Core Promoter Analysis

Uwe Ohler
Institute for Genome Sciences and Policy
Duke University
uwe.ohler@duke.edu
Computational Biology of Gene Regulation

Focus of my group:

- **Sequence Analysis**
  - In this area, we mostly work on models of (conservation of) **regulatory regions**
    - Modeling of transcription start sites
    - Condition-specific regulatory motifs
  - Also: Post-transcriptional regulation

- **Image analysis**
  - New high-throughput data source to study gene expression
  - On single gene level, but precise spatiotemporal information (in living organisms)
Steps in gene regulation

5’ promoter exon I intron exon II intron exon III 3’

TSS 5’ UTR start codon stop codon 3’ UTR

addition of 5’ cap

cleavage and addition of polyA tail

polyA signal

RNA splicing

AAAA

transport to cytoplasm

AAAA

translation

Protein
An extremely simplified view of eukaryotic transcription

- Specific information about functional context of genes: proximal promoter/enhancers
  - Binding sites of specific transcription factors confer activation at the right developmental stage or tissue

- General information: the core promoter
  - Region around the transcription start site (TSS) where RNA polymerase II (pol-II) interacts with general transcription factors
  - Potentially far away from the translation start site
Interactions in core promoters (simple „modules“)

AAACCGTTAAAAAACAGAGCAGGCAGCGTCAGCAGCAAGAGAAGAGGTCGAGGCGGAGGCGCAAGA
CGTGCTGCCTCCCAATAAAAACCGGTGCAGTGAGTCAGTGTTGTTGTTGCCCATGTCGCGAGCGGACGATC

[Other known variability: tissue-specific TAFs; TRFs]

Species specific differences

- The core protein complex is conserved, but the cis-regulatory sequences are not (*quite*)
- Example: TATA box
  - Ca 80 nt upstream in yeast, 25 nt in other eukaryotes
- Example: Initiator
  - A strong 5-6 nt motif in flies, a weak 2 nt preference in mammals
- Example: CpG islands
  - A mammalian phenomenon related to DNA methylation
  - 50-60% of genes have it
Inferring TSSs from genome wide data

- Oligo-capped cDNAs
  - 5’ mRNA cap structure is replaced by a unique synthetic oligo (RIKEN cap-trapper; Stapleton et al 2002)
  - “guarantees” that cDNA is sequenced up to the 5’ end

- 5’ SAGE/ CAGE
  - High-throughput version: sequence only the first 15-20 nt of each transcript
  - Yields a profile of TSS actually used in the cell
  - Yeast (Dietrich/Duke), Mammals (Carninci/RIKEN): > 11 mio. Tags

- Important issues: TS site vs region vs alt. TSS; definition/conservation of TSS
**High throughput pictures of TSS usage**

- High-throuput SAGE approaches (5’SAGE/CAGE) provide extensive data on individual transcription initiation events
  - Here: mouse

Is transcription initiation a sloppy event?

- CAGE data seems to indicate so
- Related: evolution of core promoters in bacteria
  - Started with a random pool of ~35nt long sequences as promoters of a selective gene
  - Selection & mutation by error-prone PCR
  - Instead of one strong promoter, the result was a set of overlapping weak initiation sites
    [Terry Hwa lab, UCSD]
- Possibility: Often, there is no strong pressure to maintain one precise start site
  - But: reproducible tissue-specific differences
    [Kawaji et al., Genome Biol 2006]
Inferring TSSs from cDNAs

- Clustering EST alignments (2001/2002)
  - 237,471 5' EST sequences aligned with sim4 (Florea et al.)
  - 1,941 cap-trapped clusters selected as follows:
    - Only if spliced or overlapping gene annotation
    - Only most 5' cluster with minimum distance 1,000 bp
    - >30% of ESTs in cluster within a 5' window of 10 bp

- Comparison with 205 known promoters
  (CPD, Kutach and Kadonaga, 2000)
  - Consensus strings allowing 1 mismatch
  - Inr: TCA(G/T)T(C/T) within −10/+10
    - CPD: 67.3%, our set: 62.8%
  - TATA box: TATAAA within −45/-15
    - CPD: 42.4%, our set: 28.3%
## Motifs found in core promoters

