ICOT Technical Report: TR-0893

TR-0893

Multiple RNA-Sequence Alignment Considering Stem Regions

by

M. Ishikawa, T. Toya, Y. Totoki
& R. Tanaka (IMS)

October, 1994

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

Multiple RNA-Sequence Alignment Considering Stem Regions

M. Ishikawa[1]
ishikawa@icot.or.jp

T. Toya[1]
toya@icot.or.jp

Y. Totoki[1]
totoki@icot.or.jp

R. Tanaka[2]
ma-tanak@icot.or.jp

[1] Institute for New Generation Computer Technology (ICOT)
1-4-28-21F Mita, Minato-ku, Tokyo 108 JAPAN

[2] Information and Mathematical Science Laboratory, Inc.
2-43-1 Ikebukuro, Toshima-ku, Tokyo 171 JAPAN

Abstract

We have developed a multiple sequence alignment system which aligns RNA sequences while estimating their stem regions. The system consists of two parts: initial and stem aligners. The initial aligner roughly aligns given RNA sequences using a parallel iterative algorithm based on dynamic programming. The stem aligner refines the rough alignment using a parallel simulated annealing algorithm taking into account connected base pairs in stem regions. In testing with tRNA sequences, the system could generate alignments which identified well-known stem sets of clover shape. We have also developed a stem specifier which monitors such stem regions using a circular representation.

1 Introduction

Most RNA molecules in biological processes are mRNAs, which intermediate the genetic information between DNA and protein. Some RNA molecules such as tRNA and rRNA work as actively as protein. Each RNA molecule is composed of a chain of four kinds of nucleotides, represented by a sequence of code letters: A, U, G and C. Because the individual nucleotides are differentiated by their base parts, the code letters are abbreviation for the names of the bases.

[1] 石川幹人、戸谷智之、十時泰：（財）新世代コンピュータ技術開発機構，〒108 港区三田 1-4-28-21F

[2] 田中令子：（株）情報数理研究所，〒171 豊島区池袋 2-43-1

An RNA chain often forms several stems. In each stem, hydrogen-bond interactions connect two sets of consecutive nucleotides with each other. Interaction works between complementary base pairs: A-U and G-C. Weak interaction exists in G-U.

Stem structure prediction of an RNA molecule is an important issue in estimating its folded conformation and its function. Algorithms have been devised which predict stems in a single RNA sequence [1, 2, 3]. The prediction ratio can be improved by gathering some RNA sequences with a similar stem structure. It is known that covariance mutations often happen at stem regions (Figure 1a). This is because a complementary mutation can compensate for a mutation at a stem region. Consideration of these covariance mutations realizes a structural multiple RNA sequence alignment (Figure 1b). Such RNA-sequence alignment increases the prediction ratio of the stem structure, and also improves the accuracy of the phylogenetic analysis among RNA sequences. Some researchers have previously discussed simultaneous analyses of RNA structure prediction and sequence alignment [4, 5, 6], but they had not considered covariance mutations.

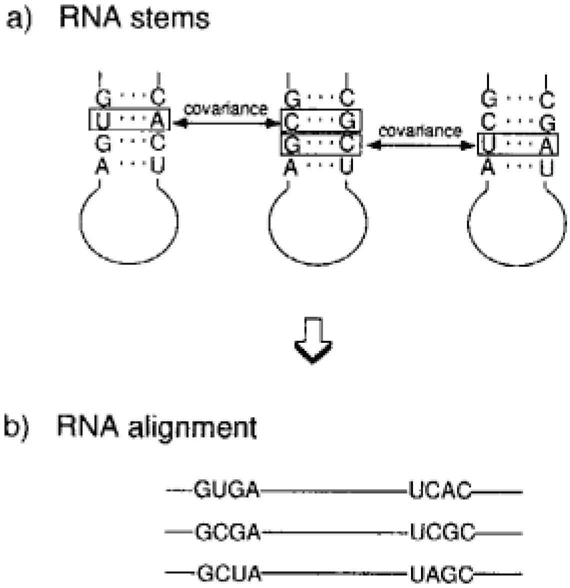

Figure 1: RNA stems and sequence alignment

In this paper, we propose a multiple sequence alignment system which aligns RNA sequences while also taking into account covariance mutations and estimating stem regions. The system was tested with tRNA sequences.

