Motif Discovery Using Chemical Reaction Optimization

1 Introduction

Motif discovery means the process of determining motifs within a set of DNA, RNA or protein sequences where a motif means a widespread amino-acid sequence or nucleotide pattern that captures a biological significance [1].Motifs are usually fixed length, short sequence patterns that represents important functional features or structural features in nucleic acid and protein sequences such as active sites, transcription binding sites, interaction interfaces or splice junctions. They can occur in an exact or approximate form within a family or a subfamily of sequences [2].In other words, a pattern common to a set of DNA, RNA or protein sequence that shares a common biological property, such as functioning as binding sites for a particular protein is called motif. So we can say that the problem of identifying short similar sequence elements shared by a set of protein or nucleotide sequences with a general biological function is defined as motif discovery [3]. For an example, n = 5 DNA sequences of length L = 35 are used to discover a motif of width W = 10 using position weight matrix (PWM), f where PWM is a generally used to representation of motifs [4].

Commented [c1]: You have to give example(s) using Figure under this paragraph.

Commented [WU2R1]: Dear Sir, we use blue color for newly added text.

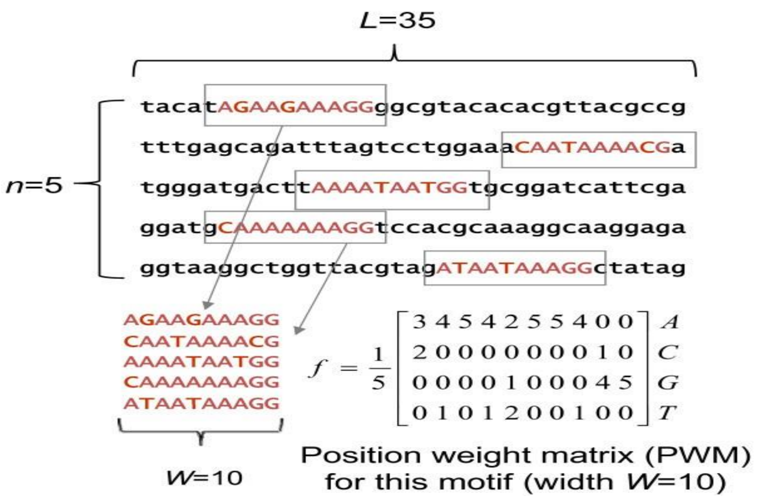
L=35
n=5
tacatAGAAGAAAGGggcgtacacacgttacgccg
tttgagcagatttagtcctggaaaCAATAAAACGa
tgggatgacttAAAATAATGGtgcggatcattcga
ggatgCAAAAAAAGGtccacgcaaaggcaaggaga
ggtaaggctggttacgtagATAATAAAGGctatag
AGAAGAAAGG
CAATAAAACG
AAAATAATGG
CAAAAAAAGG
ATAATAAAGG
W=10
$f = \frac{1}{5}\begin{array}{c|c} \begin{bmatrix} 3&4&5&4&2&5&5&4&0&0\\ 2&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&1&0&0&0&4&5\\ 0&1&0&1&2&0&0&1&0&0 \end{bmatrix} & \begin{matrix} A\\ C\\ G\\ T \end{matrix} \end{array}$
Position weight matrix (PWM)
for this motif (width W=10)

Figure 1. Motif discovery. Reprinted from Zia, Amin, and Alan M. Moses. "Towards a theoretical understanding of false positives in DNA motif finding." *BMC bioinformatics* 13.1 (2012): 151.

Now from every sequence we get a motif and then by using PWM, we discover the final motif from these motifs.

In the era of bioinformatics revolution, the volume of biological sequences is increasing in public databases. That's why motif discovery has become one of the fundamental problems in computer science and molecular biology [5]. The capability to predict the function, structure, or behavior of biological entities or motifs such as proteins and genes, additionally cooperation among them, play a major role in the analysis of information to describe biological mechanisms [6].Motif discovery is, therefore, an important field of bioinformatics.

There are two main ways to discover a motif: biological experiment and computing approaches, i.e. bioinformatics. But biological experiments are very costly and time consuming process. That's why computing approaches are extensively used to discover motifs [7]. But accurate identification of motifs is a challenging problem. Because the length of motifs are usually very short (up to 30 nucleotides), while that of the regulatory regions which contain motifs are very long (range from several hundred to several thousand nucleotides). In addition, the mutations of the actual instances of motifs are adding to the burden [5].

Motif discovery has several important application areas. It is widely used in locating regulatory sites and drug target identification [5]. It is mainly used to the analysis of information for describing biological mechanisms [6].Besides, motif discovery has become the main part of several higher-level algorithms handle with time series specially rule-discovery, compression, summarization and clustering algorithms.

Still, now many algorithms have been proposed to predict motifs, such as Gibbs sampler, MEME, GA, GARPS, ACO, ACOMotif, EMACO, MFACO, ACRI, MotifSuite, MotifSampler, Bioprospector, Iterative Algorithm, genetic algorithm based on statistical significance etc. These algorithms have their drawback of dropping into local optimum easily. The consuming time of Gibbs sampler is lower but less prediction accuracy, and MEME is superior to the other methods by its prediction accuracy but time-consuming [5]. In addition, there are many other heuristic methods to predict motifs, such as particle swarm optimization, Tabu search algorithm, and Simulated Annealing. Several fundamental limitations of these algorithms are the nucleotide level and binding site level prediction accuracy are still very low even on the prokaryotic motifs. Another limitation for transaction factors is the pattern model to capture the regularity among the binding sites [8]. To enable the biologist to determine functional motifs from statistical artifacts, many algorithms do not produce good motif statistics. For this reason valid motifs can be rejected, or time may be wasted by searching random motifs [9]. The main drawback of genetic algorithm based on statistical significance is the lack of a mechanism to identify false positives [10]. Though all algorithms have some limitations, they produce better results in some restricted inputs criteria.

