On the use of algorithms to discover motifs in DNA sequences

C. Rubio-Escudero[1], F. Martínez-Álvarez[2], M. Martínez-Ballesteros[1], J. C. Riquelme[1]

[1]*Department of Computer Science*
University of Seville, Spain
{*crubioescudero,mariamartinez,riquelme*}*@us.es*

[2]*Department of Computer Science*
Pablo de Olavide University of Seville, Spain
fmaralv@upo.es

***Abstract*—Many approaches are currently devoted to find DNA motifs in nucleotide sequences. However, this task remains challenging for specialists nowadays due to the difficulties they find to deeply understand gene regulatory mechanisms, especially when analyzing binding sites in DNA. These sites or specific nucleotide sequences are known to be responsible for transcription processes. Thus, this work aims at providing an updated overview on strategies developed to discover meaningful motifs in DNA-related sequences, and, in particular, their attempts to find out relevant binding sites. From all existing approaches, this work is focused on dictionary, ensemble, and artificial intelligence-based algorithms since they represent the classical and the leading ones, respectively.**

***Keywords*-motifs discovery, DNA sequences, regulatory mechanisms, binding sites**

I. Introduction

Genes are the main biological units of heredity and can be understood as a string of bases or sequence of chemicals whose apparently random combination encodes the hereditary information as well as genetic traits (individual characteristics). In particular, genes are composed of four different bases: Adenine (A), thymine (T), cytosine (C) and guanine (G).

Deoxyribonucleic acid (DNA) appears as a double helix of paired bases, i.e. two paired rungs of base sequences. But the matching of both rungs is not random. Instead, bases always appear matched with the same pair, adenine with thymine (A–T) and cytosine with guanine (C–G). That is, each of the two rungs consists of a sequence of bases (ATCCTG...) and the other one consists of the complementary sequence of bases (TAGGAC...).

In DNA helices it is usual to find particular sequences of bases, called binding sites, devoted to regulate transcriptions factors or factors that regulate gene expression by activating or inhibiting transcription machinery. In other words, from all the sequence of bases, these short specific sequences (typically from five to twenty pair long, versus 220 million existing base pairs long) are bound by more than one DNA-binding protein complexes.

Therefore, DNA motifs are meaningful base sequence patterns that identify binding sites responsible for transcription factors, and are known to appear in different genes and more than once within a gene. Finding motifs in DNA is a challenging task that many researchers are fielding nowadays insofar as their discovery may lead to major understanding of evolutive processes in living organisms.

Since Stormo [37] reviewed strategies to find motifs with computer algorithms, a large amount of algorithms have been developed. A major classification of these algorithms is done according to the type of DNA data sequence used to find motifs. Although there is no universal consensus on how to divide algorithms based on the input data, three major groups are generally considered [7]:

1) The algorithms that use information from coregulated genes from a single genome.
2) The algorithms that use information of a single gene in multiple species.
3) The algorithms that use information from phylogenetic footprints.

Despite the large amount of works found and grouped in classes 2) (see [31], [38], [30]) and 3) (see [1], [12], [46]), this paper is focused on providing a general overview on current algorithms that make use of the information that promoter sequences of coregulated genes generate. Actually, these methods can also be subdivided into multiple strategies, but this work only examines those based on dictionaries, ensembles and artificial intelligence-based techniques, as they represent the classical and the leading ones, respectively.

The rest of the paper is structured as follows. Section II discusses the most relevant works recently published related to dictionary-based algorithms. On the other hand, Section III presents ensemble algorithms used to find motifs in DNA. As for Section IV, it presents the latest AI works in the DNA motifs discovery field. Finally, Section V provides a brief summary of strengths and weaknesses of the reviewed strategies.

II. Dictionary-based algorithms

These are enumerative algorithms which, in contrast to heuristic methods, exhaustively cover the space of all possible motifs for a specific motif model description. The methodology progressively considers over-represented words, from short to long. The over-representativeness of a long word is computed as the weighted average of the short words in the current dictionary which could be part of the long word. Although this methodology is, in essence,

word counting based, it can filter out many spurious motifs which are over-represented owing to their overlapping with some real motifs. Therefore, it has a higher accuracy than a pure word counting method. However, computing the over-representativeness of a longer word by concatenating shorter ones is problematic, and may miss substantial over-represented motifs.

