L-SME: A System for Mining Loosely Structured Motifs

Fabio Fassetti [1], Gianluigi Greco [2], and Giorgio Terracina [2]

[1] ICAR-CNR
[2] Dep. of Mathematics, Via P. Bucci, 87036 Rende (CS), Italy
ffassetti@deis.unical.it, {ggreco,terracina}@mat.unical.it

Abstract. We present L-SME, a system to efficiently identify loosely structured motifs in genome-wide applications. L-SME is innovative in three aspects. Firstly, it handles wider classes of motifs than earlier motif discovery systems, by supporting boxes swaps and skips in the motifs structure as well as various kinds of similarity functions. Secondly, in addition to the standard exact search, it supports search via randomization in which guarantees on the quality of the results can be given a-priori based on user-definable resource (time and space) constraints. Finally, L-SME comes equipped with an intuitive graphical interface through which the structure for the motifs of interest can be defined, the discovery method can be selected, and results can be visualized. The tool is flexible and scalable, by allowing genome-wide searches for very complex motifs and is freely accessible at http://siloe.deis.unical.it/l-sme. A detailed description of the algorithms underlying L-SME is available in [1].

1 Introduction

Transcriptional control is a crucial mechanism for gene regulation, in which certain proteins, called transcription factors, bind near genes to activate or inhibit the transcription of genetic information from DNA to RNA. Transcription factors are known to have special affinity for short DNA regions called binding sites, which occur several times in the same genome and which are conserved in evolution over different organisms [5]. Thus, singling out the regions that are over-represented in suitably selected sets of DNA sequences provides us with insights on the biological functions played by the corresponding macromolecules [3]. These regions are called *motifs* in the literature.

Several discovery methods and tools have already been conceived to identify motifs conforming to some model templates that capture the similarities of diverse binding sites, and which are fixed by the biologist who is willing to corroborate his hypothesis on the co-regulation of some given genes. In its basic form, a model template is just the specification of a length l for sequences of DNA basis (called *boxes*); thus, a motif conforming to such a template is precisely a sequence of l basis that is frequently repeated over the genome at hand. More complex model templates, instead, are supported in a few state-of-the-art systems—see, e.g., the comparative analysis by [1]—in order to look for motifs *(i)* that are made of several boxes at a given distance over the gene (called *gap*) from one another [2,6,3], and *(ii)* that may be partially conserved in the repetitions, since *mismatches* are allowed.

In this paper, the L-SME motif discovery tool is presented. In addition to supporting classical model templates, the tool allows to specify box swaps and skips as well

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part III, LNAI 6913, pp. 621–625, 2011.

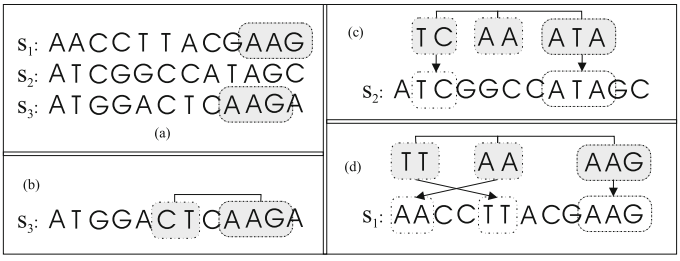

Fig. 1. Illustration of model templates and instances

as various similarity functions to handle mismatches, all of them being variabilities of interest in the context of analyzing eukaryotic transcription [4,7]. In order to handle wider classes of model templates than existing systems while still guaranteeing scalability over genome-wide applications, L-SME is founded on specialized data-structures and advanced computational methods supporting both exact and randomized searches.

