TM-1287

Constraint Based Alignment Editor with Automatic Alinger and Motif Dictionary Retriever

by
M. Hirosawa (Hitachi), Y. Totoki,
& M. Ishikawa

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5

Institute for New Generation Computer Technology

Constraint Based Alignment Editor with Automatic Alinger and Motif Dictionary Retriever

MAKOTO HIROSAWA† hirosawa@sdl.hitachi.co.jp
YASUSHI TOTOKI‡ totoki@icot.or.jp
MASATO ISHIKAWA‡ ishikawa@icot.or.jp

†ex-ICOT, presently with Hitachi System Development Laboratory
1099 Ohzenji, Asao, Kawasaki, Kanagawa 215 JAPAN [1]

‡Institute for New Generation Computer Technology (ICOT)
1-4-28 Mita, Minato, Tokyo 108 JAPAN

Abstract

We have developed an alignment editor for protein sequences and/or nucleotide sequences that runs under the on X-window system. The system is equipped with automatical aligner and motif dictionary. The system can iteratively identify known motifs stored in the motif dictionary and gradually improve the quality of the alignment to discover new motifs.

[1]元ＩＣＯＴ、現 日立 システム開発研究所： 〒 215 川崎市 麻生区 王禅寺 １０９９

1 Introduction

The Multiple alignment of protein is important for drawing the phylogenic tree of species and predicting the structure of proteins. Also the multiple alignment of protein is used to discover motifs, biologically important patterns, from protein sequences.

In our previous research project [Hirosawa *et al.* 1993a], we developed an automatic alignment and motif-discovery system (Figure 1). This iteratively refines the protein sequence alignment and gradually identifies motifs in sequences. When necessary, the system consults biological knowledge, which is composed mainly of motif knowledge that has been identified. The knowledge is written in Deductive Object-oriented database language *QUIXOTE* [Yasukawa *et al.* 1992].

This time, we took a different approach. We developed a semi-automatic alignment and motif-discovery system, *motif-aligner* (Figure 2). Figuratively, motif-aligner iteratively refines the protein sequence alignment to find motifs by consulting a motif dictionary in the system, as well as the knowledge base in the user's own brain. Seriously speaking, the system is basically an alignment editor on X-window, equipped with an automatic aligner and a motif database. In the following, the system, the motif-aligner will be explained in detail.

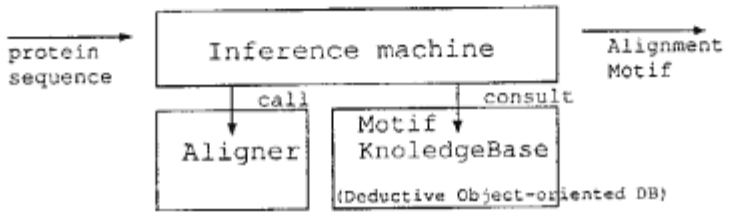

Figure 1. Automatical Aligner and Motif-discovery System

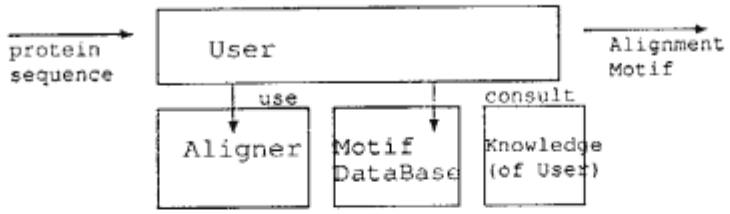

Figure 2. Semi-Automatical Aligner and Motif-discovery System

2 Overview of the system

The motif-aligner runs under the X-window system. Its characteristics are explained below.

Editing with the mouse Users can edit the alignment by using the mouse. For those users who are more familiar with using keyboard, editing with the keyboard is also supported, but the use of the keyboard is basically been dispensed.

