

# Greedy mixture learning for multiple motif discovery in biological sequences

(*Bioinformatics*, 19:607-617, 2003)

Konstantinos Blekas, Dimitrios I. Fotiadis and Aristidis Likas

Department of Computer Science, University of Ioannina 45110 Ioannina, Greece  
and Biomedical Research Institute, Foundation for Research and Technology - Hellas, 45110 Ioannina, Greece

motif	starting position	seed motifs											
		1	2	3	4	5	6	7	8	9	10		
1	1-20	***	K	L	I	M	A	T	I	S	M	A	***
2	31-50	***	P	E	G	T	H	T	I	S	M	A	***
3	61-80	***	A	R	N	D	C	Q	E	G	H	I	***
4	91-110	***	E	G	H	I	L	K	M	F	P	S	***
5	121-140	***	W	Y	V	T	R	Q	A	N	P	V	***
6	151-170	***	A	N	C	E	H	L	M	P	T	Y	***

Table 1: The motif distribution in the first series of artificial datasets. The first two motifs are the same in half of their length.

## 1 Extended experiments with artificial datasets

In the artificial datasets used in our experiments each motif has an associated randomly generated "seed substring" and copies of the motif (motif occurrences) are created by randomly performing a number of substitutions (*mutations*) on the motif's seed substring with a mutation probability  $p_m$ . In fact, the mutation operation inserts a degree of noise within the motif description and, as a result, the greater the probability value  $p_m$ , the harder the motif identification problem. For simplicity, without loss of generality, we have chosen to construct each artificial sequence using mutated copies of all the motifs (single occurrence for each one) at random positions (assuming no overlapping occurrences).

Two series of experiments have been conducted with the artificial datasets. In the first series we measured the impact of the proposed *kd*-tree approach for candidate selection on the performance of the whole algorithm. We created artificial datasets using six (6) different seed substrings of length  $W = 10$  (Table 1), where the first two substrings are identical in half length, therefore making the problem of discovering them harder. Ten such sequences of variable length (between 190 and 220) were constructed by randomly locating and mutating copies of these substrings (ensuring no overlapping), while randomly filling the rest positions with characters from the amino acids alphabet  $\Sigma$ . Assuming three different values of the mutation probability  $p_m = \{0, 0.1, 0.2\}$ , three artificial datasets were constructed with average number of substrings  $n = 1970$ .

For each dataset we run the proposed *kd*-tree technique

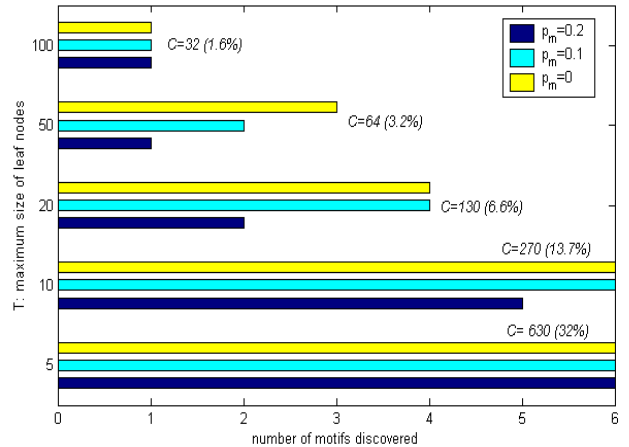


Figure 1: Estimating the impact of the proposed *kd*-tree approach using five different values of  $T$ . For small values of  $T$  the partitioning scheme produces sufficient candidates for discovering the six artificial motifs.

motif	starting position	seed motifs																					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
1	1-40	***	A	V	E	R	Y	I	N	T	E	R	E	S	T	I	N	G	V	I	E	W	***
2	61-90	***	D	E	S	T	I	N	A	T	E	D	Y	D	E	D	I	C	A	T	E	D	***
3	111-140	***	P	R	I	M	E	R	S	T	A	G	E	D	E	S	I	G	N	I	N	G	***
4	161-190	***	W	A	I	T	I	N	G	F	T	H	E	G	I	A	N	T	T	R	A	M	***
5	211-240	***	S	T	E	W	A	R	T	E	L	E	N	A	M	I	L	E	N	A	E	M	***
6	261-290	***	A	R	I	S	T	I	D	I	S	K	N	A	M	I	L	E	N	A	E	M	***

Table 2: The motif distribution in the second series of artificial datasets. Motifs 5, 6 are similar in half of their length.

considering five different values for the size of the leaf nodes ( $T = \{100, 50, 20, 10, 5\}$ ). The produced set of  $C$  candidates is then used as the reduced search space for initializing the components during the greedy mixture learning approach. The results are represented in Figure 1 with horizontal bars indicating the number of motifs discovered by the greedy algorithm for the three datasets. In addition the average number  $C$  of the produced candidates (as well as the percentage of the total number of substrings  $n$ ) is shown.

It can be observed that large values of  $T$  ( $T \geq 20$ ) result in the specification of a very small number of candidates

(less than 7% of  $n$ ), leading to unacceptable motif discovery performance. As the value of  $T$  decreases, the number of candidates increases leading to performance improvement. Best results (identification of all the six motifs) are obtained for setting  $T = 5$  that produces a search space with 630 candidate substrings (32% of  $n$ ). In the following experiments the applied Greedy EM approach uses the  $k$ -tree partitioning scheme with  $T = N/2$  (half the number of sequences).

