Computational Genomics
Prof. Ron Shamir & Prof. Roded Sharan
School of Computer Science, Tel Aviv University

גנומיקה חישובית
פרופ' רון שמיר ופרופ' רודד שרן
ביה"ס למדעי המחשב,אוניברסיטת תל אביב

Lecture 7: Gene Finding

26 Nov 2013

Gene Finding

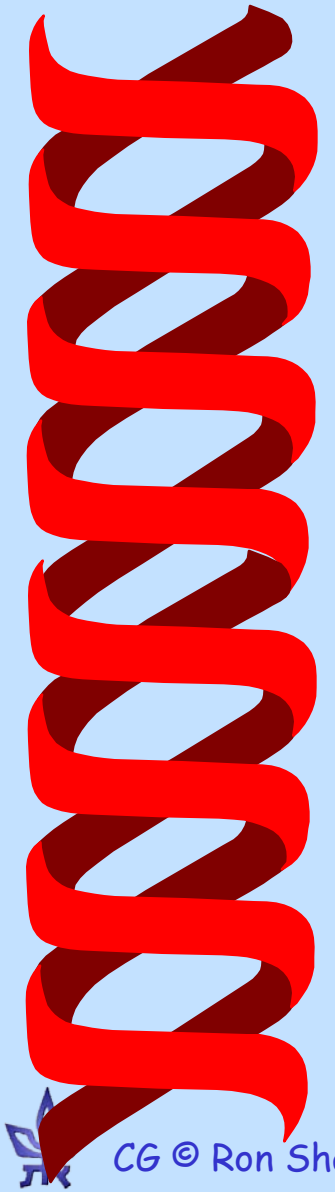
Sources:

- 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

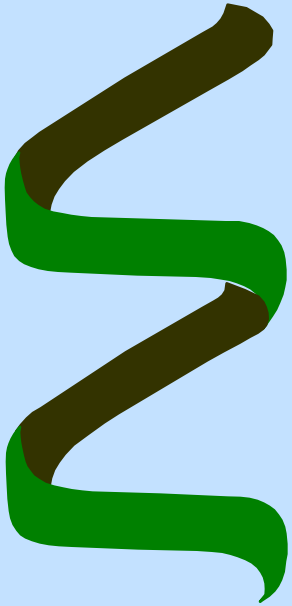
Motivation

- ~3Gb human DNA in GenBank
- Only ~1.5% of human DNA is coding for proteins
- 155,176,494,699 total bases in GenBank (10/13)
- Hundreds of species have been sequenced, thousands to follow
- Total number of species represented in UniProtKB/Swiss-Prot (11/13): 13,041
- Need to locate the genes!
- **Goal: Automatic finding of genes**

"The Central Dogma"

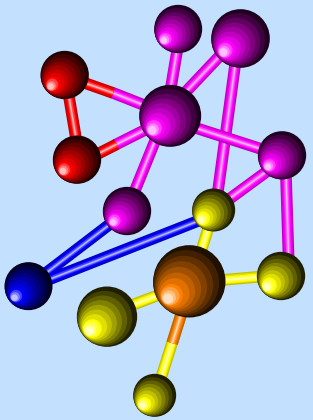


Transcription



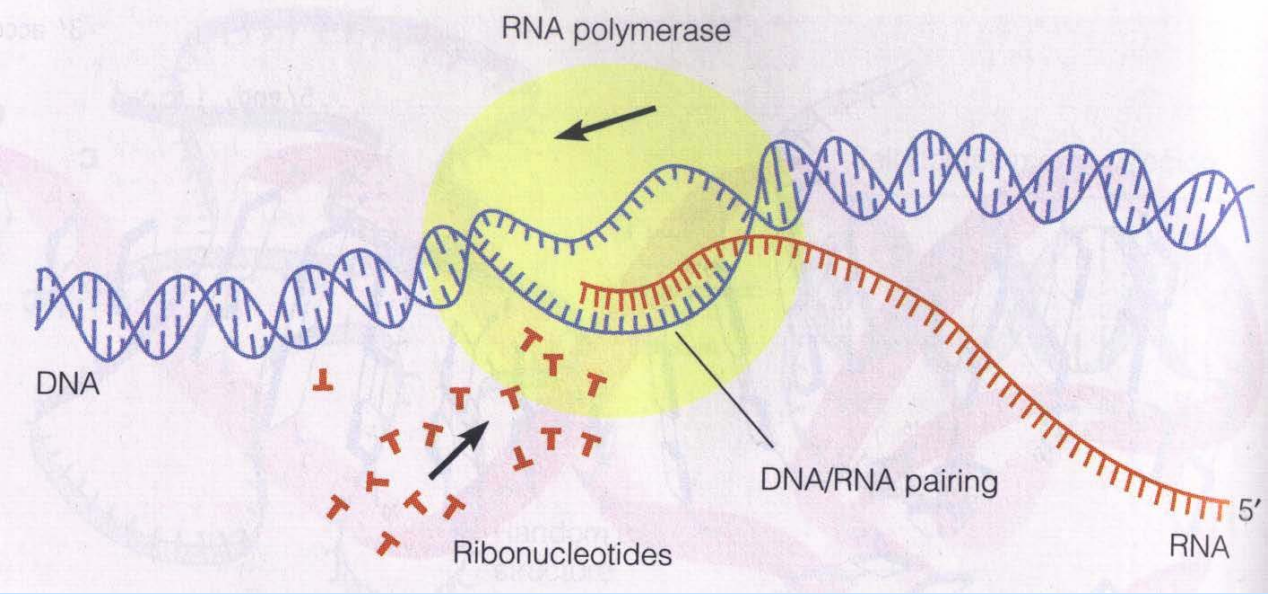
RNA

Translation

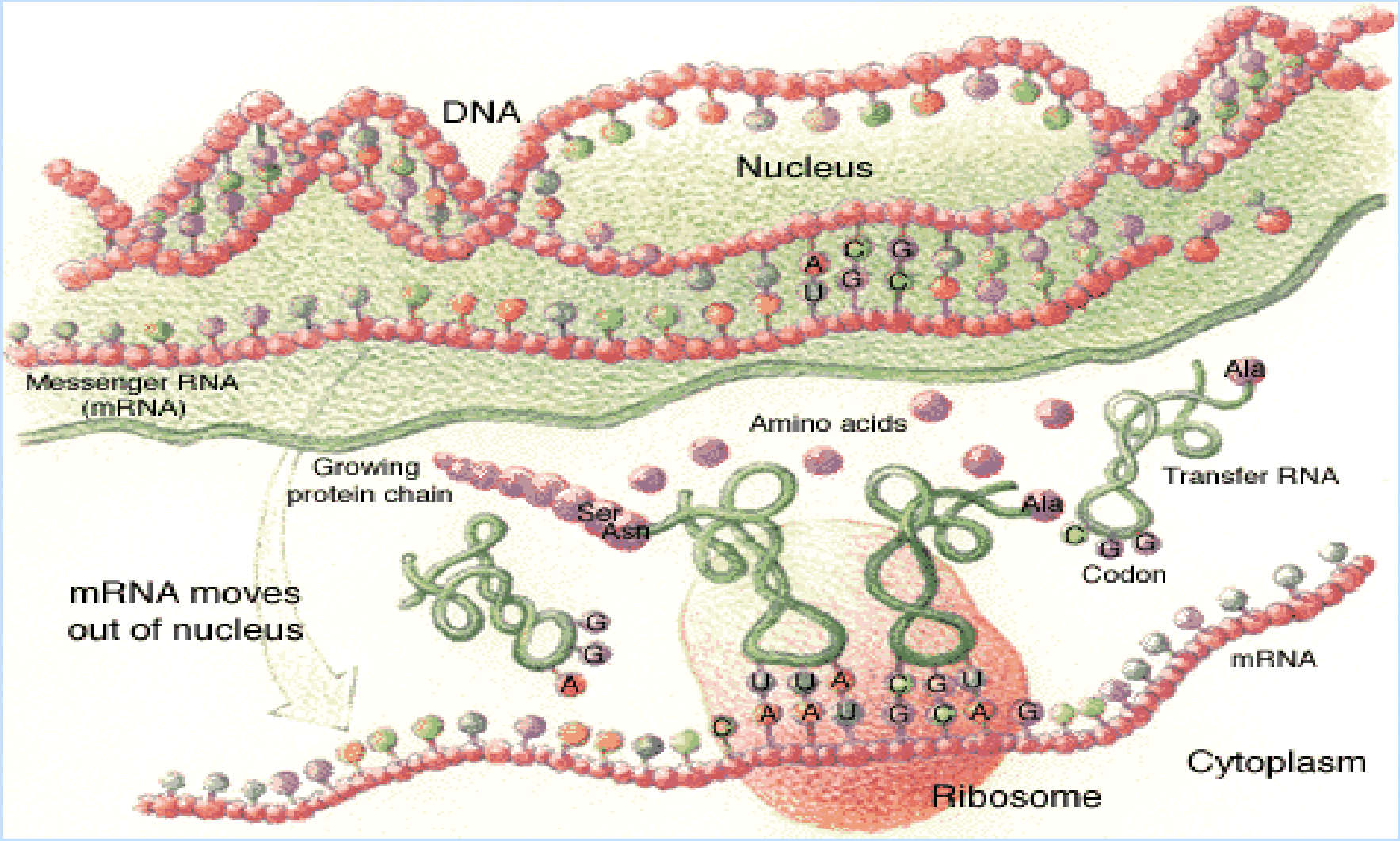


Protein

RNA Transcription

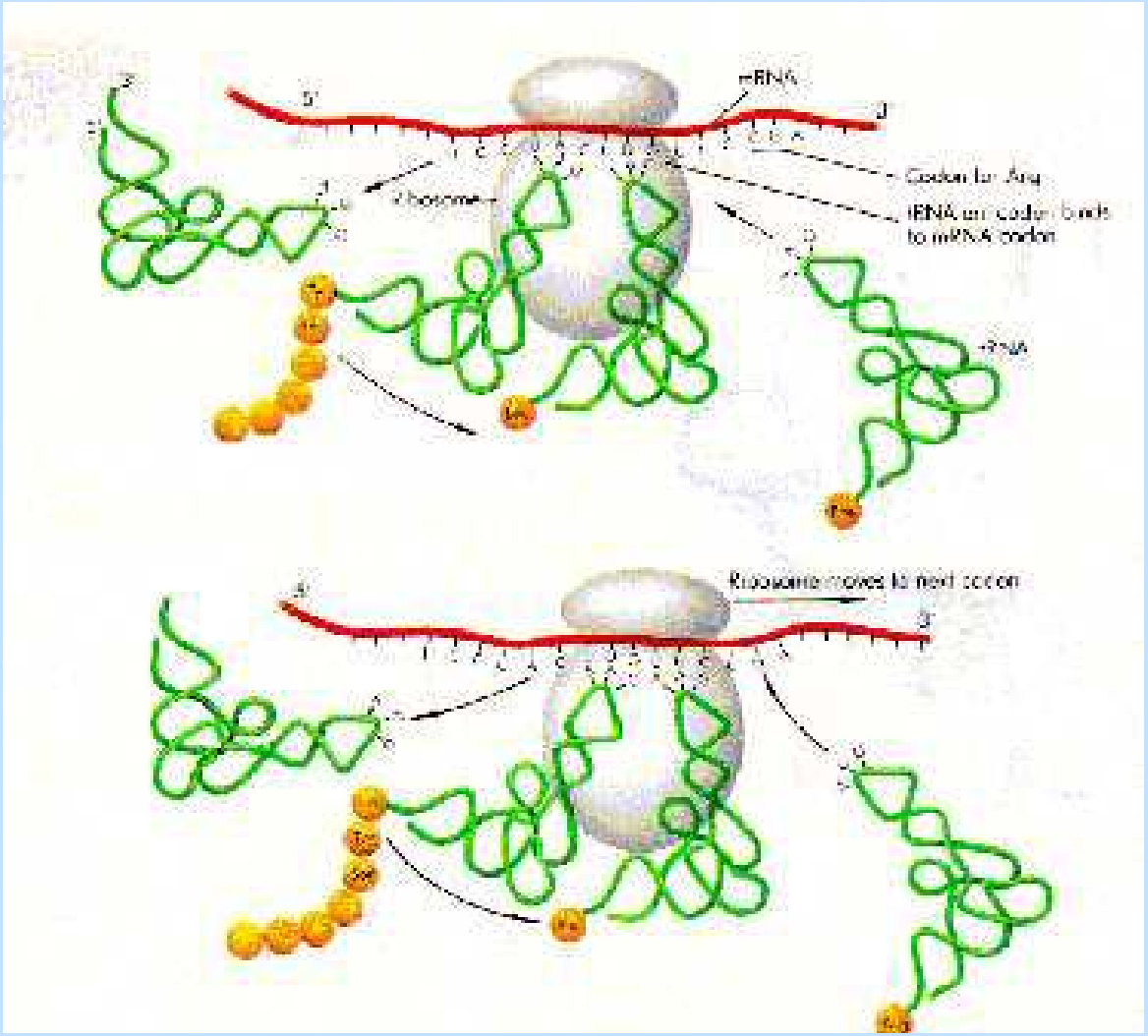


DNA→ RNA → Protein

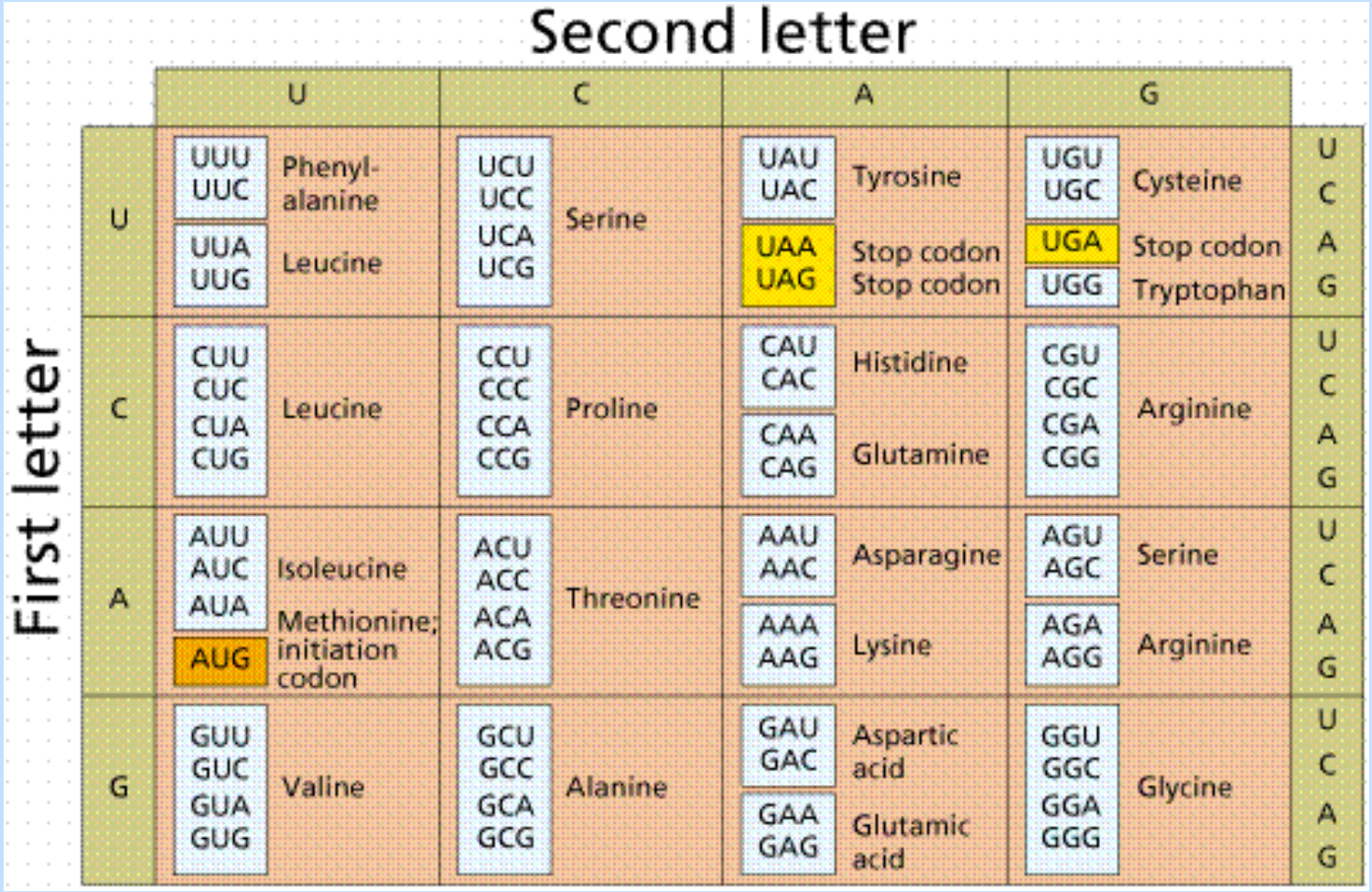


http://www.ornl.gov/hgmis/publicat/tko/index.htm

Ribosome

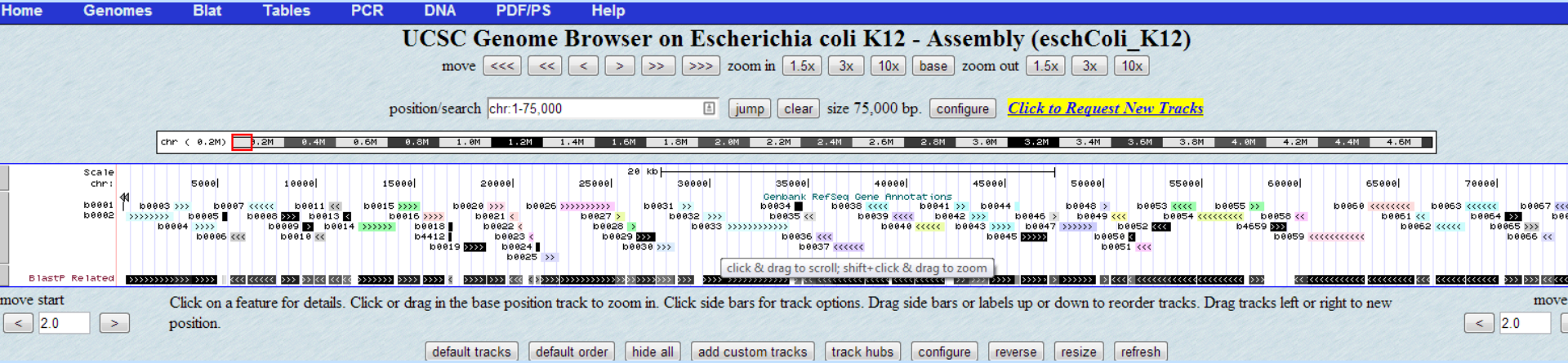

mRNA
5'
3'
Codon for Arg
tRNA anticodon binds to mRNA codon
Ribosome moves to next codon