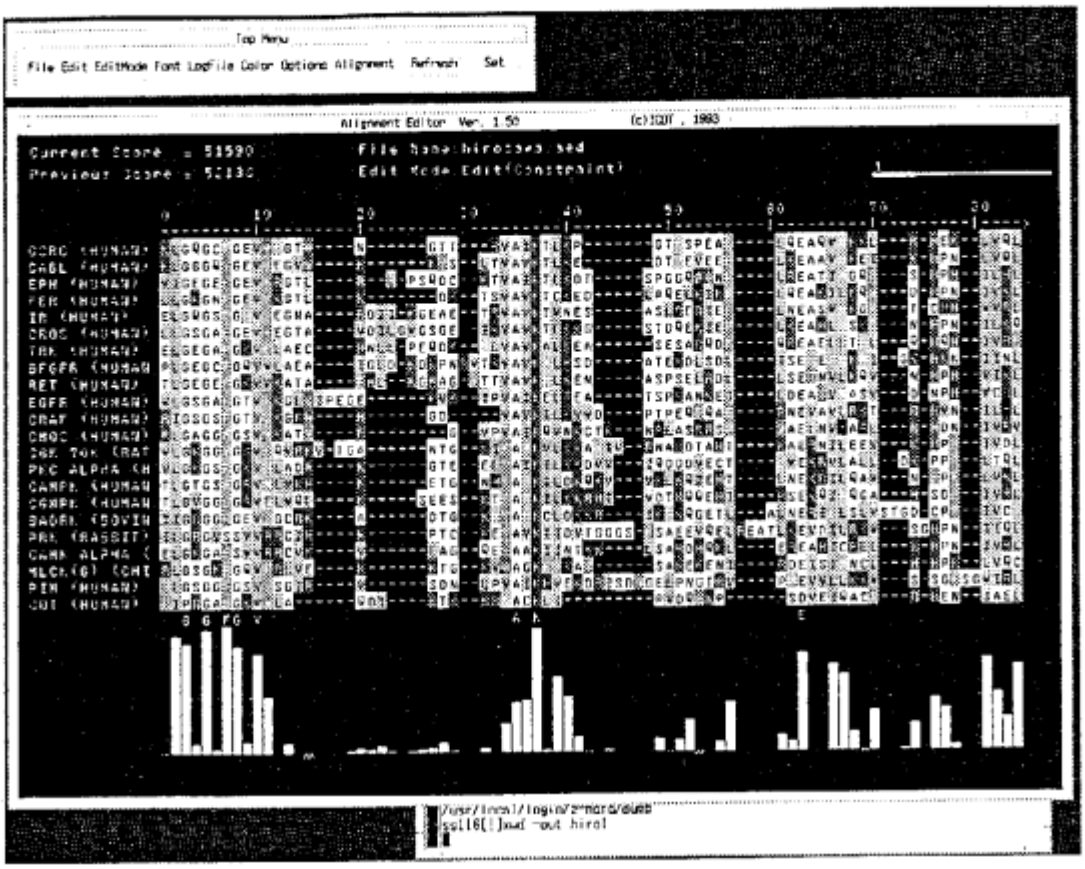

Figure 3: The initial alignment to the motif-editor

Automatic Alignment The motif-aligner automatically aligns the region of sequences specified by the user. There are three alignment algorithms that the user can select. The three differ in the quality of alignment that the selected algorithm produces, and also in the time needed to produce the alignment

1. TB (Tree-based algorithm)
2. TIR (Tree-based iterative improving method(Round Robin))
3. TIH (Tree-based iterative improving method(Hill climbing))

Among the three, TB [Barton 1990] is the fastest algorithm but offers the lowest reliability. TIH [Hirosawa 1993b] is the most reliable but the most time consuming. TIR falls between the other two, both reliability and time required. Actually, TIR is an approximation algorithm of TIH.

Users can select the algorithm according to their needs. For example, TB can be used to produce the initial alignment through sequences and TIH can be used to refine the alignment within a restricted region of sequences.

