Identification of Functional Transcription Factor Binding Sites using Closely Related Saccharomyces species

Scott W. Doniger1, Juyong Huh2, and Justin C. Fay1,2
1Computation Biology Program and 2Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA

Presented by John Brothers II
Journal Club Group Member: Rong Hu

Bioengineering and Bioinformatics Summer Institute 2005
University of Pittsburgh, Carnegie Mellon University, Duquesne University, and the Pittsburgh Supercomputing Center

Cells control the expression of genes using Transcription Factors. TFs are regulatory proteins that:
– Bind to the promoter region of DNA
– Can either promote or inhibit expression

Transcription Factor Binding Sites
– The short DNA sequences where TFs bind to the promoter

Comparative Genomics and Transcription Factor Binding Sites
– A major goal of bioinformatics and computation biology is to model the complex process of cell regulation.

Comparative Genomics
– Compares sequence conservation between species
– Quick way of identifying functional DNA

Using Comparative Genomics to identify TFBSs.
– Many TFBSs are conserved
– TFBSs make up a significant amount of functional DNA

Why Use Yeast to Find the TFBSs?
– Lots of Experimental Data about Yeast Regulatory Motifs!
– Identification of many Regulatory Motifs

Multiple Comparative Genomic Approaches!
– New regulatory motifs discovered in yeast through their conservation in other yeast species

Combining the data from both approaches
– Now possible to begin identifying functional TFBSs
**Difficulties with Identifying TFBSs**

- The sites are short, degenerative, and occur frequently.
- This leads to the identification of many false positives, e.g., it looks like a binding site but does not actually function.
- Way around this problem:
  - Functioning sites are conserved between yeast species.
  - Use comparisons to limit number of false positives found.

**Discovering Conserved Sequences**

- Molecular Evolutionary Models
  - Probabilistic Framework
  - A neutral evolutionary model estimates the rate of substitutions that would occur without selection or functional constraint.
  - Constrained and Unconstrained Sequences can then be distinguished by comparing to this model.
- Functionally constrained sequences
  - Have fewer expected substitutions.
- TFBSs in yeasts have a very low likelihood of being conserved in the absence of functional constraint.

**Calculating the Neutral Model**

- Calculating the neutral expectations for TFBS conservation
  - The neutral evolution probability of TFBS conservation is written as:

\[
P(\text{conservation}) = \sum_{j=1}^{4^w} P(X)^j P(Y)^j P(Z)^j \delta(\text{match}_j)
\]

- \( w\) is the number of bases between sequence \( s\) and the starting sequence of TFBS.
- \( j\) is the number of transitions.
- \( z\) is the number of transversions.
- \( P(W)\) is the probability of a base remaining the same over the given evolutionary distance.
- \( P(Y)\) is the probability of a transition.
- \( P(Z)\) is the probability of a transversion.
- Probabilities calculated using Kimura’s two-parameter model. \( W\) is the width of the motif.
- \( \delta(\text{match}_j)\) is a binary function which returns 1 if the sequence was a significant match to the PWM.

**Diagram of Method used to Calculate TFBSs Conservation Probability**

- This calculation shows the probability of conservation when the sequence can transition to another sequence or stay identical over evolutionary time.
The Hypothetical Two Base Pair Motif

This is calculated using an entropy equation:

\[ \text{Letter Height} = \max \text{ entropy (in this case 2)} - \text{ entropy} \]

To find entropy:

\[ \text{entropy} = -\sum P(b) \log_2(P(b)) \]

In the case of P(T) and P(C): Each will occur half the time, therefore \( P(b) = \frac{1}{2} \).
Inserting \( \frac{1}{2} \) into the equation for the bases, the entropy is 1.

Therefore the height is \( 2 - 1 = 1 \).

Results: TFBSs Conservation

- TFBSs are rarely conserved by chance
  - Neutral Evolution Model Substitution Rate: 0.83 substitutions per site across the three species of yeast (Saccharomyces cerevisiae, S. paradoxus and S. mikatae respectively)
  - The probability that a 10-base pair sequence is identical across the three species of yeast was 0.22%
- Therefore, if the possible TFBS occurs amongst the multiple yeast species, there is a significantly high chance that, according to this model, it will be functional.

Results: Identification of Functionally Constrained Ndt80 and Ume60 using neutral evolution model

- Ndt80 and Ume6 regulate meiosis-specific genes in Saccharomyces cerevisiae
- Functional Ndt80 and Ume6 binding sites in other species would be under functional constraint
- Position Weight Matrices were created to identify these TFBSs in the other two species of yeast
- For each Ume6 and Ndt80 TFBS identified, the sequence differences were counted.

Results: Identification of Functionally Constrained Ndt80 and Ume60 using neutral evolution model

- The counted results were compared to both experimental data and the neutral evolution model rate of substitution.
- The neutral evolution model expected a substitution rate of 4-5 substitutions per TFBS between the species.
What does the conservation of Ume6 and Ntd80 sites mean?

- Strong indicator that functional TFBSs are under functional constraint
- Supports this comparison model, which can be used to scan the entire genomes to discover other TFBSs

Experimental Verification

- Mutation of the Ume6 and Ntd80 binding sites
  - Significant changes in expression occurred
  - These were specifically Ume6 and Ntd80-dependent changes
  - This verifies that these are functional binding sites

Comparative Genomics Limitations for Identifying TFBSs

- The TFBSs found were a conservative estimate
  - Many false negatives
    - (e.g., functional TFBSs that were not identified)
- Model relies on assumption that functional elements will be shared between species
  - Unique Non-conserved TFBSs
  - Verified Experimentally
- Assumption of Mutational Homogeneity
  - Neutral Evolution Model relies on consistent genome-wide mutation rate
  - Some areas have higher and lower rates of substitutions

Comparative Genomics Limitations for Identifying TFBSs (continued)

- The assumption that substitutions are independent of the other positions within a TFBS
  - Position Independence is widely assumed, but has been proven false for some binding sites
- TFBSs are assumed to be conserved due to functional constraint alone, but there are other possible reasons:
  - Overlap of other functional non-coding regions exist
  - Some TFBSs even overlap each other, creating ambiguity and difficulty identifying which TFBS is causing the functional constraint
- Some errors due to misalignment are unavoidable.
Conclusions

- It is possible to discover yeast TFBSs using computational models with experimental validation.
- The model used in this paper discovered a conservative amount of TFBSs:
  - Many false negatives.
  - Less likely to have false positives.
- Comparative genomics is limited in its ability to find unique TFBSs between species:
  - Unique TFBSs are hypothesized to account for the differences between species, and this needs to be looked into.
- "The ability to identify TFBS gain, loss, and turnover between species will be crucial to our understanding of the role transcriptional regulation plays in evolution."

References