Методы предсказания структуры генов эукариот

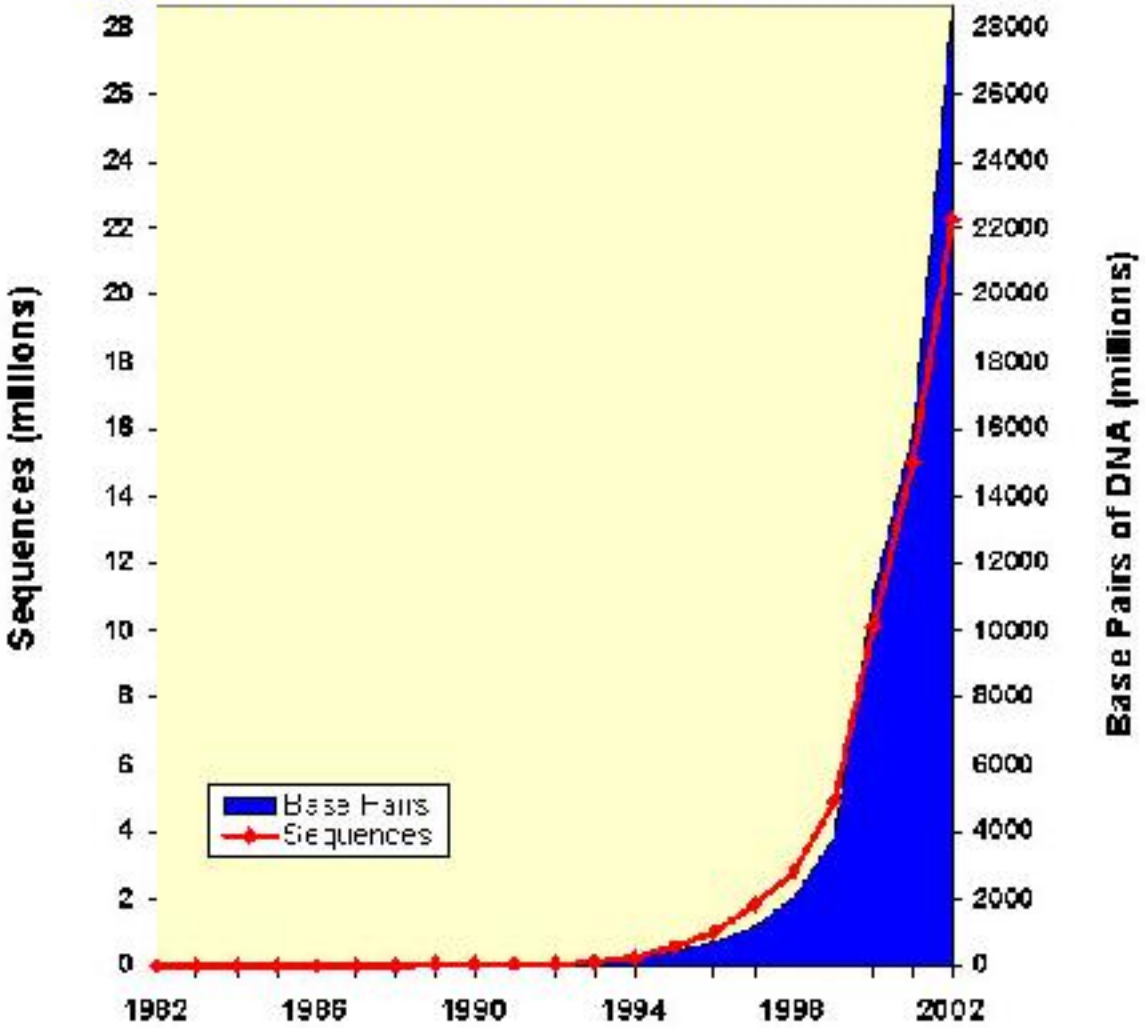
Олег Владимирович Винтевский

Кафедраинформационной биологии ФЕННГУ

Международные проекты геномных исследований

Стремительно растут темпы исследований по секвенированию геномной ДНК (Benson et al., 2000; Wheeler et al., 2000). На сентябрь 2003 г. доступны 139 полных геномов прокариот, включая 16 видов архебактерий и 123 вида бактерий.

Рост объема GenBank



http://www.ncbi.nih.gov/Genbank/genbankstats.html

GenBank за 2002 год:

28,507,990,166 п.о.

