Tools for modelling regulatory genomics data in terms of predicted regulatory sites on the DNA
How is the regulatory code in the DNA `read out’ to control cell fate and identity?

How do gene regulatory networks function as systems.

• What is a cell type?
• How is cell identity stabilized?
• Where is the information? What does not matter?

My worries

• We think we know/measure a lot, but there is orders of magnitude more we do not know.
• High-throughput measurements full of artefacts and biases that we poorly understand.
• Nowhere near the ability to meaningfully model what is going on.

What useful things can a serious computational biologist do?
Expression noise facilitates the *de novo* evolution of gene regulation

**Experimental observations**
- We evolved synthetic promoters *de novo* in *E. coli* under carefully-controlled selective conditions.
- No evidence *E. coli* promoters have been selected to lower noise.
- Promoters of regulated genes have been selected to *increase* noise.

**Theory**
- Coupling a regulator to a target promoter has two effects:
  2. Noise-propagation.
- Noise-propagation alone can act as a rudimentary form of regulation.
- Accurate regulation can evolve smoothly along a continuum in which noise-propagation and condition-response act in concert.
- Explains the general association between noise and regulation.

How is the regulatory code in the DNA `read out’ to control cell fate and identity?

How do gene regulatory networks function as systems.
- What is a cell type?
- How is cell identity stabilized?
- Where is the information? What does not matter?

My worries
- We think we know/measure a lot, but there is orders of magnitude more we do not know.
- High-throughput measurements full of artefacts and biases that we poorly understand.
- Nowhere near the ability to meaningfully model what is going on.

What useful things can a serious computational biologist do?
Develop simple, robust, and transparent methods that help guide experimental efforts.
Motif Activity Response Analysis:
Modeling gene expression and chromatin state in terms of TFBS using a linear model

Forrest et al. Nat Genet 2009
Balwierz et al. Genome Res 2014
Example: Response of Human umbilical vein endothelial cells to treatment with TNFα

Time course measurements: Wada et al. A Wave of nascent transcription on activated human genes. PNAS 2009

Top 3 most significant motifs

http://ismara.unibas.ch

Predicted regulatory interaction
Predicted interaction with experimental support.
Enriched target gene category
Completely automated prediction of regulatory interactions from high-throughput data

Upload micro-array, RNA-seq, or ChIP-seq data and predict:
- Key regulators (TFs/miRNAs) in the system.
- Regulator activities across the input samples.
- Sets of target genes and pathways for each regulator.
- The regulatory sites on the genome through which the regulators acts.
- Interactions between the regulators.

Balwierz et al. *Genome Res* 2014
Modeling TF binding specificity
Going beyond position-specific weight matrices
Probability for a set of sequences to derive from a common WM

Probability of observing the set of sequences $S$ when sampling from the known WM $w$:

$$P(S \mid w) = \prod_{i=1}^{l} P(S_i \mid w^i) = \prod_{i=1}^{l} \left[ \prod_{\alpha} (w^i_{\alpha})^{n^i_{\alpha}} \right]$$

$n^i_{\alpha} = \text{number of times letter } \alpha \text{ appears at position } i \text{ in } S$.

$w^i = (w^i_a, w^i_c, w^i_g, w^i_t)$ where $w^i_{\alpha} = \text{probability letter } \alpha \text{ appears at position } i$.

- The weight matrix $w$ is an unknown variable in our model.
- Probability theory prescribes that we should introduce a prior probability distribution for it and integrate it out of our probability.
- Using the Dirichlet prior:
  $$P(w^i) \propto \prod_{\alpha} (w^i_{\alpha})^{\lambda-1}$$

- One obtains:
  $$P(S) = \int P(S \mid w^i)P(w^i)dw^i = \frac{\Gamma(4\lambda)}{\Gamma(n + 4\lambda)} \prod_{\alpha} \frac{\Gamma(n^i_{\alpha} + \lambda)}{\Gamma(\lambda)}$$
Including pairwise dependencies

We extend the PWM to a Dinucleotide Weight Tensor (DWT) model that *allows arbitrary pairwise dependencies* between positions.

\[ P(S_i, S_j) = \int P(S_i, S_j | \omega_{ij}) P(\omega_{ij}) d\omega_{ij} = \frac{\Gamma(16\tilde{\lambda})}{\Gamma(n + 16\tilde{\lambda})} \prod_{\alpha, \beta} \frac{\Gamma(n_{ij}^{\alpha\beta} + \tilde{\lambda})}{\Gamma(\tilde{\lambda})} \]

Likelihood ratio: \[ R_{ij} = \frac{P(S_i, S_j)}{P(S_i)P(S_j)} \approx \exp(nI_{ij}) \]
Probability given a dependence tree

Sequence alignment $S$

**PWM model:**
- Each position is independent:
  $P(S) = \prod_{i} P(S_i)$

**DWT model:**
- The probability of observing a given nucleotide at a position $i$ of the alignment depends on the nucleotide at one other position $\pi(i)$.
- The set of ‘parents’ $\pi(i)$ of all positions $i$ determine a *spanning tree* of the set of positions.

