Lecture notes in computer science : multiple DNA sequence alignment using joint weight matrix

Shu, Jian Jun; Yong, Kian Yan; Chan, Weng Kong

2011

Shu, J. J., Yong, K. Y., & Chan, W. K. (2011). Lecture notes in computer science: Multiple DNA sequence alignment using joint weight matrix. International Conference on Computational Science and Its Applications (11th:2011:Santander).

<https://hdl.handle.net/10356/93996>

https://doi.org/10.1007/978-3-642-21931-3_51

© 2011 Springer-Verlag Berlin Heidelberg. This is the author created version of a work that has been peer reviewed and accepted for publication by Computational Science and Its Applications - ICCSA 2011, Springer-Verlag Berlin Heidelberg. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [DOI: http://dx.doi.org/10.1007/978-3-642-21931-3_51].

Downloaded on 26 Apr 2023 22:23:08 SGT

Lecture Notes in Computer Science: Multiple DNA Sequence Alignment Using Joint Weight Matrix

Jian-Jun Shu^{*}, Kian Yan Yong, and Weng Kong Chan

School of Mechanical & Aerospace Engineering, Nanyang Technological University,
50 Nanyang Avenue, Singapore 639798

Abstract. The way for performing multiple sequence alignment is based on the criterion of the maximum scored information content computed from a weight matrix, but it is possible to have two or more alignments to have the same highest score leading to ambiguities in selecting the best alignment. This paper addresses this issue by introducing the concept of joint weight matrix to eliminate the randomness in selecting the best alignment of multiple sequences. Alignments with equal scores are iteratively re-scored with joint weight matrix of increasing level (nucleotide pairs, triplets and so on) until one single best alignment is eventually found. This method can be easily implemented to algorithms using weight matrix for scoring such as those based on the widely used Gibbs sampling method.

Keywords: multiple sequence alignment, joint weight matrix.

1 Introduction

In the search for DNA regulatory elements such as binding sites, promoter, donor sites, TATA box and genes, the multiple sequences containing these elements have to be aligned against one another. These elements are highly but not absolutely conserved and a weight matrix is used to represent and score the multiple sequences [1]. However, the current motif discovery algorithms based on the weight matrix technique of scoring an alignment of multiple sequences in terms of information content is not without its limitations [2]. From the analysis of these algorithms [2], the highest performance coefficient on the binding site level of search is only 30.2% using Motif Sampler [3], an algorithm modified from the widely adopted Gibbs Sampling method [4]. This may be a result of randomness in selecting the best alignment from cases whereby there are multiple peaks. Hence there are rooms for improvement, which is evident from many recent methods [5-11].

In this paper, a method of removing the randomness in selection is proposed. Randomness in selection occurs when there is more than one choice of alignments with the highest information content [12]. If one peak is randomly selected, the

^{*} Corresponding author. mjjshu@ntu.edu.sg

accuracy of multiple sequence alignment is compromised. This may be the reason why methods that are based on applied information theory cannot achieve a much higher sensitivity, specificity and performance. For example, by randomly selecting two peaks of similar information content, there is a 50% chance of selecting the wrong peak.

A simple method is proposed to overcome this problem, through the use of joint weight matrix (JWM). The concept of the higher-level JWM comes from the consideration of the interdependency between neighboring bases [13-15]. This concept has been applied in transcription factor motif-finding algorithms [16] and the result is more accurate. In this paper, a JWM is employed to eliminate the randomness of peak selection and to provide the best alignment. Its flexibility means that a higher-level JWM can be used to work with cases with multiple peaks. The higher the level of JWM used, the lesser the number of peaks appeared until one is eventually singled out. In this paper, The JWM has been shown to reduce successfully the number of peaks in the alignment of multiple sequences [12].

2 Systems and Methods

JWM for Removing Ambiguity The concept of JWM is presented here to demonstrate how two or more ambiguous selections can be reduced. Two sequences are used in this example. The longer one represents the DNA sequence and the shorter one represents a motif sequence, which is aligned to the former. The motif is assumed to be a perfect weight matrix with 100% base weightage at each position. The score is then either a 100% match or mismatch at each position for simplicity of demonstration.

Since the sequence is 7 bp (base pair) long and the motif is 4 bp, the total number of possible shift positions without introducing gaps is $7 - 4 + 1 = 4$ in Figure 1. Table 1 shows the sequence alignment.