2 Problem Statement

Motif discovery problem can be explained as follows. Let a set of DNA sequences S = S_1, S_2, ..., S_n having same length .We have to find out the possible accurate motif pattern X = x_1,x_2...x_i...x_l of length *l* where $x_i \in$ {A, T,C,G} . Motif discovery is based on a defined score function that calculate the similarity of the motif pattern with its occurrences. To find intended motif pattern the objective function should have one of the following characteristics:

Commented [c3]: You have to give short description of each of the followings.

Let assume the motif is ACGTATGC.

1. **Perfect matches only (no mismatches):** The function can only allow this pattern in a sequence.
2. **Allow a given number of mismatches:** If the function allows only one mismatch, then the patterns with one position mismatch or no mismatch are allowed in a sequence such as ACATATGC, ACGTCTGC, and ACGTATGC etc.
3. **Allow a given density of mismatches:** If density of mismatch is 2, then the function allows at most 2 alongside mismatches such as ACGATTGC, CAGTATAC but it cannot allow GACTATGC or ACTAGTGC.

Now-a-days, in bioinformatics motif discovery in gene sequences is one of the most important and NP-hard problems [7] where NP-hard problems are those problems that are at least as hard as the hardest problems in NP [11]. Here NP means non-deterministic polynomial and the problems that have no solution in polynomial time is called NP problems. If we use real biological DNA sequences where the length of the nucleotide (or amino-acid) are not fixed and very large, then we cannot find the exact motif in polynomial time. That's why motif discovery is NP-hard problem and here we solve this problem using chemical reaction optimization (a meta-heuristics method) where we interested in near-optimal (or sub-optimal) solution. Several methods can be used for scoring a motif discovery such as consensus score [10] and information content [13, 14] as defined score functions.

Commented [c4]: Why is it NP-hard? You should explain it.

The candidate motif pattern can be represented by a count-based profile C where C(i, j) is the count of nucleotide i on the column j of the alignment matrix and its corresponding consensus score (CSc) is defined as:

$$CSc = \sum_{j=1}^{l} \left(\max_{i \in \{A,T,C,G\}} (C(i,j)) \right) \qquad (1)$$

The information content (IC) score function computes as follows:

$$IC = \sum_{j=1}^{i} \sum_{i \in \{A,T,C,G\}} Q(i,j) . \log_2 \frac{Q(i,j)}{B0(i)} \qquad (2)$$

Where each element Q(i, j) represents the frequency of the nucleotide i to be in position j of the motif pattern and B_0(i) indicates its background frequency[15].

References

1. Lones, Michael, and Andy Tyrrell. "Regulatory motif discovery using a population clustering evolutionary algorithm." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4.3 (2007).
2. Mohamed, Salma Aouled El Haj, Mourad Elloumi, and Julie D. Thompson. "Motif Discovery in Protein Sequences." *Pattern Recognition-Analysis and Applications*. InTech, 2016.
3. Zambelli, Federico, Graziano Pesole, and Giulio Pavesi. "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era." *Briefings in bioinformatics* 14.2 (2012): 225-237.
4. “Position weight matrix – Wikipedia, the free encyclopedia.” Web. 16 September 2017. < https://en.wikipedia.org/wiki/Position_weight_matrix >.
5. Fan, Yetian, Wei Wu, Rongrong Liu, and Wenyu Yang. "An Iterative Algorithm for Motif Discovery." *Procedia Computer Science* 24 (2013): 25-29.

6. Angela Makolo. A Comparative Analysis of Motif Discovery Algorithms. *Computational Biology and Bioinformatics*. Vol. 4, No. 1, 2016, pp. 1-9. doi: 10.11648/j.cbb.20160401.11.
7. Huan, Hoang X., et al. "An Efficient Ant Colony Algorithm for DNA Motif Finding." *Knowledge and Systems Engineering*. Springer, Cham, 2015. 589-601.
8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1199555.
9. http://hdl.handle.net/1721.1/38976.
10. Gutierrez, Josep Basha, Martin Frith, and Kenta Nakai. "A genetic algorithm for motif finding based on statistical significance." *International Conference on Bioinformatics and Biomedical Engineering*. Springer, Cham, 2015.
11. "NP-hardness – Wikipedia, the free encyclopedia." Web. 17 September 2017. < https://en.wikipedia.org/wiki/NP-hardness >.
12. Jones, Neil C., and Pavel Pevzner. *An introduction to bioinformatics algorithms*. MIT press, 2004.
13. Che, Dongsheng, Yinglei Song, and Khaled Rasheed. "MDGA: motif discovery using a genetic algorithm." *Proceedings of the 7th annual conference on Genetic and evolutionary computation*. ACM, 2005.
14. Stormo, Gary D., and George W. Hartzell. "Identifying protein-binding sites from unaligned DNA fragments." *Proceedings of the National Academy of Sciences* 86.4 (1989): 1183-1187.
15. Bouamama, Salim, Abdellah Boukerram, and Amer F. Al-Badarneh. "Motif Finding Using Ant Colony Optimization." *ANTS Conference*. 2010.