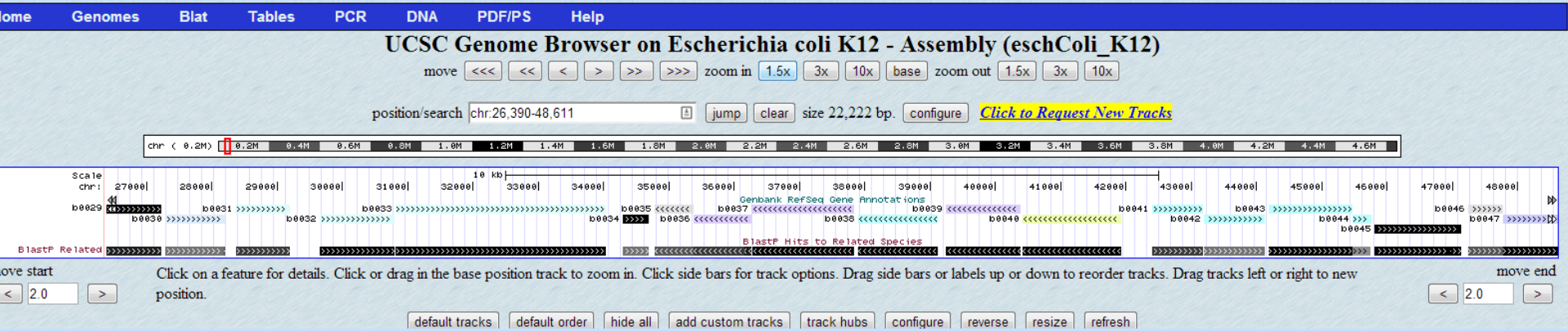
Reminder: The Genetic Code

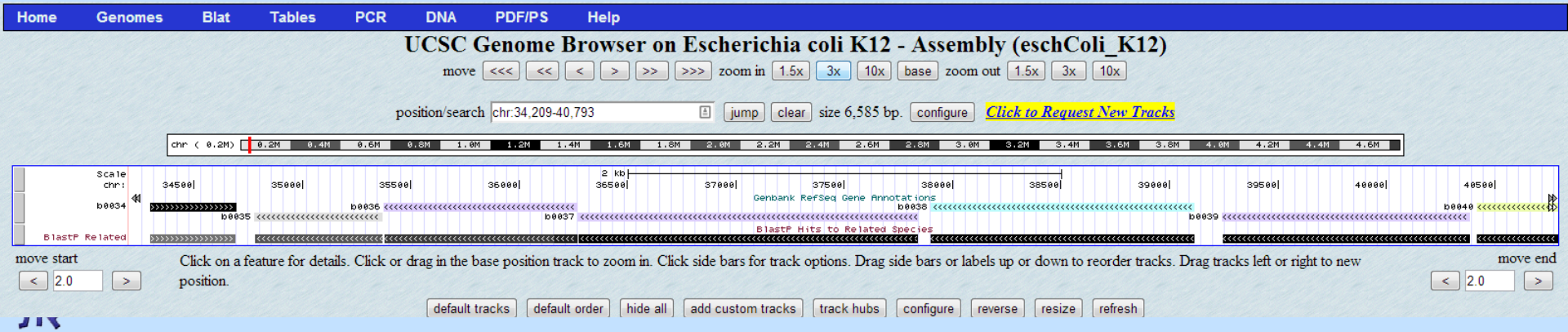


1 start, 3 stop codons

Gene Finding in Prokaryotes

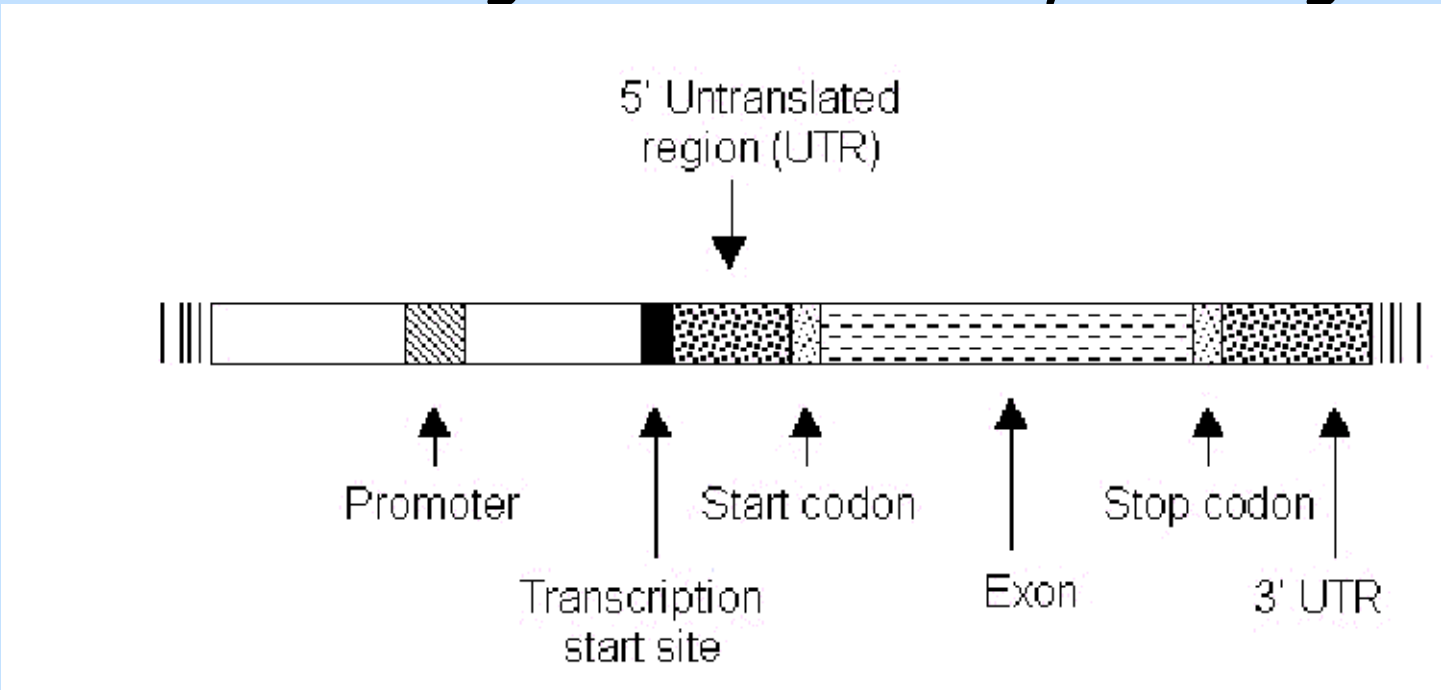

Home Genomes Blat Tables PCR DNA PDF/PS Help
UCSC Genome Browser on Escherichia coli K12 - Assembly (eschColi_K12)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr:1-75,000 jump clear size 75,000 bp. configure Click to Request New Tracks
Genbank RefSeq Gene Annotations
BlastP Related
click & drag to scroll; shift+click & drag to zoom
move start
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
move
default tracks default order hide all add custom tracks track hubs configure reverse resize refresh


Home Genomes Blat Tables PCR DNA PDF/PS Help
UCSC Genome Browser on Escherichia coli K12 - Assembly (eschColi_K12)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr:26,390-48,611 jump clear size 22,222 bp. configure Click to Request New Tracks
Genbank RefSeq Gene Annotations
BlastP Hits to Related Species
BlastP Related
move start
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
move end
default tracks default order hide all add custom tracks track hubs configure reverse resize refresh


Home Genomes Blat Tables PCR DNA PDF/PS Help
UCSC Genome Browser on Escherichia coli K12 - Assembly (eschColi_K12)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr:34,209-40,793 jump clear size 6,585 bp. configure Click to Request New Tracks
Genbank RefSeq Gene Annotations
BlastP Hits to Related Species
BlastP Related
move start
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
move end
default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

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

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

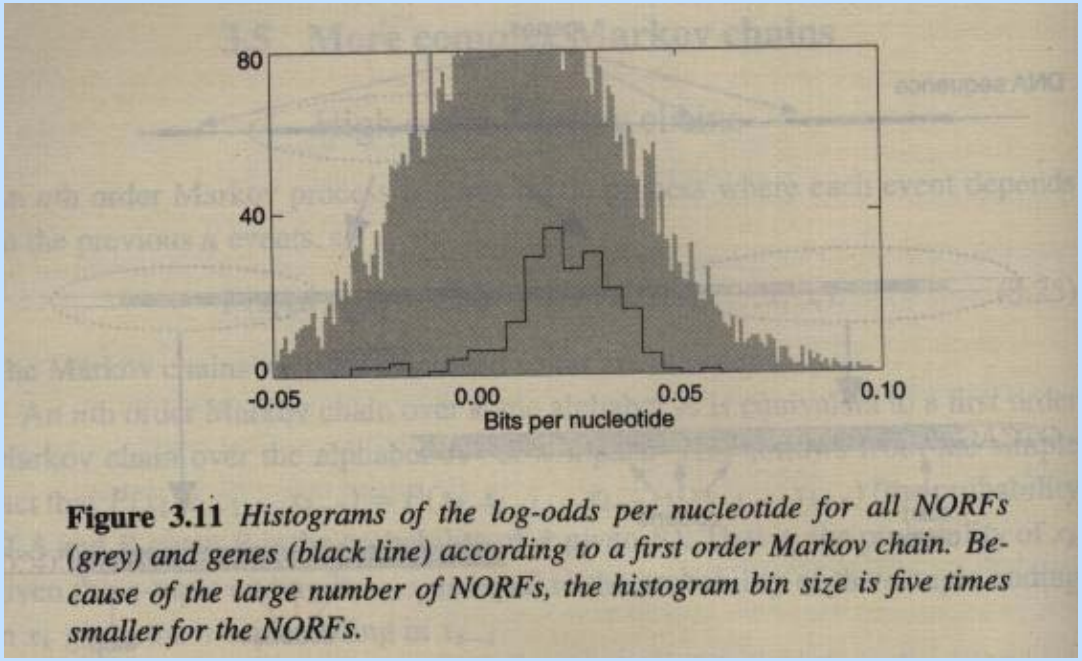


Figure 3.11 *Histograms of the log-odds per nucleotide for all NORFs (grey) and genes (black line) according to a first order Markov chain. Because of the large number of NORFs, the histogram bin size is five times smaller for the NORFs.*

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

Second Order Markov Chains

Assumption:

- X_{i+1} is independent of the past once we know X_i and X_{i-1}
- This allows us to write:

$$P(X_1,\ldots,X_n) = P(X_1)\prod_i P(X_{i+1} \mid X_1,\ldots,X_i)$$

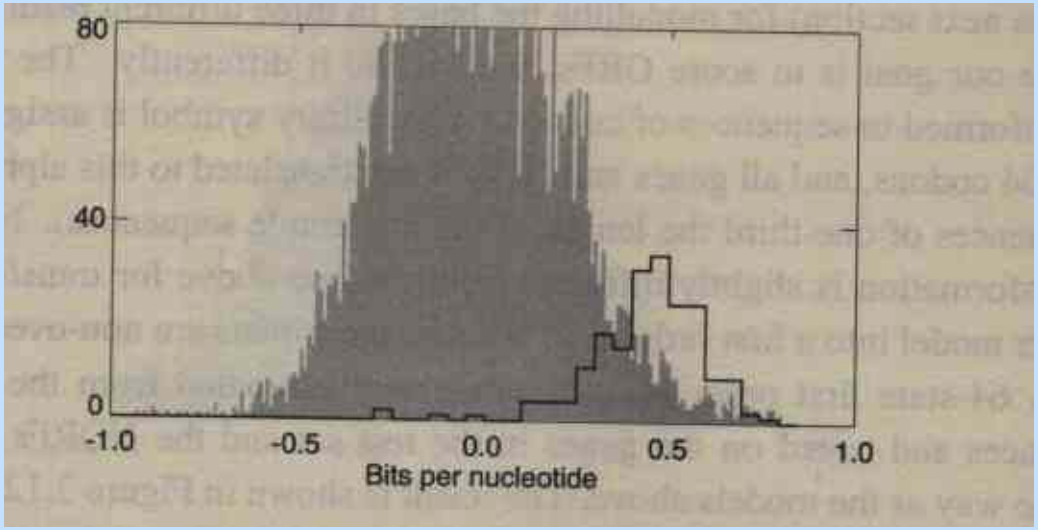
$$= P(X_1)p(X_2 \mid X_1)\prod_i P(X_{i+1} \mid X_{i-1}, X_i)$$

- Results are similar to the first order Markov chain

➔ Idea: work with codons

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

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

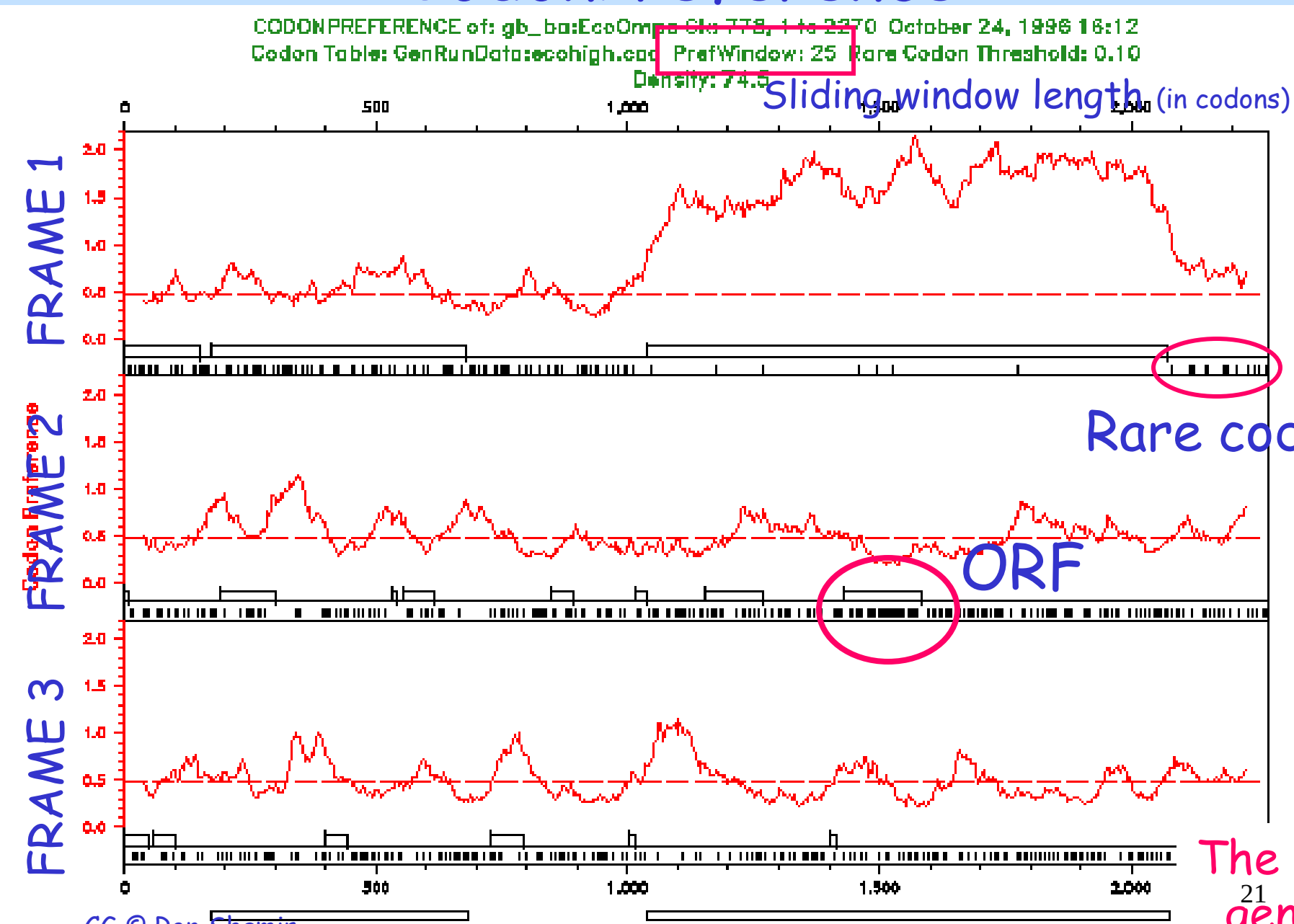
$$p_2 = f_{b_1c_1a_2} * f_{b_2c_2a_3} * \ldots * f_{b_nc_na_{n+1}}$$

