Efficient motif finding algorithms for large-alphabet inputs

Pavel P. Kuksa and Vladimir Pavlovic
Department of Computer Science, Rutgers University, Piscataway, NJ 08854
{pkuksa,vladimir}@cs.rutgers.edu

Keywords: protein, DNA, sequence motifs, algorithms, motif finding, regulatory patterns

Identifying motifs, such as regulatory patterns in DNA and protein sequences, is a fundamental problem in bioinformatics. Sequence-based computational approaches have proved to be valuable in predicting regulatory signals [1, 2, 3, 4, 5]. Many sequence-based motif algorithms (c.f., Pevzner et al. [2], MITRA [3], MEME [5], WEEDER [6]) use a technique based on searching for motifs starting from samples of fixed size present in the input sequences and exploring their neighborhood.

Method. In this work, we propose a method which significantly improves search efficiency compared to existing algorithms, makes the search feasible for large alphabet inputs (such as protein sequences, or 3D structures) and longer, less conserved motifs. The proposed *implicit* combinatorial search algorithms address the following questions: (1) For a given set of sequences find the *consensus* sequence of the motif; and (2) For a given set of sequences, containing regulatory sites, find the location of these regulatory sites.

The algorithms work by identifying a smaller feasible subset of candidate patterns $R \subset S$ in the input sample S and efficiently constructing consensus pattern from R using *wildcards*. To construct R we build the set of stems $\mathcal{H}$, common patterns of all pairs of k-mers from the set $\mathcal{I}$, the k-long substrings not more than $2m$ symbols apart. The complexity of constructing R is $O(\binom{k}{2m}nN + \binom{k}{m}|\mathcal{H}||\mathcal{I}|^2)$ for k-long patterns with up to m mismatches from N n-long sequences, and does not depend on the alphabet size Σ. Our analysis shows that, on average, both $|\mathcal{H}|$ and $|\mathcal{I}|$ are significantly lower than the corresponding factors of other exact motif search algorithms. As a consequence, the search complexity independent of the alphabet size makes the search tractable for longer, less conserved motifs. Our approach directly discovers consensus motif patterns consisting of the most likely residues at each position, the majority letters in position-specific logos used by probabilistic methods. This subsequently implies that results of our search can be used either on their own, as a part of common pattern discovery systems to improve their search efficiency, or to provide starting points for PSSM-based probabilistic motif search methods.

Results. We carried out experiments on three motif finding tasks: (1) predicting transcription factor binding sites in DNA sequences, (2) finding sequence motifs in topologically similar proteins with weak sequence-based similarity, (3) identifying protein sequence motifs corresponding to super-secondary structure (SSS) motifs. We compared our method to a number of highly efficient motif finders, including MITRA [3] and RISOTTO [4]. A sample of running time results is shown in Table 1.

Our results demonstrate that the proposed algorithms achieve significant, order-of-magnitude speedups, while making it possible to search for longer, less conserved motifs in large-$|\Sigma|$ sequences (e.g., proteins) of increased lengths.

Table 1: Algorithm performance (running time, s) on difficult motif discovery cases (protein sequences)

Dataset	**Motif**	**MITRA**	**Our algorithm**
Lipocalins	(16,6)	> 1 day	**300s**
Zinc metallopeptidase	(11,5)	> 1 day	**185s**
Super-secondary structure motifs	(20, 4)	∞	**47s**
Planted motif set	(13,4)	28905	**5.3s**

References

[1] M Tompa, N Li, T Bailey, G Church, and B De Moor. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, Jan 2005.

[2] Pavel A. Pevzner and Sing-Hoi Sze. Combinatorial approaches to finding subtle signals in dna sequences. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278. AAAI Press, 2000.

[3] Eleazar Eskin and Pavel A. Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18(suppl1):S354–363, 2002. http://www.ccls.columbia.edu/compbio/mitra/.

[4] Nadia Pisanti, Alexandra M. Carvalho, Laurent Marsan, and Marie-France Sagot. RISOTTO: Fast extraction of motifs with mismatches. In *LATIN*, pages 757–768, 2006.

[5] Timothy L. Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, 21(1-2):51–80, 1995.

[6] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17(suppl1):S207–214, 2001.