Computational methods for analyzing the architecture and evolution of the regulatory genome

Thesis Defense
Pradipta Ray
pray@cs.cmu.edu

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

11/30/2012
Thesis committee

- Eric P. Xing (co-chair)
- Veronica F. Hinman (co-chair)
- Jaime Carbonell
- Ziv-Bar Joseph
- Martin Kreitman (University of Chicago)
Organization

Motivation
Goals
CRM architecture and ghHMMs
Functional turnover and the CSMET model
Discriminative motif finding using CRFs
Evolutionary analysis of CRMs
Context in the literature: past, present and future
Diversity

- Across species, cell type, developmental stage, response to stimuli
  - Given a constant (for the same species) or a slowly evolving (for related species) genome, how is this achieved?

Courtesy: Sean Carroll
Gene regulation

Gombel et al, Nature, 2005

Courtesy:
Scott Gilbert, Sinauer, 2003
Richarddawkins.net

11/30/2012
Promoters and enhancers

- How far away from the gene body is the regulator?
- Epigenetic signature and plasticity

Ohtz Lab, WUSTL
Transcription factor binding sites

- **Transcription factors**: proteins which bind to the DNA at specific locations to aid or repress protein transcription.

- Binding sites typically occur in clusters (cis-regulatory modules or CRM) and are noisy copies of each other.

- By binding to the DNA, the transcription factors up-regulate or down-regulate the protein in question.

Courtesy: Wasserman and Sandelin, Nat Gen, 2004
The regulatory network
Annotating the regulatory genome

- Protein-DNA binding takes cues from, and is correlated with epigenomic features, chromatin states and the transcriptome (binding landscape)

- Binding sites are imperfect copies of each other: difficult to predict
The problem

5' - TCTCTCTCCACCGCTAATTAGGTGATCATGAAAAATGAAAAATTCTATGAGAAGAGTGACGACATCGAAAACATACT

5' - ATGGCGAATCTACCTTTAAACGTTGCCCCACCACGCTCGACCCTTGTGATTTTGATGTTACTGCAGAAATGACTCAACG

5' - CACATCCCAACGATCACCTCACCAGTTTACTGTACTCATCTTCTTATCGCAGCGAACAGTGCCATAAAATATATTTT

5' - TGCAGAACAAGAGAGTCTATTACAAACGAGAAATTAGAGAAGAAATGAAATAATTTCGACAAAATGTATAGTCTATTCTATC

5' - ACAAGGTACCTTCTTGGGCAATCTCACAAGATTTAATATAGTAATTTGTGATGCATATGACTCATCCGAACATGAA

5' - ATGATGACTCATTTTCTCTCTGACATCTACTCAGTCCAATGTTAGAGAAAAATAGAAAAAGCAGAAAATAATAA

5' - GGCGCACAGTCCCGGTTTGGTTATCCCGGTGACTCATTCGTACTCTTTTGGAAAGTGTTGCGATGCTTACACAC

...HIS7

...AR04

...ILY6

...THR4

...AR01

...HOM2

...PRO3
Organization

Motivation

Goals
CRM architecture and ghHMMs
Functional turnover and the CSMET model
Discriminative motif finding using CRFs
Evolutionary analysis of CRMs
Context in the literature: past, present and future
Goals

- How are cis-regulatory modules organized?
  - By performing supervised motif prediction using generalized hierarchical Hidden Markov Models
Goals

- How are cis-regulatory modules organized?
- By performing supervised motif prediction using generalized hierarchical Hidden Markov Models, modelling organization of the CRM and distance between sites.
Goals

- What processes shape the functional evolution of the regulatory regions? (especially functional turnover)
  - By performing supervised motif prediction using a graphical model that models phylogenetic mixtures in an evolutionary setting
Goals

- What processes shape the functional evolution of the regulatory regions? (especially functional turnover)
- By performing supervised motif prediction using a graphical model that models phylogenetic mixtures in an evolutionary setting
Goals

- Analyze which genetic and epigenetic features are correlated with locations of binding sites
  - By performing discriminative supervised motif prediction using Conditional Random Fields

- Analyze evolutionary dynamics in the regulatory genome
  - Using spectral clustering to group evolutionary parameters in the regulatory genome
  - Using mixed membership models to identify structure of k-mer distribution in regulatory regions, and align the dictionaries across species
Organization

Motivation
Goals

**CRM architecture and ghHMMs**
Functional turnover and the CSMET model
Discriminative motif finding using CRFs
Evolutionary analysis of CRMs
Context in the literature: past, present and future
Computational formulation of motif detection