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

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

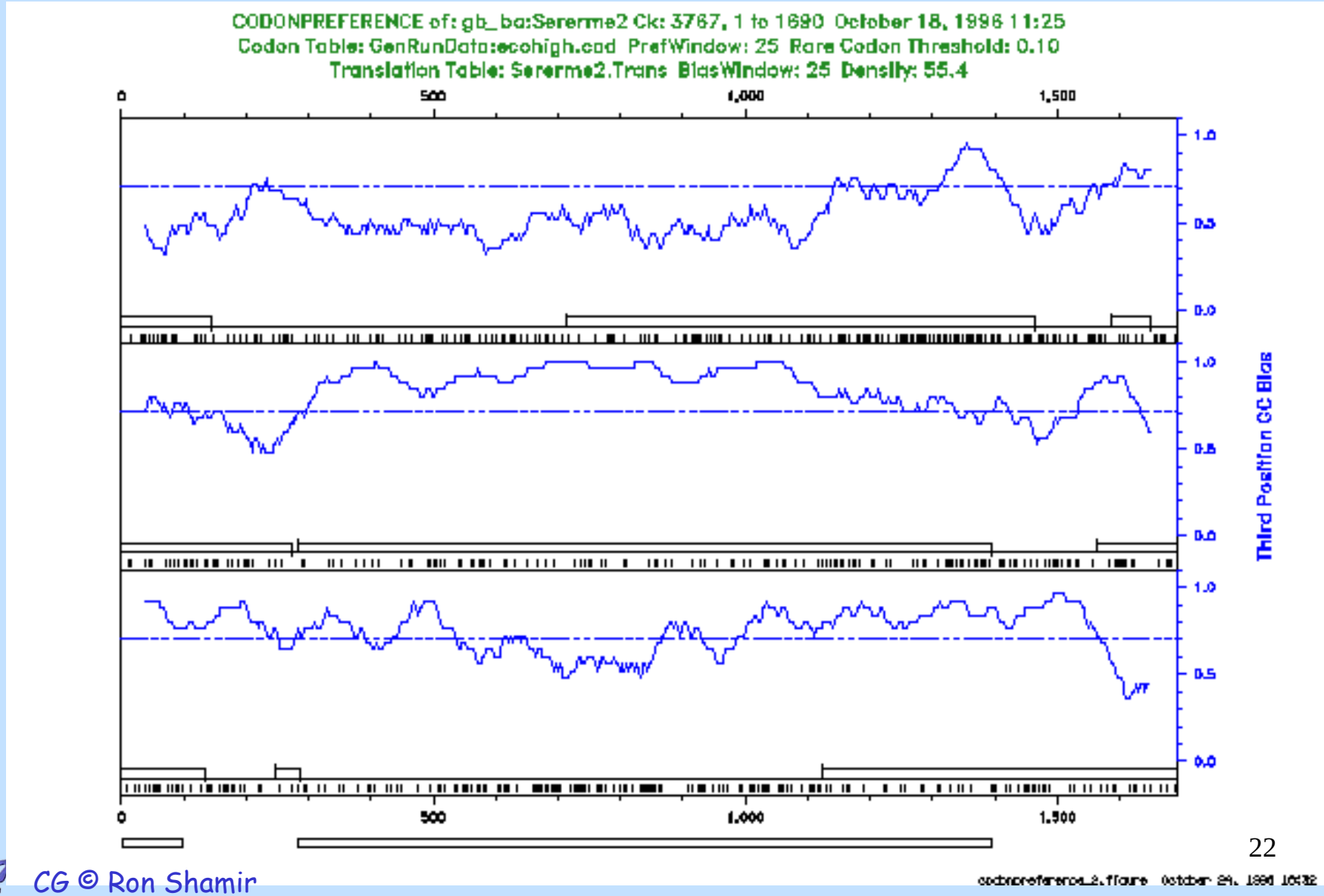
$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

CodonPreference



CG © Ron Shamir

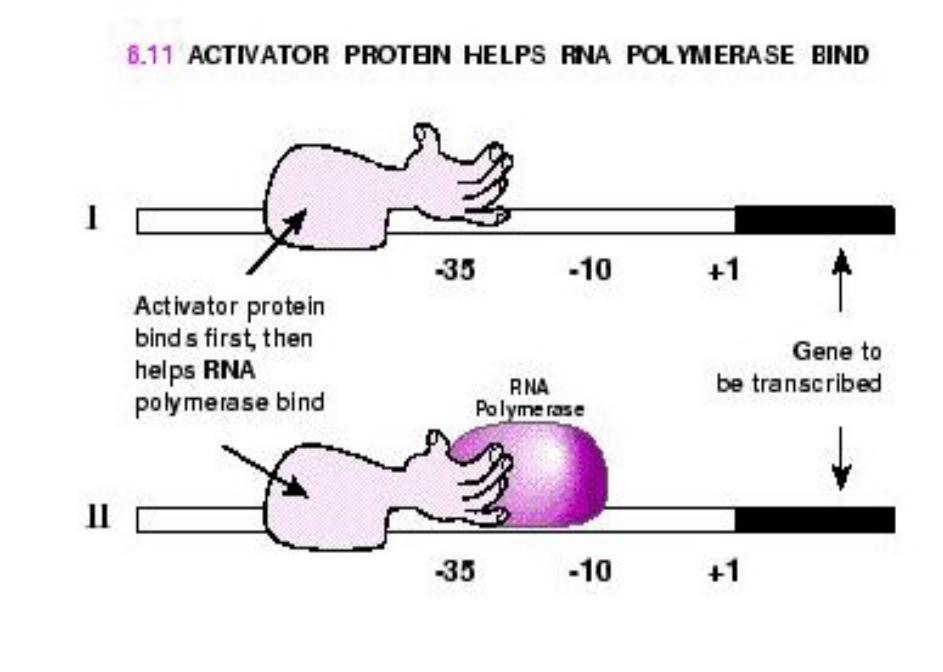
CodonPreference: 3rd position GC bias

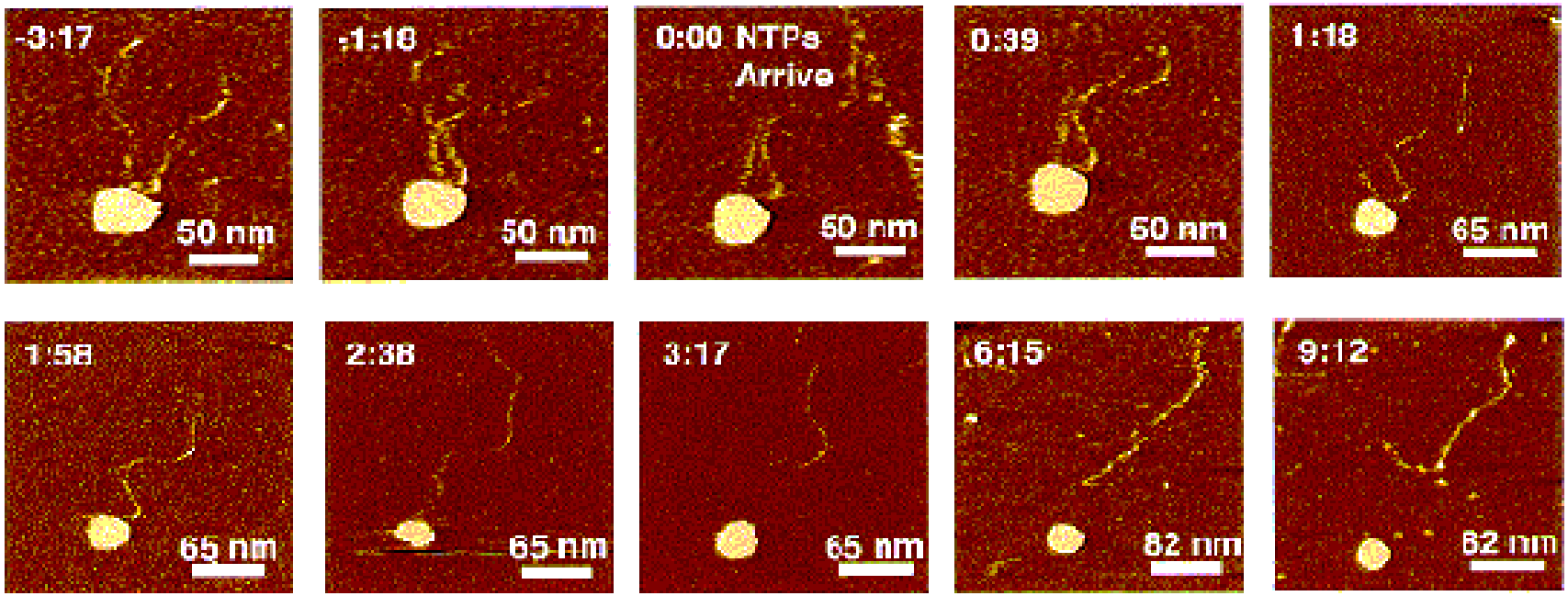


RNA Transcription

- Not all ORFs are expressed.
- Transcription depends on regulatory regions
- Common regulatory region – the promoter
- RNA polymerase binds tightly to a specific DNA sequence in the promoter called the binding site.
- "Anchor" point, pinpoints where RNA transcription should begin.
- At the termination signal the polymerase releases the RNA and disconnects from the DNA.

TF binding to the promoter





DNA being transcribed by the enzyme RNA polymerase. The enzyme (white spot) binds to the DNA (thin line) After the NTP molecules arrive in the third picture on the top row, the enzyme starts to move along the DNA . As the enzyme moves along the DNA, it uses the NTPs to make RNA (not visible) until it comes to the end of the DNA and falls off in the bottom row of pictures. The DNA continually wiggles around, as you can see from the pictures.

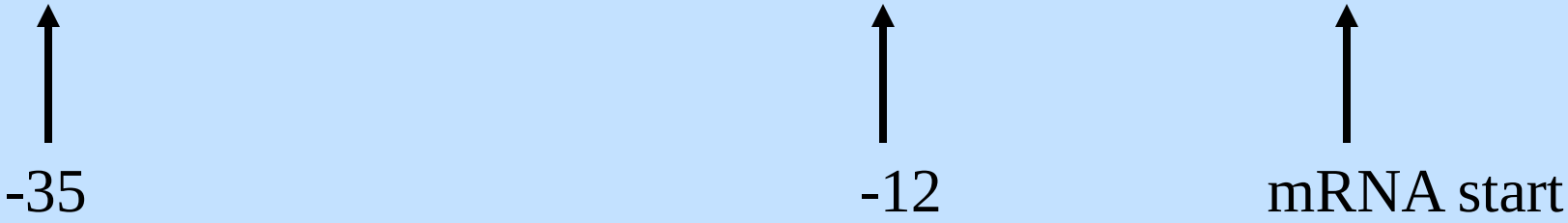
Kasas, et al 1997. *Biochemistry.* 36:461-468.

www.physics.ucsb.edu/~hhansma/ afm-acs_news.htm

E. coli promoters

consensus sequence:

nnnTTGACAnnnnnnnnnnnnnnnnnnTATAATnnnnnnNnnn



- "TATA box" (or Pribnow Box)
- Not exact
- Other common features.

Positional Weight Matrix

- $f_{b,j}$: 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 $S=B_1B_2B_3B_4B_5B_6$

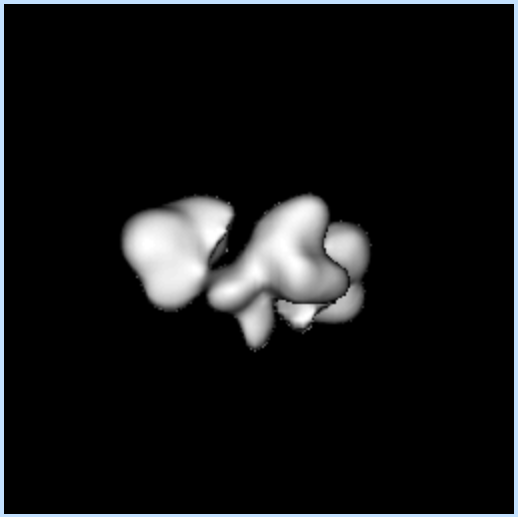
$$P(S \mid \text{promoter}) = \prod_{i=1}^{6} f_{B_i,i}$$

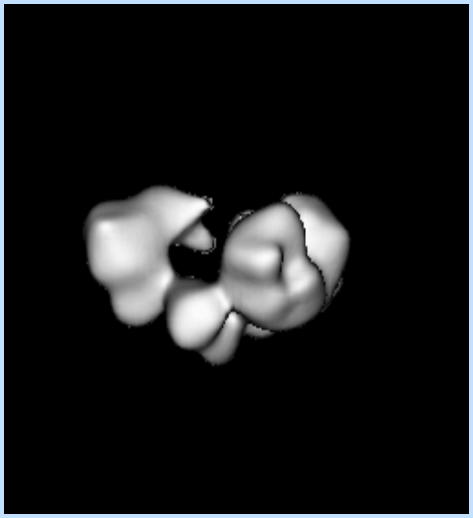
$$P(S \mid \text{non-promoter}) = \prod_{i=1}^{6} f_{B_i}$$

- Log-likelihood ratio score:

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

- Experiments show ~80% correlation of score to measured binding energy





3D reconstructions of TFIID at 35 and 30 Angstroms resolution.

TFIID

The transcription factor TFIID is localized within the nucleus of the cell and, along with other basal transcription factors, is primarily responsible for showing RNA polymerase the start of a transcription site by binding to the DNA TATA box upstream of a gene.

cryoem.berkeley.edu/ ~fandel/TFIID.html

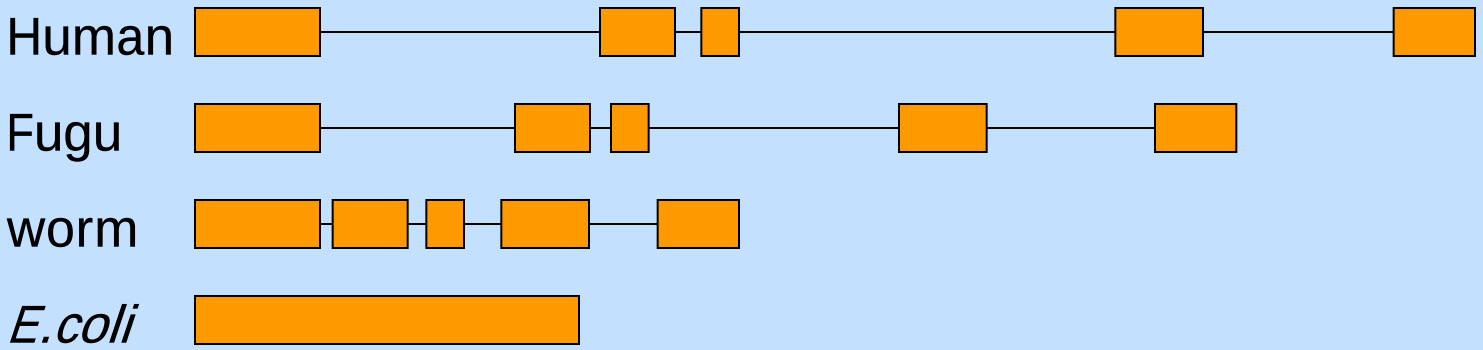
Promoter Variation

- Why do promoters vary?
  - ???
  - Specificity of promoters is responsible for transcription level: the closer the sequence to the consensus, the higher
  - This allows a 1000 fold difference between genes transcription levels.

➔ finding regulatory sequences is an inherently stochastic problem – and a hard one.