2 Motif Discovery Problem

In state-of-the-art discovery tools, a motif template $\widehat{p}$ can be viewed as a tuple $\langle l_1, d_1, l_2, d_2, ..., d_{r-1}, l_r \rangle$, where l_i indicates the length of the i-th box, d_j indicates the length of the gap separating the j-th and the $(j+1)$-th boxes, and r is the number of boxes—actually, both l_i and d_j can be (possibly degenerating) intervals of the form $l_i = [\min\_l_i : \max\_l_i]$ and $d_j = [\min\_d_j : \max\_d_j]$. A *pattern instance* p for $\widehat{p}$ is a string $p = b_{l_1}\ X(d_1)\ b_{l_2}\ X(d_2)...X(d_{r-1})\ b_{l_r}$, where b_{l_i} are strings, the length of b_{l_i} is in the range $[\min\_l_i : \max\_l_i]$, and $X(d_j)$ is a sequence of d_j special ("don't care") symbols X with length in the interval $[\min\_d_j : \max\_d_j]$. We say that the instance p *occurs* in a DNA sequence s if there is a substring s' of s that matches with p, i.e., such that the *Hamming* or *Levenshtein distance* between each box in p and the corresponding sequence of symbols in s' is below a given threshold (denoting the number of allowed mismatches). Eventually, the *motif discovery problem* over a set of DNA sequences is to find all the instances for $\widehat{p}$ that occur in at least `Q` of them, where `Q` is the *quorum* considered appropriate by the biologist for the application at hand.

As an example, over the sequences depicted in Figure 1(a), AAG is the only solution for `Q`=2, for a model template made by one box of exactly three symbols, and when no mismatches are allowed. As a further example, a more complex template composed of two boxes separated by one irrelevant symbol is shown in Figure 1(b) together with an instance for it occurring in s_3.

3 System Functionalities

L-SME is a tool for motif discovery supporting various innovative functionalities, under various different perspectives.

(1) Supported-Templates Perspective: The tool deals with a wider class of model templates than those discussed in Section 2. Indeed, in addition to those elements, L-SME

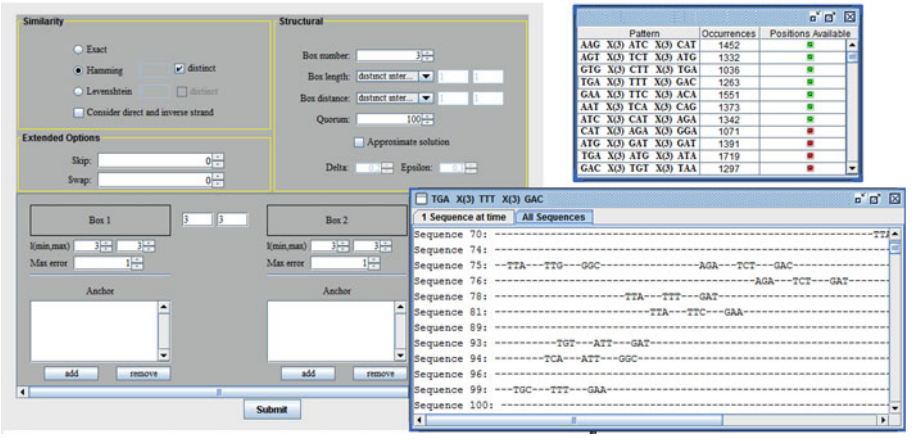

Fig. 2. Screenshots of L-SME

supports two further variabilities clearly emerged from recent studies [4,7]. Specifically, L-SME allows patterns to be matched with some given strings even though:

[*Box skips*] up to a certain user-definable number of boxes is not preserved at all. E.g., Figure 1(c) shows a pattern instance with three boxes matching with s_2 provided one box skip is allowed.

[*Box swaps*] the relative positions of two consecutive boxes is inverted, where users may specify the maximum number of allowed inversions. E.g., Figure 1(d) shows a pattern instance matching with s_1 provided one box swap.

Moreover, differently from current systems, which are designed to deal with one fixed similarity function only (either Hamming or Levenshtein distance), the similarity function to be used with L-SME can be freely selected by the biologist.

(2) Algorithmic Perspective. Given the need to handle wide classes of templates while guaranteeing scalability over genome-wide applications, L-SME supports search via randomization, in which a-priori guarantees on the quality of the results can be given based on user-definable resource (time and space) constraints. Randomization is based on using *sketches* to store pattern occurrences. Specifically, users can trim two normalized coefficients δ and ϵ to indicate the amount of space to be used for each sketch and the range of tolerance admitted over the accuracy of the solutions, respectively. Higher values reduce time/space requirements but also results quality guarantees. For $\delta = \epsilon = 0$, the randomized approach degenerates to the exact search.

