Erkennung von Protein-kodierenden Genen/Genstruktur

- Prokaryonten: Konsekutiv (keine Introns-Exons), Suche nach langen ORFs, Operons, Codon-Präferenzen
- Eurkaryonten: Intron-Exon Struktur

Using slides and figures by Rodger Staden, Ron Shamir, Jones & Pevzner, and Haixu Tang. Thanks!

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgc
taatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccga
tgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgc
ggctatgcaagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctg
ggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcat
gcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctat
gctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcgg
ctatgctaagctcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctat
gctaatgcatgcggctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatg
catgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatg
cggctatgctaagctcatgcgg

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgc
taatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccga
tgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgc
ggctatgcaagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctg
ggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcat
gcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctat
gctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcgg
ctatgctaagctcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctat
gctaatgcatgcggctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatg
catgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatg
cggctatgctaagctcatgcgg

Table 8.2. *The universal or standard genetic code*

UUU-Phe	F	UCU-Ser	S	UAU-Tyr	Y	UGU-Cys	C
UUC-Phe	F	UCU-Ser	S	UAU-Tyr	Y	UGU-Cys	C
UUA-Leu	L	UCA-Ser	S	UAA-	TER	UGA-	TER
UUG-Leu	L	UCG-Ser	S	UAG-	TER	UGG--Trp	W
CUU-Leu	L	CCU-Pro	P	CAU-His	H	CGU-Arg	R
CUC-Leu	L	CCU-Pro	P	CAU-His	H	CGC-Arg	R
CUA-Leu	L	CCA-Pro	P	CAA-Gln	Q	CGA-Arg	R
CUG-Leu	L	CCG-Pro	P	CAG-Gln	Q	CGG-Arg	R
AUU-Ile	I	ACU-Thr	T	AAU-Asn	N	AGU-Ser	S
AUC-Ile	I	ACC-Thr	T	AAC-Asn	N	AGC-Ser	S
AUA-Ile	I	ACA-Thr	T	AAA-Lys	K	AGA-Arg	R
AUG-MET	M	ACG-Thr	T	AAG-Lys	K	AGG-Arg	R
GUU-Val	V	GCU-Ala	A	GAU-Asp	D	GGU-Gly	G
GUC-Val	V	GCC-Ala	A	GAC-Asp	D	GGC-Gly	G
GUA-Val	V	GCA-Ala	A	GAA-Glu	E	GGA-Gly	G
GUG-Val	V	GCG-Ala	A	GAG-Glu	E	GGG-Gly	G

Shown are each codon and the three-letter and one-letter codes for each encoded amino acid. ATG is the usual START codon and the three TER codons cause translational termination.

Six Frames in a DNA Sequence

```
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
```

```
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
```

```
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
```

- stop codons – TAA, TAG, TGA
- start codons - ATG

Gene Prediction and Motifs

- Upstream regions of genes often contain motifs that can be used for gene prediction

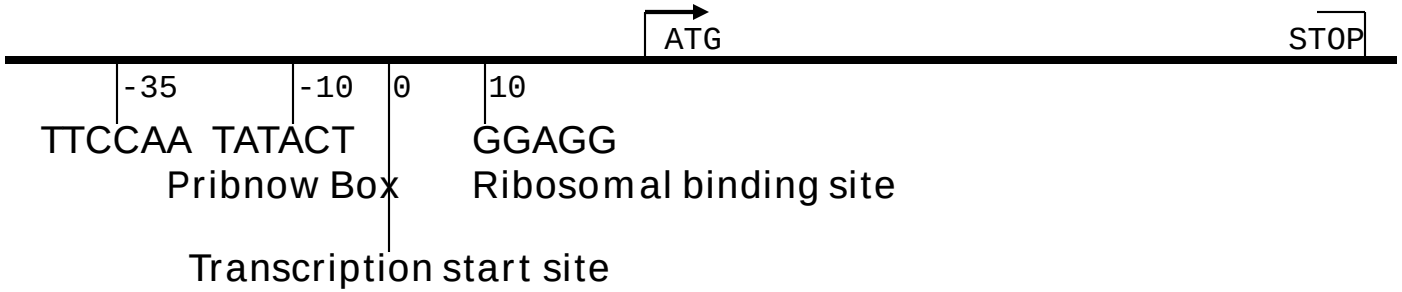


Table 8.3. *Codon usage table*

UUU-Phe	16.6	26.0	UCU-Ser	14.5	23.6	UAU-Tyr	12.1	18.8	UGU-Cys	9.7	
UUC-Leu	20.7	18.2	UCC-Ser	17.7	14.2	UAC-Tyr	16.3	14.7	UGC-Cys	12.4	
UUA-Leu	7.0	26.3	UCA-Ser	11.4	18.8	UAA-TER	0.7	1.0	UGA-TER	1.3	
UUG-Leu	12.0	27.1	UCG-Ser	4.5	8.6	UAG-TER	0.5	0.5	UGG-Trp	13.0	
CUU-Leu	12.4	12.2	CCU-Pro	17.2	13.6	CAU-His	10.1	13.7	CGU-Arg	4.7	
CUC-Leu	19.3	5.4	CCC-Pro	20.3	6.8	CAC-His	14.9	7.8	CGC-Arg	11.0	
CUA-Leu	6.8	13.4	CCA-Pro	16.5	18.2	CAA-Gln	11.8	27.5	CGA-Arg	6.2	
CUG-Leu	40.0	10.4	CCG-Pro	7.1	5.3	CAG-Gln	34.4	12.2	CGG-Arg	11.6	
AUU-Ile	15.7	30.2	ACU-Thr	12.7	20.2	AAU-Asn	16.8	36.0	AGU-Ser	11.7	
AUC-Ile	22.3	17.1	ACC-Thr	19.9	12.6	AAC-Asn	20.2	24.9	AGC-Ser	19.3	
AUA-Ile	7.0	17.8	ACA-Thr	14.7	17.7	AAA-Lys	23.6	42.1	AGA-Arg	11.2	
AUG-MET	22.2	20.9	ACG-Thr	6.4	8.0	AAG-Lys	33.2	30.8	AGG-Arg	11.1	
GUU-Val	10.7	22.0	GCU-Ala	18.4	21.1	GAU-Asp	22.2	37.8	GGU-Gly	10.9	
GUC-Val	14.8	11.6	GCC-Ala	28.6	12.6	GAC-Asp	26.5	20.4	GGC-Gly	23.1	
GUA-Val	6.8	11.7	GCA-Ala	15.6	16.2	GAA-Glu	28.6	45.9	GGA-Gly	16.4	
GUG-Val	29.3	10.7	GCG-Ala	7.7	6.1	GAG-Glu	40.6	19.1	GGG-Gly	16.5	

Shown are frequency of each codon per 100,000 codons obtained from http://www.kazusa.or.jp/cod
for *Homo sapiens*; columns 2, 5, 8, and 11, and for *Saccharomyces cerevisiae*, columns 3, 6, 9, and 12.

