Mutual enrichment in ranked lists and the statistical assessment of position weight matrix motifs

Limor Leibovich[1] and Zohar Yakhini[1,2]

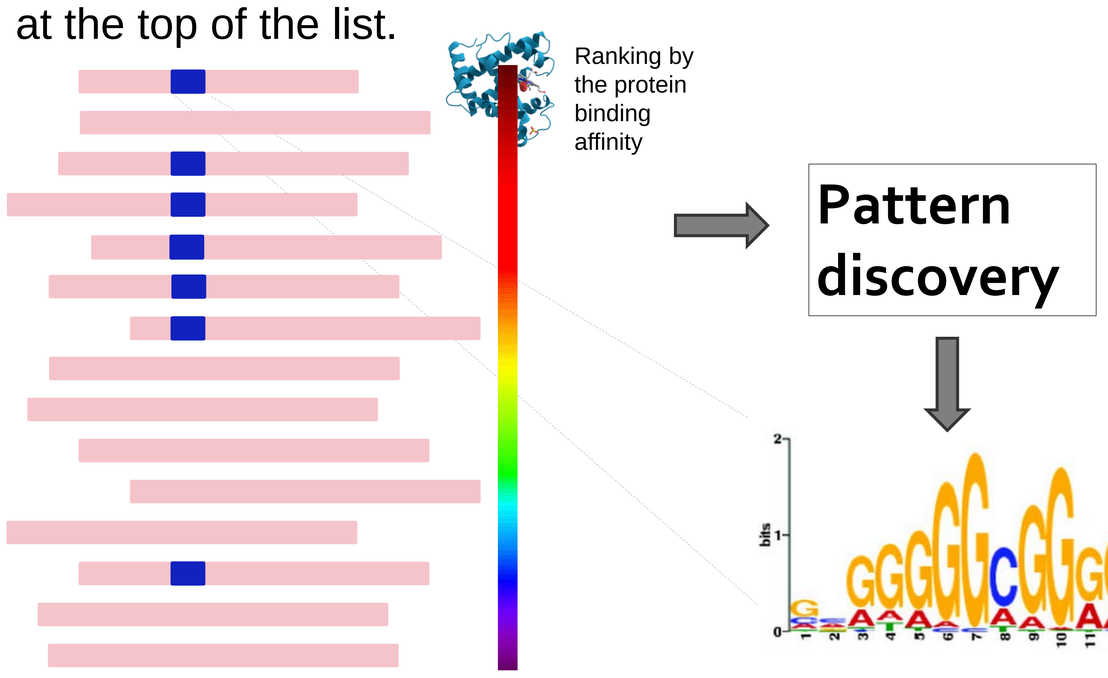
[1] Department of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel
[2] Agilent Laboratories, Petach-Tikva, Israel

Abstract

Statistics in ranked lists is important in analyzing molecular biology measurement data, such as ChIP-Seq, which yields ranked lists of genomic sequences. To assess the enrichment of a position weight matrix motif in a ranked list we use a motif induced second ranking on the same set of elements. Possible orders of one ranked list relative to the other are modeled by permutations. Due to sample space complexity, it is difficult to characterize tail distributions in the group of permutations. In this work we developed tight upper bounds on tail distributions of the size of the intersection of the top of two uniformly and independently drawn permutations and demonstrated advantages of this approach using our software implementation, mmHG-Finder, to study position weight matrix motifs in several datasets.

Motif search in ranked lists

The input consists of a ranked list of sequences, and the expected output is a collection of motifs that are enriched at the top of the list.



Position weight matrices (PWMs)

Position weight matrices are a commonly used representation of motifs in biological sequences.

	1	2	3	4	5
A	0.5	0.25	0.25	1	0.25
C	0	0.25	0	0	0.25
G	0.5	0.25	0	0	0.5
T	0	0.25	0.75	0	0

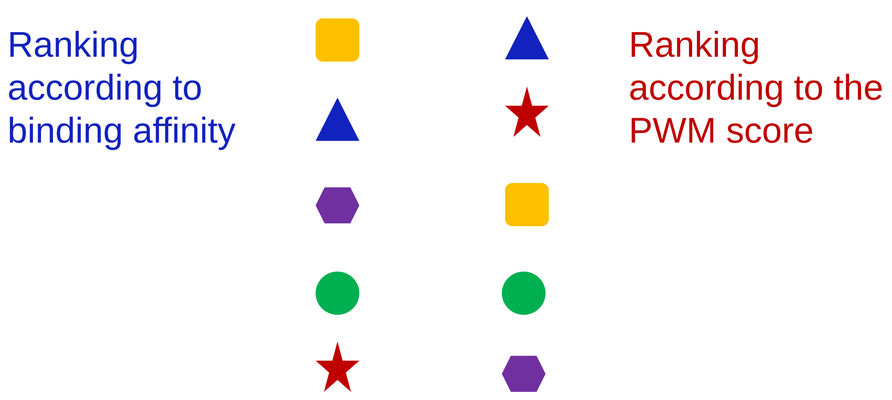
Pattern matching with a PWM gives a quantitative result (a score). For example, the score of **ACAAG** will be:

0.5 x **0.25** x **0.25** x **1** x **0.5**

This definition can be generalized to yield a score for a sequence that is longer than the motif.