In this sense, van Helden et al. [40] developed a motif finding algorithm. The program counts all nucleotide occurrences within the sequence, and estimates their statistical significance. An essential prerequisite is that the system has to be calibrated to take into account the uneven nucleotide representation in the genome. Although conceptually simple, the algorithm proved efficient for extracting motifs from most of the yeast *(Saccharomyces cerevisiae)* regulatory families analyzed. These motifs had been previously found by laboratory experimental analysis. Furthermore, putative new regulatory sites were predicted within upstream regions of coregulated genes. However, its range of detection is limited to relatively simple patterns that include short motifs with highly conserved cores.

Later, van Helden et al. [41] extended their method to find transcription factors forming a dimer, with each unit binding to a similar small element, accounting for the symmetry of the site. The fixed spacing in the DNA site is due to the existence of a linker domain in the transcription factor, separating the DNA-binding and dimerization domains. These are called spaced dyad motifs, for the detection of spaced pairs shared by a set of upstream regions. The method is based on a systematic counting of pairs of short words separated by a fixed distance (space dyads) followed by a calculation of their statistical significance. Because the spacer can be different for distinct motifs, the spacer length is systematically varied between 0 and 16. The significance of this type of motif can be computed based on the combined score of the two conserved parts in the input data or based on the estimated complete dyad frequency from a background dataset. There is a big drawback in van Helden et al. [40] approach: there are no variations allowed within an nucleotide.

This problem was addressed by Tompa [39] with a proposal of an exact dictionary-based method. Tompa took into account both the absolute number of occurrences and the background distribution and created a table that, for each sequence s of a given length, records the number n of sequences containing an occurrence of s. There is a fixed number of substitutions sb allowed for the occurrences. The existence of a motif is calculated based on the probability of having n occurrences in a random sequence according to the background distribution. Thus, Tompa proposed an efficient algorithm to estimate the probability that a single random sequence contains at least one occurrence of the sequence s from a set of background sequences based on a Markov chain.

Brazma et al. [2] used a dictionary-based approach that searches exhaustively for an a priori unknown regular expression-type patterns that are over-represented in a given set of sequences. This proposal is capable to discover various subclasses of regular expression type patterns of unlimited length common to as few as ten sequences from thousands. It was applied in two cases, (1) discovery of patterns in the complete set of > 6000 sequences taken upstream of the putative yeast genes and (2) discovery of patterns in the regions upstream of the genes with similar expression profiles. Among the highest rating patterns, most had matches to known motifs in yeast.

Sagot [35] introduced a dictionary-based approach for motif finding that is based on the representation of a set of sequences with a suffix tree. Vanet et al. [42] used suffix trees to search for single motifs in whole genomes of bacteria. Marsan and Sagot [29] extended this method to search for combinations of motifs. Representation of upstream sequences as suffix trees gave a large number of possible combinations, however, the implementation was still efficient.

Bussemaker [4] presented MobyDick, suitable for discovering multiple motifs from a large collection of sequences. This approach formalizes how one would proceed to decipher a *text* consisting of a long string of letters written in an unknown language in which words are not delineated. The algorithm is based on a statistical mechanics model that segments the string probabilistically into *words* and concurrently builds a *dictionary* of these words. MobyDick can simultaneously find hundreds of different motifs, each of them present in only a small subset of the sequences, e.g., between 10 and 100 copies within the 6000 upstream regions in the yeast genome. The algorithm does not need an external reference dataset to calibrate probabilities and finds the optimal lengths of motifs automatically. They illustrated and validated the approach by segmenting a scrambled English novel, by extracting regulatory motifs from the entire yeast genome, and by analyzing data generated from a few DNA microarray experiments.

Wang [43] approached the problem of motif finding from the perspective of steganography [44]. They thought of the sequences as if they formed a stegoscript in which functional transcription factor binding motifs were secret messages embedded in a text of background sequences.They developed WordSpy, a dictionary based motif finding algorithm, integrating a word counting method and a statistical model. The word counting method is used to examine every possible word and the statistical model to capture over-represented motifs and background words in the given sequences. The algorithm presents some advantages. First, it does not require a background sequence, because it is capable of modeling background words based on the steganographic approach to the problem. This is an important feature for applications where a true background sequence model is

hard to determine. Second, WordSpy measures the over-representativeness of a word relative to that of all the other words modeled by the statistical model, resulting in an accurate measure of the over-representativeness. This feature helps to identify motifs of exact length. Third, the algorithm can incorporate gene expression profiling information to separate biologically significant motifs from spurious ones. Fourth, WordSpy is a discriminative motif finding algorithm. It can directly take as input two sets of sequences and find motifs that are over-represented in one set of sequences but not in the other set. Finally, the algorithm can conduct a whole genome analysis on the motifs that discovers the fidelity of the motifs to the given sequences.

