Lecture 8: Gene finding and Regulatory sequence analysis

Gene Finding

Sources:

- Lecture notes of Larry Ruzzo, UW.
- Slides by Nir Friedman, Hebrew U.
- Burge, Karlin: "Finding Genes in Genomic DNA", Curr. Opin. In Struct. Biol 8(3) '98
- Slides by Chuong Huynh on Gene Prediction, NCBI
- Durbin's book, Ch. 3
- Pevzner's book, Ch. 9

Motivation

- ~3Gb human DNA in GenBank
- Only ~1.5% of human DNA is coding for proteins
- 155,176,494,699 total bases in GenBank (2013)
- Hundreds of species have been sequenced, thousands to follow
- Total number of species represented in UniProtKB/Swiss-Prot (2013): 13,041
- Need to locate the genes!
- **Goal: Automatic finding of genes**

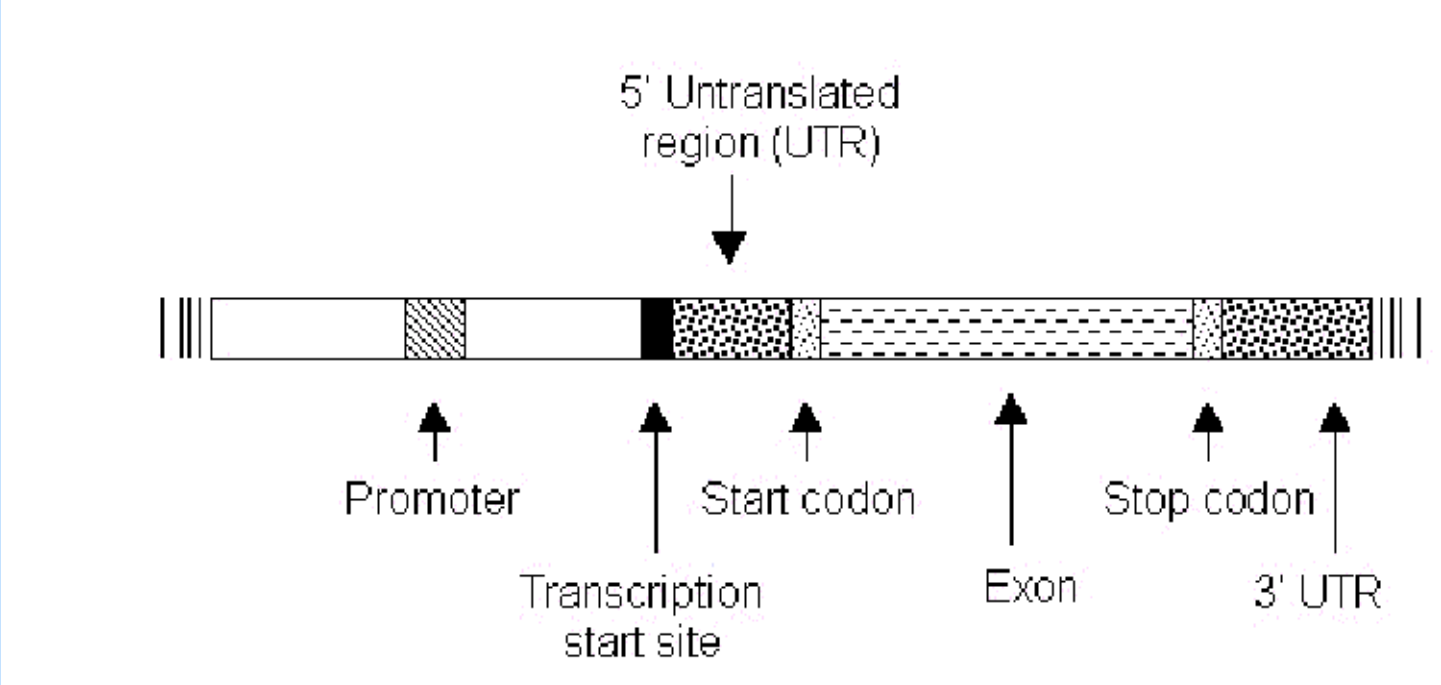
Reminder: The Genetic Code

First letter	Second letter: U		Second letter: C		Second letter: A		Second letter: G		Third letter
U	UUU UUC	Phenyl-alanine	UCU UCC UCA UCG	Serine	UAU UAC	Tyrosine	UGU UGC	Cysteine	U C
	UUA UUG	Leucine			UAA UAG	Stop codon Stop codon	UGA UGG	Stop codon Tryptophan	A G
C	CUU CUC CUA CUG	Leucine	CCU CCC CCA CCG	Proline	CAU CAC	Histidine	CGU CGC CGA CGG	Arginine	U C A G
					CAA CAG	Glutamine			
A	AUU AUC AUA	Isoleucine	ACU ACC ACA ACG	Threonine	AAU AAC	Asparagine	AGU AGC	Serine	U C A G
	AUG	Methionine; initiation codon			AAA AAG	Lysine	AGA AGG	Arginine	
G	GUU GUC GUA GUG	Valine	GCU GCC GCA GCG	Alanine	GAU GAC	Aspartic acid	GGU GGC GGA GGG	Glycine	U C A G
					GAA GAG	Glutamic acid			

1 start, 3 stop codons

Genes in Prokaryotes

- High gene density (e.g. 70% coding in H. Influenza)
- No introns
- ➔ most long ORFs are likely to be genes.



Open Reading Frames

- Reading Frame: 3 possible ways to read the sequence (on each strand).
- ACCUUAGCGUA = Threonine-Leucine-Alanine
- ACCUUAGCGUA = Proline-Stop-Arginine
- ACCUUAGCGUA = Leucine-Serine-Valine
- Open Reading Frame (ORF): Reading frame with no stop codons.
- ORF is maximal if it starts right after a stop and ends in a stop
- Untranslated region (UTR): ends of the mRNA (on both sides) that are not translated to protein.

Finding long ORFs

- In random DNA, one stop codon every 64/3 → 21 codons on average
- Average protein is ~300 AA long
- => search long ORFs
- Problems:
  - short genes
  - many more ORFs than genes
    - In E. Coli one finds 6500 ORFs but only 1100 genes.
    - Call the remaining Non-coding ORF (NORFS)
  - Overlapping long ORFs on opposite strands

Codon Frequencies

- Coding DNA is not random:
  - In random DNA, expect
    - Leucine:Alanine:Tryptophan ratio of 6:4:1
  - In real proteins, 6.9:6.5:1
  - In some species, 3rd position of the codon, up to 90% A or T
- Different frequencies for different species.

Human codon usage

frequency of usage of each codon (per thousand)

relative freq of each codon among synonymous codons

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1	Arg	CGG	10.4	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.1
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.6	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.4	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.8	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.3	0.56	Tyr	TAC	16.48	0.58	His	CAC	14	0.59
Val	GTG	28.6	0.48	Met	ATG	21.86	1	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.1	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.3	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.1	Thr	ACG	6.8	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.5	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.4	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

http://genome.imim.es/courses/Lisboa01/slide3.8.html

First Order Markov Model

- Use two Markov models (similar to CpG islands) to discriminate genes from NORFs
- Given a sequence of nucleotides $X_1,\ldots,X_n$ we compute the log-likelihood (aka log-odds) ratio:

$$\log \frac{P(X_1,\ldots,X_n \mid \mathrm{G})}{P(X_1,\ldots,X_n \mid R)} = \sum_i \log \frac{A^G{}_{X_i X_{i+1}}}{A^R{}_{X_i X_{i+1}}}$$

