A Parameterized Algorithm for Predicting Transcription Factor Binding Sites

Yinglei Song[1], Changbao Wang[1], and Junfeng Qu[2]

[1] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China
syinglei2013@163.com, wangchangbao@126.com
[2] Department of Information Technology, Clayton State University, Morrow, GA 30260, USA
jqu@clayton.edu

Abstract. In this paper, we study the Transcription Factor Binding Sites (TFBS) prediction problem in bioinformatics. We develop a novel parameterized approach that can efficiently explore the space of all possible locations of TFBSs in a set of homologous sequences with high accuracy. The exploration is performed by an ensemble of a few Hidden Markov Models (HMM), where the size of the ensemble is the parameter of the algorithm. The ensemble is initially constructed through the local alignments between two sequences that have the lowest similarity value in the sequence set, the parameters of each HMM in the ensemble are revised when the remaining sequences in the set are scanned through by it one by one. A list of possible TFBSs are generated when all sequences in the set have been processed by the ensemble. Testing results showed that this approach can accurately handle the cases where a single sequence may contain multiple binding sites and thus has advantages over most of the existing approaches when a sequence may contain multiple binding sites.

Keywords: parameterized algorithm, Hidden Markov Model (HMM), Transcription factor binding site, dynamic programming.

1 Introduction

Transcription Factor Binding Sites (TFBS) are subsequences found in the upstream region of genes in DNA genomes. A transcription factor, which is a specialized protein molecule, may bind to the nucleotides in the subsequences and thus may affect some relevant biological processes. Research in molecular biology has revealed that transcription factor binding sites are important for many biological processes, including gene expression and regulation. An accurate identification of TFBSs is thus important for understanding the biological mechanism of gene expression and regulation. Classical experimental methods are time consuming and expensive [6,7]. Recently, a few new experimental methods such as ChIP-chip and ChIP-seq have been developed for TFBS identification [17]. Although the throughput of these methods is high, processing the large amount of complex data generated by

D.-S. Huang et al. (Eds.): ICIC 2014, LNBI 8590, pp. 339–350, 2014.

these methods remains a challenging task [17]. Computational methods that can accurately and efficiently identify TFBSs from homologous sequences are thus still convenient and important alternative approaches to rapid identification of TFBSs.

Since TFBSs for the same transcription factor have similar sequence content in homologous sequences, the most often used computational approaches make the prediction by analyzing a set of homologous sequences and identifying subsequences that are similar in content. The locations of a TFBS may vary in different homologous sequences. To determine the location of a TFBS in each sequence, we need to evaluate all possible starting locations among all sequences to find the optimal solution. The total number of combinations of subsequences that need to be examined is exponential and exhaustively enumerating all of them is obviously impractical when the number or the lengths of the sequences are large. To avoid exhaustive search, a large number of heuristics have been developed to reduce the size of the search space, such as Gibbs sampling based approaches AlignACE [19], BioProspector [16], Gibbs Motif sampler [15], expectation maximization based models [1, 2], greedy approaches such as Consensus [8], and genetic algorithm based approaches such as FMGA [14] and MDGA [4].

Of all these approaches and software tools, Gibbs Motif sampler is a tool based on a stochastic approach. It computes the binding site locations by Gibbs sampling [15, 16, 19]. Consensus uses a greedy algorithm to align functionally related sequences and applies the algorithm to identify the binding sites for the E. coli CRP protein [8]. MEME+ [2] uses Expectation Maximization technique to fit a two component mixture model and the model is then used to find TFBSs. MEME+ achieves higher accuracy than its earlier version MEME [1]. However, the prediction accuracy is still not satisfactory.