- Given a sequence of nucleotides,
Computational formulation

- Given a sequence of nucleotides, identify subsequences where the transcription factor binds
Supervised vs unsupervised

- Supervised TFBS finding

- Unsupervised TFBS finding
Supervised TFBS prediction

Traditionally modelled as fixed width, table of position-specific frequencies by creating a multiple sequence alignment.
Supervised TFBS prediction

Each instance is called a motif, and the ordered frequency table is called the Position Weight Matrix (PWM)
Traditional PWM scanning

$6 \times 10^{-8} \ 2 \times 10^{-4} \ 9 \times 10^{-8}$

VS

11/30/2012
Traditional PWM scanning

- Typically generates a large number of false positives, as the motifs are noisy copies of each other
Generative models: HMMs
Modelling CRM architecture
Binding site distances
A simplified picture
Generalized and hierarchical HMM
Optimized updates

\[
\tilde{\alpha}(j) \leftarrow \sum_{i \in \tilde{Q}_1} \alpha_t(i) A_{ij}, \quad j \in \tilde{Q}_2,
\]

\[
\tilde{\alpha}(l^{(k)}) \leftarrow \alpha_t((l-1)^{(k)}), \quad 2 \leq l \leq L_k, 1 \leq k \leq K,
\]

\[
\tilde{\alpha}(l^{(k')}) \leftarrow \alpha_t((l+1)^{(k')}, \quad 1 \leq l \leq L_k - 1, 1 \leq k \leq K,
\]

\[
\tilde{\alpha}(b_d^k) \leftarrow \tilde{\alpha}(b_d^k) + \alpha_t(L_k^{(k)}) + \alpha_t(1^{(k')}), \quad 1 \leq k \leq K,
\]

\[
\alpha_{t+1}(j) \leftarrow \tilde{\alpha}(j) B_j(Y_{t+1}), \quad j \in \tilde{Q}
\]

\[
\beta_t(i) \leftarrow \sum_{j=1}^{N} A_{ij} B_j(Y_{t+1}) \beta_{t+1}(j), \quad i \in \tilde{Q}_1, j \in \tilde{Q}_2
\]

\[
\beta_t(l^{(k)}) \leftarrow B_{(l+1)^{(k)}}(Y_{t+1}) \beta_{t+1}((l+1)^{(k)}), \quad 1 \leq l \leq L_k - 1, 1 \leq k \leq K,
\]

\[
\beta_t(l^{(k')}) \leftarrow B_{(l-1)^{(k')}}(Y_{t+1}) \beta_{t+1}((l-1)^{(k')}), \quad 2 \leq l \leq L_k, 1 \leq k \leq K,
\]

\[
\beta_t(L_k^{(k)}) \leftarrow B_{b_d^k}(Y_{t+1}) \beta_{t+1}(b_d^k), \quad 1 \leq k \leq K,
\]

\[
\beta_t(1^{(k')}) \leftarrow B_{b_d^k}(Y_{t+1}) \beta_{t+1}(b_d^k), \quad 1 \leq k \leq K,
\]
Spacer approximation and non-trivial transitions
GH-HMM results

Precision = $\frac{\#TP}{\#TP + \#FP}$, Recall = $\frac{\#TP}{\#TP + \#FN}$

$F1 = \text{Harmonic mean of precision and recall}$
Organization

Motivation
Goals
CRM architecture and ghHMMs
**Functional turnover and the CSMET model**
Discriminative motif finding using CRFs
Evolutionary analysis of CRMs
Context in the literature: past, present and future
Phylogenetic models

\[ P(D|M) = P(X0, X1, X2, X3, X4, X5, X6 \mid \tau, \beta, \theta, \pi) \]

\[ = \sum_{X0, X1, X2} P(X3|X1; \text{tree}) P(X4|X1; \text{tree}) P(X5|X2; \text{tree}) P(X6|X2; \text{tree}) P(X2|X0; \text{tree}) P(X1|X0; \text{tree}) P(X0| \text{tree}) \]
Phylogenetic models

\[ P(D|M) = P(X_0, X_1, X_2, X_3, X_4, X_5, X_6 \mid \tau, \beta, \theta, \pi) \]

\[ = \sum_{X_0, X_1, X_2} P(X_3|X_1; \text{tree}) P(X_4|X_1; \text{tree}) P(X_5|X_2; \text{tree}) P(X_6|X_2; \text{tree}) \]

- **Topology** – how the observations are “tied together”: \( \tau \)
- **Branch lengths** – the length for which the CTMP runs: \( \beta \)
- **Parameters of CTMP** – characterizing the substitution model: \( \theta \)
- **Distribution at root** - maybe stationary dist of CTMP: \( \pi \)
Mixture model based approaches