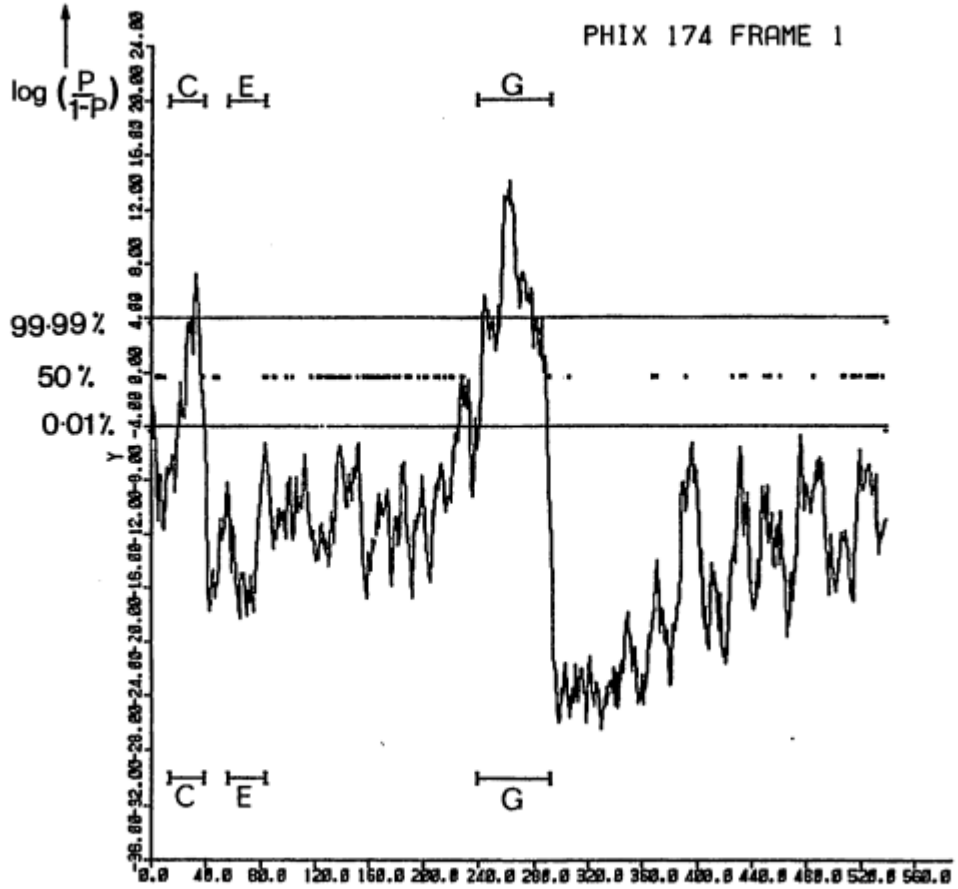

PHIX 174 FRAME 1
log (P/1-P)
99·99%
50%
0·01%
C
E
G
Y
Sequence Position x 10

Fig 2a

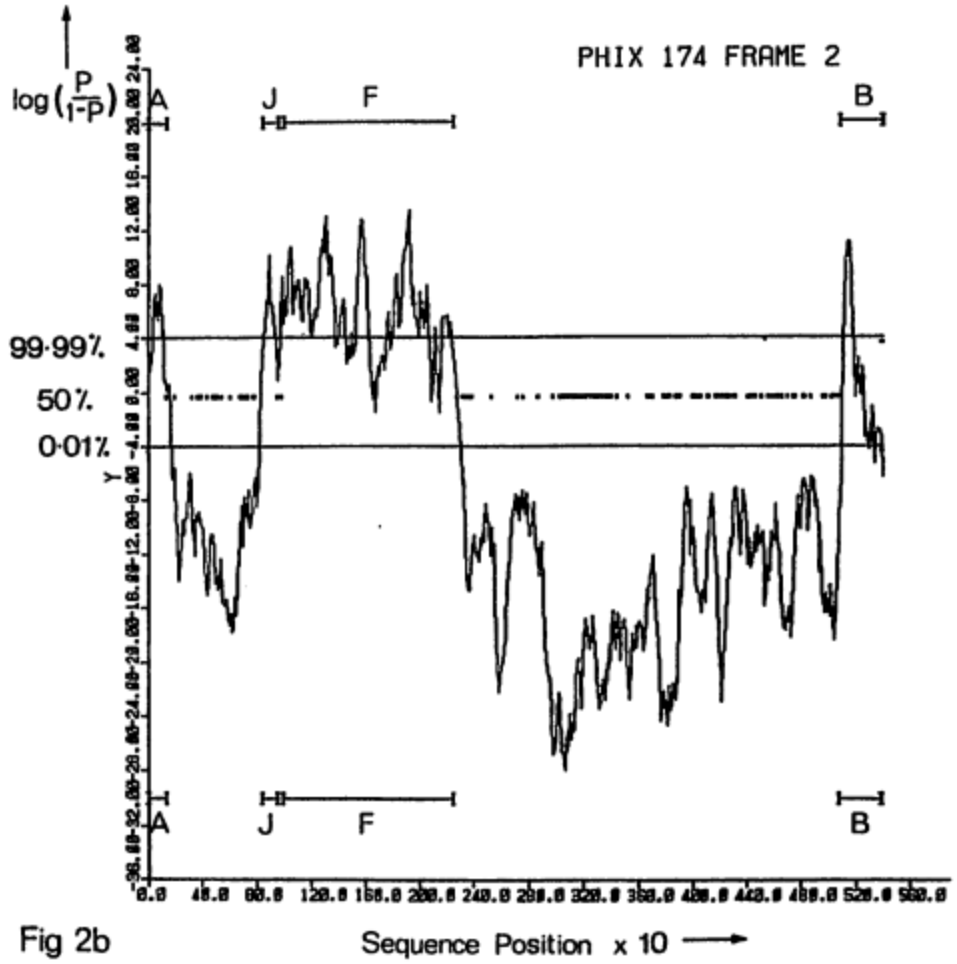

PHIX 174 FRAME 2
log(P/(1-P))
99·99%
50%
0·01%
A
J
F
B
Y
Sequence Position x 10

Fig 2b

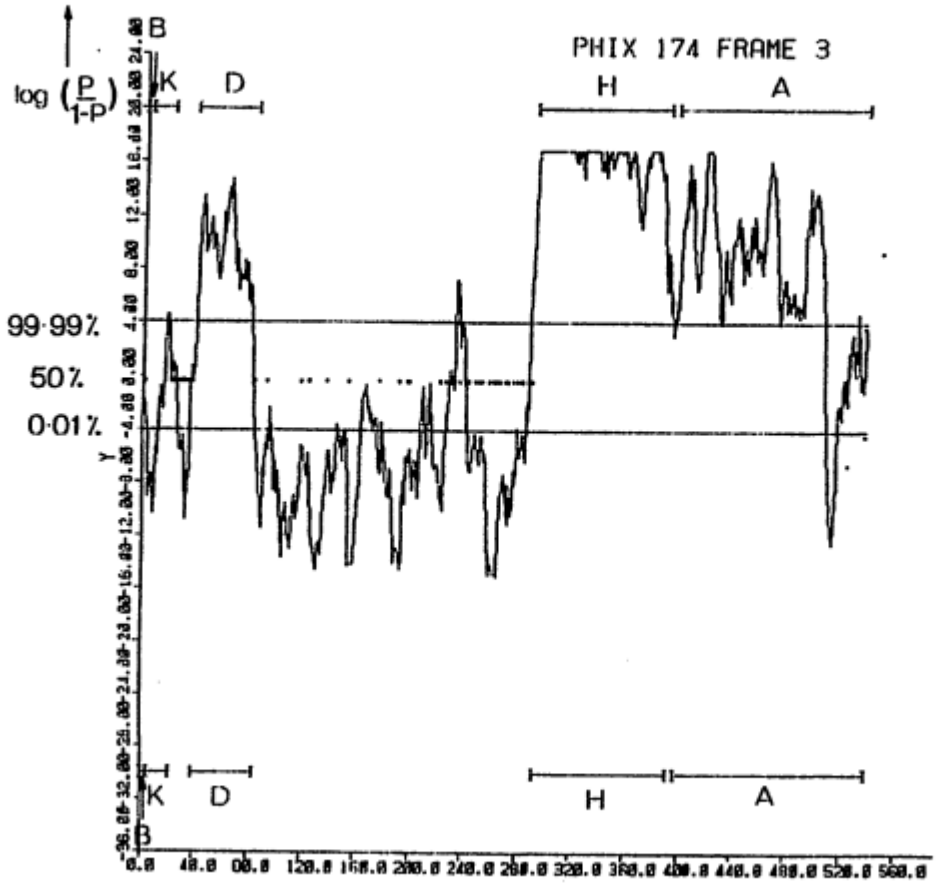

PHIX 174 FRAME 3
log (P/1-P)
99·99%
50%
0·01%
B
K
D
H
A
Y
Sequence Position x 10

Fig 2c

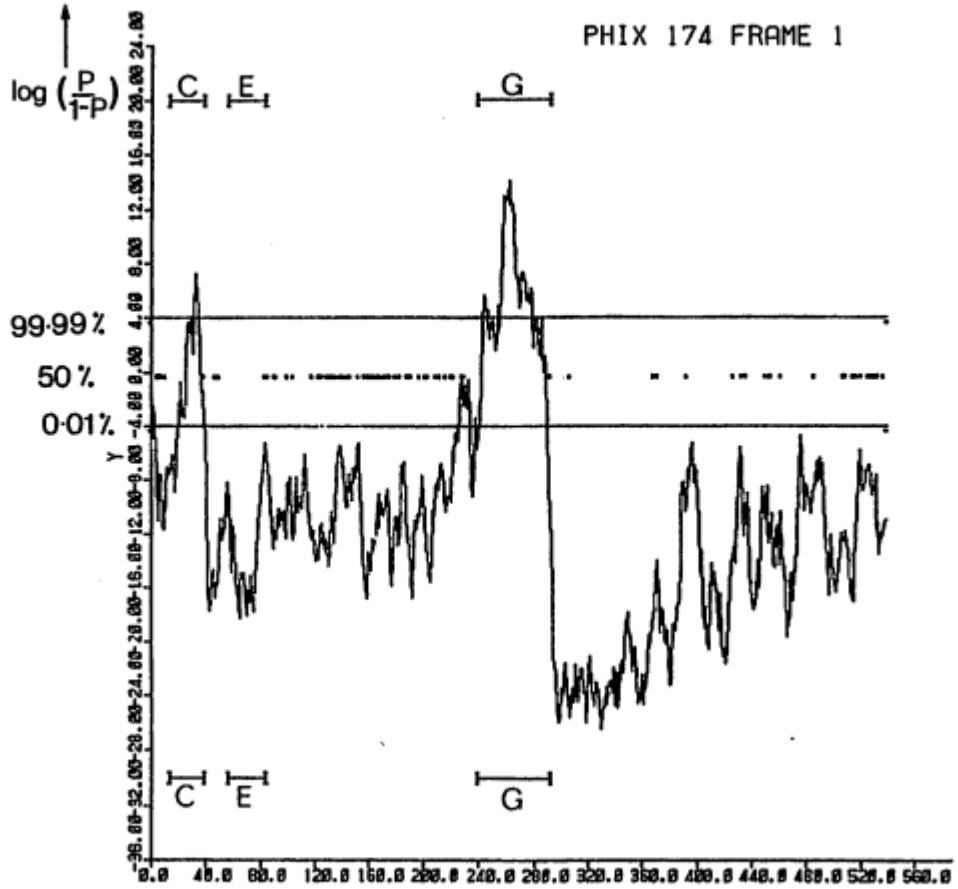

PHIX 174 FRAME 1
log (P/1-P)
99·99%
50%
0·01%
C
E
G
Y
Sequence Position x 10

Fig 2a

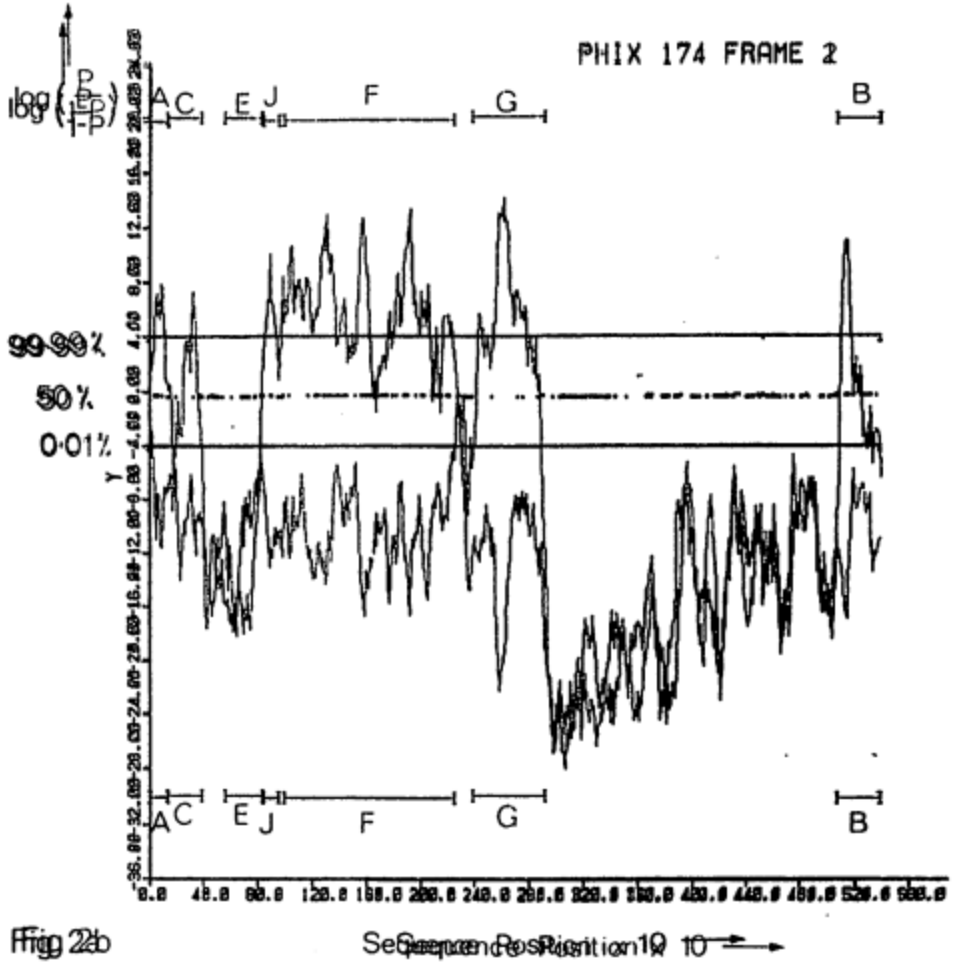

PHIX 174 FRAME 2
log (P/(1-P))
99.99%
50%
0.01%
A
C
E
J
F
G
B
Sequence Position

Fig 2b

Nucleic Acids Research

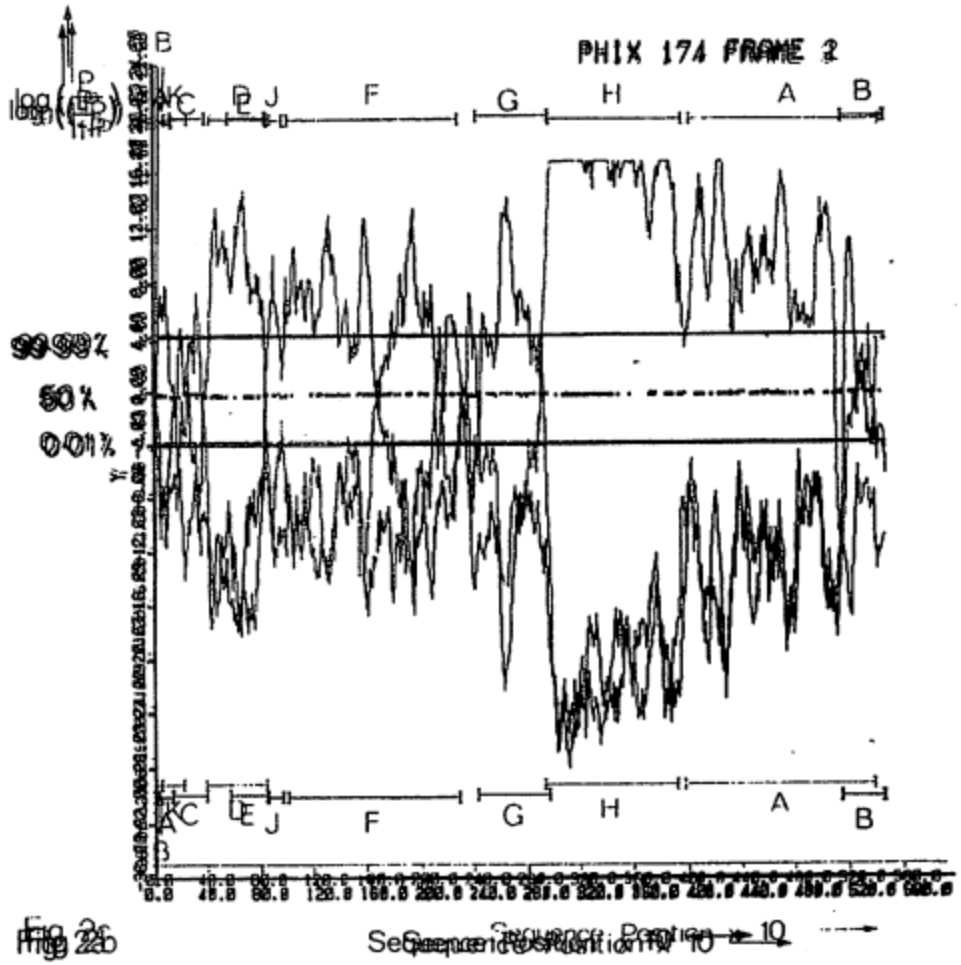
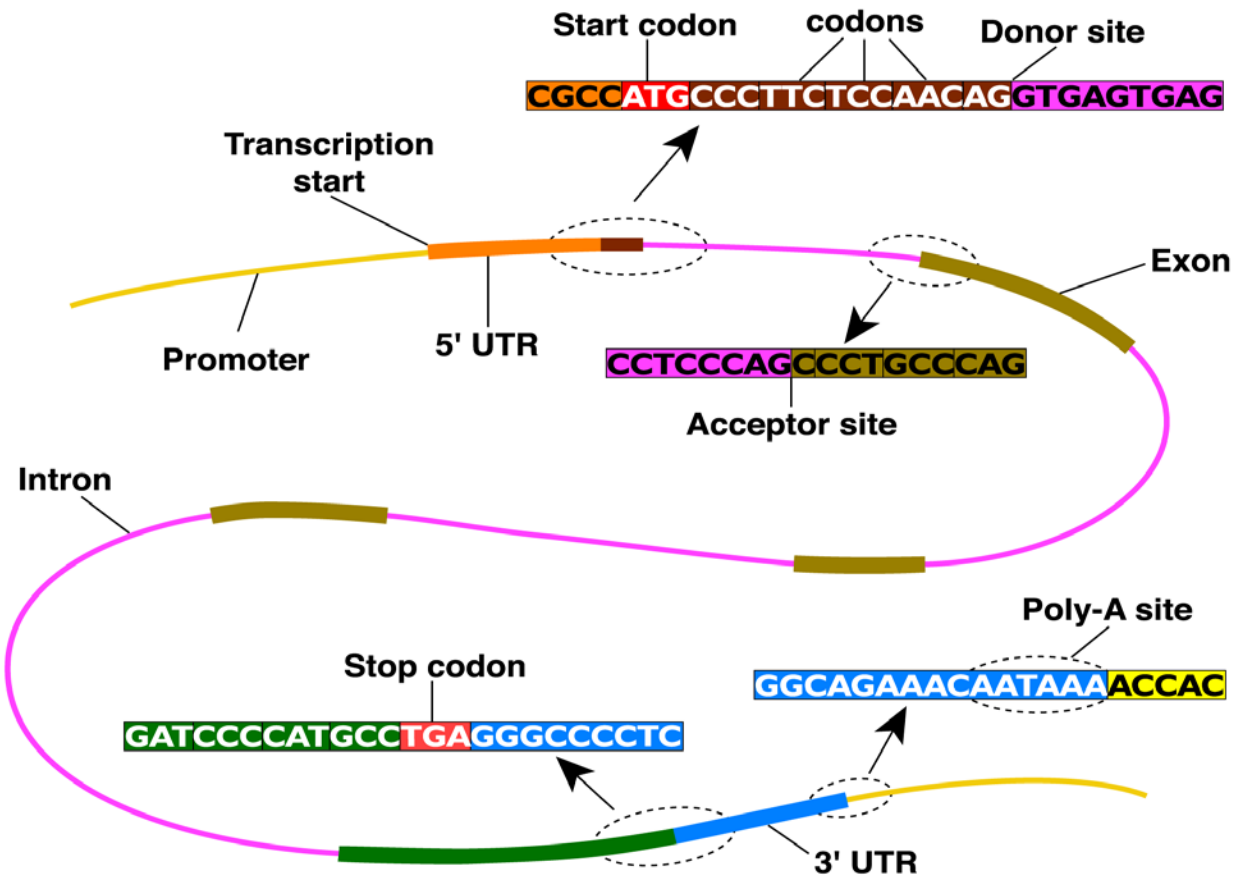


