

Core Promoter Analysis

Uwe Ohler

Institute for Genome Sciences and Policy

Duke University

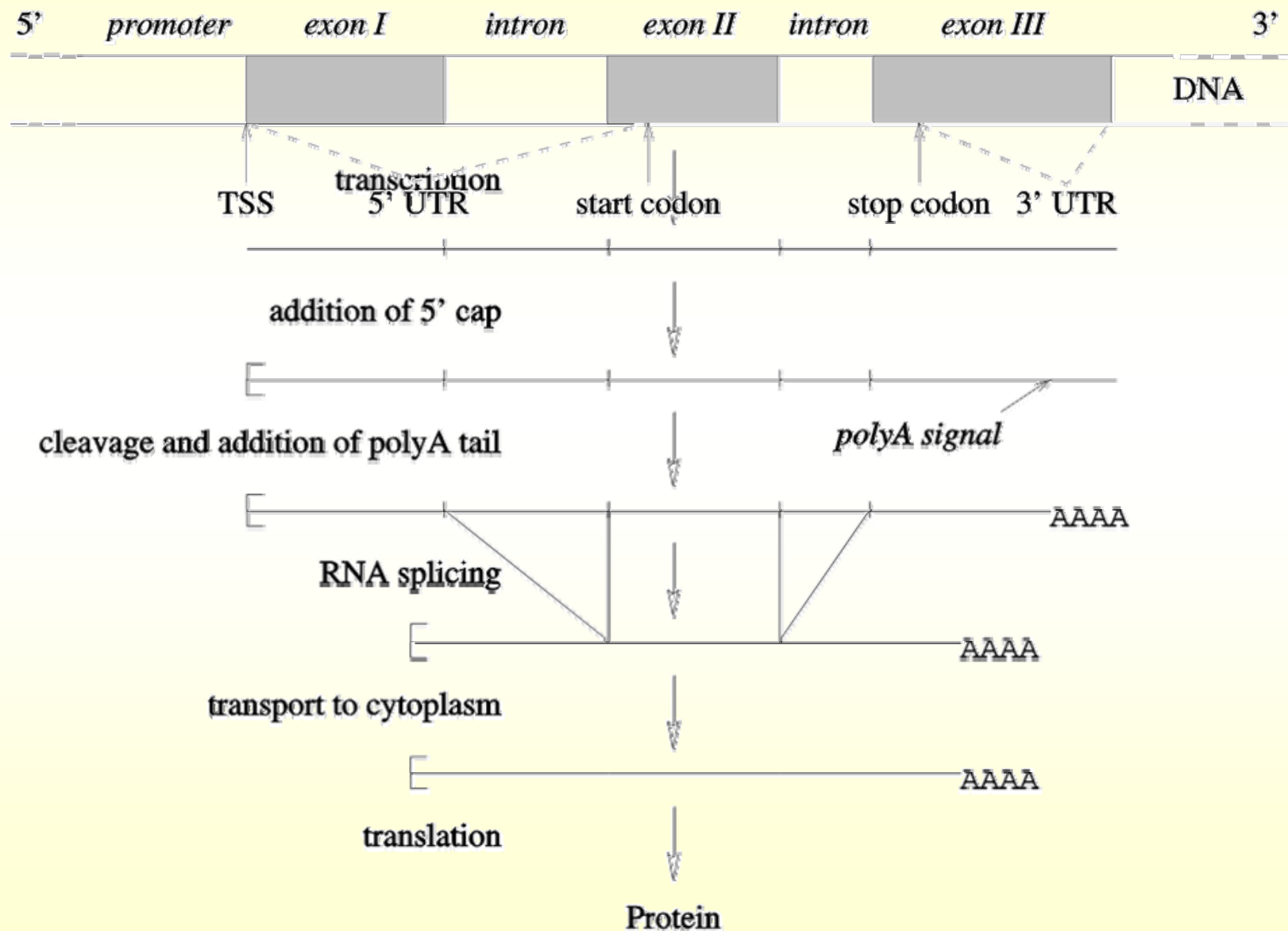
uwe.ohler@duke.edu

Computational Biology of Gene Regulation

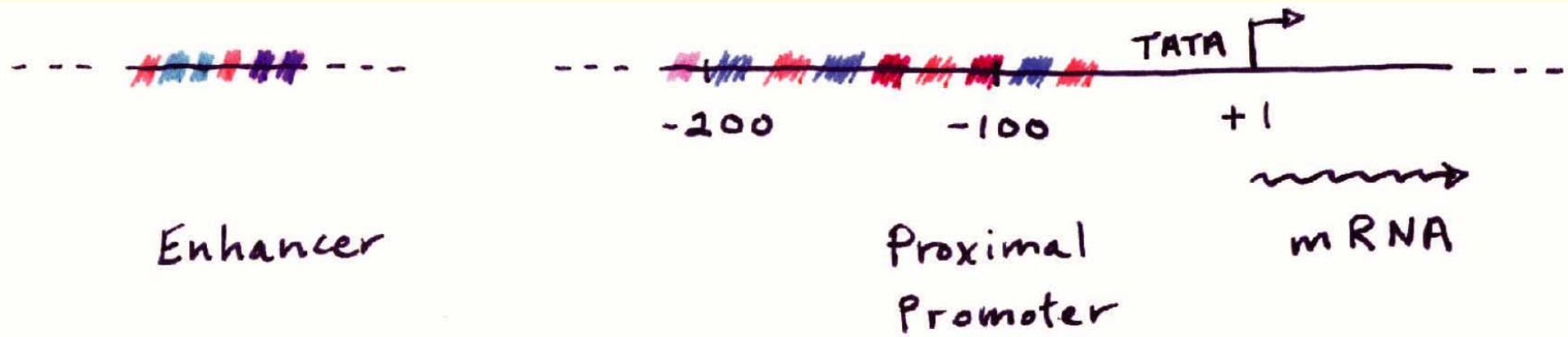
Focus of my group:

- 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

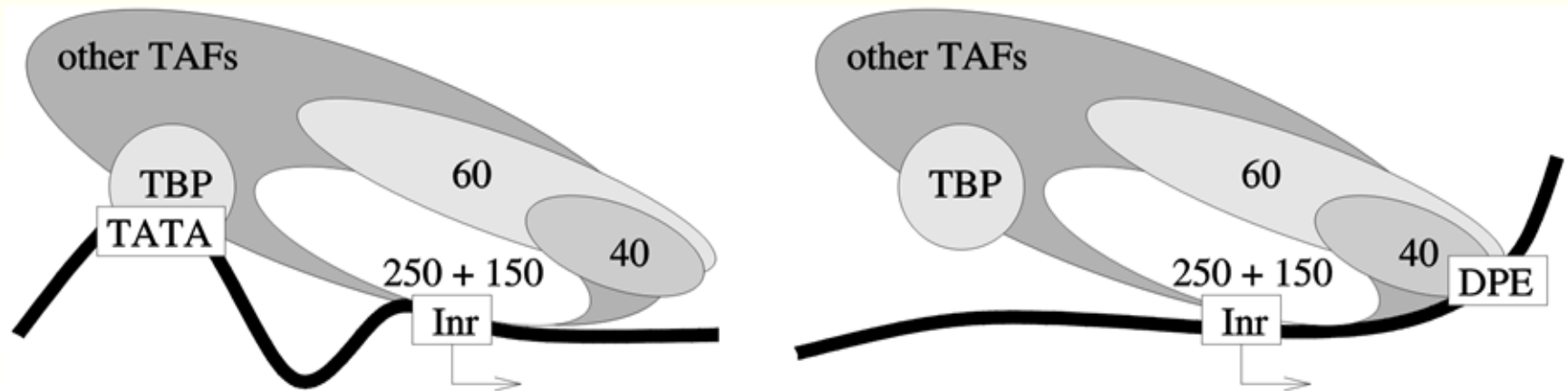


An extremely simplified view of eukaryotic transcription



- 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“)



AAACCGTAAACACAGAGCAGGCGAGCGTAAGCAAGAGAGAGGTGAAGCCAGAGGCGGAGGCGCAAGA
CGTGCTGCCTCCCAATAAACCCGGTGCAGTGAGTCAGTGTGTTGTGTGCCCCAGTCGCGAGCGGACGATC

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

Species specific differences

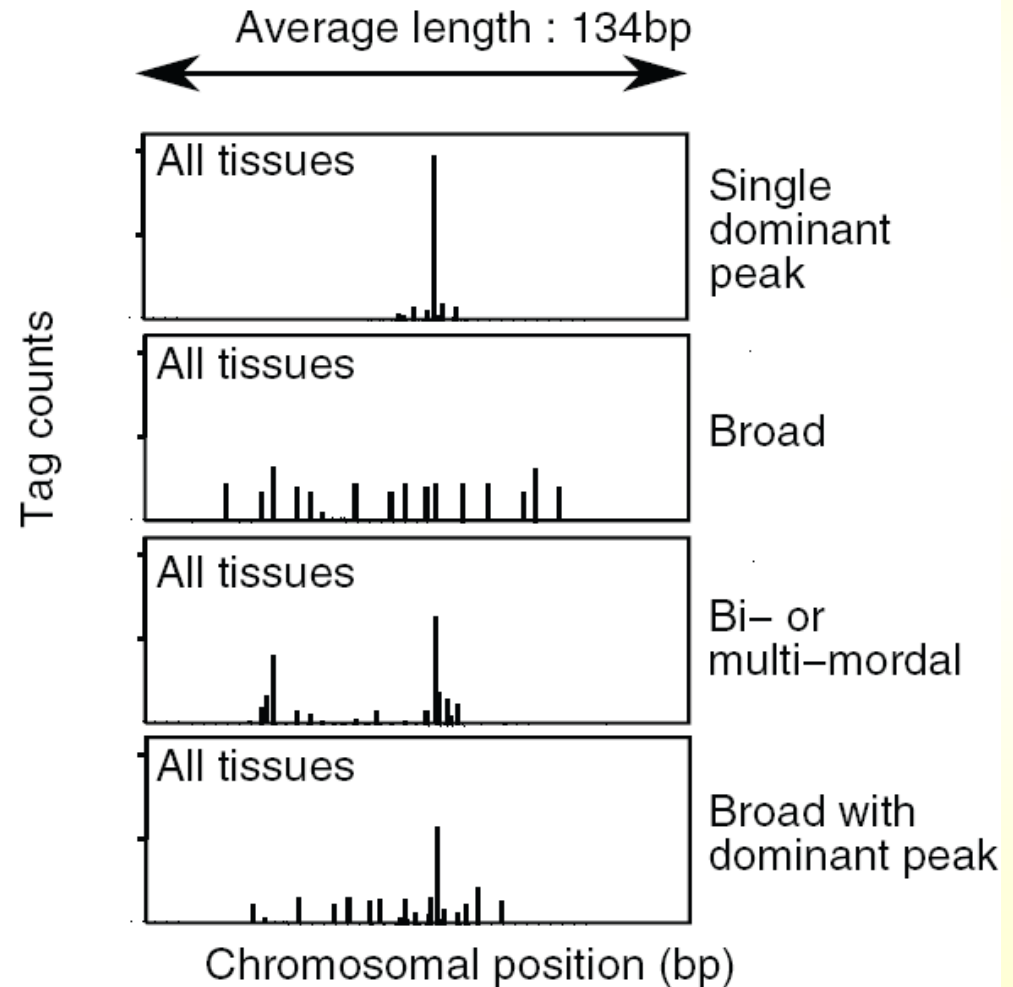
- The core protein complex is conserved, but the cis-regulatory sequences are not (*quite*)
- Example: TATA box
 - Ca 80 nt upstream in yeast, 25 nt in other eukaryotes
- Example: Initiator
 - A strong 5-6 nt motif in flies, a weak 2 nt preference in mammals
- Example: CpG islands
 - A mammalian phenomenon 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-throughput 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%