PWM motif assessment in ranked lists

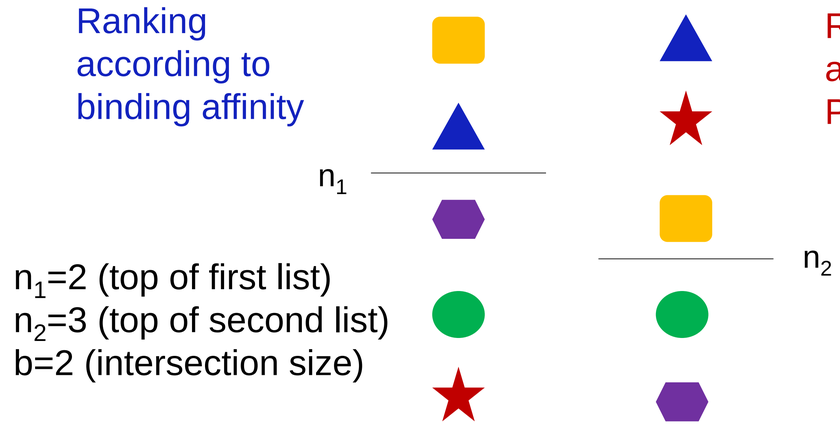
Given a PWM motif, we actually have two rank orders over the sequences.



Observation:
A significant PWM motif would yield high scores for sequences having strong binding affinities.

The mmHG statistics

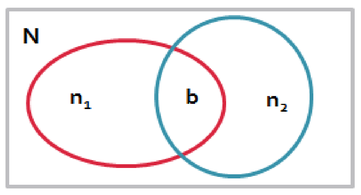
The **mmHG statistic** focuses only on agreement at the **top** of the two ranked lists.



The top of the two ranked lists is found in a data driven manner by testing all the possible partitions.

$$\text{mmHG score} = \min_{1 \le n_1 \le N} \min_{1 \le n_2 \le N} HGT(N, n_1, n_2, b(n_1, n_2))$$

intersection size

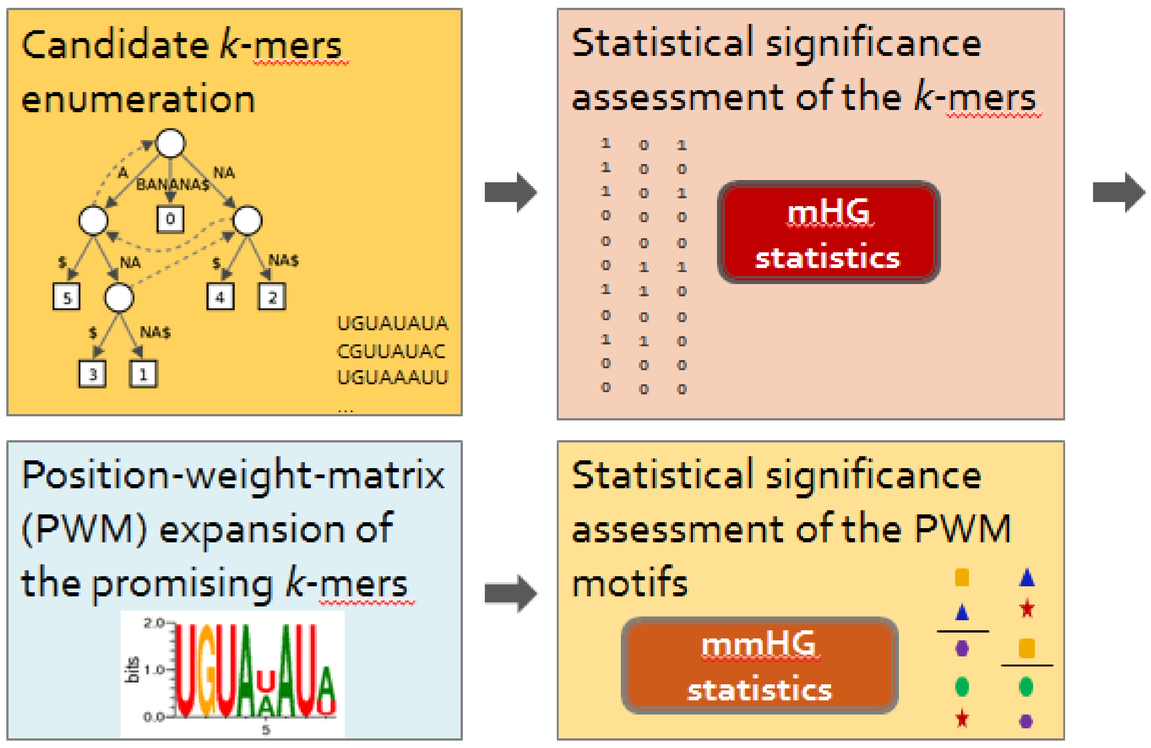


$$HGT(N, n_1, n_2, b) = \sum_{i=b}^{\min(n_1,n_2)} \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}}$$

(hypergeometric tail)

We developed tight upper bounds on the mmHG *p*-value that can be calculated in polynomial time.

Utilizing mmHG for motif search in ranked lists



References:

1. Leibovich L, Yakhini Z: **Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position Weight Matrix Motifs.** In *Algorithms in Bioinformatics. Volume* 8126. Edited by Darling A, Stoye J: Springer Berlin Heidelberg; 2013: 273-286: *Lecture Notes in Computer Science*.
2. Steinfeld I, Navon R, Ach R, Yakhini Z: **miRNA target enrichment analysis reveals directly active miRNAs in health and disease.** *Nucleic Acids Research* 2013, **41:**e45-e45.