Adding New Scores to SeedSearcher

Yoseph Barash, Aviad Rozenhek

February 6, 2008 (version 0.1)

1 Relation to Previous Score Definitions

This note describes adding new scoring schemes to SeedSearcher. In some aspects, it moves away from previous definitions because of the following reasons. First, We are interested in informative motifs with regards to some predefined measure over sequences and/or positions. Historically, we aimed to find discriminative motifs for two groups of sequences, like in a binary classification problem. However, the problem at hand may be more general then the initial settings we considered: There might be more then two groups of sequences, the meaning of the sequence and/or positions pre defined "weights" may be very different and demand different treatment, there might be only partial weights/order defined over sequence/positions, different null hypothesis about the data should be made when evaluating significance of motifs etc. As a result, the definitions of *score system* and *score function* based on the statistics T_p, T_n, F_p, F_n as previously given may not be valid. Moreover, some implicit assumptions made about how T_p, T_n, F_p, F_n are collected may not be valid either. For example, previous definitions assumed each specific sequence contributes both a positive *and* negative weight (e.g., ρ^1 *single position classification* function definition). The only way a sequence did not contribute both type of weights was when it had a $\{0, 1\}$ weight or when using the *Discrete pweight* where a certain threshold t over the weights was used to determine the weights as $\{0, 1\}$. However, in some settings (e.g. Boosting over weak learners) each sequence may have a positive or negative loss attached to it, with a matching weight. Another example is the double sided hyper-geometrical pval test we are also interested in performing.

Following is a description how the current exponential score may be used as a score for matching additive loss function, were we have both positively

and negatively labeled sequences. An adapted, more formal, definitions for all scores will follow in the future.

2 Additive, Gradient Based, Score Definition

The idea in this case is that the weight W^i given in the weights file and attached to each sequence i reflects the values of some gradient for a loss function we are trying to minimize. In this setting we are searching for motifs whose appearances my be used to minimize the loss as given by the gradient. Formally we assume for each sequence we have a value $V^i \in [-M, +M]$ where $M \in \mathcal{R}$ (i.e. the gradient is bounded by some upper/lower real number) and we look for motifs that maximize the additive gain by their presence:

$$\mathcal{L} = |\sum_{i,\text{s.t.} C_m^i \neq \Phi} V^i|$$

Where C_m^i is the set of positions in sequence i where motif m appears. This means we only "care" about a motif presence, and the fact it does not appears implicitly contributes a 0 value to the summation above. Also, note this definition only uses the weight of the sequence and no weights per position.

Since SeedSearher weights files assume $W^i \in [0, 1]$ we need some sort of an affine transformation from original values $V^i \in [-M, +M]$ to the input ones $W^i \in [0, 1]$. First we normalize the original weights to be in the range $W^i \in [-1, +1]$. Then we apply the following transformation:

$$W^i = 0.5(1 + V^i)$$

Note the above transformation holds for both $V^i \leq 0$ and $V^i \geq 0$. Now, we can use the transformed weights W^i in the SeedSearcher program using the already implemented exponential loss:

$$e_{\alpha,\beta}(T_p, T_n, F_p, F_n) = -\log_{10} Explos s_{\alpha,\beta}(T_p, F_p)$$

Where exploss is defined as:

$$Explos s_{\alpha,\beta}(T_p, F_p) = \frac{\beta^{F_p}}{\alpha^{T_p}}$$

Using any $\alpha = \beta$ would yield the desired score and using $\alpha = \beta = 10$ would give exactly $\mathcal{L}$ as defined above. Note, however, that for this to work the

definition of the T_{p_i}, F_{p_i} statistics must be changed:

$$T_{p_i} = \begin{cases} 2W^i - 1 & \text{if } W^i > 0.5 \\ 0 & \text{if } W^i \leq 0.5 \end{cases}$$

$$F_{p_i} = \begin{cases} 0 & \text{if } W^i > 0.5 \\ 1 - 2W^i & \text{if } W^i \leq 0.5 \end{cases}$$

Unlike previous definition, here each sequence contributes to F_P *or* T_p, but not both. The decision on which it belongs to is based on the "traditional" 0.5 threshold, with a contribution proportional to its given weight.

