Bioinformatics Case Study:

Genome-wide Analysis of the Distances between Human Transcription Factor Binding Sites

BIOI7712

Feb. 1, 2007

Hyunmin Kim

My Definition of BioInformatics

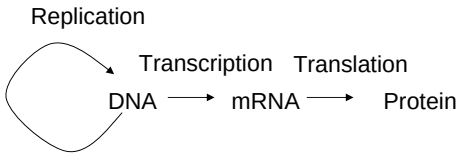
- Data
  - symbols
- Information:
  - useful data
  - what, where, when
- Knowledge
  - Application of Information and Data
  - How
- Understanding
  - Why

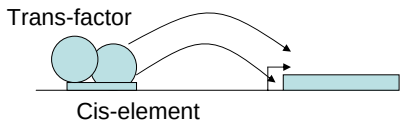
- Bio Data
  - A,C,G,T, A.A…etc
- Bio Information
  - Entities of DNA, RNA, Protein
- Bio Knowledge
  - Associations between information entities
- Understanding of Biology
  - Mechanism of life, Death, Disease

Understanding of Biology Using Information or Knowledge

Association of Bio Entities

- Flow
  - DNA->DNA: Replication
  - DNA->mRNA: Transcription
  - mRNA->Protein: Translation
- Interplay
  - Protein<->Protein: Protein-protein interaction
  - DNA<->Protein: trans-Regulation
  - DNA<->DNA: cis-Regulation

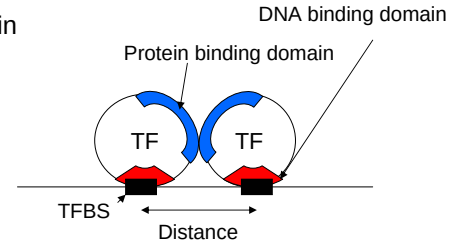


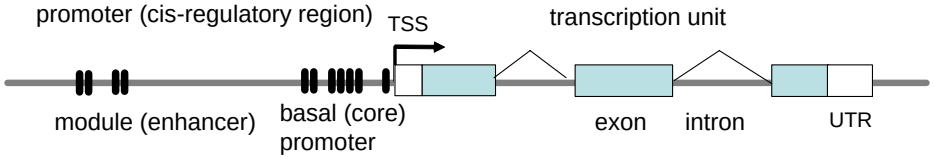


STEP1: Understanding Biology and Experiments

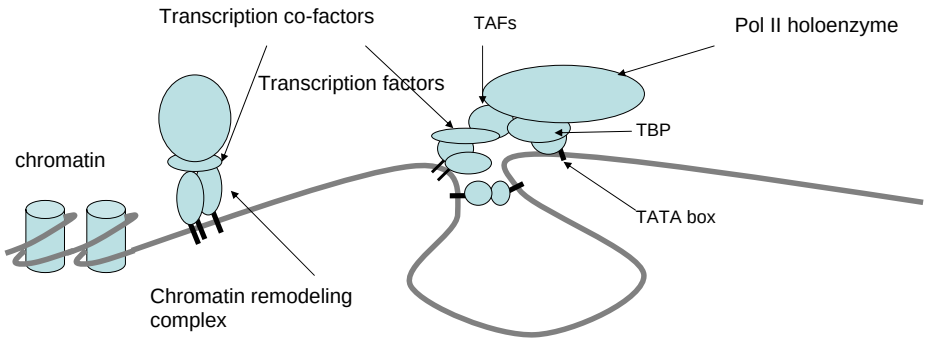
Transcription Factor Binding Sites

- ~ 1,962 DNA binding factors in Human
- Contains
  - DNA-binding domain
  - Protein-binding domain
- Function as
  - Primary transcription regulator





