Motif Discovery in Biological Data Using Simulated Annealing with Neighborhood Search and Iterative Restart Strategy

Vinita Singh
Department of Computer Science, faculty of Computer Science, University of Mumbai-400 032
Hindi Vidya Prachar Samiti's Ramniranjan Jhunjhunwala College (Empowered Autonomous),
GHATKOPAR (W), MUMBAI – 400 086

***Abstract*—Motif discovery in biological data, such as DNA, RNA, and protein sequences, plays a crucial role in understanding biological functions and regulatory mechanisms. The identification of recurring, biologically significant patterns or motifs is a challenging problem due to the vast search space and the inherent noise in biological datasets. This paper proposes a novel approach to motif discovery that combines the power of simulated annealing (SA) with a neighborhood search and iterative restart strategy to efficiently explore the solution space and avoid local optima.**

Simulated annealing is a probabilistic optimization technique inspired by the physical process of annealing, where the system gradually stabilizes to a low-energy state. We enhance this standard method by integrating a neighborhood search mechanism that dynamically explores neighboring motifs to refine the solution space. Additionally, an iterative restart strategy is introduced to overcome premature convergence and increase the likelihood of discovering global optima. The method is designed to handle the combinatorial nature of motif discovery and mitigate issues related to overfitting and underfitting by balancing exploration and exploitation during the search process.

Keywords—Motif, Nucleotide Sequences, Simulated Annealing algorithm

I. INTRODUCTION

Traditional motif discovery methods often rely on heuristic approaches that can struggle with complex datasets, where the search space is large and prone to local optima. Simulated annealing (SA), a probabilistic technique inspired by the annealing process in metallurgy, has emerged as a powerful tool for global optimization in motif discovery. However, the effectiveness of SA can be limited by the risk of getting trapped in local optima, especially in large, complex search spaces.

To address these challenges, this paper proposes an enhanced motif discovery approach that combines simulated annealing with neighborhood search and an iterative restart strategy. The neighborhood search improves the search process by exploring nearby solutions more thoroughly, while the iterative restart strategy helps to avoid local optima by periodically re-initializing the search. Together, these strategies aim to balance exploration and exploitation, ultimately leading to more accurate motif discovery in biological data.

In the following sections, we describe the methodology in detail, present experimental results, and demonstrate the effectiveness of the proposed approach in finding biologically relevant motifs across various datasets.

II. RELATED WORK

Motifs are short sequence patterns of biological significance in either DNA, RNA or protein sequences. The discovery of such motifs is an important task in molecular biology. The characterization and localization of motifs is a fundamental approach to a better understanding of the structure, function and evolutionary relationships of the corresponding genes or proteins.[1] Most of the earlier literature categorized motif finding algorithms into two major groups based on the combinatorial approach used in their design: (1) word-based (string-based) methods that mostly rely on exhaustive enumeration, i.e., counting and comparing oligonucleotide frequencies and (2) probabilistic sequence models where the model parameters are estimated using maximum-likelihood principle or Bayesian inference. [2] A classical approach to find RNA motifs is to construct a consensus RNA secondary structure from a given multiple sequence alignment based on covariation, thermodynamic stability, phylogeny, etc. [3] different motif finding Web tools provide different customizations for finding motifs as well as provide different result formats. Some Web tools have restrictions on the size of input sequence, the number of peaks, or the size of upload file.

III. METHODOLOGY

In this paper, algorithm Simulated Annealing is presented, which explores some new strategies, based on a neighborhood set concept & random search technique, which can capture the target motifs effectively.

SA method is that the temperature is gradually reduced as the simulation proceeds. Initially, T is set to a high value (or infinity), and it is decreased at each step according to some annealing schedule—which may be specified by the user but must end with T = 0 towards the end of the allotted time budget. In this way, the system is expected to wander initially towards a broad region of the search space containing good solutions, ignoring small features of the energy function; then drift towards low-energy regions that become narrower and narrower; and finally move downhill according to the steepest descent heuristic. It can be shown that for any given finite problem, the probability that the simulated annealing algorithm terminates with the global optimal solution approaches 1 as the annealing schedule is extended. This theoretical result, however, is not particularly helpful, since the time required to ensure a significant probability of success will usually exceed the time required for a complete search for the solution space.

3.1 Candidate Motifs:

From Each Sequence make a set of subsequences of length ‘l’ using neighborhood theory.

Example for DNA sequence

1.Set of candidate motifs {GAAAA, AAAAT, AAATG, AATGA, ATGAG, TGAGT, GAGTG, AGTAC, GTACA, TACAT, ACATG}

If sequence length is ‘len’ then no of candidate

Data Gathering

Data in Bioinformatics is of prime importance and is motifs=(len-l)+1.

In this case len=15, l=5 and no of candidate motifs=(15-5)+1=11.

Perform similarly process for all sequence.

Objective Function : A randomly chosen from a random sequence is compared with all the candidate motifs of length ‘l’.

Randomly selected substring GAGTG is compared with each of the candidate motif from above set of candidate motif.

3.2 Objective Function:

matching residues i.e Objective function return all the residues which matches to GAGTG. in sequence there is residue at position 7 which matches to GAGTG. So we found motif GAGTG of length ‘l’ in first sequence. Similarly for all sequence.

Annealing Process: In this Process we choose a subsequence whose mismatch parameter i.e. d=0. And accordingly, we calculate probability ’p’ for each sequence, at position 7 we got a substring GAGTG for which mismatch parameter d=0.

Calculate Probability:--

For i=0 to len-l

Calculate p=n*d;

Perform similarly process for all sequence.

Local minima: If probability calculated for particular sequence is zero then we say local minima reached. It deals with each sequence locally.