Genetic algorithm (GA) simulates the Darwin evolutionary process to find an approximate optimal solution for an optimization problem. GA based approaches have been successfully used to solve the TFBS predicting problem, such as FMGA [14] and MDGA [4]. FMGA was declared to have better performance than Gibbs Motif Sampler [15] in terms of both prediction accuracy and computation efficiency. MDGA [4] is another program that uses genetic algorithms to predict TFBSs in homologous sequences. During the evolutionary process, MDGA uses information content to evaluate each individual in the population. MDGA is able to achieve higher prediction accuracy than Gibbs sampling based approaches while using a less amount of computation time.

So far, most of the existing approaches use heuristics to reduce the size of the search space. However, heuristics employed by these approaches may also adversely affect the prediction accuracy. For example, GA based prediction tools cannot guarantee the prediction results are the same for different runs of the program. A well defined strategy that can be used to efficiently explore the search space and can generate `deterministic` and highly accurate prediction results is thus necessary to further improve the performance of prediction tools.

Recent work has shown that an ensemble of HMMs can be effectively used to improve the accuracy of protein sequence alignment [21]. In this paper, we develop a new parameterized algorithm that can predict the locations of TFBSs with an ensemble of Hidden Markov Models (HMMs), where the size of the ensemble is the parameter. The approach uses an ensemble of profile HMMs to generate a list of

positions that are likely to be the starting positions of the TFBSs. As the first step, we construct the ensemble from the local alignment of two sequences. The ensemble consists of HMMs that represent the local alignments with the most significant alignment scores. We then align each profile HMM in the ensemble to each sequence in the data set, the parameters of the HMMs are also changed to incorporate the new information from the new sequence. This procedure is repeated until all sequences in the dataset have been processed. As a parameter, the number of HMMs in the ensemble can be adjusted based on the needs of users. We have implemented this approach into a software tool EHMM and our experimental results show that the prediction accuracy of EHMM is higher than or comparable with that of the existing tools. Our testing results suggest that EHMM has the potential to provide some assistance to the ENCODE Project.

2 Algorithms and Methods

The method selects the two sequences that have the lowest similarity to initialize the ensemble. The similarity between each pair of sequences in the set is computed by globally aligning the two sequences. A local alignment of the selected sequences is then computed. The alignment results are then used to construct an ensemble that consists of k HMMs, where k is a positive integer. The algorithm selects the local alignments with the k largest alignment scores and each of such local alignments can be used to construct an HMM. An ensemble of k HMMs can thus be constructed based on the local alignments with k most significant alignment scores.

We then progressively use the HMMs to scan through each remaining sequence in the set. Each sequence segment in a sequence is aligned to each HMM in the ensemble and the alignments with the k most significant scores are selected to update

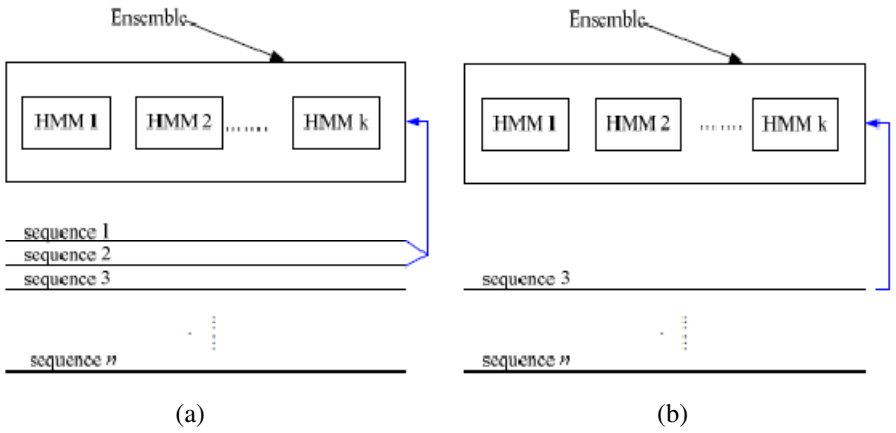

Fig. 1. (a) An ensemble is constructed from local alignments. (b) The ensemble is updated progressively.

the parameters of the HMM. This process will create up to k^2 HMMs, but only the alignments that have the k most significant alignment scores are selected to create a new ensemble of k HMMs. We repeat this procedure until all sequences in the set have been processed and the HMMs remained in the ensemble provide the candidate TFBS motifs. Figure 1 (a) shows the initialization of the ensemble and Figure 1 (b) illustrates how the ensemble is updated. Figure 2 shows the final stage of the approach, where the binding sites can be determined from the HMMs in the ensemble.