Table 1. Tabulated score for quality of match between DNA and motif based on base by base

Position	1	2	3	4	5	6	7
DNA	A	T	T	G	T	T	C
Motif		T	T	A	G		
Score		2	2	1	0		

The score for the four possible alignments presents an ambiguous choice between positions 1 and 2, which are possible alignments with the highest score of 2. Since there are more than one peak or alignment, the second-level JWM is used to score the alignment. Table 2 shows the result using the second-level comparison.

The result clearly shows that between positions 1 and 2, the better match for the motif with the DNA is at position 2 with a matching score of one compared with zero at position 1.

Table 2. Tabulated score of match between DNA and motif based on two bases

Position	1	2	3	4	5	6	7
DNA	A	T	T	G	T	T	C
Motif		T	T	A	G		
Score	0	1	0	0

Fig. 1. Scoring a DNA sequence with a motif

3 Algorithm

Steps to Implement JWM in Algorithm Malign. Here it is shown how the JWM can be implemented in a sequence alignment tool, Malign [12] to remove the randomness of selection that may surface during the alignment process. The following are the additional steps added using JWM:

Step 1: Determine a weight matrix

$$w(b, i) = \frac{n(b, i)}{\sum_{b \in \{A, T, G, C\}} n(b, i)}, \quad (1)$$

where $n(b, i)$ is the number of each base $b \in \{A, T, G, C\}$ at each position i .

Step 2: Calculate the second-level JWM

$$w_2(b_1 b_2, i) = w(b_1, i) w(b_2, i + 1). \quad (2)$$

For the second-level JWM, the number of possible combinations of the four bases are 4^2 or 16. Hence the JWM is a matrix size of 16 by length of the window.

Step 3: From the weight matrix, the uncertainty of each combination of bases is

$$Hs(i) = - \sum_{j=1}^m \sum_{b_j \in \{A,T,G,C\}} w_m \log_2 w_m. \quad (3)$$

For the second-level JWM, m is a value of 2.

Step 4: The information content for each base is then

$$R(i) = 2^m - Hs(i) - e[n(i)], \quad (4)$$

where $e[n(i)]$ is a small sample correction for $Hs(i)$ [17].

Step 5: The score for one shifting position is then

$$R_{\text{shift}}(\text{sp}) = \sum_i R(i). \quad (5)$$

The shift position (sp) ranges from negative shifting parameter to positive shifting parameter.

Step 6: Shift the JWM as predetermined to get the alignment score plot of information content versus shifting position. From the alignment score plot, choose the peak, which is highest among the ambiguous choice of the previous set of peaks generated.

Step 7: If there are still ambiguity after using the second-level JWM, a higher-level JWM (three or higher) should be calculated

$$w_m(b_1 b_2 \cdots b_m, i) = w(b_1, i) w(b_2, i+1) \cdots w(b_m, i+m-1). \quad (6)$$

Repeat the Steps 3 to 6 using the higher-level JWM in (6) when there is ambiguity in peak selection if using any lower-level JWM.

4 Implementation

An Example of Removing Ambiguous Peak Using JWM. An example of how the JWM is used to eliminate or reduce ambiguity is shown using data from 16 randomly generated sequences of 12 bp (refer to Tables 3 and 4) that bind to OxyR [12]. For illustration purpose, the centre 7th base is taken to be the start site of transcription, labeled as position 0. The alignment score is obtained by using a window of 11 bases from -6 to $+4$ with a shifting position set to the range of -5 to $+5$ with respect to the start site. The sequences are shifted one base at a time and the new alignment score is recalculated based on the simplified sequence logo [18] in Figure 2(a).

The window and shifting parameter are selected such that an ambiguous choice of more than one peak is resolved. By shifting one of the sequences from -5 to $+5$, the alignment score based on the window from -6 to $+4$ show two peaks at shift positions -3 and $+3$ in Figure 2(b). From the simplified sequence logo, the information content prior to shifting of any sequence

Table 3. Weight matrix of the sixteen OxyR binding sequences from base positions -6 to +5