Promoter Structure

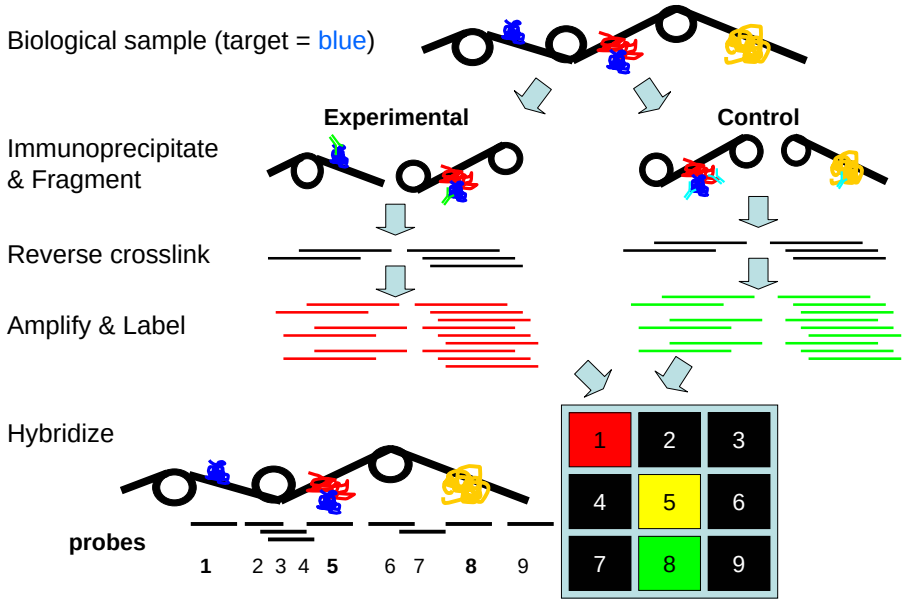


- Form composite element or *cis*-regulatory module
- Protein-protein interaction
- Chromatin remodeling
- long-distance control

Experimental TFBS assay

- *In vitro*
  - e.g., EMSA, DNaseI foot printing
  - Do not reflect complexity of the promoter structure
  - Collected in JASPAR and TRANSFAC database
- *In vivo*
  - e.g., Ligation-mediated PCR, Chromatin immunoprecipitation (ChIP)
  - Tissue-specific information
  - Inability to detect precise contacts and indirect interaction
- High-Throughput
  - e.g., ChIP-chip, DamID, PBM
  - Indirect interactions
  - Low resolution

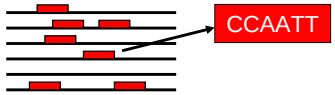
ChIP-chip Experiment



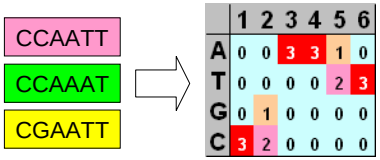
- STEP2: Problems of Computational Approaches

Computational TFBS Estimation

- Motif discovery
  - Search for frequently occurring 10-20bp segment from collection of DNA sequences
    - Coexpressed genes, orthologous genes, ChiP-chip data



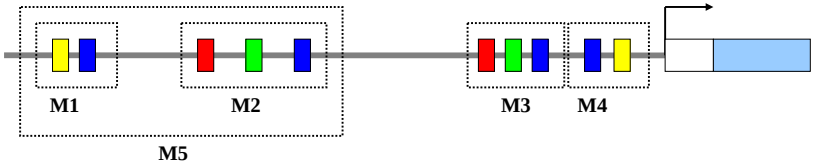
- Search for the known motifs using Position Weighted Matrices (PWMs)
  - Gather statistics of bp occurrences from collection of motifs
  - High false positive rate



- Modeling cis-regulatory dodule (CRM)

cis-Regulatory Module (CRM)