During the second series of experiments we compare the performance of our Greedy EM approach with the MEME method in terms of the ability to discover the real number of motifs in artificial datasets. For this reason we have created new artificial datasets from another set of six (6) seed substrings of length  $W = 20$ . As it is illustrated in Table 2, the last two seed substrings (5 and 6) are exactly the same in half of their length (from position 11 to 20). This fact makes the problem of identifying them as two distinct motifs very difficult. Using a similar procedure, another ten ( $N = 10$ ) artificial sequences of variable length (in the range [310, 330]) were constructed. For three different values of the mutation probability ( $p_m = \{0, 0.1, 0.2\}$ ), we have created three different training datasets used in our experimental study.

For those artificial datasets we applied both MEME and Greedy EM until at most 10 motifs were discovered in each run. For each discovered motif the number  $N_s$  of its occurrences in the training sequences was computed and we kept only the motifs occurring in at least half of the training sequences (i.e.  $N_s \geq N/2$ ). Those motifs with high coverage were finally examined and evaluated, while the other were considered as *redundant*.<sup>1</sup>

In all experiments with the three artificial datasets for  $W = 20$  (true motif length), the Greedy EM method identified six non-redundant motifs which were the same as the six true motifs of Table 2. The number of occurrences of each motif was equal to the size of the training set (i.e.  $N_s = 10$ ). On the other hand, the MEME method was not able to separate the overlapping motifs 5 and 6, always considering them as one motif with two copies per sequence (thus  $N_s = 20$ ). The remaining four true motifs were also discovered by MEME (single copy per sequence). Furthermore, in order to test the sensitivity of the results with respect to the motif length  $W$ , additional comparative experiments were carried out using the artificial dataset constructed with mutation probability 0.2.

Table 3 illustrates the non-redundant motifs discovered by both algorithms for five values of  $W$ . For each motif the number of occurrences  $N_s$  within training set and the corresponding  $IC$ -score are also presented. As mentioned previously, since in most cases the number of occurrences of the motifs obtained by both algorithms is almost the same, the  $IC$ -score constitutes a reasonable evaluation measure of motif quality. From the results in Table 3 it is clear that for values of  $W$  smaller than the true length (i.e. 15 and 18), the results of the two methods were similar, since they both considered the two overlapping motifs 5 and 6 as a single motif (indicated in Table 3 as 5\*) ap-

Length ( $W$ )	MEME			Greedy EM		
	# motif	$N_s$	$IC$	# motif	$N_s$	$IC$
15	1	10	44.7	1	10	50.8
	2	10	42.0	2	10	50.0
	3	10	40.8	3	10	49.5
	4	9	40.7	4	9	48.0
	5*	20	38.1	5*	20	45.9
	6*	7	24.1			
18	1	10	52.2	1	10	60.4
	2	10	50.3	2	10	60.2
	3	10	48.8	3	10	59.1
	4	9	48.7	4	9	56.9
	5*	20	43.1	5*	20	52.1
20	1	10	58.2	1	10	68.1
	2	10	56.6	2	10	66.8
	3	10	53.0	3	10	66.6
	4	9	54.6	4	9	63.8
	5*	20	47.7	5	10	64.8
				6	10	64.1
22	1	10	59.4	1	10	71.6
	2	10	57.6	2	10	70.1
	3	10	54.5	3	10	69.4
	4	10	54.2	4	10	67.9
	5*	20	48.7	5	10	67.5
				6	10	67.0
25	1	10	61.2	1	10	76.0
	2	10	59.8	2	10	74.7
	3	10	56.2	3	10	74.1
	4	10	56.2	4	10	72.4
	5*	20	50.1	5	10	72.3
				6	10	72.1

Table 3: Motifs discovered by the MEME and the Greedy EM approach in the artificial set of 10 sequences for five values for the motif length  $W$ . The number of occurrences  $N_s$  and the  $IC$ -score for each motif are presented.

pearing twice in each sequence. This is rather expected result since for  $W < 20$  the percentage of overlapping of the seed motifs 5 and 6 is very big, e.g. for  $W = 15$  the overlapping is 67%. The remaining four motifs discovered by the two methods were substrings of the true motifs with  $N_s \approx 10$ . In addition one false positive motif was discovered by MEME (indicated in Table 3 as 6\*) for  $W = 15$ . For the cases where  $W$  was equal or greater than the true motif length ( $W \geq 20$ ), the MEME consistently was not able to discriminate the two overlapping motifs. On the contrary, the greedy EM algorithm was able to identify all the six true motifs and locate all their occurrences within the training set of sequences.

It must also be noted that for all  $W$  values the  $IC$ -scores corresponding to the motifs discovered by greedy-EM were consistently higher than those provided by MEME. In addition no false positives were identified, since each non-redundant motif was in one-to-one correspondence with one of the true motifs (except for the case of the composite motif which corresponds to two motifs).

<sup>1</sup>In our experiment every redundant motif had at most three occurrences within the training set.