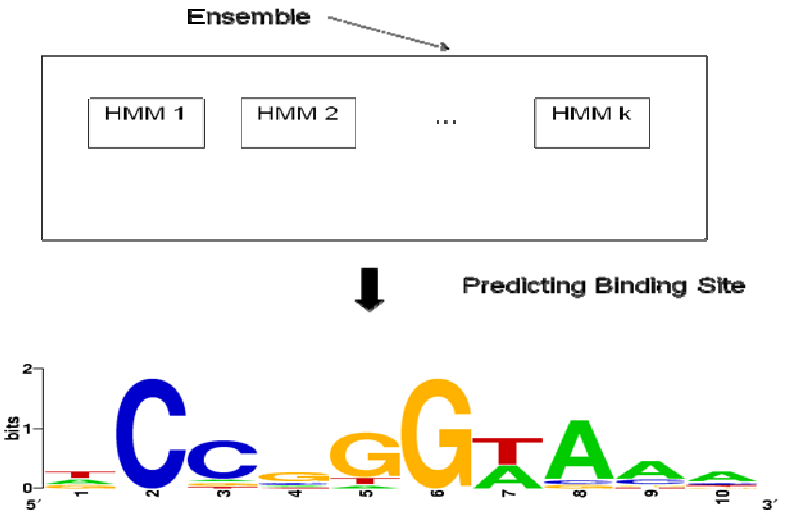

Fig. 2. Finally the binding sites can be inferred from the HMMs in the ensemble

2.1 Ensemble Initialization

The algorithm selects two sequences that are of the lowest similarity value from the set and uses Smith-Waterman local alignment algorithm [20] to obtain local alignments with significant scores. The alignment is computed based on dynamic programming. Given two sequences s and t, a dynamic programming table S, and a score matrix M that stores the fitness value to match two nucleotides together in an alignment. The recursion relation for the dynamic programming is as follows.

$$S[i,j] = \max\{0, S[i-1][j] + M[s_i, -], S[i][j-1] + M[-, t_j], \\ S[i-1][j-1] + M[s_i][t_j]\} \quad (1)$$

where $S[i,j]$ is the optimal alignment score between subsequences $s[1...i]$ and $t[1...j]$, s_i and t_j are the i th and j th nucleotides in s and t and '-' represents a possible gap that may appear in a local alignment. After S is completely

determined, the algorithm selects the alignments with the k largest alignment scores in S. An ensemble of k profile HMMs can then be constructed from these k alignments.

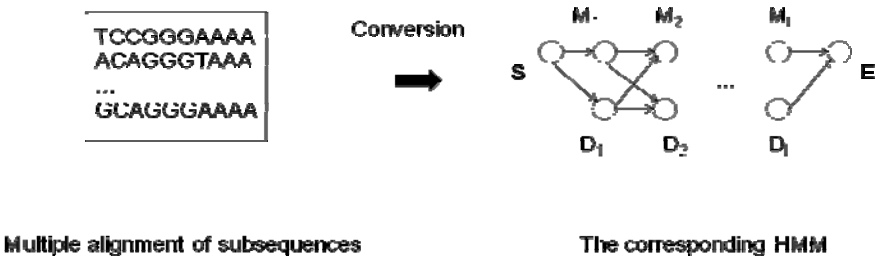

Fig. 3. A multiple alignment of subsequences can be converted into a profile HMM

Each column in an alignment contains a set of nucleotides and gaps that are aligned together. A profile HMM can be used to describe these columns. Specifically, a column i in the corresponding alignment can be modeled by two states, namely D_i and M_i, in a profile HMM. The deletion state D_i does not emit any nucleotide and is used to represent the gaps in column i; the matching state M_i emits a nucleotide and is used to describe the probabilities for each nucleotide to appear in column i. In addition, transitions in a profile HMM may occur only between states for columns consecutive in the corresponding alignment. The probabilities of emission and transition for each state can be computed from each alignment as well. Figure 3 shows that two sates can be created for each column in a multiple alignment of subsequences in the corresponding profile HMM and transitions may occur between the states for two consecutive columns. The parameters of a profile HMM can be computed as follows.