2 System and Methods

The multiple RNA-sequence alignment system consists of two parts: initial and stem aligners (Figure 2). Given RNA sequences are initially aligned by a tree-based parallel iterative algorithm which optimizes the sum-of-pair alignment score. The output of the initial aligner is merely a rough alignment, because the sum-of-pair score ignores stem interaction as too complex to implement in the framework of iterative improvement. Therefore, the rough alignment does not show complete alignment in terms of stem structures, but merely indicates some possible stem regions. The stem aligner refines this rough alignment with a temperature parallel simulated annealing algorithm which optimizes the score obtained by connected base pairs and covariance matches. This procedure gives us a proper RNA-sequence alignment, from which we can specify stem regions.

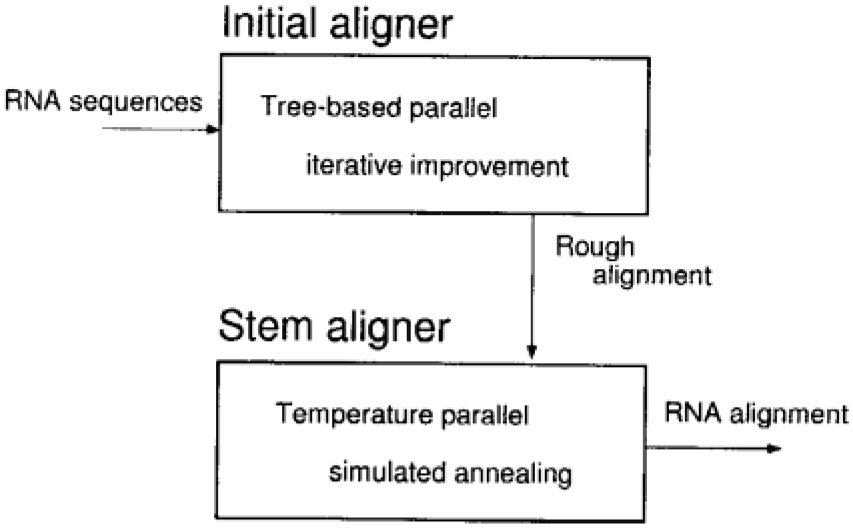

Figure 2: Multiple RNA-sequence alignment system

The system is written in the KL1 parallel logic programming language [7], and implemented on a distributed-memory parallel machine PIM/p with sixty-four processing elements [8]. Every eight processing elements in PIM/p form a cluster sharing 256 MB memory. In a cluster, processes waiting on a busy processing element are automatically allocated to idle processing elements.

2.1 Initial aligner

We chose a fast and efficient alignment algorithm, a tree-based iterative improvement algorithm [9], to obtain a rough RNA-sequence alignment. The algorithm uses 2-way dynamic programming (DP) in a group-to-group manner [10] to align two sub-alignments. In this algorithm, a tree is first drawn by the UPGMA method [11] based on previous similarity analysis by the 2-way DP on all sequence pairs. Sequences are merged based on the branching order of the tree by applying group-to-group DP, which optimizes the alignment between sequence groups (Figure 3). When the sub-alignments merge in the tree, every alignment of more than two sequences is refined by an iterative improvement algorithm.

Figure 4 shows the iterative algorithm, which refines alignment with a parallel best-first search. In this algorithm, N temporary aligned sequences are divided into a sequence and $N-1$ alignment sequences. The N different sets produced by the partitioning are then recombined in

