Multiple alignment using simulated annealing: branch point definition in human mRNA splicing

Alexander V.Lukashin+, Jacob Engelbrecht and Søren Brunak*
Department of Physical Chemistry and Computational Neural Network Center (CONNECT),
The Technical University of Denmark, DK-2800 Lyngby, Denmark

Received January 28, 1992; Revised and Accepted March 31, 1992

ABSTRACT

A method for the simultaneous alignment of a very large number of sequences using simulated annealing is presented. The total running time of the algorithm does not depend explicitly on the number of sequences treated. The method has been used for the simultaneous alignment of 1462 human intron sequences upstream of the intron-exon boundary. The consensus sequence of the aligned set together with a calculation of the Shannon information clearly shows that several sequence motives are conserved: (i) a previously undetected guanosine rich region, (ii) the branch point and (iii) the polypyrimidine tract. The nucleotide frequencies at each position of the branch point consensus sequence qualitatively reproduce the frequencies of the experimentally determined branch points.

INTRODUCTION

Multiple Alignment

The multiple sequence alignment is one of the most difficult problems in computational molecular biology. In solving the problem one needs to present the sequences in rows lying one above the other with a maximum number of matching symbols in each column. Relative shift of sequences as well as substitution and insertions of symbols are permitted at appropriate cost during the search for the best alignment. The result will show regions of sequence similarity of particular interest in the comparative analysis of protein or DNA molecules.

Since the best alignment should be established for all sequences simultaneously the exact solution of the problem using exhaustive search requires a number of operations (running time) at least of the order of the product of the sequence lengths. Such a time-consuming procedure cannot be carried out in practice for many areas of interest. Though a number of algorithms have been proposed (see, for example, (1,2,3,4,5,6,7,8)) none of them guarantees the exact solution for a difficult problem in a limited time. Hence, alternative approaches are desirable.

It is widely known that the simulated annealing method (9,10) provides a useful tool for obtaining the best heuristic solution for several hard problems in combinatorial optimization. The efficiency of the simulated annealing method cannot be evaluated *a priori*. It depends mainly on the behaviour of the cost function in a space defined by the system under consideration. In this paper we demonstrate that an approach based on the simulated annealing procedure results in a fast algorithm for the multiple sequence alignment problem.

Branch points in Human pre-mRNA Splicing

Most eukaryotic genes consist of alternating coding and non-coding regions, exons and introns. In the nucleus the introns are excised from the pre-mRNA and the mature mRNA is formed by concatenation of the exons. The splicing process recognizes sequences at the exon-intron and intron-exon borders (11) and is mediated by a group of small nuclear RNAs (snRNAs) complexed with protein as ribonucleoprotein particles (snRNPs)(12). The transition regions are known as donor and acceptor sites respectively. The excision of the introns involves a lariat formed between the initial part of the intron and a branch point nucleotide located upstream of the acceptor site (13). A basepairing interaction between the U2 snRNA and a few nucleotides surrounding the branch point nucleotide has been demonstrated genetically in mammalian systems(14,15). The branch point and the acceptor site are normally separated by a polypyrimidine tract. It has been shown that the lariat is formed with high specificity even when the acceptor site sequence is not present (16). A statistical analysis of branch points shows that the branch points themselves are characterized by a very weak consensus sequence (17,18). Branch points may therefore be defined by additional sequence characteristics. For this reason (as well as others) the general problem of prediction of exon-intron structure of human genes from the DNA sequence is difficult and cannot at present be performed without a large number of errors (19,20). The fact that the distance between the branch point and the polypyrimidine tract displays some non-

[*] To whom correspondence should be addressed

[+] Permanent address: Institute of Molecular Genetics, USSR Academy of Sciences, Moscow, 123182, Russia trivial variability motivates the use of a multiple alignment procedure to reveal sequence characteristics of the terminal part of human introns.

METHODS

The simulated annealing algorithm

Let us consider a set s of N numbered sequences of length l. The sequences in s have components (symbols) $s_i(k)$, where i and k denote sequence number and position, respectively. The symbols are drawn from a finite alphabet, *e.g.* the four nucleotides, a, t, g and c. Each sequence as a whole can be shifted relative to the rest, and gaps of different lengths can be inserted at any position. To describe the configuration of the set of sequences allowing for shifts and gaps a new set of sequences, S, is introduced. The elements in S are derived from s, but have the symbol • appearing at all positions covered by gaps and shifts. The new length L is equal to l plus the largest number of • symbols found in a sequence in S.