$$ep(M_i, a) = \frac{C_{ia}}{\sum_{b \in N} C_{ib}} \tag{2}$$

$$et(M_i, M_{i+1}) = \frac{\sum_{b \in N, c \in N} P(i, b, i+1, c)}{\sum_{b \in N, c \in N} P(i, b, i+1, c) + \sum_{b \in N} P(i, b, i+1, -)} \tag{3}$$

$$et(M_i, D_{i+1}) = 1 - et(M_i, M_{i+1}) \tag{4}$$

$$et(D_i, M_{i+1}) = \frac{\sum_{b \in N} P(i,-,i+1,b)}{\sum_{b \in N} P(i,-,i+1,b) + P(i,-,i+1,-)} \quad (5)$$

$$et(D_i, D_{i+1}) = 1 - et(D_i, M_{i+1}) \quad (6)$$

where N is the set of all types of nucleotides, C_{ia} represents the number of times that nucleotide a appears in column i, $et(M_i, a)$ is the emission probability for state M_i to emit nucleotide a. $et(M_i, M_{i+1})$ is the probability for the transition from M_i to M_{i+1} to occur; $P(i,b,i+1,c)$ is the number of times that nucleotide b appears in column i and nucleotide c appears in position $i+1$; $P(i,b,i+1,-)$ is the number of times that nucleotide b appears in column i and a gap appears in column $i+1$. $et(D_i, M_{i+1})$ is the probability for the transition from D_i to M_{i+1} to occur; $P(i,-,i+1,b)$ is the number of times that a gap appears in column i and nucleotide b appears in column $i+1$; $P(i,-,i+1,-)$ is the number of times that gaps appear in both columns i and $i+1$. More details of the algorithm can be found in [5].

2.2 Updating Ensemble

The remaining sequences in the set are processed based on the profile HMMs in the ensemble. For each of the remaining sequences, we evaluate the average similarity between it and the two sequences that have been selected to initialize the ensemble. The remaining sequences can thus be sorted based on an ascending order of this similarity value. This order is the execution order of the remaining sequences in the set.

Each of the remaining sequence is scanned through by each profile HMM in the execution order and subsequences that have the k most significant alignment scores are selected. The algorithm uses a window of certain size to slide through the sequence. The size of the window is set to be 1.5 times of the average lengths of all subsequences in the alignments used to construct the ensemble. The window moves by 1bp each time and each subsequence in the window is aligned to each HMM in the ensemble. The alignment can be computed with a dynamic programming algorithm. The recursion relation for the dynamic programming is as follows.

$$S[D_s, i, j] = \max\{et(D_s, D_{s+1})S[D_{s+1}, i, j], et(D_s, M_{s+1})S[M_{s+1}, i, j]\} \quad (7)$$

$$\begin{aligned} S[M_s, i, j] = \max\{&et(M_s, D_s)ep(M_s, t_i)S[D_s, i+1, j], \\ &et(M_s, M_{s+1})S[M_{s+1}, i+1, j]\} \end{aligned} \quad (8)$$

where $0 \le i \le j \le W$ are integers that indicate the location of subsequence t included in the window; $S[D_s, i, j]$ and $S[M_s, i, j]$ are the dynamic programming table cells that store the maximum probability for states D_i and M_i to generate the subsequence $t[i...j]$; t_i is the nucleotide at position i in t. More details of the algorithm can be found in [5].