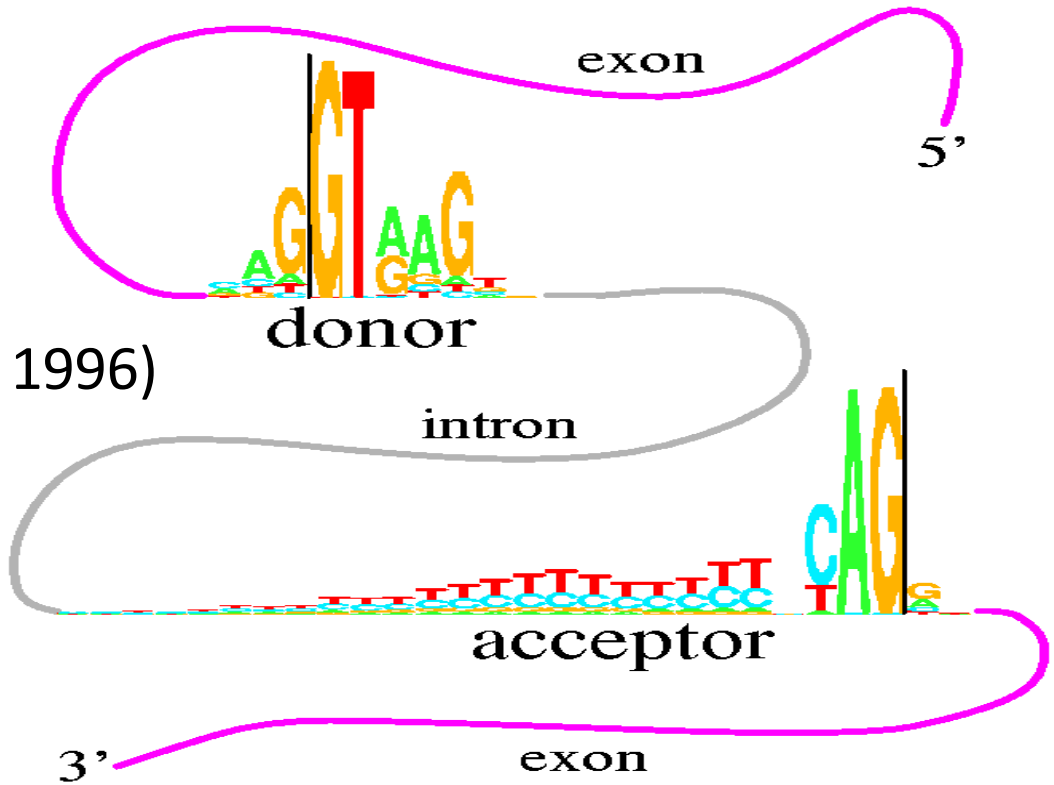
Fig 2c
Fig 2b


Start codon
codons
Donor site
CGCCATGCCCTTCTCCAACAGGTGAGTGAG
Transcription start
Exon
Promoter
5' UTR
CCTCCCAGCCCTGCCCAG
Acceptor site
Intron
Poly-A site
Stop codon
GGCAGAAACAATAAAACCAC
GATCCCCATGCCTGAGGGCCCCTC
3' UTR

Donor and Acceptor Sites: Motif Logos

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)

Donor: 7.9 bits
Acceptor: 9.4 bits
(Stephens & Schneider, 1996)

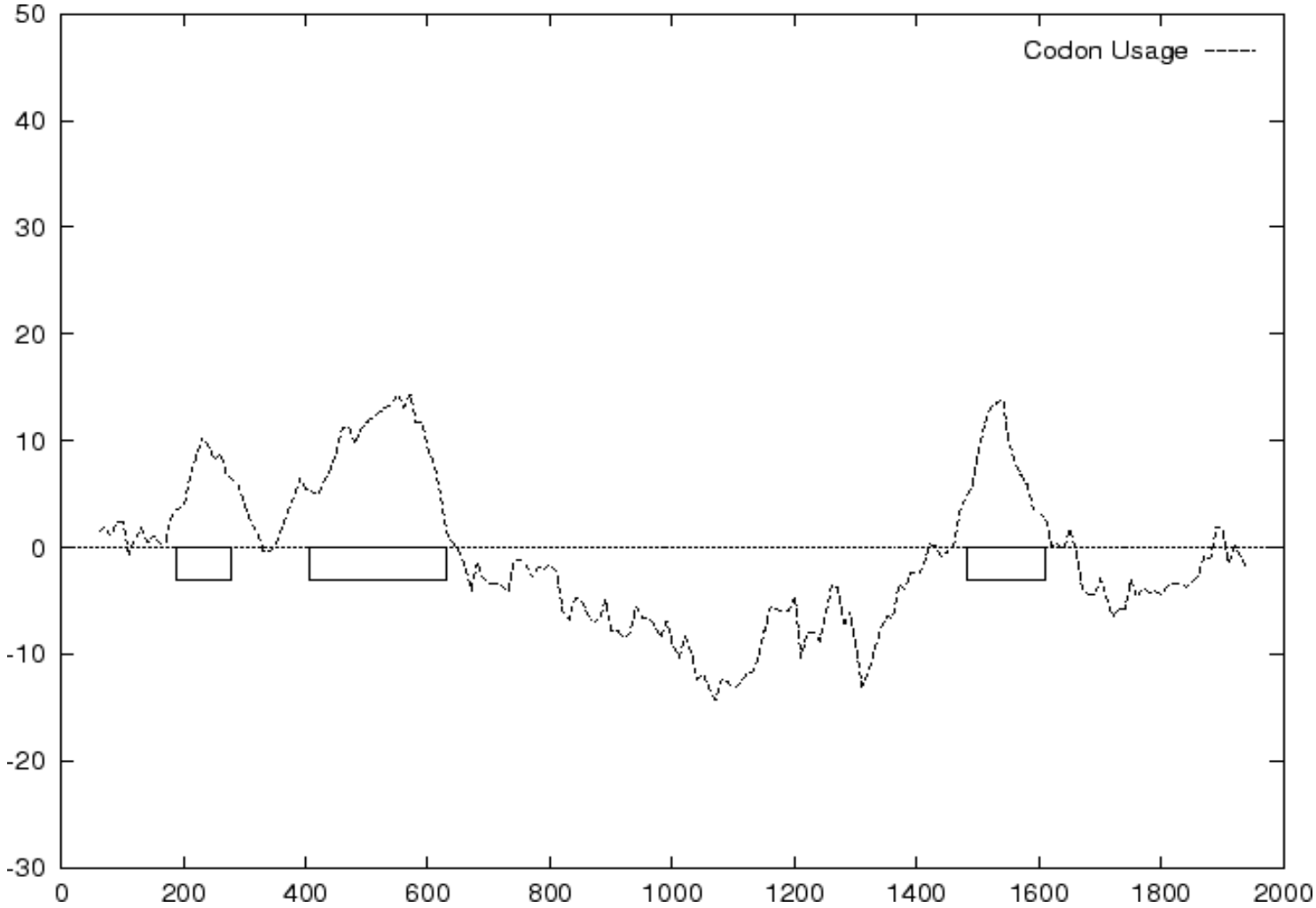


(http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

Codon Usage in Human Genome

	U			C			A			G		
U	UUU	Phe	57	UCU	Ser	16	UAU	Tyr	58	UGU	Cys	45
	UUC	Phe	43	UCC	Ser	15	UAC	Tyr	42	UGC	Cys	55
	UUA	Leu	13	UCA	Ser	13	UAA	Stp	62	UGA	Stp	30
	UUG	Leu	13	UCG	Ser	15	UAG	Stp	8	UGG	Trp	100
C	CUU	Leu	11	CCU	Pro	17	CAU	His	57	CGU	Arg	37
	CUC	Leu	10	CCC	Pro	17	CAC	His	43	CGC	Arg	38
	CUA	Leu	4	CCA	Pro	20	CAA	Gln	45	CGA	Arg	7
	CUG	Leu	49	CCG	Pro	51	CAG	Gln	66	CGG	Arg	10
A	AUU	Ile	50	ACU	Thr	18	AAU	Asn	46	AGU	Ser	15
	AUC	Ile	41	ACC	Thr	42	AAC	Asn	54	AGC	Ser	26
	AUA	Ile	9	ACA	Thr	15	AAA	Lys	75	AGA	Arg	5
	AUG	Met	100	ACG	Thr	26	AAG	Lys	25	AGG	Arg	3
G	GUU	Val	27	GCU	Ala	17	GAU	Asp	63	GGU	Gly	34
	GUC	Val	21	GCC	Ala	27	GAC	Asp	37	GGC	Gly	39
	GUA	Val	16	GCA	Ala	22	GAA	Glu	68	GGA	Gly	12
	GUG	Val	36	GCG	Ala	34	GAG	Glu	32	GGG	Gly	15

Coding Profile of ß-globin gene



Gene finding using codon frequency

Consider sequence $x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9 \ldots$
where x_i is a nucleotide

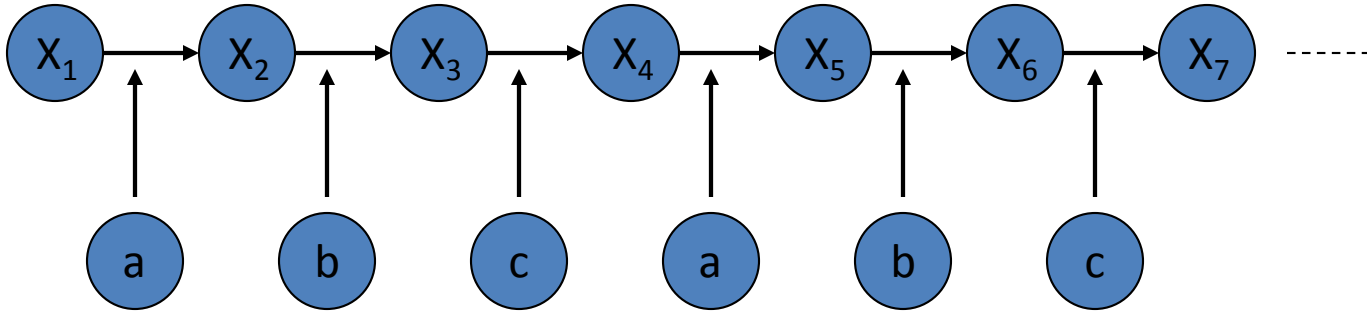
let $p_1 = p_{x1\ x2\ x3}\, p_{x4\ x5\ x6} \cdots$
$p_2 = p_{x2\ x3\ x4}\, p_{x5\ x6\ x7} \cdots$
$p_3 = p_{x3\ x4\ x5}\, p_{x6\ x7\ x8} \cdots$

then probability that ith reading frame is the coding frame is:

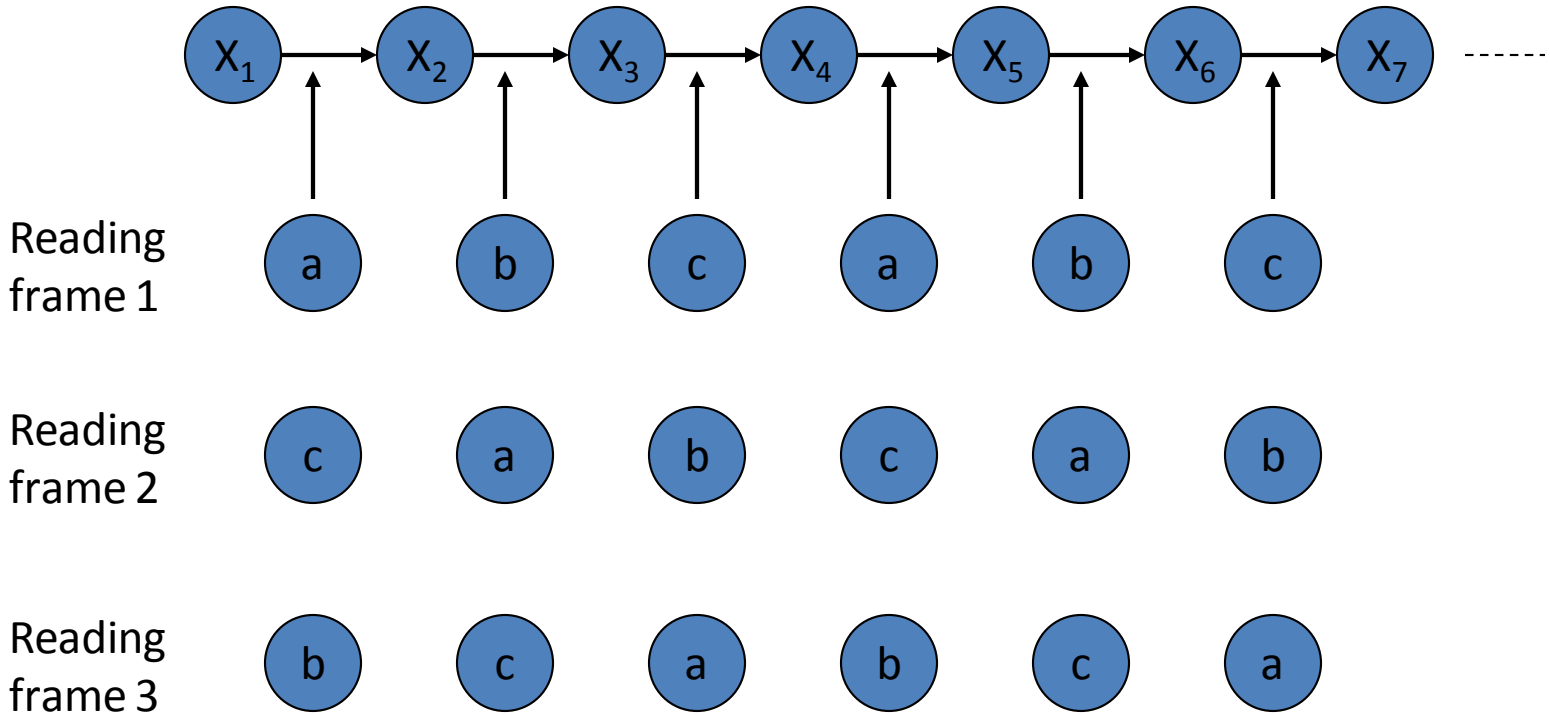
$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

slide a window along the sequence and compute P_i

Inhomogeneous Markov chain: learning



Inhomogeneous Markov chain: prediction



Gene finding using inhomogeneous Markov chain

Consider sequence $x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9 \ldots$
where x_i is a nucleotide

let $p_1 = a_{x1x2} b_{x2x3} c_{x3x4} a_{x4x5} b_{x5x6} c_{x6x7} \ldots$
$p_2 = b_{x1x2} c_{x2x3} a_{x3x4} b_{x4x5} c_{x5x6} a_{x6x7} \ldots$
$p_3 = c_{x1x2} a_{x2x3} b_{x3x4} c_{x4x5} a_{x5x6} b_{x6x7} \ldots$

then probability that ith reading frame is the coding frame is:

