BIOI7712 Case Study:
Prediction of TF-TF interactions from their distance distributions

Feb. 6, 2007

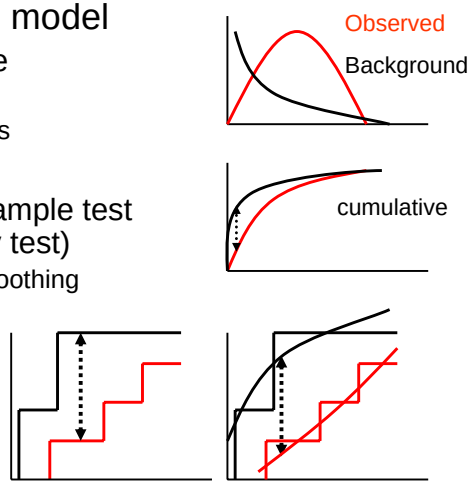
Hyunmin Kim

Specific Aims

- **Aim I-A:** To devise a calculation scheme that measures significance of a distance distribution between two TFBSs compared with background distributions
- **Aim I-B:** To predict interaction of TF-pairs from the distance patterns of the corresponding PWM-PWM hits

Aim I-A: Proposed Approach for Creating a Background Model

- Empirical background model
  - All-to-All motif distance distributions
    - Or, non-interacting pairs
  - Non-parametric two-sample test (Kolmogorov Smirnorv test)
    - With Kernel-Based Smoothing
    - With Bootstrap



Task Overview

- Extract promoter sequences
- Mapping putative TFBSs using MATCH
- Calculate pairwise distances
- Calculate empirical backgrounds
- Calculate significance of distances from the empirical backgrounds using KS statistics
- Select a p-value threshold
- Count interacting pairs satisfying the threshold

System Outline

- Home directory: $HOME=compbio:/home/hyunmink/uchsc/bioi7712_2007
- PROG: $HOME/bin
- DATA: $HOME/data
- ORIG: original resource
- SAMPLE: $HOME/sample

Extract Promoter Sequences

- Upstream regions 5' 2kb and 3' 200b from TSS
  - **ORIG:** ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver5_2/hspromoter.tab.gz
  - **PROG:**.$HOME/bin/get_upregion_from_dbtss.pl
- Get repeat masked sequences
  - **ORIG:**http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/chromFaMasked.zip
  - **PROG:** $HOME/bin/get_region.pl

Map putative TFBSs

- PROG: $HOME/bin/match/match
  - match <mxlib> <seq> <out> <mxprf>
- PARAM:
  - <seq>: fasta file
  - <mxlib>:$HOME/bin/match/match/data/matrixTF94..lib
  - <mxprf>:$HOME/bin/match/match/data/minFP94.prf

Transform MATCH to DIST file

- INPUT: dbtss_2k_200.fa.matchMinFP

```
                                    chrom#  start   end     refseqid
Inspecting sequence ID   chr17|70519366|70520566|NM_001545

 V$ELK1_02              |      799 (-) |  1.000 |  0.977 | actcTTCCGgtagc
 V$ELK1_02              |      983 (+) |  1.000 |  0.988 | cccgcCGGAAgcag
 I$HSF_01               |      226 (+) |  1.000 |  1.000 | AGAAA
 I$HSF_01               |      290 (-) |  1.000 |  1.000 | TTTCT

PWM name          start (strand)  core score   score    motif
```

- OUTPUT: dbtss_2k_200.dist

PWM-pair name [tab] #promoters[tab]distance:count[tab]....

```
V$E2A_Q2|V$E2A_Q6       127     13:1    15:2    18:1    19:3    22:1    23:1
27:1    28:1    29:1    33:1    35:1    36:2    37:2    39:1    40:1    43:1
44:3    45:1    47:1    48:2    50:1    51:1    55:1    60:2    63:1    65:1
70:1    71:1    78:1    82:1    90:1    94:1    99:1    108:1   116:1   126:2
```

See $HOME/sample

Transform MATCH to DIST file

- Sort PWM names in a pair with eliminator "|"
  - e.g., B|A => A|B
- Calculate a distance between centers of motifs
  - INPUT:
    - **V$CREB |10|...|aTGACG**
    - **V$PAX_Q6 |20|...|ctgattTCCAG**
    - **Distance = int((10+6/2) – (20+11/2)) = 12**
  - OUTPUT:
    - **V$CREB|V$PAX_Q6 ... 12:1**
- Parameters: min support, min/max distance
- **Issues: complexity with ~10k promoters and 400*400/2 pairs**