The algorithm then selects k subsequences with the largest alignment scores. We thus obtain in total k^2 candidates for updating the HMMs in the ensemble. We pick k subsequences that correspond to the largest k alignment scores from the k^2 candidates. The parameters of each profile HMM are then updated based on these additional k subsequences. Specifically, the additional subsequence changes the counts that appear in (2), (3), (4), (5), and (6). The process is applied progressively to other remaining sequences in the set until each sequence in the set has been processed. The locations of the sequence segments used to construct each HMM in the ensemble are then output as the possible binding sites.

2.3 Computation Time

We assume the set contains m sequences, each sequence contains n nucleotides, and the binding site contains l nucleotides. The construction of the initial ensemble needs $O(m^2n^2 + kn^2)$ time. The computation time needed to scan through a sequence with a single HMM is $O(l^2n)$. The total amount of computation needed by the approach is thus $O(kml^2n^2 + m^2n^2 + kn^2)$.

3 Experimental Results

We have implemented this approach and integrated it into a software tool EHMM. We tested its accuracy on a biological dataset cyclic-AMP receptor protein (CRP). This dataset consists of 18 sequences, each of which consists of 105 bps [17]. Twenty three binding sites have been determined by using the DNA footprinting method, with a motif width of 22 [16].

Figure 4 compares the prediction accuracy of EHMM with three other computational methods: Gibbs Sampler [8], BioProspector [9], and MDGA [3]. The value of the parameter is set to be $k = 10$ in all the tests. It can be seen from the table that EHMM can achieve comparable accuracy with other tools in homologous sequences that contain a single binding site. However, sequences 1, 2, 6, 9, and 17 contain two TFBSs and all three other tools fail to recognize the second one. Table 1 shows the errors of the predicted locations of the second binding site in these sequences by EHMM. For most of them, EHMM can thus accurately identify the locations of both motifs. In particular, EHMM obtains excellent prediction results on

sequence 17, where all three other methods fail to identify either of the two TFBSs. Our method is capable of identifying the locations of multiple binding sites since it uses an ensemble of HMMs to explore the alignment space of all subsequences, which significantly improves the sampling ability and the probability to accurately identify the locations of TFBSs.

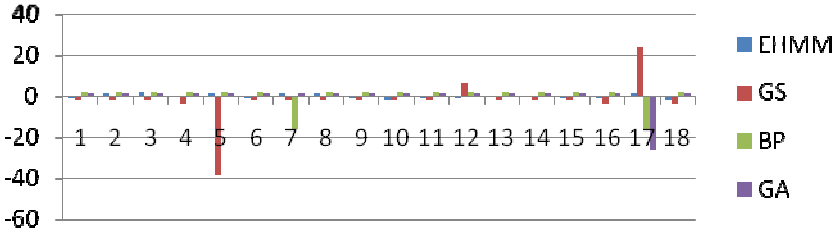

Fig. 4. The error of predicted locations of TFBSs by EHMM, GS(Gibbs Sampler), BP(BioProspector), GA(MDGA). The error is the deviation of the predicted starting positions from those obtained with fingerprint experiments.

In addition to the data set CRP, we also use EHMM ($k = 10$) and other tools to predict the binding sites for a few transcription factors including BATF [17], EGR1[9], FOXO1[3], and HSF1[18]. The prediction accuracy of a software tool is evaluated by computing its prediction accuracy on each single sequence in the set and taking the average of the prediction accuracy on all sequences in the set. The prediction accuracy on a single sequence is defined to be the percentage of correctly predicted part in the binding site. In other words, if we use B to denote the binding site and P to denote the predicted binding site, the accuracy of the prediction can be computed with

$$A = \frac{|P \cap B|}{|B|} \tag{9}$$

where $P \cap B$ denotes the intersection of P and B. For a set D of homologous sequences, the prediction accuracy of an approach on D is computed with

$$A_D = \frac{\sum_{s \in D} A_s}{|D|} \tag{10}$$

where s is a sequence in D and A_s is the prediction accuracy of the approach on s. Figure 5 (a) shows and compares the prediction accuracy of EHMM, Gibbs Sampler, BioProspector, and MDGA on the four data sets. For each of the other three tools, we test its prediction accuracy for 100 times and use the average prediction accuracy for comparison. It is not difficult to see from the figure that EHMM achieves significantly higher prediction accuracy on data sets BATF, FOXO1, and HSF1 and achieves accuracy that is comparable with other tools on data set FOXO1.

