

A Biophysical Model for Analysis of Transcription Factor Interaction and Binding Site Arrangement from Genome-Wide Binding Data

Xin He¹, Chieh-Chun Chen², Feng Hong³, Fang Fang⁴, Saurabh Sinha¹, Huck-Hui Ng⁴, Sheng Zhong^{1,2,3*}

¹ Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, ² Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, ³ Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, ⁴ Gene Regulation Laboratory, Genome Institute of Singapore, Singapore, Singapore

Received September 30, 2009; **Accepted** November 10, 2009; **Published** December 1, 2009

A Literature Review

Diyush Labhsetwar

Feb 23rd, 2010

Introduction

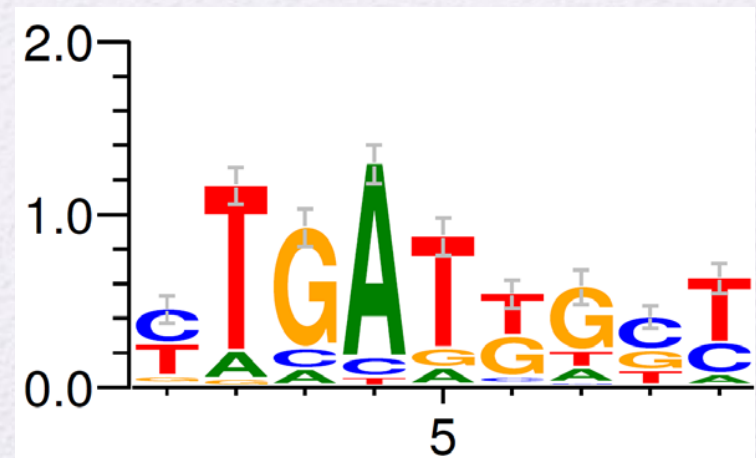
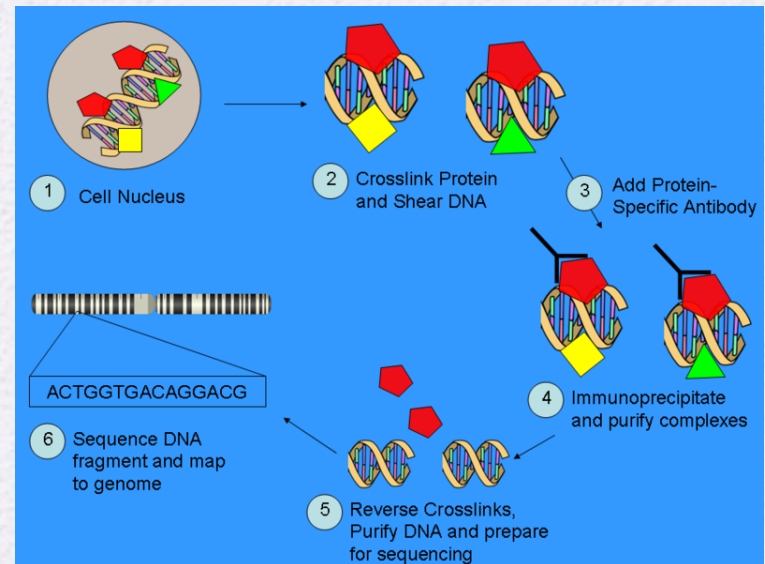
- *Cis*-regulatory modules bind to multiple transcription factors(TF) to control spatio-temporal pattern of gene expression
- Interaction among TF's and between TF and DNA key to understand gene regulation
- Modeling combinatorial TF-DNA interactions will help to understand complex transcriptional process
- Aim is to build a predictive model of TF binding affinity to DNA sequences from TF-DNA binding data, incorporating both TF-DNA and TF-TF interactions

STAP: Sequence to affinity prediction

- “ A computational method employing biophysical principles...
- which takes into account all configurations of strong and weak binding sites...
- to analyze large scale TF-DNA binding data...
- to discover cooperative interactions among TFs, infer sequence rules of interaction...
- and predict TF target genes in new conditions with no TF-DNA binding data.”

