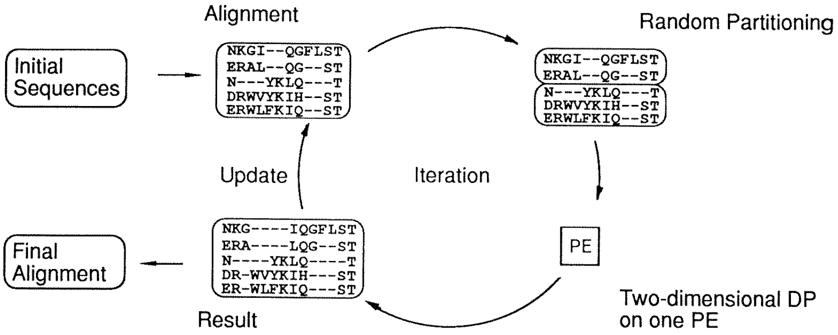


Figure 2 Original B-M algorithm

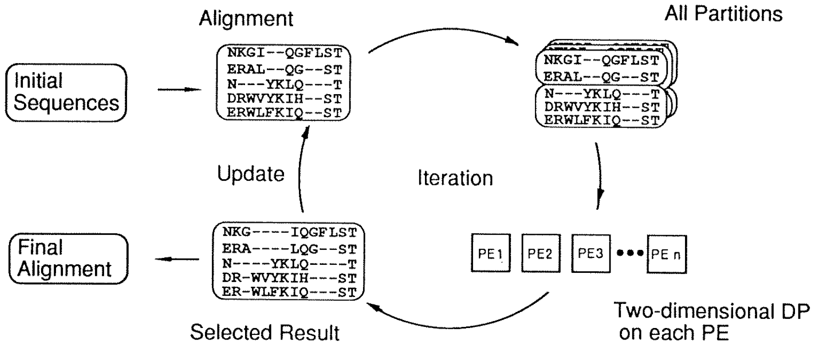


Figure 3 Parallel B-M algorithm

This iterative strategy often results in much better multiple alignments than those obtained by conventional algorithms. In Figure 4, (b) shows the performance when this B-M algorithm is applied to seven sequences. Lines (b-1), (b-2) and (b-3) are different in respect to random numbers. The quality of the results is superior to that obtained by the tree-based algorithm, shown as a horizontal line (a). However, the B-M algorithm needs a large amount of time as the number of sequences grows.

We can reduce the execution time, when a parallel machine is available. Figure 3 shows the algorithm of our parallel iterative aligner. Every possible partitioning into two groups of aligned sequences can be respectively evaluated by two-dimensional DP in a parallel way. In each iteration, the evaluation is executed in parallel and the alignment which has the best score is selected as the starting point for the next iteration. The performance of this method is shown as line (c) in Figure 4. This shows that the parallel iterative aligner performs better than the original B-M method in terms of execution time.