Factorization:

$$P(S | \pi) = P(S_r) \prod_{i \neq r} \frac{P(S_i, S_{\pi(i)})}{P(S_{\pi(i)})} = \prod_{i} P(S_i) \prod_{(i,j) \in \pi} R_{ij}$$
Summing over spanning trees

Since we do not know the spanning tree, probability theory prescribes we should sum over all possible spanning tree (with uniform prior):

\[
P(S) = \sum_T \frac{P(S | \pi)}{|\pi|} = \frac{1}{|\pi|} \prod_i P(S_i) \sum_{\pi} \left[ \prod_{(i,j) \in \pi} R_{ij} \right]
\]

**Example**: for 3 positions we would sum over the three possible spanning trees:

\[
P(S) \propto R_{12} R_{13} + R_{13} R_{23} + R_{12} R_{23}
\]

Using Kirchhoff/Matrix-tree theorem

Laplacian matrix of \( R \): \( L(R)_{ij} = \delta_{ij} \sum_k R_{ik} - R_{ij} \)

Define: \( D(R) = \) Any minor (determinant) of the \( L(R) \), then:

\[
\sum_{\pi} \left[ \prod_{(i,j) \in \pi} R_{ij} \right] = D(R)
\]

Final probability under the DWT model:

\[
P(S) = \frac{D(R)}{|\pi|} \prod_i P(S_i)
\]
Predicting TFBS and motif finding with DWTs

Probability that a sequence $s$ derives from the same motif as the set of sequences $S$:

$$P(s \mid S) = \frac{P(s, S)}{P(S)} = \frac{D(R(s, S))}{D(R(S))} \prod_i \frac{P(S_i, s_i)}{P(S_i)} = \frac{D(R(s, S))}{D(R(S))} \prod_i \frac{n_{s_i} + \lambda}{n + 4\lambda}$$

Expectation Maximization procedure for motif finding

1. Predict sites with initial motif
2. DWT defined by dinucleotide counts in sites.
3. Predict sites with current DWT.
Testing DWT performance on ChIP-seq datasets from ENCODE

- **Data**: ChIP-seq data-sets from ENCODE for 83 different human TFs.
- **Processing of each TF’s data-set:**
  - top 1000 peaks from Crunch.
  - Divide into 500 *training regions*, and 500 *test regions*.

- Fit both a PWM and DWT on the training regions.

- Calculate enrichment score on the test set mixed with background regions of equal dinucleotide composition.
An enrichment score for ChIP-seq

- **Data:** IP ‘fished’ our peak sequences from a much larger collection of DNA fragments.
- **Assumption:** The probability to fish (=IP) a sequence is proportional to the *number of copies of the TF(s) bound to it*.
- **Likelihood model:**
  - Peak sequences $P$ + Background sequences $B$ (= random seqs with same lengths and nucleotide composition).
  - Given a set of motifs $w$, and their concentrations $c$, calculate the expected number of bound TFs $n(s | w, c)$ at each sequence $s$.
  - Probability to IP sequence $s$: $P(s | w, c) = \frac{n(s | w, c)}{\sum_{s' \in P \cup B} n(s' | w, c)}$
  - Probability to IP *all* sequences in $P$ and only the sequences in $P$: $P(D | w, c) = \prod_{s \in P} P(s | w, c)$
  - Likelihood for a motif set $w$: $P(D | w) = \max_c P(D | w, c)$
DWTs often outperform PWMs and never overfit
CRUNCH: A completely automated webserver for ChIP-seq data analysis

Motivation

- For tools like MARA we would like to automatically process available ChIP-seq data to curate new motifs and annotate where they bind.