<table>
<thead>
<tr>
<th>Motif</th>
<th>Pictogram</th>
<th>Consensus</th>
<th># seq</th>
<th>E value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Motif 1 Pictogram" /></td>
<td>YGGTCACACTR</td>
<td>311</td>
<td>5.1e-415</td>
</tr>
<tr>
<td>2 DRE</td>
<td><img src="image2.png" alt="Motif 2 Pictogram" /></td>
<td>WATCGATW</td>
<td>277</td>
<td>1.7e-183</td>
</tr>
<tr>
<td>3 TATA</td>
<td><img src="image3.png" alt="Motif 3 Pictogram" /></td>
<td>STATAWAAR</td>
<td>251</td>
<td>2.1e-138</td>
</tr>
<tr>
<td>4 INR</td>
<td><img src="image4.png" alt="Motif 4 Pictogram" /></td>
<td>TCAGTYKNNT</td>
<td>369</td>
<td>3.4e-117</td>
</tr>
<tr>
<td>5 Ebox</td>
<td><img src="image5.png" alt="Motif 5 Pictogram" /></td>
<td>AWCAGCTGWT</td>
<td>125</td>
<td>2.9e-93</td>
</tr>
</tbody>
</table>
### Motifs found in core promoters

<table>
<thead>
<tr>
<th></th>
<th>Motif</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>KTYRGTATWTTT</td>
<td>107</td>
<td>1.9 e-62</td>
</tr>
<tr>
<td>7</td>
<td>KNNCAKCNCTR</td>
<td>197</td>
<td>1.9 e-63</td>
</tr>
<tr>
<td>8</td>
<td>YGGCARCGSYSS</td>
<td>82</td>
<td>5.1 e-29</td>
</tr>
<tr>
<td>9</td>
<td>CRWMGCGWKCG GTTS</td>
<td>56</td>
<td>1.9 e-12</td>
</tr>
<tr>
<td>10</td>
<td>CSARCSSAACGS</td>
<td>40</td>
<td>8.3 e-9</td>
</tr>
</tbody>
</table>

Ohler et al., Genome Biol 3:0087 (2002)
Positional distribution of motifs
Validation/definition of MTE

Analysis of Mutations in the MTE That Do Not Overlap with the DPE

Lim et al., Genes Dev 2004
## Frequency of co-occurrence

<table>
<thead>
<tr>
<th>Motif X</th>
<th>% seqs w/ X</th>
<th>% seqs with Motif X also containing Motif below</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>M1</td>
<td>25.1</td>
<td>100.0</td>
</tr>
<tr>
<td>DRE</td>
<td>26.0</td>
<td>20.6</td>
</tr>
<tr>
<td>TATA</td>
<td>19.3</td>
<td>17.1</td>
</tr>
<tr>
<td>INR</td>
<td>26.3</td>
<td>12.1</td>
</tr>
<tr>
<td>M6</td>
<td>15.8</td>
<td><strong>45.1</strong></td>
</tr>
<tr>
<td>DPE</td>
<td>7.9</td>
<td>15.6</td>
</tr>
<tr>
<td>MTE</td>
<td>8.5</td>
<td>18.2</td>
</tr>
</tbody>
</table>
A new core promoter module

- Motif 1 has a weak preference for location at the TSS
- The motif 6/1 pair is reminiscent of the TATA/Inr module
Core promoter motif modules

- TATA box/Inr: much less frequent (<25%)
- Motif 2: DNA replication element (DRE) factor binding site
  - Part of complex with TBP-replacing factor 2 (TRF2) in TATA-less promoters (Hochheimer *et al*, *Nature* 2002)
- DPE+MTE: *Two* distinct downstream motifs
- Motif 1: correlates with TSS location and motif 6
  → *several subclasses of core promoters (depending on TFIID/DNA conformation*)
McPromoter system structure
Computational approaches