First Order Markov Model

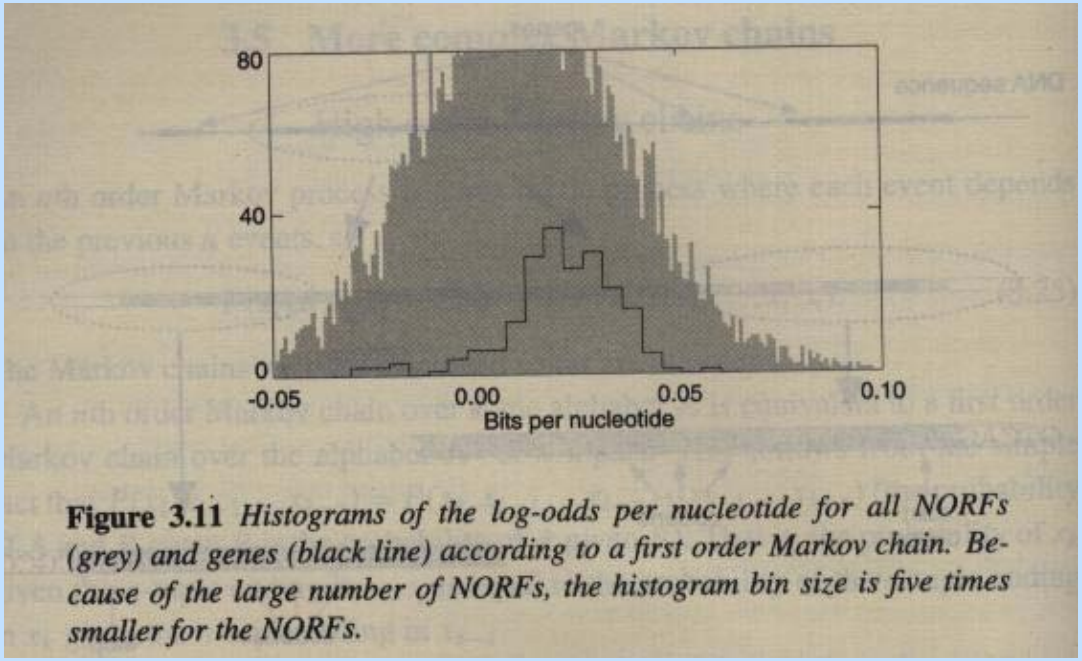


Figure 3.11 *Histograms of the log-odds per nucleotide for all NORFs (grey) and genes (black line) according to a first order Markov chain. Because of the large number of NORFs, the histogram bin size is five times smaller for the NORFs.*

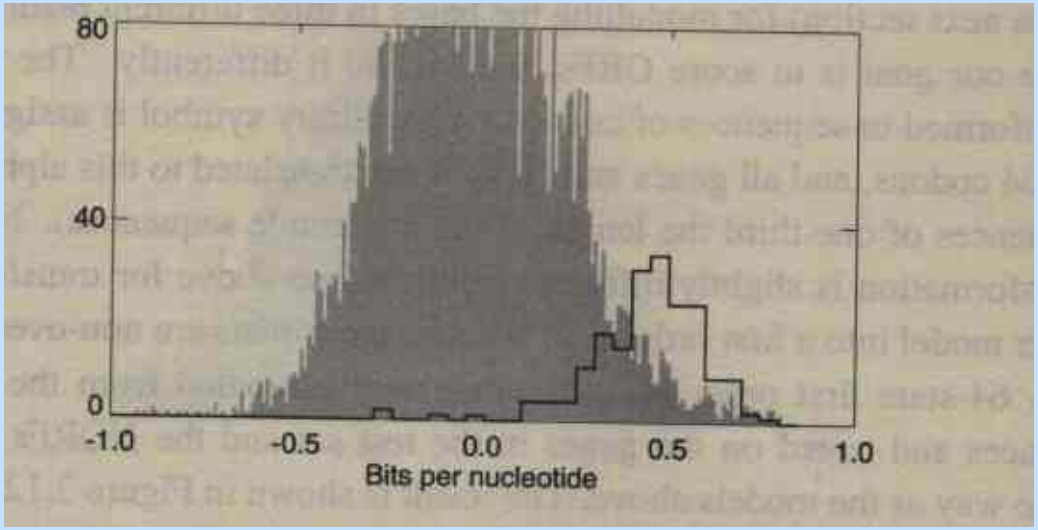
Test on E. Coli data

Durbin et al pp.74

- Average log-odds per nucleotide in genes : 0.018
- Average log-odds per nucleotide in NORFs : 0.009
- But the variance makes it useless for discrimination (similar results for 2nd-order MM)

Using codons

- Translate each ORF into a sequence of codons
- Form a 64-state Markov chain
  - Codon is more informative than its translation
- Estimate probabilities in coding regions and NORFs



Durbin et al pp.76

Using Codon Frequencies

- Assume each codon is iid
- For codon *abc* calculate frequency f_{abc} in coding region
- Given coding sequence $a_1b_1c_1,\ldots, a_{n+1}b_{n+1}c_{n+1}$
- Calculate

$$p_1 = f_{a_1b_1c_1} * f_{a_2b_2c_2} * \ldots * f_{a_nb_nc_n}$$

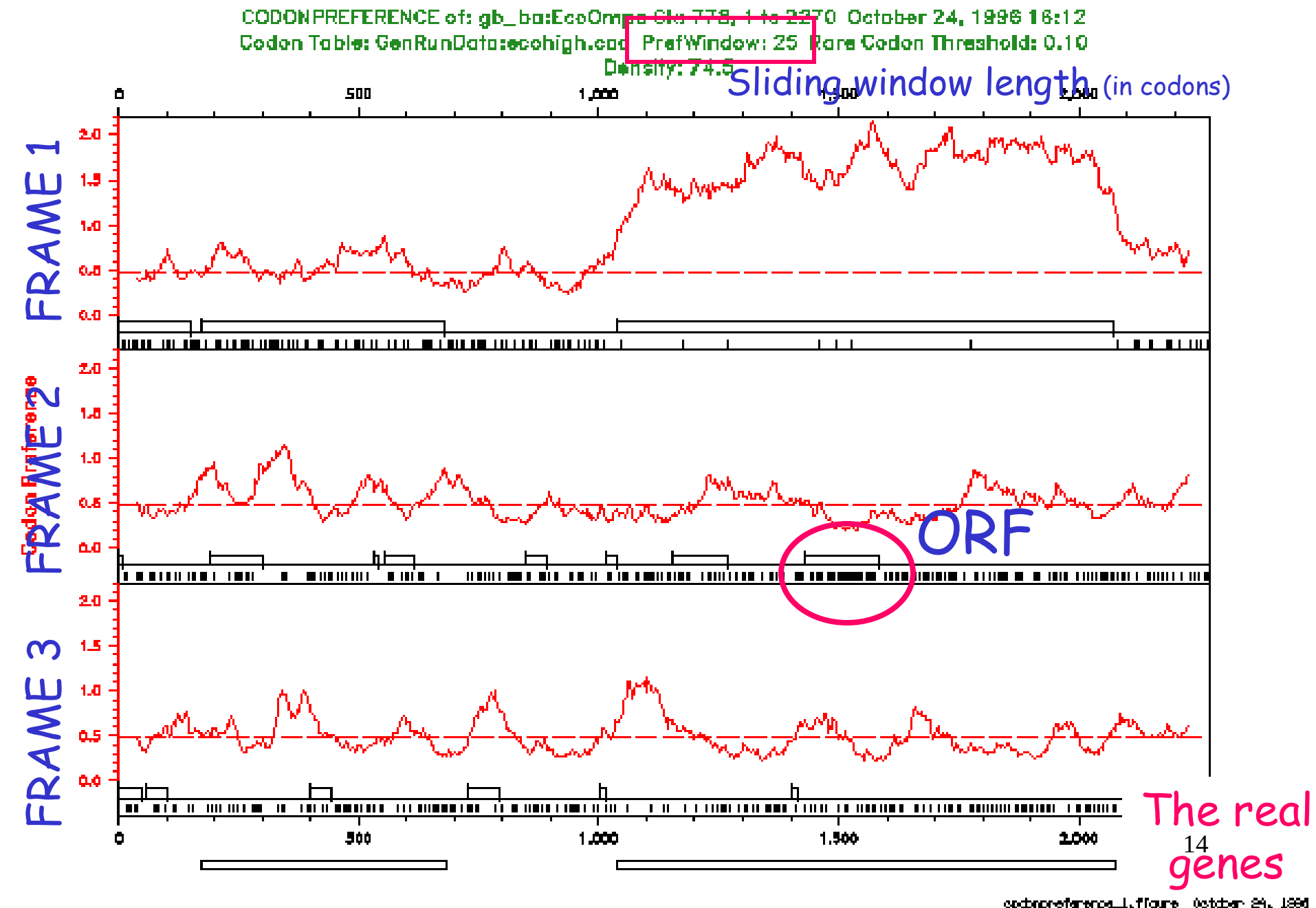
$$p_2 = f_{b_1c_1a_2} * f_{b_2c_2a_3} * \ldots * f_{b_nc_na_{n+1}}$$

$$p_3 = f_{c_1a_2b_2} * f_{c_2a_3b_3} * \ldots * f_{c_na_{n+1}b_{n+1}}$$

- The probability that the *i*-th reading frame is the coding region:

$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

CodonPreference



RNA Transcription

- Not all ORFs are expressed.
- Transcription depends on regulatory signals
- Minimal regulatory region – core promoter to which RNA polymerase and initiation factors bind to start transcription.
- At the termination signal the polymerase releases the RNA and disconnects from the DNA.