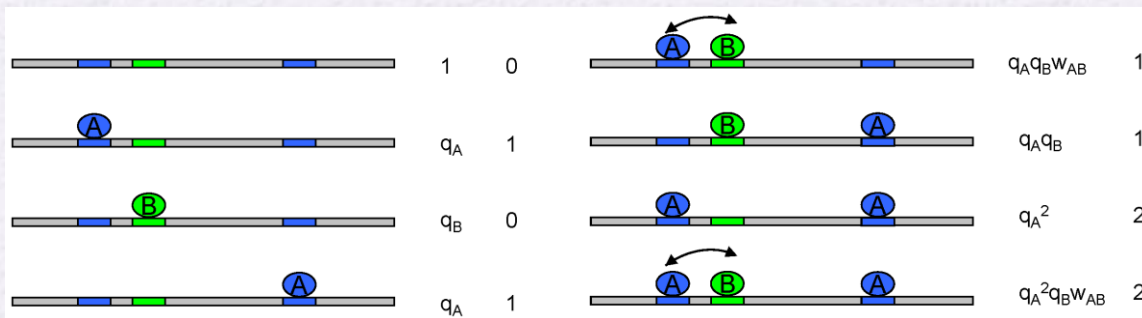
Background concepts

- **Chip-seq** : TF-DNA binding data to determine DNA regions binding to specific protein (TF)
- **TF binding motif**: Matrix representing DNA binding preference for a particular TF



Biophysical Model

- Adapted model from Hwa et al.
- Binding sites of primary TF and other cooperating TF contribute to interaction of TF with the sequence
- Binding affinity of TF to a sequence is proportional to the average number of TF occupying their sites, over all states weighted by their probabilities



$$K(S_{\max}) * [TF] = \frac{[TF] - S_{\max}}{[S_{\max}]}$$

$$q_i = [TF] * K(S_{\max}) e^{-\Delta E(S_i)}$$

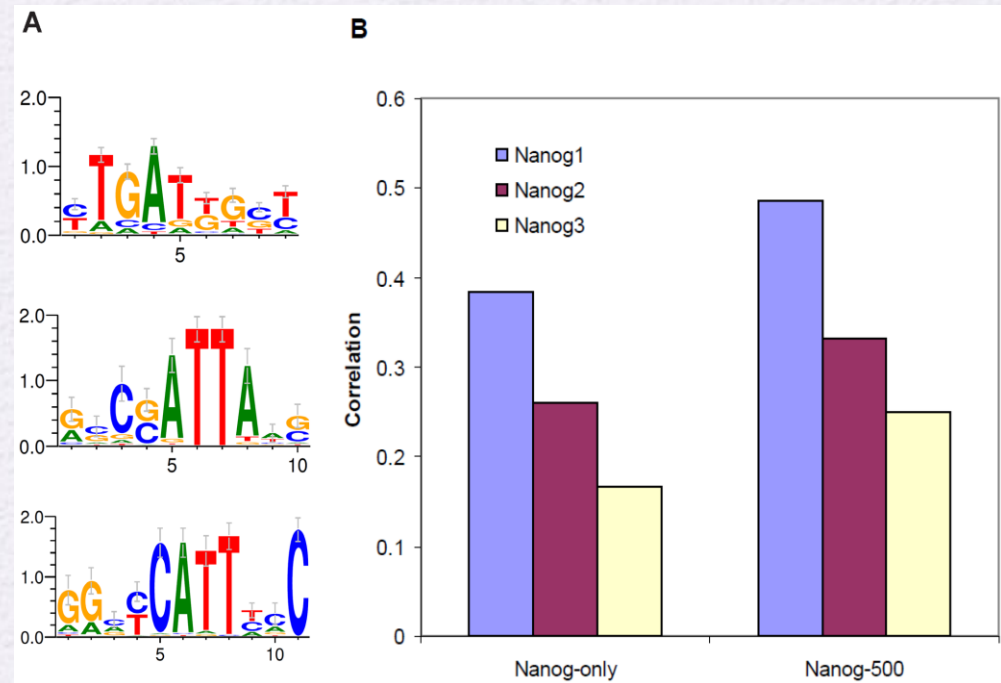
$$\bar{N}_k = \frac{\sum_{\sigma} N_k(\sigma) * W(\sigma)}{\sum_{\sigma} W(\sigma)}$$