Constraint-based alignment In the alignment that the user is currently processing, if they find a region of alignment for which they are confident of the quality, users can specify the region as a constraint region. After specifying the region as a constraint region, the user and automatic aligner of the motif-aligner cannot modify the alignment of the

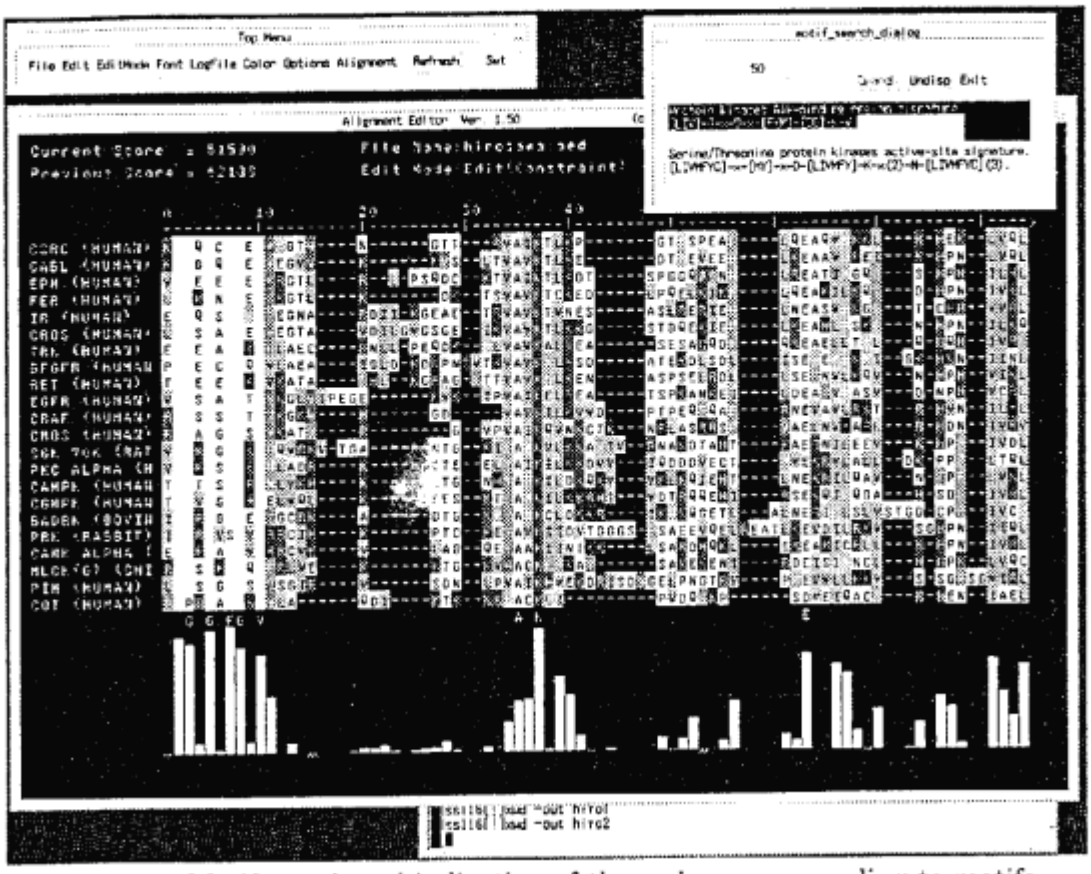

Figure 4: Motif search and indication of the region corresponding to motifs

constraint region unless they cancel the specification. The existence of the constraint region reduces the solution space of the alignment. Users can select a more reliable algorithm to automatically align sequences due to this reduction in the solution space.

Motif Dictionary The system is equipped with a motif dictionary. The system can search for motifs in the alignment and shows the users any region possibly corresponds to some registered motif. It also shows the users the significance of the motif.

If users regard the indicated region as being a motif, they can specifies the region, corresponding to some motif, as a constraint region to avoid modification of the region.

Presently, motifs registered in Prosite [Bairoch 1991] are registered in our motif dictionary (Prosite is a representative motif dictionary of protein). But the user can register motifs that they know in a simple way. All the information necessary to register the motif is its pattern and its significance.

Phylogenetic tree The system can draw the phylogenetic tree of the alignment using UPGMA (unweighted pair-group arithmatic average clustering [Sneath and Sokal 1973]). Users can refine the alignment by consulting the phylogenetic tree.

Others Users can select the font used to represent amino acid or nucleotide among the fonts stored in the workstation on which they are currently working. Also the user can select the color of amino acid or nucleotide using any of 256 supported colors. The coloring of amino acids according to property group of amino acids facilitates their alignment.

