Review Article

Review of Different Sequence Motif Finding Algorithms

Fatma A. Hashim 1, Mai S. Mabrouk 2*, and Walid Al-Atabany 1

1. Department of Biomedical Engineering, Helwan University, Egypt
2. Department of Biomedical Engineering, Misr University for Science and Technology (MUST), Egypt

Abstract

The DNA motif discovery is a primary step in many systems for studying gene function. Motif discovery plays a vital role in identification of Transcription Factor Binding Sites (TFBSs) that help in learning the mechanisms for regulation of gene expression. Over the past decades, different algorithms were used to design fast and accurate motif discovery tools. These algorithms are generally classified into consensus or probabilistic approaches that many of them are time-consuming and easily trapped in a local optimum. Nature-inspired algorithms and many of combinatorial algorithms are recently proposed to overcome these problems. This paper presents a general classification of motif discovery algorithms with new sub-categories that facilitate building a successful motif discovery algorithm. It also presents a summary of comparison between them.

Keywords: Algorithms, Bioinformatics, Consensus, Gene expression regulation, Nucleotide motif, Protein binding

Introduction

Motif discovery is one of the sequence analysis problems under the application layer and it is one of the significant difficulties in bioinformatics applications. A DNA sequence motif is a subsequence of DNA sequence that is a short similar recurring pattern of nucleotides, and it has many biological functions 1. A DNA motif refers to a short similar repeated pattern of nucleotides that has biological meaning. Sequence motifs also called regulatory elements exist in Regulatory Region (RR) in eukaryotic gene 2.

Sequence motifs have constant size and are often repeated and conserved, but at the same time, they are tiny (about 6-12 bp) and the intergenic regions are very long and highly variable that make motif discovery a problematic task. These patterns play an essential role in recognizing Transcription Factor Binding Sites (TFBSs) that help in learning the mechanisms for regulation of gene expression 3. Different types of motifs are planted motifs, structured motifs, sequence motifs, gapped motifs and network motifs 4. Motif discovery problem in a simple form can be formulated as in figure 1 where the input is DNA sequence with unknown motifs at different unknown positions with various lengths and the output is the DNA motifs. The motif discovery technique consists of three main stages 5:

A. Pre-processing

It is preparing the DNA sequences for accurate motif discovery by assembling and clean steps. In assembling step, it is advised to select as many target sequences as possible that may contain motifs, try to keep sequences as short as possible, and remove sequences that are unlikely to contain any motifs. Assembling step is done by clustering the input sequences based on some information and then extracting the desired sequences in an appropriate sequence database. Then, cleaning the input sequences to mask or remove confounding sequences is necessary.

* Corresponding author: Mai S. Mabrouk, Ph.D., Department of Biomedical Engineering, Misr University for Science and Technology (MUST), Egypt. E-mail: msm_eng@yahoo.com

Received: 12 Feb 2018
Accepted: 26 May 2018

Figure 1. General block diagram of motif discovery technique.
B. Discovering

The middle stage is the motif discovery approach that begins by representing the sequences. There are two ways to represent the motifs: consensus string and Position-specific Weight Matrices (PWM). Consensus string has the same length of DNA sequence motif; it allows to degenerate symbols in a string using IUPAC code while PWM is a matrix of 4xm where m is the motif length. Every position in the matrix represents the probability of each nucleotide at each index position of the motif. After motif representation, the suitable objective function is determined and finally appropriate search algorithm is applied. There are hundreds of algorithms for motif extraction that most of them are listed in table 1.

C. Post-processing

Post-processing evaluates the resultant motifs. This paper presents a more general classification of the sequence motifs extraction methods. Most of them are mentioned with a comparison among them.

Literature Review

There are two principal types of motif discovery algorithms: i.e. enumeration approach and probabilistic technique. Enumeration approach searches for consensus sequences; motifs are predicted based on the enumeration of words and computing word similarities so this approach is sometimes called the word enumeration approach to solve Panted (l, d) Motif Problem (PMP) with motif length (l) and a maximum number of mismatches (d). The algorithms based on the word enumeration approach exhaustively search the whole search space to determine which ones appear with possible substitutions and therefore it typically locates the global optimum. However, this also means that they are exponential-time algorithms that require a long time to detect the larger l and inefficient for handling dozens of sequences, so they are only suitable for short motifs. Moreover, these algorithms require many parameters determined by the users such as motif length, the number of mismatches allowed, and a minimum number of sequences that the motif has to appear in.

The word enumeration approach can be accelerated by using specialized data structures such as suffix trees or parallel processing. Popular algorithms based on this approach are DREME, CisFinder, Weeder, FMotif, and MCES.

A second group is a probabilistic approach. It constructs a probabilistic model called position-Specific Weight Matrix (PSWM) or motif matrix that specifies a distribution of bases for each position in TFBS to distinguish motifs vs. non-motifs and it requires few search parameters. The most popular methods based on probabilistic approach are MEME, STEME, EXTREME, AlignACE, and BioProspector.

Recently, new algorithms inspired from nature are presented that solve complex and dynamic problems with appropriate time and optimal cost. These algorithms simulate the behavior of insects or other animals for problem-solving. Evolutionary algorithms can overcome the disadvantages of local search and synthesize local search and global search. Examples of evolutionary algorithms are: Genetic Algorithm (GA), Genetic Programming (Special type of GA), Differential Evolution (DE), Evolution Strategy, Multi-modal Optimization, Cuckoo-Search (CS), Levy flight, Bacterial Colony Optimization, and Intelligent Water Drops algorithm.

Swarm intelligence is a special class of evolutionary algorithm including Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) algorithm, and Ant Colony Optimization (ACO) algorithm.

The beauty of nature-inspired algorithms is that they provide flexibility in evaluating the solutions by using fitness functions that score the solutions. These functions vary from problem to another and evaluate using different information types as biological information, functional information, etc. Moreover, these algorithms provide flexibility in motif representation.

Finally, the last category is a combinatorial algorithm that mixes multiple algorithms. The classification of motif discovery algorithms is shown in figure 2. This paper presents a classification of motif discovery algorithms and gives an overview of the most common algorithms with many examples; also, the main features while designing a new algorithm and future work are proposed.

A. Enumerative approach

The enumerative approach can be classified into many classes.

1. Simple word enumeration

The first class is based on simple word enumeration. Some existing algorithms in this class are YMF and DREME. Sinha et al developed YMF (Yeast Motif Finder) algorithm that detects short motifs with a small number of degenerate positions in yeast genomes using consensus representation. YMF enumerates all motifs in the search space approach and calculates the z-score to produce those motifs with greatest z-scores. Bailey et al proposed DREME (Discriminative Regular Expression Motif Elicitation) algorithm that also calculates the significance of motifs using Fisher’s Exact test. The algorithm starts with generating a set of short k-mers, followed by applying Fisher’s Exact test on two sets of DNA sequences (Input set and background set) using a significance threshold to calculate the significance of each word (No wildcards) of length three to eight that occurs in the positive sequences and select the 100 most significant words for being used in the inner loop where they passed as “seed” REs to perform a beam search that determines the most significant generalizations of them (One wildcard). To find multiple, non-redundant motifs in a set of sequences, outer loop determines the most significant motif using...
the heuristic search of RE motifs and the best motif found replaces its occurrences by a special letter; then the search process is repeated again many times until E-value of the new motif is less than the determined significance threshold.