E. coli promoters

consensus sequence:

nnnTTGACAnnnnnnnnnnnnnnnnnnTATAATnnnnnnNnnn

-35 -12 mRNA start

- "TATA box" (or Pribnow Box)
- Not exact

Positional Weight Matrix (PWM)

- $f_{b.j}$: frequency of base b in position j.
- Assumes independence btw positions
- For TATA box:

pos:	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

- f_b : background frequency.

Scoring Function

- For sequence $S=B_1B_2B_3B_4B_5B_6$

$$P(S \mid \text{promoter}) = \prod_{i=1}^{6} f_{B_i,i}$$

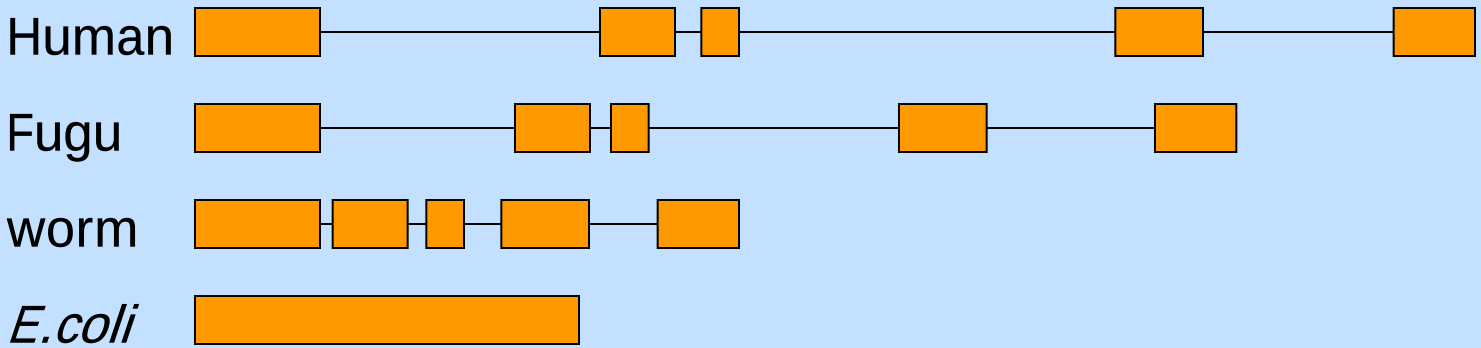
$$P(S \mid \text{non-promoter}) = \prod_{i=1}^{6} f_{B_i}$$

- Log-likelihood ratio score:

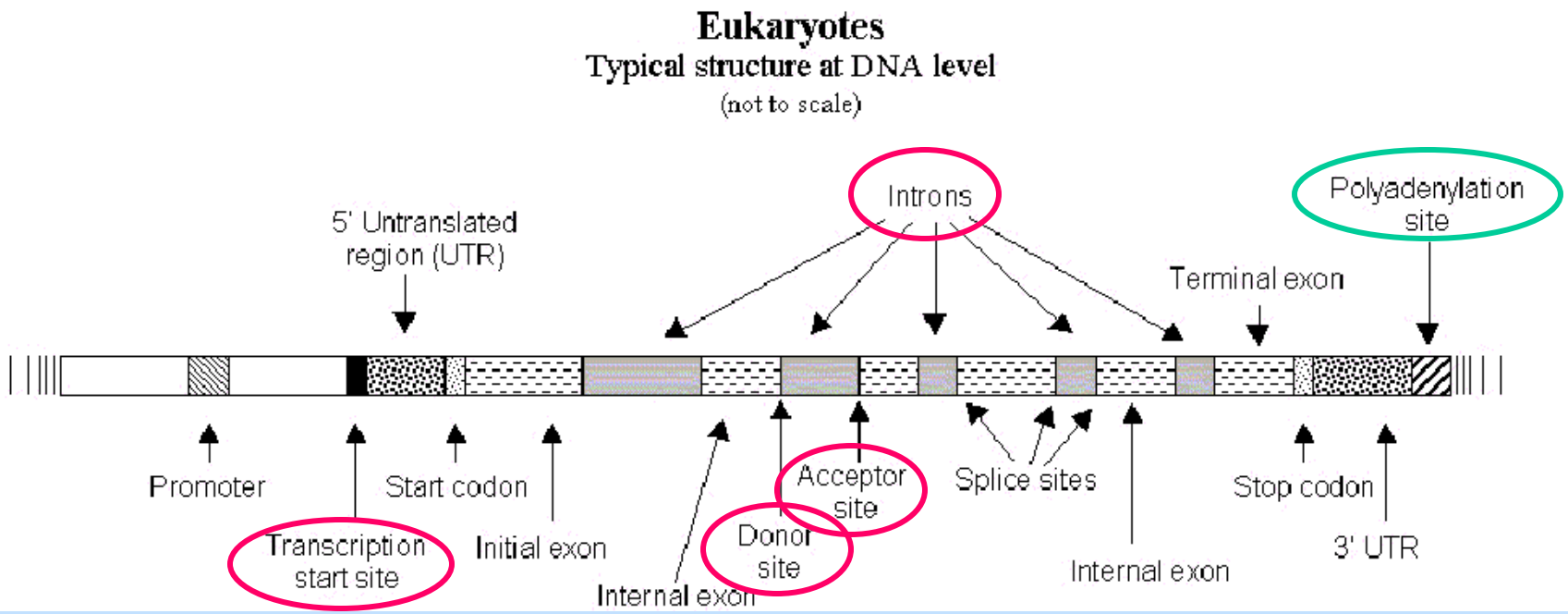
$$\log\left(\frac{P(S \mid \text{promoter})}{P(S \mid \text{non-promoter})}\right) = \log\left(\frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}}\right) = \sum_{i=1}^{6} \log\left(\frac{f_{B_i,i}}{f_{B_i}}\right)$$

Gene finding: coding density

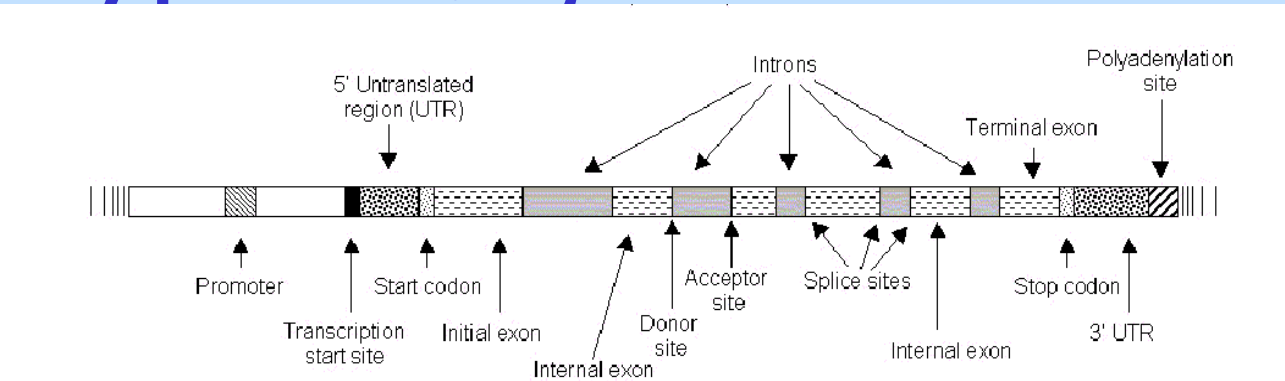
- As the coding/non-coding length ratio decreases, exon prediction becomes more complex



Eukaryote gene structure



Typical figures: verterbrates



- TF binding site: ~6bp; 0-2kbp upstream of TSS
- 5' UTR: ~750 bp, 3' UTR: ~450bp
- Gene length: 30kb, coding region: 1-2kb
- Average of 6 exons, 150bp long
- Huge variance: - dystrophin: 2.4Mb long
  - Blood coagulation factor: 26 exons, 69bp to 3106bp; intron 22 contains another unrelated gene

Splicing

- Splicing: the removal of the introns.
- Performed by complexes called spliceosomes, containing both proteins and snRNA.
- The snRNA recognizes the splice sites through RNA-RNA base-pairing
- Recognition must be precise: a 1nt error can shift the reading frame making nonsense of its message.
- Many genes have alternative splicing, which changes the protein created.