- Mixture model
- Each block of k-mer could be generated from a background model with probability $1 - \pi_m$ or from a motif model with probability $\pi_m$
- Bernoulli draw for the mixture indicator
- In the spirit of MEME
Function specific phylogenetic models

- Background model Tb
  - Topology invariant unless evidence otherwise
  - Substitution matrix invariant unless evidence otherwise
  - Branch lengths longer than functional sites
  - Root distribution: background frequency

- Motif site-specific, strand-specific model Tm, k, +/-
  - Topology invariant unless evidence otherwise
  - Substitution matrix invariant unless evidence otherwise
  - Branch lengths shorter than background sites
  - Root distribution: from PWM (site specific, strand specific)

Diagram:
- Background multinomial
- Site-specific multinomial from PWM

Lower evolutionary rate
A vanilla HMM ...

- Emits a symbol at every discrete step
- A run of the HMM outputs a sequence
- PhyloHMM outputs a vector of characters
- A run of the PhyloHMM outputs a multiple sequence alignment
Phylo-HMM

- The emission vector $O_i$ is shaded in gray

11/30/2012

Courtesy: McAuliffe
Phylo-HMM

- Emits a vector at each step, generates alignment in a run
State space comparisons

Evolutionary parameters
Associated with each state to help calculate emission probability

Standard HMM

Phylo HMM

Courtesy: Siepel
Phylo-HMM

- A normal HMM, except the emission probabilities are a multinomial distribution over the space of $[\text{ATGC}]^n$, $n$ being the no of sequences in the alignment

- $4^n$ emission probabilities can be pre computed

- But usually calculated on the fly using Felsenstein’s Pruning Algorithm - a special case of the GM Belief Propagation Algorithm on trees
  - Siepel & Haussler, RECOMB 2004, for gene finding
Analogy with HMM

- Emission probability = \( P(O_i \mid S_i = s) \)
  - \( = P(O_i \mid \text{phylogenetic model}_s) \)
  - \( = P(\text{Alignment column}_i \mid \psi_s) \)
  - \( \Rightarrow \) Calculate using the Pruning Algorithm

- Apply standard Viterbi (maximize joint) or posterior decoding on the Forward-Backward matrix

- Baum-Welch algorithm (E-M) for unsupervised settings

- Exactly analogous to single species motif finding

11/30/2012
Missing motifs

melanogaster  cgggatc–gcagttttttacgatcctcaacgg–gttttacg–acc–tccgtccgttt
sechellia  cgggatc–gctgttttttacgatcctcaacgg–gttttacg–acc–tccgtccgttt
simulans  cgggatc–gctgttttttacgatcctcaacgg–gttttacg–acc–tccgtccgttt
yakuba  cgggatc–gctgttttttacgatcctcaacgg–cttttacg–acc–tccgtccgttt
erecta  cgggatc–gccgtttttttacgatcctcaacgg–cttttacg–acc–tccgtccgttt
ananassae  cgggtttc–gctgttttt–ttctcaatcac–gtttatgggtcatac–cgcgtccgttt
pseudoobscura  cggggttccctcagcctttttttacaacactcaaaggatgcgttttgtatgcggtccgctcgggtttt
persimilis  cggggttccctcagcctttttttacaacactcaaaggatgcgttttgtatgcggtccgctcgggtttt

melanogaster  ttatttcat–cggcgaccttgaaagtggccgttttgatggggtggggggtggtttacct
sechellia  ttatttcat–ggggtccttggaatggccgttttgatggggtggggggtggtttacct
simulans  ttatttcat–ggggtccttggaatggccgttttgatggggtggggggtggtttacct
yakuba  ttatttcat–ggggtccttggaatggccgttttgatggggtggggggtggtttacct
erecta  ttatttcat–ggggtccttggaatggccgttttgatggggtggggggtggtttacct
ananassae  ttatttcat–tagccagcttgcctcagagcaggggtggagggtgttgctttacctatgt
pseudoobscura  ttatttcat–ttatttcat–ttatttcat–ttatttcat–ttatttcat–ttatttcat
persimilis  ttatttcat–ttatttcat–ttatttcat–ttatttcat–ttatttcat–ttatttcat