There are a few more things worth noting about the additive score. First, we can easily define this score sum over only positive/negative values. Second, note the description above did not contain any reference to position specific weights and assumed "gene count" rather then "total count". Although both can, in principle, be combined in the computations described above, they do not make much sense in the settings we consider here. Thus, the above computation refers to setting SeedSearcher to search with "–Sscore-partial=on" (and not "hotspots" or "off") and "–Scount=gene". Finally, the above description did not touch on the question of the statistical significance of a specific motif's score. We discuss this in the following section.

2.1 *P*-value Computation for Additive Scores

In this section we are interested in evaluating the statistical significance of a given additive score, computed as described above. To answer this, using a frequentist approach, we first need to define what is the null hypothesis. Before we do this, we start by describing the settings were the p-value of an additive score may be of interest. First, we note the difference from scoring motifs simply by additive scores as in the previous section. There, the motivation came mainly from gradient based computations. In this case we are interested in optimizing some target function and the question of the segnificance of a score in this setting is not clearly defined. However, the p-value of an additive score may be of great interest in other settings. For example, we might have a set of positively and negatively labeled sequences. Each sequence might have some weight $\in [0, 1]$ attached to it, and each position in each sequence might have some weight $\in [0, 1]$ attached to it. Such a situation might arise in a case where we have two "modules" or

regulation, and a given sequence has a prior belief of belonging to either of them. In a simpler case, this prior is $\in \{0, 1\}$, i.e. each sequence has a "hard" assignment to one of these groups. We also have some prior belief about a motif binding a specific position/area. This can come from ChIP tiling arrays, or from evaluating RNA secondary structure. In our example we look for a motif that tends to appear in the more likely binding areas in *one* of the groups, but not the other, and hence be a good candidate to explain differences in regulation. A somewhat simpler situation may involve only a single group of sequences and some weights assigned to each sequence and/or each position. This can be, for example, the results of genomic ChIP tilling arrays. In this case we are interested in a motif whose occurrences correlate well with highly weighted sequences/positions.

We now come to the question of defining a null hypothesis to test against. For this, we note the following: We search over a discrete set of motifs. Motifs occurrences are, basically, zero/one events, i.e. either the motif appears or not. However, each occurrence (i.e. a non zero event) is associated with a specific weight as we just described. A natural question to ask is therefore the following: "Given that a motif occurred K times, what is the probability of its occurrences "hitting" positions with a total additive score (as previously defined) as good as S ?" Note that if we set the weights of all positively and negatively labeled sequences/positions to 1 and 0 respectively, the problem is reduced back to the hypergeometric p-value. Also note that this is *not* the same as asking the question "What is the probability of getting a score as good as S?". To answer the later, we first need to assign a prior over any occurrence count $K = k$ and then sum over the answers to the first question we described. Another thing to note when it comes to assessing the statistical significance of a result, is that the above questions concern a *single* sample of a motif. However, just like when scoring motifs using hypergeometric pvalue, we are interested in the significance of a score when we try many different motifs (which we find during our search). This means we need to take into account multiple hypothesis testing, which is effected, via the definition of the motif search (via the motif "complexity" -length and number of wild cards, and the number of random projections), by the number of motif we evaluate and the correlations between them.

An obvious solution to the statistical significance of scores is by evaluating scores for many times from shuffled data sets. If we shuffle the labels (i.e., the weights assigned to each sequence and the weights within each sequence) we can answer the second question we postulated above, for any given motif

complexity/search definition. We will describe other, more efficient, options to estimate the p-value later.

3 Discriminative Scores for Binary Classification with Partial Weights

This section handle the case were we have weights assigned to each sequence and/or each position, and we also have a binary label attached to each sequence. The label of the sequence is deduced from the weight of the sequence, given in the input weight file, as described for the additive score above.