- Have a long history – recognizing E.coli promoters was one of the earliest “annotation” efforts
- Two (heuristic) approaches early on:
  - Signal/motif-based: explicit modeling of binding sites
  - Content-based: similar to ORF recognition
- Later: Combination
  - Probabilistic models, e.g. HMMs (generative)
  - Support vector machines (discriminative)
- TSS recognition vs. coding gene start recognition
  - Some approaches use additional gene features
Modeling promoter subclasses

- Split promoter training set in overlapping partitions defined by the presence of core promoter modules
  - ~85% of promoters have a good hit to at least one of these motifs
- Perform iterative cross-validation re-assignment (similar to k-means)
- -> Five parallel core promoter models
  - MTE does not form stable class of its own
- Performance on classification promoter/non-promoter:
  - 94% equal recognition rate (up from 89%); ROC integral 0.98 (1.0 means perfect classification)
Clustering of core promoters

![Bar chart showing the number of promoters with different initial motifs for various motifs and their initial and final states.](attachment:image.png)
Modeling promoter subclasses

- Inr/DPE
- Inr only
- TATA/Inr
- DRE
- Motif 1/6
5 subclasses of Drosophila core promoters
Comparison of results, Adh region

92 promoters from full-length cDNA alignments
- Positive region: -500/+50
  (Sn: sensitivity; Sp: specificity; AP: addtl predictions/nt)

<table>
<thead>
<tr>
<th>McPromoter 2002 (one model)</th>
<th>Sharan &amp; Myers 2005</th>
<th>McPromoter 2006 (five models)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sn</td>
<td>Sp</td>
<td>Sn</td>
</tr>
<tr>
<td>20</td>
<td>69</td>
<td>20</td>
</tr>
<tr>
<td>37</td>
<td>51</td>
<td>35</td>
</tr>
<tr>
<td>52</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>67</td>
<td>29</td>
<td>65</td>
</tr>
</tbody>
</table>
Alternative transcription start sites

- A large fraction of genes has more than one TSS
- Here, we mean distinctly separate TSS (~100 nt or more apart, not small scale fluctuation)
  - Alternative 5’ UTRs
  - Alternative translation start sites
  - Tissue-specific promoters
- Prominent example: e.g. protocadherin genes
Evolution/turnover of TSS

- If core promoter motifs are only there to define a TSS, they should frequently turn over
  - Position changes
  - Motif changes, i.e. TATA box replaced by DPE
- If they however provide *context* information, this should not be the case
  - Core promoter/enhancer interaction
  - Tissue-specific activation of alternative TSS
Scenario I: Conservation

- Alignment of human and mouse promoters
  - TSS is inferred in one species and mapped to other species by genomic alignments

Jin et al., *BMC Bioinformatics* 2006
Scenario II: Turnover of TSS

Frith et al., *Genome Res* 2006
Revisiting TSS

- Refined cluster protocol for ESTs
  - Large groups: Separated by > 100 nt
  - Enough tags available: Determine TSS positions
  - Requirements:
    - TSS defined by >=2 tags, with >=3 tags within 10 nucleotides;
    - Upstream of annotated ATG
    - Library-specific information
- Two RIKEN libraries: embryo and adult head
  - Embryo: 2,872 genes w/4,046 TSS
  - Head: 1,682 genes w/2,144 TSS
- Total: 3,683 genes w/6,190 TSS
Current dataset

- More stringent criteria to include TSS from other libraries
- Example:

Corresponding_TSS_frequencies [(4)(3)(4)(7)]
Number_of_tags_from_RE_RIKEN_EMBRYO [(0)(0)(0)(0)]
Number_of_tags_from_RH_RIKEN_HEAD [(4)(0)(0)(7)]
Number_of_tags_from_LD_EMBRYO [(0)(0)(0)(0)]
Number_of_tags_from_GM_OVARY [(0)(1)(0)(0)]
Number_of_tags_from_HL_ADULT_HEAD [(0)(0)(0)(0)]
Number_of_tags_from_GH_ADULT_HEAD [(0)(1)(0)(0)]
Number_of_tags_from_LP_Larvae_Pupae [(0)(0)(0)(0)]
Number_of_tags_from_SD_SCHNEIDER CELLS [(0)(1)(0)(0)]
Number_of_tags_from_AT_ADULT_TESTES [(0)(0)(4)(0)]
Number_of_tags_from_UT_ADULT_TESTES [(0)(0)(0)(0)]
Number_of_tags_from_OTHERS [(0)(0)(0)(0)]
Example of a complex TSS arrangement in Drosophila