Table 1. The errors of the locations of the second TFBS predicted by EHMM

Seq.#	1	2	6	9	17
Error	-1	-1	-1	1	-4

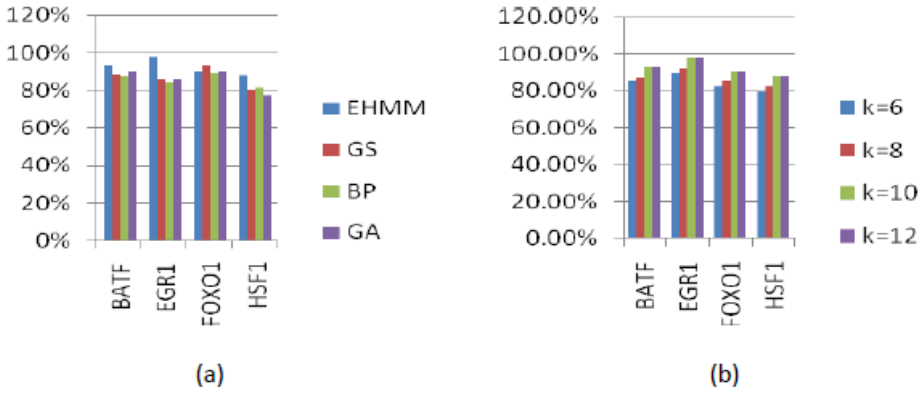

Fig. 5. (a) Prediction accuracy of the EHMM, GS, BP, GA on data sets BATF, EGR1, FOXO1, and HSF1. (b) Prediction accuracy of the EHMM when k is 6,8,10, and 12 respectively.

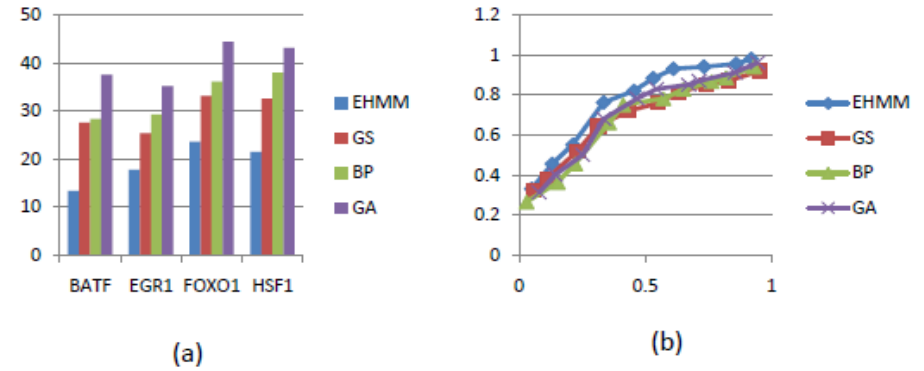

Fig. 6. (a) Computation time needed by the four programs on all data sets. (b) The ROC curve for the four programs.

The size of the ensemble can be changed by the user to balance the prediction accuracy and the computation time needed for prediction. Figure 5 (b) shows the prediction accuracy on data sets BATF, EGR1, FOXO1, and HSF1 when the value of the parameter k is 6,8,10, and 12. It can be seen from the figure that the prediction

accuracy improves when the size of the ensemble increases and the prediction accuracy becomes steady when the value of the parameter is 10. The testing results also show that a parameter value of 10 is thus sufficient to achieve satisfactory prediction accuracy in practice.