2. Clustering-based method

Instead of using two loops for finding multiple motifs, the second class was proposed. Sharov et al. proposed word clustering method called CisFinder to detect short motif with high processing speed in large se-

---

**Table 1. Motif discovery algorithms**

<table>
<thead>
<tr>
<th>No.</th>
<th>Algorithm</th>
<th>Operating principle</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YMF</td>
<td>Simple word-based</td>
<td>(29)</td>
</tr>
<tr>
<td>2</td>
<td>DREME</td>
<td>Simple word-based</td>
<td>(9)</td>
</tr>
<tr>
<td>3</td>
<td>oligonucleotide analysis</td>
<td></td>
<td>(30)</td>
</tr>
<tr>
<td>4</td>
<td>CisFinder</td>
<td>Simple word-based with Clustering technique</td>
<td>(31)</td>
</tr>
<tr>
<td>5</td>
<td>By Thomas et al</td>
<td>Simple word-based with Clustering technique</td>
<td>(32)</td>
</tr>
<tr>
<td>6</td>
<td>POSMO</td>
<td>Tree-based</td>
<td>(7)</td>
</tr>
<tr>
<td>7</td>
<td>Weeder</td>
<td>Tree-based</td>
<td>(11)</td>
</tr>
<tr>
<td>8</td>
<td>FMotif</td>
<td>Tree-based</td>
<td>(33)</td>
</tr>
<tr>
<td>9</td>
<td>By G. Pavesi</td>
<td>Tree-based</td>
<td>(10)</td>
</tr>
<tr>
<td>10</td>
<td>MITRA</td>
<td>Tree-based</td>
<td>(34)</td>
</tr>
<tr>
<td>11</td>
<td>CENSUS</td>
<td>Tree-based</td>
<td>(35)</td>
</tr>
<tr>
<td>12</td>
<td>RISOTTO</td>
<td>Tree-based</td>
<td>(36)</td>
</tr>
<tr>
<td>13</td>
<td>SLI-REST</td>
<td>Tree-based</td>
<td>(37)</td>
</tr>
<tr>
<td>14</td>
<td>DRIMust</td>
<td>Tree-based</td>
<td>(38)</td>
</tr>
<tr>
<td>15</td>
<td>MCES</td>
<td>Tree-based</td>
<td>(12)</td>
</tr>
<tr>
<td>16</td>
<td>WINNOWER</td>
<td>Tree-based</td>
<td>(39)</td>
</tr>
<tr>
<td>17</td>
<td>Pruener</td>
<td>Tree-based</td>
<td>(40)</td>
</tr>
<tr>
<td>18</td>
<td>cWINNOWER</td>
<td>Tree-based</td>
<td>(41)</td>
</tr>
<tr>
<td>19</td>
<td>By Sze et al</td>
<td>Tree-based</td>
<td>(42)</td>
</tr>
<tr>
<td>20</td>
<td>RecMotif</td>
<td>Tree-based</td>
<td>(43)</td>
</tr>
<tr>
<td>21</td>
<td>ListMotif</td>
<td>Tree-based</td>
<td>(44)</td>
</tr>
<tr>
<td>22</td>
<td>TreeMotif</td>
<td>Tree-based</td>
<td>(45)</td>
</tr>
<tr>
<td>23</td>
<td>GWM</td>
<td>Tree-based</td>
<td>(46)</td>
</tr>
<tr>
<td>24</td>
<td>GWM2</td>
<td>Tree-based</td>
<td>(47)</td>
</tr>
<tr>
<td>25</td>
<td>Voting</td>
<td>Tree-based</td>
<td>(48)</td>
</tr>
<tr>
<td>26</td>
<td>PMS1</td>
<td>Tree-based</td>
<td>(49)</td>
</tr>
<tr>
<td>27</td>
<td>PMS2</td>
<td>Tree-based</td>
<td>(49)</td>
</tr>
<tr>
<td>28</td>
<td>PMS3</td>
<td>Tree-based</td>
<td>(49)</td>
</tr>
<tr>
<td>29</td>
<td>By Sze et al</td>
<td>Tree-based</td>
<td>(50)</td>
</tr>
<tr>
<td>30</td>
<td>PMSi</td>
<td>Tree-based</td>
<td>(51)</td>
</tr>
<tr>
<td>31</td>
<td>PMSP</td>
<td>Tree-based</td>
<td>(51)</td>
</tr>
<tr>
<td>32</td>
<td>Stemming</td>
<td>Tree-based</td>
<td>(52)</td>
</tr>
<tr>
<td>33</td>
<td>PMS4</td>
<td>Tree-based</td>
<td>(53)</td>
</tr>
<tr>
<td>34</td>
<td>PMS5</td>
<td>Tree-based</td>
<td>(54)</td>
</tr>
<tr>
<td>35</td>
<td>PMS6</td>
<td>Tree-based</td>
<td>(55)</td>
</tr>
<tr>
<td>36</td>
<td>ParMotif</td>
<td>Tree-based</td>
<td>(56)</td>
</tr>
<tr>
<td>37</td>
<td>zTriplet</td>
<td>Tree-based</td>
<td>(57)</td>
</tr>
<tr>
<td>38</td>
<td>PMSPPrune</td>
<td>Tree-based</td>
<td>(58)</td>
</tr>
<tr>
<td>39</td>
<td>Pampa</td>
<td>Tree-based</td>
<td>(59)</td>
</tr>
<tr>
<td>40</td>
<td>PMS3p</td>
<td>Tree-based</td>
<td>(60)</td>
</tr>
<tr>
<td>41</td>
<td>Provable</td>
<td>Fixed candidates</td>
<td>(61)</td>
</tr>
<tr>
<td>42</td>
<td>qPMSPrune</td>
<td>Modified candidate</td>
<td>(62)</td>
</tr>
<tr>
<td>43</td>
<td>qPMS7</td>
<td>Modified candidate</td>
<td>(62)</td>
</tr>
<tr>
<td>44</td>
<td>By Tanaka et al</td>
<td>Modified candidate</td>
<td>(63)</td>
</tr>
<tr>
<td>45</td>
<td>Random projection</td>
<td>Hashing</td>
<td>(64)</td>
</tr>
<tr>
<td>46</td>
<td>Uniform projection</td>
<td>Hashing</td>
<td>(65)</td>
</tr>
<tr>
<td>47</td>
<td>Low-dispersion projection</td>
<td>Hashing</td>
<td>(66)</td>
</tr>
<tr>
<td>48</td>
<td>MULTIPROFILER</td>
<td>Extended sample-driven (ESD)</td>
<td>(67)</td>
</tr>
<tr>
<td>49</td>
<td>Pattern Branching</td>
<td>Reference selection</td>
<td>(68)</td>
</tr>
<tr>
<td>50</td>
<td>Ref Select</td>
<td>Reference selection</td>
<td>(69)</td>
</tr>
</tbody>
</table>

**Enumerative approach**

1. YMF
2. DREME
3. oligonucleotide analysis
4. CisFinder
5. By Thomas et al
6. POSMO
7. Weeder
8. FMotif
9. By G. Pavesi
10. MITRA
11. CENSUS
12. RISOTTO
13. SLI-REST
14. DRIMust
15. MCES
16. WINNOWER
17. Pruener
18. cWINNOWER
19. By Sze et al
20. RecMotif
21. ListMotif
22. TreeMotif
23. GWM
24. GWM2
25. Voting
26. PMS1
27. PMS2
28. PMS3
29. By Sze et al
30. PMSi
31. PMSP
32. Stemming
33. PMS4
34. PMS5
35. PMS6
36. ParMotif
37. zTriplet
38. PMSPPrune
39. Pampa
40. PMS3p
41. Provable
42. qPMSPrune
43. qPMS7
44. By Tanaka et al
45. Random projection
46. Uniform projection
47. Low-dispersion projection
48. MULTIPROFILER
49. Pattern Branching
50. Ref Select

**Probabilistic approach**

51. MEME
52. STEMME
53. EXTREME
54. Profile Branching
55. APMotif
56. AlignACE
57. SPWDVM
58. By Lawrence et al
59. Motif- Sampler
60. BioProspector
Hashim FA, et al

Table 1. Count.