Exon-intron junctions

AGGUAAGU.....CTGAC.....NCAGG.....

62 77 100 100 60 74 84 50 [63 -91] - 78 100 100 55 freq(%)

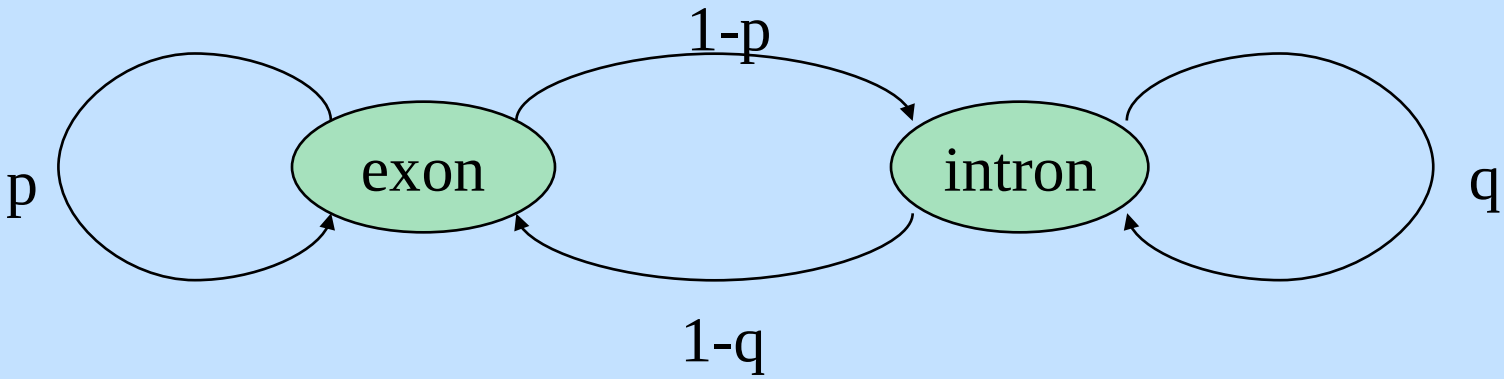
Donor site branchpoint <-15bp -> acceptor site

pyrimidine [C,T] rich

- 1st approach: positional weight matrices
  - Problematic with weak/short signals
  - Does not exploit all info (reading frames, intron/exon stats…)

➔ try integrated approaches!

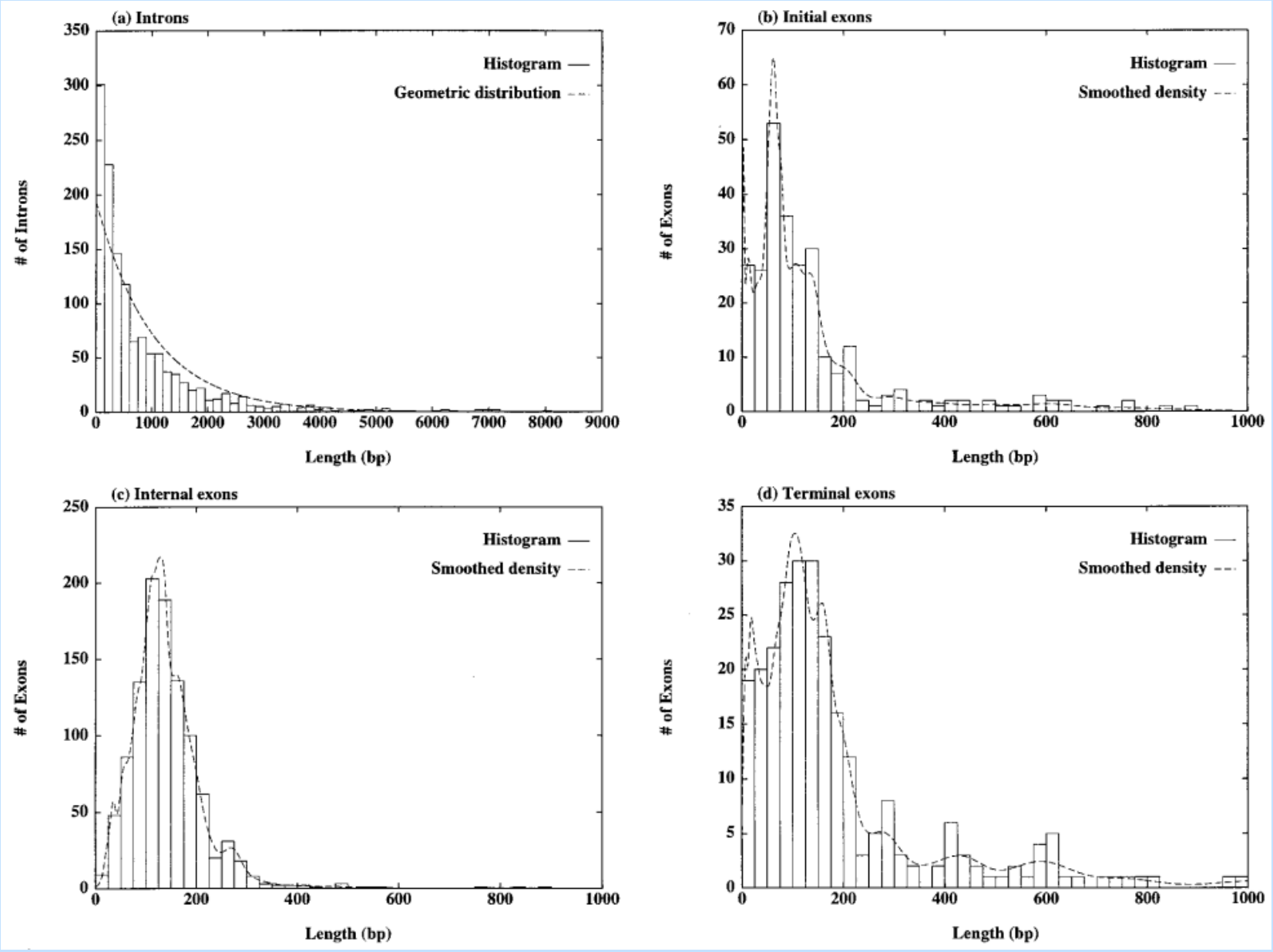
Length Distribution



- Above is a simple HMM for gene structure
- The length of each exon (intron) has a geometric distribution:

$$P(\text{exon of length } k) = p^k (1-p)$$

Since an HMM is a memory-less process, the only length distribution that can be modeled is geometric.