Gene finding: coding density

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

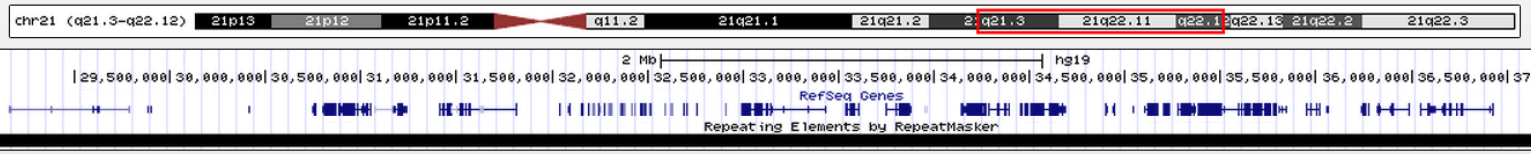


Gene Finding in Eukaryotes

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr21:28,580,448-37,557,047 8,976,600 bp. enter position, gene symbol or search terms go



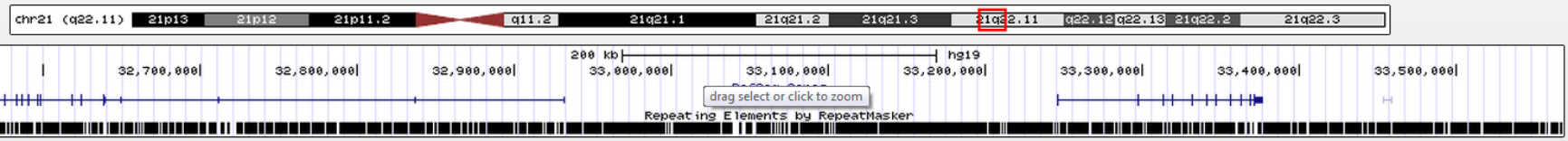
move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 2.0 >

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

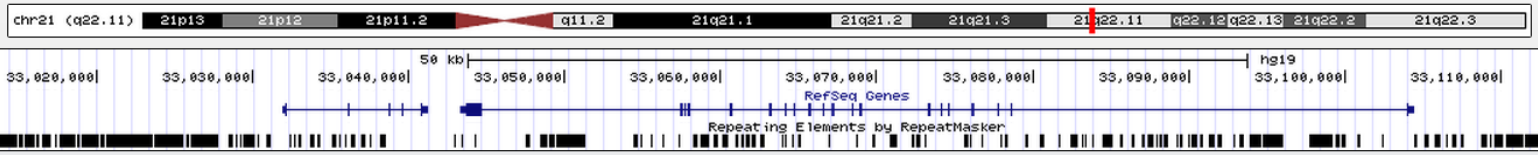
chr21:32,570,048-33,567,447 997,400 bp. enter position, gene symbol or search terms go



move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 2.0 >

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

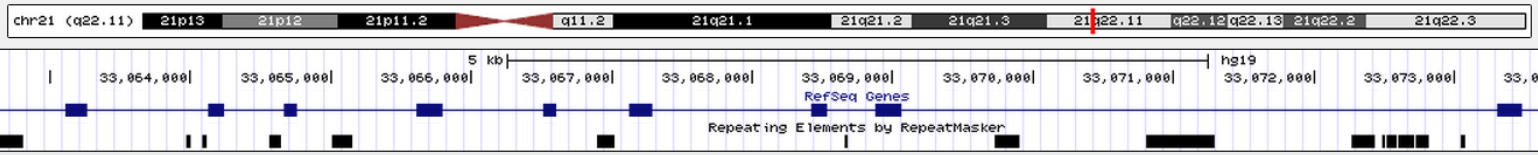
chr21:33,013,337-33,124,158 110,822 bp. enter position, gene symbol or search terms go



move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 2.0 >

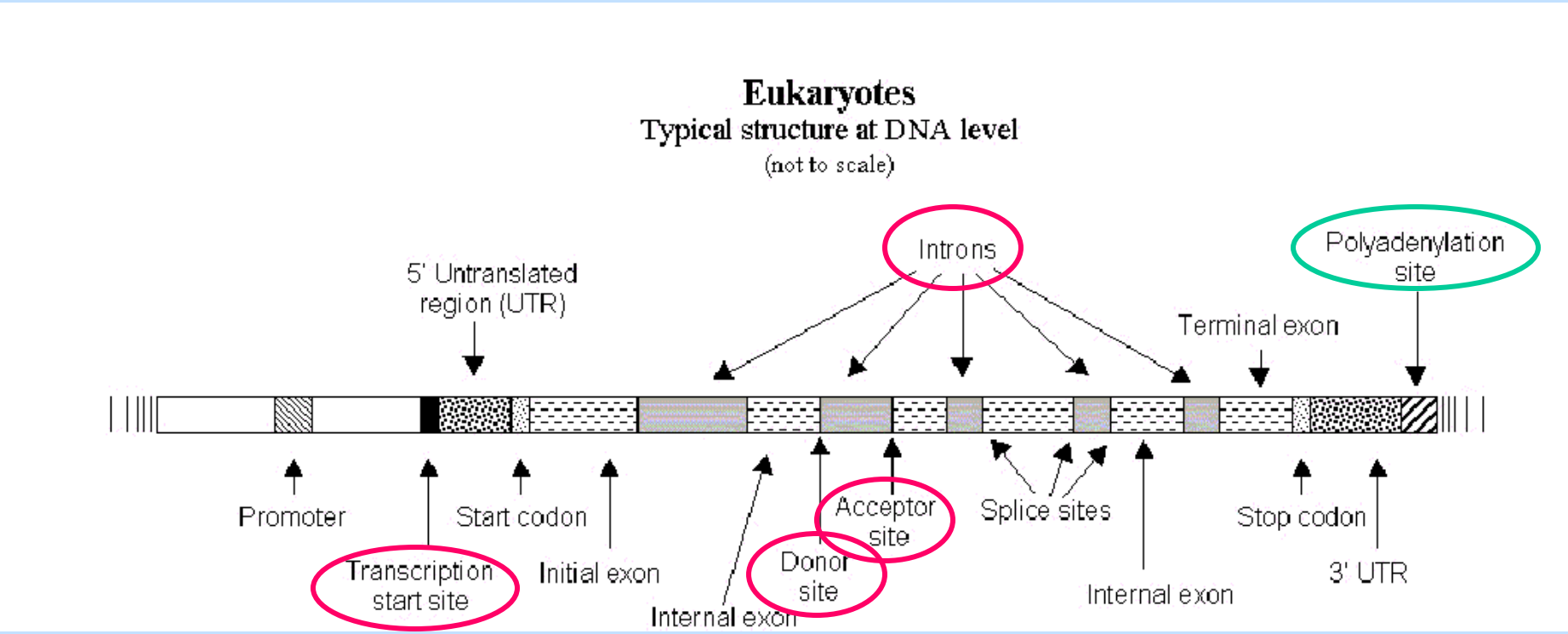
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr21:33,062,591-33,074,904 12,314 bp. enter position, gene symbol or search terms go

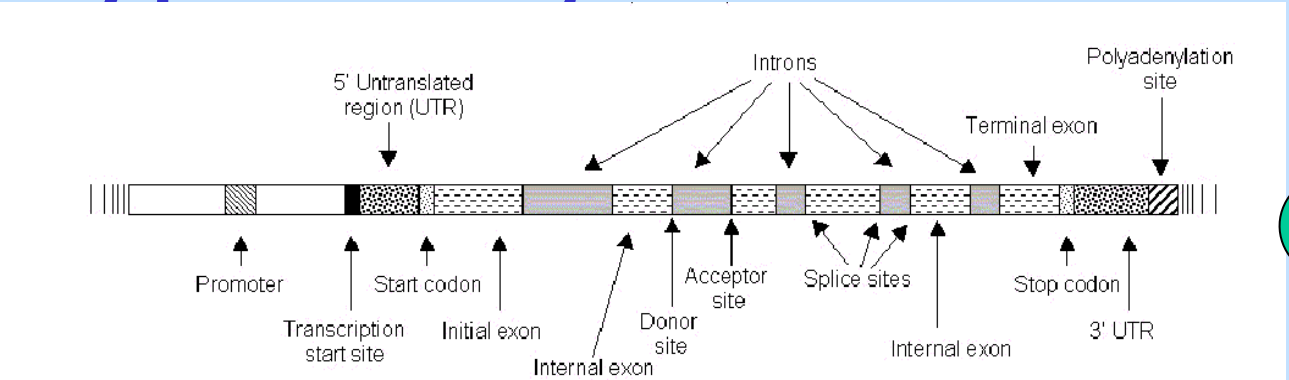


move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 2.0 >

Eukaryote gene structure



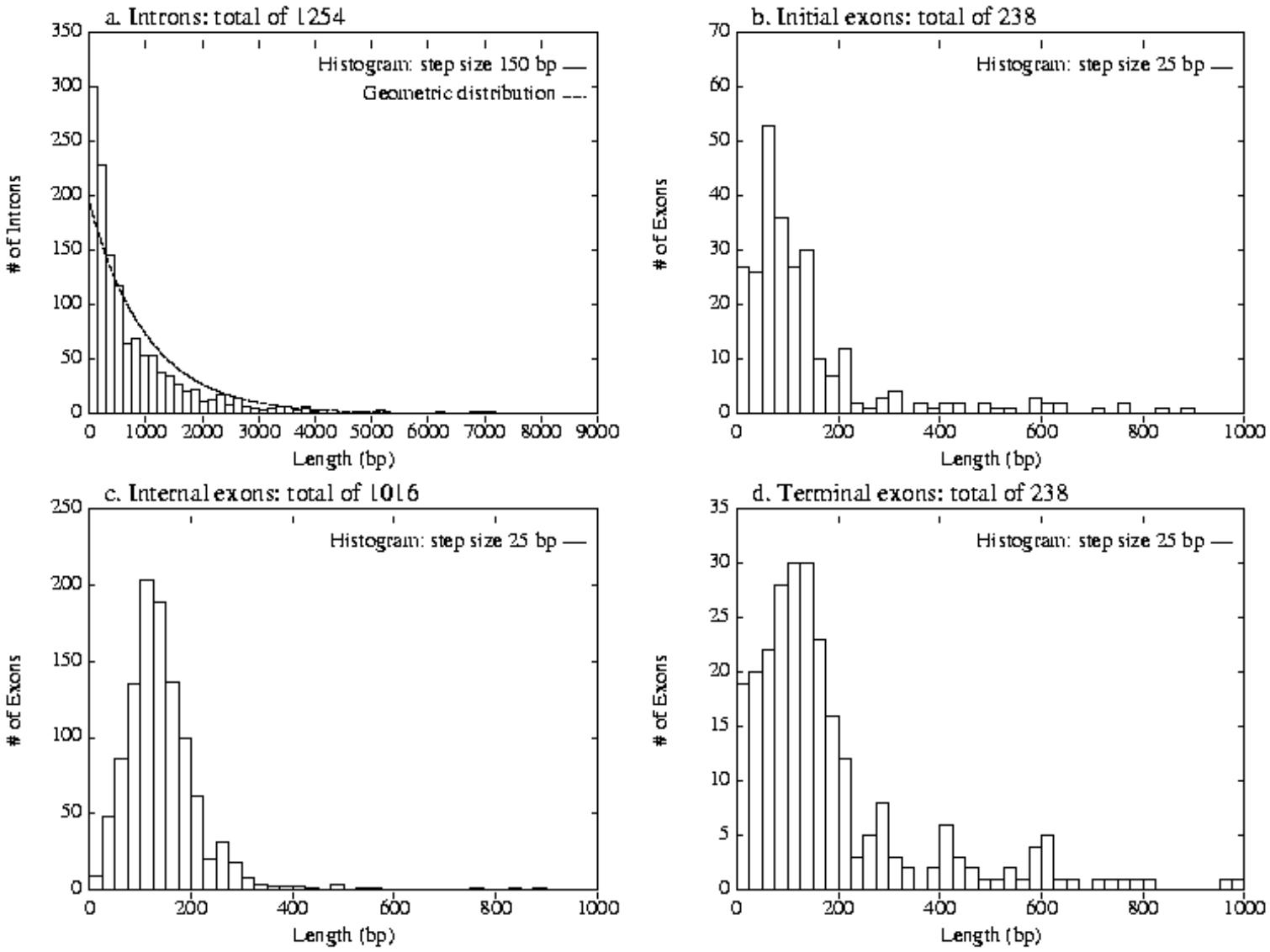
Typical figures: verterbrates



- Transcription rate: <50 b/sec
- Splicing rate: minutes

- 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

Fig. 1. Length distributions of introns and exons in human genes

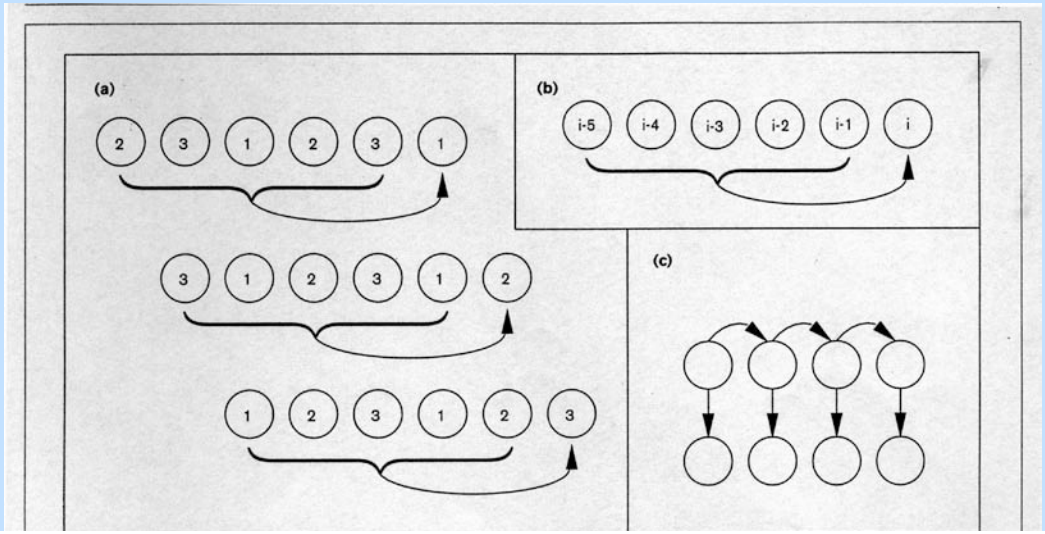


Legend. Intron, exon length data from 238 multi-exon genes of GENSCAN learning set (Appendix A).

Burge & Karlin JMB 97

Markov Sequence Models

- Key: distinguish coding/non-coding statistics
- Popular models:
  - 6-mers (5th order Markov Model)
  - Homogeneous/non-homogeneous (reading frame specific)

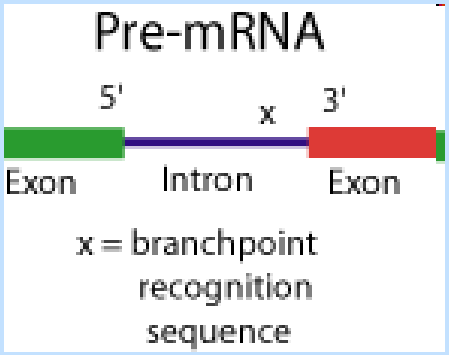


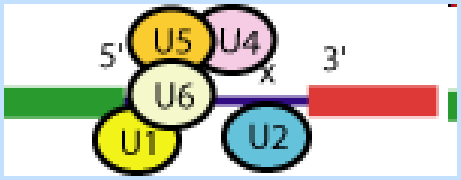
Not sensitive enough for eukaryote genes: exons too short, poor detection of splice junctions

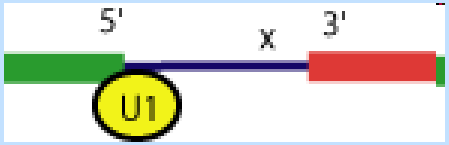
Splicing

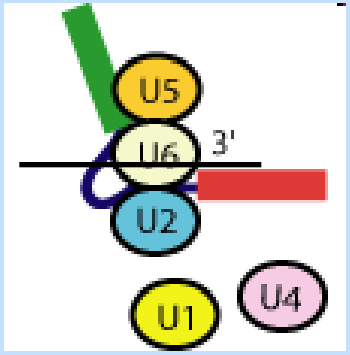
- Splicing: the removal of the introns.
- Performed by complexes called spliceosomes, containing both proteins and snRNA.
- The snRNA recognizes the splice sites through RNA-RNA base-pairing
- Recognition must be precise: a 1nt error can shift the reading frame making nonsense of its message.
- Many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