The problem of multiple alignment consists in the search for a configuration S that maximizes the matching score

$$M(S) = \sum_{k+1}^{L} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{Comp}(S_i(k),S_j(k)) - \sum_{i=1}^{N} P(S_i). \quad (1)$$

The function Comp(X,Y) compares two symbols X and Y. The value of Comp(X,Y) can be chosen to be 1 for a match and 0 for a mismatch, or can reflect the degree of similarity between X and Y. A gap penalty function, $P(S_i)$, of any form is also introduced.

One can reformulate the alignment problem by treating $-M(S)$ as the cost or 'energy' of the configuration S and then search for the global minimum of the system. In the framework of this treatment the many local minima of the energy make it difficult to obtain the exact solution. Simulated annealing (9,10) is an optimization scheme designed to cope with this problem. It is based on the standard Metropolis-Monte Carlo algorithm (21) which accepts not only changes in the configuration that lower the energy, but also changes that raise it. The probability of the latter event is chosen such that the system eventually obeys the Boltzmann distribution at a given temperature. The simulated annealing procedure is initialized at a sufficiently high temperature, at which a relatively large number of state changes are accepted. The temperature is then decreased according to a cooling schedule. If the cooling is slow enough for equilibrium to be established at each temperature, the global minimum is reached in the limit of zero-temperature. However this cannot not be guarantied in practice when the optimal cooling rate is unknown.

For the problem in question the simulated annealing algorithm can be reduced to the following procedure. Suppose that the current configuration S and the matching score $M(S)$ be known.

Table 1. Parameters and matching score for *E.coli* promoters

#	$1-\beta$	steps	M_r
0			0.2762
1	10^{-6}	5×10^{7}	0.3380
2	10^{-6}	2×10^{8}	0.3390
3	10^{-7}	3×10^{8}	0.3396

Row 0 shows M_r for the set of unaligned sequences s. Rows 1–3 show the resulting M_r values for the sets of aligned sequences. The results in rows 1–3 were obtained with different series of random numbers in the course of the simulated annealing procedure. The values of the other parameters were: $n_{\max}=5$, $m_{\max}=2$ and $l_{\max}=5$. The latter set of parameters gave the best alignment, and in the same time small variations in parameters did not change the results significantly.

Table 2. Consensus sequences of aligned *E.coli* promoters

#	consensus sequence
0	a-at--t-tTt-a-att----------t-tgttAaatT-----ca
1	Ag-TcA---TTGAcAt-tta-ca--att-tG-TAtAAT----acA
2	Ag-TcAg--TTGAcAt--ta-ca--att-tG-TAtAaT-a--acA
3	Ag-TcA---TTGAcAt-tta-c---att-tG-TAtAAT--a-acA

The parameters and values of matching score are shown in the corresponding rows in table 1. Row 0 shows consensus sequence for the unaligned set.

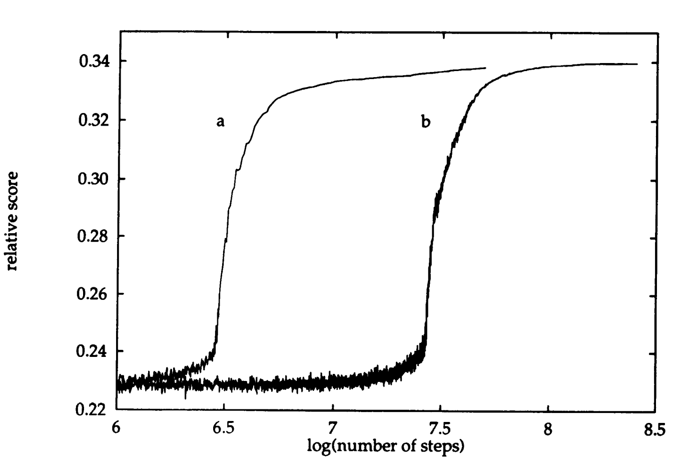

Figure 1. The dependence of the matching score on the number of steps in the course of simulated annealing. Curves *a* and *b* correspond to the β values shown in rows 1 and 3 in table 1 respectively.