- Find re-occurring groups of motifs and PWM hits



- Features:
  - Combination of TFBS elements (Segal and Sharan 2005; Zhou and Wong 2004)
  - Ordering (Li, Cheng et al. 2006)
  - Compactness (Hannenhalli and Levy 2002; Long, Liu et al. 2004; Rateitschak, Muller et al. 2004)
  - Spacing between Composite Elements (CEs) (Diamond, Miner et al. 1990; Kel-Margoulis, Romashchenko et al. 2000)
  - Distance preference (Yu, Lin et al. 2006a; Yu, Lin et al. 2006b) between TFBS elements

Emerging Issues in Characterizing Distance Features between TFBSs

- Distance distribution between TFBS groups
  - PWM similarity (e.g., homotypic, heterotypic)
  - Class type of corresponding TF (e.g., bZip, Homeodomain)
- Incorporation of distance distributions into a CRM model
  - Characteristics of promoters (tissue-type, function, etc.)
  - ChIP-chip experimental results
  - Correlation of TFBSs with complementary genomic contents

- STEP 3: Generate Testable Hypotheses and Corresponding Aims
  - Significance
  - Novelty
  - Feasibility

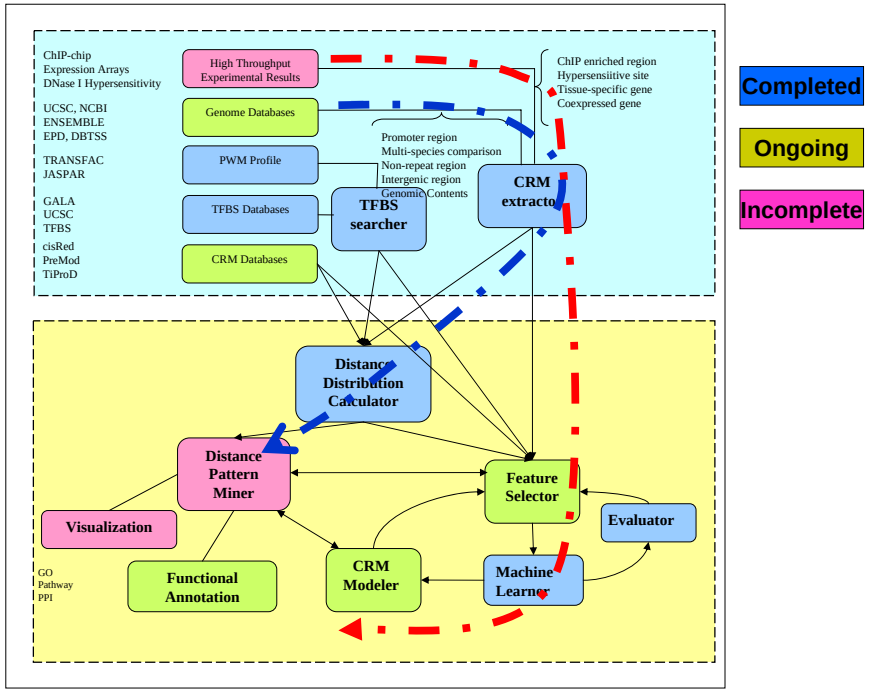
Hypotheses

- There are specific distance distributions (between TFBS) that are functionally important in transcriptional control.
- These distributions can be characterized from the experimentally proven data.
- The characterizations can be used to predict novel transcriptional phenomena.

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 II-A:** To create a scoring scheme combining distance preference of TFBS-pairs
- **Aim II-B:** To discriminate functional binding sites from the false positives
- **Aim III:** To discover long-distance DNA-DNA interaction thru TF-TF interactions

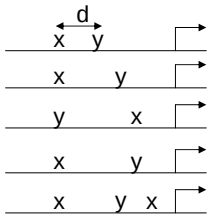
Database	Category and Contents
	Promoter
EPD	experimentally determined 1871 (total 4809) human promoters sized from -499 to 100 bp http://www.epd.isb-sib.ch/ (Schmid, Perier et al. 2006)
DBTSS (v5.2.0)	30,964 human promoters (425,117 TSSs) http://dbtss.hgc.jp/ (Suzuki, Yamashita et al. 2004)
PromoSer	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://triprod.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/bcmge/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 (hg17mn5Rn3Canfam1) 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