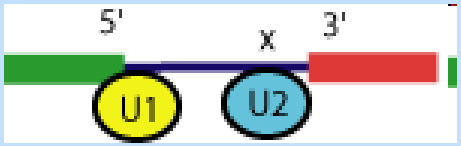
Spliceosome - path

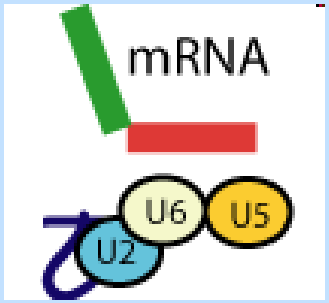






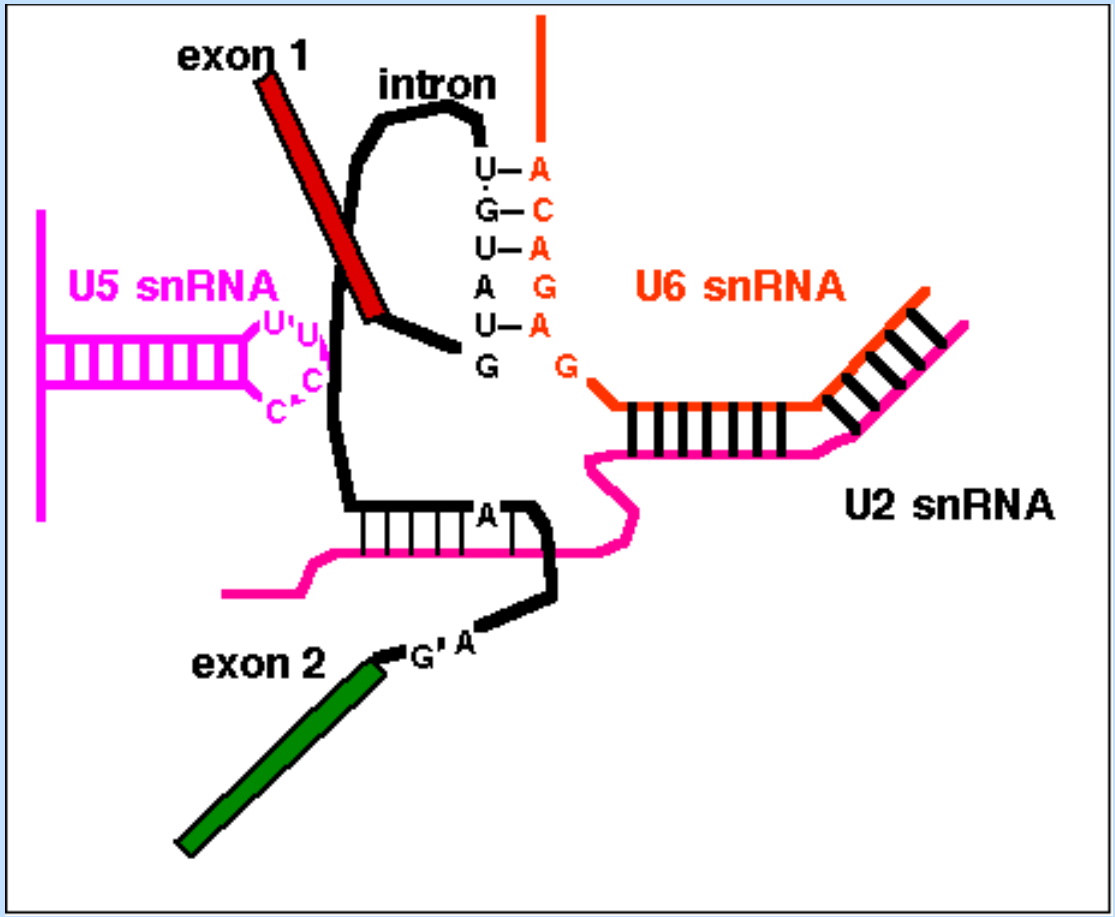






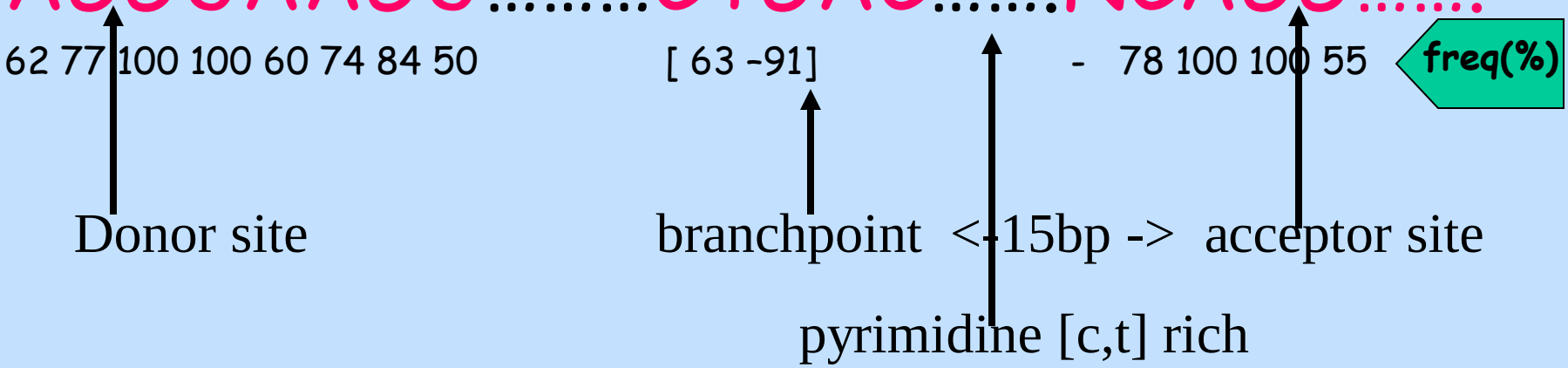
http://www.neuro.wustl.edu/neuromuscular/pathol/diagrams/splicefunct.html

Spliceosome - mechanism



http://www.neuro.wustl.edu/neuromuscular/pathol/diagrams/splicemech .html

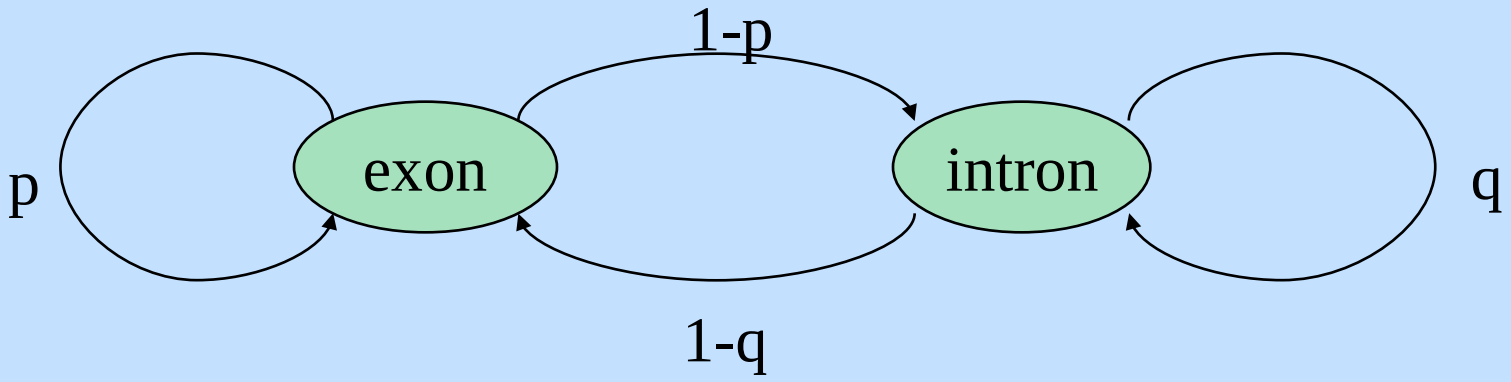
Exon-intron junctions



- 1st approach: position specific 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.

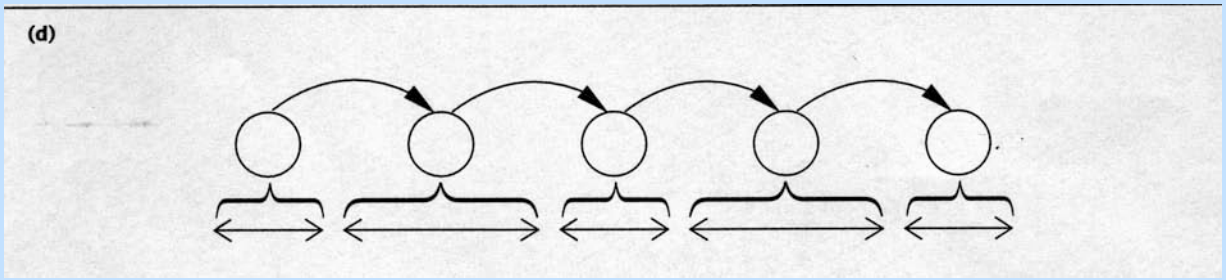
Exon Length Distribution

- 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.

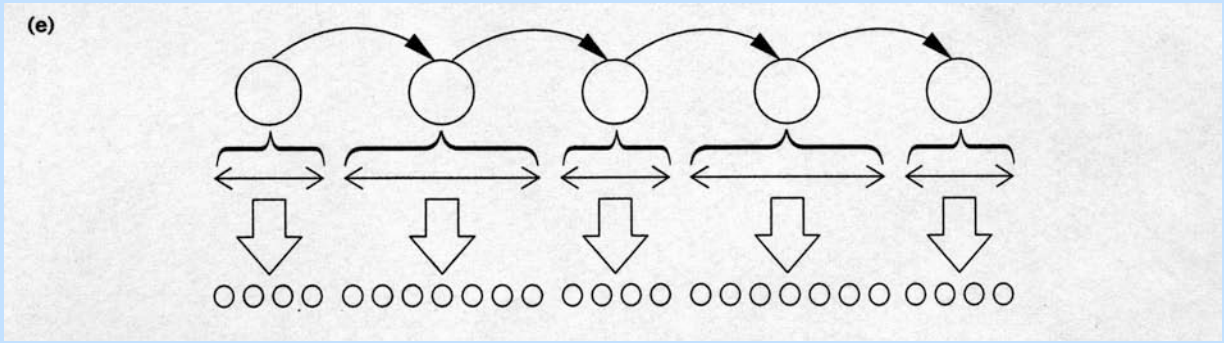
➔ Need a different model for exons.

Generalized HMM

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



- Semi-Markov model with different output length at each node



- HMM with different output length and different output distribution at each node

Generalized HMM

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

- Overview:
  - 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 model, HMM…)
  - Initial state probability distribution π
  - State transition probabilities T_{ij}

GenScan Model

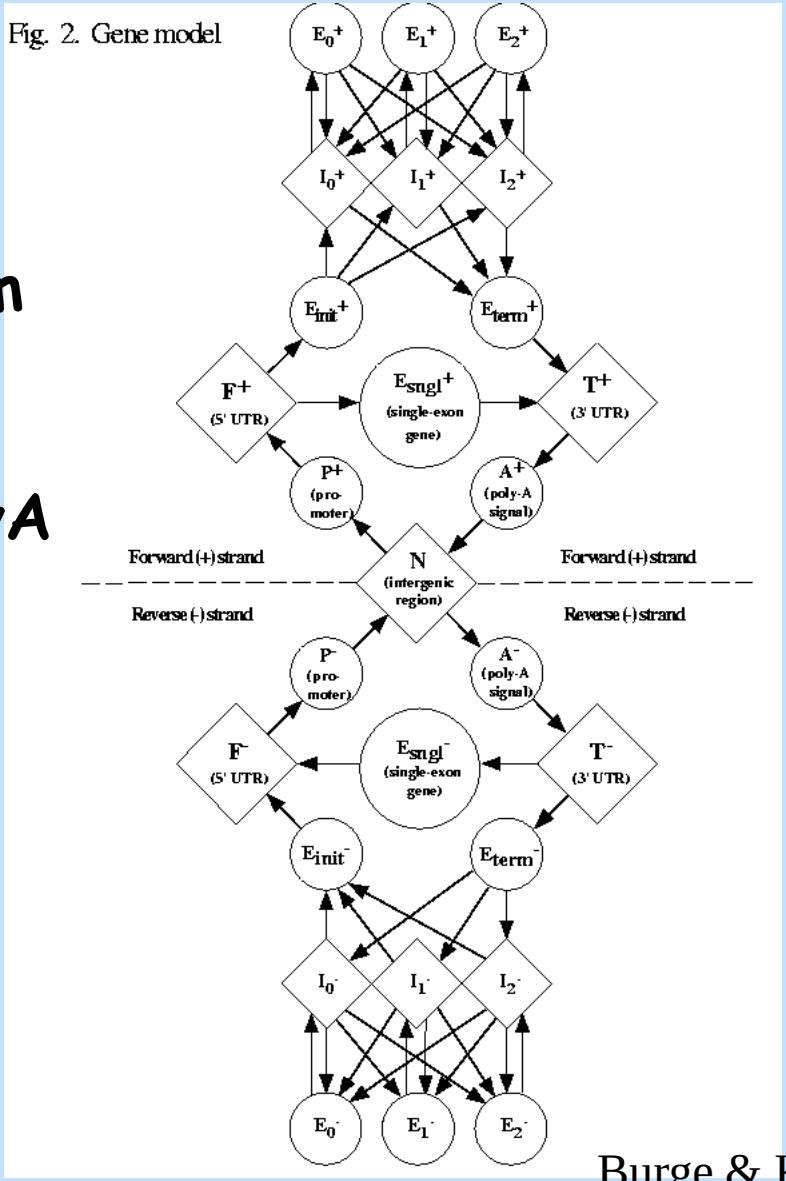
Exon

Intron

Exon init/term

5'/3' UTR

Promoter/PolyA

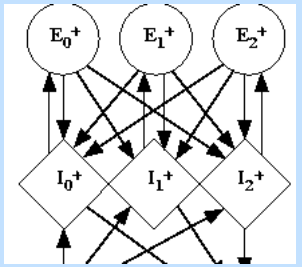


Forward strand

Backward strand

Burge & Karlin JMB 97

GenScan model

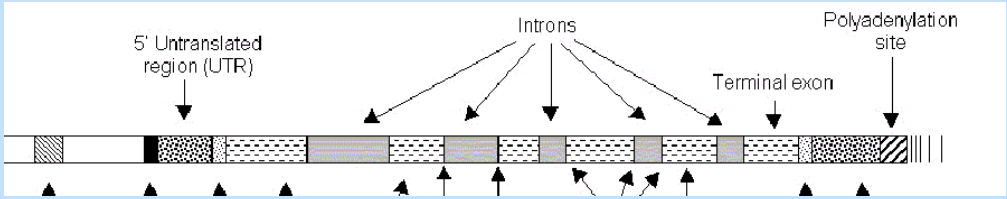


- states = functional units on a gene
- The allowed transitions ensure the order is biologically consistent.
- As an intron may cut a codon, one must keep track of the reading frame, hence the three *I* phases:
  - phase I_0: between codons
  - phase I_1: introns that start after 1st base
  - phase I_2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

Prediction

- A parse Φ of a sequence S with $|S|=L$: ordered sequence of states $(q_1,\ldots,q_t)$; associated durations d_i for each state.

$$\sum_{i=1}^{t} d_i = L$$



- Parse = annotation
- Given a parse Φ and a sequence S:
  - the probability the model went through states Φ to create S is:

$$P(\Phi,S) = \pi_{q_1} f_{q_1}(d_1) P_{q_1}(S_1 \mid d_1) \prod_{k=2}^{t} T_{q_{k-1}q_k} f_{q_k}(d_k) P_{q_k}(S_k \mid d_k)$$

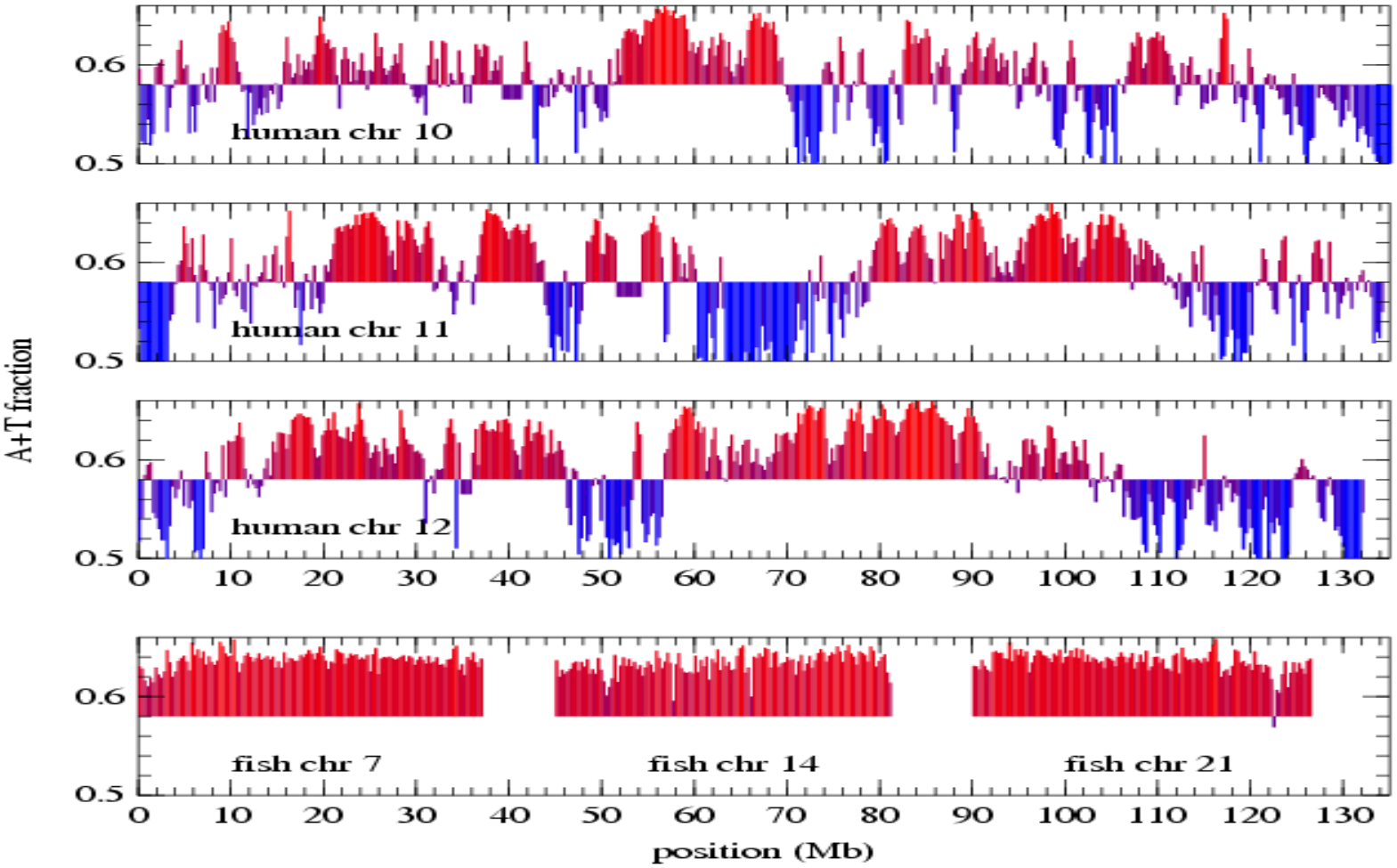
Prediction

- probability of a specific parse given the sequence:

$$P(\Phi \mid S) = \frac{P(\Phi, S)}{P(S)} = \frac{P(\Phi, S)}{\sum_{\Phi_i \text{ is a parse of length } L} P(\Phi_i, S)}$$

  - Can compute Φ_{opt} by Viterbi-like algorithm.
  - *Can compute P(S)* by forward-like alg.