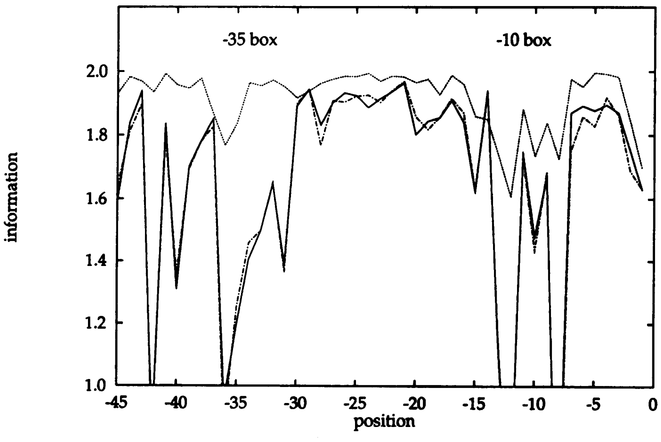

Figure 2. Shannon information as a function of the position for the set of *E.coli* promoters. The dotted curve represents the set of unaligned sequences (row 0 in table 1 and 2). The solid and dashed curves correspond to consensus sequences for the aligned sets shown in rows 1 and 3 in table 2. Position −1 corresponds to the very right nucleotide in table 2.

In one step of the recurrent procedure a new configuration S' is chosen (*e.g.* as described below) and the corresponding value $M(S')$ is calculated. The new configuration is accepted and used as the starting point for the next step if $M(S') \geq M(S)$. If $M(S') < M(S)$ then the probability that S' is accepted is equal to $\exp(-(M(S)-M(S'))/T)$, where the parameter T is the 'temperature'. We have used the standard exponential cooling schedule (9) used in simulated annealing: $T_{n+1} = \beta T_n$, where n is the step number and the value $1-\beta$ is positive and close to zero. The initial temperature T_0 is chosen such that almost all suggested changes are accepted.

To generate a new configuration, S', as a trial configuration in the annealing procedure, a sequence i is chosen at random, and substituted by the i'th sequence from s (the original set of sequences) with a randomly chosen shift and randomly chosen insertions. The parameters describing this choice are defined as follows: $p(n)$ is the probability that the length of the shift for the i'th sequence will be equal to n; $q(m)$ is the probability that m insertions will occur in the sequence; and $r(l)$ is the probability that an insertion will have the length l.

In the present work we have used the specific variant of the above approach. We have used the simplest definition of the Comp(X,Y) function:

$$\mathrm{Comp}(X,Y) = \begin{array}{l} 1 \text{ if } X = Y \text{ and } X \neq \bullet \\ 0 \text{ otherwise,} \end{array} \qquad (2)$$

and have not penalized gaps, *i.e.* $P(S_i) = 0$. In this case equation (1) reduces to

$$M(S) = \frac{1}{2} \sum_{k=1}^{L} \sum_{\alpha \neq \bullet} n_\alpha(k)(n_\alpha(k) - 1), \qquad (3)$$

where $n_\alpha(k)$ is the number of symbols α in the k'th column of the set of sequences S. Here α runs through the alphabet, in the nucleotide case a, t, g, c and •. To compare the results obtained for different sets of sequences it is convenient to use the *relative* matching score:

$$M_r(S) = \frac{M(S)}{lN(N-1)/2}, \qquad (4)$$

where $M_r(S)$ falls in the interval between 0 and 1.

We have used the simplest definition of the above probabilities $p(n)$, $q(m)$ and $r(l)$. The numbers n, m and l are uniformly distributed up to the values $n_{\max}$, $m_{\max}$ and $l_{\max}$. Beyond these values the probabilities are equal to zero. This implicitly includes a gap penalty since we have a maximum number of allowed gaps and maximum length of the gaps.