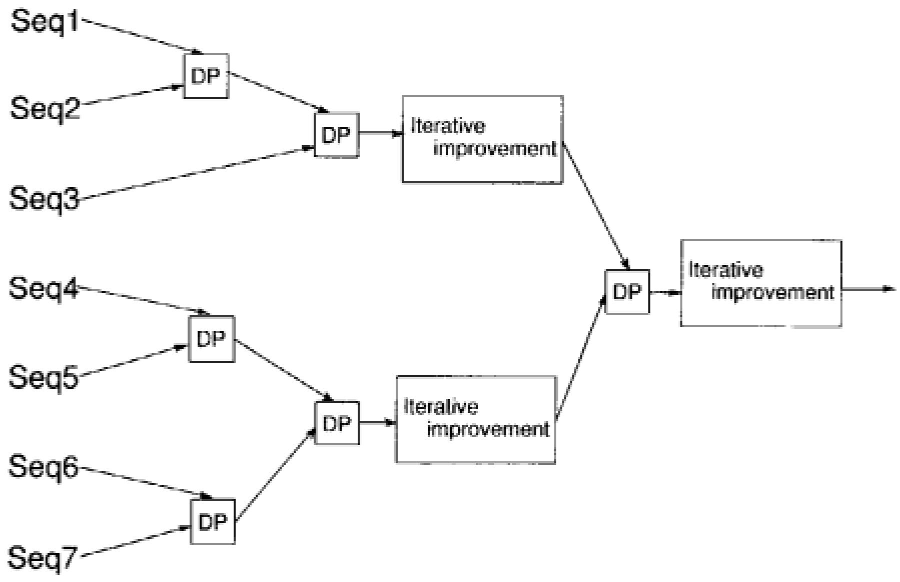

Figure 3: Tree-based iterative alignment algorithm

parallel by group-to-group DP. The results are compared and the best-score alignment becomes the starting point for the next iteration cycle. In this way, the cycle gradually improves the alignment of all the sequences. Iteration terminates when none of the N partitions can be improved any further. This parallel routine requires N processing elements.

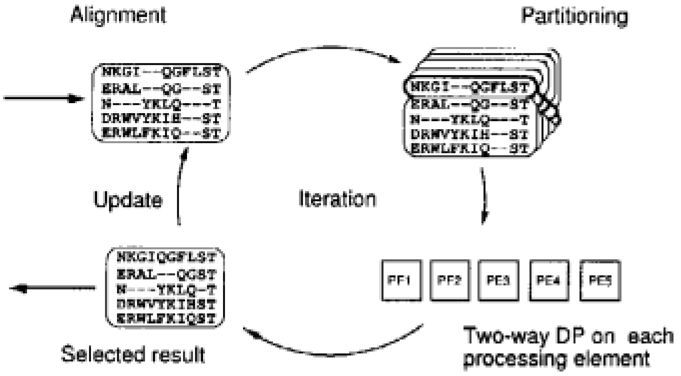

Figure 4: Parallel best-first iterative improvement

The score S to be optimized is the summation of all pairwise alignment scores (sum-of-pair score [12]). The pairwise alignment score is derived from match and mismatch scores and a linear gap relationship whose default values are shown below.

$$S = \sum_{i<j}^{seq.pair} \sum_{k}^{column} MatchScore(B_{ik}, B_{jk})$$

$$MatchScore(B_{ik}, B_{jk}) = \begin{cases} 4 & \text{if } Bs \text{ are the same base} \\ -2 & \text{if } Bs \text{ are different bases} \\ 0 & \text{if } Bs \text{ are gaps} \\ -7 & \text{if } Bs \text{ are the opening gap and a base} \\ -1 & \text{if } Bs \text{ are the extending gap and a base} \end{cases}$$

2.2 Stem aligner

We chose a temperature parallel simulated annealing algorithm [13] to align the stem regions of RNA sequences. This algorithm can deal with a sufficiently complex scoring function that remote interactions between RNA segments can be adequately taken into account. It can start working from a partially solved state of solution.

