A novel algorithm for discovering transcription factor binding sites

Robert Gardner*, Sascha Ott* and Vicky Buchanan-Wollaston*^
*Warwick Systems Biology, The University of Warwick, Coventry, CV4 7AL, ^Warwick HRI, Wellesbourne, Warwickshire, CV35 9EF

1. Background & Objectives

- Since the advent of the microarray, gene expression over the course of biological processes can be seen.
- Experimental determination of transcription factors is possible, but slow and expensive: computational methods usually search motifs or gene profiles.(1, 2)
- We aim to develop a new Matlab tool to search for sites (including pairs) that explain co-regulation.
- Search for *Arabidopsis* leaf senescence motifs (3).

2. Transcriptional Modules (TMs)

- Groups of genes that are co-regulated in a biological process, by a set of transcription factors.
- Determined by both their promoter sequence information and time series expression profile.

ATCAGTGACTTTAACACGGAT
CGAACGTTGATTACGTTGCGT
GAAGCTTCTCGGCATTATGCA

3. Statistical Sampling

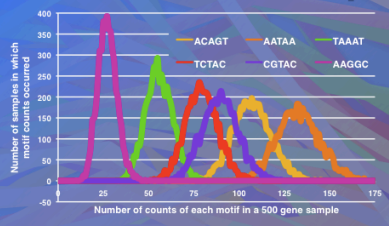
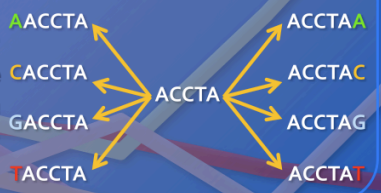


Fig. 1. Bootstrapping returns for six 5-mers in 500 gene proposed TMs.

- 5000 samples were taken of the sizes of the TMs generated, to give bootstrap statistics for mean and variance of each motif for each cluster size.

4. Elongation & Repetition

- Determine the most significant motifs via normal approximation (as per Fig. 1.) and extend these in each direction.
- Use sparsity indicator vectors to reduce the size of the search when considering the longer motifs.



INITIATE: Propose all motifs of a given length *k*, as initial putative binding sites.

→ Propose new TMs from the coupled data, using the adapted Kundaje algorithm (4).

→ Sample from full pool of motif count data sets of proposed transcriptional module sizes.

→ Compare real motif counts for each of the proposed TMs, with bootstrapped values.

→ **TERMINATE:** Retain putative motifs with highest relevance to the latest proposed TMs.

← Remove all bar the *N* motifs with greatest specificity for certain proposed TMs.

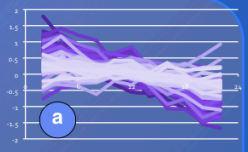
← Generate new proposals for motifs via the extension and pairing of the retained motifs.

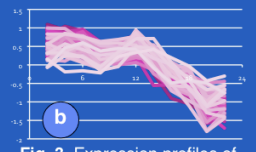
← Calculate the number of each proposed motif present in the promoter of each gene considered.

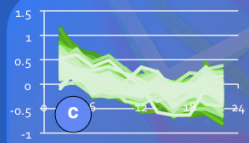
Fig. 2. A flow diagram of the steps taken by the algorithm, to determine the putative transcription factor binding site motifs

5. Results

- Returned TMs were based upon many micro-satellite repeats.
- Removing pair-repeat 5-mers did not prevent this, but proposed AACCGGTTT as a membrane transport related motif.
- GOstat (5) study of TMs (Fig. 3a, b) revealed more relevance for *k*-means (Fig. 3c, d) clustering than our algorithm.







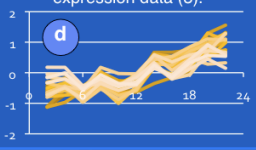


Fig. 3. Expression profiles of (a,b) Proposed TMs; (c,d) *k*-means clusters, for *Arabidopsis* leaf senescence time course expression data (3).

6. References

1. Spellman, P.T. *et al.* (1998) Comprehensive Identification of Cell-Cycle Regulated Genes of the yeast *Saccharomyces cerevisiae*. *Mol. Bio. of the Cell* **9** (12) 3273 - 3297.
2. Tavazoie, S. *et al.* (1999) Systematic Determination of Genome Network Architecture *Nature Genetics* **22** 281-285.
3. Thanks to the Buchanan-Wollaston Warwick HRI group for their *Arabidopsis* leaf senescence gene expression data set.
4. Kundaje, A. *et al.* (2005) Combining Sequence and Time Expression Data to Learn Transcriptional Modules. *IEEE/ACM* **2** (3) 194-202.
5. Beissbarth, T. & Speed, T. P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20** (9) 1464-1465.

Poster and project work by:
Robert Mark Gardner
r.m.gardner@warwick.ac.uk
Systems Biology MSc Student

THE UNIVERSITY OF WARWICK
SYSTEMS BIOLOGY DOCTORAL TRAINING CENTRE
Warwick HRI
EPSRC Engineering and Physical Sciences Research Council
bbsrc

Project supervised by:
Sascha Ott and Vicky Buchanon-Wollaston.