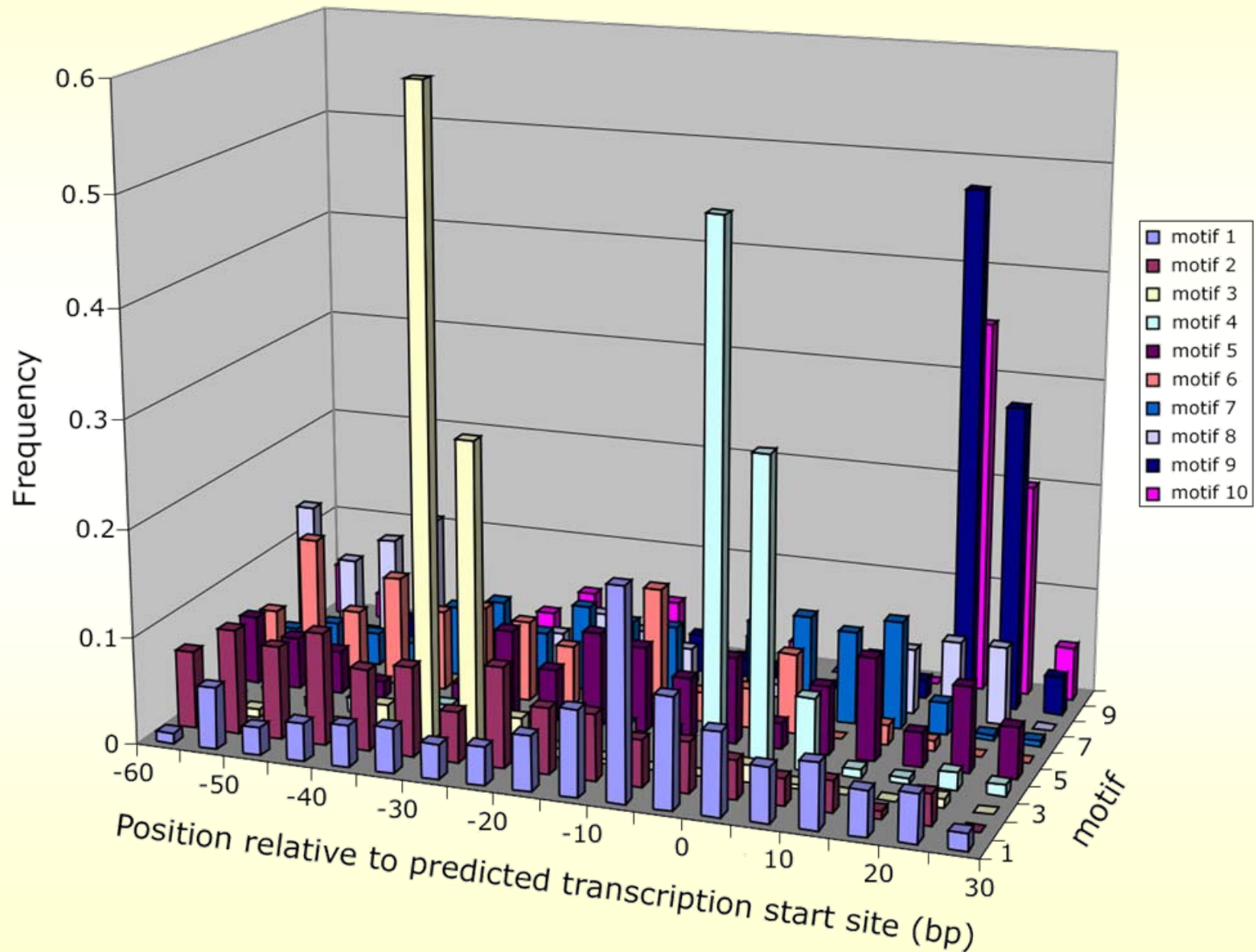
Motifs found in core promoters

| Motif | Pictogram | Consensus | # seq | E value |
|-----------|---|-------------|-------|--------------|
| 1 |  | YGGTCACACTR | 311 | 5.1 e-415 |
| 2 DRE |  | WATCGATW | 277 | 1.7 e-183 |
| 3 TATA |  | STATAWAAR | 251 | 2.1 e-138 |
| 4 INR |  | TCAGTYKNNNT | 369 | 3.4 e-117 |
| 5 Ebox |  | AWCAGCTGWT | 125 | 2.9 e-93 |

Motifs found in core promoters

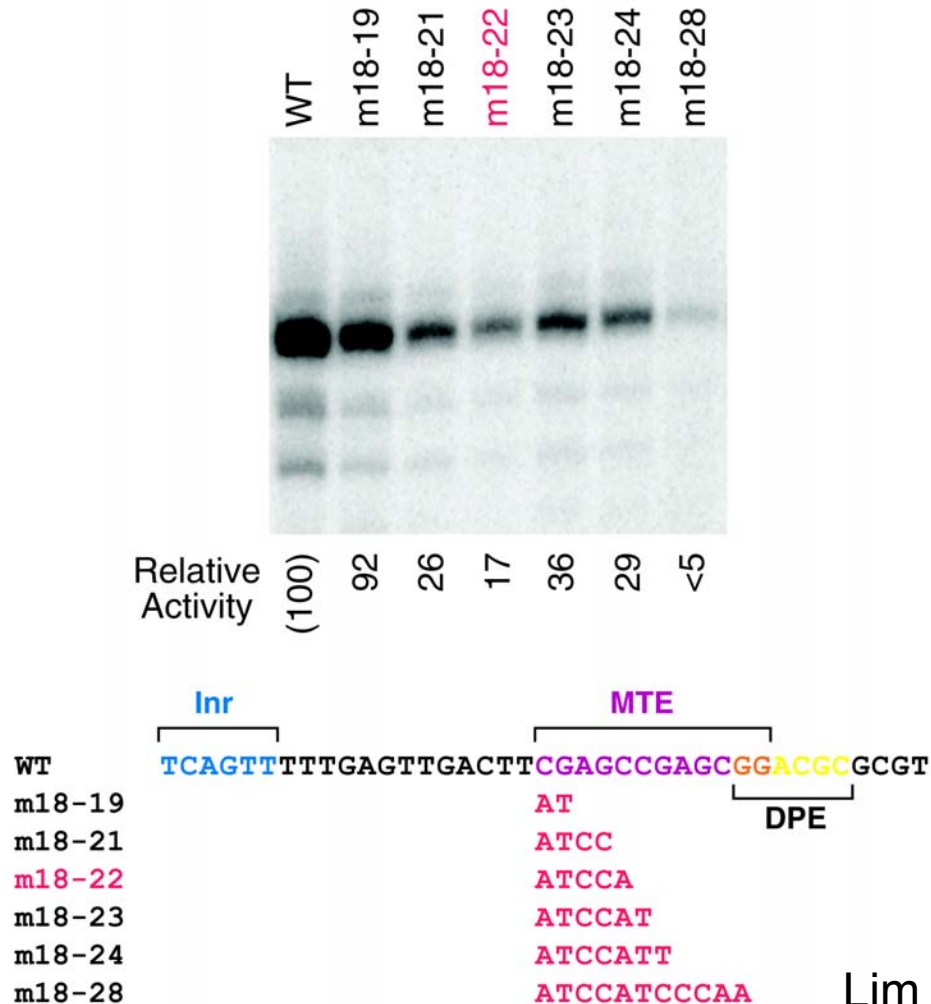
| | | | | |
|-----------|---|---------------------|-----|-------------|
| 6 |  | KTYRGTATWTTT | 107 | 1.9 e-62 |
| 7 |  | KNNCAKCNCTR | 197 | 1.9 e-63 |
| 8 |  | YGGCARCGRSYSS | 82 | 5.1 e-29 |
| 9 DPE |  | CRWMGCGWKCG GTTS | 56 | 1.9 e-12 |
| 10 MTE |  | CSARCSSAACGS | 40 | 8.3 e-9 |