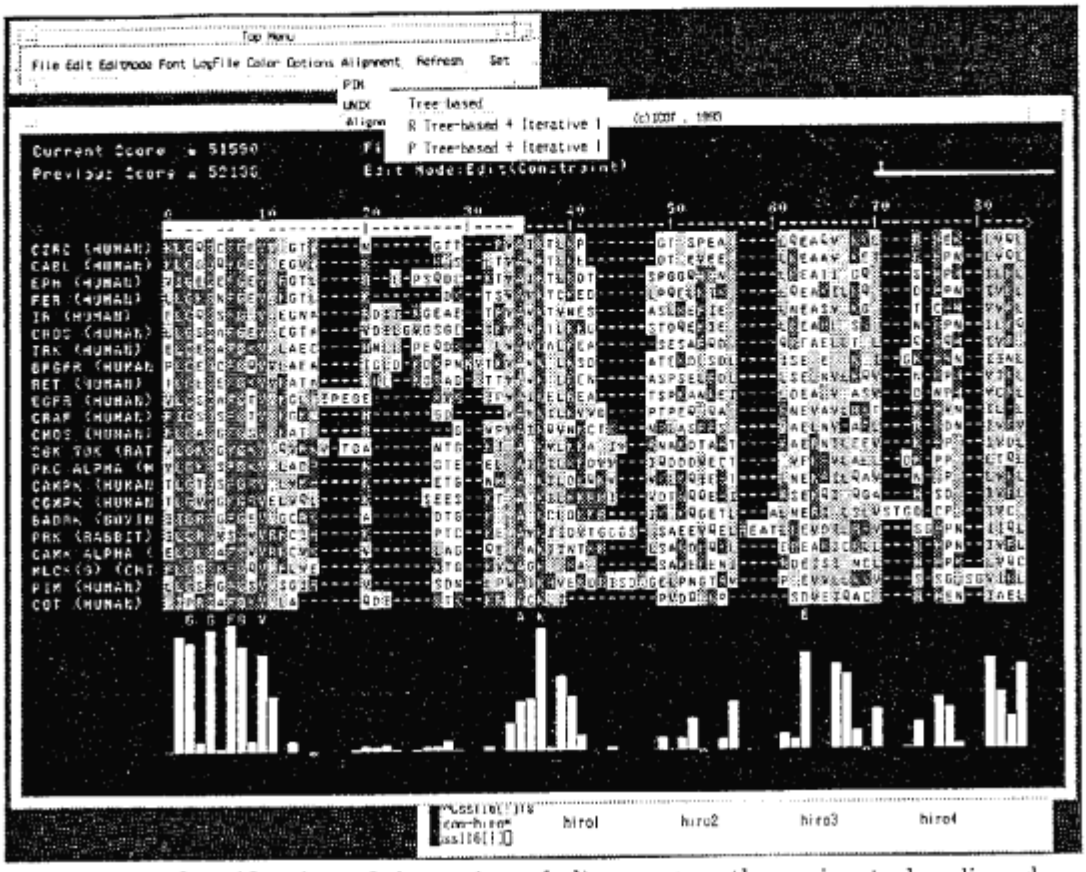

Figure 5: Specification of the region of alignment as the region to be aligned

3 Example application

An example application of motif-editor is given below with actual image of the screen display. Figure 3 shows the initial multiple alignment of 22 protein sequences. Amino acids are colored according to their class they belong to. The bar graph below the alignment indicates the similarity of amino acids in the same column. As measure of similarity score Dayhoff's matrix is used. We can edit the alignment by referencing the bar graph.

Firstly, we search known motifs, registered in Prosite, in the alignment (Figure 4). Two motifs are found in the alignment and the region of alignment corresponding to one of the motifs, protein-kinase-ATP-binding-signature, is indicated as a whitened area.

Secondly, the region of alignment to be refined is specified. In Figure 5, the region between the former motif and subsequence which is characterized by pattern "A-x-K" is specified to be aligned. As algorithm for aligning the sequences is to be selected. In this case, we use TIH (corresponding to "Tree-based + Iterative 1" in the figure) as the algorithm.

In Figure 6, the alignment of the selected region is displayed, together with the present alignment. The score of the present alignment in the region, 5377, and that of newly created alignment, 5950, are compared. Because the score of the newly created alignment is better, the new alignment is substituted into the present alignment. Then, the updated alignment is displayed in Figure 7. In the figure, phylogenetic tree of the alignment is also shown.