- CG33113: Chr 2L
- TSS position/#tags/array support:
  - 5006561 (15) 1-2
  - 5004921 (5) 8
  - 5000362 (21) 3-6
  - 4999500 (4)
  - 4997377 (10) 5-8
## Related work

<table>
<thead>
<tr>
<th>Sequence logo</th>
<th>Consensus sequence</th>
<th>Name</th>
<th>Common name</th>
<th>Other #</th>
<th>8-mers in consensus</th>
<th>Peak bps from TSS</th>
<th>CF+</th>
<th>CF-</th>
<th>Pooled peaks</th>
<th>Unique genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TATAAAA</strong></td>
<td>STATAAA</td>
<td>DMP1</td>
<td>TATA</td>
<td>3</td>
<td>30</td>
<td>-32</td>
<td>24</td>
<td>2</td>
<td>48-49</td>
<td>511</td>
</tr>
<tr>
<td><strong>TCAGTY</strong></td>
<td>TCAGTY</td>
<td>DMP2</td>
<td>INR</td>
<td>4</td>
<td>101</td>
<td>-2</td>
<td>29</td>
<td>2</td>
<td>49-51</td>
<td>1,501</td>
</tr>
<tr>
<td><strong>TCATTTCG</strong></td>
<td>TCATTTCG</td>
<td>DMP3</td>
<td>INR1</td>
<td>5</td>
<td>2</td>
<td>15</td>
<td>3</td>
<td>50-51</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td><strong>CGGTTT</strong></td>
<td>KCGGTTT</td>
<td>DMP4</td>
<td>DPE</td>
<td>9</td>
<td>10</td>
<td>+25</td>
<td>14</td>
<td>4</td>
<td>51-52</td>
<td>147</td>
</tr>
<tr>
<td><strong>CGGACCGT</strong></td>
<td>CGGACGT</td>
<td>DMP5</td>
<td>DPE1</td>
<td>11</td>
<td>+26</td>
<td>18</td>
<td>3</td>
<td>51-52</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td><strong>CAACCCCT</strong></td>
<td>CARCCCT</td>
<td>DMV1</td>
<td>5</td>
<td>-60 to -41</td>
<td>5</td>
<td>5</td>
<td>47-51</td>
<td>311</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TGGYAAC</strong></td>
<td>TGGYAACR</td>
<td>DMV2</td>
<td>8</td>
<td>-20 to -1</td>
<td>11</td>
<td>13</td>
<td>46-51</td>
<td>311</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CATC CTA</strong></td>
<td>CAYCNCTA</td>
<td>DMV3</td>
<td>7</td>
<td>+1 to +20</td>
<td>11</td>
<td>18</td>
<td>46-52</td>
<td>604</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GGTCAAC</strong></td>
<td>GGYCACAC</td>
<td>DMV4</td>
<td>1</td>
<td>-20 to -1</td>
<td>42</td>
<td>23</td>
<td>46-51</td>
<td>649</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TGGTAT</strong></td>
<td>TGGTATTT</td>
<td>DMV5</td>
<td>6</td>
<td>-60 to -41</td>
<td>3</td>
<td>11</td>
<td>45-51</td>
<td>287</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence logo</th>
<th>Consensus sequence</th>
<th>Name</th>
<th>Common name</th>
<th>Other #</th>
<th>8-mers in consensus</th>
<th>Peak bps from TSS</th>
<th>CF+</th>
<th>CF-</th>
<th>Pooled peaks</th>
<th>Unique genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GAGAGCG</strong></td>
<td>GAGAGCG</td>
<td>NDM1</td>
<td>GAGA</td>
<td>2</td>
<td>-100 to -81</td>
<td>6</td>
<td>11</td>
<td>44-47</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td><strong>CGCTCC</strong></td>
<td>CGMYGYCR</td>
<td>NDM2</td>
<td>3</td>
<td>-80 to -61</td>
<td>6</td>
<td>3</td>
<td>45-47</td>
<td>424</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GAAAGCT</strong></td>
<td>GAAAGCT</td>
<td>NDM3</td>
<td>2</td>
<td>-60 to -41</td>
<td>9</td>
<td>5</td>
<td>44-47</td>
<td>215</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ATCGATA</strong></td>
<td>ATCGATA</td>
<td>NDM4</td>
<td>DRE</td>
<td>2</td>
<td>48</td>
<td>-60 to -41</td>
<td>13</td>
<td>12</td>
<td>45-51</td>
<td>1,593</td>
</tr>
<tr>
<td><strong>CAGCT</strong></td>
<td>CAGCTSW</td>
<td>NDM5</td>
<td>E-box</td>
<td>5</td>
<td>5</td>
<td>-20 to -1</td>
<td>10</td>
<td>9</td>
<td>46-52</td>
<td>1,184</td>
</tr>
</tbody>
</table>