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

M. Bodorovsky, Genemark (commonly used gene finder for bacterial genomes)

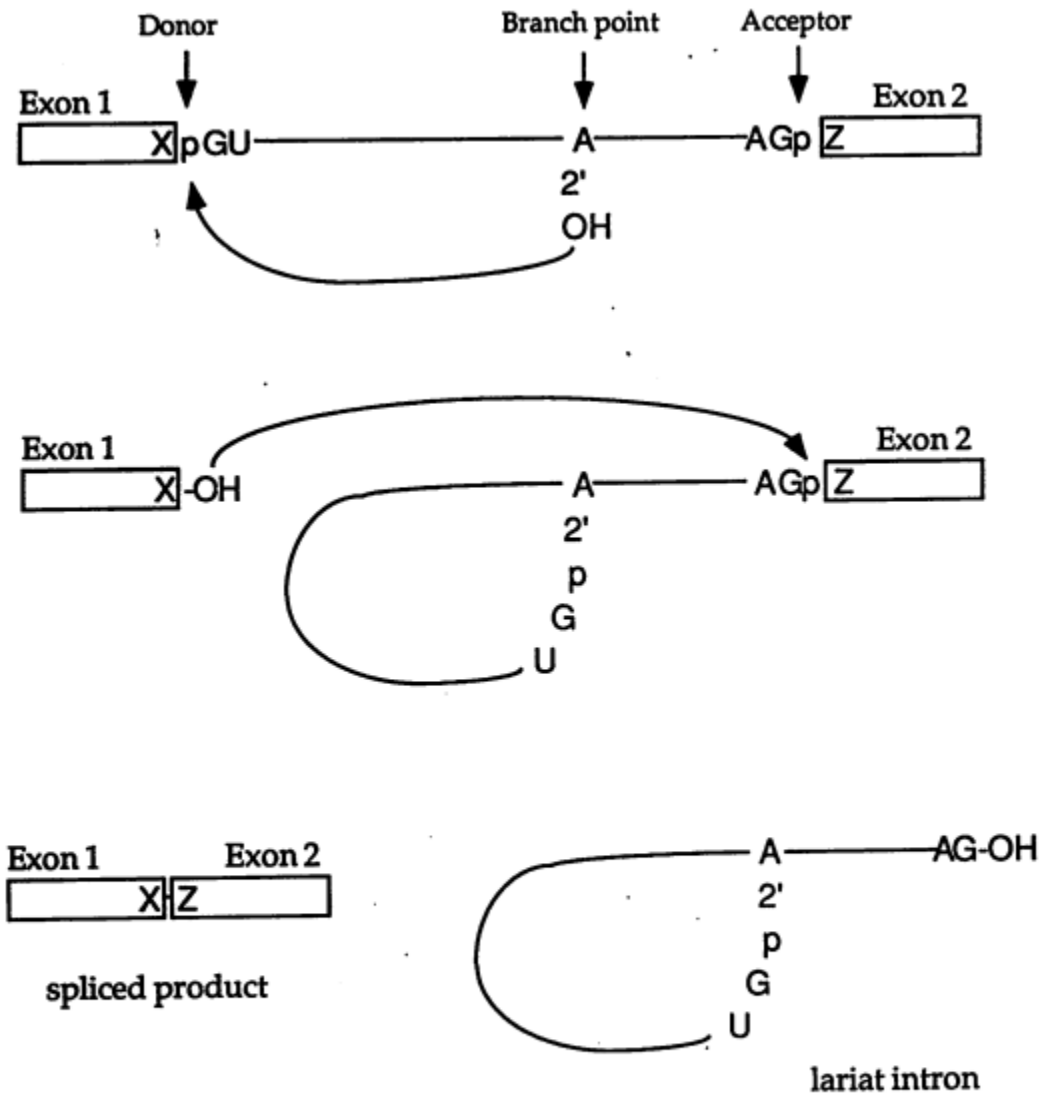


Fig. 1. The essential steps in the splicing of a pre-mRNA containing two exons and one intron are illustrated. The letter "p" is used to represent the two phosphodiester linkages which are broken in the course of splicing; the letters "OH" represent hydroxyl groups, e.g., the 2′ hydroxyl of the branch point adenosine. The conserved donor (GU), acceptor (AG) and branch point (A) nucleotides are indicated by the corresponding letters; X and Z represent the last nucleotide of exon 1 and the first nucleotide of exon 2, respectively.

Table 5. Base composition around intron/exon junctions

a. Branch point region, $[-38, -21]$

Pos	-38	-37	-36	-35	-34	-33	-32	-31	-30	-29	-28	-27	-26	-25	-24	-23	-22	-21
A%	22	20	22	24	21	21	20	22	23	22	21	21	22	23	21	23	20	20
G%	25	26	25	22	23	22	22	21	23	20	20	18	20	16	17	18	17	16
C%	28	28	26	28	28	29	29	29	29	30	30	31	30	31	30	29	31	34
T%	26	27	26	26	28	28	29	28	25	28	28	30	28	31	33	30	32	30
Y%	54	54	52	55	56	57	57	57	55	58	50	61	58	61	63	59	63	64

b. Pyrimidine-rich region, $[-20, -5]$

Pos	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5
A%	20	16	15	14	14	12	9	9	8	8	8	8	8	9	6	7
G%	16	18	18	18	15	12	13	13	12	12	13	13	12	10	6	6
C%	31	32	32	31	35	37	35	34	34	33	33	38	41	41	44	38
T%	34	33	35	37	35	39	42	45	46	47	46	42	39	41	44	48
Y%	65	66	66	68	71	76	78	79	80	80	80	80	80	82	88	87

c. Acceptor site region, $[-4, +3]$

Pos	-4	-3	-2	-1	+1	+2	+3
A%	22	4	100	0	25	25	27
G%	22	0	0	100	52	22	24
C%	33	74	0	0	13	21	27
T%	22	21	0	0	9	32	23
Y%	55	96	0	0	23	53	50

Legend. Compositional data for 1,254 acceptor sites from the 238 multi-exon genes of the learning set (Appendix A). The letter Y indicates either pyrimidine nucleotide (C or T).

A natural approach to building gene recognition algorithms is to first construct component algorithms that recognize the major features of genes: statistical bias in exon sequence, the patterns at intron junctions, promoters, enhancers, etc., and then to build a combined algorithm that that recognizes when all these component patterns occur in a pattern consistent with with that present in a gene.

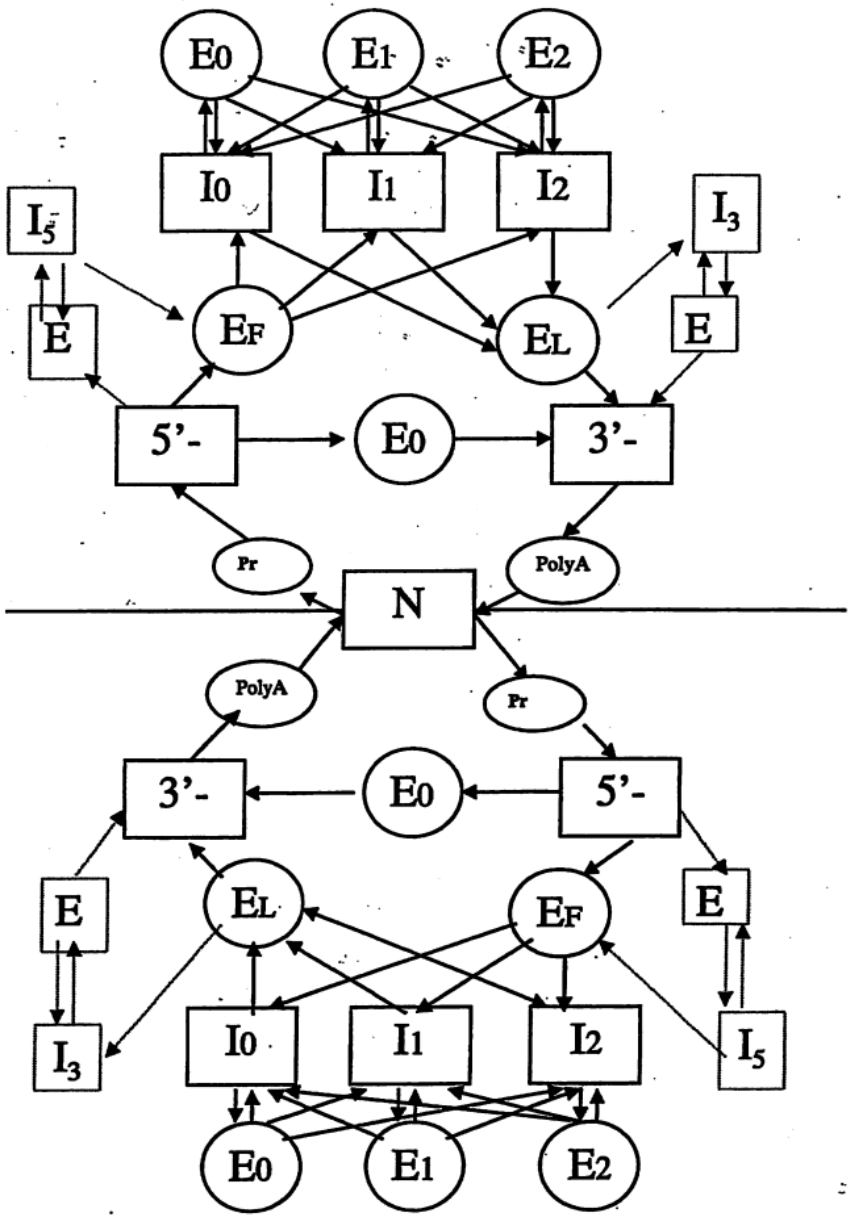
[Fickett & Tung, Nucl. Acid Res., 20:6441-6450, 1992]

Eukaryontische Genvorhersage

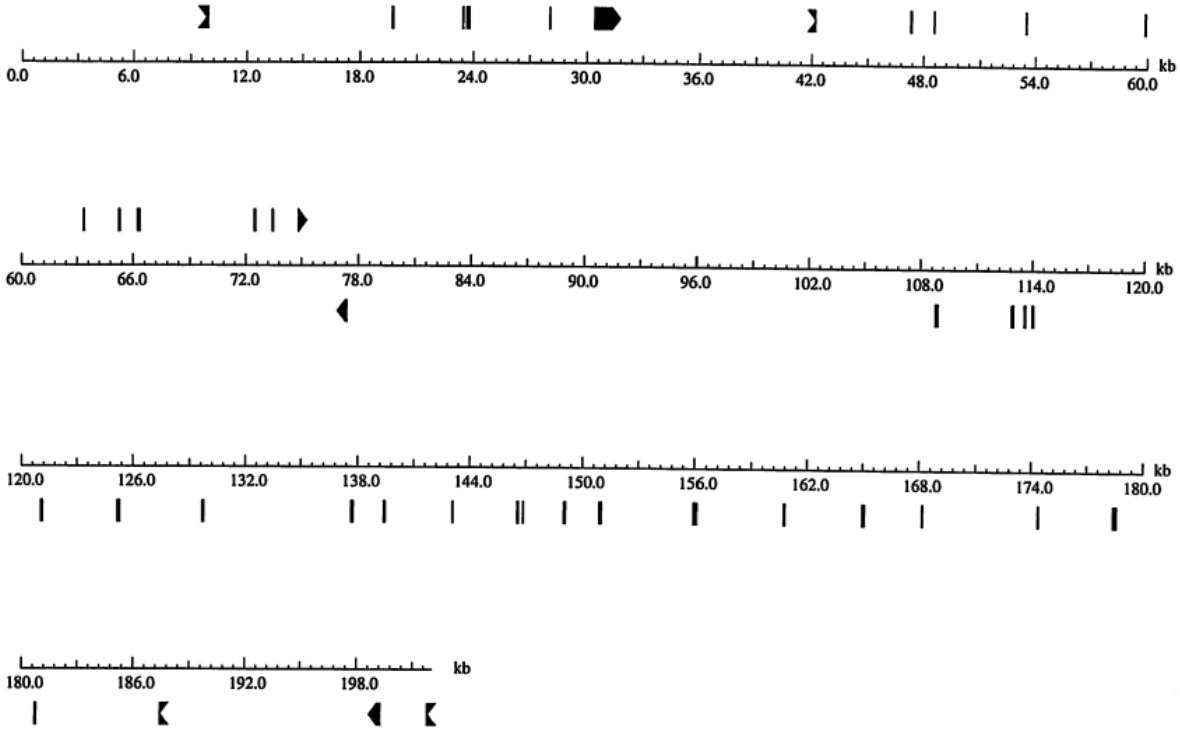
See:

Gene finding: putting the parts together

Anders Krogh


E0
E1
E2
I0
I1
I2
I5
I3
E
EF
EL
E
5'-
E0
3'-
Pr
PolyA
N
PolyA
Pr
3'-
E0
5'-
E
EL
EF
E
I3
I0
I1
I2
I5
E0
E1
E2

GENSCAN predicted genes in sequence 02:01:14



Key: Initial exon; Internal exon; Terminal exon; Single-exon gene; Optimal exon; Suboptimal exon

GENSCANW output for sequence 02:01:14

GENSCAN 1.0 Date run: 24-Oct-102 Time: 02:03:04

Sequence 02:01:14 : 202056 bp : 44.59% C+G : Isochore 2 (43 - 51 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	9824	9943	120	1	0	39	113	126	0.524	10.04
1.02	Intr	+	19649	19777	129	1	0	82	92	18	0.833	2.49
1.03	Intr	+	23395	23494	100	0	1	95	86	62	0.975	6.38
1.04	Intr	+	23607	23802	196	1	1	64	61	117	0.313	5.07
1.05	Intr	+	28007	28083	77	2	2	36	106	12	0.218	-3.04
1.06	Term	+	30359	31323	965	0	2	91	54	388	0.776	27.92
1.07	PlyA	+	31358	31363	6							1.05
2.00	Prom	+	37105	37144	40							-5.16
2.01	Init	+	42199	42206	8	0	2	110	106	2	0.946	4.56
2.02	Intr	+	47279	47384	106	2	1	96	116	41	0.954	7.92
2.03	Intr	+	48522	48578	57	2	0	78	86	40	0.785	1.88
2.04	Intr	+	53429	53518	90	1	0	32	111	79	0.962	4.79
2.05	Intr	+	59812	59910	99	0	0	124	95	-28	0.362	1.81
2.06	Intr	+	63313	63405	93	0	0	112	37	79	0.859	5.36
2.07	Intr	+	65204	65334	131	1	2	105	92	20	0.880	3.59
2.08	Intr	+	66207	66391	185	0	2	86	86	56	0.818	4.63
2.09	Intr	+	72429	72549	121	1	1	84	10	125	0.562	3.85
2.10	Intr	+	73403	73493	91	2	1	26	116	26	0.210	-0.80
2.11	Term	+	74801	74836	36	1	0	87	48	39	0.391	-2.86
2.12	PlyA	+	75397	75402	6							1.05
3.24	PlyA	-	77218	77213	6							1.05
3.23	Term	-	77420	77314	107	0	2	93	47	82	0.456	3.17
3.22	Intr	-	108960	108765	196	0	1	104	99	158	0.997	17.49
3.21	Intr	-	113011	112848	164	2	2	120	64	73	0.853	7.79
3.20	Intr	-	113669	113533	137	1	2	56	110	33	0.953	2.51
3.19	Intr	-	114103	113961	143	1	2	92	93	80	0.971	8.05
3.18	Intr	-	121158	121003	156	0	0	85	94	83	0.995	8.81
3.17	Intr	-	125318	125131	188	0	2	129	72	235	0.998	25.41
3.16	Intr	-	129778	129649	130	1	1	72	105	114	0.787	11.77
3.15	Intr	-	137798	137637	162	2	0	59	45	185	0.454	11.47
3.14	Intr	-	139516	139381	136	0	1	118	99	-1	0.718	4.57
3.13	Intr	-	143134	143061	74	1	2	78	105	-21	0.525	-3.20
3.12	Intr	-	146649	146532	118	2	1	102	86	112	0.980	12.87
3.11	Intr	-	146903	146845	59	2	2	106	75	25	0.921	0.68
3.10	Intr	-	149129	148989	141	2	0	76	67	60	0.889	3.25
3.09	Intr	-	151051	150884	168	1	0	107	100	125	0.885	15.74
3.08	Intr	-	156149	155934	216	2	0	112	82	257	0.984	26.20
3.07	Intr	-	160859	160736	124	1	1	111	87	103	0.984	13.09
3.06	Intr	-	165138	164957	182	0	2	70	-76	176	0.161	-1.93
3.05	Intr	-	168253	168153	101	2	2	93	100	-42	0.665	-2.67
3.04	Intr	-	174458	174355	104	1	2	85	93	69	0.880	6.92
3.03	Intr	-	178629	178394	236	0	2	103	76	333	0.918	30.09
3.02	Intr	-	180812	180681	132	2	0	44	73	261	0.591	21.04