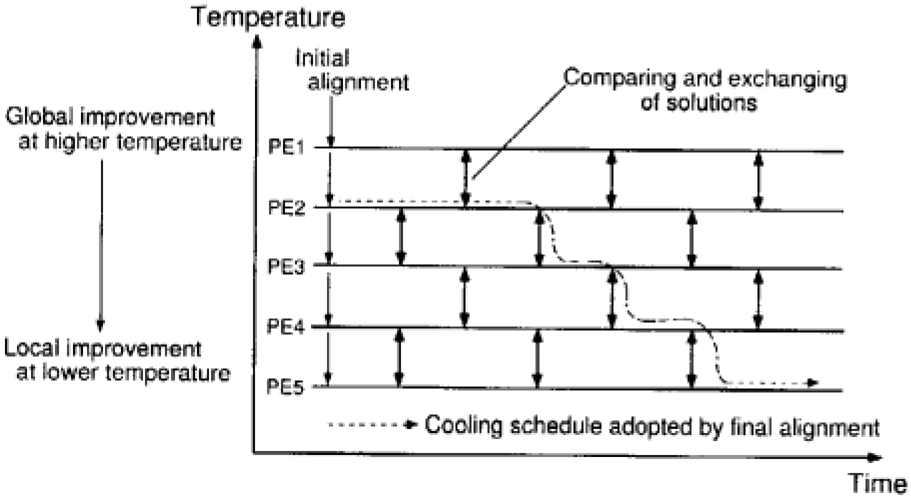
Temperature
Initial alignment
Comparing and exchanging of solutions
Global improvement at higher temperature
PE1
PE2
PE3
PE4
PE5
Local improvement at lower temperature
Cooling schedule adopted by final alignment
Time

Figure 5: Temperature parallel simulated annealing algorithm

The outline of the algorithm is as follows. Each processing element (PE) maintains one solution and performs improvement processes concurrently at a constant temperature that differs between PEs (Figure 5). Periodically, each pair of PEs with adjacent temperatures performs a probabilistic exchange of solutions. The algorithm can be stopped at any time after a large number of steps and a well-optimized solution can be found in the PE with the lowest temperature.

Each improvement step executed at temperature T is as follows. E is energy, the scoring function, to be minimized:

(i) Modify the current solution x_n randomly and get a candidate for the next solution x'_n.

(ii) Calculate the change in the scoring function: $\Delta E = E(x'_n) - E(x_n)$.

(iii) When $\Delta E \leq 0$, accept the candidate: $x_{n+1} = x'_n$. When $\Delta E > 0$, accept the candidate with the probability $p = \exp(-\Delta E/T)$; and reject it otherwise: $x_{n+1} = x_n$.

The probability of the exchange $p(T, E, T', E')$ between two solutions, one with energy E at temperature T and the other with energy E' at temperature T', is defined by the following:

$$p(T,E,T',E') = \begin{cases} 1 & \text{if } \Delta T \cdot \Delta E < 0 \\ \exp(-\frac{\Delta T \cdot \Delta E}{TT'}) & \text{otherwise} \end{cases}$$

$$\text{where} \qquad \Delta T = T - T', \qquad \Delta E = E - E'.$$

We formulated the alignment modification for each improvement step as follows. Focusing on one sequence in the alignment, the system selects a gap and a column (horizontal location) randomly in that sequence. The gap moves to the selected column. In some cases, a set of gaps move at the same time. The scoring function E is calculated using the following formula:

$$E = -\sum_{k+4<l}^{column} positive[Stem(k,l) + Stem(k+1,l-1) + Stem(k-1,l+1)]$$

$$Stem(k,l) = \sum_{i}^{seq} PairMatch(ik, il) + CovarianceMatch(k,l)$$

$$PairMatch(B_{ik}, B_{il}) = \begin{cases} 1 & \text{if } Bs \text{ are A\&U or G\&C} \\ 0 & \text{if } Bs \text{ are G\&U} \\ -2 & \text{otherwise} \end{cases}$$

$$\begin{aligned} CovarianceMatch(k,l) &= \#AU(k,l) + \#UA(k,l) + \#GC(k,l) + \#CG(k,l) \\ &\quad -max[\#AU(k,l), \#UA(k,l), \#GC(k,l), \#CG(k,l)] \end{aligned}$$

(ex. $\#AU(k,l)$: number of sequences each of which has A at column k and U at column l.)

3 Experimental Results

We sampled dozens of tRNA sequences from the Genbank database and generated several test data sets. Figure 6 shows an example of rough alignment done by the initial aligner. Twenty-two Leucine-tRNA sequences were roughly aligned in nine minutes. The first eight sequences are mitochondrial and the others are nucleic.