<table>
<thead>
<tr>
<th>No.</th>
<th>Algorithm</th>
<th>Operating principle</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>MITSU</td>
<td>Stochastic Expectation Maximization (sEM)</td>
<td>(74)</td>
</tr>
<tr>
<td>63</td>
<td>MCEMDA</td>
<td>(75)</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>SEAM</td>
<td>(76)</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>By Jensen et al</td>
<td>(13)</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>LOGOS</td>
<td>(77)</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>BaMM</td>
<td>(78)</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>By Jaaskinen et al</td>
<td>(79)</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>By Frith et al</td>
<td>(80)</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>SBaSeTraM</td>
<td>Bayesian approach</td>
<td>(81)</td>
</tr>
<tr>
<td>71</td>
<td>By Wakefield et al</td>
<td>(82)</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>MotifCut</td>
<td>(83)</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>MCL-WMR</td>
<td>Graphic based</td>
<td>(84)</td>
</tr>
<tr>
<td>74</td>
<td>EPP</td>
<td>Entropy-based position projection</td>
<td>(6)</td>
</tr>
<tr>
<td>75</td>
<td>CONSENSUS</td>
<td>Greedy Algorithm</td>
<td>(85)</td>
</tr>
<tr>
<td>76</td>
<td>By Huang et al</td>
<td>heuristic algorithm</td>
<td>(86)</td>
</tr>
<tr>
<td>77</td>
<td>St-GA</td>
<td>(87)</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>GAMI</td>
<td>(88)</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>FMGA</td>
<td>Simple GA</td>
<td>(89)</td>
</tr>
<tr>
<td>80</td>
<td>MDGA</td>
<td>(90)</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>By Paul et al</td>
<td>(91)</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>By Vijayvargiya et al</td>
<td>Clustering</td>
<td>(92)</td>
</tr>
<tr>
<td>83</td>
<td>By Gutierrez et al</td>
<td>(93)</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>GARPS</td>
<td>(94)</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>GAEM</td>
<td>(95)</td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>GADEEM</td>
<td>(96)</td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>CompareProspector</td>
<td>(97)</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>By Fatemeh et al</td>
<td>Hybrid</td>
<td>(98)</td>
</tr>
<tr>
<td>89</td>
<td>GEMFA</td>
<td>(99)</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>MRPGA</td>
<td>(100)</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>By Xiaochn et al</td>
<td>(101)</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>GAME</td>
<td>(19)</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>By Yetian et al</td>
<td>(102)</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>By Li et al</td>
<td>Others</td>
<td>(103)</td>
</tr>
<tr>
<td>95</td>
<td>MOGAMOD</td>
<td>(104)</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>PMbPSO</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>LFBS</td>
<td>Standard PSO</td>
<td>(105)</td>
</tr>
<tr>
<td>98</td>
<td>PSOMF</td>
<td>(106)</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>Lei et al</td>
<td>(107)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Lei et al</td>
<td>(108)</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>DSAPSO</td>
<td>Modified PSO</td>
<td>(109)</td>
</tr>
<tr>
<td>102</td>
<td>By Karabulut et al</td>
<td>(110)</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>Lei et al</td>
<td>(111)</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>Hardin et al</td>
<td>(112)</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>GSA-PSO</td>
<td>Hybrid</td>
<td>(113)</td>
</tr>
<tr>
<td>106</td>
<td>SPLO-Lk</td>
<td>(114)</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>Multiobjective ABC</td>
<td>(115)</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>MO-ABC/DE</td>
<td>(116)</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>Consensus ABC</td>
<td>ABC</td>
<td>(8)</td>
</tr>
<tr>
<td>110</td>
<td>Machhi et al</td>
<td>(117)</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>MFACO</td>
<td>ACO with Gibbs sampling</td>
<td>(118)</td>
</tr>
<tr>
<td>112</td>
<td>Cheng et al</td>
<td>ACO with EM</td>
<td>(119)</td>
</tr>
<tr>
<td>113</td>
<td>MACS</td>
<td>CS</td>
<td>(120)</td>
</tr>
<tr>
<td>114</td>
<td>STGEMS</td>
<td>Enumerative and probalistic approaches</td>
<td>(121)</td>
</tr>
<tr>
<td>115</td>
<td>MDIScan</td>
<td>Probabilistic and machine learning approaches</td>
<td>(122)</td>
</tr>
<tr>
<td>116</td>
<td>MUSA</td>
<td>Probabilistic and machine learning approaches</td>
<td>(123)</td>
</tr>
<tr>
<td>117</td>
<td>EMD</td>
<td>Multiple algorithms</td>
<td>(124)</td>
</tr>
<tr>
<td>118</td>
<td>MobyDick</td>
<td>(125)</td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>WordSpy</td>
<td>Dictionary</td>
<td>(126)</td>
</tr>
</tbody>
</table>

Table 1 continued...

quences (up to 50 Mb). Firstly, one should define nucleotide substitution matrix for each n-mer word, then calculate Position Frequency Matrices (PFMs) for n-mer word counts with and without gaps in both test and control sets. To generate non-redundant motifs, PFMs are extended over flanking and gap regions followed by means of clustering. Thomas et al \(^{31}\) extended the CisFinder technique to deal with whole ChIP-seq peak data sets.
3. Tree-based method
The third class is a tree-based search to accelerate the word enumeration technique. Pavesi et al. presented Weeder algorithm based on count matching patterns with specific and most extreme mismatches. At first, the motifs are represented using consensus sequence and based on the difference between the k-mers of the input sequences and the consensus under a limited number of substitutions, k-mers are assembled and each group is evaluated with a specific measure of significance. The constructed suffix tree was proposed in FMotif algorithm for finding long (l, d) motifs in large DNA sequences under the ZOMOPS (Zero, one or multiple occurrence(s) of the motif instance(s) per sequence) constraints. This proposed tree is faster than standard suffix tree as it avoids a large number of repeated scans of sequences on the suffix tree. Afterward, Qiang et al. proposed MCES algorithm for a PMP that used both suffix tree and parallel processing to deal with large datasets. MCES algorithm starts with mining step that constructs the Suffix Array (SA) and the Longest Common Prefix array (LCP) for the input datasets. At that point, combining step clusters substrings of various lengths to get predicted motifs.

4. Graph theoretic-based method
The graph-theoretic method represents a motif instance, as a clique; the graph G is built by representing each l-mer in the input sequences by vertex and the edge between a pair of vertices representing a pair of l-mer in different input sequences having the Hamming distance between the substrings which is less than or equal to 2d. Then, cliques of size N are searched for in this graph. Popular graph-theoretic methods are WINNOWER, Pruner, and cWINNOWER.

5. Hashing-based method
Buhler et al. developed random projection algorithm for a PMP that projects every l-mer in the input data into a smaller space by hashing. Initially, a projection of l-dimensional space onto a k-dimensional subspace for all subsequences in the input set is developed, and random projection is constructed by choosing random k positions from l position. Using this projection, each l-mer is hashed to its corresponding bucket. After projections, each bucket contains l-mer more than a threshold and this is called qualified bucket. Random hashing is repeated n times to ensure the qualified bucket at least more than once. Finally, profile for each of them should be computed to get the most probable l-mer in the sequence that was represented as consensus sequences. In previous studies, random projection was developed using uniform projection and low-dispersion projection algorithms, respectively.

6. Fixed candidates and modified candidate-based methods
The sixth class is fixed candidates that select candidate motifs from input sequences and use them for motif scanning while the seventh class is modified candidate that selects one candidate from the input sequence and modifies it letter by letter. Finally, there is a proposed algorithm called Ref-Select to select reference sequences for PMP. The reference sequences are the sequences that don’t contain motif instances, so, this method tries to select the reference sequences that generate a small number of candidate motifs as possible. The algorithm consists of two steps; firstly, for every two sequences in input dataset D, the number of candidate motifs generated from them should be computed using the Hamming distance between every two l-mers. Then, the set with candidate
motifs, as small as possible, is selected as a reference set.

B. Probabilistic approach

Deterministic approach: Expectation-Maximization (EM) is the famous example of deterministic approach. EM for motif finding was first introduced by Lawrence et al. and consists of two main steps, the first called "Expectation step" that estimates the values of some set of unknowns based on a set of parameters. The second step is "Maximization step" that uses those estimated values to refine the parameters over several iterations. EM is used to identify conserved areas in unaligned DNA and proteins with an assumption that each sequence must contain one common site, the parameters; in this case, they are the entries in the PWM and the background nucleotide probabilities while our unknowns are the scores for each possible motif position in all of the sequences.

