MINING DNA NETWORK EXPRESSIONS WITH EVOLUTIONARY MULTI-OBJECTIVE PROGRAMMING

Manuel Belmadani, M. Computer Science (specialization in Bioinformatics) Candidate
mbelm006@uottawa.ca

Abstract

High-throughput sequencing experiments, such as ChIP-seq, identify protein interactions within large samples of DNA. Such experiments may identify and output thousands of peak scores where an interaction of interest is regulated by that region of DNA. Computational biologists often require a de novo analysis of these peaks to mine short overrepresented patterns, frequently found amongst the sequences. This task is referred to as motif discovery. Existing tools struggle at motif discovery if the number of input sequences increases. The time and complexity costs of calculating pattern occurrence often compromises on runtime, quality, or length of the output predictions.

The proposed solution (MotifGP) is a motif discovery tool driven by a Strongly Typed Genetic Programming (STGP) algorithm equipped with multi-objective optimization. STGP is a stochastic search process based off evolutionary algorithms which evolves candidate solutions as typed checked programs. Multi-objective optimization allows the algorithm to evolve according to multiple metrics used to build predictions that are high in both support and specificity. The software produces a collection of non-dominated multi-objective solutions, which can then be aligned to reference DNA motif databases (such as JASPAR, TRANSFAC, and Uniprobe) in order to validate/identify predictions.

Methodology

To determine if MotifGP can outperform DREME, we compared both software against 13 datasets from mouse Embryonic Stem Cells ChIP-Seq experiments [1]. The recorded runtimes for each DREME run were used as time limits for MotifGP. Each dataset may contain one or more motif patterns. Desirable solutions should report local overlap and the lowest E-values once aligned to known motifs.

MotifGP matches candidate solutions (network expressions) against a set of input and control sequences. We use the match count on each set to compute two metrics of discrimination and coverage:

\[
\text{Discrimination} = \frac{P - C}{k + C}
\]

\[
\text{Coverage} = \frac{P}{k + C}
\]

For a given pattern, let P be the set of matched input sequences, C be the set of matched control sequences, and k be the total number of input sequences.

Discrimination favors solutions that are not highly represented in the control sequences, while Coverage maintains solutions that are found in multiple input sequences.

Experimental Results

A state-of-the-art motif discovery tool, DREME, predicts DNA motifs by mining and refining a smaller set of patterns (words) from an input set of sequences [2]. The predictions are compared to known motif databases with TOMTOM [5]. Reported alignments are listed and scored by E-value.

While DREME predictions have significant E-values, the pattern length is limited to 8 characters. Some known motifs are more than twice the size DREME can predict. This is a restriction set to keep low runtimes under real ChIP-seq datasets. We propose our software, MotifGP, as an alternative to this limitation. Our goal is to explore the potential of multi-objective evolutionary computing in finding significant motif predictions on large datasets, under runtimes comparable to DREME.

As shown by the Klf4 dataset prediction, MotifGP has the ability to overlap an entire target motif. It even finds additional nucleotides attached to the left end of the predicted motif, found in the input sequences.

Conclusion and Future Work

Combining STGP and multi-objective optimization directs the solution search towards motifs that align with high confidence and fully overlaps known motifs from databases.

Other existing ChIP-seq experiments can be revisited using MotifGP as it may be able to uncover subtle patterns that other tools could not.

Future development of MotifGP will investigate different multi-objective fitness function variants and explore other evolutionary algorithms.

Background and Motivation

A di de novo motif discovery pipeline

DNA Motif Discovery

Motif discovery: Mining DNA sequences for motifs. Can be used to identify TF binding sites.

References


DISTRIBUTED EVOLUTIONARY ALGORITHMS IN PYTHON

Motif Comparison Tool

Tomtom 13.03.15 13:33

MotifGP's motif fully overlaps the 19 bases of CTCF. Each tool's prediction targeted a different strand (orientation), showing a complementation between each methods.