We also developed a stem specifier using OSF/Motif to identify some possible stem regions in an RNA sequence alignment. The specifier displays connectable pairs of base locations with lines using a circular representation (Figure 7). Each number surrounding the circle corresponds to a location (column) number of the alignment. A brighter line means that there are some covariance matches between the pair of base locations.

Figure 7 shows five possible stem regions with sixty percent consensus from the rough alignment. Only sets of at least three consecutive parallel lines are displayed. Three sets of brighter parallel lines correspond to real stem regions. The other two sets of parellel lines are pseudo signals. The three regions are also shown in Figure 6 with white-colored characters identifying paired bases.

After a twelve-hour annealing process (ninety-two thousand moves) beginning with the rough alignment, the stem aligner generated the final RNA alignment shown in Figure 8. The stem specifier displayed eight possible stem regions with sixty percent consensus for the final alignment shown in Figure 9. Five regions indicated by sets of parallel brighter lines form the complete set of stems giving a clover shape typical of tRNA structure. The white-colored sections in Figure 8 correspond to the five stem regions.

4 Conclusions

We have succeeded in developing a multiple RNA-sequence alignment system, in which a parallel iterative improvement using dynamic programming and a parallel simulated annealing are successively applied. In the former process, RNA sequences are rapidly and roughly aligned without considering RNA stem structure. The latter process refines the alignment in greater detail, taking into account possible base pairs and covariance mutations. As a result, some stem regions can be estimated in a final alignment.

In testing with tRNA sequences, the system could generate alignments which identified well-known stem sets of clover shape. It is important when specifying stem regions correctly to take into account the covariance mutations when aligning RNA sequences, because most stem regions include many such mutations.

In this experiment, the simulated annealing execution took a long time to solve an alignment problem despite using a parallel machine. We need to devise more effective heuristic moves to reduce the execution time, in order to solve more sophisticated alignment problems.

Acknowledgement

We would like to thank Profs M. Lynch and J. Conery at the University of Oregon for their discussion. This work was done in collaboration with the Genome Informatics Research Project of the Ministry of Education, Science and Culture of Japan.