Sharov and Minoru [36] presented CisFinder, a software that generates a comprehensive list of motifs enriched in a set of DNA sequences and describes them with position frequency matrices. A new algorithm was designed to estimate these matrices directly from counts of words with and without gaps; then the matrices are extended over gaps and flanking regions and clustered to generate non-redundant sets of motifs. The algorithm successfully identified binding motifs for twelve transcription factors in embryonic stem cells based on published chromatin immunoprecipitation sequencing data. Furthermore, CisFinder successfully identified alternative binding motifs of transcription factors and motifs for known and unknown co-factors of genes associated with the pluripotent state of ES cells. CisFinder also showed robust performance in the identification of motifs that were only slightly enriched in a set of DNA sequences.

III. Ensemble algorithms

Many motif finders have been proposed using different approaches. They have shown to be effective for discovering motifs in small living organisms, such as yeast [10]. However, their effectiveness remains unproven when dealing with huge DNA sequences belonging to more complex living forms.

In an attempt to solve this problem, Burset and Guigó [3] presented the idea of combining the outputs of several gene finding algorithms. Each algorithm typically covers only a small subset of the known binding sites, with relatively little overlap between the algorithms. They analyzed 9 motif finding programs with 570 DNA sequences. The dataset contained 2649 exons, and 174 of them where predicted by all programs and only 33 of them where not predicted by any of them. It is therefore advised to combine the results from multiple motif discovery tools, ideally covering a range of motif descriptions and search algorithms.

Harbison et al. [8] observed that different motif finders have different strengths. They successfully identified more binding sites by combining results of six motif finders compared to using only single finder. In fact, the benchmark datasets from Tompa et al. [10] also support this. By simply taking the union of all binding sites predicted by 10 selected motif finders, the sensitivity can be increased by more than double over each selected motif finder. However, the union of all predicted sites could contain a lot of noise therefore decreasing specificity. It is not trivial to distinguish the real binding sites from the noise.

Hu and Kihara, [16] proposed an algorithm which systematically combines predictions from five popular motif discovery algorithms. All the possible combinations of one to five component algorithms are examined. To be able to combine predictions of different runs from different component algorithms an algorithm termed EMD was developed. They tested their approach on two different types of datasets. One dataset is generated from the intergenic regions of the *E. coli* genome, and the other is comprised of the input sequences of different lengths generated by adding margins of different sizes to each known site. The best ensemble algorithm performed 22.4% better than the best single component algorithm in terms of the nucleotide level accuracy.

Wijaya et al. [45] presented MotifVoter, which, given a set of sequences, executes m different motif finders, each reporting n motifs. Note that each motif also defines a set of predicted binding sites. MotifVoter comprises two stages. (1) Motif filtering: in the first stage, MotifVoter processes the candidate motifs predicted by the m motif finders and attempts to remove the spurious motifs. The main idea is to find a cluster of motifs with high conformity based on a certain motif similarity measure. (2) Sites extraction: based on the candidate motifs retained in Stage 1, MotifVoter then identifies a set of sites with high confidence that they are real. They evaluated their approach on Tompa's benchmark and obtained a 95.2% of accuracy in the sensitivity. In E.coli dataset, MotifVoter achieved 95.7%.

Liu et al. presented EVIGAN [25] (EVidence Integration for Genome ANnotation using a Network). EVIGAN employs a dynamic Bayes net (DBN), a type of probabilistic graphical model that can accommodate multiple (possibly incomplete) gene predictions and other lines of evidence, yielding consensus gene models that maximize the probability of the evidence provided. The DBN model supports a wide variety of evidence types, including computational gene predictions, sequence homology search results, EST alignments and splice site predictions and it is easily extensible to incorporate other evidence types, such as proteomics hits, predicted domain architecture, SAGE tags, or Affymetrix tiling array data. EVIGAN's annotation process simulates an idealized human curator: different evidence sources are compared, those that tend to agree in particular contexts are assigned higher confidence and a consensus model is then created that reflects those confidence estimates. EVIGAN can produce a single consensus gene model or an ordered list of the n-best gene models, along with associated posterior probabilities for each. They applied EVIGAN to three large-scale datasets:

The ENCODE regions of the human genome [33], and the genomes of *Plasmodium vivax* and *Arabidopsis thaliana* [9]. These experiments demonstrate that for all three species, EVIGAN achieves better performance than any individual data source used as evidence.