There are several algorithms based on EM. MEME (Multiple EM for Motif Elicitation) is a popular motif recognition program that optimizes PWMs using the EM algorithm. It has several versions. The idea of MEME algorithm is to find an initial motif and then use expectation and maximization steps to improve the motif until the values in the PWM do not improve or the maximum number of iterations is reached. The MEME algorithm starts from a single site, i.e. k-mer (Random or specified) and estimates motif model (PWM). For every possible location in every input sequence, the probability, given the PWM model, should be identified to detect examples of the model; then, the motif model should be re-estimated by calculating a new PWM. EM alternates between examples of the model step and re-estimates the motif model step. A single iteration for each k-mer in target sequences should be performed and the best motif from this site needs to be selected and then only the one to converge should be iterated. The algorithm searches for new motifs after erasing the old discovered motif. It defines all three types of motif discovery sequence model: OOPS, ZOOPS, and TCMs corresponding to one occurrence per sequence, zero or one occurrence per sequence, and zero or more occurrences per sequence, respectively. Reid et al. presented STEME (Suffix Tree EM for Motif Elicitation) algorithm to accelerate the MEME algorithm and the first application of suffix trees to EM algorithm was considered. Quang et al. also tried to accelerate the MEME algorithm by developing EXTREME (Online EM algorithm for motif discovery) algorithm. EXTREME is an online web server that implements the MEME algorithm and can work on a large dataset without discarding any sequences.

Stochastic approach

Gibbs sampling is a famous stochastic approach, similar to EM algorithm. Pseudocode of the Gibbs sampling algorithm for motif detection follows these steps:

- Random initializing of motif positions in the input N sequences with an assumption of the presence of one motif per sequence,
- Choosing one sequence at random,
- Computing PWM for the other N-1 sequences using starting positions of motifs and background probabilities for each base using the non-motif positions,
- Calculating probability of each possible motif location in the removed sequence using PWM and background probabilities,
- For the removed sequence, choosing a new starting position based on step 4.

Steps 2-5 should be iterated until the values in the PWM do not improve or the maximum number of iterations has been reached.

Many methods have been developed that implement the concept of Gibbs’ sampling to extend its functionality. Hughes et al. proposed Align ACE (Aligns Nucleic Acid Conserved Elements) algorithm based on Gibbs’ sampling with some improvements: (1) The motif model was changed to fit the source genome because the base frequencies for non-site sequence is fixed, (2) Both strands of the input sequence are considered and no circumstance overlapping is allowed, (3) Iteratively, aligned sites were masked out to find multiple different motifs, and (4) It uses an improved near-optimum sampling method.

BioProspector algorithm is also based on Gibbs’ sampling with several improvements: (1) It uses a Markov model estimated from all promoter noncoding sequences to represent the non-motif background in order to improve the motif specificity, (2) It can find two-block motifs with variable gap, and (3) Sampling with two thresholds allows every input sequence to include zero or multiple copies of the motif.

3. Advanced approach

Different algorithms were proposed based on Bayesian approach. Jensen et al. proposed an algorithm based on Bayesian approach with Markov chain Monte Carlo. Xing et al. proposed LOGOS (Integrated Local and Global motif sequence model) algorithm that combines between HMDM (Hidden Markov Dirichlet-Multinomial) for local alignment model for each different motif and HMM (Hidden Markov model) for global motif distribution model for the occurrence of multiple motifs.

Recently, Siebert et al. developed a Bayesian Markov Model (BaMM) approach that trains higher order Markov models to build the dependency model. BaMM algorithm is more complex than PWMs wherein the PWMS cannot model correlations among nucleotides because PWMS nucleotide probabilities are independent of nucleotides at other positions. In the proposed algorithm, Bayesian approach using Markov models makes optimal use of the available information while avoids training by the decrease in number of parameters.
Other different algorithms 79,80 were presented like clustering methods based on Bayesian approach.

4. Others

EPP (Entropy-based position projection) 6 algorithm was proposed to escape from local optima. This algorithm is based on projection process depends on the relative entropy in each position of motif instead of random projection.

C. Nature-inspired algorithms

Nature-inspired algorithms are classified according to the sources of inspiration into three main categories 138. They are swarm-intelligence, non swarm-intelligence and physics/chemistry as shown in figure 2. Popular evolutionary algorithms used in motif discovery are GA and PSO and few of them are ABC, CS, and ACO.

1. GA

GA is a probabilistic optimization algorithm based on evolutionary computing. GA is inspired from biological evolution processes like selection, crossover, and mutation. The motivation for using GA comes from the idea of reducing the number of searches in a high number of DNA sequences. The basic structure of GA consists of a population of candidate solutions throughout several generations to find the best solution or set of possible solutions. The algorithm starts with the random generation of individuals that are then evaluated by a fitness function. At that point, a selection process selects new individuals called offspring that have some features of the parents and the others are discarded, then the genetic operators are applied on offspring. The parents are selected based on a probabilistic process biased by their fitness value using specified selection techniques that keep the diversity of the population large and prevent premature convergence on poor solutions. The most common technique for parent selection is the roulette-wheel method. In a roulette wheel selection, based on the number of individuals say n, the circular wheel is divided into n pies. The size of each pie is proportional to the fitness of the element. The roulette wheel is spun randomly and the element where it stops is chosen as the parent. The selected two individuals are used to produce offspring; this means the elements of higher chance for selection have higher fitness.

After two selections of parents, a crossover is applied to select a random site, and the rightmost strings are swapped to produce new children, followed by applying the mutation process by changing the value of some selected position. The process is iteratively repeated until some stop criterion is reached on the satisfactory fitness level 139. Some methods 89,90 use standard GA. Liu et al 89 introduced FMGA algorithm that is based on simple GA. The genetic operators were specialized for motif discovery problem; the mutation operator was applied using PWM to reserve the completely conserved positions and the crossover operator was implemented with specially-designed gap penalties that optimize a scoring function. Che et al 90 proposed MDGA algorithm that is also based on simple GA. Other methods were proposed and most of them 91-93 used population clustering technique that partitions population into subpopulations before mating. Clustering is done by the similarity between solutions using data clustering algorithms.

Using population clustering technique, Paul et al 91 proposed an algorithm for PMP to detect multiple and weak motifs. Firstly, population initialization by random selecting of subsequences of motif length used to form a candidate consensus motif is done and then all input sequences are scanned to detect all similar sub-strings followed by sorting them according to a number of mismatches of each substring from the candidate motif. Next, scoring function is applied on them. This method was extended to detect weak motifs using alignment score metric and clustering technique. Finally, fitness of an individual (Cluster) is calculated to select the parents for using in GA. Vijayvargiya et al 92 proposed an algorithm with the position based representations of individual and clustering of population scheme. Gutierrez et al 93 proposed a new statistical GA that mixes GA structure with several statistical coefficients. In this method, all the input sequences are joined in a single super-sequence and the individuals are represented by a single position value; at that point, the super-sequence is separated into subsequences of any length regardless of the length of each sequence. For each generation of the population, the fitness value will be calculated against one of the subsequences. The fitness function is a combination of different functions applied at different moments of the process. The algorithm is started by checking if they are overrepresented in the given subsequence by calculating the difference of similar words between the candidate motif and the background motif. Then, the selection process is applied consisting of two steps; the first is using the Fluffiness Coefficient by the mean and standard deviation of the simple overrepresentation values and the individuals with the lowest Fluffiness value are eliminated from the population. The second step is using different coefficients to decide if the candidates are the final solutions of the problem or not for an individual has survived for at least 10 generations. The first coefficient is Thinness coefficient where the individuals with a Thinness coefficient above 0.6 are eliminated from the population. Mann-Whitney is the second coefficient for two populations to quantify, if one of them has a tendency to have larger values than the other. The first sample corresponds to the vector with the values stored for the number of similar words for the candidate motif in each generation, and the second sample is formed by the same values for the background motif. If the probability of both data samples coming from the same population is lower than 0.05, at that point, the motif is considered as a possible final solution. The
third step is the creation of offspring by applying genetic operators, one-point crossover, and random mutation, on the selected parents. Finally, filtering and clustering of solutions is another method for every given motif width to generate the final solutions.

Some methods enhance the GA by using hybrid methods that combine GA with another technique \(^{94, 96, 140}\) or propose additional operators in addition to basic genetic operators \(^{19, 102}\).

Huo et al \(^{94}\) proposed a new algorithm (GARP) that optimizes GA based on the random projection strategy (RPS) to identify planted \((l, d)\) motifs. The idea behind using RPS before GA is to find good starting positions for being used in simple GA as an initial population instead of random population. Wang et al \(^{95}\) presented GAEM algorithm that combines GA and EM for planted edited \((l, d)\)-motif finding problem. In planted edited \((l, d)\) -motif finding problem, the mutation in motifs due to substitution, deletion, and insertion, causes the motifs to have different lengths ranging from \(l-d\) to \(l+d\).