Positional distribution of motifs



Validation/definition of MTE

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



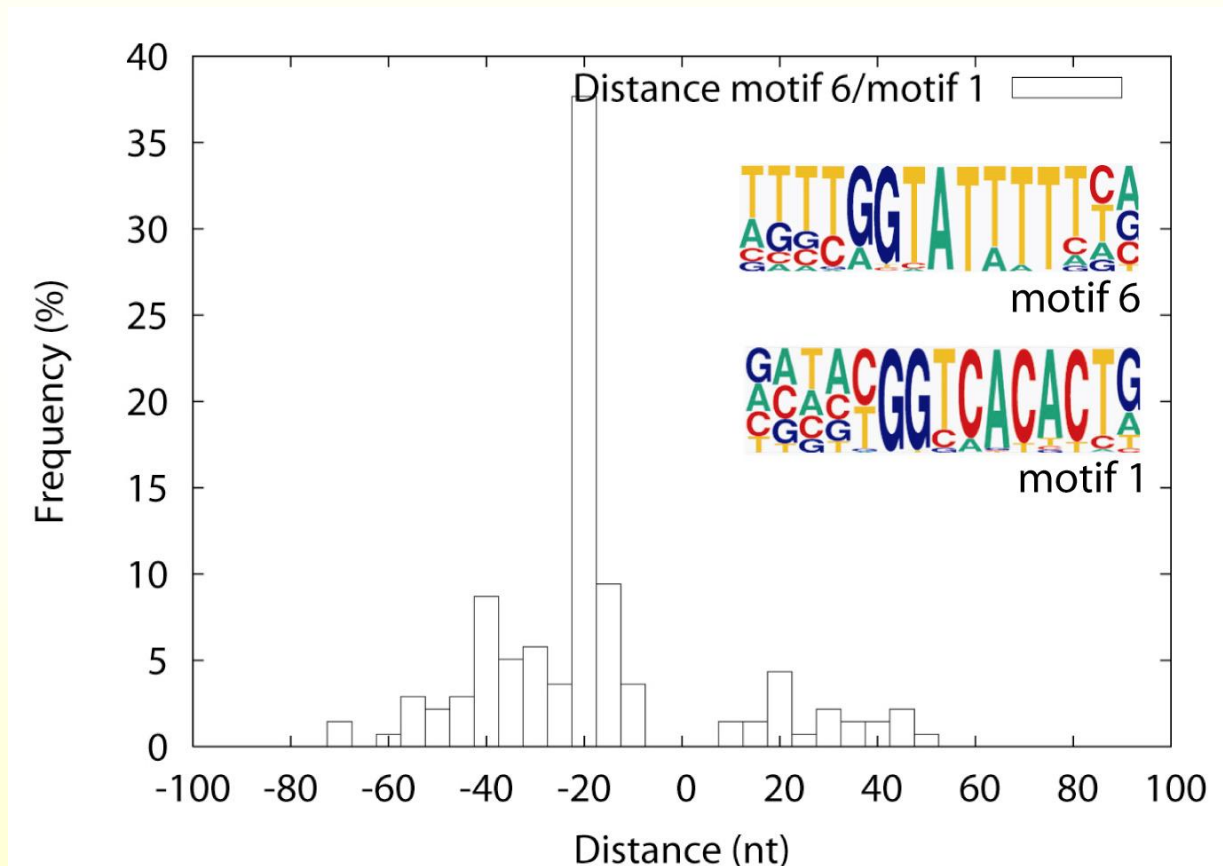
Lim et al., *Genes Dev* 2004

Frequency of co-occurrence

| Motif X | % seqs w/ X | % seqs with Motif X also containing Motif below | | | | | | |
|-------------|-------------------|---|-------|-------------|-------------|-------------|-------------|-------------|
| | | M1 | DRE | TATA | INR | M6 | DPE | MTE |
| M1 | 25.1 | 100.0 | 21.3 | 13.1 | 12.7 | 28.3 | 4.9 | 6.1 |
| DRE | 26.0 | 20.6 | 100.0 | 14.9 | 16.8 | 14.1 | 5.7 | 6.9 |
| TATA | 19.3 | 17.1 | 20.1 | 100.0 | 28.9 | 14.4 | 4.8 | 9.4 |
| INR | 26.3 | 12.1 | 16.6 | 21.1 | 100.0 | 12.1 | 14.9 | 12.9 |
| M6 | 15.8 | 45.1 | 23.2 | 17.6 | 20.3 | 100.0 | 4.6 | 4.2 |
| DPE | 7.9 | 15.6 | 18.8 | 11.7 | 49.4 | 9.1 | 100.0 | 8.4 |
| MTE | 8.5 | 18.2 | 21.2 | 21.2 | 40.0 | 7.9 | 7.9 | 100.0 |

A new core promoter module

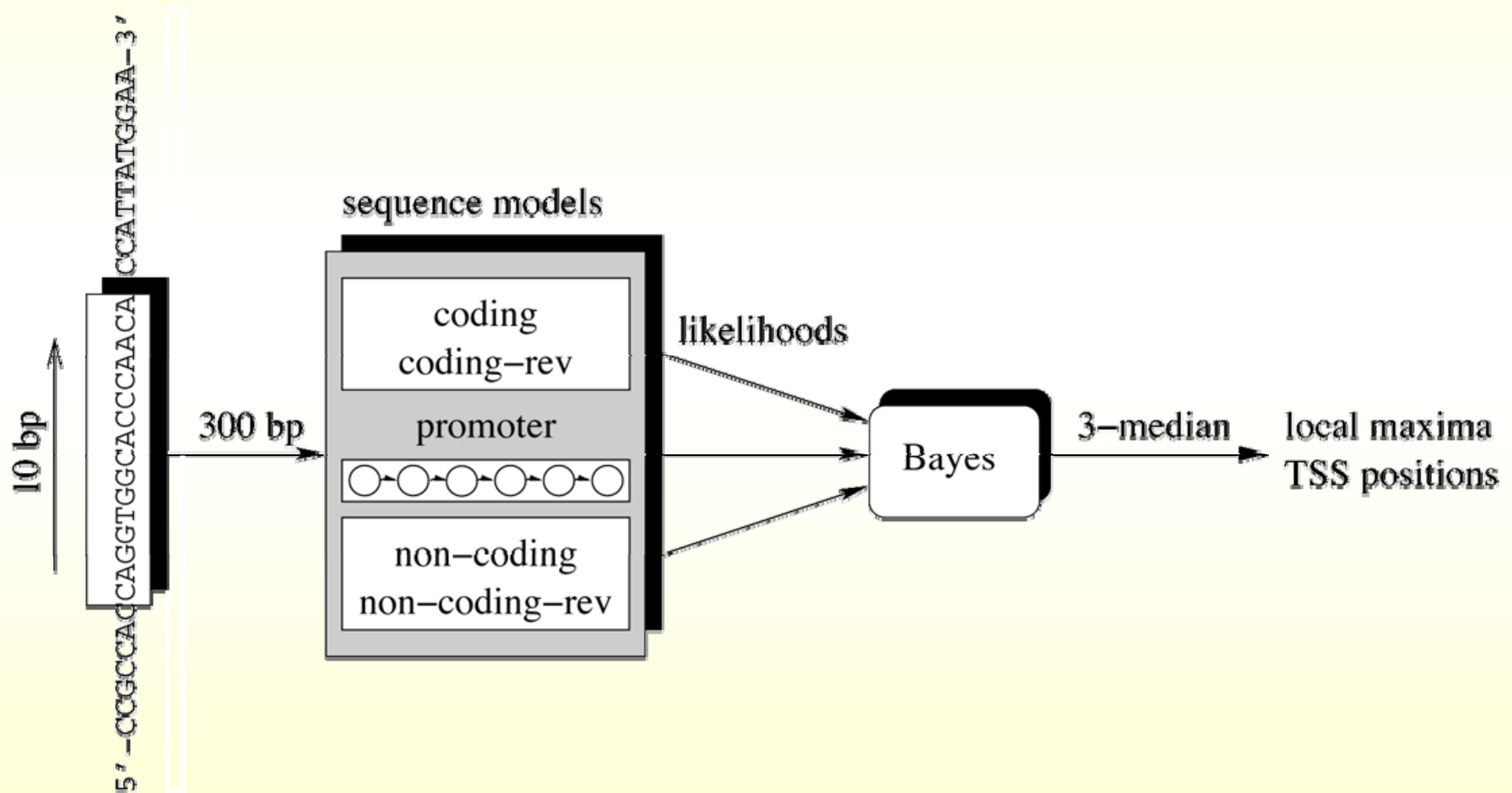
- Motif 1 has a weak preference for location at the TSS
- The motif 6/1 pair is reminiscent of the TATA/Inr module