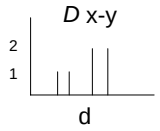


- STEP 3: Study with Smallest Samples

Aim I-A: Measure Significance of Distance Distributions

- Formulation: Binding motif *x* of TF *X* and *y* of TF *Y* occur in a same *s*-sized promoter of a gene group G.
- P1: What is <u>the probability of observing x and y with a distance *d*</u> assuming independence x and y in a random sequence?
- P2: If we observe an occurrence of *x* and *y* in *d*, what is <u>statistical significance of observing >*d* and <*d*</u>?
- P3: If we observe <u>distribution *D*</u> between *x* and *y* along the different *d*'s, what is statistical significance over a given null model?



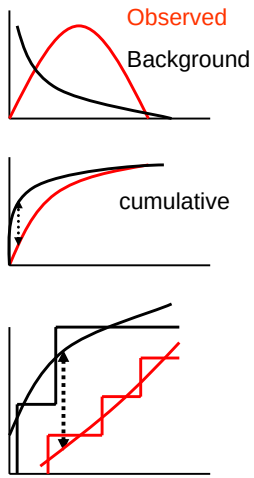


Aim I-A: Previous Approaches

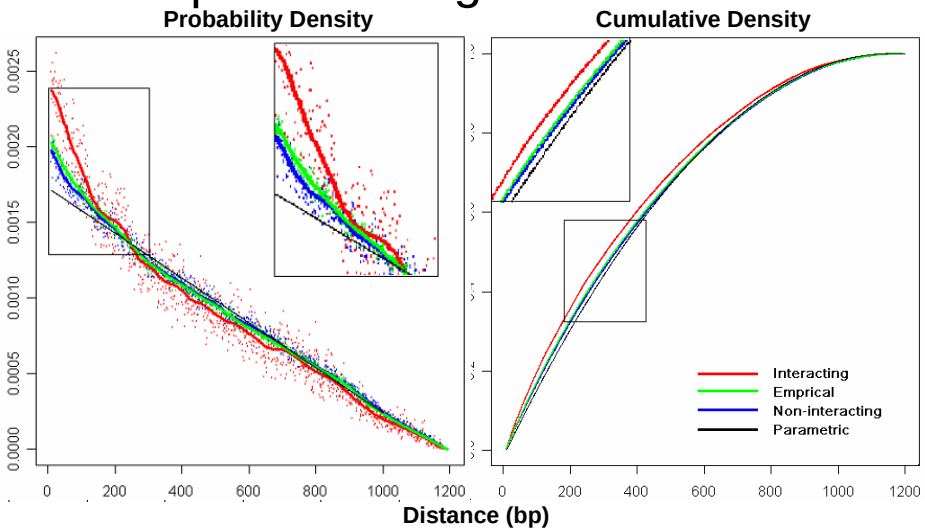
- P1: probability of x and y at dist d
  - Function of distance, motif length for *x* and *y* and size of promoter (Yu, Lin et al. 2006a; Yu, Lin et al. 2006b)
  - Additional term for range between *d* and *d'* (Smith, Sumazin et al. 2005a)
  - Recall: assumes random sequence model yet evidence that background sequence is not random and background distribution is not random
- P2: probability of x and y at dist <d
  - Count number of promoters with motifs *x* and *y* in sliding window of size *d*, compute log odds ratio versus background
    - Background by shuffling (Hannenhalli and Levy 200)
    - Background by marginal distribution (Long, Liu et al. 2004Rateitschak; Muller et al. 2004)
  - Use hypergeometric distribution so can calculate p-values (Yu, Lin et al. 2006a; Yu, Lin et al. 2006b)
  - Performance depends selection of *d* (different lengths have different backgrounds)
- P3: distributions vs given null model; motivated by above observations