FitzGerald et al., Genome Biol 2006
More data does not equal good data

Berendzen et al., BMC Bioinformatics 2006
Key points

- Core promoters are quite variable
  - Diverse set of core promoter modules
    - New (fly) core promoter elements: MTE, DRE, M1/6
    - Scenario I: Specific enhancer/TF interactions; tissue-specific regulation
    - Scenario II: Alternative options, no functional correlation

- Computational *Drosophila* promoter recognition currently most accurate
  - Models of core promoter subclasses improve success of computational strategies
  - Mammalian promoters lack most of these motifs; instead, CpG islands dominate

- Conservation/alternative TSSs
Evolution of regulatory regions

- A popular area: comparative analysis of regulatory regions
- Current Problem: accurate evolutionary models for non-coding sequences
- Many comparative genomics algorithms involving TF binding sites assume perfect alignments
  - But: How do we know how well our algorithms deal with TF evolution?
  - How often do alignment/motif finding programs lead to a comprehensive picture?
- -> Simulate complex regulatory regions to evaluate/design (new) algorithms
This is really not new...

- Has been done quite extensively
- Key assumption: TFBS are islands of conservation within larger not-so-conserved region -> use two sets of rates [Pollard et al., BMC Genomics 2006]
  - What about turnover events?
- Instead: Model evolution with one rate, but subject to constraints
  - Assuming neutral evolution/stabilizing selection – which other sequences are possible?
- Bad stuff upfront:
  - Ignores trans-factor and adaptive evolution
  - Ignores population genetics
The framework

- Simulate 1,000 ancestor sequences
  - 3rd order background, human upstream sequences
- Evolve each one 1,000 times
  - Get a distribution of features in the evolved set

<table>
<thead>
<tr>
<th>Sequence length</th>
<th>250 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution Model</td>
<td>HKY85</td>
</tr>
<tr>
<td>Transition: Transversion</td>
<td>20:1</td>
</tr>
<tr>
<td>Point substitutions : insertion/deletion</td>
<td>10:1</td>
</tr>
<tr>
<td>InDel length model</td>
<td>Geometric (p=0.5)</td>
</tr>
</tbody>
</table>
A simple example

Set of constraints:
- This is the difference to related efforts, e.g. Pollard et al. 2006

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GC content</td>
<td>45%-55%</td>
</tr>
<tr>
<td>Number of E2F sites</td>
<td>1</td>
</tr>
<tr>
<td>E2F location relative to TSS</td>
<td>[-50, -100]</td>
</tr>
<tr>
<td>DNA strand of E2F site</td>
<td>+</td>
</tr>
<tr>
<td>Cutoff threshold of E2F site</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Not thought to be a precise model
Rather, to get some idea how frequent
- current alignment algorithms work
- more complex turnover events may happen
Results: E2F site turnover