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



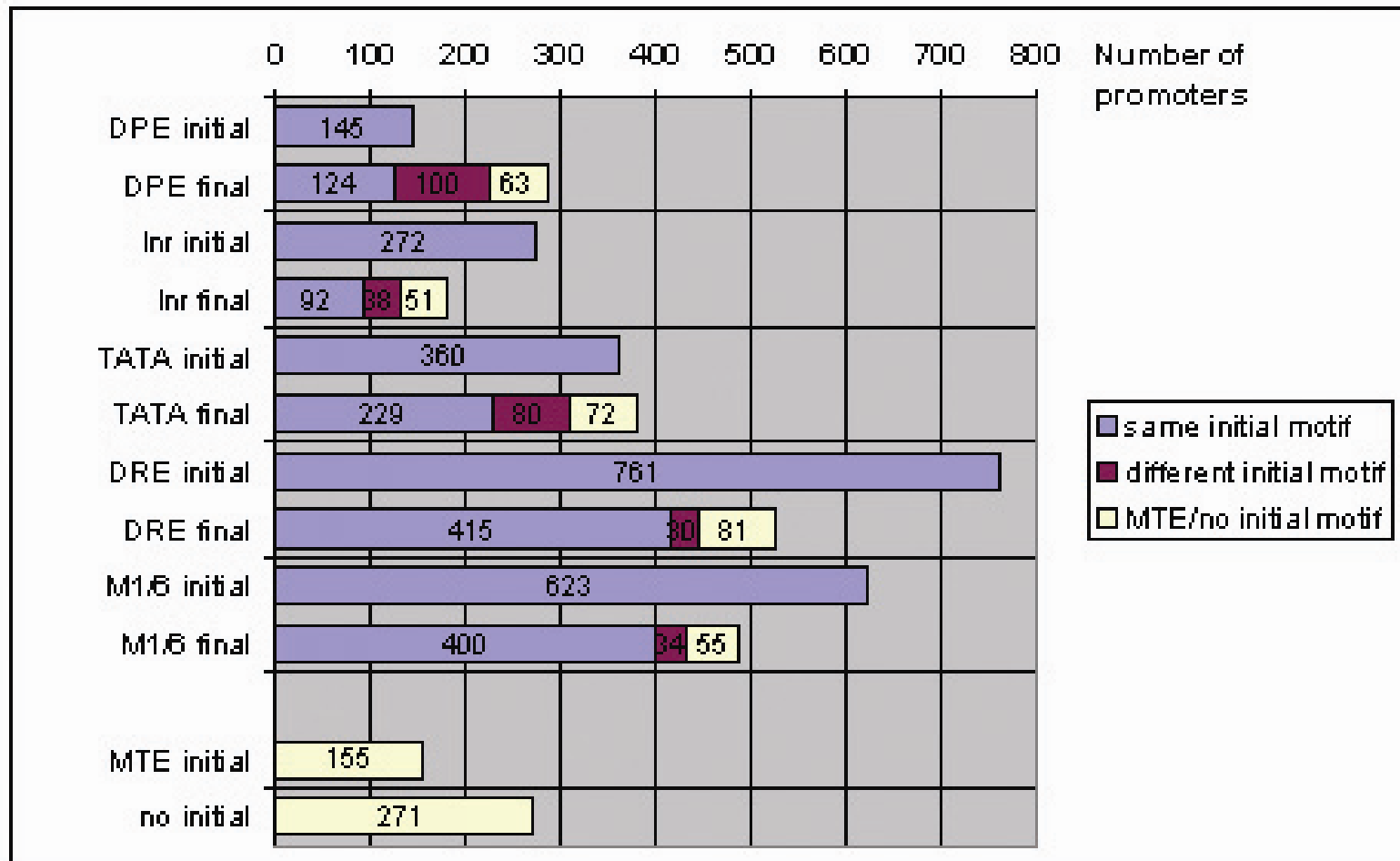
Computational approaches

- 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

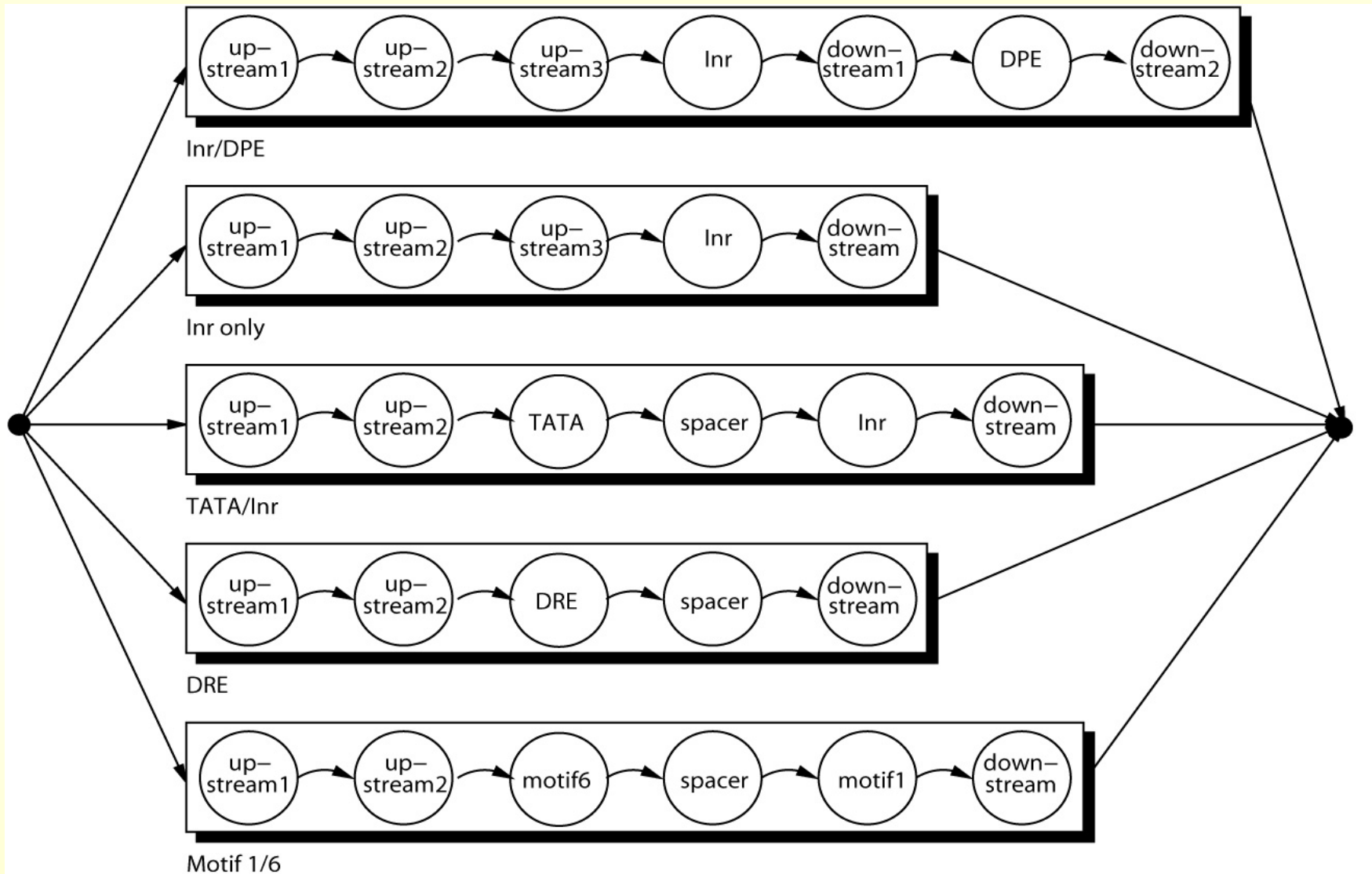
Modeling promoter subclasses

- 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)

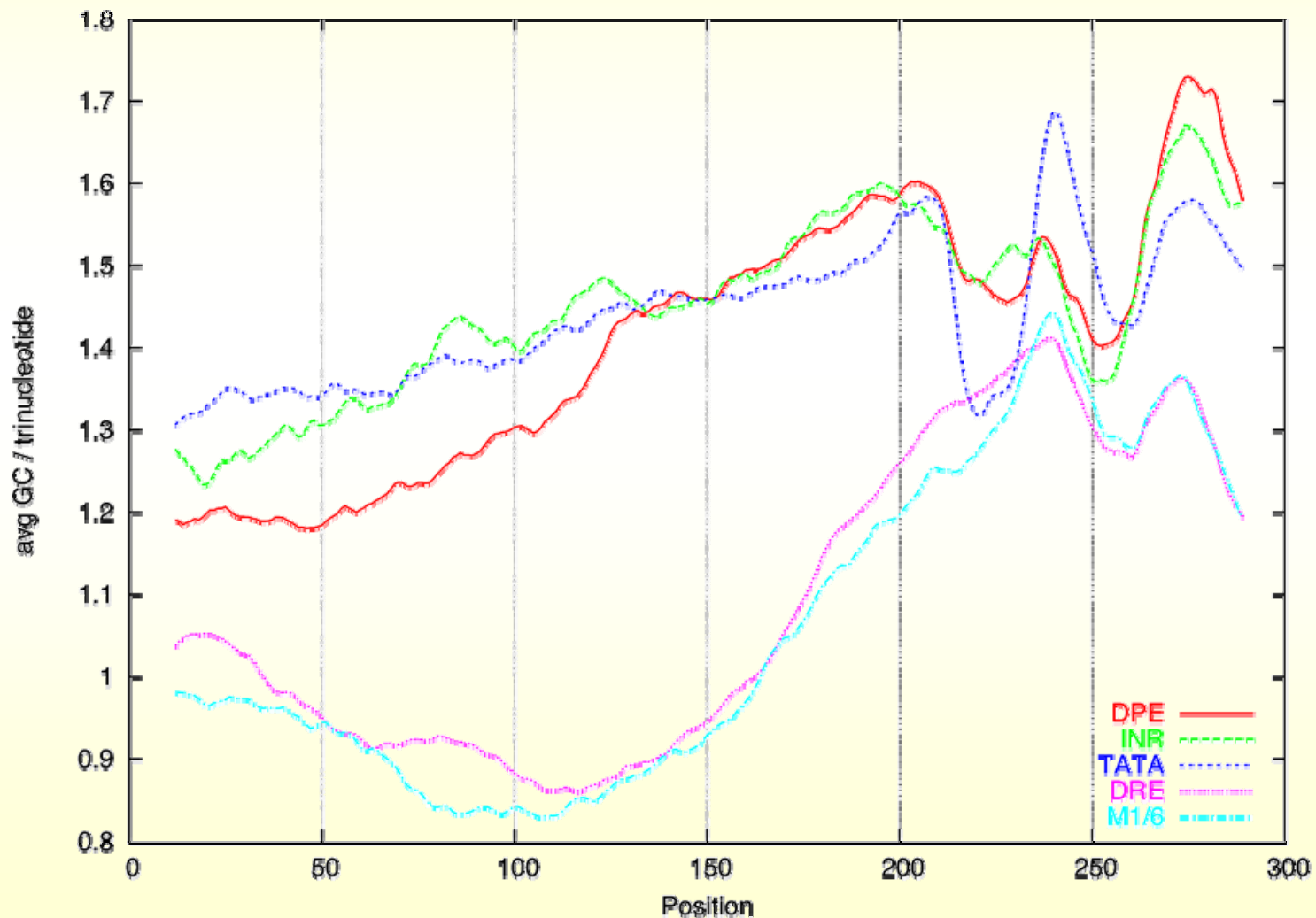
Clustering of core promoters



Modeling promoter subclasses



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)

| McPromoter 2002 (one model) | | Sharan & Myers 2005 | | McPromoter 2006 (five models) | | |
|--------------------------------|-----------|------------------------|-----------|----------------------------------|-----------|----------------|
| <i>Sn</i> | <i>Sp</i> | <i>Sn</i> | <i>Sp</i> | <i>Sn</i> | <i>Sp</i> | <i>AP rate</i> |
| 20 | 69 | 20 | 79 | 23 | 91 | 1/426,590 |
| 37 | 51 | 35 | 53 | 36 | 79 | 1/94,797 |
| 52 | 40 | 50 | 33 | 50 | 47 | 1/16,097 |
| 67 | 29 | 65 | 20 | 64 | 36 | 1/8,203 |

Alternative transcription start sites

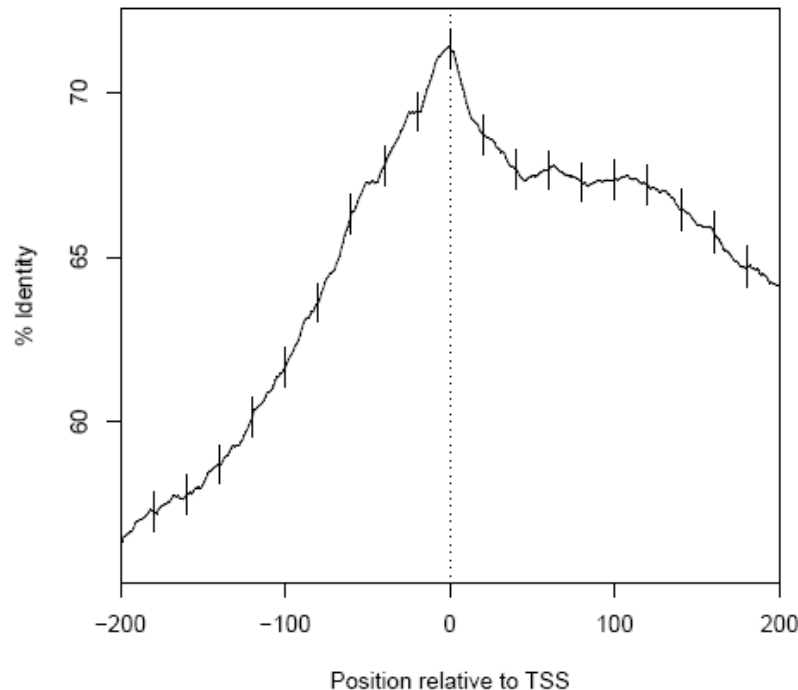
- A large fraction of genes has more than one TSS
- Here, we mean distinctly separate TSS (~100 nt or more apart, not small scale fluctuation)
 - Alternative 5' UTRs
 - Alternative translation start sites
 - Tissue-specific promoters
- Prominent example: e.g. protocadherin genes

