index. The variable gap constraints are also considered at the same time as the joins, resulting in considerable efficiency. In order to save time and space, we only keep the start positions of each intermediate pattern during the positional join.

Related work

Many simple motif extraction algorithms have been proposed primarily for extracting the transcription factor binding sites, where each motif consists of a unique binding site [4-10] or two binding sites separated by a fixed number of gaps [11-13]. A pattern with a single component is also called a *monad pattern*. Structured motif extraction problems, in which variable number of gaps are allowed, have attracted much attention recently, where the structured motifs can be extracted either from multiple sequences [14-21] or from a single sequence [22,23]. In many cases, more than one transcription factor may cooperatively regulate a gene. Such patterns are called *composite regulatory patterns*. To detect the composite regulatory patterns, one may apply single binding site identification algorithms to detect each component separately. However, this solution may fail when some components are not very strong (significant). Thus it is necessary to detect the whole composite regulatory patterns (even with weak components) directly, whose gaps and other possibly strong components can increase its significance.

Several algorithms have been used to address the composite pattern discovery with two components, which are called *dyad patterns*. Helden et al. [11] propose a method for dyad analysis, which exhaustively counts the number of occurrences of each possible pair of patterns in the sequences and then assesses their statistical significance. This method can only deal with fixed number of gaps between the two components. MITRA [12] first casts the composite pattern discovery problem as a larger monad discovery problem and then applies an exhaustive monad discovery algorithm. It can handle several mismatches but can only handle sequences less than 60 kilo-bases long. Co-Bind [24] models composite transcription factors with Position Weight Matrices (PWMs) and finds PWMs that maximize the joint likelihood of occurrences of the two binding site components. Co-Bind uses Gibbs sampling to select binding sites and then refines the PWMs for a fixed number of times. Co-Bind may miss some binding sites since not all patterns in the sequences are considered. Moreover, using a fixed number of iterations for improvement may not converge to the global optimal dyad PWM.

SMILE [14] describes four variants of increasing generality for common structured motif extraction, and proposes two solutions for them. The two approaches for the first problem, in which the structured motif template consists of two components with a gap range between them, both start by building a generalized suffix tree for the input sequences and extracting the first component. Then in the first approach, the second component is extracted by simply jumping in the sequences from the end of the first one to the second within the gap range. In the second approach, the suffix tree is temporarily modified so as to extract the second component from the modified suffix tree directly. The drawback of SMILE is that its time and space complexity are exponential in the number of gaps between the two components. In order to reduce the time during the extraction of the structured motifs, [18] presents a parallel algorithm, PSmile, based on SMILE, where the search space is well-partitioned among the available processors.

RISO [15-17] improves SMILE in two aspects. First, instead of building the whole suffix tree for the input sequences, RISO builds a suffix tree only up to a certain level l, called a *factor tree*, which leads to a large space saving. Second, a new data structure called *box-link* is proposed to store the information about how to jump within the DNA sequences from one simple component (box) to the subsequent one in the structured motif. This accelerates the extraction process and avoids exponential time and space consumption (in the gaps) as in SMILE. In RISO, after the generalized factor tree is built, the box-links are constructed by exhaustively enumerating all the possible structured motifs in the sequences and are added to the leaves of the factor tree. Then the extraction process begins during which the factor tree may be temporarily and partially modified so as to extract the subsequent simple motifs. Since during the box-link construction, the structured motif occurrences are exhaustively enumerated and the frequency threshold is never used to prune the candidate structured motifs, RISO needs a lot of computation during this step.

For repeated structured motif identification problem, the frequency closure property that "all the subsequences of a frequent sequence must be frequent", doesn't hold any more since the frequency of a pattern can exceed the frequency of its sub-patterns. [22] introduces an closure-like property which can help prune the patterns without missing the frequent patterns. The two algorithms proposed in [22] can extract within one sequence all frequent patterns of length no greater than a length threshold, which can be either manually specified or automatically determined. However, this method requires that all the gap ranges $[l_i, u_i]$, between adjacent *symbols* in the structured motif be the same, i.e., $[l_i, u_i] = [l, u]$ for all $i \in [1, k - 1]$. Moreover, approximate matches are not allowed for the structured motif.