11/30/2012
Functional turnover

<table>
<thead>
<tr>
<th></th>
<th>kr – 6</th>
<th>kr – 5</th>
<th>kr – 4</th>
<th>kr – 3</th>
<th>kr – 2</th>
<th>kr – 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>melanogaster</td>
<td>ATAACCAAT</td>
<td>TTAATCGTT</td>
<td>ACC—GGTTGC</td>
<td>GAAGGGATTAG</td>
<td>ACTGGTTAT</td>
<td>TTAACCGGTTT</td>
</tr>
<tr>
<td>yakuba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>erecta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>. . AA . .</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>bcd – 5</th>
<th>bcd – 4</th>
<th>bcd – 3</th>
<th>bcd – 2</th>
<th>bcd – 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>melanogaster</td>
<td>GTTAATCGG</td>
<td>GAGATTATT</td>
<td>TATAATCGG</td>
<td>GGGATTAGC</td>
<td>GAAGGGATTAG</td>
</tr>
<tr>
<td>yakuba</td>
<td></td>
<td>C. . . . . . . G</td>
<td>. . . . . . . .</td>
<td>. . . . . . . .</td>
<td>. . . . . . . .</td>
</tr>
<tr>
<td>erecta</td>
<td></td>
<td>C. . . . . . . GT</td>
<td>. . . . . . . .</td>
<td>. . . . . . . .</td>
<td>. . . . . . . .</td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>A. . . . . . . A</td>
<td>N/A</td>
<td>A. . . . . . . A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>hb–3</th>
<th>hb–2</th>
<th>hb–1</th>
<th>gt–3</th>
<th>gt–1</th>
</tr>
</thead>
<tbody>
<tr>
<td>melanogaster</td>
<td>CATAAAA–ACA</td>
<td>TTATTTTTTTT</td>
<td>CGATTTTTTTT</td>
<td>CGAGATTATTAGTCATATTG—CAGTTGC</td>
<td></td>
</tr>
<tr>
<td>yakuba</td>
<td></td>
<td></td>
<td></td>
<td>A. . . . . . . A</td>
<td></td>
</tr>
<tr>
<td>erecta</td>
<td></td>
<td></td>
<td></td>
<td>C. . . . . . . C</td>
<td></td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>. . C. . . . .</td>
<td>N/A</td>
<td>N/A</td>
<td>. . . . . . . . TTCATATTTC . C.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>gt–2</th>
<th>gt–1</th>
</tr>
</thead>
<tbody>
<tr>
<td>melanogaster</td>
<td>GACTTTATTGCAGCATCTTG—AACAAATCGTC–GCAGTTTTGTAACAC</td>
<td>GAAAGTCATAAAA–ACACATAATA</td>
</tr>
<tr>
<td>yakuba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>erecta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>T. . . . . . .</td>
<td>AA. T. G. A. . T</td>
</tr>
</tbody>
</table>
CSMET: Phylogenetic mixtures of phylogenies

- Mixture models for evolutionary model selection (EMnEM) →
- Bernoulli draw for mixture variable

- What if the mixture variables are phylogenetically related?
- Output of a “functional” phylogeny
Likelihoods on partial phylogenies

- Marginalize out observed nucleotides present in parts of the phylogeny we are not interested in.
- Turns out to be equivalent to calculating the likelihood of the data on the subtree!

\[
P(A'_i | T^{(l)}) = \sum_{A''_i} P(A'_i, A''_i | T^{(l)}) = \sum_{A''_i} \sum_{v_1:K'} P(V_1:K' = v_1:K', V = A''_i, V' = A'_i)
\]
CSMET: toolkit for calculations

- Calculating likelihoods on the nucleotide phylogeny and functional phylogeny

Nucleotide phylogeny:
F84 model – simplest arbitrary stationary distribution

\[ Q_N = \begin{pmatrix}
(1 + \frac{\kappa}{\pi_Y})\pi_T & \pi_A & \pi_G \\
(1 + \frac{\kappa}{\pi_Y})\pi_T & \pi_A & \pi_G \\
\pi_T & \pi_A & (1 + \frac{\kappa}{\pi_Y})\pi_T \\
\pi_T & \pi_A & (1 + \frac{\kappa}{\pi_Y})\pi_T \\
\end{pmatrix} \]

Functional phylogeny:
Jukes Cantor model

\[ P_F = \begin{pmatrix}
\frac{1}{2} + \frac{1}{2} e^{-2\beta} & \frac{1}{2} - \frac{1}{2} e^{-2\beta} \\
\frac{1}{2} - \frac{1}{2} e^{-2\beta} & \frac{1}{2} + \frac{1}{2} e^{-2\beta} \\
\end{pmatrix} \]

Likelihoods on partial phylogenies:

\[ P(A'_i | T'''(l)) = \sum_{A''_i} P(A'_i, A''_i | T'''(l)) = \sum_{A''_i} \sum_{v_{1:k'}} P(v_{1:k'}, v = A''_i, v' = A'_i) \]
CSMET: emission probabilities