Table 3. Parameters and matching score for human intron sequences

#	k_a	w	M_r
0			0.3010
1		1	0.3243
2	−29	2	0.3218
3	−29	5	0.3188
4	−29	10	0.3189
5	−35	10	0.3143
6	−23	10	0.3126
7	−23/−29	10	0.3108

Row 0 shows M_r for the set of unaligned sequences s. Rows 1−7 show the resulting M_r values for the sets of aligned sequences. The value k_a indicate the position(s) of the adenosine nucleotide counted with weight w. The position of the terminal nucleotide in each intron, guanosine, is −1. The values of the other parameters were: $1-\beta = 10^{-6}$, number of steps 4×10^7, $n_{\max}=0$, $m_{\max}=5$ and $l_{\max}=5$. The latter set of probability parameters gave the best alignment, and in the same time small variations in parameters did not change the results significantly.

It is important to note that recalculation of the score M does not depend on the number of sequences, N. The recalculation of the set $\{ n_\alpha(k)\}$ only requires of the order of L operations. The recalculation of $\Delta M = M(S') - M(S)$ can be performed in a simple manner. Let the randomly selected sequence S_i be given and $\tilde{n}_\alpha(k)$ be the number of α symbols in the k'th column of S omitting the sequence S_i. Then

$$\Delta M = \sum_{k=1}^{L} (\tilde{n}_{\alpha'(k)}(k)) - (\tilde{n}_{\alpha'(k)}(k)), \qquad (5)$$

($\tilde{n}_\bullet(k)=0$), $\alpha(k)$ and $\alpha'(k)$ are the symbols found at the k-th position in the old and in the new version of S_i. The independence of N will be valid for any definition of the Comp(X,Y) function. Thus the total running time of the algorithm is in the order of KL where K is the number of steps required for the annealing procedure.

Quantification of the alignment

The Shannon information (22,23) was used and estimated by the following formula:

$$I(k) = -\sum_{\alpha \neq \bullet} \frac{n_\alpha(k)}{N} \log_2 \frac{n_\alpha(k)}{N} \qquad (6)$$

This value quantifies the similarity (ordering) in the k-th position of the set S. $I(k)$ falls in the range between 0.0 (one and the same letter in the all sequences) and 2.0 (for a random distribution with a four letter alphabet). The value $I(k)$ is useful for quantitative description of the similarity of a set of sequences in addition to the description in terms of consensus sequences.

Sequence data

The algorithm was tested on two sets of sequences. The first set consisted of *E.coli* promoter sequences (where the principal features are known), and was used to verify the algorithm. The

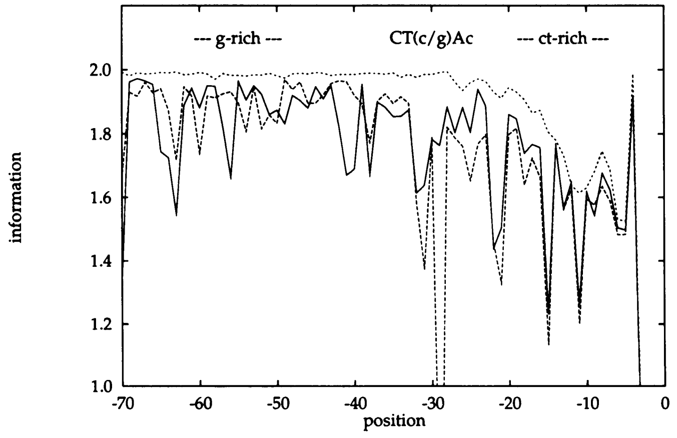

Figure 3. Shannon information as a function of the position for the set of human intron sequences. The dotted curve represents the set of unaligned sequences (row 0 in table 3 & 4). The solid and dashed curves correspond to consensus sequences for the aligned sets shown in rows 1 and 4 in table 4. Position −1 corresponds to the very right nucleotide in table 4.

Table 4. Consensus sequences for the aligned human introns

#	consensus sequence
0	------------g-----------------------------cc---ctttcccttttttccct-CAG
1	-ctG----ctGg-g--tG-c--ggc-gc---cTG-C--gctCTgaccct--CTgtctcTcTcTcTccct-CAG
2	-ctG----ctG--g-cagggcctg---ctctg--cc-gg--CTcAc--t--CT--ctcTcTcTctccct-CAG *
3	--tcag-gctgcaG-ca-gg-ctg-g-ctg----cct-g--CTcAc-----CT--c-cTcTcTctccct-CAG *
4	-cTGag---tG--Gcc-ggg-ctg---ct-----cct-g--CTcAc-----CT--ctcTcTcTctccct-CAG *
5	g--g-ctG---tg-g--ctg--ctg----c--g--cTcA--------CT-aCt-tctcTcTcTctctct-CAG *
6	-ctG-cag-----gga--tg--Ctgg---cTG-c-------c-ct---CTcAc----ctcTtcTcTCcct-CAG *
7	-ctgag---tG--g----ggcctg-g-c-------c-----CTcAc--t-Act----cTcTcTctcctt-CAG * *

The parameters and values of matching score are as shown in the corresponding rows of table 3. Row 0 shows the consensus sequence for the unaligned set. The * symbol indicates the value of k_a.