3.01	Init	-	187463	187425	39	2	0	91	110	11	0.359	3.79
3.00	Prom	-	196987	196948	40							-6.56
4.03	PlyA	-	197231	197226	6							1.05
4.02	Term	-	199318	199101	218	2	2	41	37	154	0.921	2.91
4.01	Init	-	201857	201785	73	2	1	55	94	110	0.961	7.53

Click here to view a PDF image of the predicted gene(s)

Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

```
>02:01:14|GENSCAN_predicted_peptide_1|528_aa
MVKLSIVLTPQFLSHDQGQLTKELQQHVKSVTCPCEYLRKVINTLADHHHRGTDFGGSPW
LHVIIAFPTSYKVVITLWIVYLWVSLLKTIFWSRNGHDGSTDVQQRAWRSNRRRQEGLRS
ICMHTKKRVSSFRGNKIVLKDVITLRRHVETKVRAKIRKRKVTTKINHHDKINGKRKTAR
KQLSQHSISHVLAFSDPPFCKKGSLQLAPPSADDNIKIPAERLRIPLPPSADDNLKTPSE
RQLTPLPPSAPPSADDNIKTPAERLRGPLPPSADDNLKTPSERQLTPLPPSAPPSADDNI
KTPAERLRGPLPPSADDNLKTPSERQLTPLPPSAPPSADDNIKTPAERLRGPLPPSADDN
LKTPSERQLTALPPSAPPSADDNIKTPAERLRGPLPPSADDNLKTPPLATQEAEAEKPRK
PKRQRAAEMEPPPEPKRRRVGDVEPSRKPKRRRAADVEPSSPEPKRRRVGDVEPSRKPKR
RRAADVEPSSPEPKRRRVGDVEPSRKPKRRRAADVEPSLPEPKRRRLS

>02:01:14|GENSCAN_predicted_peptide_2|338_aa
MPGISNMRALENDFFNSPPRKTVRFGGTVTEVLLKYKKGETNDFELLKNQLLDPDIKDDQ
IINWLLEFRSSVMYLTKDFEQLISIILECYVHNLLRISVYFPTLRHEILELIIEKLLKLD
VNASRQGIEDAEETANQTCGGTDSTEGLFNMGFAEAFLEHLWKNLQDPSNPAIIRQAAGN
YIGSFLARAKFISLITVKPCLDLLVNWLHIYLNNQDSGTKAFCDVALHGPFYSACQAVFY
TFVFRHKQLLSGNLKEVSLMTEHLAGDGKRCSTEHHPNITRASADPQLTADEEAQPRELS
WQMGTVSSLGKGHRRLCILRTHNHNGKKLHKDIPAPSA

>02:01:14|GENSCAN_predicted_peptide_3|1070_aa
MECRVIQCQIPGRAAVENHLEQRLHQPQKLLEDLRKTDAQQFRTAMKCLLEDKKDGLDLK
DIIIDLGEIRERALQSPGVNRSLFLITLERCFQMLNSLECVEILGKVLRGSSGSFLQPDI
TERLPRDLREDAFKNLSAVFKDLYDKTSAHSQRALYSWMTGILQTSSNATDDSASWVSAE
HLWVLGRYMVHLSFEEITKISPIEIGLFISYDNATKQLDMVYDITPELAQAFLERISSSN
FNMRNTSTIHRQAHELWALEPFPKMLGLLVCFYNDLELLDATVAQVLLYQMIKCSHLRGF
QAGVQKLKAELLDIAMENQTLNETLGSLSDAVVGLTYSQLESLSPEAVHGAISTLNQVSG
WAKSQVIILSAKYLAHEKVLSFYNVSQMGALLAGVSTQAFCSMKRKDISQVLRSAVSQYV
SDLSPAQQQGILSKMVQAEDTAPGIVEIQGAFFKEVSLFDLRRQPGFNSTVLKDKELGRS
QALFLYELLLKTTRRPEELLSAGQLVKGVTCSHIDAMSTDFFLAHFQDFQNNFALLSPYQ
VNCLAWKYWEVSRLSMPPFLLAALPARYLASVPASQCVPFLISLGKSWLDSLVLDSHKKT
SVLRKVQQCLDDSIADEYTVDIMGNLLCHLPAAIIDRGISPRAWATALHGLRDCPDLNPE
QKAAFPEILLQAASKMARTLPTKEFLWAVFQSVRNSSDKIPSYDPMPGCHGVVAPSSDDI
FKLAEANACWALEDLRCMEEDTFIRTVELLGAVQGFSRPQLMTLKEKAIQVWDMPSYWRE
HHIVSLGRIALALNESELEQLDLSSIDTVASLSWQTEWTPGQAESILQGYLDDSGYSIQD
LKSFHLVGLGATLCAINITEIPLIKISEFRVVVARIGTLLCSTHVLAEFKRKAEVVFGDP
TEWTSSVLQELGTIAAGLTKAELRMLDKDLMPYFQPSAIKCLPDEIFKVGAQFFKEKWEL
DPISNHTGKQELSAEQIASLGPENAAAVTHAQRRRLSPLQLQSLQQALDGAKTHSWQDAP
ASAGPTRTSSSRSPAEVLLPNESFAFYPGMRFSENPTNMYSISCTQSSKD

>02:01:14|GENSCAN_predicted_peptide_4|96_aa
MSPAAPPCATRPPVGLSAAKPGSAVGESFSGEGVWPIEKAIDLQIIIALPSPFSPFDQSA
KKVTQSAFNENQILMGCSFSLEKGHYAQQKARRLFQ
```

Explanation

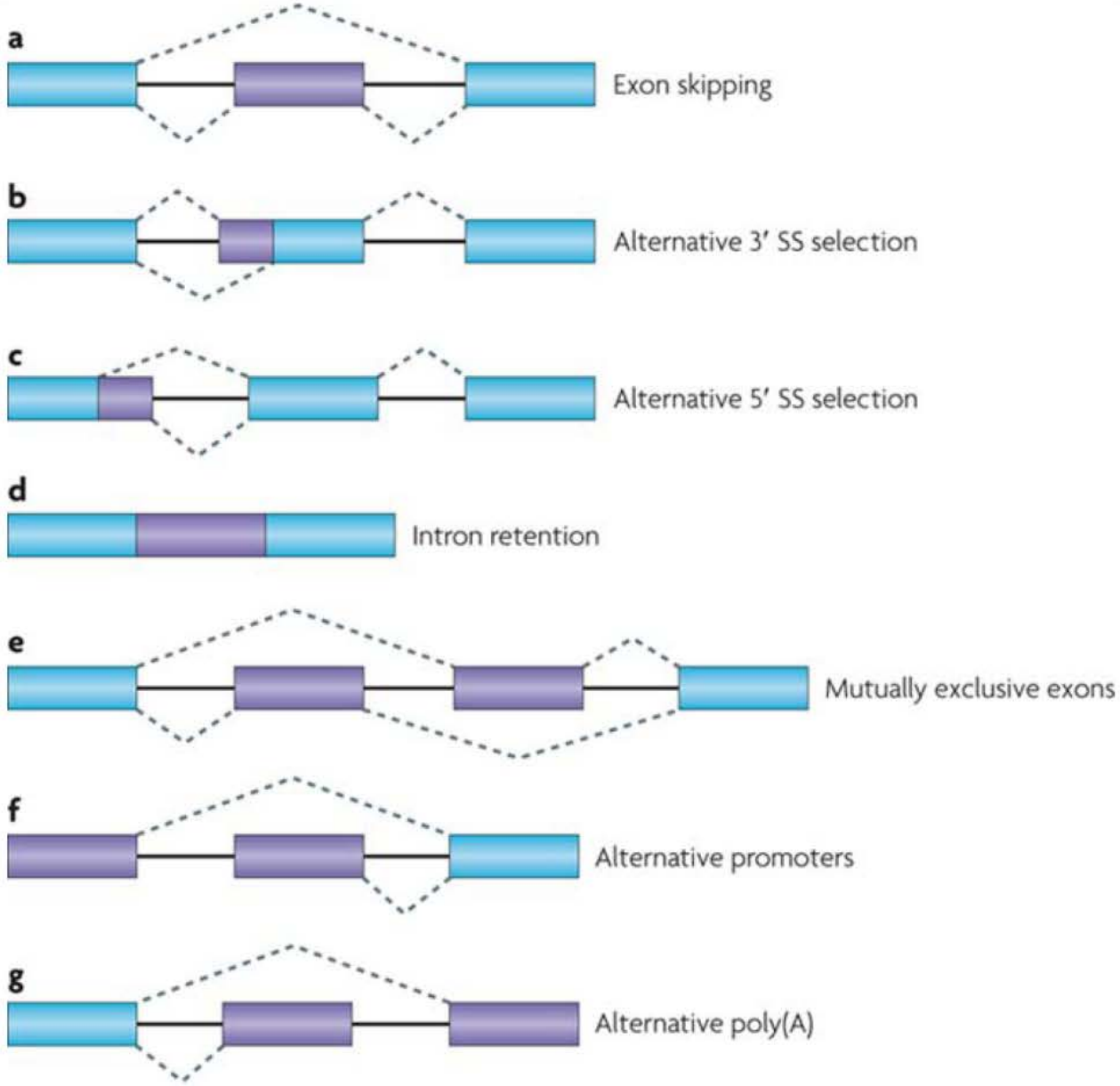
Gn.Ex : gene number, exon number (for reference)

```
Type  : Init = Initial exon (ATG to 5' splice site)
        Intr = Internal exon (3' splice site to 5' splice site)
        Term = Terminal exon (3' splice site to stop codon)
        Sngl = Single-exon gene (ATG to stop)
        Prom = Promoter (TATA box / initation site)
        PlyA = poly-A signal (consensus: AATAAA)
S     : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exon or signal (numbered on input strand)
End   : end point of exon or signal (numbered on input strand)
Len   : length of exon or signal (bp)
Fr    : reading frame (a forward strand codon ending at x has frame x
mod 3)
Ph    : net phase of exon (exon length modulo 3)
I/Ac  : initiation signal or 3' splice site score (tenth bit units)
Do/T  : 5' splice site or termination signal score (tenth bit units)
CodRg : coding region score (tenth bit units)
P     : probability of exon (sum over all parses containing exon)
Tscr  : exon score (depends on length, I/Ac, Do/T and CodRg scores)
```

Comments

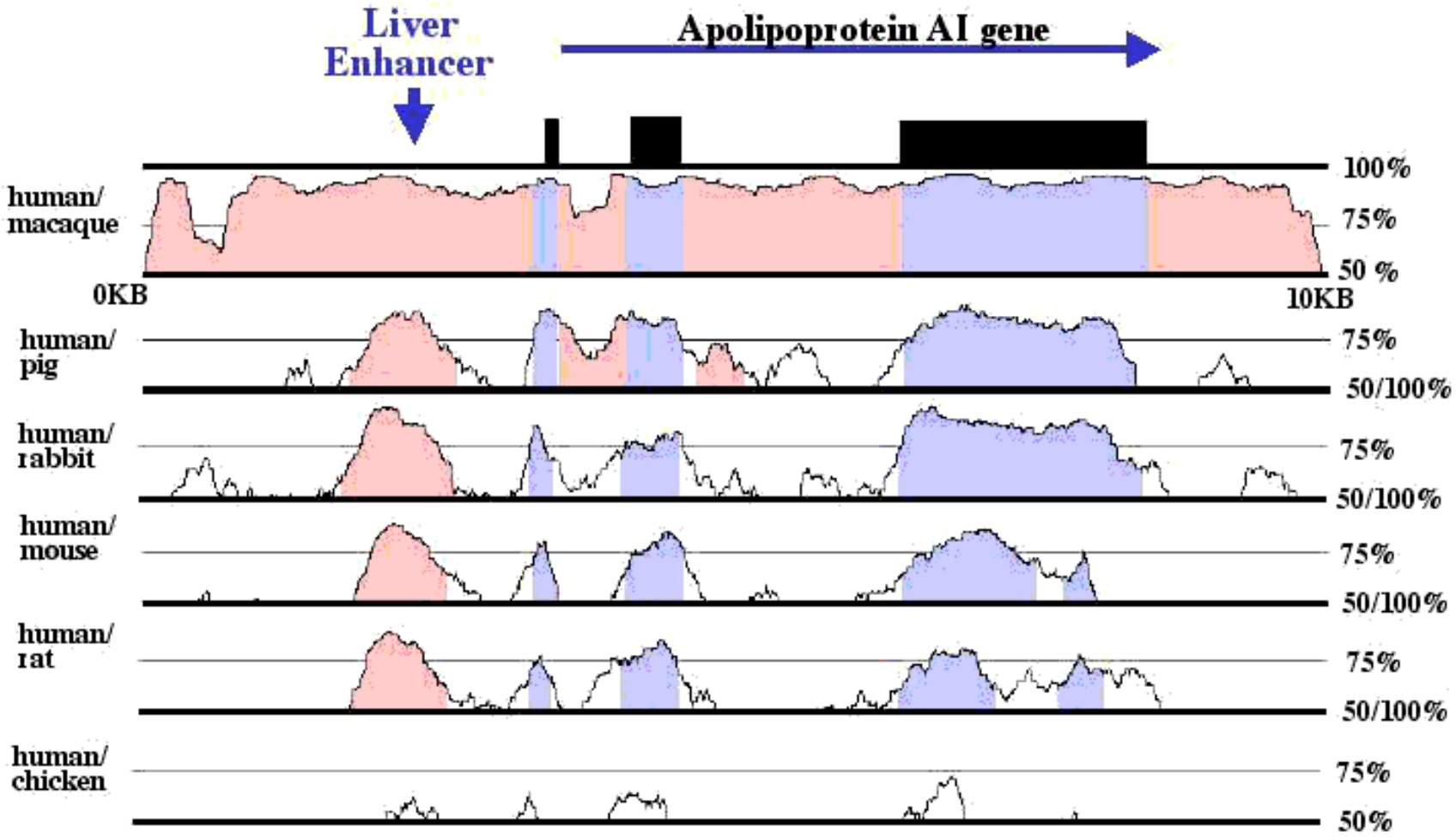
The SCORE of a predicted feature (e.g., exon or splice site) is a log-odds measure of the quality of the feature based on local sequence properties. For example, a predicted 5' splice site with score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor (more than likely not a real donor site).