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

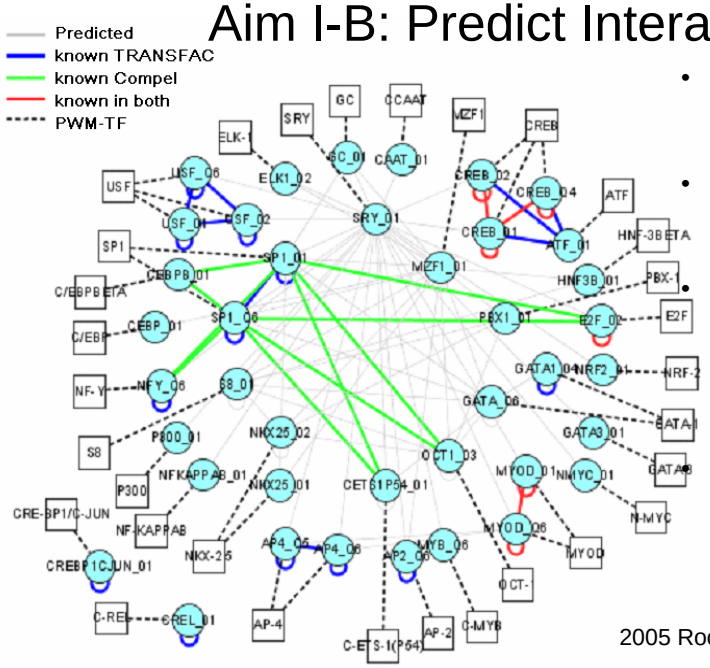


Aim 1-A: Rationale for Importance of Empirical Background Distribution



DATA: 14,604 RefSeqs TSS from DBTSS 5.2v; 462 PWMs in TRANSFAC 9.4v

Aim I-B: Predict Interactions



- Pilot study of 19k promoters, 300 PWMs
- Create distance distributions for each PWM pair
- Using our empirical background distribution, test significance of each pair
- Of top 988 predictions, 23% (101/437) are known interacting pairs

2005 Rocky Conference

Aim I-B: PWM-pairs in a Tissue-Specific Promoters

- Previous Approaches:
  - Rediscover 70% of the known protein-protein interactions (PPIs) in Yeast (Yu et al 2006a)
  - 40% of the known PPIs in Tissue-specific Human genes (Yu et al 2006b)
  - Use combination of **co-occurrence** and distance distribution with **parametric** background model
- Pilot study of Muscle-specific PWM-pairs in 46 muscle-specific genes:
  - Choose 69 out of 3,222 pairs (p-value threshold 0.0052)
  - Rediscover 34.8% (24/69) the known PPI
  - Rediscover most of known muscle-specific pairs except Mef2-Mef2

pwm1	pwm2	# BS	p-val	#gene
Myf	Sp1	1844	1.35E-09	30
Myf	**Myf**	456	4.98E-08	22
Sp1	Srf	622	1.56E-05	24
Srf	**Srf**	72	4.22E-05	13
Sp1	Sp1	8330	5.39E-05	35
Mef2	**Myf**	190	7.79E-05	26
Myf	Tef	200	0.004599	20
Tef	Tef	84	0.008883	12
Sp1	Tef	782	0.030254	24
Myf	Srf	131	0.058355	19
Mef2	**Mef2**	118	0.132712	17
Mef2	Srf	70	0.229636	20
Mef2	Sp1	620	0.27522	30
Srf	Tef	64	0.441749	17
Mef2	Tef	82	0.999958	20

Aim I-B: Consideration of PWM Similarity

- Improve the performance by distinguishing PWM types: homotypic & heterotypic using TRANSFAC v9.4
  - Homotypic pairs have similar PWM matrices
  - MatCompare program with p-value cutoff 0.05
    http://rulai.cshl.edu/MatCompare/
- Compute empirical background model specific for homotypic or heterotypic
  - Overall 31% rediscovery rate of interacting pairs
  - Rediscover 53.5% of homotypic pairs