- Emission prob: Prob of block surrounding particular aligned site
- Again, analogous to an HMM, with one twist: \( Z_i \) s not observed

**Joint Probability for an instantiated block**

\[
P(A_t, z_t, z_t^r) = P(A_t | Z_t = z_t, T_m, T_b) P(Z_t = z_t | Z_t^r = z_t^r, T_f)
\]

**Conditional probability for the block**

\[
P(A_t | Z_t = z_t, T_m, T_b) = P(A_t' | T_m') P(A''_t | T_b') P(z_t | z_t^r, T_a) P(z_t^r).
\]

**Emission probability for the block (marginalized)**

\[
P(A_t | z_t^r) = \sum_{z_t} P(A_t, z_t | z_t^r) = \sum_{z_t} P(A_t'(z_t) | T_m'(z_t)) P(A''_t(z_t) | T_b'(z_t)) P(z_t | T_a, z_t^r).
\]
CSMET

- CSMET-HMM:
  - An HMM with emission vector \([A,T,G,C]^*\)
  - Each vector is the output of a generative process involving a mixture of trees
  - Mixture indicator variables themselves generated by a phylogeny
  - Similar scheme to PhyloHMM, except for calculating emission probs

- Schematic of generation and ML inference

- CSMET-HMM:
  \[(S, TN, TF, N, F, n, T, b)\]

Set of nucleotide phylogenetic models corr to each annotation

Corr character sets

Set of functional phylogenetic models corr to each state
CSMET

- To calculate emission probabilities:
  - Calculate likelihoods of nucleotide data for each subtree of the nucleotide phylogeny
  - Calculate likelihood of functional indicators for the functional phylogeny
  - Putting the likelihoods together using conditional independences
  - Marginalize out hidden variables
- The rest would be analogous to an HMM!
Simulating is easy
Estimating is tedious
Varying degrees of turnover

Figure 5. Performance under varying degrees of functional conservation.
doi:10.1371/journal.pcbi.1000090.g005
Varying degrees of motif – bkg contrast

Figure 6. Performance under varying degrees of motif/background contrast.
doi:10.1371/journal.pcbi.1000090.g006
Varying functional turnover rates in single sequence

Figure 7. Effect of varying motif turnover rates across sequence. In the pair of barplots of each method, the left bar corresponds to performance with varying turnover rates ranging from 20% to 32%; the right bar corresponds to performance under a fixed turnover rate at 25%.
doi:10.1371/journal.pcbi.1000090.g007
Real data

- Laws of diminishing return

Figure 9. Comparison of algorithms on motif search performance over 5 motifs on real CRMs.
doi:10.1371/journal.pcbi.1000090.g009
Identified turnover examples
Generative models : HMMs

- May tune to the noise instead of the signal, for noisy data

- Performance saturation to some degree: sophisticated models like Baycics [Lin et al ‘08] have not improved on the state of the art significantly

- Difficult to include a variety of evidence, especially continuous ones
Organization

Motivation
Goals
CRM architecture and ghHMMs
Functional turnover and the CSMET model
**Discriminative motif finding using CRFs**
Evolutionary analysis of CRMs
Context in the literature: past, present and future
Diverse sources of evidence

- Non PWM Genetic data
  - Distance from TSS [Sinha et al 2008]
  - Distance between TFBSs [Ray et al 2008]
- Epigenetic data
  - Nucleosome binding scores [Segal et al 2006, Narlikar et al 2007]
- Combining features [Sharon and Segal 2007, Ernst 2008]
DISCOVER
DIScriminative COnditional random field for motif recoVERy in metazoan genomes

- Conditional random fields: powerful integrational device where features (evidence like phylogeny, proximity to transcribed regions, nucleosome binding affinity score, etc) can be added at will
- Discriminative model
  - Maximizes the conditional probability of the label given the sequence, and not the joint probability of both
- Feature set: carefully chosen from the literature and modelled accordingly
Input & Output

- **Estimation:**
  Input: A set of sequences which have corresponding feature values for each nucleotide and positions of TFBS
  Output: Feature weights (i.e., the learnt conditional probability model)

- **Inference:**
  Input: A set of sequences which have corresponding feature values for each nucleotide and learnt model
  Output: A sequence of ML labels each corresponding to a nucleotide specifying its state: i.e., binding site, CRM-background or background
11/30/2012
Estimation