	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
1	A	C	A	C	C	G	A	C	T	T	G	A
2	C	A	C	A	A	G	T	C	G	G	T	C
3	T	T	A	T	C	G	A	T	C	C	G	T
4	T	G	C	G	G	A	T	C	G	A	T	T
5	T	T	A	A	C	A	A	T	A	G	G	T
6	G	C	C	C	T	A	T	T	G	T	T	G
7	C	G	A	T	A	A	T	A	G	G	C	C
8	T	T	G	C	C	T	A	T	T	A	T	T
9	T	A	C	A	T	T	A	T	C	C	A	T
10	T	A	T	G	G	A	T	A	A	T	G	T
11	A	T	T	A	T	T	G	T	A	A	C	A
12	C	T	G	T	T	A	C	A	A	T	A	C
13	T	T	T	C	C	C	A	G	A	G	T	T
14	G	A	A	C	T	C	T	G	G	G	A	G
15	G	A	G	A	T	C	G	C	T	C	T	G
16	T	T	A	G	A	G	C	G	A	T	C	T
A	2	5	6	5	3	6	6	3	6	3	3	2
T	8	7	3	3	6	3	6	6	3	5	6	8
G	3	2	3	3	2	4	2	3	5	5	4	3
C	3	2	4	5	5	3	2	4	2	3	3	3
% A	0.13	0.31	0.38	0.31	0.19	0.38	0.38	0.19	0.38	0.19	0.19	0.13
% T	0.50	0.44	0.19	0.19	0.38	0.19	0.38	0.38	0.19	0.31	0.38	0.50
% G	0.19	0.13	0.19	0.19	0.13	0.25	0.13	0.19	0.31	0.31	0.25	0.19
% C	0.19	0.13	0.25	0.31	0.31	0.19	0.13	0.25	0.13	0.19	0.19	0.19

Table 4. The second-level JWM for the sixteen OxyR binding sequences

	-6 -5	-4 -3	-2 -1	0 +1	+2 +3	+4 +5
AA	0.0391	0.1172	0.0703	0.0703	0.0703	0.0234
AT	0.0547	0.0703	0.0352	0.1406	0.1172	0.0938
AG	0.0156	0.0703	0.0469	0.0703	0.1172	0.0352
AC	0.0156	0.1172	0.0352	0.0938	0.0703	0.0352
TA	0.1563	0.0586	0.1406	0.0703	0.0352	0.0469
TT	0.2188	0.0352	0.0703	0.1406	0.0586	0.1875
TG	0.0625	0.0352	0.0938	0.0703	0.0586	0.0703
TC	0.0625	0.0586	0.0703	0.0938	0.0352	0.0703
GA	0.0586	0.0586	0.0469	0.0234	0.0586	0.0313
GT	0.0820	0.0352	0.0234	0.0469	0.0977	0.1250
GG	0.0234	0.0352	0.0313	0.0234	0.0977	0.0469
GC	0.0234	0.0586	0.0234	0.0313	0.0586	0.0469
CA	0.0586	0.0781	0.1172	0.0234	0.0234	0.0234
CT	0.0820	0.0469	0.0586	0.0469	0.0391	0.0938
CG	0.0234	0.0469	0.0781	0.0234	0.0391	0.0352
CC	0.0234	0.0781	0.0586	0.0313	0.0234	0.0352

Fig. 2. Graphical results of OxyR binding sites (a) Simplified sequence logo. (b) Information content of each shift positions of a randomly selected sequence using conventional weight matrix. (c) Simplified sequence logo (JWM). (d) The same scanning process in (b) using JWM.

$$R_{\text{shift}}(0) = 0.2194 + 0.2038 + 0.0637 + 0.0456 + 0.1171 + 0.0637 \\ + 0.1887 + 0.0637 = 1.1922 \text{ bits.}$$

One of the sequences is randomly selected and shifted about its position. In this case, sequence number 16 is shifted from its position by a shifting parameter of 5, *i.e.*, from -5 to $+5$. The weight matrix is calculated for each new position and by the end of the shift, a set of $R_{\text{shift}}(\text{sp})$ is obtained. The amount of shift required for the 16 sequences to produce a highest value of R_{shift} or the peak value can be found on the R_{shift} plot in Figure 2(b), which two peaks locate at shift positions -3 and $+3$. The situation is that of an ambiguous one and requires of higher-level search using JWM. The weight matrix is replaced by the second-level JWM in the new search. The new R_{shift} plot based on the higher-level JWM is shown in Figure 2(d).

The new alignment score using JWM clearly shows that the shift of 3 positions to the right has higher information content, compared with the shift of -3 (or 3 positions to the left). Hence, the better alignment is the shift of the active sequence rightward by 3 bases. Instead of randomly selecting one of the peaks, the criteria in which one can be used to select the peaks with higher information content is proposed.