The PROBABILITY of a predicted exon is the estimated probability under GENSCAN's model of genomic sequence structure that the exon is correct.
This probability depends in general on global as well as local sequence
properties, e.g., it depends on how well the exon fits with neighboring
exons. It has been shown that predicted exons with higher probabilities
are more likely to be correct than those with lower probabilities.



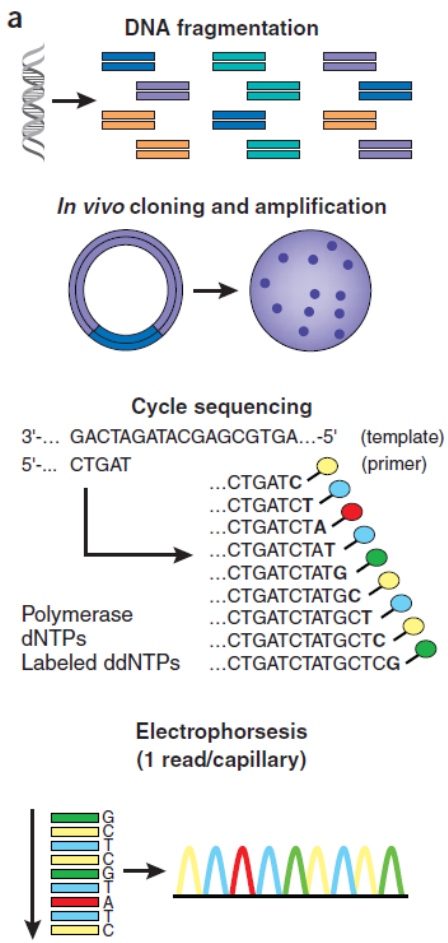
Nature Reviews | Genetics

Multi-Species Comparative Analysis



Sequenzierung

Sanger sequencing



- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags

Kurze Geschichte

- Sequenzierung (klonierter) genomischer Abschnitte
- Sequenzierung von cDNA
- Sequenzierung kompletter Genome
  - Hefe (S. cervisiae), Wurm (C. elegans), Fliege (Drosophila melanogaster), Maus, Mensch, …
- EST Sequenzierung: EST = expressed sequence tag, Sequenzierung von Bruchstücken der mRNAs

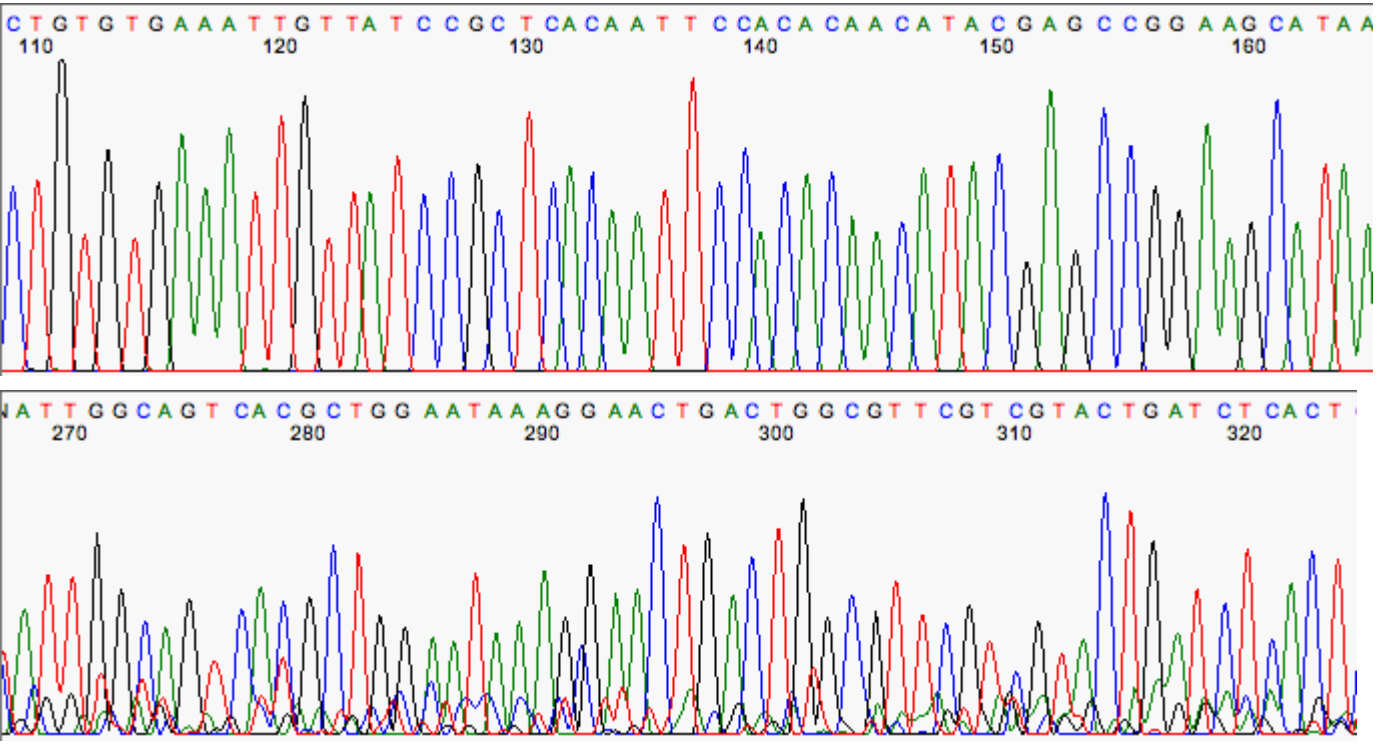
Shotgun sequencing & Assembly

- Sequence reads ca 500-800 Basen lang
- Große DNA Stücke, z.B. BACs, Bacterial artifical chromosome. Länge 100-300 kb.
- Zerlegen und klonieren: Clone. Insert einige 1000 bp. Von einer oder von beiden Seiten ansequenzieren.

- Wikipedia: „Shotgun sequencing“, „DNA sequencing theory“

Read quality

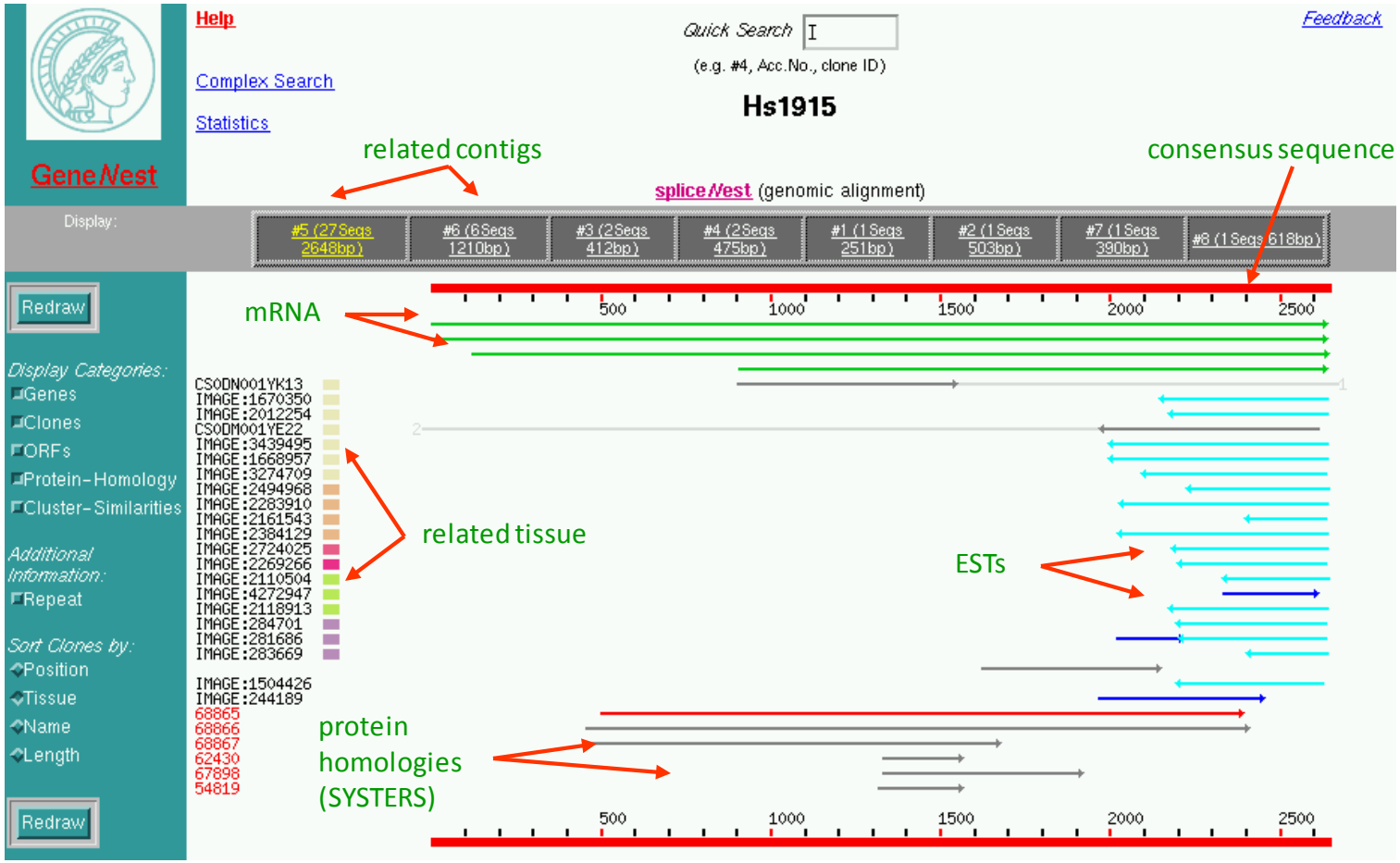
- Fehler am Ende eines reads (-> „clipping“)
- Schlechte Auflösung von Homopolymer-runs



http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.html

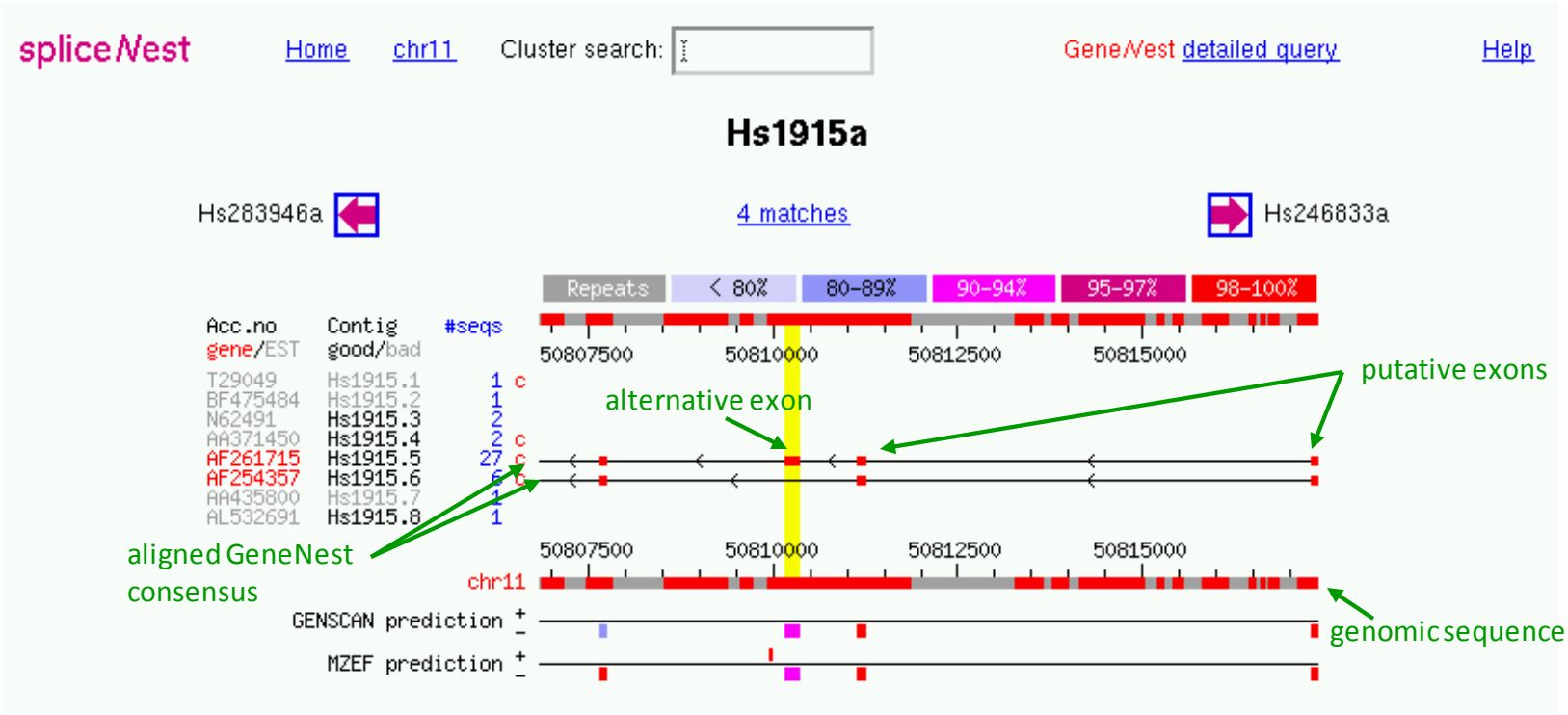
GeneNest visualization

(http://GeneNest.molgen.mpg.de)



SpliceNest

(http://SpliceNest.molgen.mpg.de)