The EM algorithm is used after the random initial population to get the best starting positions to be used as a seed to GA. Wei et al \(^{19}\) introduced GAME (Genetic Algorithm for Motif Elicitation) that proposes two operators called ADJUST and SHIFT to escape from local optima. The ADJUST operator escapes from a local optimum that occurs if any of the motif sites has not been aligned correctly by checking every possible site position in a sequence and choosing the best match to the sites in the other sequences. But SHIFT operator escapes from local optima that occur when all motif sites are slightly misaligned by shifting the subsequences in the direction which gives the best fitness. Yetian et al \(^{102}\) also presented a new operator called addition to the standard GA operators. The algorithm started with a motif length of three and then the proposed operator is used until the length of the optimal motif reaches to the standard level. The result achieves a higher score than the other three methods: Gibbs Sampler \(^{141}\), GA \(^{142}\) and GARPS \(^{94}\) algorithm.

2. PSO

PSO is a new global optimization technique \(^{143}\) for solving continuous optimization problems. PSO algorithm is characterized by its simple computations and information sharing within the algorithm. It simulates the social behaviors of organisms movements in flocks of birds or schools of fish to find food sources and defenses against predators. Each particle uses its own flying experience and flying experience of other particles to adjust its "flying" so it combines self-experiences with social experiences. In self-experiences, the particle tries to get local best particle position and this is done by the particle stores. The best solution visited so far in its memory is called pbest, and it has an attraction towards this solution as it navigates through the solution search space. Social experiences are utilized to get best global particle position through the particle stores and the best solution visited by any particle and attraction towards this solution is called gbest.

For an \(n\)-dimensional search space, the position and velocity of the \(i\)th particle are represented by \(Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})\) and \(V_i = (v_{i1}, v_{i2}, v_{i3}, …, v_{in})\), respectively. The previous best position is denoted as \(P_i = (p_{i1}, p_{i2}, \ldots, p_{in})\). \(Pg\) is the global best particle in the swarm. For the swarm \(S\), the new velocity of each particle is calculated according to the following equation:

\[
V_i(t+1) = v_i(t) + c1 r1 (p_{i} - y_i) + c2 r2 (p_{g} - y_i)
\]

(1) The position is updated using:

\[
y_i(t+1) = y_i(t) + V_i(t+1)
\]

(2) Where \(i = 1, 2, \ldots, S\) represents the particle index and \(n=1, 2, \ldots, N\) represents the dimension. \(c1\) and \(c2\) are cognitive and social scaling parameters, respectively.

At each iteration, pbest and gbest are updated for every particle as per their fitness values. The procedure is iteratively repeated until some stop criterion is reached or satisfactory fitness level has been reached.

PSO has wide applications and has been proven to be effective in motif finding problems \(^{112}\). In recent years, there are few numbers of researches which utilized PSO to solve different types of motif finding problems. Chang et al \(^{144}\) proposed a modified PSO algorithm where the position and velocity are adjusted to escape from local optima. Hardin et al \(^{112}\) proposed a hybrid motif discovery approach based upon a combination of PSO and EM algorithm and PSO was utilized to generate a seed for the EM algorithm. In previous studies, modified PSO algorithm was proposed based on word dissimilarity graph \(^{107, 108}\). Modified PSO algorithm began to break all input sequences into \(l\)-mers and develop a novel mapping scheme that was used in the fitness function. Each particle kept track of a vector of locations in each given sequence and formed a consensus sequence. Then, the updated policy of PSO was modified where the new and current motif positions must be in the upper and lower bounds of the velocity. To escape from local optima, the algorithms scan all input sequences after gbest value reaches to a certain threshold to check if \(l\)-mer has a fitness value in comparison to gbest. Also, the method used "reset move" that moves all the current solution, pbest and gbest by a random distance. Finally, the termination criteria are determined automatically using repeat -based method. Reddy et al \(^{4}\) developed PMbPSO (PSO-based algorithm for Planted Motif Finding) algorithm. PMbPSO algorithm selects initial positions for all motifs by random and generates ten children for each parent (Motif) and then computes the fitness function for each parent and its children to get the best position; at that point, the best position from all particles is got followed by updating velocity and position for each particle for a number of iterations. Haruna et al \(^{105}\) integrate Linear-PSO with binary search technique (LPBS) to minimize the execution time and increase the validity in motif discovery of DNA sequence for specific species. LPBS algorithm starts with initializing the population by selecting the target motif from the reference set and
searching for similar motifs using the binary search, then applying the standard PSO algorithm to discover motifs. Recently, Etehkal et al. introduced a hybrid GSA-PSO algorithm that combines local search capabilities of the Gravitational Search Algorithm (GSA) and global search capabilities of PSO algorithm.

3. **ABC algorithm**

ABC algorithm is a type of swarm-based algorithm proposed by Karaboga. It simulates the behavior of honey bees to find a food source. Two fundamental properties to obtain swarm intelligent behavior in honey bee colonies are self-organizing and division of labor. The bee colony contains two main groups which are employed and unemployed foragers. Employed bees are going to the food source which is visited previously and they are responsible for giving information to unemployed foragers about the quality of the assigned nectar supply. Many factors determine the value of a food source like its distance from the nest, its energy concentration and the degree of difficulty to extract this energy. Unemployed bees are categorized to scout and onlooker bees. Scout bee searches around the nest randomly to find new food sources while onlooker bee uses the information shared by employed foragers to establish a food source.

Employed bees’ numbers are the same for food sources’ numbers around the hive. At first, scout bees initialize all positions of food sources that represent possible solutions to the problem. Then, employed bees and onlooker bees exploit the nectar of food sources that corresponds to the quality (Fitness) of the associated solution, and this continual exploitation will finally cause them to become exhausted. The employed bee which turned into exploiting the exhausted food source turns into a scout bee looking for other food sources. The ABC algorithm involves four fundamental phases: A neighbor food source FS is determined and calculated by the following equation:

\[ FS_i = f_i + \text{rand}(-1, 1) \ast (f_i - f_{best}) \]  

Where \( i \) is a randomly selected parameter index, \( f_i \) is a randomly selected food source, \( \text{rand}(-1, 1) \) is a random number between -1, 1.

The fitness of this new food source is calculated and a greedy selection is applied between it and its parent i.e. between FS and F. From that point, employed bees share information about their food source via dancing on the dancing area with onlooker bees waiting inside the hive.

Gonzalez-Alvarez et al. developed multiobjective ABC algorithm that aims to adapt the ABC algorithm to multiobjective context. The multiobjective optimizes more than one objective function at the same time to get a set of optimal solutions known as Pareto set. This algorithm defines three conflicting objectives as motif length, support, and similarity and multiobjective adaptations of ABC including multi-term fitness function, ranking, and sorting methodology are used. González et al. proposed Multiobjective Artificial Bee Colony with Differential Evolution algorithm (MO-ABC/DE) that combines the general schema of ABC with Differential Evolution. Recently, Karaboga et al. proposed consensus ABC algorithm. This is a discrete model based on a similarity value between consensus sequences of the ABC algorithm. The algorithm starts with random initialization, and then calculates the similarity value of the resultant consensus sequence followed by a new neighborhood selection method that is based totally on the similarity values of the consensus sequences.

4. **ACO algorithm**

The ACO algorithm is a metaheuristic optimization technique that mimics the behavior of real ants, which try to find the shortest path to the food from their nest. The ants explore randomly the area surrounding their nest, and while moving, they leave a chemical pheromone trail on the ground that helps them to go to the nest. Ants interact with each other through this chemical component. The quantity of pheromone is proportional to the quantity and the quality of the food and this pheromone will be guided to other ants for the food source. When evaporation occurs, it reduces the attractive strength of pheromone. The evaporation takes a long time in the shorter path than the longest. The ants choose their way with strong pheromone concentrations. The characteristics of ACO algorithm are: (1) It depends on two variables including the amount and the evaporation of pheromone. The amount of pheromone is directly proportional to the richness of the food and inversely to evaporation: evaporation factor avoids the convergence to a locally optimal solution. (2) Ants act concurrently and independently, and (3) The behavior is stigmergy i.e. the interaction between ants is indirect. (4) Ants can explore vast areas without global view of the ground. (5) Starting point is selected at random.