C+G Content variability



www.nr.com/bio/**Isochores**and**Genes**Ver4.ppt

C+G Content

- C+G content ("isochore") has strong effect on gene density, gene length etc.
  - < 43% C+G : 62% of genome, 34% of genes
  - >57% C+G : 3-5% of genome, 28% of genes
- Gene density in C+G rich regions is 5 times higher than moderate C+G regions and 10 times higher than rich A+T regions
  - Amount of intronic DNA is 3 times higher for A+T rich regions. (Both intron length and number).
  - Etc…

C+G Content statistics

Table 3. Gene density and structure as a function of C + G composition: derivation of initial and transitio probabilities

Group	I	II	III	IV
C + G% range	<43	43-51	51-57	>57
Number of genes	65	115	99	101
Est. proportion single-exon genes	0.16	0.19	0.23	0.16
Codelen: single-exon genes (bp)	1130	1251	1304	1137
Codelen: multi-exon genes (bp)	902	908	1118	1165
Introns per multi-exon gene	5.1	4.9	5.5	5.6
Mean intron length (bp)	2069	1086	801	518
Est. mean transcript length (bp)	10866	6504	5781	4833
Isochore	L1 + L2	H1 + H2	H3	H3
DNA amount in genome (Mb)	2074	1054	102	68
Estimated gene number	22100	24700	9100	9100
Est. mean intergenic length	83000	36000	5400	2600
Initial probabilities:				
Intergenic (N)	0.892	0.867	0.540	0.418
Intron (I_0^+, I_1^+, I_2^+, I_0^-, I_1^-, I_2^-)	0.095	0.103	0.338	0.388
5′ Untranslated region (F^+, F^-)	0.008	0.018	0.077	0.122
3′ Untranslated region (T^+, T^-)	0.005	0.011	0.045	0.072

Estimates by Duret et al. 95

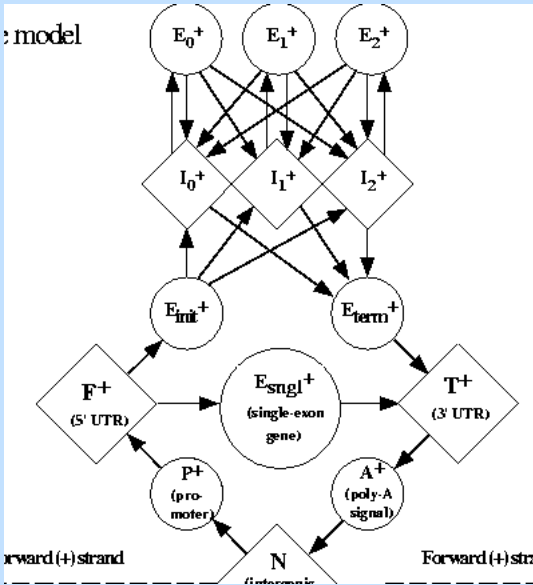
Burge & Karlin JMB 97

Handling diverse C+G Content

- The training set was divided into 4 categories:
  - < 43% C+G
  - 43-51% C+G
  - 51-57% C+G
  - >57% C+G
- separate initial state probabilities, transition probabilities, and state length distributions for each category
- Initial, terminal, internal exons treated separately

The Gory Details

- Initial State Probabilities:
  - Proportional to the frequencies at which various functional units occur in actual genomic data.
    - Used not only training set of genes but all of Genbank
- Transition Probabilities
  - Estimated frequencies of all biologically permissible transitions.
- The diamond shaped states are regular HMM states emitting the background distribution



Exon States

- Length Distribution
  - Varies great between initial, internal and terminal exons, separate density for each
  - Small variance with C+G content, pooled the different sets for larger sample size
  - Used a smoothed empirically calculated distribution
  - Length of exon needs to be consistent with phase of its adjacent introns
    - preceding state I_2 succeeding state I_1 then length is $3n+2$ for some randomly generated n.
- Emission probabilities:
  - Based on base frequencies in all exons.

Signal Models

- Genscan uses different models to model the different biological signals
  - WMM (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 WAM (Weighted Array Model)
- A generalization of PWM that allows for dependencies between adjacent positions
- Much effort went to modeling these splice sites
- This gave GenScan a substantial improvement in performance.

GenScan Performance

- Features
  - Identification of complete intron, exon structures
  - Handles both multiple and partial genes
  - Ability to predict on both strands of the DNA
  - Predicts both optimal annotation and sub-optimal exons

GenScan Performance

sensitivity
true positive rate
TP/(TP+FN)

positive predictive value
TP/(TP+FP)

Accuracy of GENSCAN for different signal and

(a) Prediction of individual splice sites and translatio

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

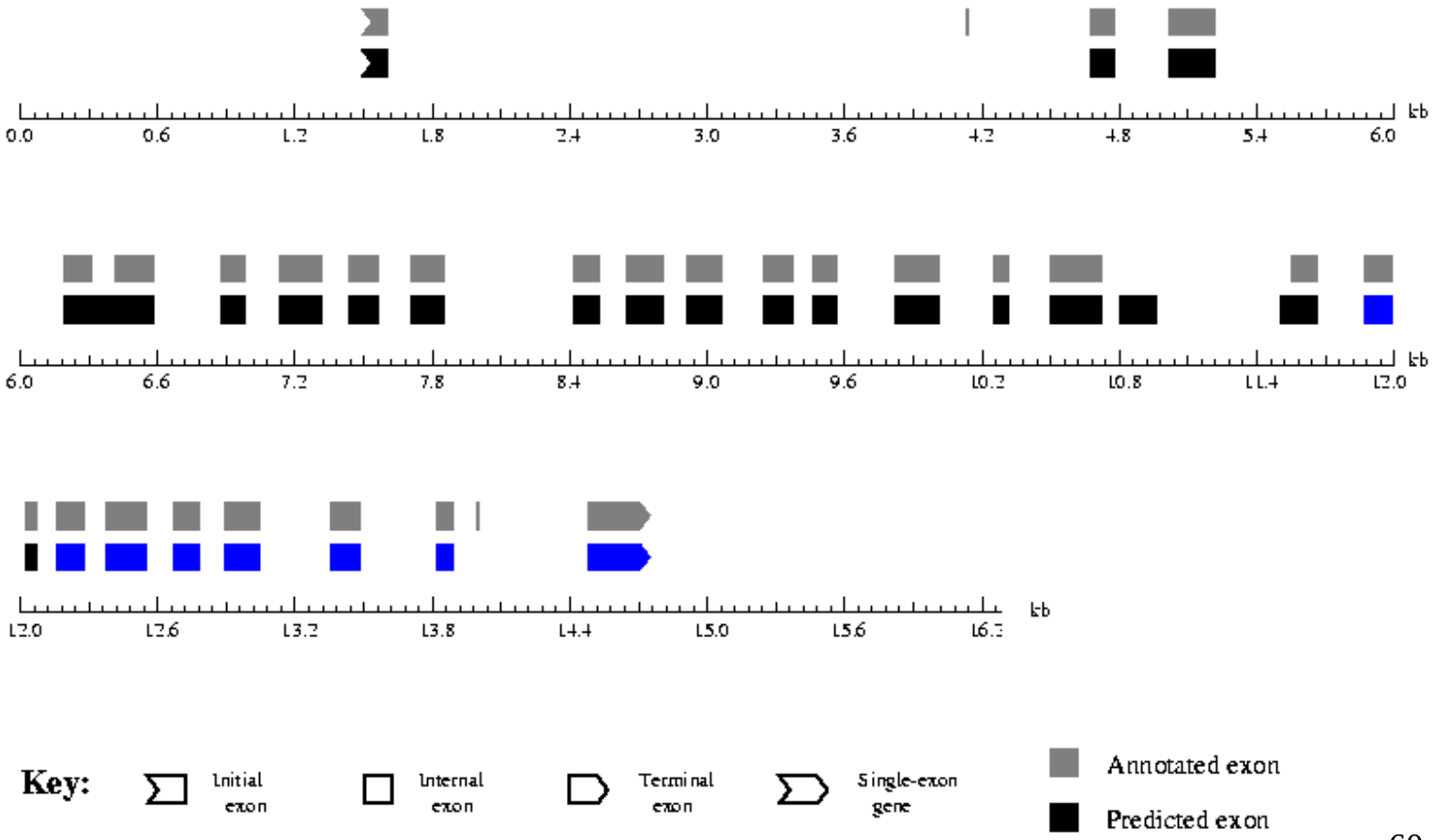
(b) Accuracy for initial, internal and terminal exons.

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

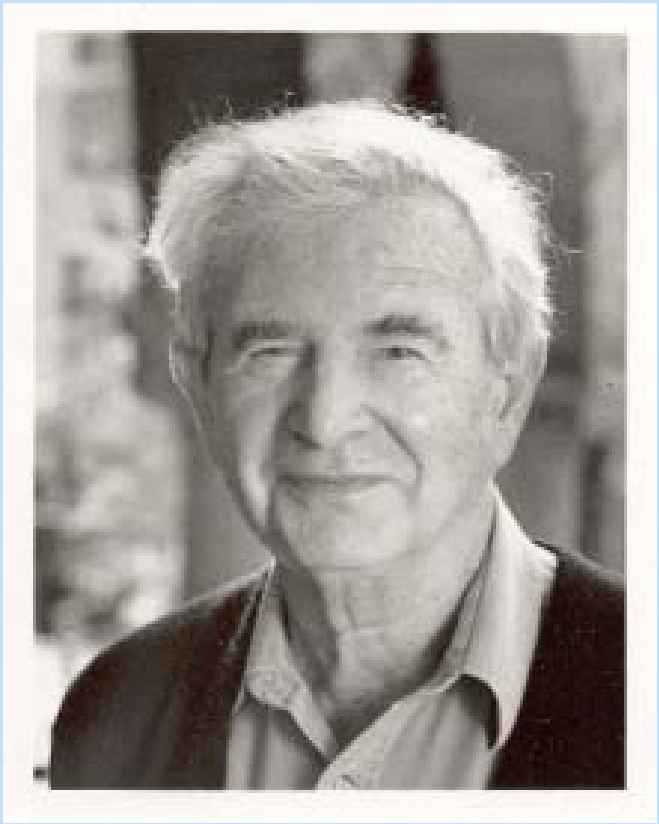
- Predicts correctly 80% of exons
  - with multiple exons probability declines…
- Prediction accuracy per bp > 90%