- Conditional likelihood based framework for estimating the model parameters: the weights corresponding to each field
- Converted to a convex optimization problem
  \( x : \) nucleotides, \( y : \) state labels
- A quasi-Newton method is applied

\[
\hat{\lambda} = \arg \max_{\lambda} L(\lambda \mid y, x)
\]

where

\[
L(\lambda \mid y, x) = P(y \mid x, \lambda)
\]

\[
L(\lambda \mid y, x) = P(y \mid x, \lambda) = \prod_{k=1}^{m} P(y^{(k)} \mid x^{(k)}, \lambda)
\]
Inference

- Use the parameters learnt in the estimation stage for predicting the binding sites
- Corresponding analogs of HMM inference algorithms
- We use marginal decoding, as it allows us to find overlapping motifs

$$\hat{y}_i = \arg \max_{y_i} P(y_i | x, \lambda)$$

- Rank decoding: a fixed fraction of the sequence is classified as predicted binding sites
Sequence-specificity

- PWM based score

- Self Complementarity
  - How similar is it to the reverse complement?
Cis-regulatory grammar

- Transition probabilities between states of the model: modelling the grammar
- Along with the PWM score, correspond to the traditional sources of evidence used by HMM based models
Evolutionary data

- Conservation: function implies conservation from the Neutral Theory

- Presence of repeats: duplicated binding sites or functional turnover?
Chromatin stability & accessibility

- G-C content and melting temperature
  - Correlates with chromatin stability, which facilitates TFBS binding

- Nucleosome binding
  - Nucleosomes affect DNA wrapping and thus the accessibility of DNA to TFs
Distance and location based

- Distance to TSS
- Presence in 5’UTR, or 3’UTR, or intron
Is a feature discriminative?

![Graph showing comparison of mean feature scores for different categories.](image)

- Background
- CRM-background
- Motif

**X-axis:** G–C content vs. PhastCons score

**Y-axis:** Mean feature score
Could it be more discriminative?
How well do the features predict?
Simulation Results
Datasets used

Legend:

- CIS REGULATORY MODULE

11/30/2012
LOOCV Results: Drosophila data

Precision = \#TP / (\#TP + \#FP), Recall = \#TP / (\#TP + \#FN)

F1 = Harmonic mean of precision and recall

Rank decoding
DISCOVER predictions
Cluster Buster predictions

DISCOVER predictions

ClusterBuster predictions
Organization

Motivation
Goals
CRM architecture and ghHMMs
Functional turnover and the CSMET model
Discriminative motif finding using CRFs
**Evolutionary analysis of CRMs**
Context in the literature: past, present and future
Sliding windows characterize evolutionary “profile”

- Which part of the regulatory regions have similar evolutionary parameters?
Does it make sense?

- How well does it cluster based on the overall evolutionary rate?
Heatmap based plot of a regulatory region

- Hairy Stripe 7
Clustering groups of k-mers together

- Only a few clusters represent binding sites, but not statistically significant
- Interpreting the trees are difficult: wide confidence intervals due to a small number of samples for estimating tree
How about selection?

- Proportional splits: only shows constraints on each TFBS, rather than species specific branch elongation/reduction

Motif evolutionary rates

- moj
- vir
- pse
- ana
- yak
- sim
- mel

11/30/2012
Selection

- Varying lengths of phylogenetic tree branches [Eisen 07]
  - Very prone to error under low data scenario (okay for whole genes, not for small sets of binding sites)
- Kreitman-Aguade test for selection in coding sequences, comparing synonymous vs non-synonymous substitutions in cross-species vs population genetic data
  - Important: how to calibrate neutral evolution?
- Alternatively, we may think about a broader idea of selection: evolutionary model preferring one kind of change over another
  - Strongly invariant nucleotide components thus resist change, under negative selection
Basis of the bag of words model

- Each stochastic dictionary is a probability distribution over oligomers of a particular length.

- Multiple stochastic dictionaries are used to model the observation that regulatory modules tend to be involved in multiple regulatory activities.