(a) Introns
Histogram
Geometric distribution
# of Introns
Length (bp)
(b) Initial exons
Histogram
Smoothed density
# of Exons
Length (bp)
(c) Internal exons
Histogram
Smoothed density
# of Exons
Length (bp)
(d) Terminal exons
Histogram
Smoothed density
# of Exons
Length (bp)

Exon Length Distribution

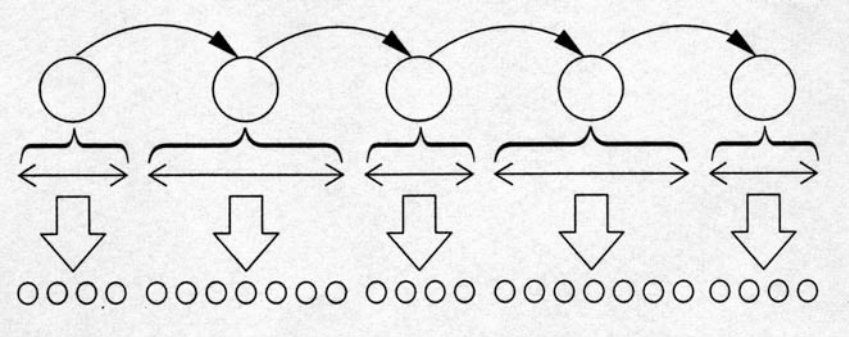
- Intron length distribution seems approximately geometric
- This is not so for exons.
- Length seems to have a functional role on the splicing itself:
  - Too short (under 50bps): the spliceosomes have no room
  - Too long (over 300bps): ends have problems finding each other.
  - But as usual there are exceptions.

➔ Need a different model for exons.

Generalized HMM

(Burge & Karlin, J. Mol. Bio. 97 268 78-94)

- Hidden Markov states $q_1, \ldots q_n$
- State q_i has output length distribution f_i
- Output of each state can have a separate probabilistic model (weight matrix, MM...)
- Initial state probability distribution π
- State transition probabilities T_{ij}



GenScan Model

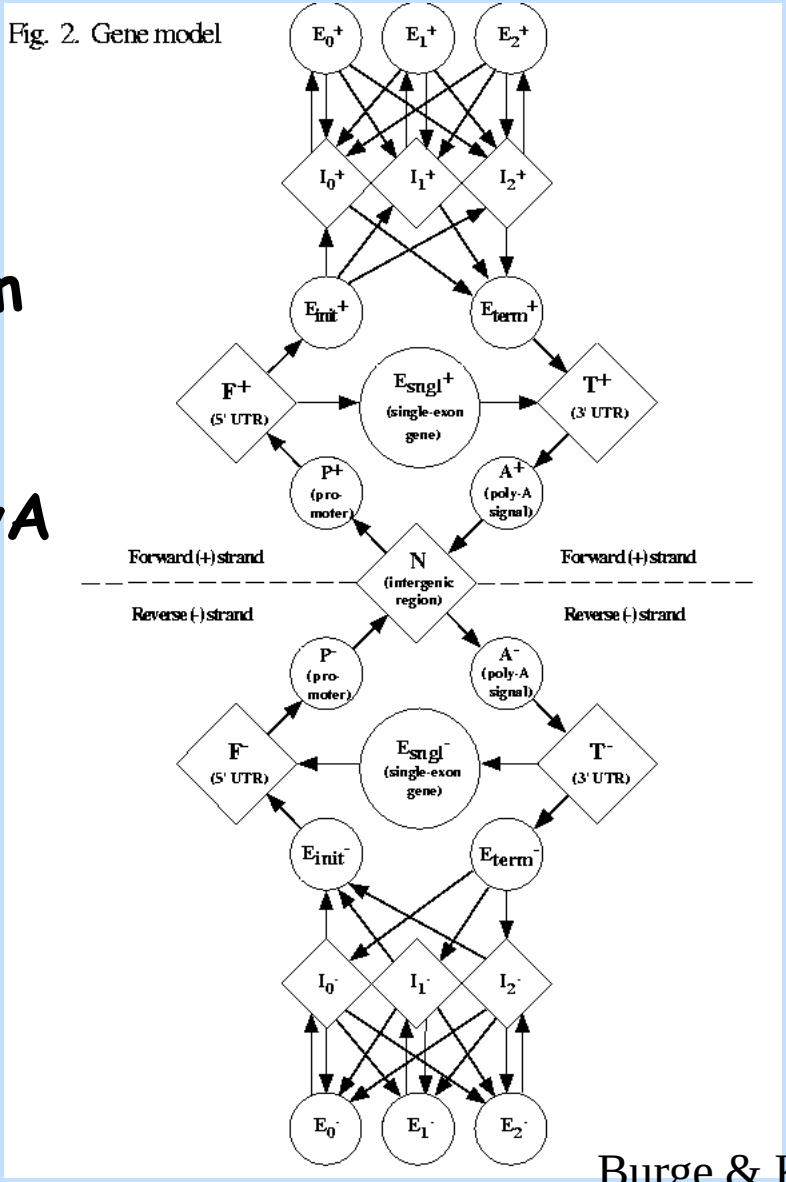
Exon

Intron

Exon init/term

5'/3' UTR

Promoter/PolyA

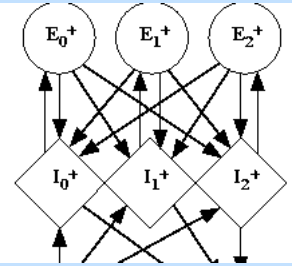


Forward strand

Backward strand

Burge & Karlin JMB 97

GenScan model



- states = functional units on a gene
- The allowed transitions ensure the order is biologically consistent.
- As an intron may cut a codon, one must keep track of the reading frame, hence the three *I* phases:
  - phase I_0: between codons
  - phase I_1: introns that start after 1st base
  - phase I_2 : introns that start after 2nd base

Signal Models

- Genscan uses different models to model the different biological signals
  - Weight Matrix Model
    - Position specific distribution.
    - Each column is independent
  - Used for
    - Translation initiation signal
    - Translation termination signal
    - promoters
    - polyadenylation signals

Splice Sites

- Correct recognition of these sites greatly enhances ability to predict correct exon boundaries.
- Used Weighted Array Model: a generalization of PWM that allows for dependencies between adjacent positions
- Accurate modeling of these sites led to substantial improvement in performance.

GenScan Performance

Accuracy of GENSCAN for different signal and exon types.

(a) Prediction of individual splice sites and translational signals.

Type of signal	Type of exon	Annotated exons		Predicted exons	
		Number	% Correctly predicted	Number	% Correctly predicted
Initiation	Initial only	570	66	450	84
Termination	Terminal only	570	78	487	91
5′ splice site	Initial only	570	88	450	89
5′ splice site	Internal only	1510	93	1682	89
5′ splice site	Initial and internal	2080	91	2132	89
3′ splice site	Terminal only	570	81	487	92
3′ splice site	Internal only	1510	92	1682	83
3′ splice site	Internal and terminal	2080	89	2169	85

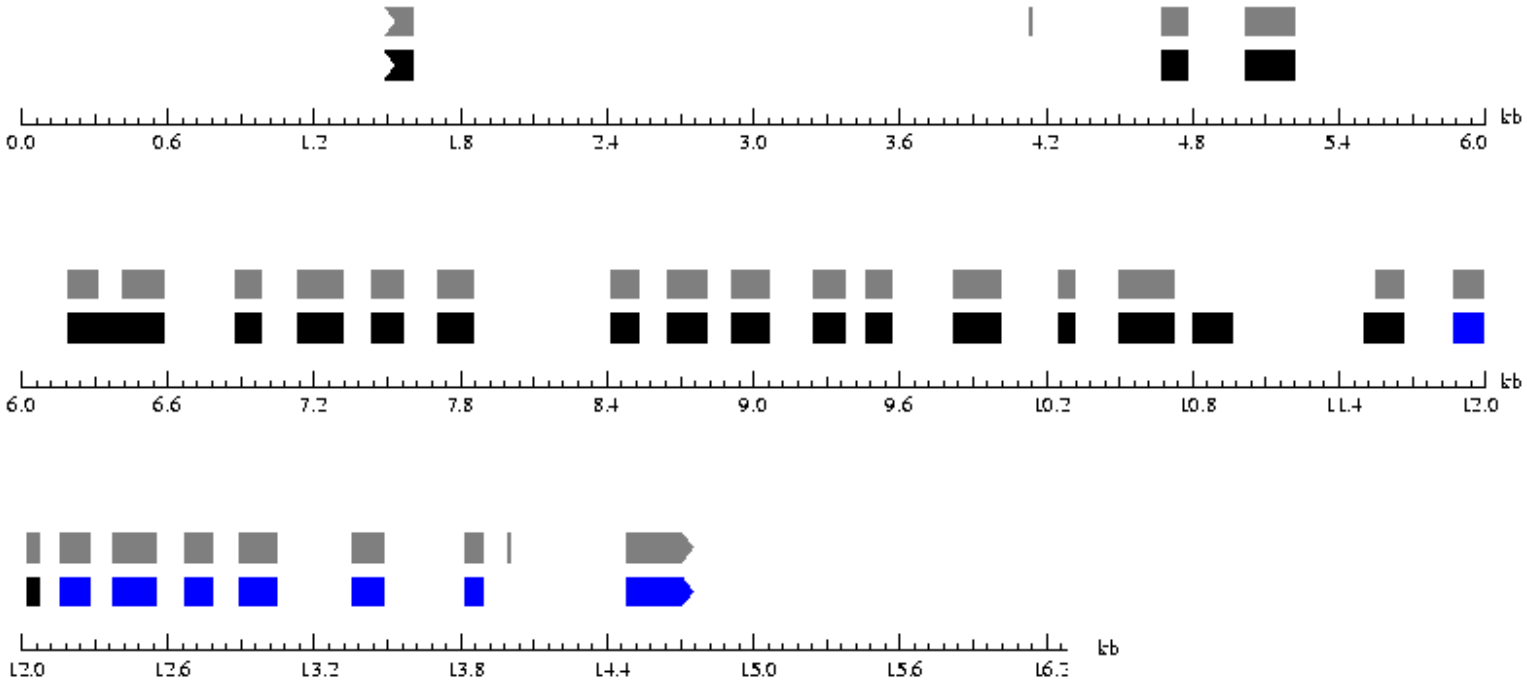
(b) Accuracy for initial, internal and terminal exons.