Implementation and model fitting

- Model fit is done by maximizing Pearson's correlation coefficient between predicted binding affinity and Chip-seq counts
- Model consists of
 - Cooperative factors
 - Free parameters: TF specific constant and interaction parameters
- Scans the sequence for all putative binding sites using motifs of various TF's
- Cooperative binding sites identified by checking for improvement in correlation coefficient by its inclusion in the model
 - Significance checked by comparison with inclusion of random motif
- All cooperating motifs combined into single model and free parameters calculated by global optimization algorithms to maximize correlation

Results: Novel motif for Nanog

- Chip-seq data of 12 TF active in embryonic stem cells of mouse used.
- Motifs identified using MEME suite on top 100 regions detected in Chip-seq experiments.
- Nanog motif different from previously reported ones, fits chip-seq data well.
- Novel motif confirmed by experimentation
 - Electrophoretic Mobility Shift Assay
 - Mutation of “TGA” core



Results: Co-operativity among TFs

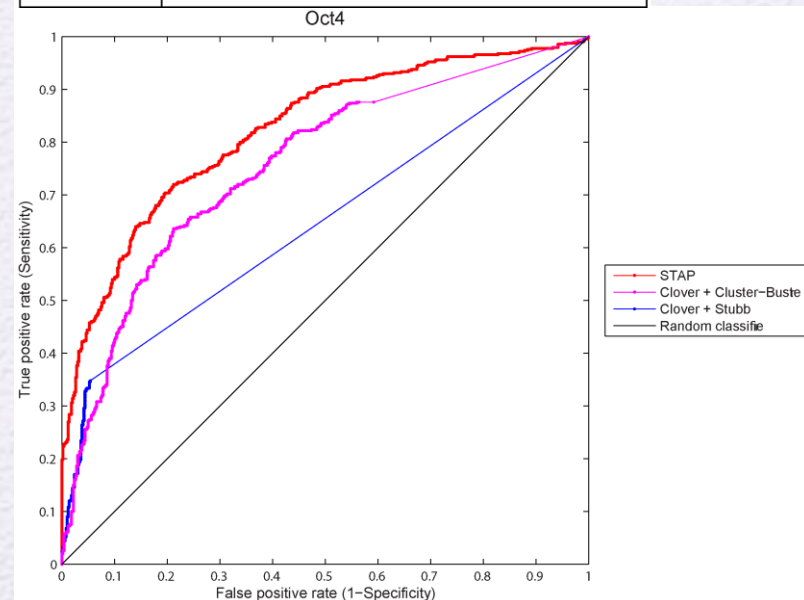
- Reproduced known interacting pairs
- Klf4, which facilitates self renewal of ESCs, found to cooperatively bind with many TFs
- Significant improvement over non-cooperative model
 - Self-cooperativity in case of Zfx and CTCF

Factor	Non-coop. Model	Coop. Model	Improvement	Significant Coop. Factor (p-value)
cMyc	0.57	0.82	44%	E2f1(0.004), Klf4(0.04), Zfx(0.033)
CTCF	0.75	0.81	7%	
E2f1	0.50	0.66	31%	Nanog(0.048)
Esrrb	0.62	0.78	26%	Zfx(0.003)
Klf4	0.58	0.74	28%	CTCF(0)
Nanog	0.24	0.50	107%	Sox2(0), Klf4(0.012), Zfx(0.05)
nMyc	0.67	0.83	23%	E2f1(0.005)
Oct4	0.45	0.56	22%	E2f1(0.029), Klf4(0.032), Zfx(0.017)
Sox2	0.50	0.62	24%	Klf4(0.014), Oct4(0.039), Zfx(0.045)
STAT3	0.52	0.65	24%	Klf4(0.004), E2f1(0.049), Zfx(0.039)
Tcfcp2l1	0.74	0.76	3%	Esrrb(0.121)
Zfx	0.70	0.71	1%	