- The set of stochastic dictionaries are shared across the whole collection of regulatory sequences since functionally related genes share common regulatory roles and are often co-regulated by the same TFs.
Getting rid of sequence alignment

Many to many map defined by regulatory n/w & binding sites

Similar computational framework as single species model

Set of evolving stochastic dictionaries

Star phylogeny

Identification of evolutionary selection forces

Aligned stochastic dictionaries

Comparing across species: aligned bag of words model
Dirichlet distribution

- Distribution over the simplex: can model distribution over categorical / multinomial distributions

\[
\frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i-1}
\]

Wikipedia
1: for each stochastic dictionary $t = 1 \ldots T$ do
2: \hspace{1em} $\phi_t \sim \text{Dirichlet}(\beta)$
3: for each regulatory sequence $d = 1 \ldots D$ do
4: \hspace{1em} $\theta_d \sim \text{Dirichlet}(\alpha)$
5: \hspace{1em} for each motif $i = 1 \ldots N_d$ do
6: \hspace{2em} $z_i \sim \text{Multinomial}(\theta_d)$
7: \hspace{2em} $w_i | z_i \sim \text{Multinomial}(\phi_{z_i})$
Inference

- Nested Gibbs Sampler

- Dirichlets not explicitly modelled, to obtain $P(z \mid w)$, the marginal $P(w)$ is intractable and needs to be sampled

\[
P(w \mid z) = \int \limits_{\phi} P(w \mid z, \phi) P(\phi) \, d\phi
\]

\[
= \left( \frac{\Gamma(W \beta)}{\Gamma(\beta)^W} \right)^T \prod_{t=1}^{T} \prod_{v=1}^{W} \frac{\Gamma(n_{tv} + \beta)}{\Gamma(n_{t*} + W \beta)}
\]

\[
P(z) = \int \limits_{\theta} P(z \mid \theta) P(\theta) \, d\theta
\]

\[
= \left( \frac{\Gamma(T \alpha)}{\Gamma(\alpha)^T} \right)^D \prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\Gamma(m_{dt} + \alpha)}{\Gamma(m_{d*} + T \alpha)}
\]
Function composition vectors: Predicting binding in yeast
Are the dictionaries indicative of regulatory function?
Are they better than PWMs for identifying CRMs?
Function composition vector
Aligned dictionaries across species
Organization

Motivation
Goals
CRM architecture and ghHMMs
Functional turnover and the CSMET model
Discriminative motif finding using CRFs
Evolutionary analysis of CRMs

Context in the literature: past, present and future
Does CHIP-Seq spell the death of motif analysis?

- On the contrary, deconvoluting motif signals from the peaks (cofactors) are now a primary challenge.
Context

Placing the thesis in the context of the literature:

Context

- Hierarchical HMMs for identifying genetic signatures
- HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models, Kadri et al, APBC 2009

[ Disclosure: inbred ]
Context

- Statistical models of turnover for identifying regulatory regions (with simultaneous alignment)
  - Alignment and Prediction of *cis*-Regulatory Modules Based on a Probabilistic Model of Evolution – He et al, 2009
Context

- Functional turnover modelled in an HMM
- Statistical models of binding site loss and gain, Diehl & Siepel, 2010
Context

• Modelling the evolution of functional indicators
• Towards an Evolutionary Model of Transcription Networks, Xie et al, PloS Comp Bio 2011
Context

- Ensemble learner based prediction of binding sites
  - Computational localization of transcription factor binding sites using extreme learning machines, Wang and Do, Springer SC 2012
Context

- Epigenetic feature prediction by integrative modelling using CRFs, WUSTL
Where is the field headed?

- Integrative models: competitive binding, chromatin structure and epigenetic modifications
  - in the end chromatin modifications, chromatin configuration, DNA methylation, TF binding and transcriptional machinery are correlated
  - goal is to distil such correlations into more “theoretical biology”
Where is the field headed?

- Making sense of the transcriptomic, epigenetic and regulatory landscape during differentiation
Thanks

- My advisors Eric Xing and Veronica Hinman
- My committee members Jaime Carbonell, Ziv-Bar Joseph and Martin Kreitman
- Co-authors: Le Song, Mladen Kolar, Suyash Shringarpure, Geir Kjetil Sandve, Selen Uguroglu, Wenjie Fu, Henry Lin
- Every member of the SAILING Lab and Hinman Lab
- Friends and family
Thanks!

http://www.sailing.cs.cmu.edu
Normalizing the scores
Multi-species data pooling

- Simply pool together regulatory regions in related species
  - Hunchback TFBS in Drosophila species demonstration:

  CACCACCTTTTTATGCCGAGTTAAT  D. melanogaster
  GGTCTTTTTCGGATTTAATCGGTATA  D. yakuba
  AGTTCAGCTTTACCGTATTATTTTAAC  D. persimilis
  GCATTATCTCTTTTTTATAAGGTTT  D. mojavensis

- What could be problematic with this approach?
Multi-species data pooling

- Biases analyses towards motifs in a bunch of closely related species – no explicit phylogenetic information used

- No distinction between paralogs and orthologs

- Variation in number of binding sites in orthologous CRMs much less than in CRMs of coregulated genes in same species
  - Signals in one species may be drowned out by cross species signals, or vice versa
Orthologous sequence analysis

- What if the sequences are orthologous?