The pheromone probability in terms of motif discovery is given by:

\[
P_c(\mathbf{ic}) = \frac{(\omega_{ic}(t))^{\beta} (\eta_{ic})^\alpha}{\sum_{\forall \mathbf{ic} : \mathbf{c} \in \mathbf{ic}} (\omega_{ic}(t))^{\beta} (\eta_{ic})^\alpha} 
\]

Where \( P_c(\mathbf{ic}) \) is the probability of ant \( a \) choosing \( c \) in position \( l \), \( \eta_{ic} \) is the heuristic information to measure the frequency of letter \( c \) in input sequences, i.e. the weight of the letter \( c \), \( \beta \) and \( \alpha \) represent the influence of the pheromone trails, \( c \) is the character set of input sequences, \( \omega_{ic}(t) \) is the amount of pheromone on the character \( c \) at position \( l \) at time \( t \), and \( \omega_{ic}(t) \) is the amount of pheromone on neighborhood at position \( l \) which ant \( a \) has not visited yet at time \( t \). An updated pheromone trial is:

\[
\omega_{ic}^+ (t + 1) = (1 - \rho) \cdot \omega_{ic}(t) + \sum_{k=1}^n \Delta \omega_{ic}^k(t) 
\]

Where \( \omega_{ic} \) is the total number of ants, which carry the character \( c \) at position \( l \), \( \rho \) is the rate of the pheromone trails evaporation (0<\( \rho \)<1), and \( \Delta \omega_{ic}^k(t) \) is the variable of pheromone deposited by kth ant on the character \( c \) at the hive.

This algorithm defines three conflicting objectives as motif length, support, and similarity and multiobjective adaptations of ABC including multi-term fitness function, ranking, and sorting methodology are used. González et al. proposed Multiobjective Artificial Bee Colony (MO-ABC/DE) that combines the general schema of ABC with Differential Evolution. Recently, Karaboga et al. proposed consensus ABC algorithm. This is a discrete model based on a similarity value between consensus sequences of the ABC algorithm. The algorithm starts with random initialization, and then calculates the similarity value of the resultant consensus sequence followed by a new neighborhood selection method that is based totally on the similarity values of the consensus sequences.
Position I. A lot of methods used ACO algorithm in motif discovery. Machhi et al. 109 used ACO algorithm with a Gibbs sampling algorithm. An ACO algorithm finds better starting positions of the sequences provided as starting position for the Gibbs sampler method instead of random initialization. This algorithm starts with each ant choosing the path to construct a sample with motif length (m) and that depends on pheromone probability. Then, each ant is compared between the selected sample (m) and each substring in input sequences to get the set that represents the best matching substrings. Next, fitness function for each selected set is calculated. After that, the amount of pheromone is updated and finally iterated until no change.

5. CS algorithm

CS is a new simple heuristic search algorithm that is more efficient than GA and PSO 116-118. CS is inspired from brood parasitism reproduction behavior of some cuckoo species in combination with Lévy flight behavior 149. The cuckoos lay their eggs in nests of the other birds with the abilities of selecting the latest spawned nests and removing existing eggs to increase the hatching probability of their eggs. If host birds discover these eggs, they either throw them away or abandon the nest and build a new nest. CS is characterized by the subsequent rules:

Each cuckoo lays one egg at a time and disposes its egg in a random selected nest.

The nests that have high quality of eggs (Solutions) are the best and will continue to the following generations.

The number of accessible host nests is fixed, and a host bird can discover a parasitic egg with a probability $P_a \epsilon [0,1]$.

To simulate the behavior of cuckoo reproduction, each egg in a nest is a solution and each cuckoo’s egg is a new solution. The aim is to supplant a not-so good solution in the nests with newer and better solutions by Lévy flights:

$$y_{i}^{(t+1)} = y_{i}^{(t)} + \alpha \times \text{Levy}(\lambda)$$ \hspace{1cm} (6)

Where $y_{i}^{(t+1)}$ is a new solution, $y_{i}^{(t)}$ is the current location, $\alpha$ is the step size and Levy($\lambda$) is the transition probability or random walk based on the Lévy flights. CS is an effective global optimization algorithm and has many applications in different fields 150. Etehadi et al. 120 applied CS and Modified Adaptive Cuckoo Search (MACS) algorithm on PMP. The MACS algorithm enhances the basic CS algorithm by grouping parallel, incentive, information and adaptive strategies.

### DNA Motif Databases

#### Synthetic DNA sequences

The simulated data set was first introduced by Pevzner et al. 10 to create planted (l, d) motif in n sequences selected randomly from a set of N sequences where $n \leq N$. There are online sites 151,152 to automatically create planted (l, d) motif.

#### Real DNA sequences

There are several online databases of DNA motifs listed in table 2 with a short description of each one.

### Discussion

As sequencing technology has improved, the volume of biological sequence data in public databases increases and this increases the importance of motif discovery in computer science and molecular biology 153. Motif discovery has some difficulties: (1) Motifs are not identical to each other, (2) The motif sequence is unknown, (3) Motif location is unknown, (4) Existing of random motifs, and (5) The location of a motif in each sequence is unrelated to other ones.

The motif discovery algorithms are classified into two major groups as enumerative approach and probabilistic approach. As the name of the first class imparts, it is counting and comparing oligonucleotide frequencies for all possible motifs, based on specific motif model description. It has some advantages: (1) Global optimality, (2) It is appropriate for short motifs; therefore, it is useful for motif finding in eukaryotic genomes, (3) With optimized data structures, it becomes
fast, and (4) Can find totally constrained motifs. However, the problems of this approach are: (1) It needs to be post-processed with some clustering systems as the typical transcription factor motifs often have several weak constrained positions, (2) It suffers from the problem of producing too many spurious motifs, (3) It represents certain numbers of motifs (11 wild cards), and (4) Long time processing is another problem as it checks every possible substring in the input dataset.

There are many algorithms based on sub-categories of this approach. The first sub-category is based on the enumerative approach without any enhancement; this leads to getting all possible motifs but at the same time there are many disadvantages which were mentioned above. YMF is designed for yeast genomes, it can’t detect motifs with large length or the number of degenerate positions is significant. DREME is a discriminative motif discovery tool to discover multiple, short, non-redundant, statistically significant motifs in short runtime using simplified form of regular expression words (11 wildcard characters). DREME algorithm was tested on ChIP-seq datasets [13 mouse ES Cell (mESC), 3 mouse erythrocytes and one human cell line (ChIP-seq datasets)]. DREME is compared to MEME algorithm and the results show that DREME algorithm can correctly predict motifs on ChIPseq experiment sequences in a shorter runtime than MEME. However, it can only find short motifs (from 4 to 8 bp) and an occurrence of the motif is well defined, as is the number of possible motifs of a given width.

The second sub-category is based on simple enumerative approach but it can discover multiple and weak motifs at the same time so, it is considered as the small enhancement of simple-based method. CisFinder technique tested on ChIP-seq data of TFs was expressed in ES cells. CisFinder can accurately identify PFM of TF binding motifs and it is faster than MEME. CisFinder can find motifs with a low level of enrichment, but it does not support outputting motifs of a specified length.

The algorithm proposed is similar to CisFinder algorithm, but it supports outputting motifs of a specified length. They reported that this algorithm combined accuracy and low computational time, but it is limited to short (l, d) motifs.

The first two sub-categories have nearly the same disadvantages, but they are time-consuming. There are many suggestions to improve this problem; in tree-based methods, it is also based on enumerating all possible motifs but it is time-consuming using suffix tree. Weeder algorithm accelerates the word enumeration technique by using a suffix tree, but it operates with a low efficiency for long motifs. FMotif algorithm could identify unknown motif lengths on ChIP-enriched regions. Next, MCES algorithm is a more powerful algorithm and there are two contributions in the mining step; it uses an adaptive frequency threshold for each possible length and it is based on MapReduce strategy to deal well with very large datasets. The MCES is tested on simulated data and real data (ChIP-seq) and the results show that MCES can find the motif like the published one and run in a short time. MCES has many advantages like identifying motifs without OOPS constraint, handling very large data sets, handling the full-size input sequences and making full use of the information contained in the sequences, completing the computation with a good time performance and good identification accuracy.