Table 5. Nucleotide frequencies (in %) for branch points

position	consensus	a	t	g	c	●
−3	C	8	15	13	56	5
−2	T	8	65	7	12	5
−1	c/g	12	11	32	39	4
0	A	94	1	0	1	1
1	c	14	19	16	40	9

The set of parameters used for alignment and total consensus sequence are shown in rows number 4 in tables 3 and 4 respectively. Position 0 in the table corresponds to position −29 in table 4 when the very right nucleotide is numbered as −1.

set of *E.coli* promoters used was taken as in (24), and consisted of 219 sequences 45 bp in length with known transcription start sites, that are recognized by the common form of *E.coli* RNA polymerase.

The set of sequences selected for analysis containing branch points was generated as follows: All complete human introns from GenBank release 66.0 were extracted, as indicated by the *intron* feature key in the feature table. The terminal part (70 bp) of each intron was used for alignment. Only introns terminated by the consensus dinucleotide AG were used. After removal of all duplicate sequences, 1462 sequences remained.

Consensus sequences

In the consensus sequences given below an upper case letter indicates that the frequency of occurrence of the corresponding nucleotide exceeds ½, while a lower case letter means that the frequency exceeds ⅓.

RESULTS AND DISCUSSION

Verification of the algorithm. *E.coli* promoters

Before using the algorithm for proper biological sequences it was verified on examples with artificially constructed sequences where the best alignment could be established by full enumeration (6 sequences 30 letters in length with a four letter alphabet). In all cases the global maximum was successfully found by the simulated annealing procedure (data not shown).

The efficiency of the method was assessed by alignment of the 219 *E.coli* promoter sequences. Table 1 gives the set of parameters used for several realizations of the alignment procedure and resulting values of matching score. Figure 1 shows how the relative score $M_r(S)$ changes during the alignment procedure. Figure 1 and the last column of table 1 show that the resulting matching score did not depend strongly on either the value of β or on the number of steps, if $1-\beta$ was 10^{-6} or less and the number of steps was 10^7 or greater. Consensus sequences for the aligned sets (rows 1−3) are given in table 2. The corresponding values for the Shannon information are shown in figure 2. The well established common features of *E.coli* promoters are clearly seen in both figure 2 and table 2. These are the −35 box with consensus **TTGAcA** and the −10 box with consensus **TAtAAT**. The nucleotide frequencies at each positions of these regions are very close to the frequencies obtained by other methods (25) (data not shown).

Human Introns

For alignment of the terminal part of the 1462 human introns we have used two approaches: The first one has been described in previous sections and was used without modification. The second approach incorporates prior knowledge of specific sequence features in the alignment process. It is known from experimental data that the branch point signal is normally positioned between 10 and 50 nucleotides upstream from the acceptor site and contains an adenosine residue at which the lariat is formed. In order to include this additional condition into the computation of the score we strengthen with weight w the contribution of the adenosine nucleotide at a fixed position, k_a. (The basic algorithm implies that w is equal to 1.0 for all nucleotides in all positions).

Table 3 shows several positions k_a and w values used for aligning. Row 1 gives the result obtained in the framework of the basic algorithm. Relative score values shown in rows 2−7 are recalculated for the aligned sets of sequences without additional weight and can directly be compared with the score in row 1. Corresponding consensus sequences and Shannon information values are presented in table 4 and figure 3.

Inspection of the number and size of gaps in the final alignments showed that most that these typical were smaller than the maximal values allowed. Very few sequences had gaps with the maximal size of five. The reproducibility of the results for different series of random numbers used in the simulated annealing procedure as well as for variation of the β parameter is as good as for the *E.coli* promoters.