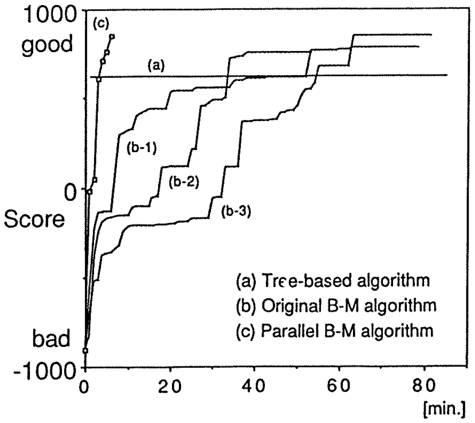


Figure 4 Performance comparison

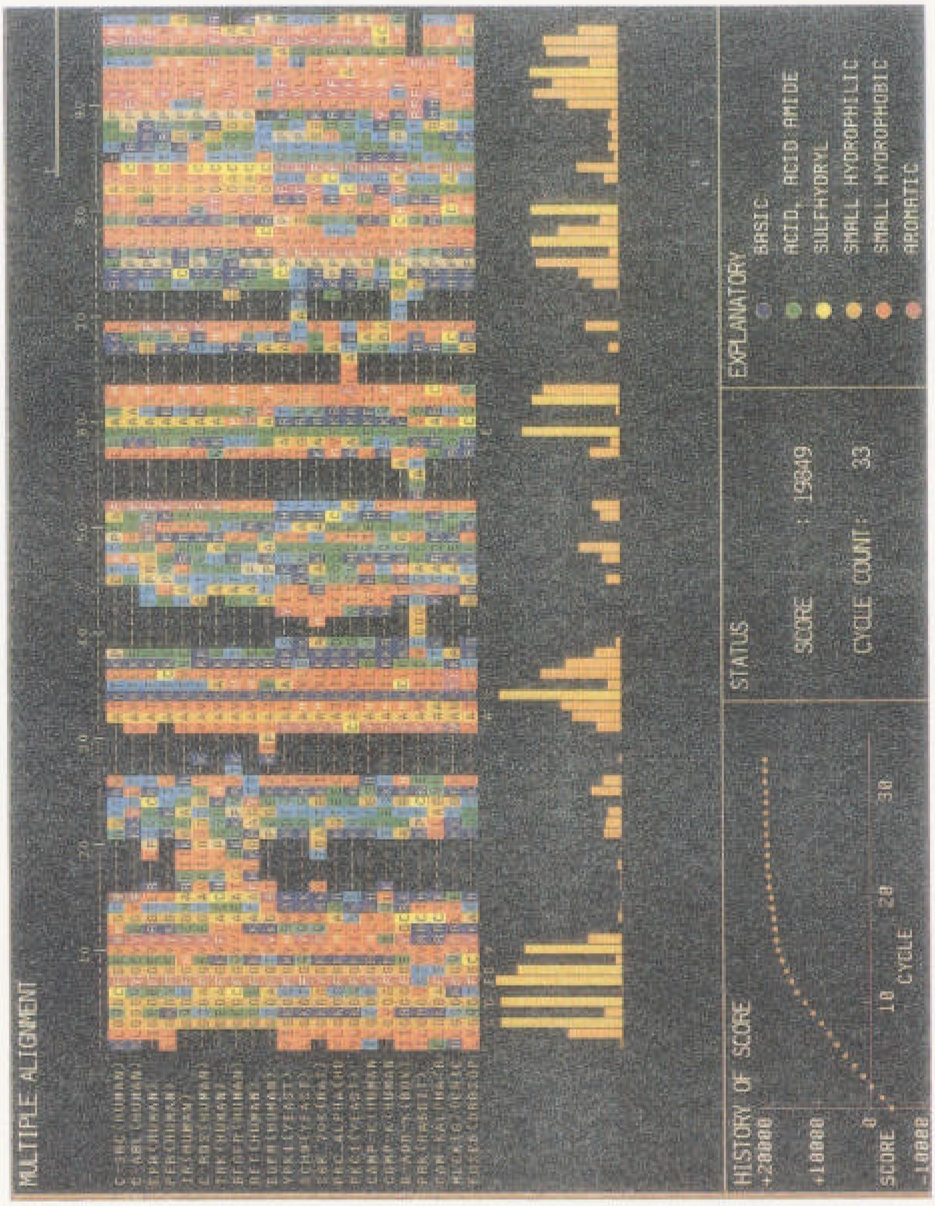
Furthermore, we have developed an effective heuristic search, *restricted partitioning technique.* Applying the iterative strategy, we realized that the number of sequences in the divided groups is important. As partitioning divides N sequences into k sequences and $N-k$ sequences, a smaller k tends to provide a larger improvement when using two-dimensional DP. The restricted partitioning technique preferentially selects partitionings which have a small k such as one or two. It can restrict the search space and reduce the execution time remarkably. Parallel iterative aligners with this technique can manage more sequences at the same time than those without it.

Demonstration

The demonstration system solves an alignment of 22 sequences using the parallel iterative aligner. It is executed on PIM/m, which consists of 256 processing elements (PEs).

We use the restricted partitioning technique; the number of sequences in the smaller divided group is restricted to one or two. 253 PEs ($_{22}C_1 = 22$ plus $_{22}C_2 = 231$) are necessary to execute all restricted partitioning in parallel. Additionally, one processor is used for a manager process which selects the best alignment in every iteration. In total, 254 PEs are employed in this demonstration.

The demonstration shows gradual improvement of the alignment in each iteration. Code letters for amino acids in the display are colored according to their properties (see next page). The score for each column using a bar graph under the alignment is also shown. High-score columns, indicated by tall bars, correspond to well-conserved sites in evolution. Representative sequence patterns in the sites are regarded as motifs. In the final resultant data of this demonstration, we can recognize a number of motifs. One of them is a famous motif, the site of which binds ATP.



SCORE : 19849

CYCLE COUNT: 33

Figure 5 Demonstration display