Graph-based techniques are the same simple-based techniques but they represent the motif-instance by a graph to facilitate the search strategy.

There are many search strategies to enhance time operator.

Sub categories from five to seven try to enhance time operator using random concept whether random searching about the motif like hashing strategy or selecting random candidate motif or many candidate motifs.

However, the random projection algorithm takes long time operations as it depends on random initialization and it repeats the process for n times.

In fixed candidates and modified candidate-based techniques, the technique scans all input sequences to get the matched motifs.

In the probabilistic approach, the probability of each nucleotide base to be present in that position of the sequence is multiplied to yield the probability of the sequence. PWM is an appealing model due to its simplicity and wide application and it can represent an infinite number of motifs but it has some problems. (1) It scales poorly with dataset size, (2) PWM representation assumes the independence of each position within a binding site, while this may be not true in reality, and (3) It converges to locally optimal solution.

EM algorithm is a popular example of probabilistic approach, but it has some limitations: (1) It converges to a local maximum, (2) It is extremely sensitive to initial conditions, (3) It assumes one motif per sequence, and (4) The running time of EM is linear with the length of the input sequences. MEME added several extensions to overcome these limitations where MEME runs the EM algorithm many times from different starting points using every existing 1-mer in the sequence dataset. However, the running time scales poorly with the size of the dataset so, MEME algorithm can discover motifs in a satisfactory time by a compromising strategy such as discarding a majority of the sequences but discarding data is far from ideal as it can decrease the chance of discovering motifs corresponding to infrequent cofactors. STEME runs faster than the MEME algorithm, but with a large dataset, it finds motifs up to width 8 as its efficiency decreases quickly as the motif width increases. EXTREME achieves better running time, but at the same time, it requires too much storage space for processing large data. Gibbs sampling algorithm is another example of a probabilis-
tic approach. It converges to local optimum, and is less dependent on initial parameters, but more dependent on all sequences exhibiting the motif. Advanced methods based on Bayesian technique are a subclass of probabilistic approach; examples of this class are the speedy algorithms with better objective function and BaMM algorithm \(^{13}\). BaMM was tested on 446 human ChIP-seq datasets and the results show that the precision increases by 30-40% compared to PWM. However, BaMM algorithm has some limitation: (1) It uses EM algorithm which has some disadvantages as described above, (2) The running time is longer than EM algorithm which makes it difficult to apply on a big data set, (3) It defines ZOOPS model only, and (4) It requires motif length as an input parameter. Finally, EPP algorithm can be applied to OOPS, ZOOPS, and TCM sequence models and the results indicate that it can efficiently and effectively recognize motifs.

Evolutionary algorithms have been recognized due to their advantages of synthesizing local search and global search \(^{91}\). GA and PSO are the famous evolutionary algorithms. GA is a discrete technique, but PSO is a continuous technique that must be modified to handle discrete design variables. GA and PSO are similar in the presence of interaction between the population’ members, but unlike GA that changes the population from generation to another, PSO keeps the same population; moreover, PSO does not have genetic operators and no notion of the “survival of the fittest”.

Based on GA, a lot of algorithms are proposed. The simple GA favors selection of the fittest, which may be a biologically meaningless solution and this tends to remove the diversity of the population. Simple GA identifies OOPS model only and ignores ZOOPS and TCM models which are not correct because some sequences contain multiple or other weak motifs that also need to be identified \(^{157}\). FMGA and MDGA algorithms are examples on simple GA. FMGA performs better than MEME and Gibbs sampler algorithms. MDGA algorithm is compared with a Gibbs sampling algorithm when tested on real datasets \(^{158}\) and the results showed that it achieves higher accuracy in short computation time; the computation time does not explicitly depend on the sequence length. To overcome simple GA problems, clustering approach was presented. Clustering scheme enables to retain the diversity of population over the generations and it can find various motifs. Based on clustering technique, a new scoring function was developed that takes some consideration like, the number of mutations, and the number of motifs per sequence \(^{91}\). The authors reported that it gives effective result when it’s applied to simulated and real data. Previous methods presented by Vijayavargiya et al \(^{83}\) could identify multiple motifs of the same length and discover long motifs when tested on synthetic and real datasets \(^{92}\). Gutierrez et al \(^{93}\) presented a new algorithm that has many advantages; there are no assumptions about the presence of the motifs in the input sequences, it is a heuristic algorithm, variable-length gaps can be predicted, and the background set of sequences is generated by shuffling the candidate motifs instead of shuffling the sequences themselves. The algorithm was tested on 52 data sets of four different organisms and the efficiency was compared with 14 methods using eight statistics. The proposed method gives its best results with fly and mouse data sets. However, when connecting all input sequences into one sequence and then dividing it, it may cause loss of motifs. Based on hybrid methods, GARPS algorithm is better than the projection algorithm when tested on both simulated and real biological data, while GAEM algorithm tested on a simulated dataset, gives a success rate higher than HIGEDA \(^{140}\) and when tested on a real dataset, it performs better than HIGEDA, GLAM2, and MEME. Finally, based on algorithms that enhance genetic operators, the GAME algorithm performs better than MEME and BioProspector when applied on simulation and real-data; the algorithm presented by Fan et al \(^{102}\) has a higher score than the other three methods of Gibbs Sampler \(^{141}\), GA \(^{142}\), and GAR-PS \(^{94}\).

From the proposed algorithms based on GA, it can be concluded that the methods presented previously \(^{19,90,92,94,102}\) used simple fitness function although there are a lot of suggestions to improve this function. Though the methods proposed by Paul et al, Vijayavargiya et al, and Gutierrez et al \(^{91,93}\) can identify ZOOPS model, they cannot identify multiple motifs of variable lengths instead of multiple motifs of the same lengths. Methods by Paul et al and Vijayavargiya et al \(^{91,92}\) ignore the TCM model, while a method by Gutierrez et al \(^{93}\) identifies it. The methods of Paul et al, Vijayavargiya et al, and Gutierrez et al \(^{91,93}\) enhance the selection strategy by proposing new fitness function. The methods proposed by Huo et al and Wang et al \(^{94,95}\) enhance the GA algorithm by combining it with another algorithm to get the starting positions to be used as seed to the GA algorithm, but this increases the computational time also. The selection of EM in the method by Wang et al \(^{95}\) is a bad choice as EM has many limitations as mentioned above. The proposed methods by Wei et al and Fan et al \(^{19,102}\) try to enhance the GA algorithm in another aspect by adding operators to escape from local optimum, but they also didn’t overcome the limitations of the simple GA algorithm. All discussed methods based on the GA algorithm require some parameters determined by the user as motif length. It can be said that the GA algorithm can be enhanced by using a new method that can identify OOPS, ZOOPS, and TCM models, escape from local optimum, improve the fitness function, have good starting positions instead of random initialization, detect multiple motifs with variable lengths, and have intelligent operators in addition to selection, crossover and mutation operators.

PSO is an exploration-exploitation trade off. Exploration is the ability to get the global optimum by search.
in various regions while the exploitation is the ability to locate the optimum by concentrating the search around a promising candidate solution \cite{159}, PSO is a simple concept, easily programmable \cite{160}, but it can easily fall in a local optimum and low convergence rate \cite{161}.

Based on PSO algorithm, Chang \textit{et al} \cite{144} obtained global optimum in protein sequences when the results were compared with the PROSITE database while Hardin \textit{et al}'s algorithm \cite{112} suffers from local solution. The methods in Lei \textit{et al} \cite{107,108} were tested on both simulated (PMP) and real biological data (\textit{E. coli}); they are efficient and accurate in motif discovery, but they suffer from a long time delay due to full scan on all sequences to check the value of gbest and the repeat-based method was used to automatically terminate the program. PMbPSO algorithm is tested on simulated and real biological datasets \cite{162}, the results indicate that PMbPSO algorithm performs better than MbGA and PbGA \cite{163} and it is able to find longer size motifs with a minimum number of mismatches. LPBS algorithm searches for motifs based on reference set. LPBS was tested on Genbank (COI) and there were only two selected species; BosTaurus (Cow-10 DNA sequences with lengths between 658 and 715 bp) and GallusGallus (Chicken-9 DNA sequences with lengths between 537 and 699 bp). The longest motifs which are able to discover LPBS algorithm are 294 and 261 bp in Bos-Taurus and GallusGallus, respectively. The authors did not evaluate the LPBS method with other previous algorithms of motif discovery. GSA-PSO algorithm was tested on synthetic \cite{164} and real data (TRANSFAC) \cite{165} and the authors reported that the GSA-PSO algorithm performs better than AlignACE, MEME, and GALF \cite{166}. GSA-PSO algorithm was compared with old algorithms, although there are a lot of recent algorithms for motif discovery. The presented methods based on the PSO algorithm have some limitations and the methods proposed by Reddy \textit{et al} and Elewa \textit{et al} \cite{4,111} used very simple fitness function and the program terminated manually by determining the numbers of iterations by the user. All the PSO-based methods start with random initialization except the method proposed by Abdullah \textit{et al} \cite{105}; this leads to time consuming operations and may not provide any correct solution. Moreover, PSO-based algorithms detect the motifs with OOPS model and ignore ZOOPS, and TCM models, and require motif length as the user input.