Results: Improvement over existing methods

- Clover used to find over-represented motifs (motif set) in TF bound sequences for each TF
 - Results largely parallel to Co-localization results
- Cluster-Buster and Stubb used to predict presence of motif set in target sequence
- Existing methods based on Co-occurrence
 - Ignore TF binding intensities
 - Need explicit annotation of binding site, which is known for inaccuracy

cMyc	cMyc, nMyc
CTCF	CTCF, Nanog
E2f1	N/A
Esrrb	Esrrb
Klf4	Klf4, Sox2, Esrrb
Nanog	Nanog, Sox2, Oct4, Esrrb
nMyc	cMyc, nMyc
Oct4	Oct4, Sox2, Nanog, Esrrb
Sox2	Sox2, Oct4, Nanog, Esrrb
STAT3	STAT3, Klf4, Sox2, Nanog, Oct4, Esrrb
Tefcp211	Tefcp211, Sox2, Esrrb
Zfx	Zfx

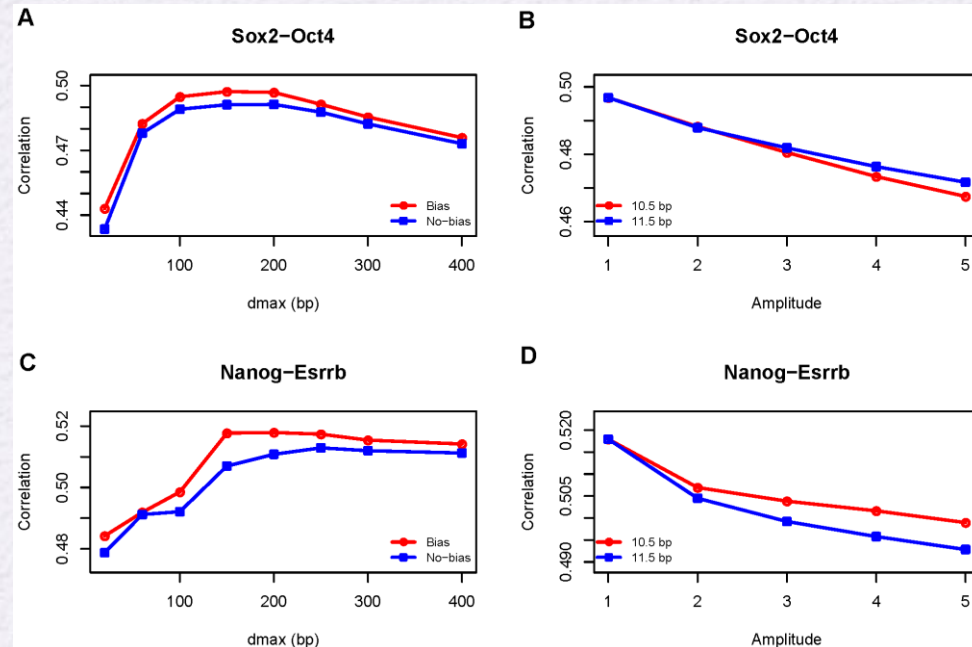


Combinatorial gene regulation

- 3 types of relationships
 - Co-localization of two factors
 - Direct binding of factors to neighboring sites (co-binding)
 - Co-operative interaction of two factors in neighborhood
- Represent hierarchy in relationship of TFs
- This may be reason for difference of results from STAP and Clover-Cluster-Buster/Stubb

Results: Effect of binding site arrangement

- d_{\max} : Maximum distance beyond which no interaction possible
- Binary and linear model have orientation bias
- Periodic model has periodic strength of interaction corresponding to helical period of DNA
- Linear model didn't improve predictability implying no decrease in interaction with distance
- Results indicate no strict rules followed during TF interaction.