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

2 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

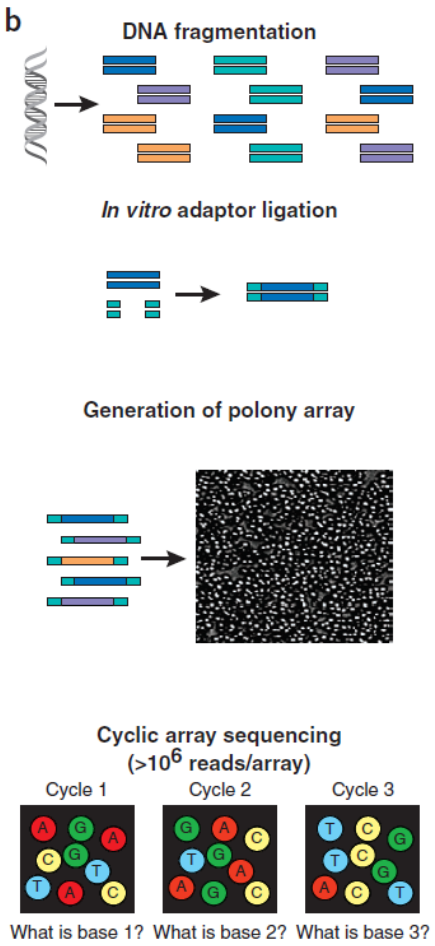
Multiz Alignments of 46 Vertebrates

Next Generation Sequencing

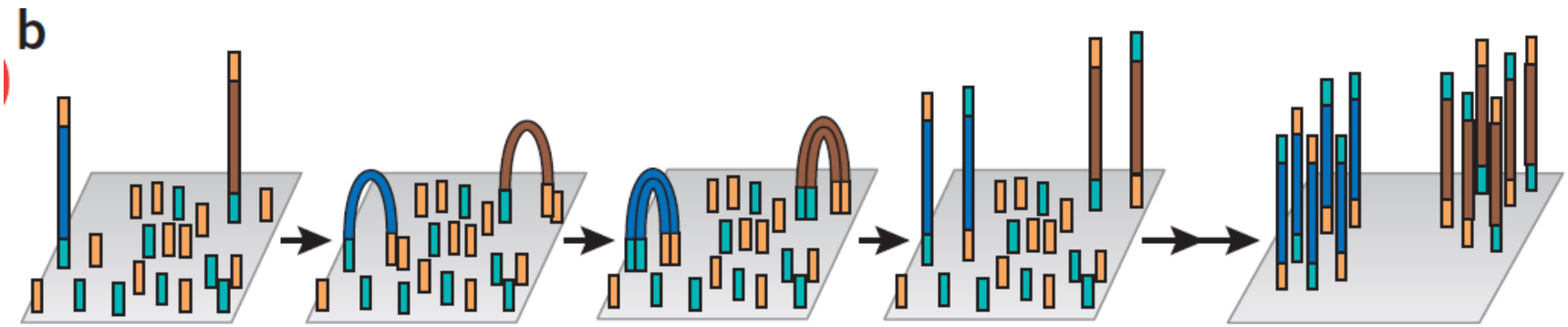
- Illumina, ABI 454, Solid (Roche)
- Read length: ~100nt, possibly paired end
- 100 million reads in one experiment

Cyclic-array methods

- DNA is fragmented
- Adaptors ligated to fragments
- Several possible protocols yield array of PCR colonies.
- Enyzmatic extension with fluorescently tagged nucleotides.
- Cyclic readout by imaging the array.

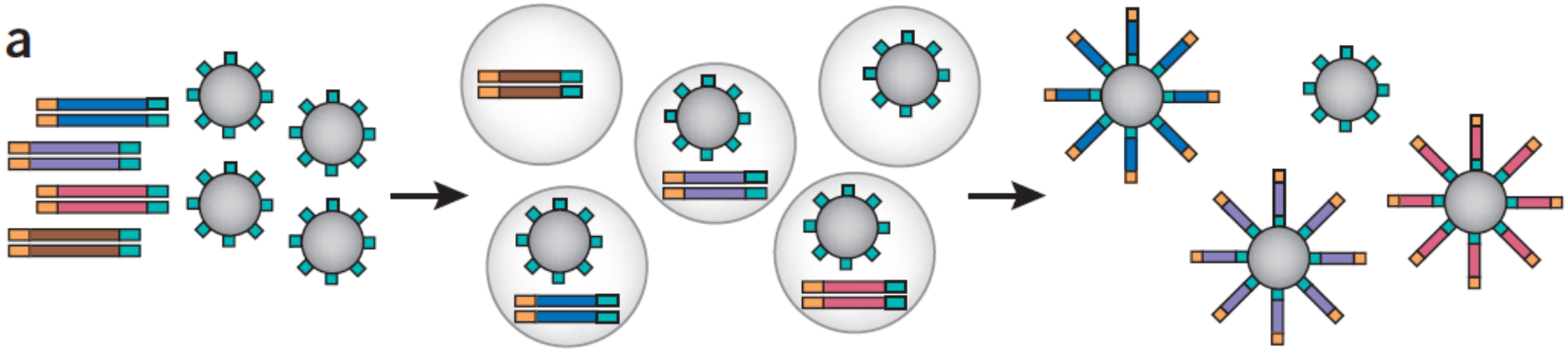


Bridge PCR



- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa/Illumina.

Emulsion PCR



- Fragments, with adaptors, are PCR amplified within a water drop in oil.
- One primer is attached to the surface of a bead.
- Used by 454, Polonator and SOLiD.

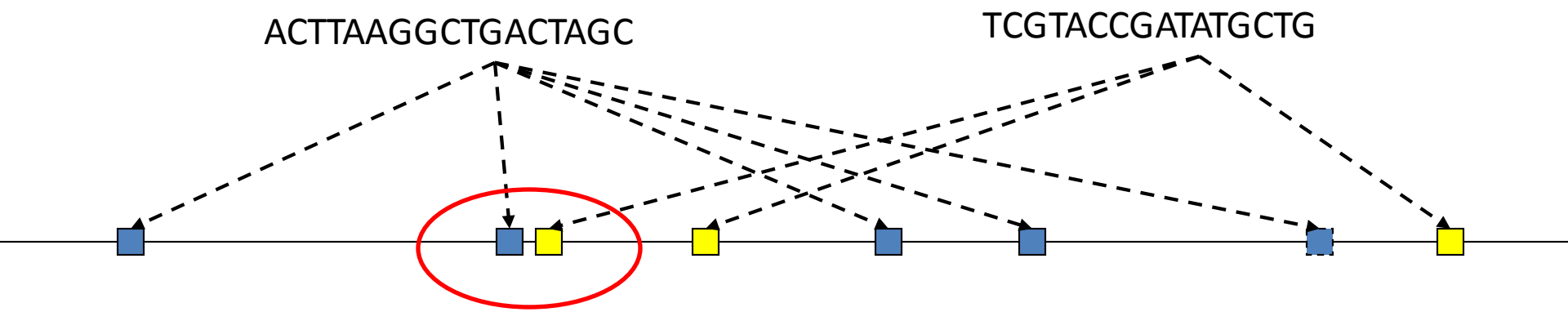
Resultierende Daten 2008-heute

- Anfangs ca 30bp lange reads
- 30bp paired-end reads
- Dann 70-100bp
- Heute: 100bp, paired end, 70+Mio reads in einem Experiment (=1 flow cell, Illumina), Dauer mehrere Tage
- Mehr Fehler als bei Sanger Sequenzierung – kompensiert durch höhere Abdeckung

Resultierende Verarbeitungsprobleme 2008-heute

- Ca 30bp ---- Assembly fast unmöglich, stattdessen mapping auf bekanntes Genom
- 30bp paired end reads --- Assembly immer noch schwierig, paired ends machen mapping besser
- 70-100bp
- Heute: 100bp, paired end, 70Mio reads in einem Experiment (=1 flow cell, Illumina), mehrere Tage ---- Assembly schwer, aber möglich. Mapping mit mismatches, Repeats zum Teil auflösbar.

Read length and pairing



- Short reads are problematic, because short sequences do not map uniquely to the genome.
- Solution #1: Get longer reads.
- Solution #2: Get paired reads.

Mapping Software

- BLAST zu langsam (Vorverarbeitung der query)
- Hashing: k-mer index for seeds.
- Suffix trees, suffix arrays: Vorverarbeitung des Textes. Speicherbedarf ist ein Mehrfaches des Genoms.
  - Suffix tree: 10-20fach; suffix array: 8fach
  - Beispiel: Humangenom 3 GB, Suffix tree mehr als 30GB, suffix array 24GB.
  - Wieviel RAM hat Ihr Computer?

Reminder: Secondary Storage Data Structures

- Data structure resides on disk
- B-trees (1972), string B-tree (1996)
- Suffix arrays were designed to reside on disk (not any more)

- Secondary Storage Data Structures sind nicht schnell genug für read mapping! Datenstruktur muss in RAM passen.

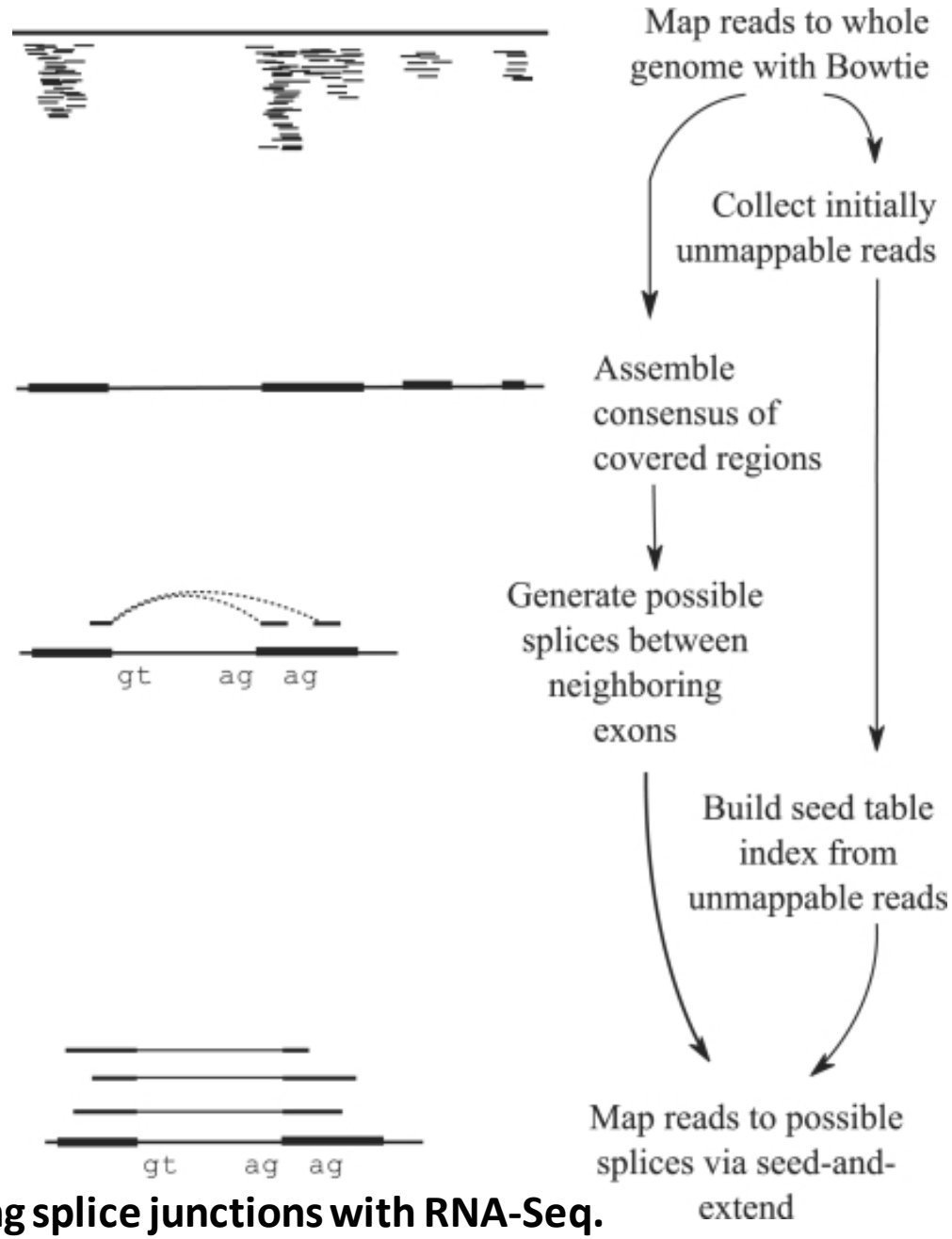
Software

- Erste Generation: eland (hashing), vmatch, …
- SOAP, MAQ (hashing)
- Bowtie, SOAP2, BWA … Burrows-Wheeler transform

- Bowtie uses as little as 1.3GB of RAM for the index of the human genome (according to the authors, see Table 5)
- See: “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, by Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg. Genome Biology 2009