Rubio-Escudero et al. [34] formulated the motif finding as a classification problem. It was interpreted as a decision between which section of a sequence is protein coding and which is not. The methodology uses a multi-objective approach to extract the best methods aggregations by maximizing the specificity and sensitivity of their predictions individually. It was applied to the EGASP sets from the ENCODE Genome Annotation Assessment Project (EGASP) [13], [11]. These datasets contain manually curated fragments of the human genome originating from the ENCODE project [33]. This dataset was selected by the EGASP assessment because the genes encoded in these regions were not used to train any particular gene predictor. Therefore, it is not a biased dataset. The aggregation of the results from various methods is accomplished using the union and intersection operators [14]. The methodology obtained successful results and consistently outperformed even the best individual approach and, in some cases, produced dramatic improvements in sensitivity and specificity. Moreover, they observed that even the worst methods contributed to the aggregation with more accurate programs.

IV. Artificial Intelligence-based techniques

This section explores the latest AI-based works published in the field of DNA motifs discovery. In particular, this section reviews approaches based on evolutionary techniques, self-organized maps, clustering and support-vector machines techniques.

Thus, a new multi-objective genetic algorithm (MO-GA) was introduced in [47] for the dyad motif discovery issue. In particular, the authors focused on optimizing three features: The sum of pairs, the number of matches and the information content. They also proposed new genetic operators to carry out such a task. Another GA-based approach was described in [26]. However, this time, the authors preferred to adopt a mechanism to regulate the concentration so that both the population diversity and vaccine mechanism are maintained in order to inhibit degeneracy during the evolutive process. Also in 2010, a GA was developed in [24]. This algorithm used a stochastic optimization technique based on particle swarm optimization (PSO). In particular, they proposed a modification of the standard PSO algorithm to adapt it to the discrete values that DNA sequences exhibit. The authors claim that the approach is especially useful when gaps are present in the motifs.

The application of self-organizing maps (SOM) can also be found in the literature. Hence, a SOM-based clustering algorithm was presented in [23], in which the authors extracted binding sites in DNA sequences. The main novelty of this work was to consider two different types signals in DNA sequences, showing that treating them separably better results can be achieved. On the contrary, three self-organizing neural networks were presented in [27] to find short motifs. Another SOM-based technique called SOMIX was introduced in [22] to discover binding sites in a set of regulatory regions. The tool proposed a intra-node soft competitive procedure in each node model to achieve maximum discrimination of motif from background signals, by weighting two different models: position specific scoring matrix and Markov chain. As it happened in [23] and [27], this method was inserted in another SOM-based approach, called SOMBRERO [28], that constructed models for motifs that were structurally similar.

The use of clustering techniques is also a usually strategy among researchers in this area. Thus, a hierarchical model with variable number of clusters was described in [17]. In particular, they used the Gibbs sampling strategy to allow width variation for each of the motifs. Moreover, a tool called Matlign based on hierarchical clustering was presented in [19]. The authors claimed that the tool was capable of post-processing large collections of DNA sequence motifs and of providing a non-redundant set of motifs, which could be further associated to known regulatory elements. Also, the well-known Fuzzy C-means (FCM) algorithm was applied in [20] to identify motifs in some particular regions of DNA sequences. The authors also tested K-means and Expectation-Maximization algorithms, showing that the fuzzy solution outperformed all others. The use of the K-means, and in particular an improved version, has been also explored in [5]. Thus, based on a previous enhancement by Zhong et al. [48] to overcome the random initialization problem associated to the original K-means version, the authors proposed two granular computing models that use FCM to split the dataset into smaller ones. Once divided, they applied their own K-means clustering algorithm version to every set to extract meaningful knowledge, reducing thus time costs.