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

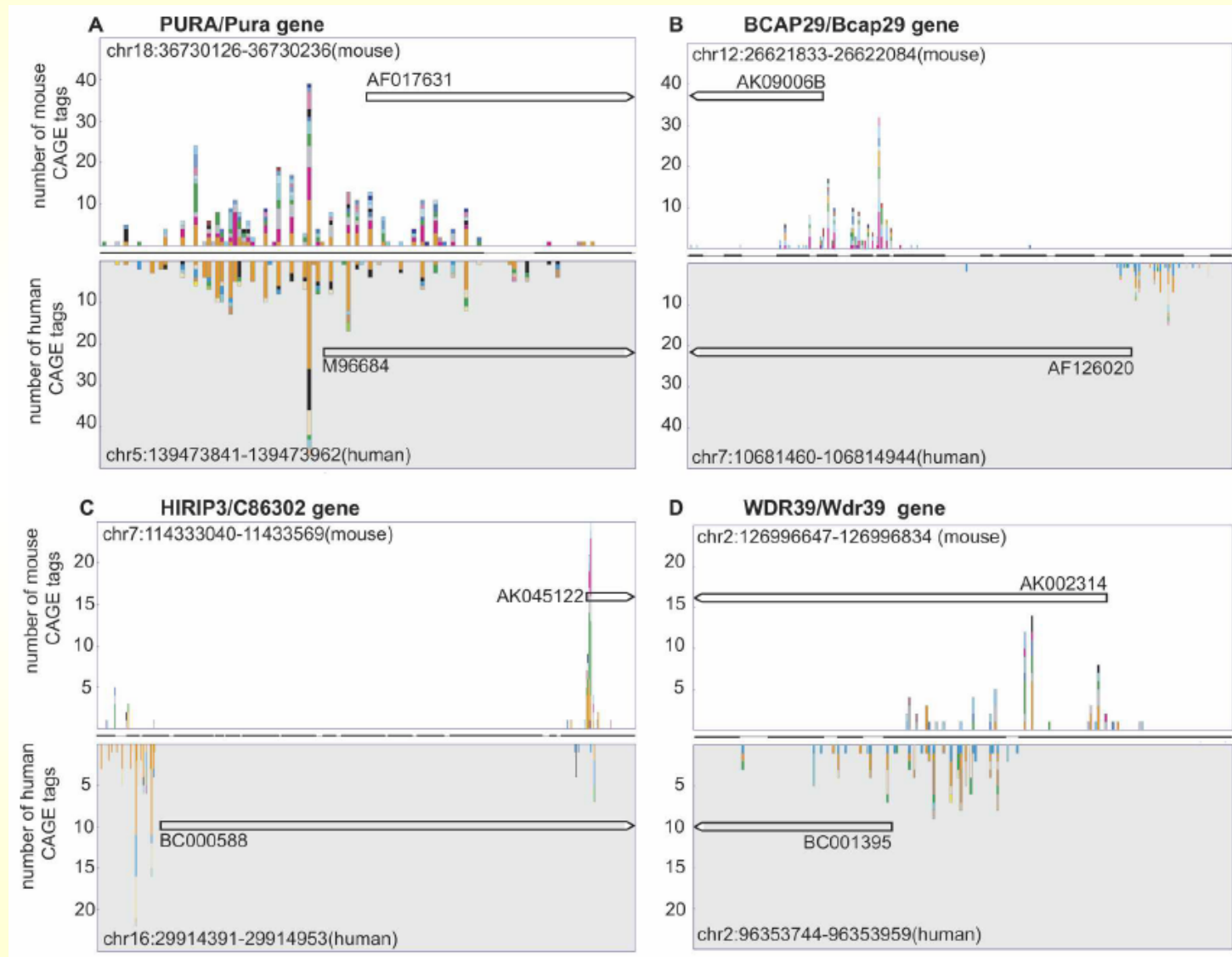
Scenario I: Conservation

- 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



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)]

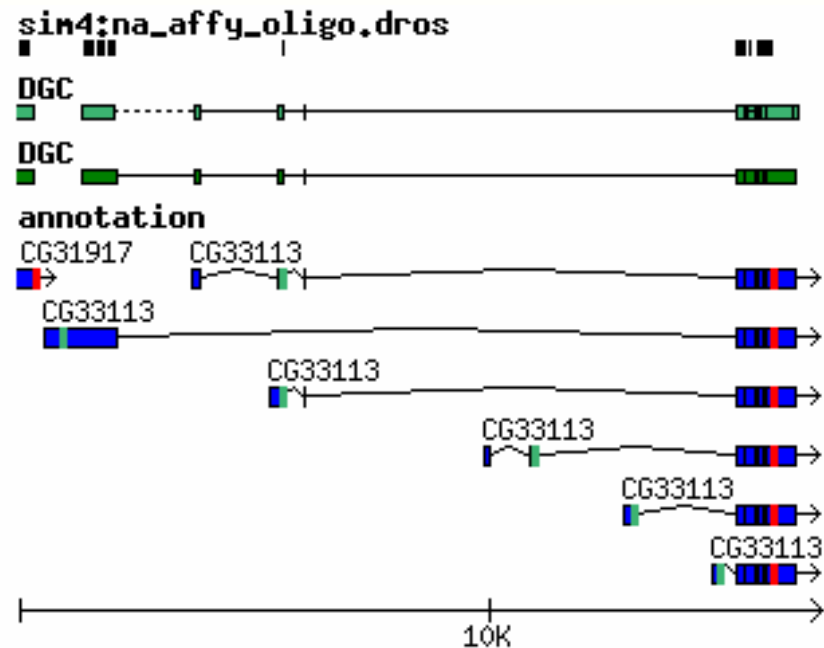
Number_of_tags_from_AT_ADULT_TESTES [(0)(0)(4)(0)]

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



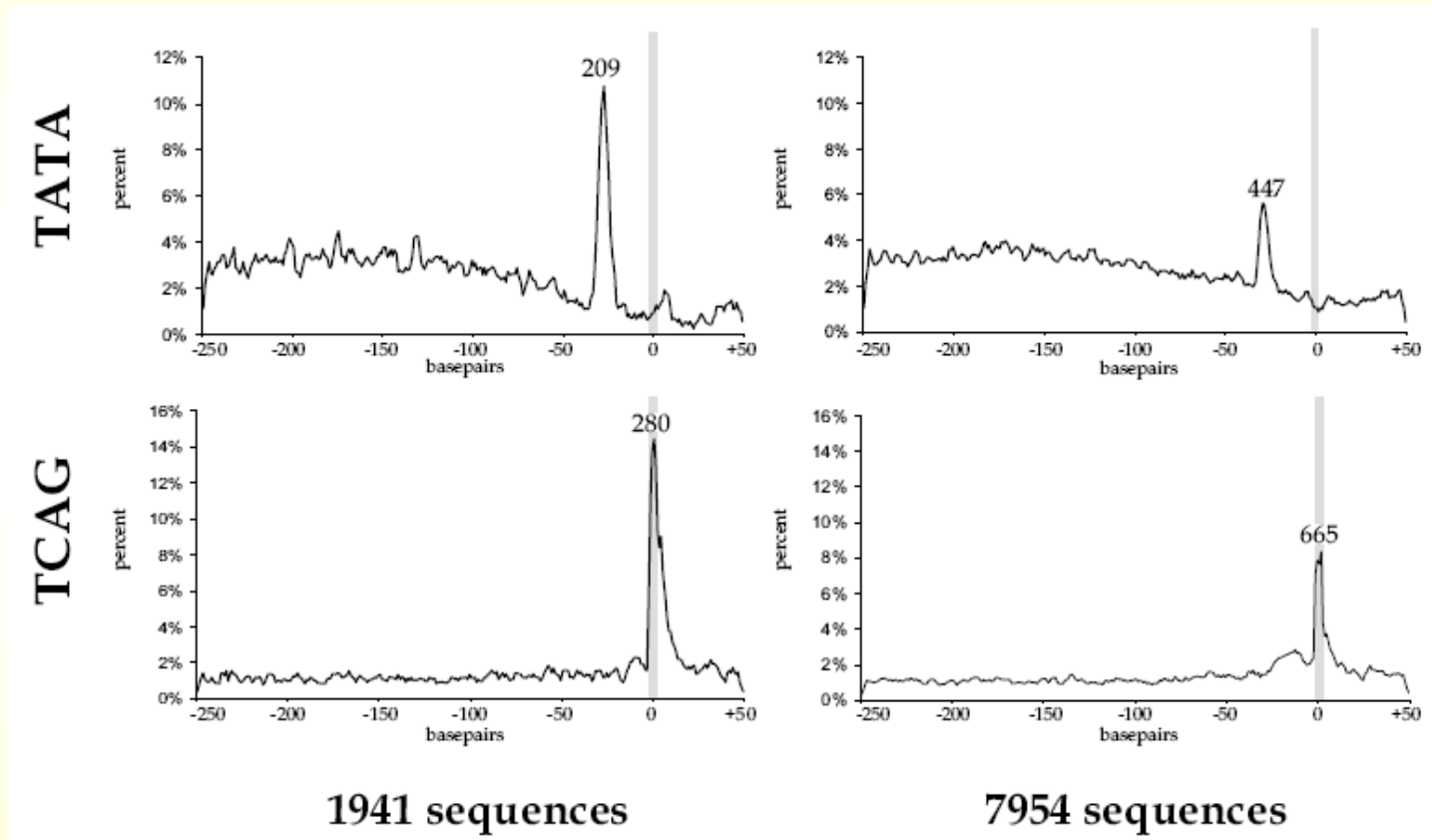
Related work

| Sequence logo | Consensus sequence | Name | Common name | Ohler # | 8-mers in consensus | Peak bps from TSS | CF+ | CF- | Pooled peaks | Unique genes |
|---|--------------------|------|-------------|---------|---------------------|-------------------|-----|-----|--------------|--------------|
|  | STATAAA | DMp1 | TATA | 3 | 30 | -32 | 24 | 2 | 48-49 | 511 |
|  | TCAGTY | DMp2 | INR | 4 | 101 | -2 | 29 | 2 | 49-51 | 1,501 |
|  | TCATTCG | DMp3 | INR1 | | 5 | -2 | 15 | 3 | 50-51 | 113 |
|  | KCGGTTSK | DMp4 | DPE | 9 | 10 | +25 | 14 | 4 | 51-52 | 147 |
|  | CGGACGT | DMp5 | DPE1 | | 11 | +26 | 18 | 3 | 51-52 | 80 |
|  | CARCCCT | DMv1 | | | 5 | -60 to -41 | 11 | 5 | 47-51 | 311 |
|  | TGGYAACR | DMv2 | | 8 | 11 | -20 to -1 | 13 | 5 | 46-51 | 311 |
|  | CAYCNCTA | DMv3 | | 7 | 11 | +1 to +20 | 18 | 4 | 46-52 | 604 |
|  | GGYCACAC | DMv4 | | 1 | 42 | -20 to -1 | 23 | 7 | 46-51 | 649 |
|  | TGGTATTT | DMv5 | | 6 | 3 | -60 to -41 | 11 | 5 | 45-51 | 287 |

| Sequence logo | Consensus sequence | Name | Common name | Ohler # | 8-mers in consensus | Peak bps from TSS | CF+ | CF- | Pooled peaks | Unique genes |
|---|--------------------|------|-------------|---------|---------------------|-------------------|-----|-----|--------------|--------------|
|  | GAGAGCG | NDM1 | GAGA | | 2 | -100 to -81 | 6 | 11 | 44-47 | 360 |
|  | CGMYGYCR | NDM2 | | | 3 | -80 to -61 | 6 | 3 | 45-47 | 424 |
|  | GAAAGCT | NDM3 | | | 2 | -60 to -41 | 9 | 5 | 44-47 | 215 |
|  | ATCGATA | NDM4 | DRE | 2 | 48 | -60 to -41 | 13 | 12 | 45-51 | 1,593 |
|  | CAGCTSWW | NDM5 | E-box | 5 | 5 | -20 to -1 | 10 | 9 | 46-52 | 1,184 |

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

| | |
|---|-----------------------|
| Sequence length | 250 nt |
| Substitution Model | HKY85 |
| Transition: Transversion | 20:1 |
| Point substitutions : insertion/deletion | 10:1 |
| InDel length model | Geometric ($p=0.5$) |

