Assessment of Genome-Scale Predictions of the Transcription Factor Binding Sites of Cys$_2$His$_2$ Zinc Finger Proteins in Yeast Using the FOOTER Algorithm

Presented by John Brothers II
Mentor: Takis Benos

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

Zinc Finger DNA-Binding Domains

• Cys$_2$His$_2$ zinc finger proteins are located on a majority of transcription factors.

• A major goal of bioinformatics and computational biology is to model the complex process of cell regulation

• Key to that understanding is being able to computationally predict zinc finger binding sites
Zinc Finger Binding Sites

- Example of a 3-finger zinc finger transcription factor

- The critical amino acids are at the -1, +3, and +6 sites on each zinc finger domain

Zinc Finger Binding Sites

- 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 is non-functional

- Way around these problems:
  - Functioning sites are possibly conserved between yeast species
  - Use comparisons to limit number of false positives found
The FOOTER Algorithm

• Uses comparative genomics to predict regulatory regions
  – Originally used to find DNA binding sites between the human and mouse genomes

• Now being used to find the regulatory regions for 2- and 3- zinc finger proteins between multiple yeast species
  – *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Candida glabrata*

The FOOTER Algorithm

• Is a quantitative phylogenetic footprinting algorithm
  – Phylogenetic footprinting is a comparative genome analysis used to identify evolutionary conserved “signals”

• Compares the promoter regions of the orthologous proteins between the yeast species

• Assigns probability scores for each binding site based on the:
  – Distances between the positions of the binding sites between species
  – Generated PSSM models based on binding preferences for each zinc finger transcription factor.
Finding the PSSM Models

- First: Found identities of amino acids at the -1, +3, and +6 positions on the zinc finger domain

\[
\text{ACE2\_YEAST/ } \text{K.AFVRN...HDLIRHKIS......H} \\
\downarrow -1 \downarrow +3 \downarrow +6
\]

- The three contact C2H2 model developed by Benos et al was used to determine the position specific preferences of the protein

How FOOTER Works

Species 1 promoter

TCAAGTTACTTTGAAACCTCTTTTAGG
ATTAACATTCAAGGGACAGCCACCAA...

Compare to PSSM models of zinc finger binding preferences to compute Binding Sites

Species 2 promoter

AGGGAACATTTAGAGGAAAATTTAGG
AATATTTAAAAGATTCATCAAAAATTT...

Identification of conserved regions

S. cere. S. pombe

S. cerevisiae putative TF binding site S. pombe putative TF binding site

Compare to PSSM models of zinc finger binding preferences to compute Binding Sites
Methods

• All yeast zinc finger protein domain sequences obtained from PFAM database (http://pfam.wustl.edu/)

• PSSMs generated only for 2- and 3-zinc fingered transcription factor binding sites

Methods

• Identified proteins that contained only 2 or 3 zinc finger domains

  Good -> ACE2\_YEAST P21192 Metallothionein expression activator. [770 residues]

  Not Used -> AZF1\_YEAST P41696 Asparagine-rich zinc finger protein AZF1. [914 residues]

• Created the PSSMs
Methods

- Downloaded the proteomes from GenBank through the NCBI

- Needed to identify the orthologues between the species to be used by FOOTER

- Orthologues are functionally related genes with sequence similarity which indicates they had a common ancestor

Methods

- Method to find the orthologues between the three yeast species
  - Used BLAST
  - An *S. cerevisiae* protein was blasted against the *S. pombe* proteome.
  - The top hit returned was blasted back against the *S. cerevisiae* proteome
  - If the top hit returned was the original *S. cerevisiae* protein, they were considered orthologues.
  - This was then done between *S. cerevisiae* and *C. glabrata*
Methods

• Promoters were considered to be the sequence starting 1500 base pairs upstream from the transcription start site

• Downloaded the GenBank genomes and used tBLASTn to blast both possible DNA strands for the location of each protein

• Recorded the promoter data upstream from the protein

• Checked promoter regions against S. cerevisiae promoter database and confirmed that the correct promoters were being located

Methods

• Total amount of orthologues found in each species: 2890

• Found the promoter regions where the binding sites would exist

• Fed this data into the FOOTER algorithm
FOOTER Results

• Noticed that FOOTER was returning a lot of binding sites with very low scores, but confirmed that it was running correctly

• The binding sites and the promoter regions between the yeast species seemed to be less conserved than originally thought

• Lowered the scoring threshold to find more binding site matches, but increased the risk of having more false positives

Comparing Footer to ChIP data

• What is ChIP-chip data?
  - Chromatin Immuno-Precipitation (ChIP) is an assay used to determine whether proteins including transcription factors bind to a particular region on the chromosome

• Specifically, it links transcription factors to specific promoter regions and while it does not tell what exactly the binding site sequence is, it can be used to verify the FOOTER results
ChIP

1. Use formaldehyde to fix transcription factors to DNA binding sites
2. Sonicate the cells to lyse the cells and break up the DNA into 500 bp fragments
3. Immunoprecipitate using antibodies that attach to the specific transcription factor you are analyzing
4. Use PCR to quantify how much DNA you have after removing the TFs.
5. Label
6. View and Analyze Results

Results

• Low amount of positives matches to the known ChIP-chip data.

• Only 24 out of the 347 sites from the S. cerevisiae ChIP-chip data were identified by FOOTER.

• FOOTER had computationally identified over 5000 possible binding sites on the 2890 promoter sequences for S. cerevisiae, but only 24 of those matched with experimentally confirmed sites
Discussion

• The data was not expected to have so few positive matches

• Possible reasons why data was so poor:
  – The three yeasts species not as closely related compared to human and mouse genomes
  – FOOTER was not optimized to work with species that have a greater distance between the comparative binding sites.
  – Results need to be compared to other experimentally known binding sites for *S. cerevisiae*

Discussion

• Not enough time to compare FOOTER to other known positive results to see if more binding sites would be found.

• Also possible that tweaking the thresholds as well as the FOOTER algorithm itself to better account for the lower amount of conservation between the yeast species would probably find better results.
Future Possibilities

• Expand FOOTER to do multiple comparisons for multiple species

• Optimize the program to work with the less conserved binding sites between these three yeast species

• Possibly update the Benos et al C2H2 model used to make the PSSMs for the zinc finger binding preferences to better account for the differences in yeast species

Acknowledgements

• Takis Benos

• David Corcoran

• Rajan Munshi

• NIH-NSF

• BBSI @ Pitt 2005
References