D. melanogaster
CTTTACGTATTTTAGTTATCGAGTTTTATCTTTCTGCTTGCTATCTCGCGC

D. yakuba
T--TTACGTATTTTAGTTATCGAGTTTTATCTTTCTGCTTGCTATCTCGCGC

D. persimilis
GTTTACGTATTTTAGTTATCGAGTTTTATCTTTTCGCTT------TCTCGC

D. mojavensis
CTTTACGTATTTTAGTTATCAACTTTTGT--TTTGCTT--TGCTTTTTCGC

- Functional regions like TFBSs are more conserved than background
- Phylogenetic dependencies between orthologs may be modelled to get more accurate scores for \( P(\text{data} \mid \text{model}) \)
- Paralogous sequence analysis also possible
Chronology: from one to many

<table>
<thead>
<tr>
<th>Method</th>
<th>Single species</th>
<th>Multispecies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PhyloCon (2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PhyME (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSMET (2008)</td>
</tr>
<tr>
<td>Ensemble models</td>
<td>EMD (2006)</td>
<td>-</td>
</tr>
</tbody>
</table>

- First usage of term phylogenetic footprinting: Tagle *et al*, J Mol Biol 1988: Regulatory regions of paralogous gamma and epsilon globin genes in Galago

- Google scholar hits for “phylogenetic footprinting”
  - 1991 – 2000: 141 hits
  - 2001 – 2009: 1850 hits

- Gibbs Sampling particularly easily adapted to incorporate phylogenetic footprinting
Motif Sampler + footprinting

- A simple idea that works reasonably well
- Wasserman et al, 2000
  - Given an input multiple alignment $A$
  - Compute a score for conservation across the alignment
  - Filter out all regions of the alignment with a score below a threshold $t$
  - Perform Gibbs Motif Sampling on the remaining alignment $A'$

D. melanogaster
CTTTACGTATTAGTTATCGATTATTATTCTGCTTGCTATCTCGCGC

D. yakuba
T--TACGTATTAGTTATCGATTATTATTCTGCTTGCTATCTCGCGC

D. persimilis
GTTTACGTATTAGTTATCGATTATAGTTTTTGCTTTTTCTTCTCGCGC

D. mojavensis
CTTTACGTATTGAGTATTCAATTTTTTGGTTTTTGGCTTTTTGCTTT--TGCTTTTTTCGC
rVISTA

- Select motifs with a PWM score greater than a threshold
- Screens for motifs above a certain threshold for nucleotide conservation
- Two step screening a common way to capture both overrepresentation & conservation
  - Loots et al, Gen Res 2002 [rVISTA]
  - Kellis et al, Nature 2003
  - Wang & Stormo, Bioinf 2003 [PhyloCon]
EMnEM: Expectation maximization on mixtures of phylogenies

\[ L = \prod_{i=0}^{N-w} \sum_{m_i} p(m_i) \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k | A_{kb}, m_i) p(A_{kb} | m_i) \]

• E-step:
  - Mixture parameter:
    \[ \langle m_i \rangle = p(m_i | X, Y) = \frac{p(m_i) p(X, Y | m_i)}{p(X, Y)} \]
    \[ p(X, Y | m_i) = \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k | A_{kb}, m_i) p(A_{kb} | m_i) \]
    \[ p(X, Y) = \sum_{m_i} p(X, Y | m_i) p(m_i) \]
  - Ancestral nucleotide:
    \[ \langle A_{ib} \rangle = p(A_{ib} | X_i, Y_i) = \sum_{m_i} p(A_{ib} | X_i, Y_i, m_i) p(m_i) = \sum_{m_i} \frac{p(A_{ib}) p(X_i, Y_i | A_{ib}, m_i)}{p(X_i, Y_i | m_i)} p(m_i) \]

• M-step:
  \[ \langle \ln L_c \rangle = \sum_{i=0}^{N-w} \sum_{m_i} \langle m_i \rangle \left[ \ln \tau_m + \sum_{k=i}^{i+w-1} \sum_{b=0}^{3} \langle A_{kb} \rangle (\ln p(X_k, Y_k | A_{kb}, m_i) + \ln f_{mb}) \right] \]
  \[ \frac{\partial \langle \ln L_c \rangle}{\partial \tau_m} = 0, \quad \frac{\partial \langle \ln L_c \rangle}{\partial f_{mb}} = 0 \quad \text{and} \quad \frac{\partial \langle \ln L_c \rangle}{\partial X_{mk}} = 0 \]