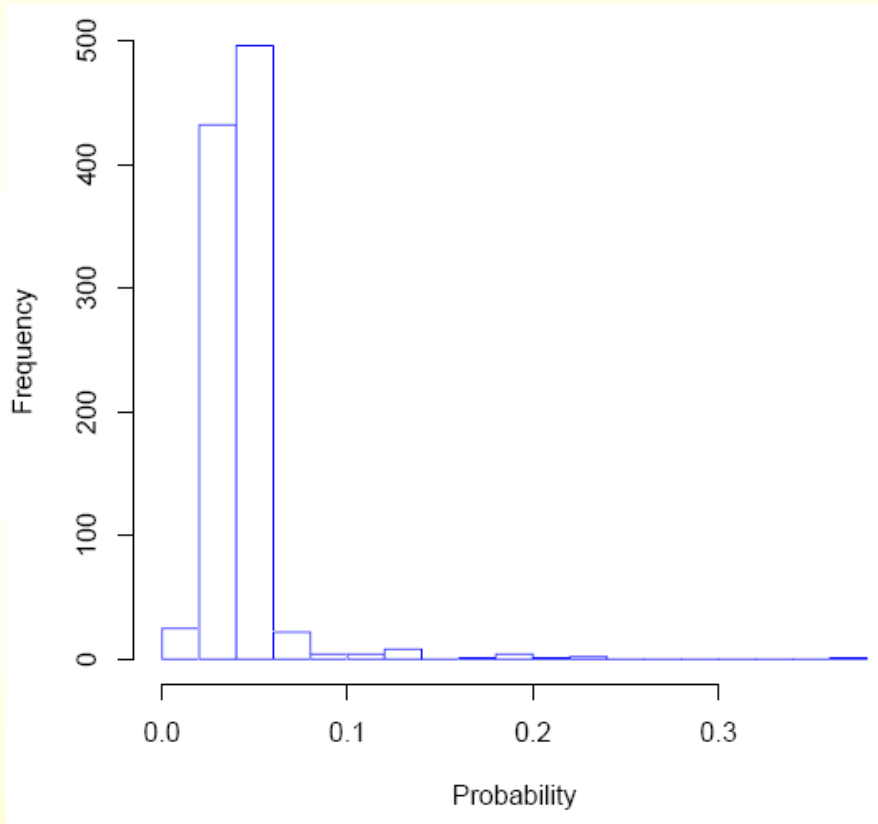
A simple example

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

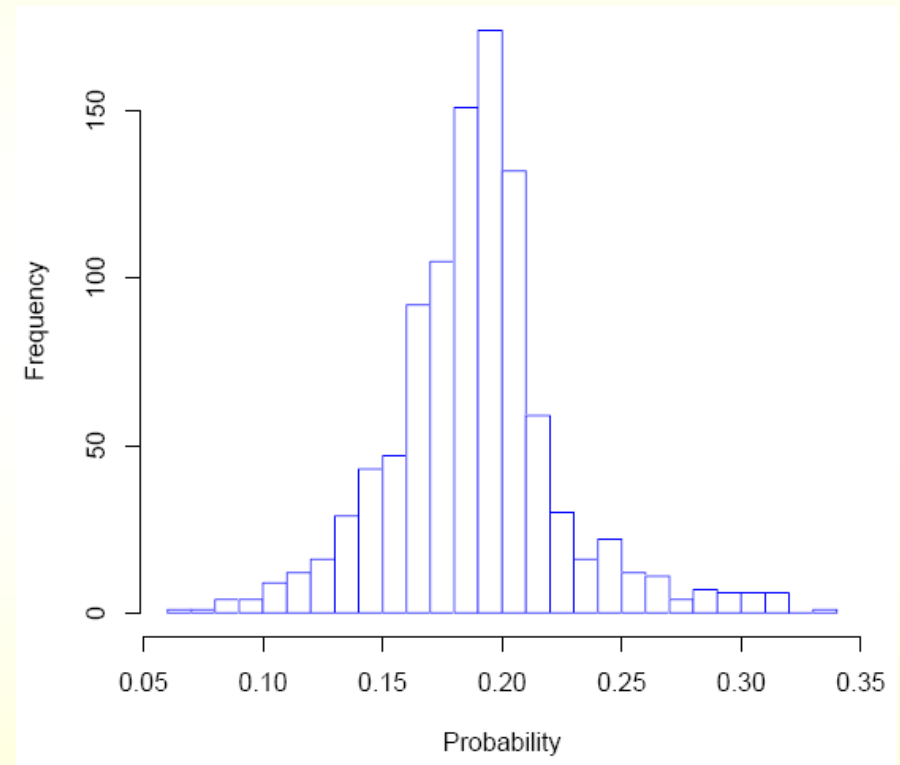
| | |
|------------------------------|-------------|
| GC content | 45%-55% |
| Number of E2F sites | 1 |
| E2F location relative to TSS | [-50, -100] |
| DNA strand of E2F site | + |
| Cutoff threshold of E2F site | 0.90 |

- 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

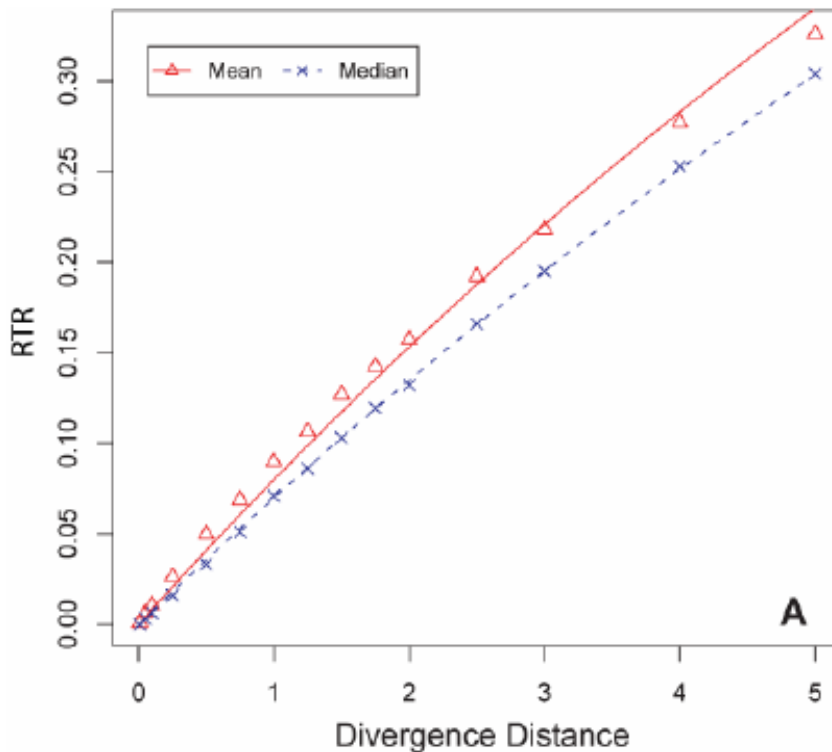


0.5 substitutions/site

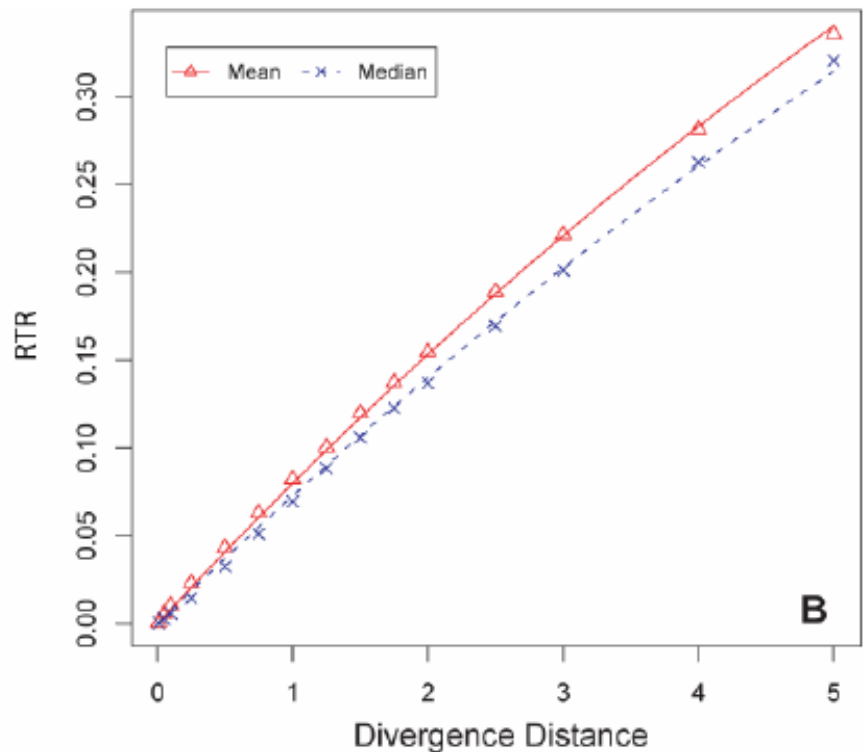
Turnover at various distances

Poisson distribution for # turnovers:

