White Paper: Motif Discovery Methods

Introduction

This white paper describes the methods used in Partek® Genomics Suite™ (Partek GS) to search for motifs in a set of genomic regions. Partek includes tools to search for known motifs and discover new motifs de novo. The sections below describe each of these tools. For information on how to apply these tools to your data, see the Chip-Seq tutorial located under the Regulation tab on the Partek Tutorials website.

Motif Search

Given a set of genomic regions, Partek can search for patterns provided by the user. These patterns can be either a user-provided string or a user-provided alignment matrix, such as those available in the *JASPAR* database (Bryne, et al.).

String Search

The string search tool will return all positions in the set of genomic regions that match the given string. The string match is case insensitive, meaning if you search for *ATCG,* you may get *atcG* as a match. Nucleotides that are lower case have been "repeat masked", meaning they are located in a repetitive region of the genome. Your search string may contain any of the characters from the IUPAC nucleotide code (http://www.bioinformatics.org/sms/iupac.html). For example, if you search for *WAAA,* you may get back *AAAA* or *TAAA* (or any variation of upper and lower cases), since *W* represents *A* or *T*.

Alignment Matrix Search

Alignment matrices are matrices of nucleotide counts per position (Hertz & Stormo). They are often used in literature to model transcription factor binding sites. Each instance of the motif is aligned to each other and the number of nucleotides at each position is counted and summarized in an alignment matrix. All positions from the set of genomic regions are scored against the alignment matrix. The score represents how likely the position is an instance of the motif. A quality cutoff is used to determine which sequences in the regions are instances of the motif. The scoring scheme and quality cutoff are similar to Schug & Overton and is briefly described below:

Let *M* be a motif of length *L* consisting of *N* motif instances. Let *A* be a *4xL* alignment matrix such that $a_{i,j}$ is the count of letter *i* at position *j*. Let B_i be the background frequency of letter *i* (calculated as the number of nucleotides *i* in the regions divided by the total nucleotides in the regions). Let *S* be a sequence of length *L*. The score of *S* given the alignment matrix is

$$L_A(S) = \sum_j \left[\ln\left(\frac{a_{S(j),j} + b_{S(j)}}{N+1}\right) - \ln(b_{S(j)}) \right] \quad (1)$$

Let *h* be the maximum of L_A. The quality score of a sequence is calculated as $Q_A(S) = L_A(S)/h$. A quality score of 1 corresponds to a sequence with the most likely base at each position of the alignment matrix. In the Partek dialog, you are asked to specify a threshold Q_A. All sequences that have a score $T_A > Q_A * h$ will be returned in the resulting spreadsheet.

The probability $P_{Expected}$ of a sequence having a score above T_A is calculated under the assumption that the bases are *i.i.d.* according to the background distribution *B*. Let N_{Trials} be the number of sequences compared to the alignment matrix. The expected number of occurrences of the motif in the regions is $P_{Expected} * N_{Trials}$. The p-value of observing N_{Actual} instances with a score above T_A is calculated based on the binomial distribution, where N_{Trials} is the number of trials and $P_{Expected}$ is the probability of success. A low p-value indicates that the regions are enriched with instances of the motif.

De Novo Motif Discovery

Motif discovery is done using Gibbs motif sampling. Partek's implementation of the Gibbs motif sampling is based on Neuwald, et al. The Gibbs sampling method is a stochastic procedure that attempts to find the subset of sequences within the regions that maximizes the log likelihood ratio (LLR)

$$LLR = \sum_i \sum_j \left[\ln\left(\frac{a_{i,j} + b_i}{N+1}\right) - \ln(b_i) \right] \quad (2)$$

This is done by repeating the below two steps until convergence:

A. Given the alignment matrix from Step B, search for locations in the set of regions that score highly compared to the alignment matrix using equation (1).
B. Create a new alignment matrix from the set of high scoring positions from Step A and return to step A.

The Gibbs sampler is run on a range of motif sizes specified by the user. The motif with the greatest average *LLR* (*LLR* divided by length) is returned. To find *N* motifs in a set of regions, the Gibbs sampling method is run *N* times. The motif instances found from the previous run of the Gibbs sampler are removed before performing the next run.

References

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., Sandelin, A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 2008 Jan; 36(Database issue):D102-6.

Hertz, G.Z., & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999, 15, 563-577.

Schug, J., & Overton, C.G. TESS, Transcription Element Search Software on the WWW. *http://www.cbil.upenn.edu/tess/techreports/1997/CBIL-TR-1997-1001-v0.0.pdf*.

Neuwald, A.F., Liu, J.S., & Lawrence, C.E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science* 1995, 4: 1618-1632.