- However, ChIP-seq data analysis is still *wild-west*:
  - Almost no standardized procedures even within consortia like ENCODE.
  - Cannot meaningfully compare results from different studies.
### Overview of CRUNCH analysis steps

<table>
<thead>
<tr>
<th>Preprocessing</th>
<th>Peak Calling</th>
<th>Regulatory Motif Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. BED and WIG Extraction</td>
<td></td>
<td>12. Motif Scoring and Annotation</td>
</tr>
<tr>
<td>5. Fragment Size Estimation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Detecting enriched regions

**Preprocessing**
1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

**Peak Calling**
6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

**Regulatory Motif Analysis**
9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

- Slide 500 bp window across genome.
- Quantify significance of the enrichment of ChIP-seq over input DNA.
Bayesian model for identifying enriched regions

Noise model for read-counts in un-enriched windows

• Multiplicative noise plus Poisson sampling, i.e. as previously developed in:

**Balwierz** PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, **van Nimwegen** E.

Variables:

• \(n,m\) = reads in ChIP/input sample.
• \(N,M\) = total reads in ChIP/input sample.
• \(\sigma\) = standard-deviation of the multiplicative noise.
• \(\mu\) = shift in average log read-density.

**Probability of observing** \(x\):

\[
P(x | \mu, \sigma) \propto \exp \left[ -\frac{(x - \mu)^2}{2 \left( 2\sigma^2 + \frac{1}{n} + \frac{1}{m} \right)} \right]
\]

**Enrichment** \(x\):

\[
x = \log \left( \frac{n}{N} \right) - \log \left( \frac{m}{M} \right)
\]

Mixture model

• The enrichment \(x_i\) for each window \(i\) derives from either the noise model or a uniform distribution (= ‘something else’):

\[
P(D | \mu, \sigma, \rho) = \prod_i \left[ P(x_i | \mu, \sigma) \rho + \frac{1 - \rho}{x_{\text{max}} - x_{\text{min}}} \right]
\]

• We fit \(\mu, \sigma, \text{ and } \rho\) to maximize \(P(D | \mu, \sigma, \rho)\), and calculate an enrichment z-score for each window.
The noise model accurately captures the observed genome-wide enrichment statistics.

Z-statistic for each window:

\[
z_i = \log\left(\frac{n_i}{N}\right) - \log\left(\frac{m_i}{M}\right) - \mu
\]

\[
\sqrt{2\sigma^2 + \frac{1}{n_i} + \frac{1}{m_i}}
\]

As far as we are aware, ours is the only peak-finder that demonstrably matches the data's statistics.
Overview of the analysis steps

**Preprocessing**
1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

**Peak Calling**
6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

**Regulatory Motif Analysis**
9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation
De novo motif finding

Top 1000 peak sequences

500 ‘Training set’
For de novo motif finding

500 ‘Test set’
For downstream motif selection

1. Align with orthologous regions (7 mammals/10 Drosophilids)

…acccattctacggagctgactcagatcagatacagtcg…
…accaattctacggagcttagattgagtaacacagatcag…
…acccattctacggagctgactcagatcagatacagtcg…
…acccattctacggagctgactcagatcagatacagtcg…
…acccattctacggagctgactcagatcagatacagtcg…
…acccattctacggagctgactcagatcagatacagtcg…

2. Identify motifs with PhyloGibbs

PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.
Siddharthan R¹, Sigia ED, van Nimwegen E.

3. Refine motifs with MotEvo

MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences.
Arnold P¹, Erb J, Pachkov M, Molina N, van Nimwegen E.

4. Result
Up to 24 candidate de novo motifs
Library of known motifs

Library of 2325 known motifs (position-specific weight matrices) from:

**HOCOMOCO**

**HOMER**

**UniPROBE**

**ENCODE**

**HTSELEX**

**SwissRegulon**

**Task**

Find a set of complementary known/de novo motifs that jointly explain the observed binding peaks of the test set.
Sorted list of most enriched motifs

**Final enrichment score**: Per sequence likelihood ratio relative to *randomly selecting sequences*:

\[ E_w = \left[ \frac{P(D \mid w, c)}{P(D \mid \text{random})} \right]^{1/|P|} = \left[ \prod_{s \in P} \frac{n(s \mid w, c)}{\langle n \rangle_B} \right]^{1/|P|} \]

\[ |P| = \text{Number of binding peaks.} \]

We sort all known and *de novo* motifs by their enrichment.