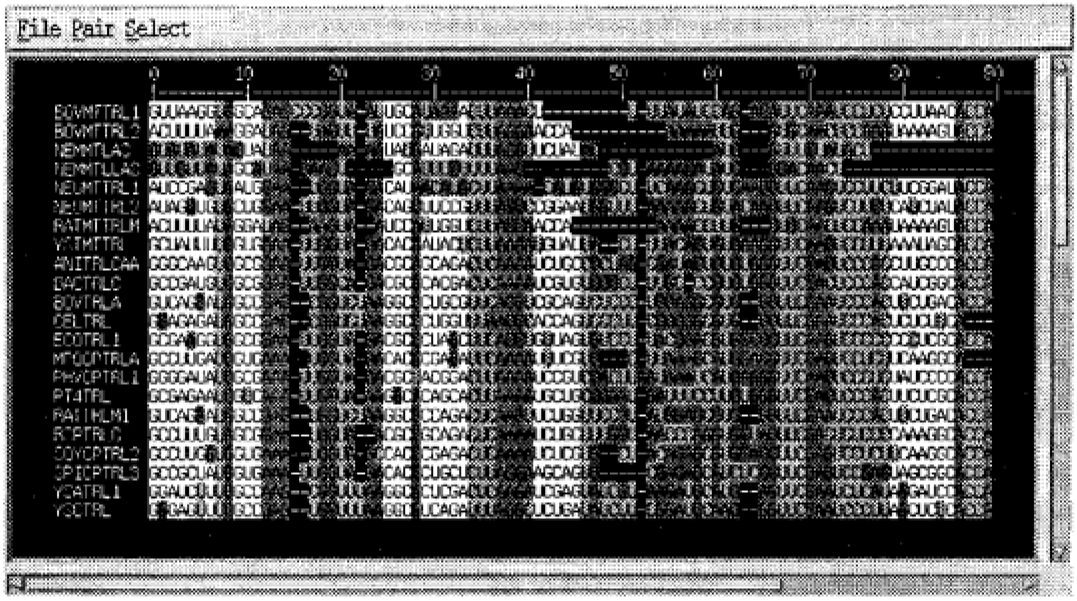

Figure 6: Example of a rough RNA alignment

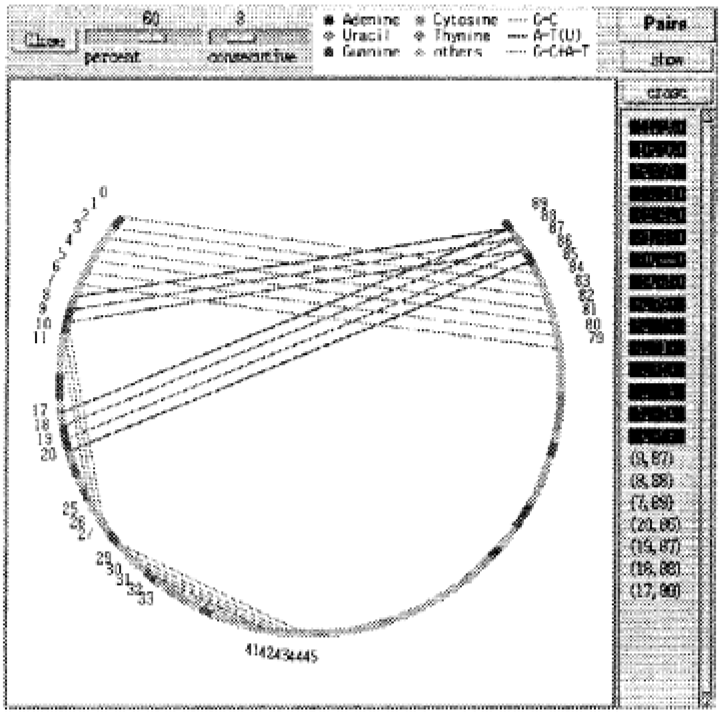

Figure 7: Stem specification with the rough alignment

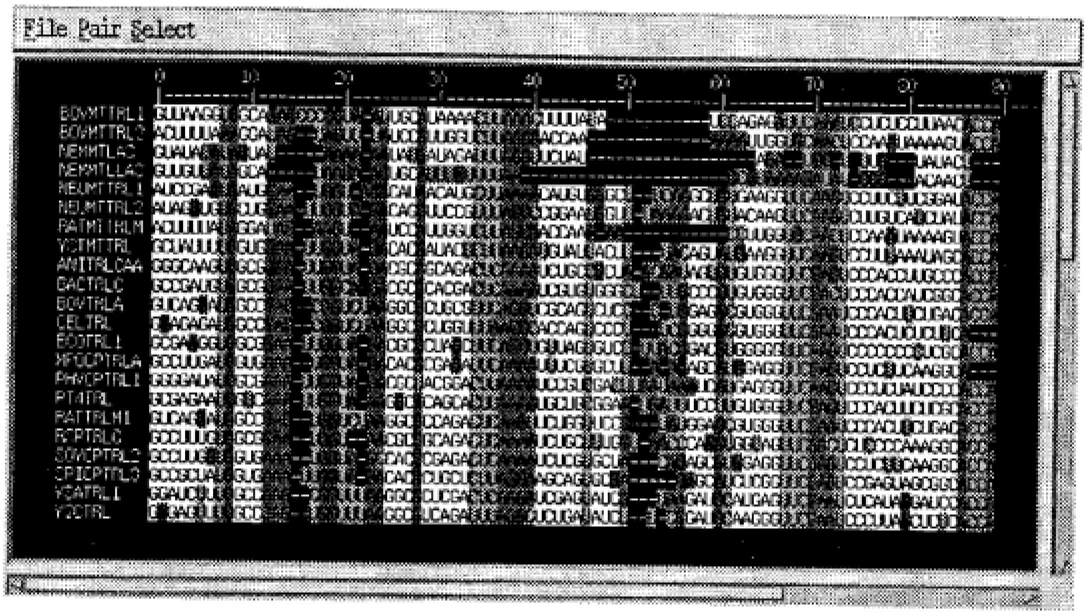

Figure 8: Example of a final RNA alignment

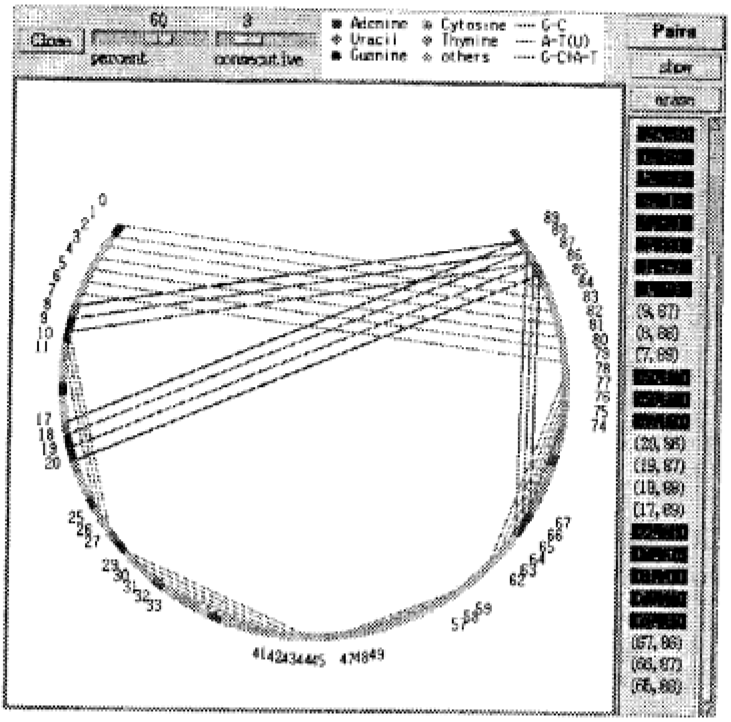

Figure 9: Stem specification with the final alignment

References

[1] J. M. Pipas, and J. E. McMahon, "Method for Predicting RNA Secondary Structure," *Proc. Natl. Acad. Sci.,* Vol.72, pp.2017-2021, 1975.

[2] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Research*, Vol.9, pp.133-148, 1981.

[3] K. Yamamoto and H. Yoshikura, "An improved algorithm for the prediction of optimum and suboptimum folding structures of long single-stranded RNA," *Comput. Applic. Biosci.*, Vol.3, pp.31-35, 1987.

[4] D. Sankoff, "Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems," *SIAM J. Appl. Math.*, Vol.45, pp.810-825, 1985.

[5] M. S. Waterman, "Mathematical Methods for DNA Sequences," CRC Press, pp.185-224, 1989.

[6] Y. Sakakibara *et al.,* "Stochastic context-free grammars in computational biology: application to modeling RNA," *Proc. Genome Informatics Workshop IV*, pp.36-45, 1993.