It is important to emphasize the following fact with respect to the applicability of the simulated annealing approach. As can be observed in tables 3 and 4 and figure 3: when the local consensus sequence around strengthened position(s) becomes stronger the global alignment becomes worse, because potential alignment of sequence downstream the strengthened position in one intron with sequence upstream the strengthened position in others no longer can contribute to the score. This holds for all lengths of the maximal gap. If the algorithm is able to find the best heuristic solution for the problem under consideration any additional constraint should lower the global relative matching score. The relative score decreases monotonously towards saturation as the *w* value raises (rows 1−4 in table 3). Our calculations using different series of random numbers confirm the statistical significance of the dependence. The computing time was in the order of hours when the alignment algorithm was executed on a workstation with a computational power of 4 million floating point operations per second (4 Mflops).

Branch points

As can be observed from table 3 the alignment of human introns does not cause a very large increase in the relative matching score due to the variability of the set of sequences used. However, table 4 (row 1) and the solid curve in figure 3 clearly show the existence of three conserved regions: A guanosine rich region (from position −70 to −40), a region with consensus sequence **CTgac** (from position −32 to −28) and a pyrimidine rich region (from position −27 to −5). The **CAG** trinucleotide on the far right is the acceptor site. The site **CTgac** resembles the experimentally established branch point site (17), but it suffers from the very low frequency (46%) of the 'invariant' adenosine residue, which is the actual point of lariat formation (17). The possibility that the branch point can in some cases be positioned at a distance larger than 70 bp from the acceptor site (16) do not seem to be important for the set of sequences used. The low frequency of adenosine residues is caused by the strong competition from other conserved sequence parts in the region. When the sequences were aligned using a window of 5 bp the branch points did not emerge (data not shown). To overcome this problem we chose to strengthen the alignment of the experimentally determined 'invariant' adenosine residue as described in the previous section. In order to determine the optimal position k_a a range of different positions was tested. The results for several trials are shown in table 3 and 4 (rows 4−7). The maximal relative score was obtained when k_a was equal to −29. This is also the position where the conserved adenosine is found without additional weight (compare rows 1 and 4 in table 4).

The nucleotide frequencies at each position of the branch point consensus signal found is shown in table 5. Qualitatively the distribution of frequencies reproduces the distribution for experimentally determined branch points as reported in (17). 1% of the 1462 sequences had gap symbols at the branch point. This pathological feature may well be caused by the fact that some branch points are known to be positioned as far as 150 nucleotides upstream from the acceptor site, and thus not included within the 70 nucleotides used in the alignment. Analysis of the correlations among the symbols in the branch point signals indicated that they are rather weak a feature also noted in (17). The fact that the branch point consensus is very weak and actually can be found in many other places (17,18) indicates that some additional factors are necessary to biologically define the exact point of lariat formation. Our results strongly indicate that a branch point is defined by three conserved regions in concert: (i) a guanosine rich region, (ii) the branch point consensus and (iii) the polypyrimidine tract. RNAase protection experiments have shown that the part of the pre-mRNA which is protected by U2 snRNPs not only includes the branch point, but also a rather large region (35−42 nucleotides) adjacent to it(26). In comparison the region at the donor site that is protected by U1 snRNPs is much smaller covering around 15 nucleotides. Interactions between the branch structure and the human U2 snRNA sequence(27) may include base-pairing between the guanosine rich region and the cytidine rich region of the U2. 55% of the cytidine is located in the initial third of the U2 sequence, which also hold the binding site for core proteins(28).

The importance of the guanosine rich region in selecting the proper nucleotide as target for the lariat formation could be demonstrated by the case of the human growth hormone pre-mRNA. Two of the four introns contain very unusual branch points(29). In intron A the major branch acceptors are cytidine and uridine and not the usual adenosine. In intron D neither of the two adenosine branch points fits the consensus sequence. The complementarity to the standard U2 sequence is extremely poor. However both introns have regions with high g concentrations, intron A at positions −45 to −65 and intron B at −40 to −60. Both regions contain triple g's.

Alignment of the U2 RNA sequence with the consensus sequences for the guanosine rich region given in table 4 suggests that an extended U2 RNA-RNA interaction may include the U2 triple cytidine and the triple guanosine present in most of the intron sequences within 70 nucleotides from the acceptor site.