Calculate Empirical Backgrounds

- Sum up distributions of all-to-all pairs
- $HOME/sample/all.dist:

```
all|all 19484554    10:41916    11:41267    12:41411    13:41177
        14:40910    15:40668    16:40975    17:41138    18:40865
        19:40848    20:40552    21:40496    22:40095    23:39847
        24:39720    25:39404    26:39367    27:39103    28:38972
        29:38792    30:38900    31:38476    32:38620    33:38673
```

Two-sample Test with KS Statistic

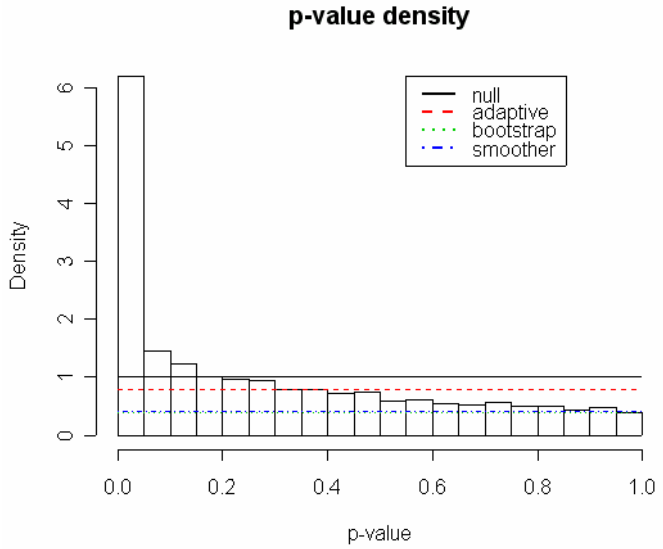
```
parseLine<-function(strx){
    strx<-gsub("\r","",strx);
     strx<-gsub("\n","\t",strx);
    strx<-gsub(":","\t",strx);
     x<-strsplit(strx,"\t")[[1]];
    pwm<-sprintf("%s",as.character(x[1]));
    n<-as.integer(x[2]);
     x<-as.numeric(x[3:length(x)]);
    m<-matrix(x,ncol=length(x)/2);
    return(list(x=m[1,],y=m[2,],pwm=pwm,
    n=n));
}

con<-file("dist background","r");
string<-readLines(con=con,n=1);
yy<-parseLine(string); # parse the line
close(con);
# sampling background distribution
y<-sample(yy$x, size=2000,
    prob=yy$y ,replace=T)
```

```
con<-file("dist file","r");
while(length(strings)==1){
  readLines(con=con,n=1);
  xx<-parseLine(strings);
  x<-rep(xx$x, xx$y);
  ks.test(x,y);
  ....
  string<-
    readLines(con=con,n=1);

}
close(con)
```

