Can we improve on the identification of Transcription Factor Binding Sites?

Hugh Shanahan & Piotr Byzia,
Department of Computer Science,
Royal Holloway,
University of London

StReBio 2009
28th June 2009
Outline I

• The underlying Biology

  • Why is regulation important?

  • Transcriptional regulation - basic mechanism

  • Old data - *in vivo* (co-expressed genes) and *in vitro* (Sage-Selex)

  • New data - ChIP-chip, Next Generation Sequencing
Outline II

• Present methods
  • Over-representation of words.

• String kernel approach
  • Simple classifier

• Where to go from here.
Regulation - very general comments

- In essence Systems Biology = Understanding Regulation (personal opinion)

- For eukaryotes, increasingly complicated mechanisms of regulation
  - Transcriptional: regulate transcription of genes to mRNA
  - Post-transcriptional: alternative splicing, siRNA’s, ....
  - Translational: controlling amounts of tRNA (very exotic !)
  - Post-Translational: Degradation of proteins, phosphorylation, ...
  - Metabolic: hormone signalling, biosynthesis through pathways, ....
Underlying Biology of Transcription

• Other mechanisms are important (typically faster) but regulation at transcription is “lowest level”

• Proteins called Transcription Factors bind to (normally) regions upstream of coding regions.

  • Transcription Factor Binding Sites

• Binding Regions 5-20 bases in length

• 10-15 TFBS per gene (up to 50) for Eukaryotes

• transcription factors can be behave cooperatively or antagonistically.
Model for transcription
Note

• Will not discuss other proteins that bind to DNA called histone deacetylases which can open up non-exposed DNA for binding to transcription factors and transcription mechanisms.

• Potentially, much of the techniques we learn for transcription factors could be applied for the above.
Crystal structure of transcription factor-DNA complex
Trascription Factor Binding Sites

- Protein-DNA complex is a very detailed arrangement of many atoms

- No one clear combination of bases that will bind to a transcription factor

- Binding sites typically represented as propensities.

- Represent as single site frequencies.

- Does not take into account of inter-site correlation.

- At the very least, approximately palindromic solutions.

- But also direct correlations possible.
Position Weight matrix for Zinc Finger
Example of resulting Position Weight Matrix

Nuclear Receptor
Fundamental difficulty

- Why not use evolution?

- Non-coding regions evolve much more quickly than coding regions.

- Regions upstream of closely orthologous regions bear very little resemblance to each other.

- Proteins stay very much the same, but their regulation changes quickly.

- Standard sequence alignment methods do not work.

- Looking at closely related sequences (phylogenetic footprinting) is making progress here.
Multiple alignment of upstream regions with known promoter region.
Present methods

• Search for over-represented words.

• Models are based on
  
  • looking at background frequency of nucleotides and then asking what words are over-represented in a set of sequences. Construction of Position Weight Matrix.

  • Some use of phylogenetic information plus extras (e.g. methylation patterns)

• Huge number of different implementations (over one hundred different methods suggested)
Data sets (up to recently)

- Most work is based on co-expressed data.
  - Genes which exhibit co-expression in a particular experiment.
  - Assumption: Co-Expression = Co-Regulation = Same set of TFBS’s.
- Sage-Selex data
  - *In vitro* data where single transcription factor binds to specific regions.
  - Small data set (10’s of transcription factors)
Example of co-expression
Results of co-expression

• Huge number of false positives (10^6 identified 10^3 actual)

• Futility Theorem :- any computationally derived TFBS has no functional significance

• Trying to disentangle
  • multiple TFBS’s
  • cut-off criteria for significant word

• At present most methods do not take into account of inter-site correlations.
New Data Sets - here comes the flood....

- Chromatin Immunoprecipitation (ChIP) allows us to examine (in vivo) which transcription factors bind to what regions.
  - We do not know what other transcription factors are co-regulating.

- Next Generation sequencing has substantially reduced the cost of sequencing.
  - Gathering genomes will become easier and easier.
    - 1000 Genomes project :- Genomes of approximately 1000 humans from all around the world (genetically diverse).
    - Upstream regions can be explored of very closely related species.
Challenges

• Much more data coming in the near future.

• Accurate estimates of binding regions for small but growing set of transcription factors - ChIP-chip data

• Should have sets of sequences which bind to a specific transcription factor or related due a similar phenotype (next generation sequencing)

  • Can we identify other co-regulating TFBS’s?
String Kernel approach

• Basic idea has been around since turn of the century.

• Applied in Bioinformatics from 2003 in

  • protein family assignment
  
  • splice site identification
  
  • siRNA identification
Feature vector

• Define Alphabet of nucleic acids

• Determine all k-mers of length from the above alphabet of fixed length N.

• Allow for up to one mismatch - i.e. construct sets of k-mers that differ by only letter (allows for variability in sequence).

• Determine frequency of k-mers (with or without mismatch) for each sequence. Use this as the feature vector.

• At present do not include positional data.

• Gaussian/Linear Kernels give similar results.
Classification

- Use Sage-Selex data as clear benchmark.

- Generate negative data set by taking random upstream regions from human genome.

- Construct SVM to classify data.

- Compare with results using
  - Hidden-Markov-Model.
  - Standard motif detection technique (MEME/MAST).
which was the main motivation for the SKM development. The aim is to provide a set of simple steps to generate a classifier from a training set and run the classification on a set of sequences.

**asic workflow steps in SKM**

1. Fix (vide input data into proper groups, positive and negative training sets and positive and negative test sets)
2. Create permutations with replacement from alphabet and of k-mers length
3. Calculate which and how many of permutations are present in our training set and store that information in global motifs vector
4. Perform the same calculation between global motifs vector and each of our test sequence
5. Train SVM with training set
6. Prediction phase and accuracy calculation

Workflow might change depending on what data is provided. Whether or not user has his own test set or wants to use precomputed SVM model. Use SKM user guide for the reference.

**METHODS**

SKM is a package written in the Python language and uses jupyter notebooks to handle sequence data. SVM classification is accomplished by LibSVM package from Lin and others. Libsvm version uses 5 types of SVM, from which linear and classification SVM type has been used in this study. There are also 5 types of kernel functions, two of them used in this study are linear and radial. In case of radial kernel, there has been cross validation run to obtain best parameters in SVMs and gamma in radial kernels. After training our SVM, a prediction phase is run on the test set and accuracy is calculated. Accuracy = (TP + TN) / (TP + FP + TN + FN)

This will be the main indicator in our benchmark. Additionally, all accuracies are mean of 5 runs of workflow.

There are several options to choose from in the SKM's workflow:

- fixed length k-mer
- range of k-mer length
- include mismatch
- size of test set when using shuffling TP and or TN to obtain it
- alphabet
- precomputed model file

All intermediate steps are logged so that any concerns should arise it is convenient to investigate the whole process. Results are stored in a separate file where final scores are provided including ST description together with SVM classification verdict. In case where no user's test set was provided there is accuracy for each kernel provided.

Package is easily extendible and well documented if anyone wish to prepare own workflow or add functionality.
Where do we go from here?

- Simple model shows that string kernel approach is more accurate than standard approaches.

- Key element is picking the right data set. If we started out with standard co-expressed sets, unclear what the results would be. This provides a clear benchmark.

- Choice of Kernel at present does not consider longer-range correlations.

- Only examined case with single Transcription Factor Binding Site.

- Multiple transcription factor binding sites are crucial - kernel will be updated accordingly.