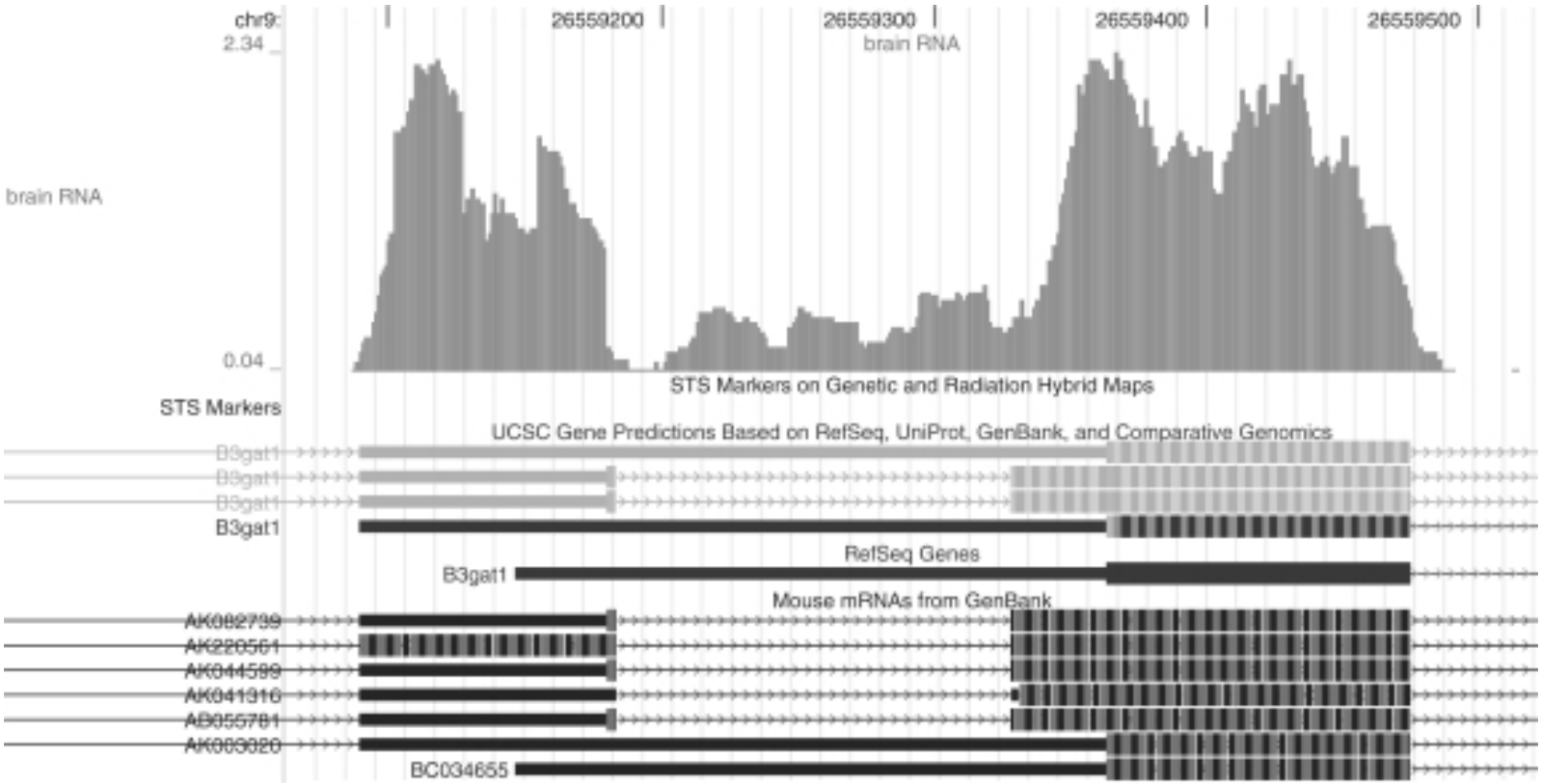
Burrows-Wheeler transform & FM index

- BW Transform is a string (of equal length to the text).
  - BWT can be transformed back into the text
  - BWT can be compressed efficiently
- FM Index: Allows counting and searching of strings in the BWT. By Ferragina and Manzini (2000), but FM stands for „Full text index in Minute space“
- See Intro be Ben Langmead: „Introduction to the Burrows-Wheeler Transform and FM Index”, bwt_fm.pdf

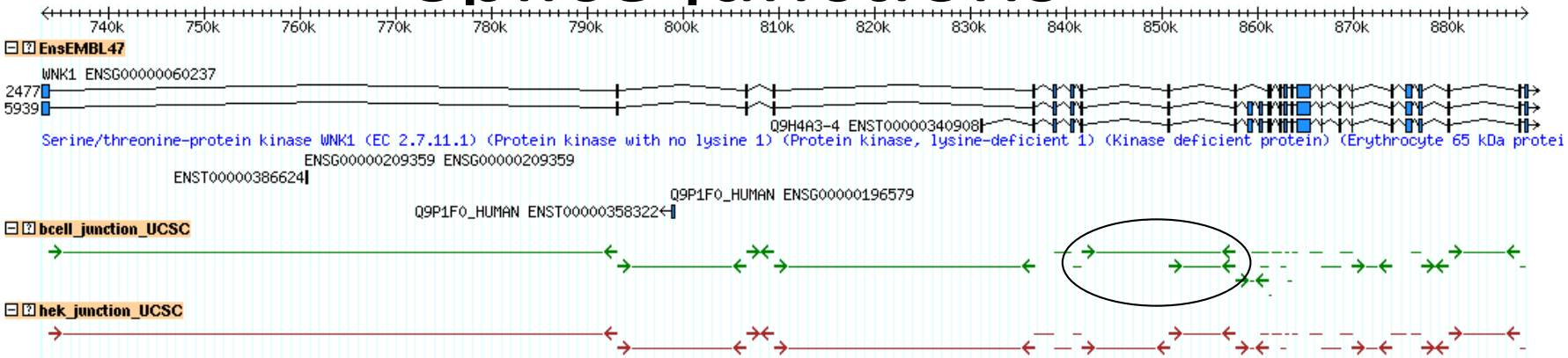


TopHat: discovering splice junctions with RNA-Seq.

Trapnell C[1], Pachter L, Salzberg SL.

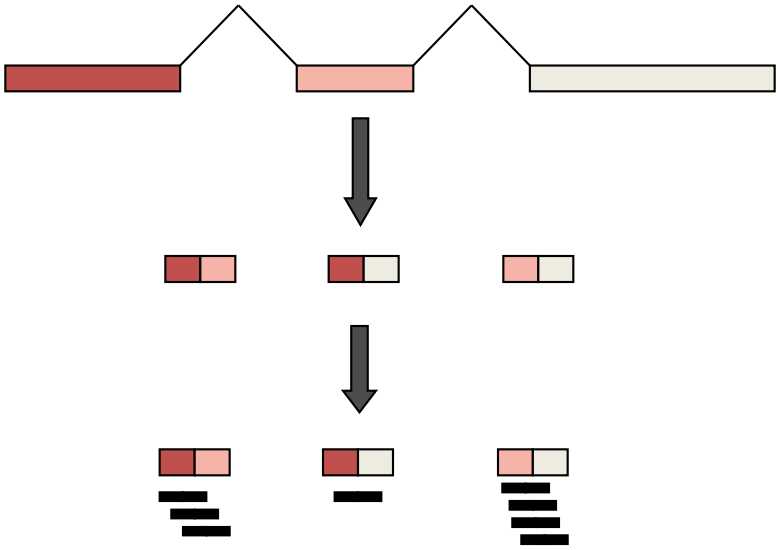

chr9:
26559200
26559300
26559400
26559500
2.34
brain RNA
brain RNA
0.04
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
B3gat1
B3gat1
B3gat1
B3gat1
RefSeq Genes
B3gat1
Mouse mRNAs from GenBank
AK082739
AK220561
AK044599
AK041916
AB055781
AK003020
BC034655

Splice junctions



Align unmatched reads to artificial junctions

(~ 2,8 x 10^6 artificial junctions)

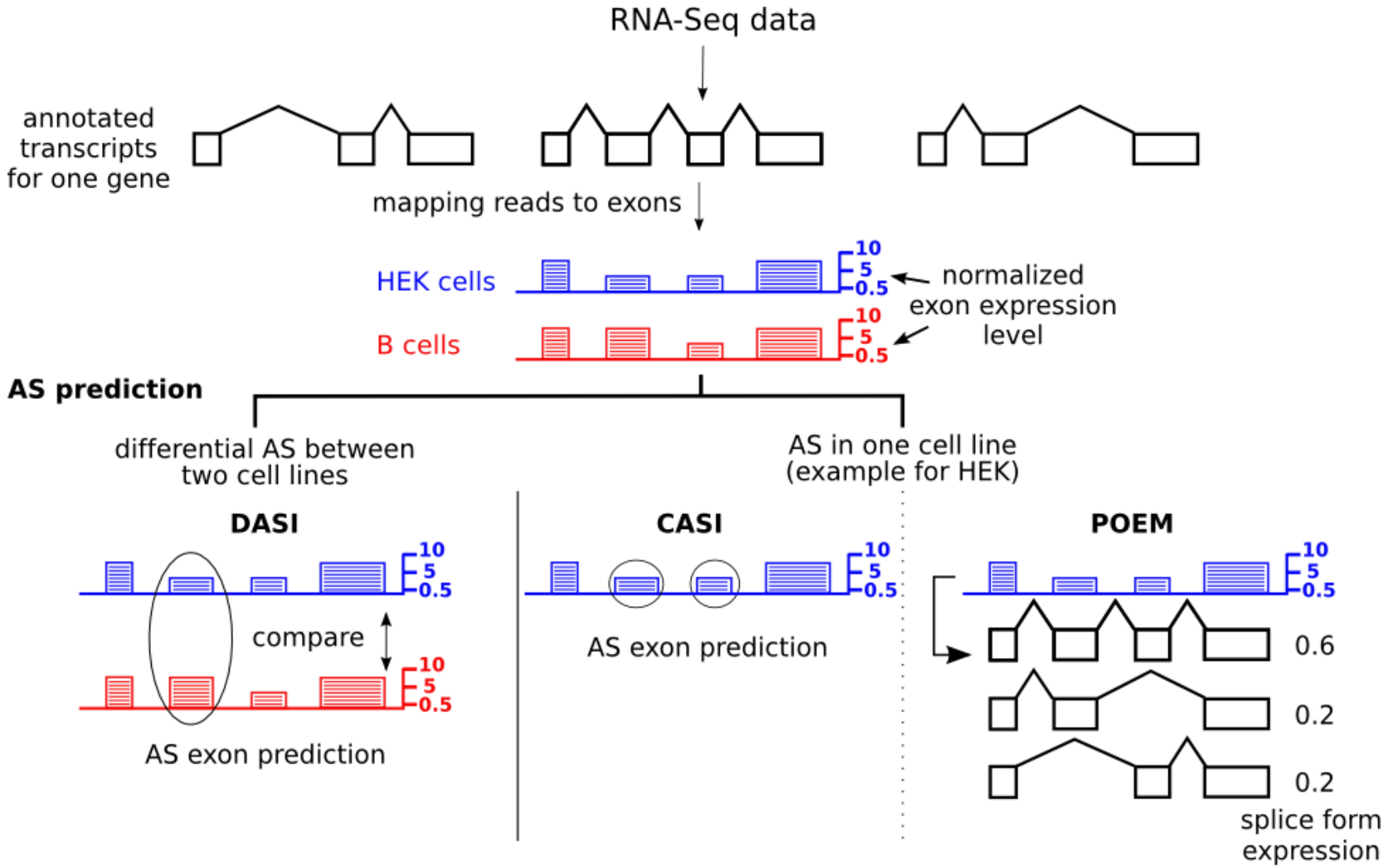


Sultan et al. (2008) A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956-960

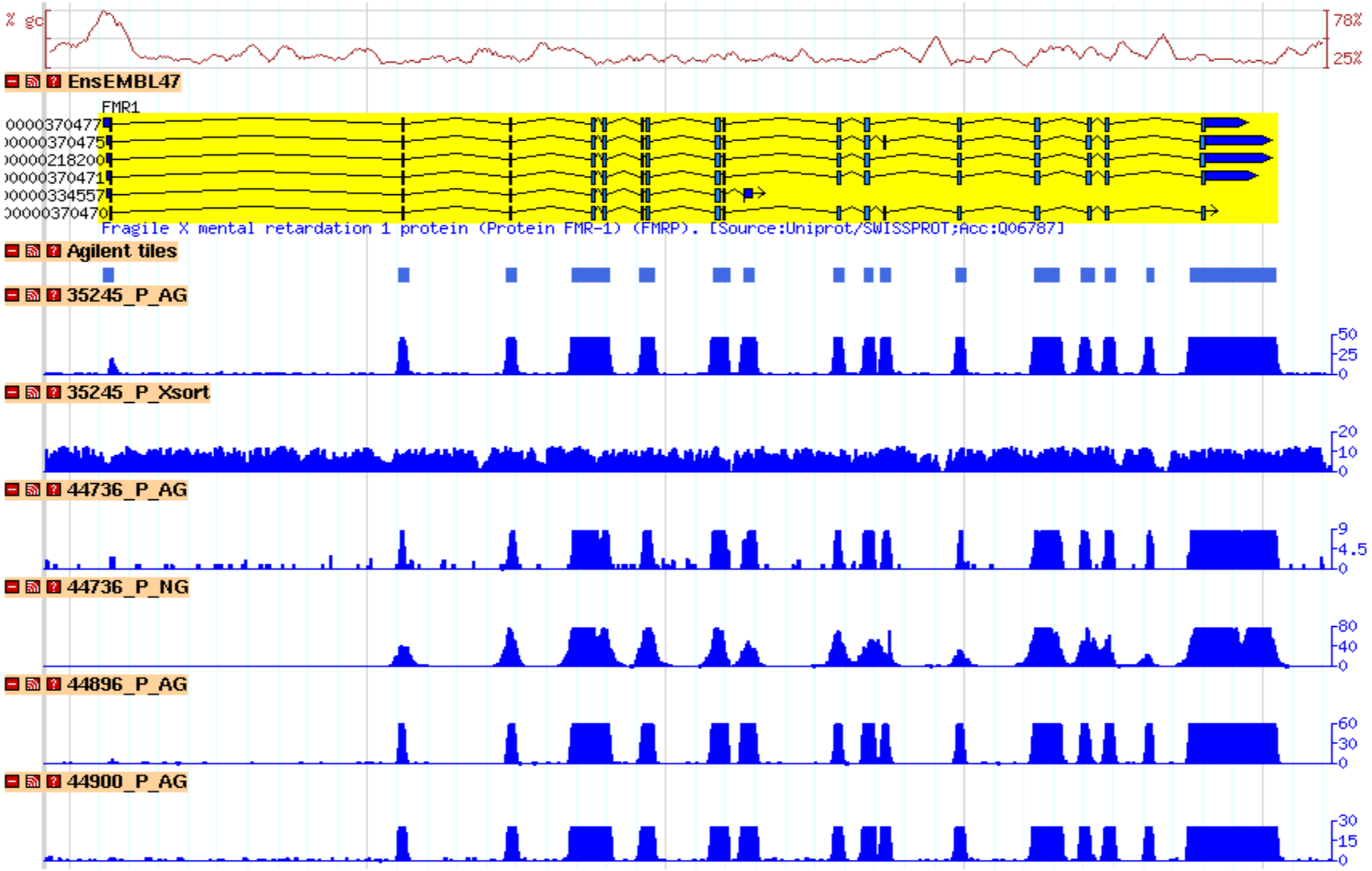
Quantifizierung und Sampling

- Angenommen, es sind ca 1/3 aller Gene in einer Zelle exprimiert. Manche häufig (viele mRNA Moleküle), andere gering (wenige mRNA Moleküle)
- ESTs: ca 100K reads aus einer cDNA Bibliothek
- RNA-seq: 100 Mio reads

Detecting alternative splicing events



FMR1


% gc
78%
25%
EnsEMBL47
FMR1
00000370477
00000370475
00000218200
00000370471
00000334557
00000370470
Fragile X mental retardation 1 protein (Protein FMR-1) (FMRP). [Source:Uniprot/SWISSPROT;Acc:Q06787]
Agilent tiles
35245_P_AG
50
25
0
35245_P_Xsort
20
10
0
44736_P_AG
9
4.5
0
44736_P_NG
80
40
0
44896_P_AG
60
30
0
44900_P_AG
30
15
0