GenScan Output

Fig. 12. GENSCAN PostScript output for sequence HSNCAMX1

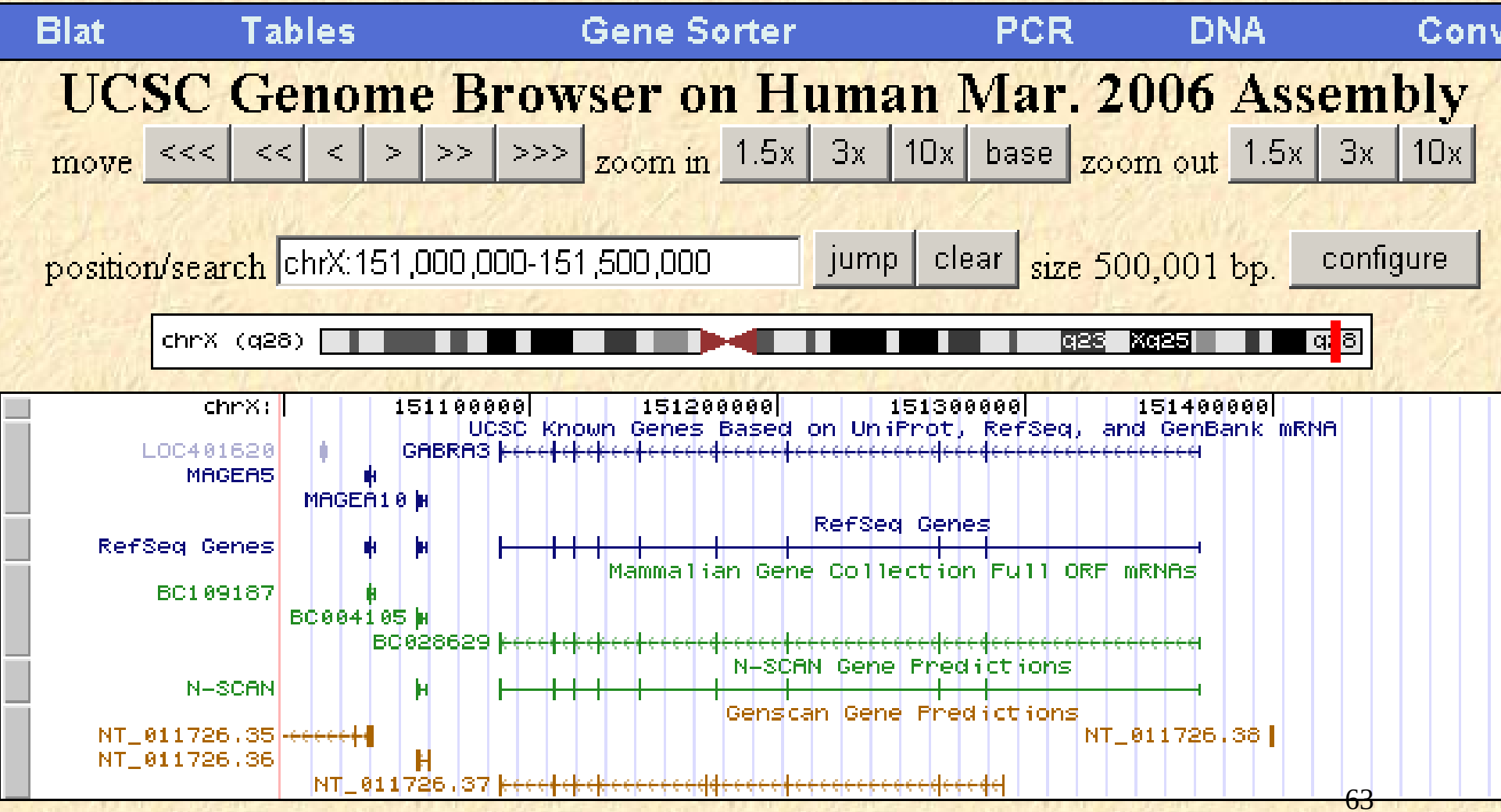


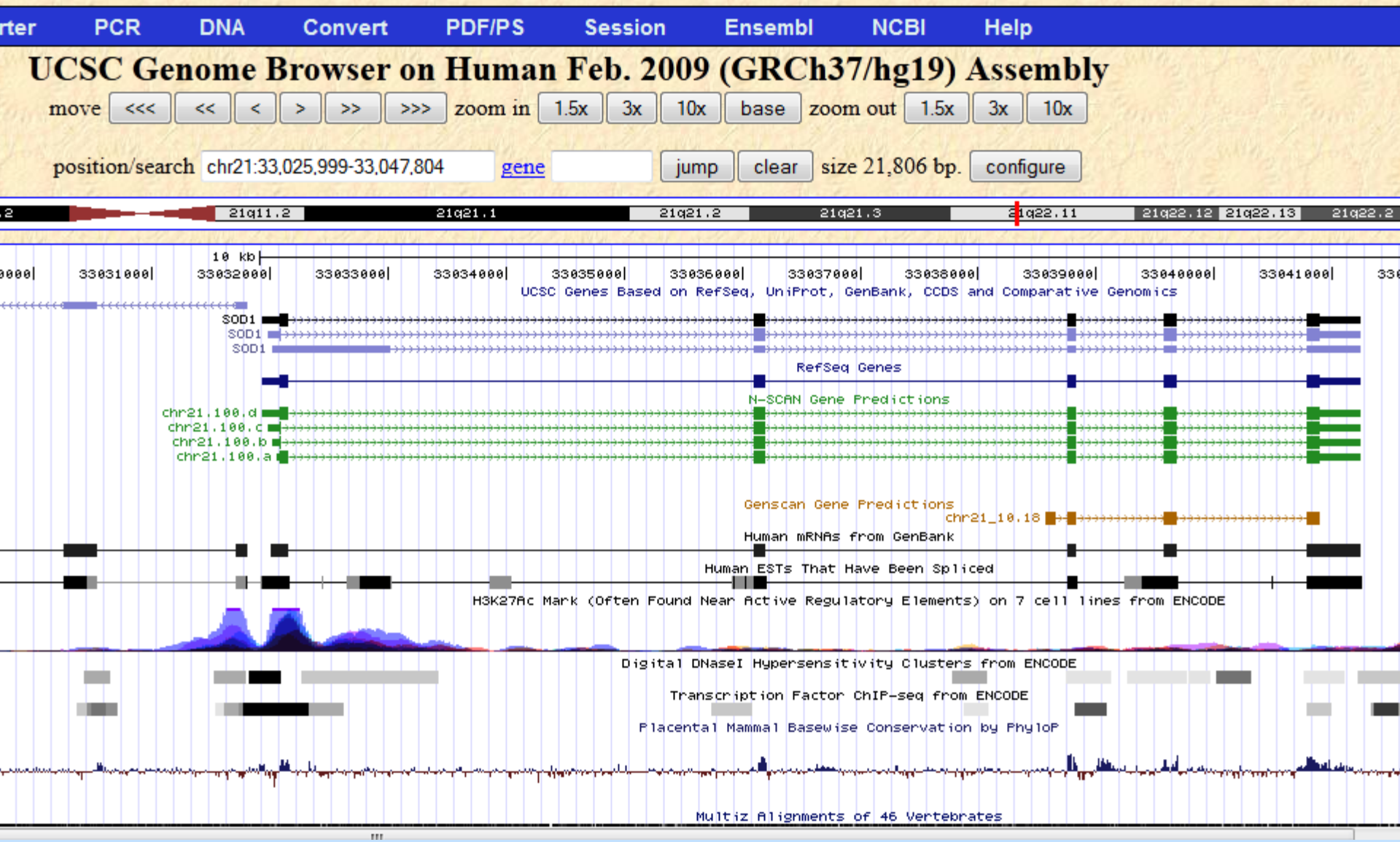
Sam Karlin, Chris Burge



Many prediction Tools

- Many prediction tools:
- dynamic programming to make the high scoring model from available features.
  - e.g. Genefinder (Green)
- Running a 5'→ 3' pass on the sequence through a Markov model based on a typical gene model
  - e.g. TBparse (Krogh), GENSCAN (Burge) or GLIMMER (Salzberg)
- Running a 5'→3' pass on the sequence through a neural net trained with confirmed gene models
  - e.g. GRAIL (Oak Ridge)
- Tools are usually used in combination.


Blat
Tables
Gene Sorter
PCR
DNA
Conv
UCSC Genome Browser on Human Mar. 2006 Assembly
move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
position/search
chrX:151,000,000-151,500,000
jump
clear
size 500,001 bp.
configure
chrX (q28)
q23
Xq25
chrX:
151100000
151200000
151300000
151400000
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
LOC401620
GABRA3
MAGEA5
MAGEA10
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC109187
BC004105
BC028629
N-SCAN Gene Predictions
N-SCAN
Genscan Gene Predictions
NT_011726.35
NT_011726.38
NT_011726.36
NT_011726.37


rter PCR DNA Convert PDF/PS Session Ensembl NCBI Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr21:33,025,999-33,047,804 gene jump clear size 21,806 bp. configure
21q11.2 21q21.1 21q21.2 21q21.3 21q22.11 21q22.12 21q22.13 21q22.2
10 kb
33031000 33032000 33033000 33034000 33035000 33036000 33037000 33038000 33039000 33040000 33041000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
SOD1
SOD1
SOD1
RefSeq Genes
N-SCAN Gene Predictions
chr21.100.d
chr21.100.c
chr21.100.b
chr21.100.a
Genscan Gene Predictions
chr21_10.18
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates

Comparative Gene Finding

An end to *ab initio* prediction?

- *ab initio* gene prediction has limited accuracy
- High false positive rates for most predictors
- Exon prediction sensitivity can be good
- Rarely used as a final product
  - Human annotators run multiple algorithms and score exon predicted by multiple predictors.
  - Used as a starting point for refinement / verification
- Prediction need correction and validation
- ➔ build gene models by comparative means!

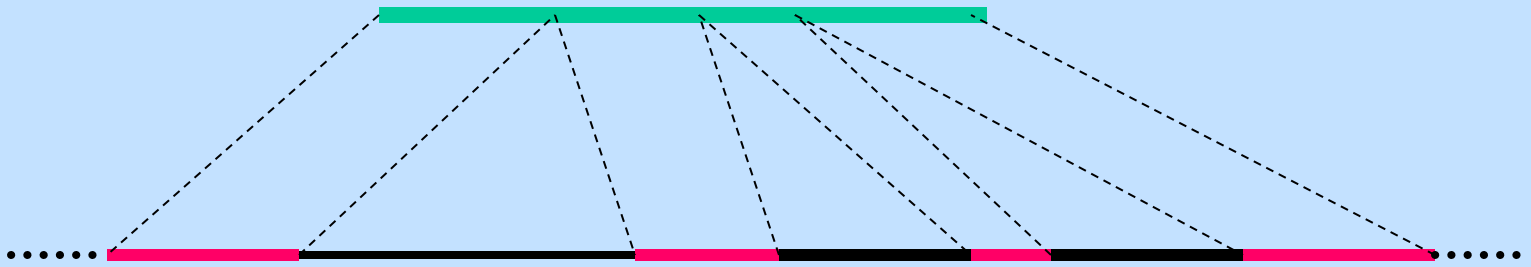
Scenario

- We have the coding sequence T of a protein from species A, and the DNA sequence G of species B.
- We think that a homolog of T appears somewhere in G, possibly interrupted by introns
- Want to find the best alignment of T to G

Spliced Alignment

Gelfand, Mironov, Pevzner PNAS '93 9061-6

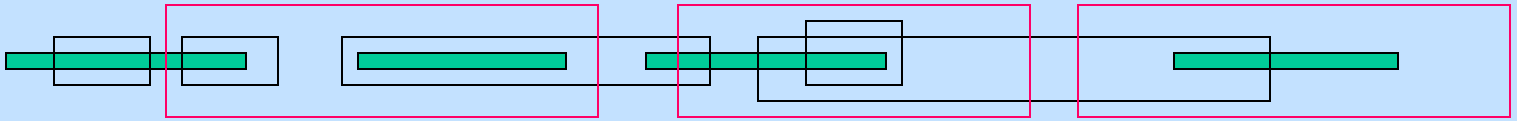
- Given **G** genomic seq, **T** reference seq (DNA seq of a related protein)
- Want to find the best match of T to G, skipping introns in G when necessary



- Need to identify alignment and splicing pattern.

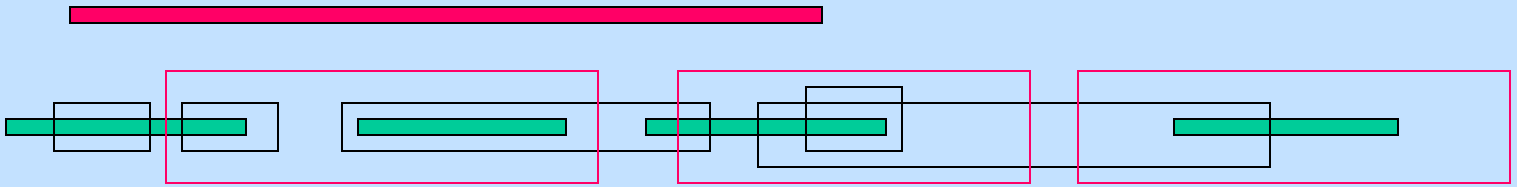
Spliced alignment: defs

- $G=g_1,\ldots,g_n$: underlying sequence
- $B=g_i,\ldots,g_j$ $B'=g_{i'},\ldots,g_{j'}$ blocks (candidate exons)
- $B \le B'$ if $j \le i'$
- $C=\{B_1,\ldots B_k\}$ is a chain if $B_1 \le \ldots \le B_k$
- C^* - concatenation of $B_1 * B_2 * \ldots B_k$
- $S(A,B)$ – score of opt. global alignment of sequences A,B

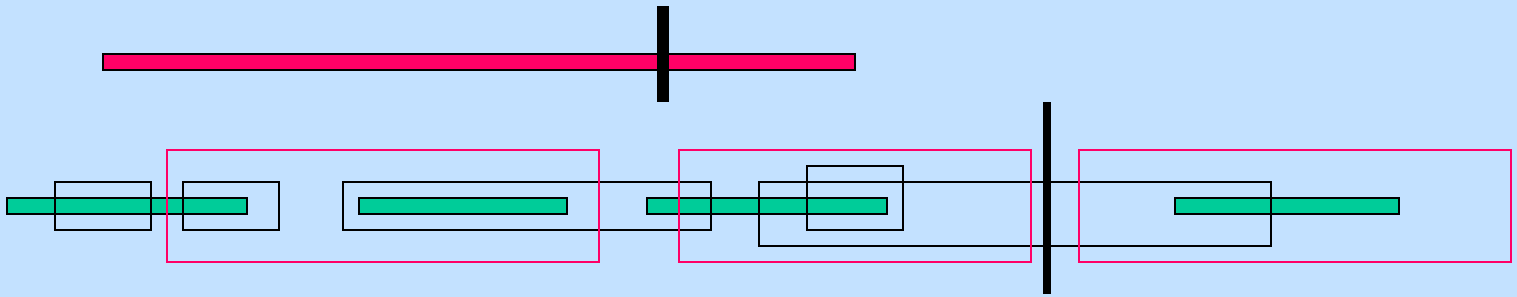


Spliced Alignment Problem

- $G=g_1,\ldots,g_n$ genomic seq
- $T=t_1,\ldots,t_m$ reference seq
- $B=\{B_1,\ldots B_b\}$ set of blocks in G
- Goal: Find a chain C of blocks from B such that $S(C,T)$ is maximum

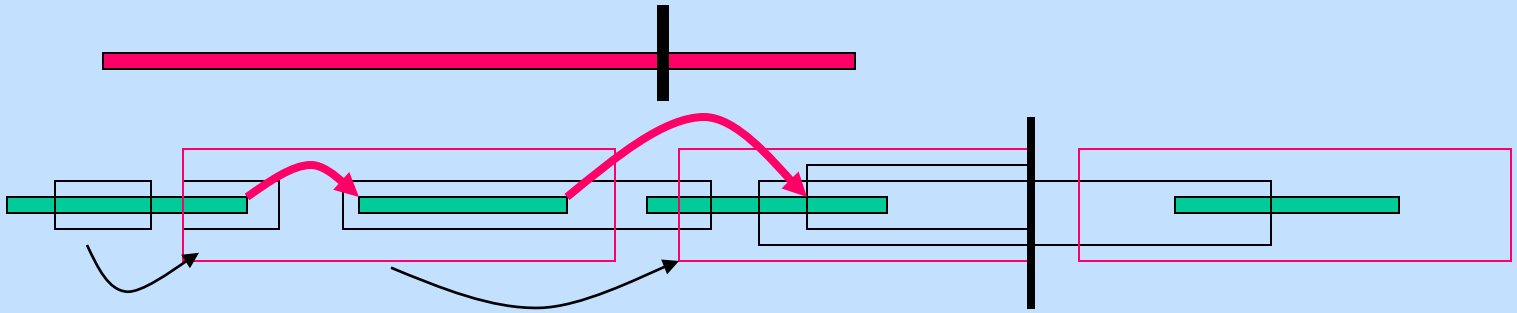


- j-prefix of $g_i,\ldots g_j,\ldots,g_n$: $A(j) = g_i,\ldots g_j$
- In block $B = g_i,\ldots g_j$ first(B) =i, last(B) =j
- Chain $F = B_1*\ldots*B_k$ ends at last(B_k),
- F ends before position i if last(B_k)<i
- If B_k contains the position i, i-prefix of $C = B_1*\ldots*B_k$ is $C^*(i) = B_1*\ldots*B_k(i)$



Network formulation

- Blocks=paths
- connect block B to B' if $B \leq B'$
- seek best alignment of T to a path in the network



- i: position contained in block B_k
- B[i] = set of blocks ending before i
- S(i,j,k) = max S(C*(i),T(j)) over all chains C containing block B_k.

|T|=m, |G|=L, N blocks
Complexity:
time: $O(mLN^2)$
space: O(mLN)

(Best score matching $t_1,\ldots,t_j$ to a chain $B_1*\ldots*B_k(i)$ where i belongs to block B_k)

- S(i,j,k) = Max {
  - $S(i-1,j-1,k) + \delta(g_i,t_j)$ if $i \neq first(k)$
  - $S(i-1,j,k) + \delta_{indel}$ if $i \neq first(k)$
  - $\text{Max}_{l \in B[i]}\ S(last(l),j-1,l) + \delta(g_i,t_j)$ if i=first(k)
  - $\text{Max}_{l \in B[i]}\ S(last(l),j,l) + \delta_{indel}$ if i=first(k)
  - $S(i,j-1,k) + \delta_{indel}$ }
- Final score: $\text{Max}_k\ S(last(k),m,k)$

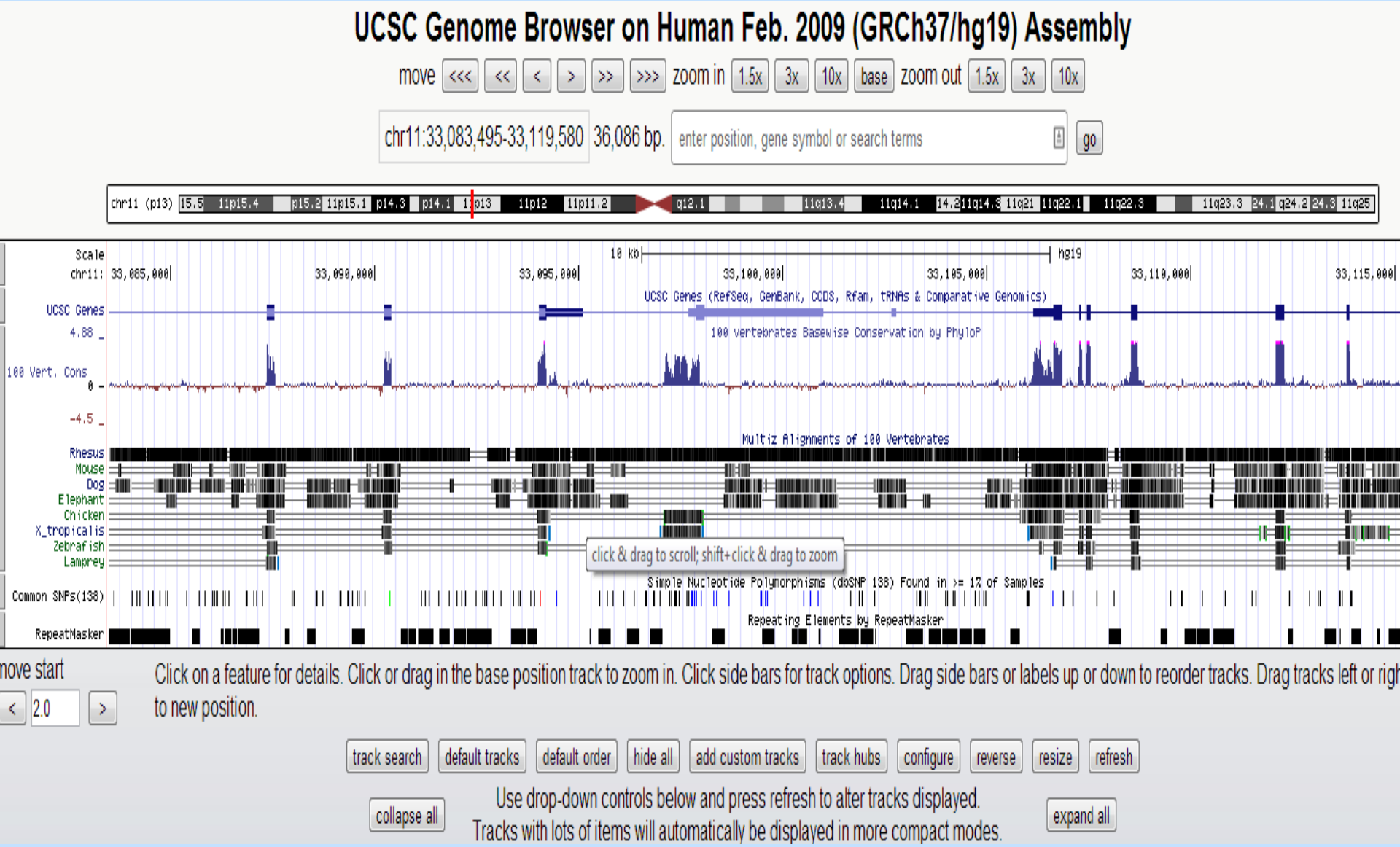
Improvement: Reducing the Number of Edges

|T|=m, |G|=L,
N blocks
time: O(mLN)
space: O(mLN)
much smaller
in practice

- $P(i,j) = \max_{l \in B[i]} S(last(l),j,l)$

(Best score matching $t_1,\ldots,t_j$ to a chain of full blocks that ends before i)