P-value Threshold



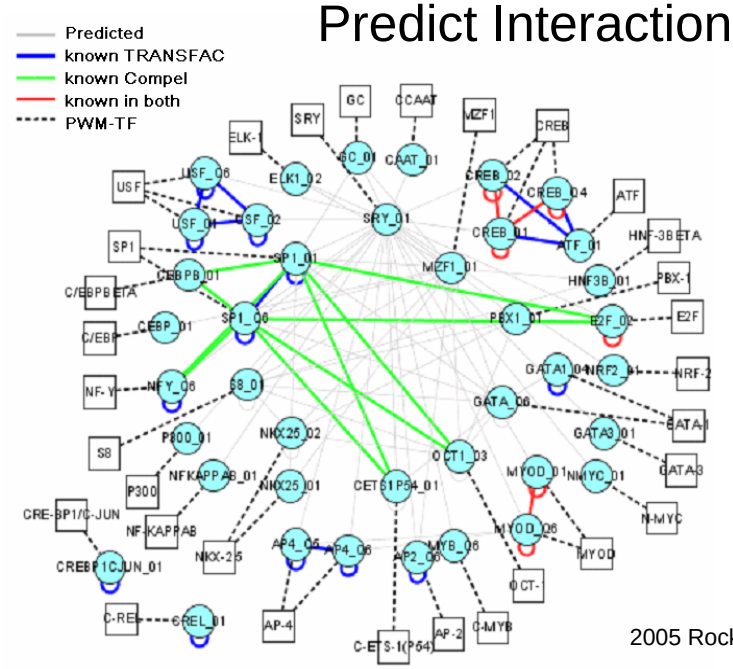
Interaction

- **ORIG:**
  - $HOME/data/TRANSFACv9.4/data/factor.dat
  - IN fields

```
FF  promotes development of sensory organ mother cel
her AS-C genes [7];
FF  mRNA accumulates at blastoderm stage;
FF  antagonized by emc [3];
XX
IN  T01034; Da; fruit fly, Drosophila melanogaster.
IN  T00204; E12; human, Homo sapiens.
IN  T00675; E12; rat, Rattus norvegicus.
```

- **DATA:**
  - $HOME/data/INTER_PWM.INFO
    - PWM1[tab]PWM2[tab]Similarity

Predict Interactions



2005 Rocky Conference

Database	Category and Contents
	Promoter
EPD	experimentally determined 1871 (total 4809) human promoters sized from -499 to 100 bp http://www.epd.isb-sib.ch/ (Schmid, Perier et al. 2006)
DBTSS (v5.2.0)	30,964 human promoters (425,117 TSSs) http://dbtss.hgc.jp/ (Suzuki, Yamashita et al. 2004)
PromoSer	http://biowulf.bu.edu/zlab/PromoSer/ (Halees, Leyfer et al. 2003)
	CRM
cisRed (human v2)	promoter regions sized from 1.5k to 200bp containing ~381k conserved motifs in ~18k human target genes (Ensembl v31/NBCI 35) including 357 ENCODE genes; ~4.5K motifs discovered in 366 of the ~640 ENCODE Stanford promoters. http://www.cisred.org/ (Robertson, Bilenky et al. 2006)
PReMod	~100,000 computational predicted CRMs using 481 TRANSFAC 7.2 PWMs http://genomequebec.mcgill.ca/PReMod/pages/welcome.jsp (Ferretti, Poitras et al. 2007)
TriProD	15,384 promoter sequences for tissue-specificity or according to Gene Ontology terms http://tiprod.cbi.pku.edu.cn:8080/index.html. (Chen, Wu et al. 2006)
High-confidence Coexpression Data	As part of the cisRED project 92472 coexpressed gene pairs for 7447 genes http://www.bcgsc.ca/project/bomge/coexpression (Griffith, Pleasance et al. 2005)
	PWM profile
TRANSFAC v9.4	774 PWMs http://www.biobase-international.com/pages/index.php?id=transfac (Wingender, Chen et al. 2001)
JASPAR	non-redundant set of 123 profiles from published articles http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl (Sandelin, Alkema et al. 2004)
	genome-wide putative TFBS collections
GALA	2,963,975 conserved (hg17mm5Rn3Canfam1) http://gala.cse.psu.edu/gala/downloads/hg17/conserved_tfbs/hg17Mm5Rn3Canfam1/ (Giardine, Elnitski et al. 2003)
UCSC	695,221 conserved in the human/mouse/rat alignment
	ENCODE ChIP-chip http://genome.ucsc.edu/ENCODE/encode.hg17.html
Uppsala	Hnf3β, Hnf4α, USF1 (HepG2) (Rada-Iglesias, Wallerman et al. 2005)
UT-Aus	c-Myc, E2F4 (HeLa, 2091 fibroblasts, FBS stim.) (Kim, Bhinge et al. 2005)
Yale	STAT1, c-Fos, c-Jun, BAF155, BAF170, TAF1 (HeLa) (Trinklein, Murray et al. 2004)
Stanford	Sp1, Sp3 (HCT116, Jurkat, K562) (Cawley, Bekiranov et al. 2004)
UC Davis	E2F1,c-Myc (HeLa)
GIS	p53 (HCT116), STAT1 (HeLa), c-Myc (P493 B) (Ng, Wei et al. 2005)
Affy	CEBPe, CTCF, P300, PU1, RARA, SIRT1, Brg1

Due Date

- 2/15/07