There are also some relevant works that made use of support-vector machines (SVM) techniques. Thus, an approach that used one-class SVM algorithms to recognize transcription factor binding sites was proposed in [18]. Its main feature lied on the assumption that there exist correlations between transcription factors. The use of SVM combined with evolutionary processes can be found in [21]. This time the authors developed a method to predict binding proteins in DNA sequences. Thus, they created several SVM modules that were successfully combined with position specific scoring matrix (PSSM) profiles, a sort of evolutionary information. Pavesi and Valentini [32] formalized the problem of predicting genes' functional information as a classification problem, by using SVM with non-linear kernels. The training of such SVMs were carried out by means of both some particular DNA motifs and statistical

procedures. Finally, another SVM classifier can be found in [6], in which the authors used stochastic grammar rules to find regulatory DNA sequences that were indeed evaluated by means of SVM.

V. Conclusions

Gene expression regulatory mechanisms are widely studied among biologists and computer scientists. Particularly, most of their efforts are directed towards analyzing protein generation in the so-called binding sites. The discovery of these particular sequence of bases (or motifs) has generated numerous works insofar as they provide meaningful information on evolutive processes.

Many different strategies and subsequent approaches have been published. Consequently, experts need to have a piece of advice when selecting one or another algorithm in order to find motifs the best possible way. Although all of them are limited in what they can find [15], and it has been a challenging task to conduct studies on performance comparisons of motif finding tools, the scientific community agrees in labeling ensemble algorithms as the most effective ones owing to their capability of retrieving results from cooperative different methods.

In contrast, algorithms based on dictionaries have proven to be useful when analyzing small organisms but insufficient in big organisms. However, their inherent simplicity makes them as popular as widespread, and many experts continue conducting research on this topic nowadays.

Despite dictionary-based and ensembles approaches provides the researcher with reasonably good results, the discovery of DNA motifs has to deal with enormous amounts of data, being difficult to mine them with classical methodologies. Therefore artificial intelligence techniques have turned into necessary tools to speed up the full analysis of such data, as the other ones usually are able to partially face the complexities associated to such a problem.

None of these algorithms claim to be the *panacea*, and they are not indeed. However, it is an undeniable fact that the combination of all their strengths are leading to important discoveries that are helping to better understand transcription mechanisms in genes and therefore in human beings.

Acknowledgements

The research has been partially supported by the Spanish Ministry of Science and Technology under project TIN2007-68084-C-00, and Junta de Andalucía under project P07-TIC-02611.

References

[1] T. L. Bailey, Mikael Bodén, T. Whitington, and P. Machanick. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*, 11(179), 2010.

[2] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Research*, 8:1202–1215, 2000.

[3] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–357, 1996.

[4] H. J. Bussemaker, H. Li, and E. D. Siggia. From the cover building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Science of the USA*, 97(18):10096–10100, 2000.

[5] B. Chen, S. Pellicer, R. Harrison, P. C. Tai, and Y. Pan. Novel efficient granular computing models for protein sequence motifs and structure information discovery. *International Journal of Computational Biology and Drug Design*, 2(2):169–196, 2009.

[6] R. Damasevicius. Structural analysis of regulatory DNA sequences using grammar inference and support vector machine. *Neurocomputing*, 73(4-6):633–638, 2010.

[7] M. K. Das and H. K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21, 2007.

[8] C. Harbison et al. Transcription regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[9] J. E. Allen et al. Computational gene prediction using multiple sources of gene evidence. *Genome Research*, 14, 2004.

[10] M. Tompa et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, 2005.

[11] R. Guigó et al. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biology*, 7(1):S2, 2006.

[12] V. Storms et al. The effect of orthology and coregulation on detecting regulatory motifs. *PloS ONE*, 5(2):e8938, 2010.

[13] R. Guigó and M. Resse. EGASP: Collaboration through competition to find human genes. *Nature Methods*, 2:575–577, 2006.

[14] P. Halmos. *Naive set theory*. Princeton, NJ: D. Van Nostrand Company, 1960.

[15] J. Hu and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33:4899–4913, 2005.

[16] J. Hu and D. Kihara. EMD: An ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics*, 7(342):4899–4913, 2006.

[17] S. T. Jensen and J. S. Liu. Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association*, 103(481):188–200, 2008.

[18] B. Jiang, M. Q. Zhang, and X. Zhang. OSCAR: One-class SVM for accurate recognition of cis-elements. *Bioinformatics*, 23(21):2823–2828, 2007.

[19] M. Kankainen and A. Lytynoja. MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics*, 8(8):189, 2007.