**Example (NRF1 ChIP-seq):**

<table>
<thead>
<tr>
<th>Motif Name</th>
<th>Sequence Logo</th>
<th>Enrichment (log-Likelihood Ratio)</th>
<th>Precision and Recall</th>
<th>Prediction - Observation Correlation</th>
<th>Enrichment at Binding Sites</th>
<th>Number of Positively Predicted Peaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTSELEX.NRF1.NRF.full.dimeric.wm1</td>
<td><img src="image" alt="Logo" /></td>
<td>38.364 (1823.56)</td>
<td>0.9271</td>
<td>0.6756</td>
<td>9.423</td>
<td>3977/9227</td>
</tr>
<tr>
<td>denovo_WM_17</td>
<td><img src="image" alt="Logo" /></td>
<td>33.838 (1760.787)</td>
<td>0.9226</td>
<td>0.6441</td>
<td>8.7474</td>
<td>4102/9227</td>
</tr>
<tr>
<td>denovo_WM_23</td>
<td><img src="image" alt="Logo" /></td>
<td>21.864 (1542.42)</td>
<td>0.9217</td>
<td>0.6572</td>
<td>7.6023</td>
<td>4749/9227</td>
</tr>
<tr>
<td>NRF1.p2</td>
<td><img src="image" alt="Logo" /></td>
<td>17.218 (1422.981)</td>
<td>0.8688</td>
<td>0.6509</td>
<td>8.1677</td>
<td>4290/9227</td>
</tr>
</tbody>
</table>
Selecting an optimal set of complementary motifs

Initialize motif set \{w\} with best motif w.

Iterate:
1. For each of the remaining motifs \(w'\), add \(w'\) to \{w\}, and calculate new \(E_{\{w\}}\).
2. Select \(w'\) that maximizes \(E_{\{w\}}\) and add to the set \{w\}.

Stop when the enrichment increases by less than 5%.

Example: ATF2 from ENCODE

<table>
<thead>
<tr>
<th>Motif Name</th>
<th>Sequence Logo</th>
<th>Motif Ensemble Enrichment (Motif Ensemble log-Likelihood Ratio)</th>
<th>Enrichment (log-Likelihood Ratio)</th>
<th>Precision and Recall</th>
<th>Prediction - Observation Correlation</th>
<th>Enrichment at Binding Sites</th>
<th>Number of Positively Predicted Peaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>denovo_WM_16</td>
<td><img src="image" alt="" /></td>
<td>1.648 (305.878)</td>
<td>1.648 (305.878)</td>
<td>0.4515</td>
<td>0.0609</td>
<td>1.159</td>
<td>25751/29180</td>
</tr>
<tr>
<td>HTSELEX.CREB3.bZIP.full.dimeric.wm1</td>
<td><img src="image" alt="" /></td>
<td>2.746 (503.08)</td>
<td>1.303 (131.994)</td>
<td>0.2624</td>
<td>0.1125</td>
<td>2.2613</td>
<td>655/29180</td>
</tr>
<tr>
<td>HCMC.BATF_s1.wm</td>
<td><img src="image" alt="" /></td>
<td>3.381 (606.679)</td>
<td>1.353 (150.403)</td>
<td>0.2871</td>
<td>0.1028</td>
<td>1.7715</td>
<td>5218/29180</td>
</tr>
<tr>
<td>HCMC.SP1_f2.wm</td>
<td><img src="image" alt="" /></td>
<td>4.2 (714.638)</td>
<td>1.323 (139.465)</td>
<td>0.2733</td>
<td>-0.0401</td>
<td>0.6875</td>
<td>5619/29180</td>
</tr>
</tbody>
</table>

Motif combination better explains the binding data.
We observe two types of TFs:
Solitary binders vs. TFs co-binding with other TFs
Top motifs for a TF are consistent across experiments

- Top enriched motifs for a TF are highly consistent across different cell lines/experiments.
- Even when motifs are extremely similar!

This suggests we can select a ‘best’ motif for each solitary TF in a meaningful way.
Summary and acknowledgments

Crunch:
• Automated webserver for comprehensive ChIP-seq analysis.
• Realistic statistical model.
• Explain the binding peaks in terms of a complementary set of motifs.

Check BioRxiv in the coming days for the papers!

Dinucleotide Weight Tensors:
• Rigorous Bayesian model allowing arbitrary dependencies.
• Zero tunable parameters.
• DWTs never overfit and outperform PWMs for many TFs.
• Source code for motif finding and TFBS prediction using DWTs.

Severin Berger
CRUNCH

Lukas Burger
Original DWT model

Saeed Omidi
DWTs for TFBS prediction