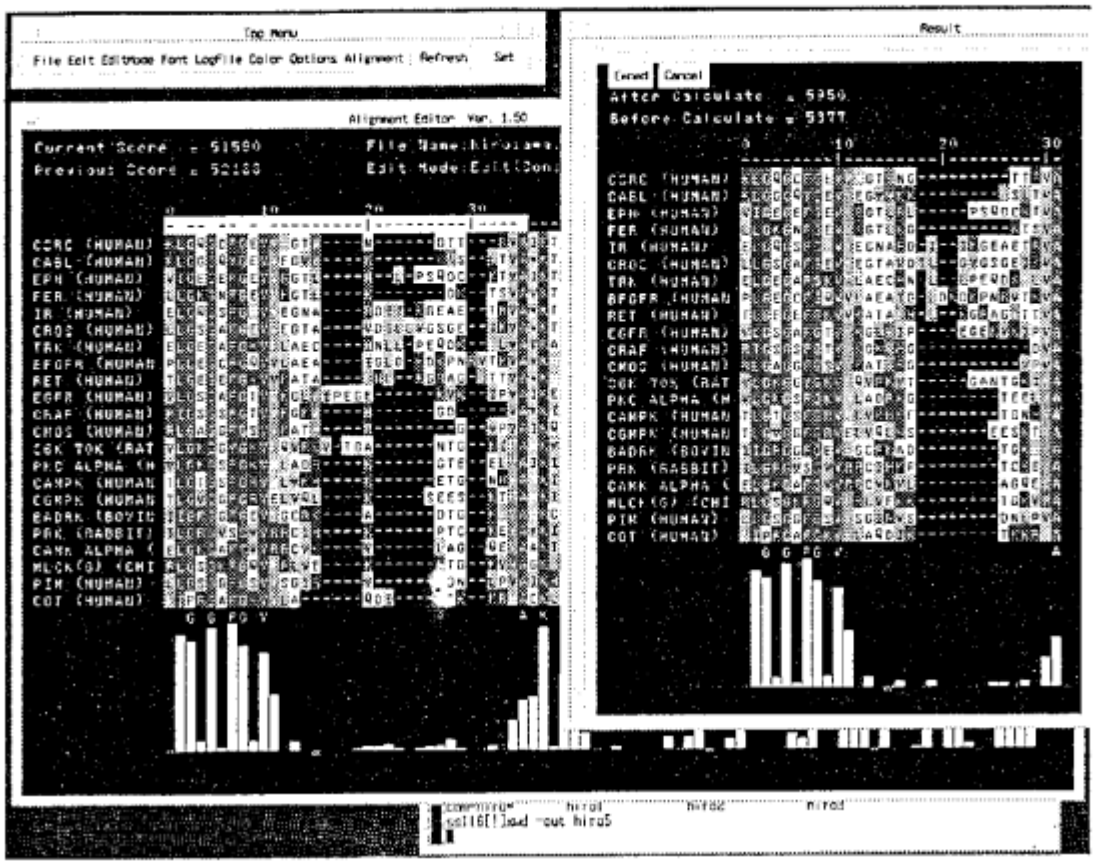

Figure 6: Newly created alignment of the region

4 Discussion

The alignment produced by the motif-editor is almost the same as that produced by the automatic alignment system [Hirosawa 1993b]. The difference is that in motif-editor, we can select any quality of alignment according to the time available, whereas in the automatic system we cannot. If the user doesn't want to be bothered by the detailed aspects of alignment, the automatic aligner is suitable. Otherwise, the alignment produced by motif-editor is preferable.

5 Obtaining the program

The motif-aligner program is registered as *ICOT free software* and can be obtained by using of ftp. The internet address to be accessed is 192.26.9.333 and the program is stored in /ifs/exper-apps/pimos/editalign.tar.Z .

References

[**Bairoch 1991**] Bairoch, A. Prosite : A dictionary of protein site and pattern : User manual Release 7.00, May 1991.

[**Barton 1990**] Barton, G.J. Protein Multiple Sequence Alignment and Flexible Pattern Matching. *Methods in Enzymology Volume 183* Academic Press, 1990, pp. 403–428.