Few methods based on ABC, ACO, and CS are the most recent techniques in motif discovery.

ABC algorithm is a popular stochastic algorithm due to it is simple mechanism, and it requires few parameters which is easy to implement \cite{166}. Based on ABC algorithm, MO-ABC/DE algorithm is the same as the multiobjective ABC algorithm except for the generation of new candidate solutions; the DE operator was used to generate new candidates that combine existing ones according to a set of simple crossover-mutation schemes. The consensus ABC algorithm, another example of ABC technique, was tested on three different data sets and compared with a GA-based technique (MOGAMOD), a DE-based technique (DEFT) \cite{167}, and a genetic operator-based ABC algorithm (MO-ABC) \cite{168}. Consensus ABC algorithm achieves the highest similarity values of the motifs with different lengths; it used the neighbor selection strategy based on similarity values between consensus sequences and this improved the probability of the selection of a food source.

ACO algorithm has some advantages as; it usually avoids fault convergence due to distributed computation and can be used in dynamic applications.

From ACO algorithm, it can be concluded that there are some disadvantages: (1) It can easily trap local optimum, (2) The consuming time is uncertain, \textit{i.e.} it may take short/long time to converge, (3) The code is not obvious, (4) The ants must visit all points to get a good result that is unsuitable for motif discovery as it takes long time to check every substring in the input sequences and finally, (5) ACO requires many parameters. Based on ACO, Machhi \textit{et al} \cite{117} reported that the total required computing time is reduced.

Finally, there are several advantages for CS algorithm: (1) It usually converges to the global optimality, (2) It combines local and global capabilities and local search takes a quarter of the total search time and the remaining time is for global search which makes CS algorithm more efficient on the global scale, (3) Levy flight is used in its global search instead of standard random walks, so the CS can explore the search space more efficiently, and (4) It is easy to implement, compared with another metaheuristic search which essentially depends on only a single parameter (pa). In a previous study, it was reported that the CSO and MACS can be effectively used to obtain global optimum motif patterns for all used input sequences and they are faster than other nature inspired algorithms \cite{120}. Recently, Grey Wolf Optimization for motif finding (GWOMF) \cite{169} was applied to get motifs in DNA input sequences.

### Conclusion

The motif discovery algorithms are classified into four classes of enumerative, probability, nature inspired and combinatorial ones and each one has many subclasses. The comparison between them is listed in table 3. The enumerative technique is an exhaustive search with a simple concept, and it is the only technique that ensures to find all motifs (Except weak motifs). However, it is very slow, and requires a lot of parameters; as a result, it becomes difficult to deal with either long motifs or big data. Moreover, the degenerative positions are limited because of restricted representation of motifs.

Probability approach overcomes many weak points of the enumerative approach like speed, dealing with long motifs and big data, numbers of required parame-
Hashim FA, et al

Table 3. Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Can deal with a big data set</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Is it fast?</td>
<td>×</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Can it find long motif?</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Is it a global search?</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Low number of required parameters</td>
<td>×</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Representation</td>
<td>Consensus</td>
<td>PWM</td>
<td>Flexible</td>
<td>Flexible</td>
</tr>
<tr>
<td>Can it find all motifs?</td>
<td>✓</td>
<td>×</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Can it find weak motifs?</td>
<td>×</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Degenerate positions</td>
<td>Limited</td>
<td>Flexible</td>
<td>Flexible</td>
<td>Flexible</td>
</tr>
<tr>
<td>Objective function</td>
<td>Flexible</td>
<td>Flexible</td>
<td>Flexible</td>
<td>Flexible</td>
</tr>
<tr>
<td>Is it a simple concept?</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Derivation-free mechanism</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>?</td>
</tr>
</tbody>
</table>

Hint: Question mark means maybe yes or no. Enum., prob., Nat., and Com. are standing for Enumerative, Probabilistic, Nature-inspired and Combinatorial, respectively.

References


45. Sharma D, Rajasekaran S. A simple algorithm for (l, d) motif search. Proceedings of the 6th Annual IEEE con-


61. Congdon CB, Fizer CW, Smith NW, Gaskins HR, Aman J, Nava GM, Mattingly C. Preliminary results for GAM: A genetic algorithms approach to motif inference. Computational Intelligence in Bioinformatics and Comput-
olution strategy and

rithm for

proceed

eering 2017;20(1):33

Gibbs sampling and genetic algorithm. J Cluster Com-


Wei Z, Jensen ST. GAME: detecting cis-regulatory ele-

ments using a genetic algorithm. Bioinformatics 2006;22

(13):1577-1584.


motif discovery. Procedia Computer Science 2013;24:

25-29.

Li X, Wang D. (2009) An Improved Genetic Algo-

rithm for DNA Motif Discovery with Public Domain

Information. In: Köppen M., Kasabov N., Coghill G.

(eds) Advances in Neuro-Information Processing.

ICONIP 2008. Lecture Notes in Computer Science, vol

5506. Berlin, Heidelberg: Springer.

Kaya M. MOGAMOD: Multi-objective genetic algo-

rithm for motif discovery. Expert Systems Applications


Abdullah SLS, Harun H. SPECIES MOTIF EXTRA-

CTION USING LPBS. Proceedings of the 4 th Inter-

national Conference on Computing and Informatics,

ICO CI 2013, Sarawak, Malaysia. Universiti Utara Mal-

aysia 2013.

Zare-Mirakabad F, Ahrabian H, Sadeghi M, Mohammad-

zadeh J, Hashemifar S, Nowzari-Dalini A, et al. PSOMF:

An algorithm for pattern discovery using PSO. Pro-

ceedings of the Third IAPR International Conferences

on Pattern Recognition in Bioinformatics, Melbourne.

Australia; 2008: 61-72.

Lei C, Ruan J. A particle swarm optimization algorithm

for finding DNA sequence motifs. 2008 IEEE Inter-
national Conference on Bioinformatics and Biomedicine

Workshops, 166-173.

Lei C, Ruan J. A novel swarm intelligence algorithm for


Verma RS, Singh S, Kumar S. DNA sequence assembly

using particle swarm optimization. Int J Comput Appl


Karabulut M, Ibric k T. A Bayesian scoring scheme

based particle swarm optimization algorithm to identify

transcription factor binding sites. Appl Soft Comput


Lei C, Ruan J. A particle swarm optimization-based

algorithm for finding gapped motifs. BioData Min 2010;

3:9.

Hardin CT, Rouchka EC. DNA motif detection using

particle swarm optimization and expectation-maximiz-

ation. Proc IEEE Swarm Intell Symp 2005;2005:181-

184.

Elewa ES, Abdelhalim MB, Mabrouk MS. An efficient

system for finding functional motifs in genomic DNA

sequences by using nature-inspired algorithms. In:

Hassanien A, Shaalan K, Gaber T, Azar A, Tolba M,

(eds). Proceedings of the International Conference on

Advanced Intelligent Systems and Informatics 2016.

AISI 2016. Advances in Intelligent Systems and


Akbari R, Zeighami V, Ziarati K, Akbari I. Develop-

ment of an efficient hybrid method for motif discovery

in DNA sequences. AUT J Electrical Engineering 2012;

44(1):63-75.


156. Li L. Graphic network based methods in discovering TFBS motifs [master's thesis]. [Ohio (USA)]: The Ohio State University; 2012. 38 p.