$$Pr(N>0) = 1 - Pr(N=0) = 1 - \text{Exp}(-\lambda t); \quad \lambda \sim 0.08$$



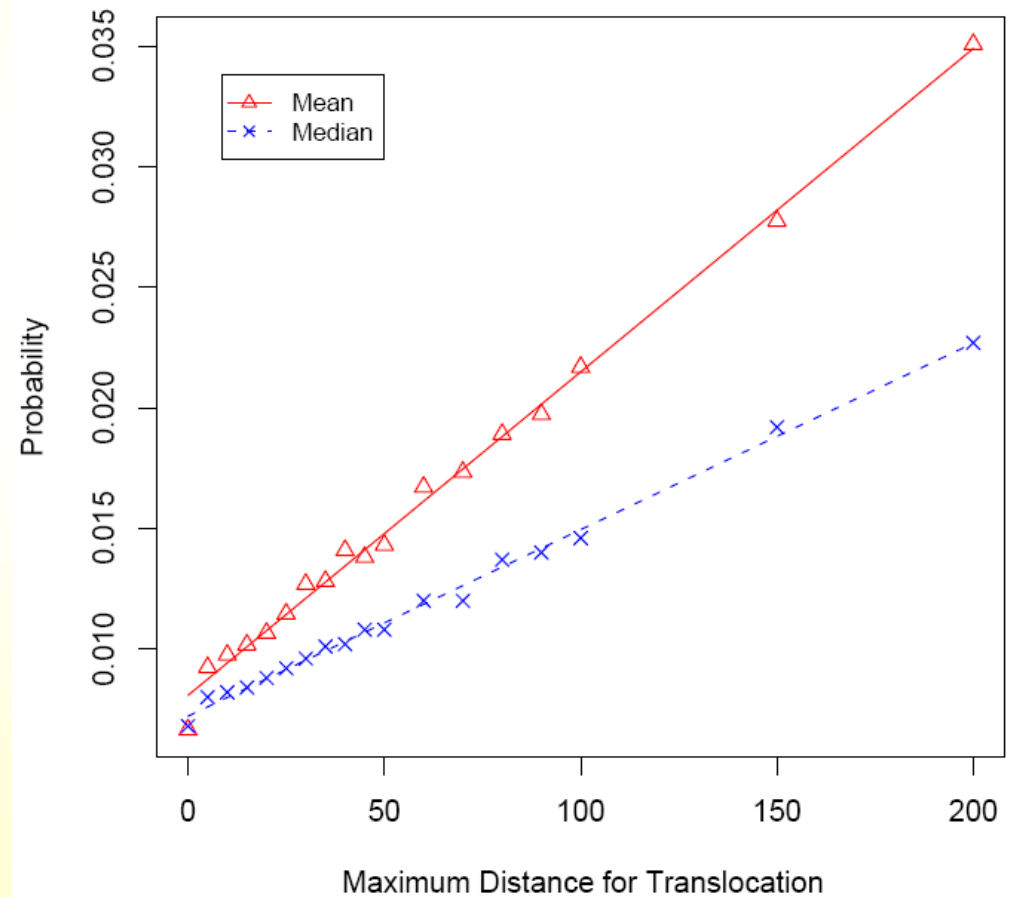
Simulated starting set



E2F promoters as starting set

Evolving along two branches

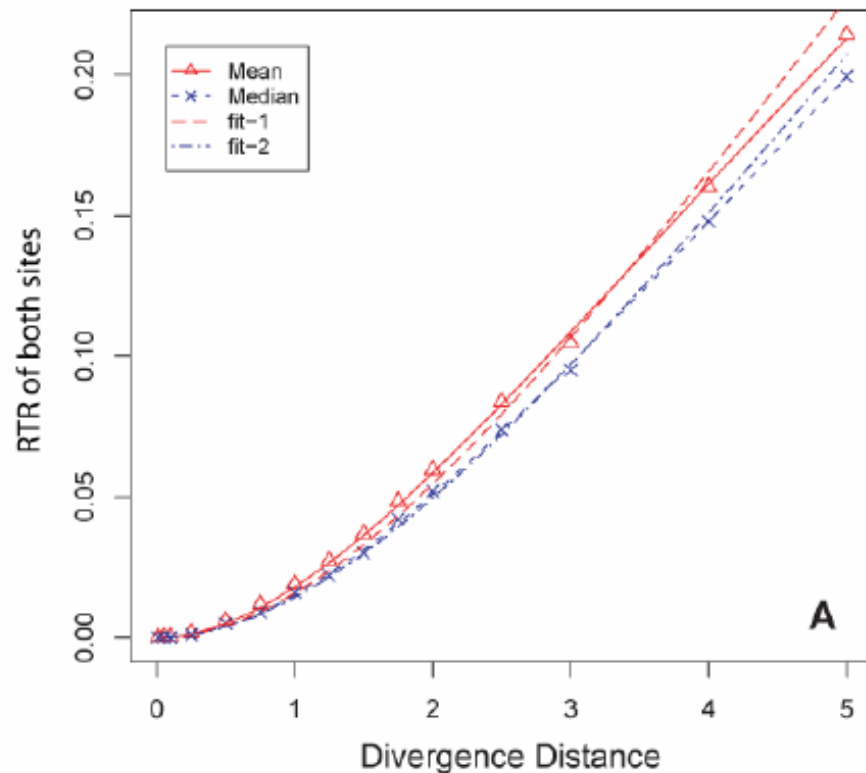
- Now: distance fixed for human/mouse
- free parameter shown: spacer E2F/TSS
- Prob. for turnover in *both* species



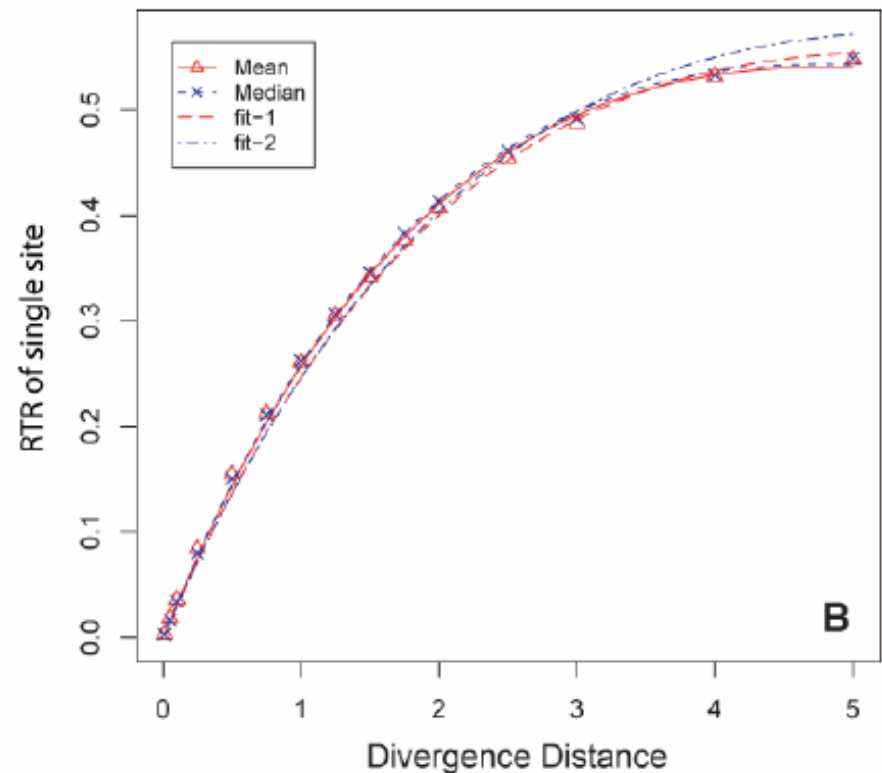
Pair of E2F/myc

| | |
|---|----------------|
| E2F location relative to TSS | [-50, -100] |
| Myc location relative to TSS | [-100, -150] |
| Copy number of E2F | 1 |
| Copy number of Myc | 1 |
| DNA strand of E2F site | + |
| DNA strand of Myc site | + |
| Additional space constraint between Myc and E2F sites | [50, 60] |