5 Discussion and Conclusions

In the selection of the best alignment among multiple sequences using weight matrix, it is assumed that the probability of each base is independent of its

neighboring one. By comparing two bases at one time, the probability of the next base is affected by what appears before it. In fact, there are 16 probabilities of a pair of bases compared with just 4 probabilities if only one base is considered. This increases the depth of search, and the result reduces the number of peaks.

Under the Implementation section, it is shown how the second-level JWM can identify the highest peak when a conventional weight matrix could not. This reduces the error that may occur when "conflicts are resolved" by making a "pseudo-random choice" [12].

The higher-level JWM can be used depending on the level of accuracy required. For example, the second-level JWM may be able to reduce the number of peaks from 5 to 3. The randomness is reduced when one is choosing the best peak from 3 instead of 5 possible sites. However, if the application requires a level of match to be of greater accuracy, the user is able to proceed and use a higher-level JWM. The higher-level JWM can further filter out more peaks till only one obvious choice is left. Although the higher-level JWM may require more computation time and additional scan, this may be compensated by the faster convergence of results as a better alignment is selected early in the iterations. This is true especially for cases whereby a large number of iterations are required before a satisfactory convergence can be found.

The JWM can be used to improve applications using conventional weight matrix system in bioinformatics. Besides aligning DNA sequences, the JWM can also be implemented in the alignment of protein sequences.

References

1. Stormo, G.D., Hartzell, G.W.: Identifying Protein-Binding Sites from Unaligned DNA Fragments. *Proceedings of the National Academy of Sciences of the United States of America* 86(4), 1183–1187 (1989)
2. Hu, J.J., Li, B., Kihara, D.: Limitations and Potentials of Current Motif Discovery Algorithms. *Nucleic Acids Research* 33(15), 4899–4913 (2005)
3. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., Moreau, Y.: A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. *Journal of Computational Biology* 9(2), 447–464 (2002)
4. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C.: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 262(5131), 208–214 (1993)
5. Liu, Y.Y., Liu, X.S., Wei, L.P., Altman, R.B., Batzoglou, S.: Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics. *Genome Research* 14(3), 451–458 (2004)
6. Shu, J.-J., Ou, L.S.: Pairwise Alignment of the DNA Sequence Using Hypercomplex Number Representation. *Bulletin of Mathematical Biology* 66(5), 1423–1438 (2004)
7. Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., Mironov, A.A., Makeev, V.J.: A Gibbs Sampler for Identification of Symmetrically Structured, Spaced DNA Motifs with Improved Estimation of the Signal Length. *Bioinformatics* 21(10), 2240–2245 (2005)

8. Kuo, L., Yang, T.Y.: An Improved Collapsed Gibbs Sampler for Dirichlet Process Mixing Models. *Computational Statistics & Data Analysis* 50(3), 659–674 (2006)
9. Cattani, C.: Fractals and Hidden Symmetries in DNA. *Mathematical Problems in Engineering* 507056, 1–31 (2010)
10. Li, M.: Fractal Time Series-A Tutorial Review. *Mathematical Problems in Engineering*, 157264, 1–26 (2010)
11. Shu, J.-J., Li, Y.: Hypercomplex Cross-correlation of DNA Sequences. *Journal of Biological Systems* 18(4), 711–725 (2010)
12. Schneider, T.D., Mastronarde, D.N.: Fast Multiple Alignment of Ungapped DNA Sequences Using Information Theory and a Relaxation Method. *Discrete Applied Mathematics* 71(1-3), 259–268 (1996)
13. Benos, P.V., Bulyk, M.L., Stormo, G.D.: Additivity in Protein-DNA Interactions: How Good an Approximation is It? *Nucleic Acids Research* 30(20), 4442–4451 (2002)
14. Eden, E., Brunak, S.: Analysis and Recognition of 5' UTR Intron Splice Sites in Human Pre-mRNA. *Nucleic Acids Research* 32(3), 1131–1142 (2004)
15. Osada, R., Zaslavsky, E., Singh, M.: Comparative Analysis of Methods for Representing and Searching for Transcription Factor Binding Sites. *Bioinformatics* 20(18), 3516–3525 (2004)
16. Zhou, Q., Liu, J.S.: Modeling within-Motif Dependency for Transcription Factor Binding Site Predictions. *Bioinformatics* 20(6), 909–916 (2004)
17. Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A.: Information Content of Binding Sites on Nucleotide Sequences. *Journal of Molecular Biology* 188(3), 415–431 (1986)
18. Schneider, T.D., Stephens, R.M.: Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Research* 18(20), 6097–6100 (1990)