Figure 6 (a) shows the computation time needed by the four programs on all data sets in seconds. It can be seen from the figure that EHMM is computationally more efficient than the other three programs. Figure 6 (b) shows the ROC curve of all four programs computed based on the four testing data sets. The horizontal axis in the figure is the value of 1-specificity and the vertical axis represents the sensitivity. It is also clear from the figure that EHMM is on average the most accurate program of all four programs.

4 Conclusions

In this paper, we developed a new parameterized approach that can accurately and efficiently identify the binding sites with an ensemble of HMMs. Experimental results show that this approach can achieve higher or comparable accuracy on sequences with a single binding site while its accuracy on sequences with multiple binding sites is significantly higher than that of other tools. Our approach thus may provide a useful computational tool for the ENCODE project [32], whose goal is to identify all functional elements in human genome sequences.

Our previous work has demonstrated that introducing additional parameters to the algorithms for some bioinformatics problems may significantly improve the accuracy of the results [10-13, 22-29]. Our future work will focus on the development of new approaches that can exploit these parameters to further improve the accuracy of binding site prediction.

Acknowledgments. Y. Song's work is fully supported by the University Fund of Jiangsu University of Science and Technology, under the Number 635301202 and 633301301.

References

1. Bailey, T.L., Elkan, C.: Unsupervised Learning of Multiple Motifs In Biopolymers Using Expectation Maximization. Technical Report CS93-302, Department of Computer Science, University of California, San Diego (August 1993)
2. Bailey, T.L., Elkan, C.: Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. In: Proceedings of the Second International Conference on Intelligient Systems for Molecular Biology, pp. 28–36 (1994)
3. Brent, M.M., Anand, R., Marmorstein, R.: Structural Basis for DNA Recognition by Foxo1 and its Regulation by Posttranslational Modification. Structure 16, 1407–1416 (2008)
4. Che, D., Song, Y., Rasheed, K.: MDGA: Motif Discovery Using a Genetic Algorithm. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 447–452 (2005)

5. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1998)
6. Galas, D.J., Schmitz, A.: A DNA Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity. Nucleic Acids Research 5(9), 3157–3170 (1978)
7. Garner, M.M., Revzin, A.: A Gel Electrophoresis Method for Quantifying He Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia Coli Lactose Operon Regulatory Systems. Nucleic Acids Research 9(13), 3047–3060 (1981)
8. Hertz, G.Z., Stormo, G.D.: Identifying DNA and Protein Patterns with Statistically Significant Alignments of Multiple Sequences. Bioinformatcs 15(7), 53–577 (1999)
9. Hu, T.C., et al.: Snail Associates with EGR-1 and SP-1 to Upregulate Transcriptional Activation of P15ink4b. The FEBS Journal 277, 1202–1218 (2010)
10. Liu, C., Song, Y., Shapiro, L.W.: RNA Folding Including Pseudoknots: A New Parameterized Algorithm and Improved Upper Bound. In: Proceedings of the 7th Workshop on Algorithms in Bioinformatics, pp. 310–322 (2007)
11. Liu, C., Song, Y., Burge III, L.L.: Parameterized Lower Bound and Inapproximability of Polylogarithmic String Barcoding. Journal of Combinatorial Optimization 16(1), 39–49 (2008)
12. Liu, C., Song, Y.: Parameterized Dominating Set Problem in Chordal Graphs: Complexity and Lower Bound. Journal of Combinatorial Optimization 18(1), 87–97 (2009)
13. Liu, C., Song, Y.: Parameterized Complexity and Inapproximability of Dominating Set Problem in Chordal and Near Chordal Graphs. Journal of Combinatorial Optimization 22(4), 684–698 (2011)
14. Liu, F.F.M., Tsai, J.J.P., Chen, R.M., Chen, S.N., Shih, S.H.: FMGA: Finding Motifs by Genetic Algorithm. In: IEEE Fourth Symposium on Bioinformatics And Bioengineering, pp. 459–466 (2004)
15. Liu, J.S., Neuwald, A.F., Lawrence, C.E.: Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. J. Am. Stat. Assoc. 90(432), 1156–1170 (1995)
16. Liu, X., Brutlag, D.L., Liu, J.S.: Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes. In: Pacific Symposium of Biocomputing, vol. 6, pp. 127–1138 (2001)
17. Quigley, M., et al.: Transcriptional Analysis of HIV-Specific CD8+ T Cells Shows That PD-1 Inhibits T Cell Function by Upregulating BATF. Nature Medicine 16, 1147–1151 (2010)
18. Rigbolt, K.T., et al.: System-Wide Temporal Characterization of the Proteome and Phosphoproteome of Human Embryonic Stem Cell Differentiation. Science Signaling 4, RS3–RS3 (2011)
19. Roth, F.R., Hughes, J.D., Estep, P.E., Church, G.M.: Finding DNA Regulatory Motifs Within Unaligned Non-Coding Sequences Clustered by Whole-Genome Mrna Quantitation. Nature Biotechnology 16(10), 939–945 (1998)
20. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology 147, 195–197 (1981)
21. Song, J., Liu, C., Song, Y., Qu, J., Hura, G.: Alignment of Multiple Proteins With an Ensemble of Hidden Markov Models. International Journal of Bioinformatics and Data Mining 4(1), 60–71 (2010)
22. Song, Y., Liu, C., Huang, X., Malmberg, R.L., Xu, Y., Cai, L.: Efficient Parameterized Algorithm for Biopolymer Structure-Sequence Alignment. In: Casadio, R., Myers, G. (eds.) WABI 2005. LNCS (LNBI), vol. 3692, pp. 376–388. Springer, Heidelberg (2005)