No spatial constraint

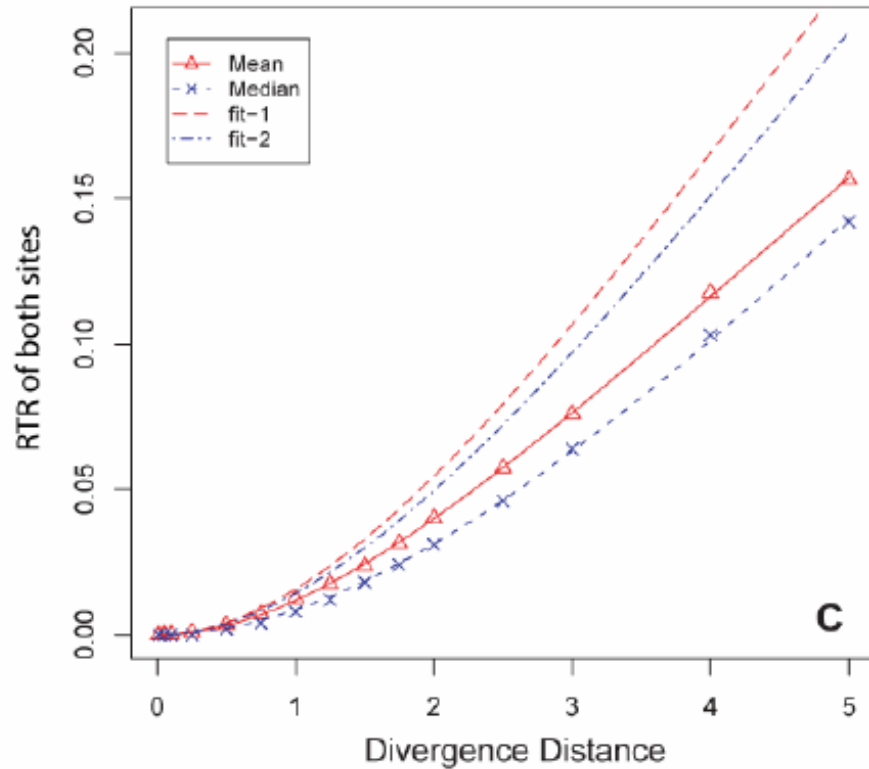


Both sites

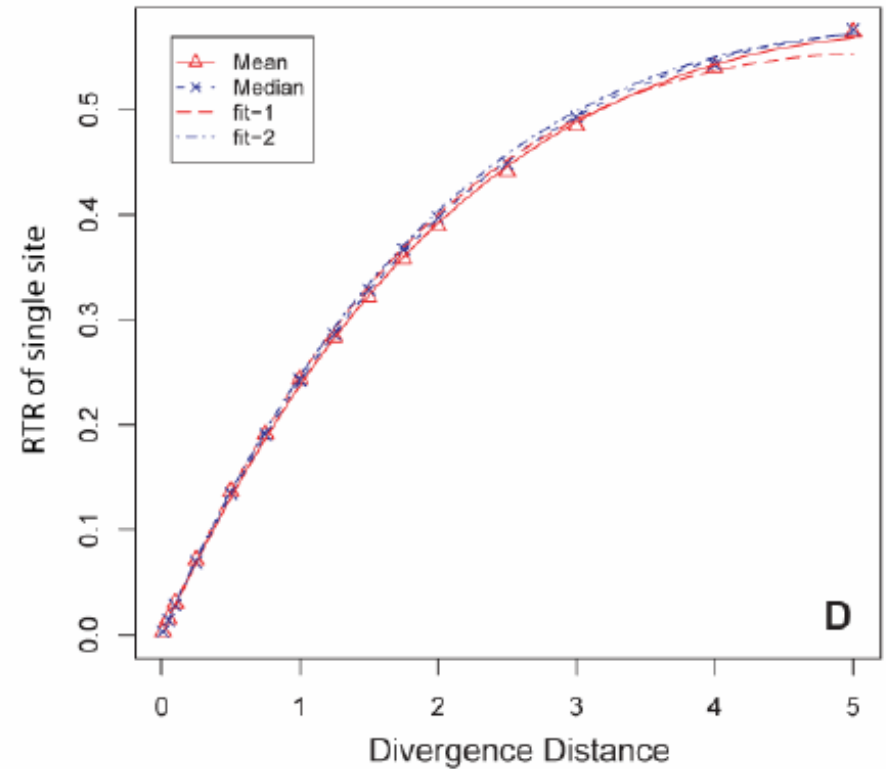


One site

With spatial constraint



Both sites



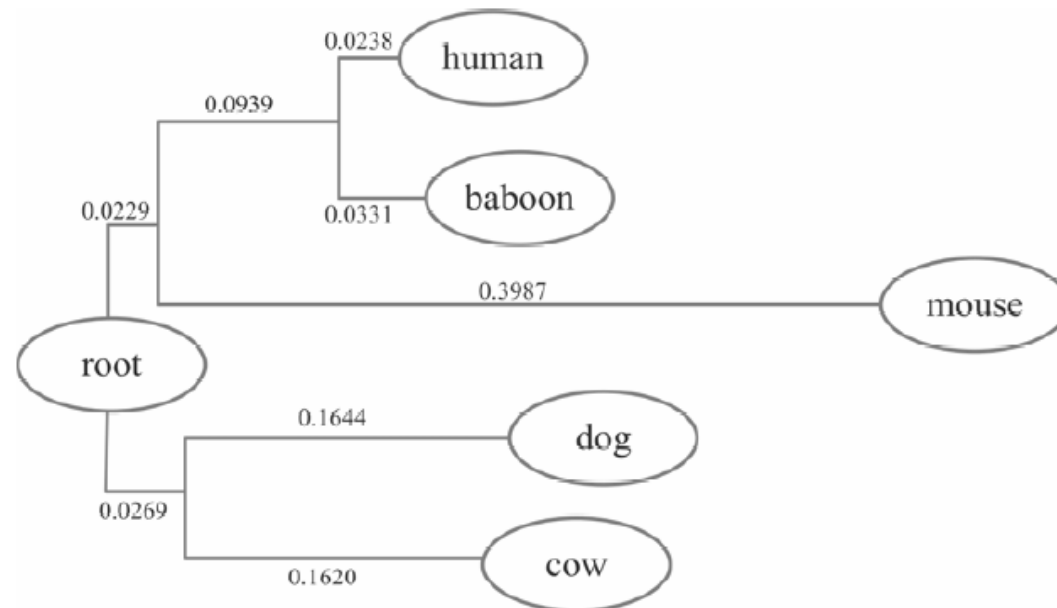
One site

Evaluate alignment accuracy

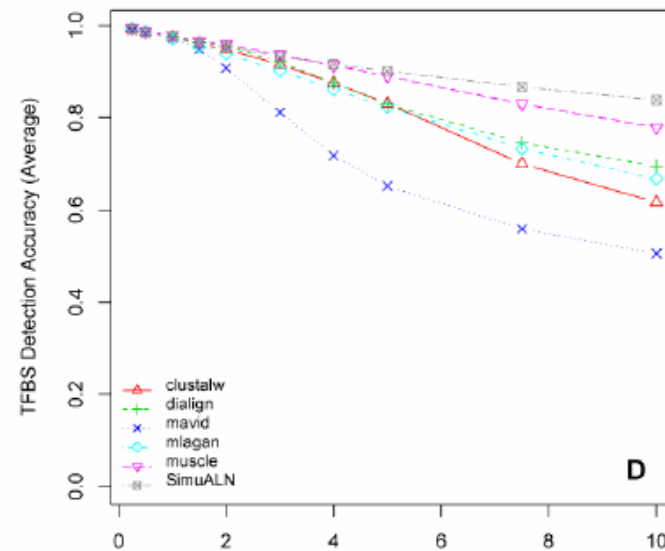
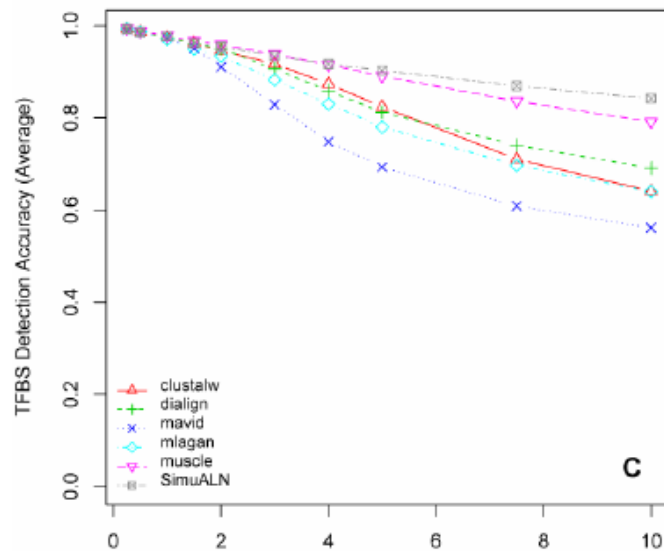
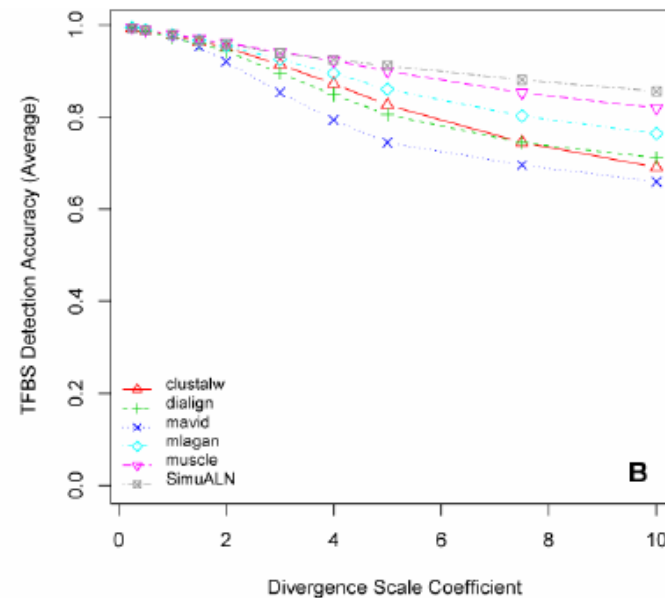
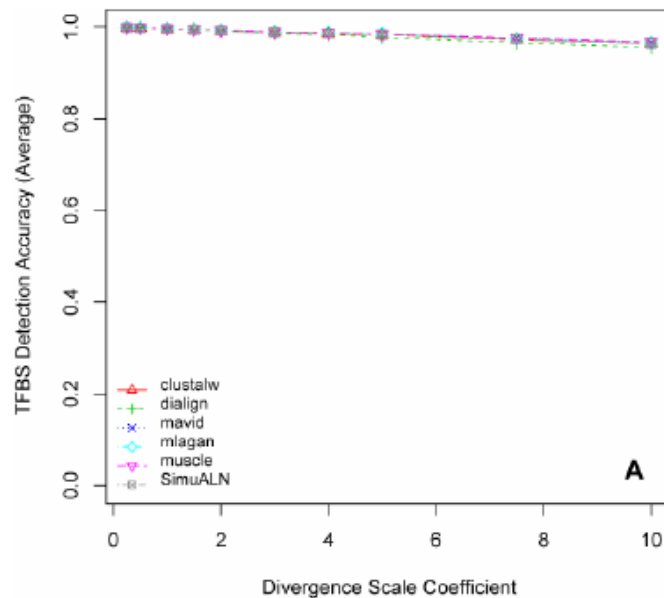
- Simulate evolution over various scaled trees
- Once simulated, run global multiple aligners
 - mlagan, mavid, muscle, dialign, clustalw
- We can then trace back which sites did not turn over and should be aligned
 - Neutral evolution -> we know all sites are there
- We are nice (of course :)
 - Turnover,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

Mammalian sequences

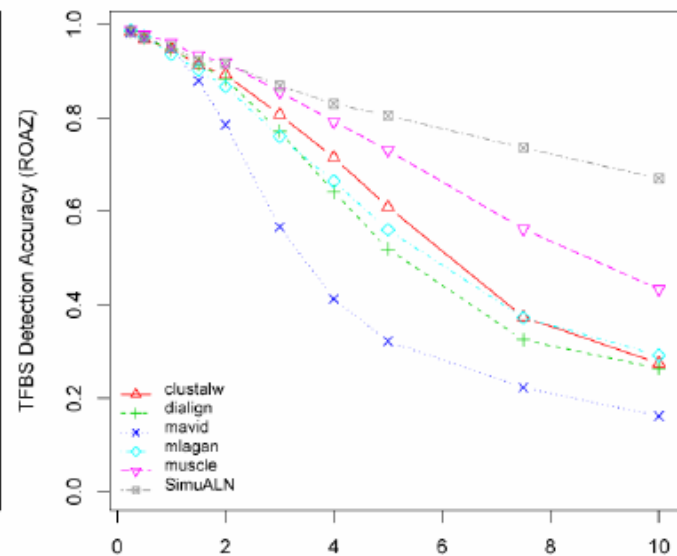
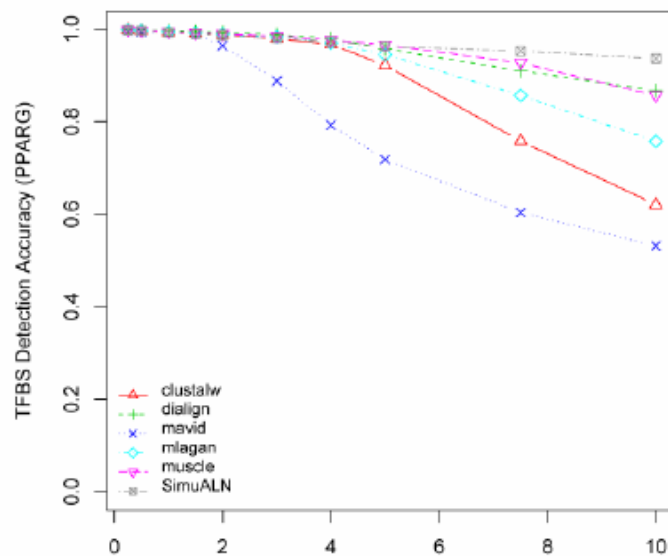
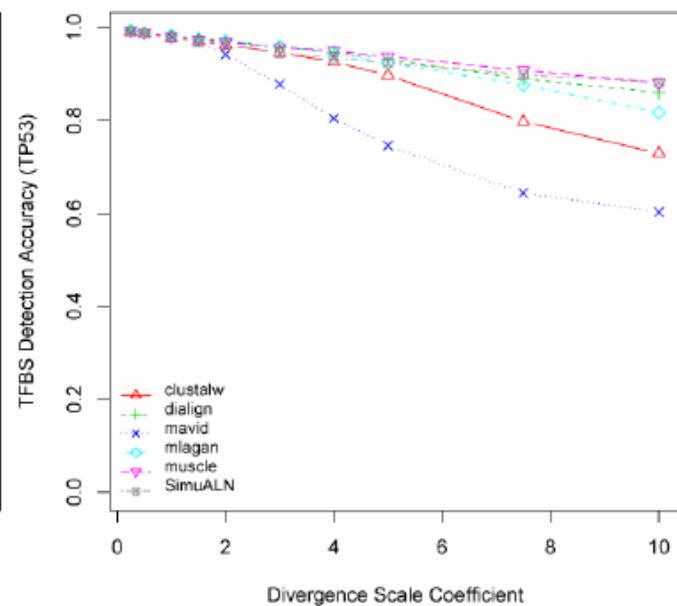
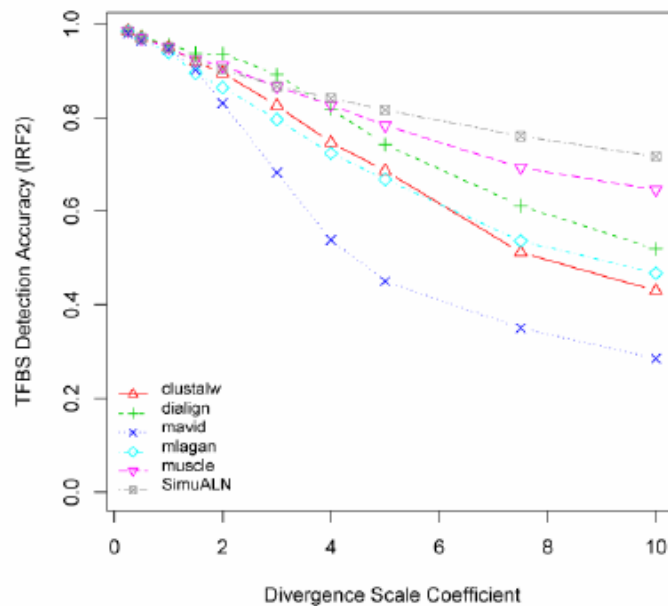
| Name | Accession# | Len | Strand | Location (min, max) | Copy # (min, max) | Cutoff |
|--------|------------------------------|-----|--------|---------------------|-------------------|--------|
| YY1E2F | MA0095 (YY1) MA0024 (E2F) | 13 | + | (20, 30) | (1, 1) | 0.90 |
| Pax6 | MA0069 | 14 | + | (50,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



Accuracy w/increasing #species



Accuracy for individual factors



Summary pt II

- 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

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

Images

Dan Mace

Phil Benfey Lab
Ji-Young Lee
Todd Twigg
Juliette Colinas

Rob Clark Lab