0.1 substitutions/site

0.5 substitutions/site
Turnover at various distances

Poisson distribution for # turnovers:
\[ Pr(N>0) = 1 - Pr(N=0) = 1 - \text{Exp}(-\lambda t); \quad \lambda \sim 0.08 \]

Simulated starting set
E2F promoters as starting set
Evolving along two branches

- Now: distance fixed for human/mouse
- Free parameter shown: spacer E2F/TSS
- Prob. for turnover in both species
## Pair of E2F/myc

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F location relative to TSS</td>
<td>[ -50, -100 ]</td>
</tr>
<tr>
<td>Myc location relative to TSS</td>
<td>[ -100, -150 ]</td>
</tr>
<tr>
<td>Copy number of E2F</td>
<td>1</td>
</tr>
<tr>
<td>Copy number of Myc</td>
<td>1</td>
</tr>
<tr>
<td>DNA strand of E2F site</td>
<td>+</td>
</tr>
<tr>
<td>DNA strand of Myc site</td>
<td>+</td>
</tr>
<tr>
<td>Additional space constraint</td>
<td></td>
</tr>
<tr>
<td>between Myc and E2F sites</td>
<td>[ 50, 60]</td>
</tr>
</tbody>
</table>
No spatial constraint

Both sites

One site
With spatial constraint

Both sites

One site
Evaluate alignment accuracy

- Simulate evolution over various scaled trees
- Once simulated, run global multiple aligners
  - mlagan, mavid, muscle, dialign, clustalw
- We can then trace back which sites did not turn over and should be aligned
  - Neutral evolution -> we know all sites are there
- We are nice (of course :)
  - Turnover, but no change in order of sites
  - Accuracy: averaged over pairwise alignments
## Mammalian sequences

<table>
<thead>
<tr>
<th>Name</th>
<th>Accession#</th>
<th>Len</th>
<th>Strand</th>
<th>Location (min, max)</th>
<th>Copy # (min, max)</th>
<th>Cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>YY1E2F</td>
<td>MA0095 (YY1)</td>
<td>13</td>
<td>+</td>
<td>(20, 30)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>MA0024 (E2F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pax6</td>
<td>MA0069</td>
<td>14</td>
<td>+</td>
<td>(50, 70)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
<tr>
<td>TP53</td>
<td>MA0106</td>
<td>20</td>
<td>+</td>
<td>(360, 400)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
<tr>
<td>IRF2</td>
<td>MA0051</td>
<td>18</td>
<td>+</td>
<td>(420, 480)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
<tr>
<td>PPARG</td>
<td>MA0066</td>
<td>20</td>
<td>+</td>
<td>(2000, 2080)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
<tr>
<td>ROAZ</td>
<td>MA0116</td>
<td>15</td>
<td>+</td>
<td>(2100, 2200)</td>
<td>(1, 1)</td>
<td>0.90</td>
</tr>
</tbody>
</table>

![Gene tree](image-url)
Accuracy w/increasing #species
Accuracy for individual factors
There is an open issue with aligning non-coding sequences
  • Current aligners do not scale well with increasing number of species
  • Alignment accuracy suffers
  • Assessing site turnover may be lost in the noise
    [Pollard et al., 2006; Moses et al., 2006]

Developed a general tool to simulate non-coding evolution
  • Based on constraints and not on a different evolutionary model
  • Next step: TSS evolution
Thanks to...

Motifs
Elizabeth Rach
Weichun Huang
Bill Majoros

UC Berkeley/BDGP
Gerry Rubin
Martin Reese
Guo-chun Liao
Josh Kaminker

UC San Diego
Jim Kadonaga Lab

Images
Dan Mace

Phil Benfey Lab
Ji-Young Lee
Todd Twigg
Juliette Colinas

Rob Clark Lab

Alfred P Sloan Foundation, NSF
Computational Biology & Bioinformatics Program
http://tools.genome.duke.edu/generegulation