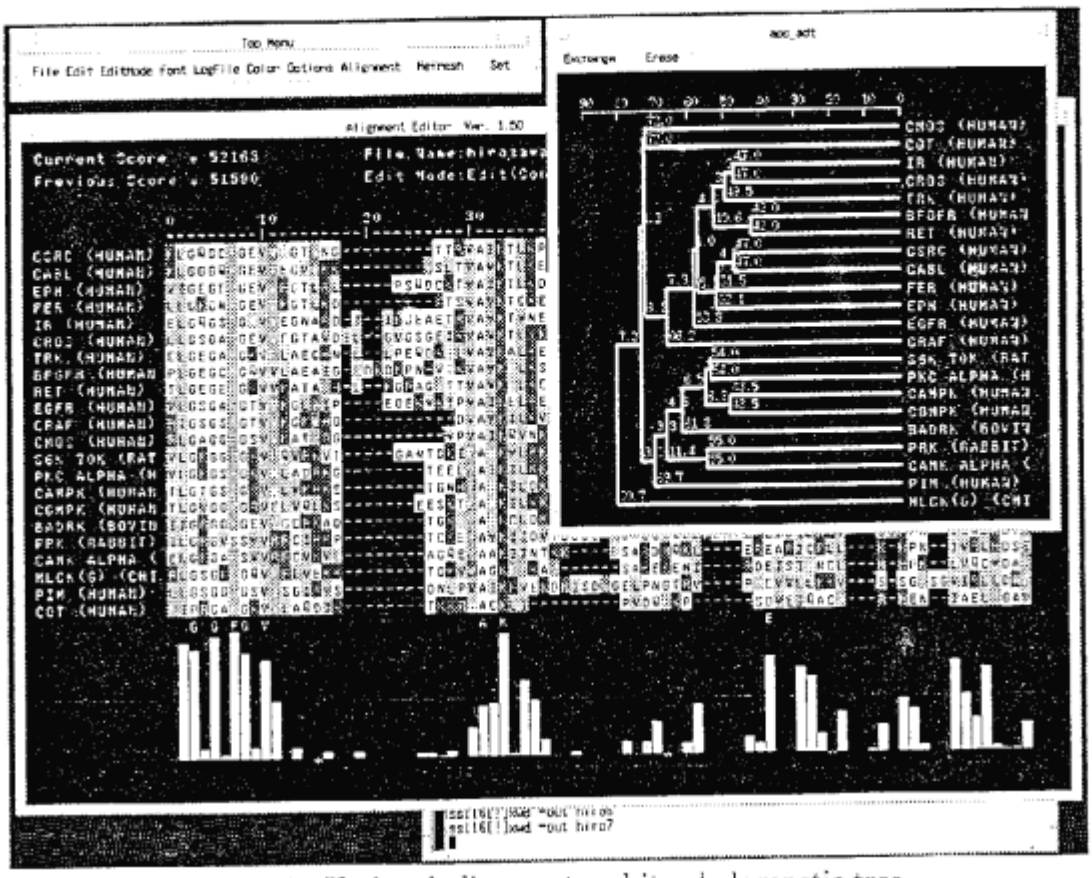

Figure 7: Updated alignment and its phylogenetic tree

[**Hirosawa** *et al.* **1993a**] Hirosawa, M. Hoshida, M., M, Ishikawa. Protein Multiple Sequence Alignment using Knowledge. *Proc. of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*,**Vol.1** pp 803-812,1993.

[**Hirosawa** *et al.* **1993b**] Hirosawa, M., Hoshida, M., Ishikawa, M. Iterative Multiple Alignment with Similarity Consideration. *Proc. of the Thirty-Sixth Convention of Information Processing Society of Japan*,**Vol.2** pp 289-290, 1993.

[**Sneath and Sokal 1973**] Sneath, P.H.A., Sokal, R.R. Numerical Taxonomy, W.H.Freeman and company, S.F. 1973.

[**Yasukawa** *et al.* 1992] H. Yasukawa *et al.* Objects, Properties, and Modules in Quixote. *Proc. of FGCS92, pp 89-112, 1992*