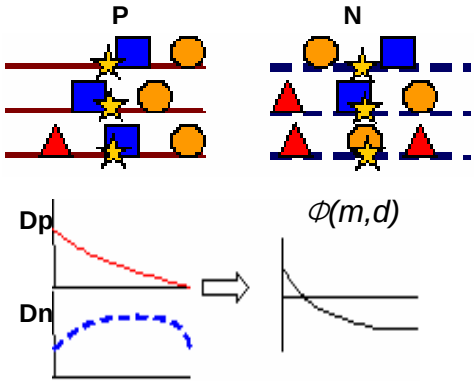
	All pairs total	All pairs significant	Interacting Significant	Interacting total	p-value threshold
Heterotypic	15227	716	29	368	6.36168e-05
Homotypic	7465	651	204	381	0.000127623

Conclusion of Pilot Study on Aim I

- Distance distribution of TFBS-pair is useful to predict interaction of TF-pair
- Empirical background model is reasonable

Aim II-A: Create Scoring Scheme Combining Distance Preferences

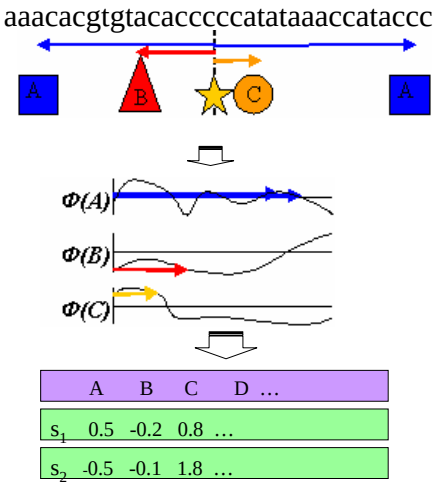
- Problem Formalization: Given a positive set *P* and a negative set *N* of DNA sequences, find occurrences of a target motif *t* in every sequence
- Using *P* and *N*, calculate distance distributions *Dp* and *Dn* of other motifs *m* to *t*
- Compute a 'distance preference' $\Phi(m,d)$ of *Dp* versus *Dn* for each distance *d* for motif m relative to target motif t



$$\phi(m,d) = \log \frac{\Pr(d \mid m, P)}{\Pr(d \mid m, N)}$$

Aim II-A: Create Distance Features

- Given a sequence *s*, with target t and motif m, represent the context
  $\boldsymbol{X}=(x_{1s},x_{2s},\ldots,x_{ls})$
  where x_{ms} is distance preference score of m in s.
- Use context vectors as features to learn classifier on P and N
- Score new sequences



Aim II-A: Classification

- Using novel distance preference features
  - Rule based ensemble learners
    - Random forest – decision trees built from random samples (randomForest library in the R2.4.0 package)
    - RuleFit – ensemble of decision rules on random samples (http://www-stat.stanford.edu/~jhf/R-RuleFit.html)
  - Avoid overfitting with
    - Out-of-back (OOB) and panelizing large values of the coefficients
  - Easier interpretation
  - Importance measurement to define most influential variables
- Cross validation evaluation scheme
  - Average of Classification Errors=Average false positive and false negative rate

Aim II-B: Discriminating functional TFBSs from false positives

- Functional TFBSs from ChIP-chip experimental results
- False positives occur because:
  - Putative PWM scoring is error prone
  - Resolution of ChIP-based TFBSs assay (~ 500bp)
  - Indirect interactions
- False negatives occur because:
  - Too constrict threshold for scoring PWMs
  - Non-consensus motifs in a high peak region
    - Novel motifs?
    - Indirect interactions?