In conclusion, the set of results obtained in the present paper demonstrates the applicability of the simulated annealing algorithm for the search for the best alignment of a very large number of nucleotide sequences. It should, however, be emphasized that the running time of the simulated annealing algorithm depends on the energy landscape of a particular system. For instance, the applicability of the simulated annealing procedure for the multiple alignment of protein sequences is still an open question.

REFERENCES

1. Murata, M., Richardson, J.S. and Sussman, J.L. (1985) Proc. Natl. Acad. Sci. USA **82**, 3073−3077.
2. Waterman, M.S. (1986) Nucl. Acids Res. **14**, 9095−9102.
3. Corpet, F. (1988) Nucl. Acids Res. **16**, 10881−10890.
4. Bacon, D.J. and Anderson, W.F. (1990). In Methods in Enzymology, 'Computer Analysis of Protein and Nucleic Acid Sequences'. R.F. Doolittle Ed. **183**, 438−447.
5. Vihinen, M. (1990). In Methods in Enzymology, 'Computer Analysis of Protein and Nucleic Acid Sequences'. R.F. Doolittle Ed. **183**, 447−456.
6. Taylor, W.R. (1990). In Methods in Enzymology, 'Computer Analysis of Protein and Nucleic Acid Sequences'. R.F. Doolittle Ed. **183**, 456−474.
7. Argos, P., Vingron, M. and Vogt, G. (1991). Protein Engineering **4**, 375−383.
8. Vingron, M. and Argos, P. (1991) J. Mol. Biol. **218**, 33−43.
9. Kirkpatrick, S., Gelatt, C. D. and Vechi, M. P. (1983) Science **220**, 671−680.

10. Aart, E.H.L. and van Laarhoven, P.J.M. (1987) Simulated Annealing: A Review of the Theory and Applications (Kluwer Academic Publishers)
11. Green, M.R. (1986) Ann. Rev. Genet. **20**, 671–693.
12. Steitz, J.A., Wolin, S.L., Rinke, J., Petterson, I., Mount, S.M., Lerner, E.A., Hinterberger, M. and Gottlieb, E. (1983) Cold Spring Harbor Symp. Quant. Biol. **47**, 893–900.
13. Keller, E.B. and Noon, W.A. (1984) Proc. Natl. Acad. Sci. USA **84**, 7417–7420.
14. Zhuang, Y. and Weiner, A.M. (1989) Genes Dev. **3**, 1545–1552.
15. Wu, J. and manley, J.L. (1989) Genes Dev. **3**, 1553–1561.
16. Smith, C.W.J., Porro, E.B., Patton, J.G. and Nadal-Ginard, B. (1989) Nature **342**, 243–247.
17. Harris, N.L. and Senapathy, P. (1990) Nucl. Acids Res. **18**, 3015–3019.
18. Parker, R., Siliciano, P.G. and Guthrie, C. (1987) Cell **49** 229–239.
19. Senapathy, P., Shapiro, M.B. and Harris N.L. (1990) In Methods in Enzymology, 'Computer Analysis of Protein and Nucleic Acid Sequences'. R.F. Doolittle Ed. **183**, 252–281.
20. Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) J. Mol. Biol. **220**, 49–65.
21. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) J. Chem. Phys. **21**, 1087–1092.
22. Shannon, C.E. (1948) A Mathematical Theory of Communication. Bell System Tech. J. **27**, 379–423, 623–656.
23. Schneider, T.D., Stormo, G.D. and Gold, L. (1986) J. Mol. Biol. **188**, 415–431.
24. Lukashin, A.V., Anshelvich, V.V., Amirikian, B.R., Gragerov A.I. and Frank-Kamenetskii, M.D. (1989) J. Biomol. Struct. Dyn. **6**, 1123–1133.
25. Harley, C.D. and Reynolds, R. (1987) Nucl. Acids Res. **15**, 2343–2361.
26. Black, D.L., Chabot, B. and Steitz, J.A. (1985) Cell **42**, 737–750.
27. Van Arsdell, S.W. and Weiner, A.M. (1984) Mol. Cell Biol. **4**, 492–499.
28. Guthrie, C. (1991) Science **253**, 157–163.
29. Hartmuth, K. and Barta, A. (1988) Mol Cell. Biol. **8**, 2011–2020.