Shortcomings:

- Cooperative function may be much more complex
- Only neighboring interaction considered

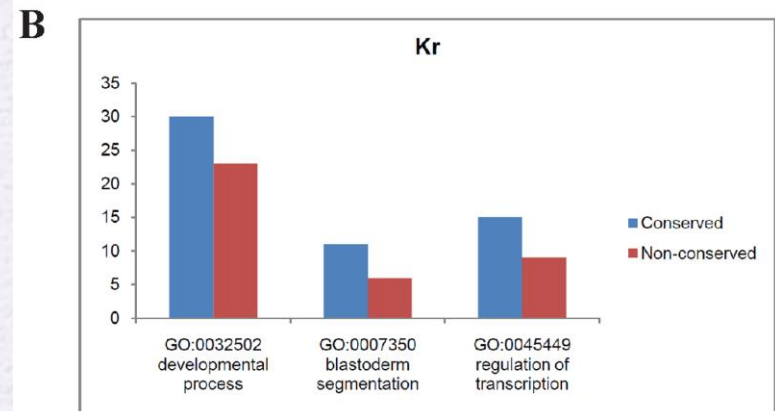
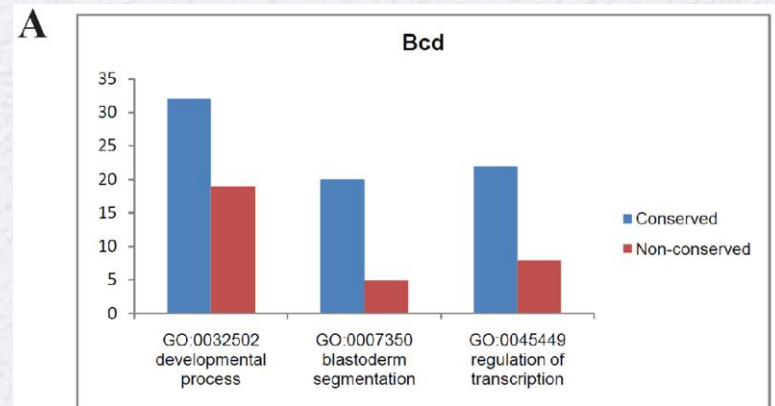
Results: Cross species extrapolation

- Trained binding models with Chip-chip data of 2 TFs from *D. melanogaster* (Mel) and applied them to genome of *D. pseudoobscura* (Pse)
 - Training on 500 TF bound and 500 random unbound sequences
 - Applied on Pse orthologs of all bound sequences and 250 random unbound sequences
- No conservation in binding affinity even for transcribed genes

Sequences	Bcd	Kr
Random	0.13 (32/250)	0.22 (54/250)
Enhancers	0.83 (29/35)	0.48 (16/33)
Bound (1% FDR)	0.45 (310/692)	0.34 (685/2001)
Bound (1% FDR) and expressed	0.43 (141/331)	0.33 (205/621)

Results: Predicting functionality

- Able to filter the non-functional sequences from all the bound ones by testing binding affinities of the orthologous sequences
- Bound sequences (adjacent to expressed genes) in *Mel* classified into 2 groups:
 - With high predicted affinities in *Pse*
 - With low predicted affinities in *Pse*
- Predicted affinities of orthologous sequences can be used to filter out non-functional sequences in results from genome wide binding experiments



Summary

- Development of biophysical model for TF-DNA interaction with inclusion of TF-TF inclusion to analyze large scale TF binding data
 - Explicitly expressed expected number of TF bound to sequence of interest
 - STAP software
- STAP can be used to compare several putative motifs and identify TF-TF interactions
- STAP can be used to check for regulatory rules in binding site arrangement
- STAP can be used to make TF target predictions in a specie for which binding data is not available

THANK YOU