Exon type	Annotated exons				Predicted exons			
	Number	% Exactly	% Partially	% Missed	Number	% Exactly	% Partially	% Wrong
Initial	570	65	25	9	457	81	9	10
Internal	1510	90	5	4	1707	80	11	8
Terminal	570	76	8	15	509	84	6	8
All types	2650	81	10	8	2678	81	10	9

- Predicts correctly 80% of exons
- Prediction accuracy per bp > 90%

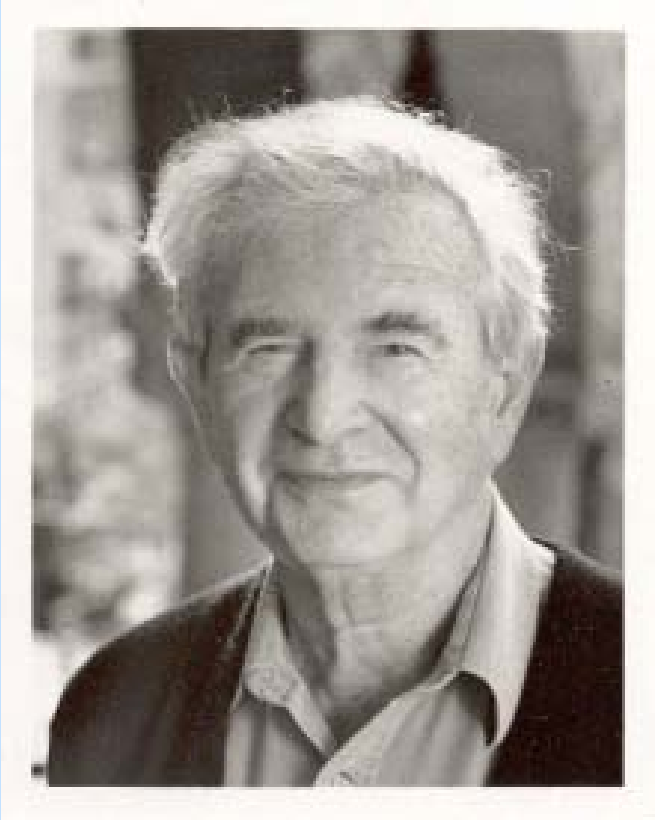
GenScan Output

Fig. 12. GENSCAN PostScript output for sequence HSNCAMX1



Key: Initial exon; Internal exon; Terminal exon; Single-exon gene; Annotated exon; Predicted exon

Sam Karlin, Chris Burge

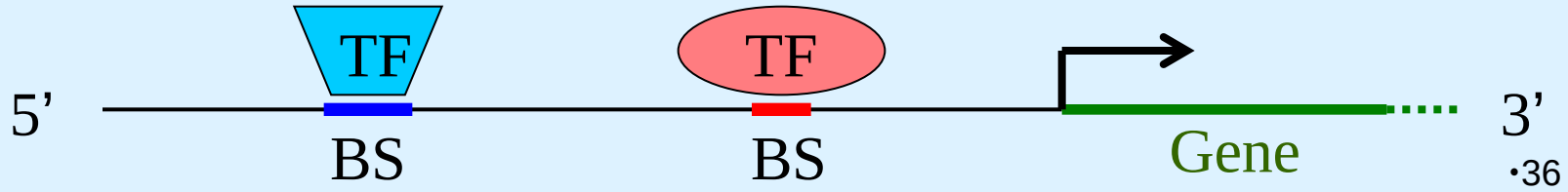


Regulatory sequence analysis

Slides with Chaim Linhart

Regulation of Transcription

- A gene's ranscription regulation is mainly encoded in the DNA in a region called the promoter
- Each promoter contains several short DNA subsequences, called binding sites (BSs) that are bound by specific proteins called transcription factors (TFs)



Regulation of Transcription (II)

Assumption:

Co-expression

↓

Transcriptional co-regulation

↓

Common BSs

WH-questions

- √ Why are we looking for common BSs?
- *What* exactly are we trying to find?
- *Where* should we look for it?
- *How* can we find it?

Promoter Region (Where?)

What is the promoter region?

- Upstream Transcription Start Site (TSS)
  - Too short → miss many real BSs (false negatives)
  - Too long → lots of wrong hits (false positives)
  - Length is species dependent (e.g., yeast ~600bp, thousands in human)
  - Common practice: ~ 500-2000bp
- Consider both strands?
  - Common practice: Yes

What: Models for Binding Sites

(I) Exact string(s)

Example:

BS = TACACC , TACGGC

CAATGCAGGATACACCGATCGGTA

GGAGTACGGCAAGTCCCCATGTGA

AGGCTGGACCAGACTCTACACCTA

In red: hits

(II) String with mismatches

<u>Example:</u>

BS = TACACC + 1 mismatch

CAATGCAGGATTCACCGATCGGTA

GGAGTACAGCAAGTCCCCATGTGA

AGGCTGGACCAGACTCTACACCTA

(III) Degenerate string

			T		
		G	G		
T	A	C	A	A	C

Example:

BS = TASDAC (S={C,G} D={A,G,T})

CAATGCAGGATACAACGATCGGTA

GGAGTAGTACAAGTCCCCATGTGA

AGGCTGGACCAGACTCTACGACTA

(IV) Position Weight Matrix (PWM)

a.k.a Position Specific Scoring Matrix (PSSM)

Example:

Score: product of base probabilities. Need to set score threshold for hits.