Termination Condition: The process is terminated once the global optimal solution approaches. When probability ‘p’ for each sequence is zero then global optimal solution approached.

For i=0 to n

If p=0 then global optimal solution;

3.3 Steps involved in simulated annealing for finding motif:

Input the n sequences and motif length l.

Select any random subsequence of specified length from randomly selected sequence.

Create a set of candidate motifs.

Check each candidate motif with randomly selected subsequence.

If missmatch(d) is 0 for all sequence, then stop the process.

Else repeat the steps from 2-5.

The initialization process randomly selects a DNA sequence and choose the core foundation in the process of Development of any Biological Automated System.

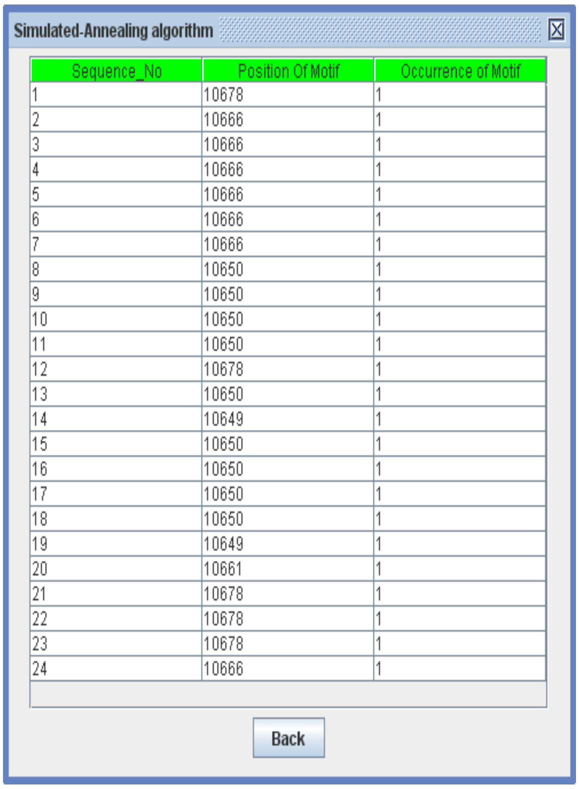
The data dealt in this project are primarily nucleotide sequences of Flaviviridae family strains which includes genes as Flavivirus, Pestivirus and Hepacivirus. This data has been widely gathered from the websites: www.ncbi.nlm.nih.gov/. The proposed algorithm for the Motif finding consists of the initialization process, candidate motif generation process, Objective function, and the Termination process as shown in the flowchart of the proposed algorithm. Detailed steps of the proposed Simulated Annealing algorithm for motif finding are described as follows:

IV.IMPLEMENTATION

The initialization process randomly select motif of specified length and compare it with all the candidate motif of all sequences. An iterative restart strategy is used, by which we can use several motif's information to detect the Identical motif.

V.RESULT AND DISCUSSION

I have found Motif of different length on DNA sequences of different viruses. All these sequences are complete genome of length >=10,000.



Sequence_No	Position Of Motif	Occurrence of Motif
1	10678	1
2	10666	1
3	10666	1
4	10666	1
5	10666	1
6	10666	1
7	10666	1
8	10650	1
9	10650	1
10	10650	1
11	10650	1
12	10678	1
13	10650	1
14	10649	1
15	10650	1
16	10650	1
17	10650	1
18	10650	1
19	10649	1
20	10661	1
21	10678	1
22	10678	1
23	10678	1
24	10666	1

a substring of length *'l'* from this sequence with random starting position. Where *'l'*=motif length.
Consider 8 DNA sequences each of length 15 nucleotide. *Let n=8 and l=5 where n=no of DNA sequences. And l=motif length.*

GAAAATGAGTGCATG
AAGAGTGAGTGGTGT
AGTGAACGTGAGTGC
GAGAGTGAGAATCCA
GAGTGTCAGGCCTGA
AGTGTCGTGGAGTGA
ACACCTTGAGTGAAT
CTGTGAGTGCTTCCG

Let randomly selected sequence=4 and randomly selected position=3.Selected Substring=GAGTG.

VI.CONCLUSION

Given a set of nucleotide sequences we can efficiently find identical motifs using Simulated Annealing Algorithm.
The algorithm uses a heuristic function which allows algorithm to run faster and also allows analysis of biological sequences.

REFERENCES AND BIBLIOGRAPHY

[1] A survey of DNA motif finding algorithms, *BMC Bioinformatics 2007, 8(Suppl 7):S21 doi:10.1186/1471-2105-8-S7-S21*

[2] CMfinder—a covariance model based RNA motif finding algorithm, *bioinformatics Vol. 22 no. 4 2006, pages 445–452 doi:10.1093/bioinformatics/btk008*.

[3] A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data, *Tran and Huang Biology Direct 2014, 9:4 http://www.biologydirect.com/content/9/1/4.*

[4] Motif Finding Using Ant Colony Optimization, *M. Dorigo et al. (Eds.): ANTS 2010, LNCS 6234, pp. 464–471, 2010. c Springer-Verlag Berlin Heidelberg 2010*

[5] Hashim FA, Mabrouk MS, Al-Atabany W. Review of Different Sequence Motif Finding Algorithms. *Avicenna J Med Biotechnol. 2019 Apr-Jun;11(2): 130-148. PMID: 31057715; PMCID: PMC6490410*

[6] Bouamama, S., Boukerram, A., Al-Badarneh, A.F. (2010).

[7] Motif Finding Using Ant Colony Optimization. In: *Dorigo, M., et al. Swarm Intelligence. ANTS 2010. Lecture Notes in Computer Science, vol 6234. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15461-4_45*