Aim II-B: Genome Contents

- Conventional computational algorithms to discriminating FP in a ChIP-chip dataset do not consider relation of TFBSs with complementary genomic contents (Smith, Sumazin et al. 2005b; Jin, Rabinovich et al. 2006; Macisaac, Gordon et al. 2006)
- Such as
  - Nucleosome positioning codes
  - DNase Hypersensitive sites
  - Multi species conserved regions
- Performed pilot study using genome contents on the hepatocyte transcriptional regulators (HNF4α, HNF3β, and Usf1)

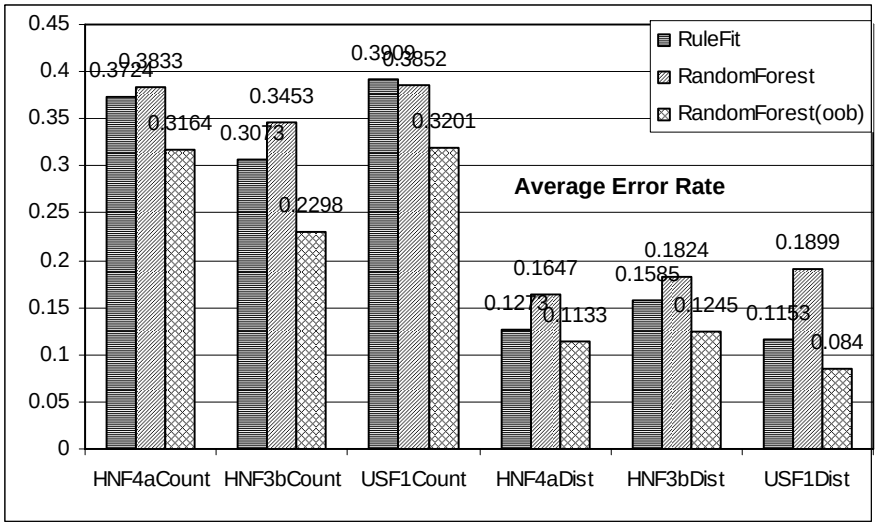
Aim II-B: Data from ENCODE (ENCyclopedia Of DNA Elements)

- ChIP-chip experimental data
  - ENCODE ChIP-chip track in hg17 UCSC genome repository (Rada-Iglesias, Wallerman et al. 2005)
- Genome contents
  - Nucleosome occupancy estimation: (Segal, Fondufe-Mittendorf et al. 2006)
  - Dnase hypersensitive (HS) sites http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/database/
  - Conserved regions: http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/database/
- Putative TFBSs in ENCODE regions
  - MATCH with TRANSFAC 9.2
  - Hg17 ENCODE regions in UCSC genome browser