A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

ATGCAGGATACACCGATCGGTA 0.0605

GGAGTAGAGCAAGTCCCGTGA 0.0605

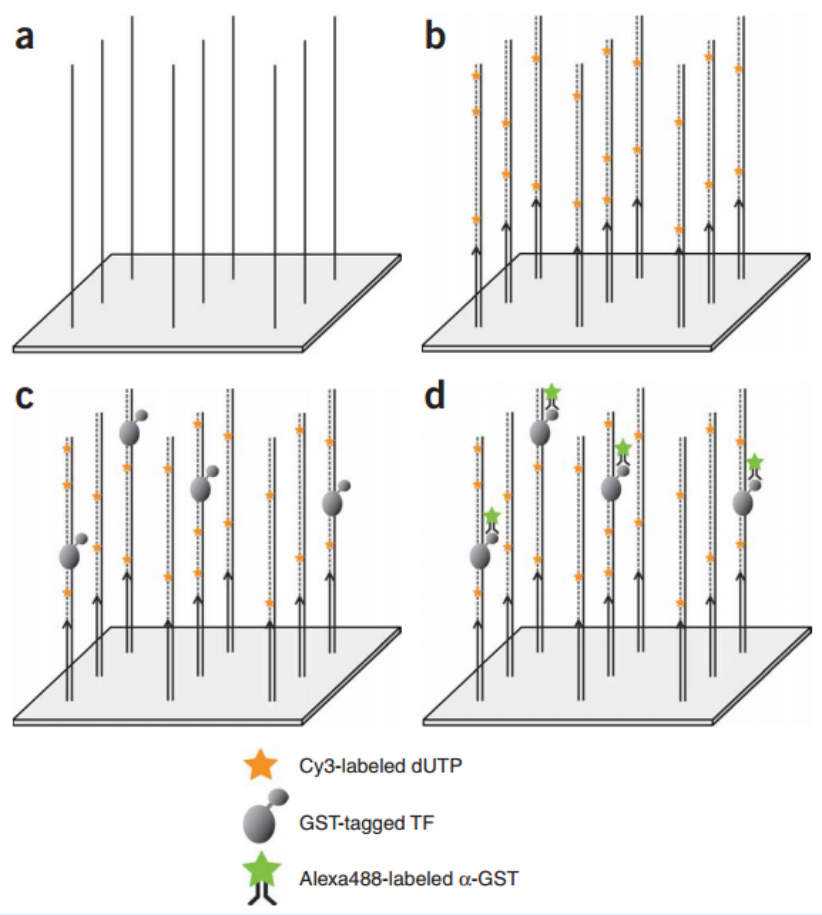
AAGACTCTACAATTATGGCGT 0.0151

How: Experimental techniques

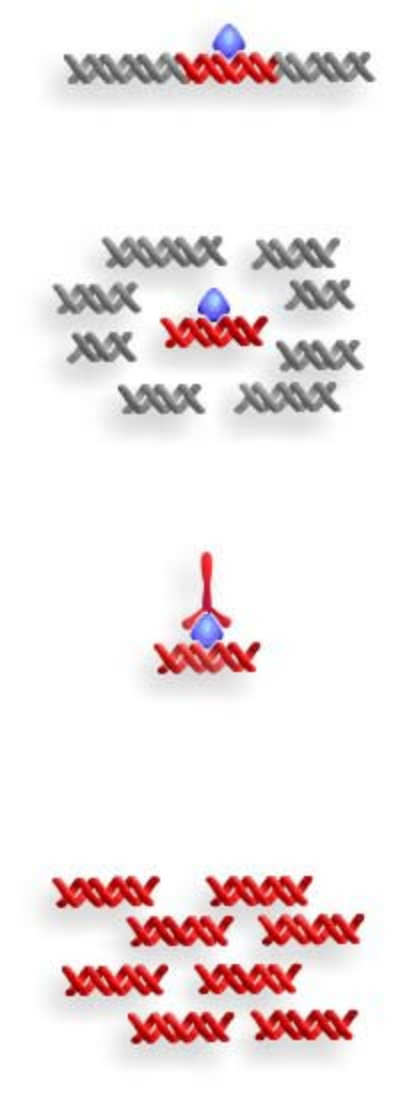
Protein Binding Microarrays

Berger et al, Nat. Biotech 2006

- Generate an array of double-stranded DNA with all possible k-mers
- Detect TF binding to specific k-mers



Chromatin Immunoprecipitation (ChIP)



DNA-binding proteins are crosslinked to DNA with formaldehyde in vivo.

Isolate the chromatin. Shear DNA along with bound proteins into small fragments.

Bind antibodies specific to the DNA-binding protein to isolate the complex by precipitation. Reverse the cross-linking to release the DNA and digest the proteins.

Identify bound DNA via microarray hybriziation or sequencing

http://www.bio.brandeis.edu/haberlab/jehsite/chip.html

How: Analyzing known motifs

PRIMA

PRomoter Integration in Microarray Analysis (Elkon et al. '03)

Goal: Identify enriched TFs = TFs whose BSs are over-represented in promoters of co-regulated genes

- Prepare a dictionary of motif hits
- Compute enrichment of hits in the given promoter set compared to a background set.

Computation of Motif Hits

Computing a threshold for a PWM:

- Compute 2nd-order Markov-Model of background sequences
- Generate random sequences using MM (e.g., 1,000 sequences of length 1,000)
- Set threshold s.t. PWM has ~5% hits at random.

This "ensures" a pre-defined false-positive rate,

but no guarantee on false-negative rate.

Motif Enrichment

Each promoter is hit or not.

Let: B = # of BG promoters

T = # of target-set promoters

b = # of BG promoters that are hit

t = # of target-set promoters that are hit

Then (hypergeometric distribution assumption):

Prob. for t hits in target-set:

$$P(t) = \binom{b}{t}\binom{B-b}{T-t} \Big/ \binom{B}{T}$$

Prob. for at least t hits:

$$p-value = \sum_{i=t}^{\min\{b,T\}} P(i)$$

TF Synergism

Find pairs of TFs that tend to occur in the same promoters

Let: T = # of promoters in target-set

t_1 , t_2 = # of promoters hit by TF 1,2

t_{12} = # of promoters hit by *both* TFs (w/o overlaps!)

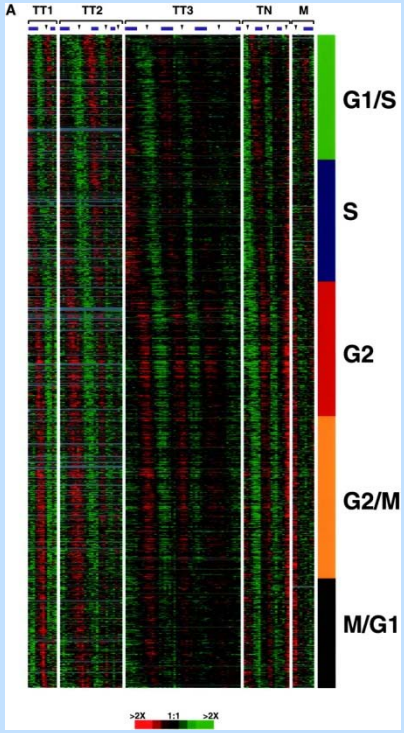
Then:

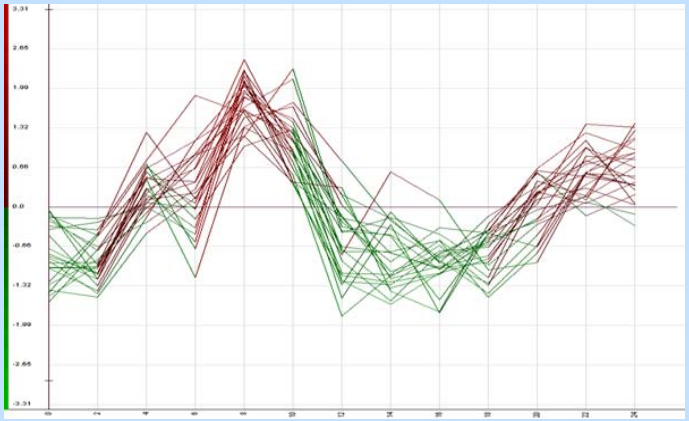
Prob. for co-occurrence of at least t_{12}:

$$synergism \;\; p-value = \frac{\sum_{i \geq t_{12}} \binom{t_1}{i}\binom{T-t_1}{t_2-i}}{\binom{T}{t_2}}$$