- $S(i,j,k) = \max \{$
  - $S(i-1,j-1,k) + \delta(g_i,t_j)$ if $i \neq first(k)$
  - $S(i-1,j,k) + \delta_{indel}$ if $i \neq first(k)$
  - $P(first(k),j-1) + \delta(g_i,t_j)$ if $i=first(k)$
  - $P(first(k),j) + \delta_{indel}$ if $i=first(k)$
  - $S(i,j-1,k) + \delta_{indel} \}$
- $P(i,j) = \max \{P(i-1,j), P(i,j-1) + \delta_{indel}, \max_{k:\, last(k)=i-1} S(i-1,j,k)\}$

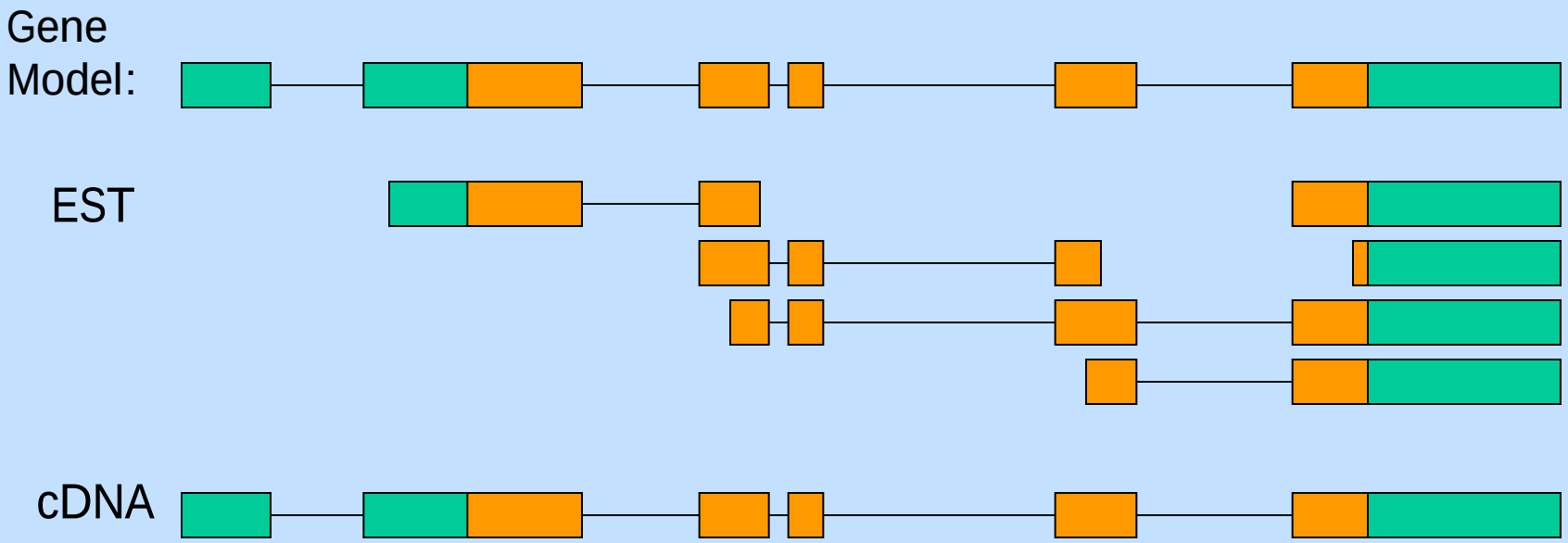


Transcript based prediction (1995-2008 style)

- sources:
  - ESTs (short mRNA fragments, must be assembled first)
  - cDNAs (longer fragments, up to full transcript length)
- Idea: align transcripts to genome, jumping over introns

Transcript-based prediction: How it works

Align transcript data to genomic sequence using pair-wise sequence comparison



Transcript based prediction using NGS (2009+ style)

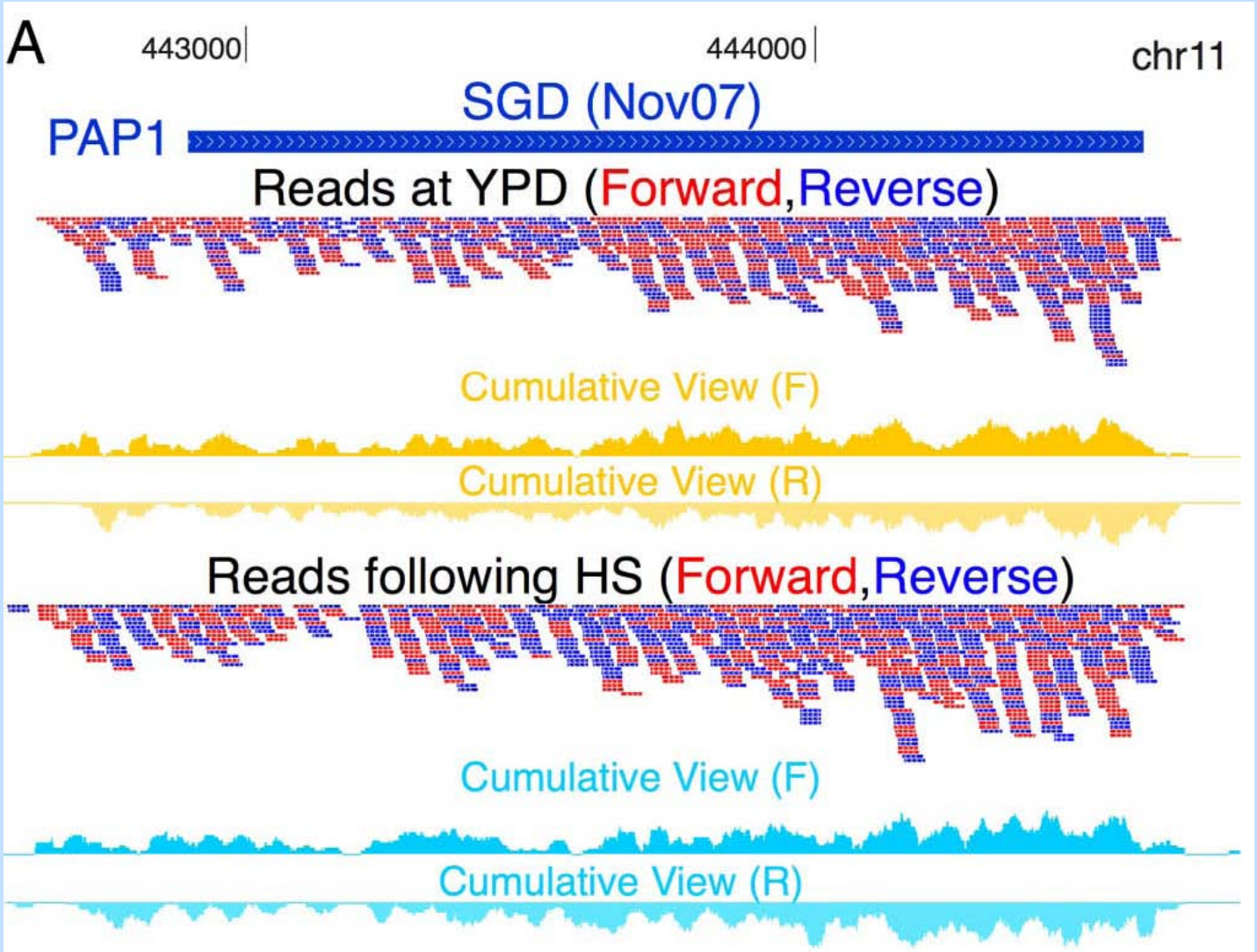
- Extract mRNA; break randomly into short segments (20-100bp)
- Sequence 100K-~~1~~100M segments
- Map segments to the known gene sequences (← suffix trees here!)
- Obtain counts how many copies of each gene were found

ABI SOLID 3

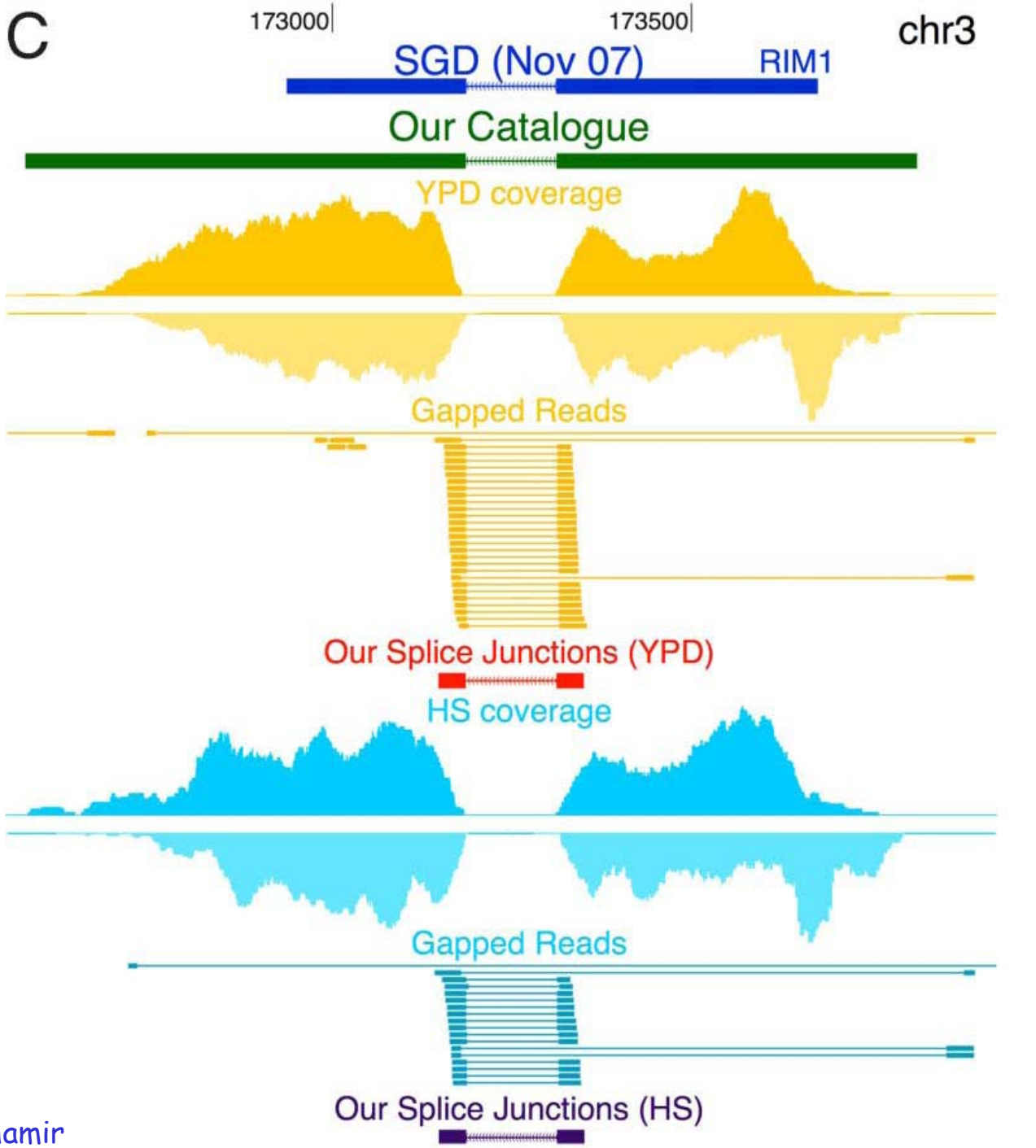
Roche 454

Illumina Genome Analyzer II

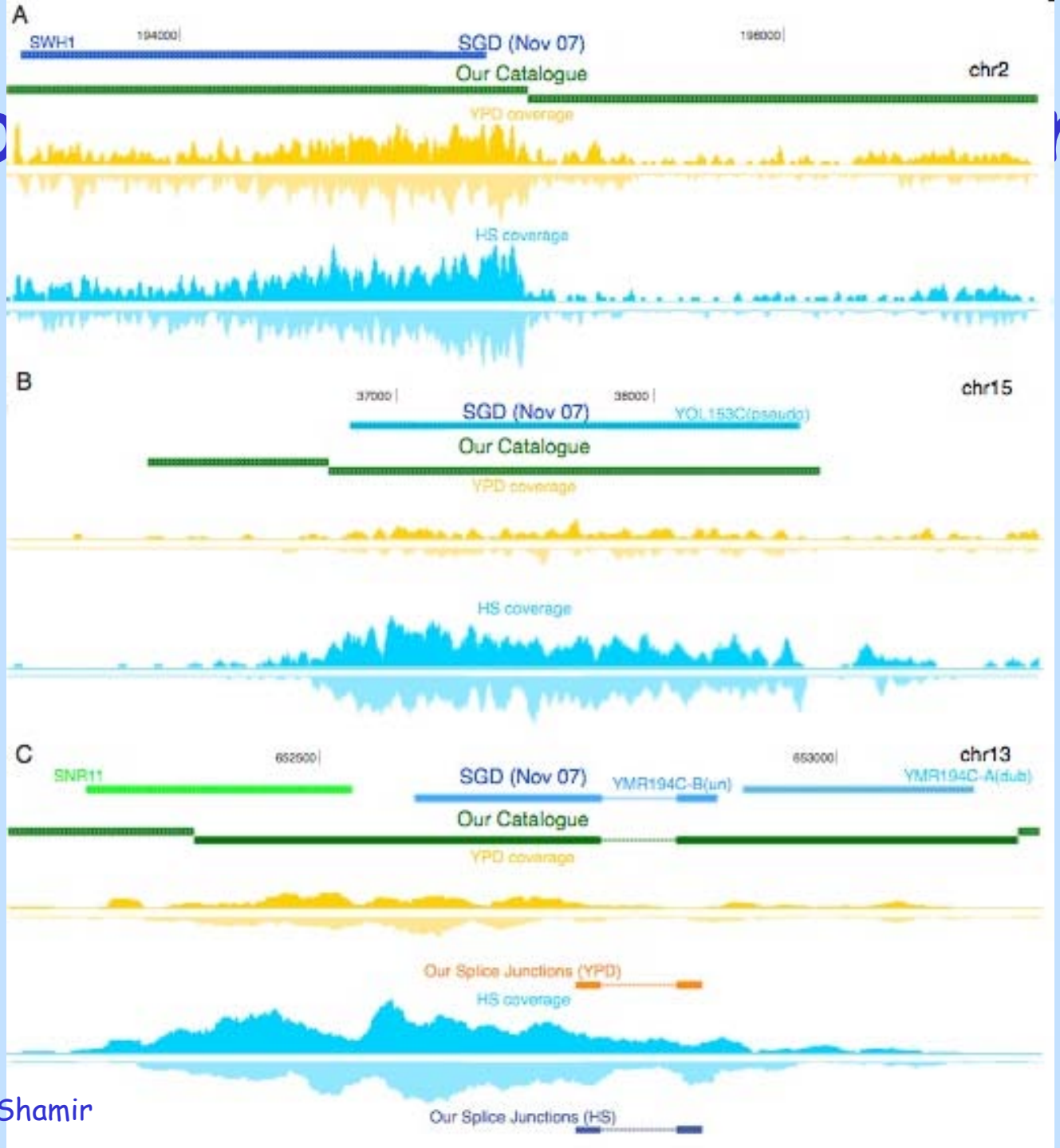
NGS transcript based gene prediction



Yassour M, et al. Ab initio Construction of a Eukaryotic Transcriptome by Massively Parallel mRNA Sequencing. PNAS 09


C
173000
173500
chr3
SGD (Nov 07)
RIM1
Our Catalogue
YPD coverage
Gapped Reads
Our Splice Junctions (YPD)
HS coverage
Gapped Reads
Our Splice Junctions (HS)

Co ns


A
SWH1
194000
SGD (Nov 07)
196000
Our Catalogue
chr2
YPD coverage
HS coverage
B
37000
38000
chr15
SGD (Nov 07)
YOL153C(pseudo)
Our Catalogue
YPD coverage
HS coverage
C
652500
653000
chr13
SNR11
SGD (Nov 07)
YMR194C-B(un)
YMR194C-A(dub)
Our Catalogue
YPD coverage
Our Splice Junctions (YPD)
HS coverage
Our Splice Junctions (HS)

FIN