23. Song, Y., Liu, C., Malmberg, R.L., Pan, F., Cai, L.: Tree Decomposition Based Fast Search of RNA Structures Including Pseudoknots in Genomes. In: Proceedings of IEEE 2005 Computational Systems Bioinformatics Conference, pp. 223–234 (2005)
24. Song, Y., Zhao, J., Liu, C., Liu, K., Malmberg, R.L., Cai, L.: RNA Structural Homology Search with a Succinct Stochastic Grammar Model. Journal of Computer Science and Technology 20(4), 454–464 (2005)
25. Song, Y., Liu, C., Huang, X., Malmberg, R.L., Xu, Y., Cai, L.: Efficient Parameterized Algorithms for Biopolymer Structure-Sequence Alignment. IEEE/ACM Transactions on Computational Biology and Bioinformatics 3(4), 423–432 (2006)
26. Song, Y., Liu, C., Malmberg, R.L., He, C., Cai, L.: Memory Efficient Alignment Between RNA Sequences and Stochastic Grammar Models of Pseudoknots. International Journal on Bioinformatics Research and Applications 2(3), 289–304 (2006)
27. Song, Y.: A New Parameterized Algorithm for Rapid Peptide Sequencing. Plos ONE 9(2), E87476 (2014)
28. Song, Y., Chi, A.Y.: A New Approach for Parameter Estimation in the Sequence-Structure Alignment of Non-Coding Rnas. Journal of Information Science and Engineering (in press, 2014)
29. Song, Y.: An Improved Parameterized Algorithm for the Independent Feedback Vertex Set Problem. Theoretical Computer Science (2014), doi:10.1016/J.Tcs.2014.03.031
30. Stormo, G.D.: Computer Methods for Analyzing Sequence Recognition of Nucleic Acids. Annu. Rev. Biochem. 17, 241–263 (1988)
31. Stormo, G.D., Hartzell, G.W.: Identifying Protein-Binding Sites from Unaligned DNA Fragments. Proc. of Nat. Acad. Sci. 86(4), 1183–1187 (1989)
32. `https://Www.Genome.Gov/Encode/`