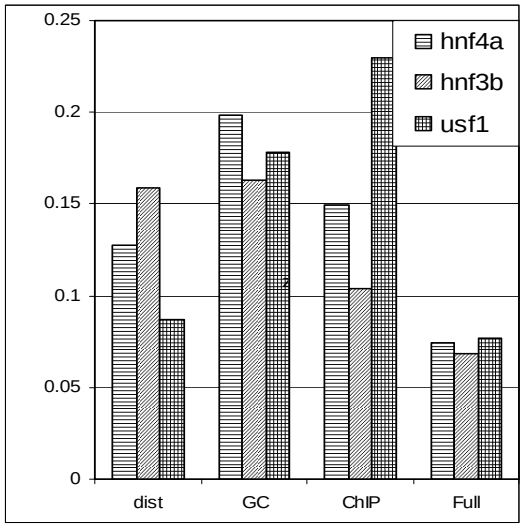
Aim II-B: Methods

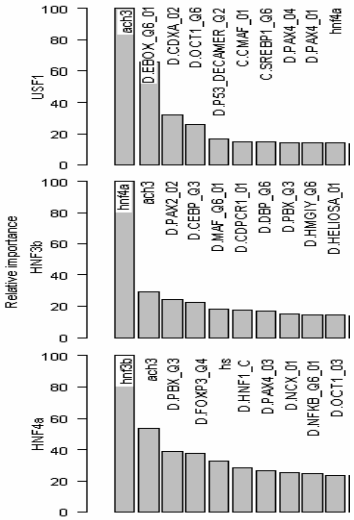
- From Aim II-A:
  - Represent and score sequences using TFBS distance preference context vectors
  - Train classifier using the ChIP enriched regions thru a context of putative TFBSs using ensemble learning algorithms
- Test and compare:
  - The performance of proposed distance feature sets
  - With the reference sets: (1) count feature, (2) genome content, and other related ChIP experimental results

Aim II-B: Pilot Study Using Distance Features (without Genome Contents)



Aim II-B: Pilot Study Using Distance features (RuleFit with genomic contents)



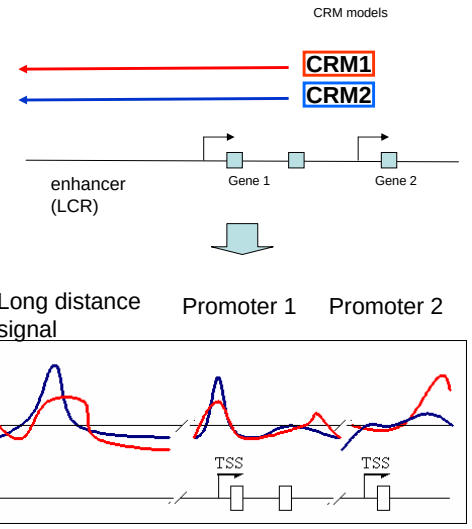


Conclusion of Pilot Study on Aim II

- Consideration of distance patterns improve the performance of discriminating TFBSs in ChIP-enriched regions
- Genomic features are useful, but their effects depends on TFs
- This characterization would be useful to capture other ChIP-chip based signatures

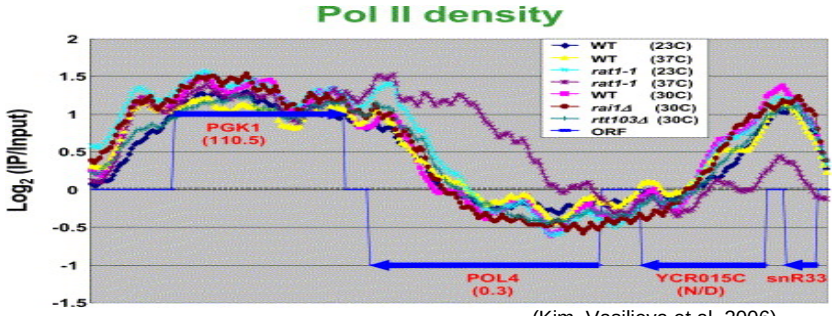
Aim III: Discovering Long-distance Signals

- Facts
  - Interactions of TFs play a crucial role in facilitating long-distance communication between Locus Control Region and promoters (Bondarenko, Liu et al. 2003; Vieira, Levings et al. 2004)
- Proposed Test Suite:
  - Tissue-specific genes of the HOXA cluster
  - And, beta and alpha globin gene clusters
  - 921 predicted promoters in ENCODE regions (Cooper, Trinklein et al. 2006)



Extensions to Pilot Studies

- Method Extension
  - Aim I-A: consideration of types of PWM-pairs
  - Aim I-B: prediction of missing CRMs without consensus motifs of s target TF
  - Aim I-C: N-way interaction
- Dataset Extension
  - Aim II-A: consideration of promoter types
  - Aim II-B: remaining ChIP experiments
  - Aim II-C: Beyond ENCODE, Mouse Genome with human CRMs;
- Integration of Diverse ChIP-chip signals
  - other genome content like PolyII elongation and termination in the yeast genome (with David Pollock, Ph.D and David Bentley, Ph.D)



(Kim, Vasiljeva et al. 2006)