[7] K. Hirata *et al.,* "Parallel and distributed implementation of concurrent logic programming language KL1," *Proc. Fifth Generation Computer Systems '92,* pp.436-459, 1992.

[8] K. Kumon *et al.* "Architecture and implementation of PIM/p," *Proc. Fifth Generation Computer Systems '92,* pp.414-424, 1992.

[9] M. Ishikawa, Y. Totoki, R. Tanaka and M. Hirosawa, "Multiple sequence alignment editor featured by constraint-based parallel iterative aligner," *Proc. 3rd Int'l Conf. Bioinformatics and Genome Research,* (forth-coming).

[10] O. Gotoh, "Optimal Alignment between Groups of Sequences and its Application to Multiple Sequence Alignment," *Comput. Applic. Biosci.*, Vol.9, pp.361-370, 1993.

[11] P. H. A. Sneath and R. R. Sokal, "Numerical Taxonomy," Freeman and Company, 1973.

[12] S. F. Altschul and D. J. Lipman, "Trees, stars and multiple biological sequence alignment," *SIAM J. Appl. Math.,* Vol.49, pp.197-209, 1989.

[13] M. Ishikawa, T. Toya, M. Hoshida, K. Nitta, A. Ogiwara and M. Kanehisa, "Multiple sequence alignment by parallel simulated annealing," *Comput. Applic. Biosci.,* Vol.9, pp.267-273, 1993.