PRIMA: Human Cell Cycle

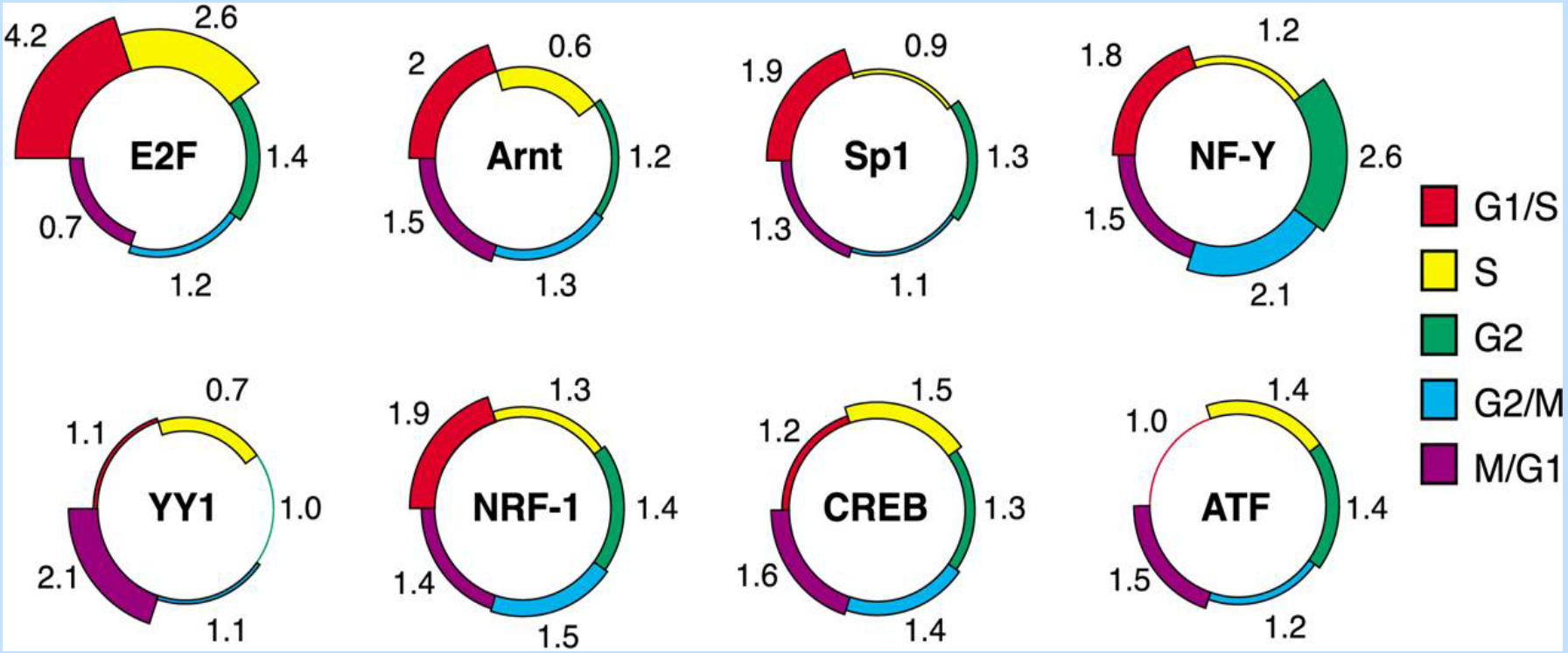
Whitfield et al. (’02) identified 568 genes that are periodically expressed in the human cell-cycle and partitioned them into the 5 phases of the cell-cycle





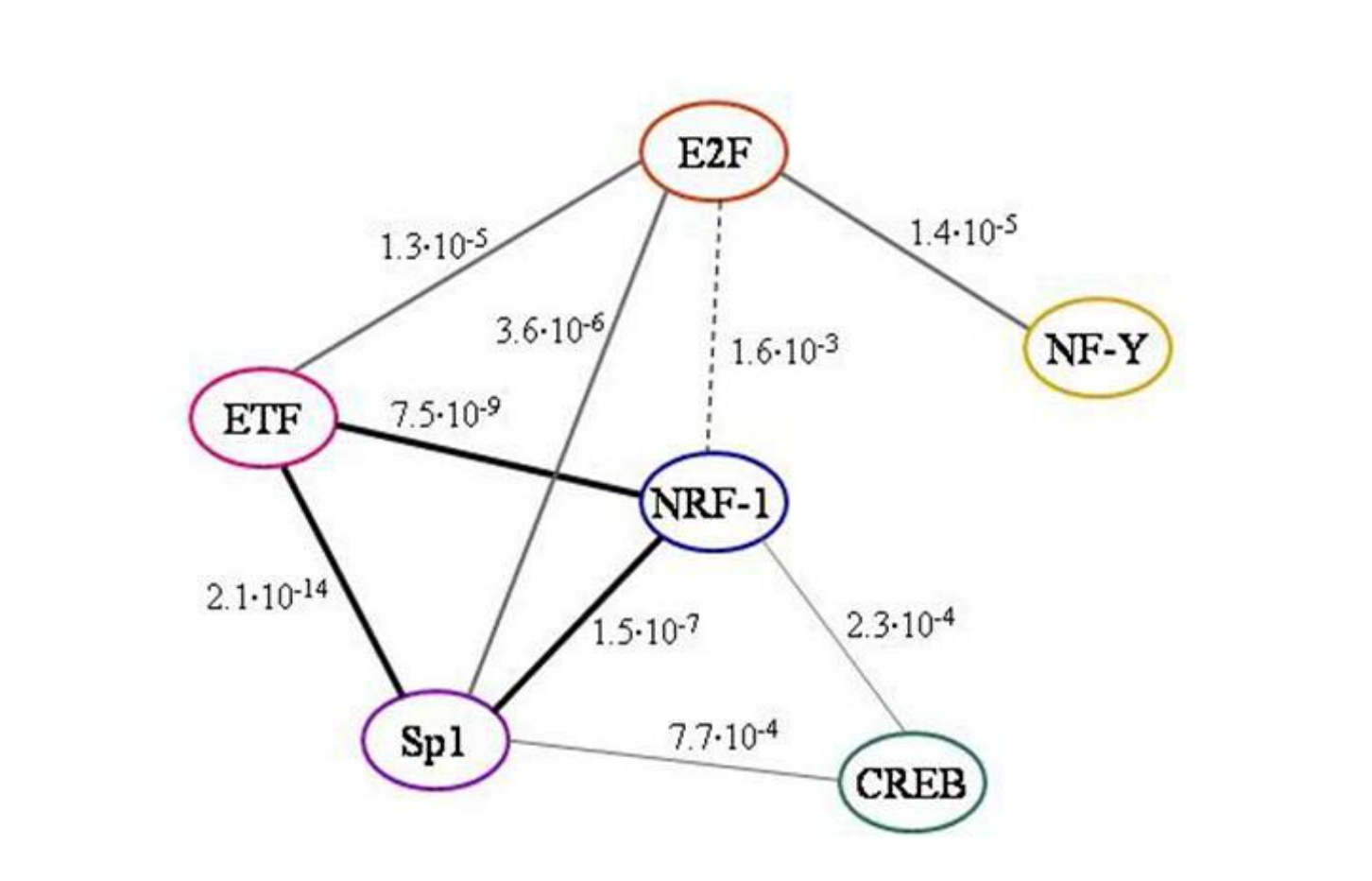
PRIMA: results on HCC

PRIMA found 8 enriched TFs in the 568 HCC genes (w.r.t. 13K BG promoters):



Results on HCC (III)

Co-occurring pairs of TFs:



How: Motif finding

Bailey & Elkan ZOOPS model

- n sequences, m possible motif positions per sequence.
- <u>Assumption:</u> each sequence contains zero or one occurrence of the motif.
- Prior probability for one occurrence – γ
- Prior probability for motif in position j – $\lambda = \gamma/m$
- *What is the hidden data?*
- *What is the Q function?*

Bailey & Elkan ZOOPS (cont.)

- Z_{ij} indicator for motif at sequence i, position j.
- Q_i indicator for motif at sequence i.

$$
\begin{aligned}
&\log Pr(X, Z|\theta, \gamma) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j} \log Pr(X_i|Z_{i,j} = 1, \theta) \\
&\quad + \sum_{i=1}^{n}(1 - Q_i) \log Pr(X_i|Q_i = 0, \theta) \\
&\quad + \sum_{i=1}^{n}(1 - Q_i) \log(1 - \gamma) + \sum_{i=1}^{n} Q_i \log \lambda
\end{aligned}
$$

$$
Z_{i,j}^{(t)} = \frac{f_i}{f_0 + \sum_{k=1}^{m} f_k}, \text{ where}
$$

$$
\begin{aligned}
f_0 &= Pr(X_i|Q_i = 0, \theta^{(t)})(1 - \gamma^{(t)}), \text{ and} \\
f_j &= Pr(X_i|Z_{i,j} = 1, \theta^{(t)})\lambda^{(t)}, \ 1 \le j \le m
\end{aligned}
$$