[20] M. Karabulut and T. Ibrikci. Fuzzy C-Means based DNA motif discovery. *Lecture Notes in Computer Science*, 5226:189–195, 2008.

[21] M. Kumar, M. M. Gromiha, and G. P. S. Raghava. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, 8:463, 2007.

[22] N. K. Lee and D. Wang. SOMIX: Motifs discovery in gene regulatory sequences using self-organizing maps. *Lecture Notes in Computer Science*, 6444:242–249, 2010.

[23] N. K. Lee and D. Wang. SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model. *BMC Bioinformatics*, 12(Suppl 1):S16, 2011.

[24] C. Lei and J. Ruan. Finding gapped motifs by a novel evolutionary algorithm. *Lecture Notes in Computer Science*, 6023:50–61, 2010.

[25] Q. Liu, A. J. Mackey, D. S. Roos, and F. C. N. Pereira. Evigan: A hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, 24(5):597–605, 2008.

[26] J. W. Luo and T. Wang. Motif discovery using an immune genetic algorithm. *Journal of Theoretical Biology*, 264(2):319–325, 2010.

[27] S. Mahony, P. V. Benosa, T. J. Smith, and Aaron Golden. Self-organizing neural networks to support the discovery of DNA-binding motifs. *Neural Networks*, 19:950–962, 2006.

[28] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–1814, 2005.

[29] L. Marsan and M. F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Molecular Biology*, 287:345–362, 2000.

[30] T. T. Nguyen and I. P. Androulakis. Recent advances in the computational discovery of transcription factor binding sites. *Algorithms*, 2(1):582–605, 2009.

[31] M. J. Palumbo and L. A. Newberg. Phyloscan: Locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Research*, 38(2):W268–W274, 2010.

[32] G. Pavesi and G. Valentini. Classification of co-expressed genes from DNA regulatory regions. *Information Fusion*, 10(3):233–241, 2009.

[33] ENCODE project consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306:636–640, 2004.

[34] C. Rubio-Escudero, R. Romero-Záliz, I. Zwir, and C. del Val. Optimization of multiclassifiers for computational biology: Application to gene finding and gene expression. *Theoretical Chemical Accounts*, 125(3-6):599–611, 2010.

[35] M. Sagot. Spelling approximate repeated or common motifs using a suffix tree. *Lecture Notes in Computer Science*, 1380:111–127, 1998.

[36] A. A. Sharov and S. H. Minoru. Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Research*, 16:261–273, 2009.

[37] G. D. Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[38] D. Straume, R. F. Johansen, M. Bjrs, I. F. Nes, and D. B. Diep. DNA binding kinetics of two response regulators, PlnC and PlnD, from the bacteriocin regulon of Lactobacillus plantarum c11. *BMC Biochemistry*, 10(17), 2009.

[39] M. Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the Seventh International Conference on Intelligent Systems on Molecular Biology*, pages 262–271, 1999.

[40] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281:827–842, 1998.

[41] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28:1808–1818, 2000.

[42] A. Vanet, L. Marsan, A. Labigne, and M. F. Sagot. Inferring regulatory elements from a whole genome. an analysis of Helicobacter pylori σ^{80} family of promoter signals. *Journal of Molecular Biology*, 297:335–353, 2000.

[43] G. Wang, T. Yu, and W. Zhang. WordSpy: Identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Research*, 33:412–416, 2005.

[44] P. Wayner. *Disappearing Cryptography*. Morgan Kaufmann Publishers, 2002.

[45] E. Wijaya, K. Rajaraman, S. M. Yiu, and W. K. Sung. Detection of generic spaced motifs using submotif pattern mining. *Bioinformatics*, 23(12):1476–1485, 2007.

[46] K. J. Won, B. Ren, and W. Wang. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, 11(1):R7, 2010.

[47] F. Zare-Mirakabad, H. Ahrabian, M. Sadeghi, S. Hashemifar, A. Nowzari-Dalini, and B. Goliaei. Genetic algorithm for dyad pattern finding in DNA sequences. *Genes and Genetic Systems*, 84(1):81–93, 2009.

[48] W. Zhong, G. Altun, R. Harrison, P. C. Tai, and Y. Pan. Improved kmeans clustering algorithm for exploring local protein sequence motifs representing common structural property. *IEEE Transactions on Nanobioscience*, 4(3):255–265, 2005.