(3) Interfacing Perspective. Given the wide range of parameters handled by L-SME, its user interface is carefully designed so as to simplify the setting-up phases of biological experimentations. In fact, differently from most of the other tools in the literature, L-SME is equipped with a web-based interface where both the process of specifying the model template with the parameters of interest, and the navigation of results can be carried out in a visual and interactive manner. In particular, for each motif that is discovered, L-SME allows the user to visualize its occurrences over the input sequences, which is often very helpful for the biologist.

Table 1. Binding sites information for UASH and URS1H in Saccharomyces cerevisiae. Here ORF stands for Open Reading Frame.

Gene #	ORF	Gene ID	Mapped Site for UASH	Gap	Mapped Site for URS1H
1	YDR285W	ZIP1	GATTCGGAAGTAAAA	5	TCGGCGGCTAAAT
2	YER044C-A	MEI4	TCTTTCGGAGTCATA	8	TGGGCGGCTAAAT
3	YER179W	DMC1	TTGTGTGGAGAGATA	17	AAATAGCCGCCCA
4	YHR014W	SPO13	TAATTAGGAGTATAT	4	AAATAGCCGCCGA
5	YNL210W	MER1	GGTTTTGTAGTTCTA	22	TTTTAGCCGCCGA
6	YHR153C	SPO16	CATTGTGATGTATTT	96	TGGGCGGCTAAAA
7	YHR157W	REC104	CAATTTGGAGTAGGC	74	TTGGCGGCTATTT
8	YLR263W	RED1	ATTTCTGGAGATATC	173	TCAGCGGCTAAAT
9	YMR133W	REC114	GATTTTGTAGGAATA	179	TGGGCGGCTAACT
10	YOR351C	MEK1	TCATTTGTAGTTTAT	179	ATGGCGGCTAAAT
11	YIL072W	HOP1	TGTGAAGT	-323	ATGGCGGCTAAAT

(4) Computation Perspective. Finally, since motif discovery is a computationally intensive task, L-SME is designed to incrementally produce results. In fact, each request is immediately answered with an *url* where discovered results are visualized as soon as they are discovered by internal algorithms, and remain available for some days.

Example Usage. As an example of the results that can be obtained with our system, coupled with a Z-score analysis, we consider here the Saccharomyces cerevisiae, for which several (single) transcription factors are well-known, with some of them being recognized to cooperatively regulate the corresponding genes. For instance, the transcription factors URS1H and UASH are involved in early meiotic expression during sporulation and are known to cooperate for the expression of 11 genes. Table 1 summarizes the genes involved, the transcription factors and the relative positions, which were annotated by biologists and confirmed by our system. The negative gap reported for HOP1 indicates that, in this gene, the relative positions of the two factors are actually swapped w.r.t. all the other genes; this occurrence would not be derived without the support of box swaps and, hence, in current systems available in the literature.

References

1. Fassetti, F., Greco, G., Terracina, G.: Mining loosely structured motifs from biological data. IEEE Transaction on Knowledge and Data Engineering 20(11), 1472–1489 (2008)
2. Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. Journal of Molecular Biology 296(5), 1205–1214 (2000)
3. Marsan, L., Sagot, M.-F.: Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. Journal of Computational Biology 7(3-4), 345–362 (2000)
4. Osanai, M., Takahashi, H., Kojima, K.K., Hamada, M., Fujiwara, H.: Essential motifs in the 3' untranslated region required for retrotransposition and the precise start of reverse transcription in non-long-terminal-repeat retrotransposon SART1. Mol. Cell. Biol. 24(19), 7902–7913 (2004)
5. Sandve, G.K., Drabls, F.: A survey of motif discovery methods in an integrated framework. Biology Direct 1(11), 1–16 (2006)

6. Sinha, S., Tompa, M.: YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acid Research 31(13), 3586–3588 (2003)
7. Tu, Z., Li, S., Mao, C.: The changing tails of a novel short interspersed element in aedes aegypti: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(da) tail. Genetics 168(4), 2037–2047 (2004)