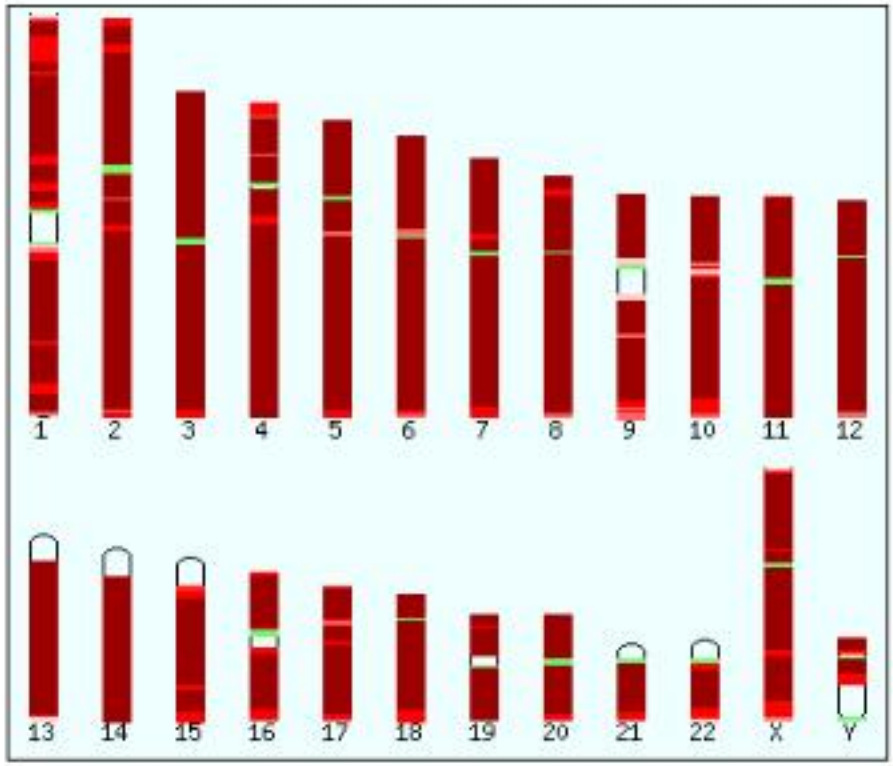
22,318,883 последовательностей

Полностью секвенированы эукариотические геномы мышевидного салата *Arabidopsis thaliana*, червя *Caenorhabditis elegans*, плодовой мушки *Drosophila melanogaster*, дрожжей *Saccharomyces cerevisiae* и *Schizosaccharomyces pombe*, некоторых внутриклеточных паразитических организмов (*Plasmodium falciparum*, *Encephalitozoon cuniculi*).

Ensembl Stats: Chromosome status

The following images indicate the status of the Golden Paths

NCBI34

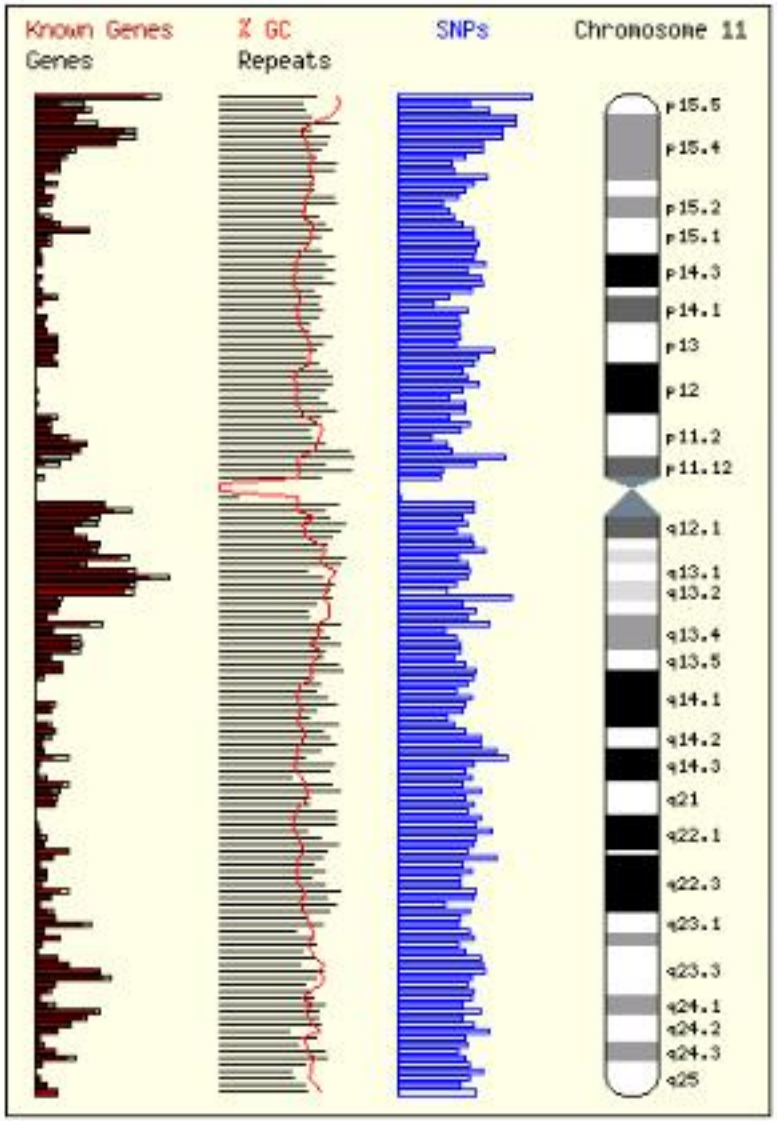


Key

- N50 length > 5mb
- N50 length > 1mb
- N50 length > 500kb
- N50 length > 100kb
- N50 length > 10kb
- More than 50% of bin
- Less than 50% of bin
- -Clear- None of bin in golden

Stats

Freeze date:
July 2003
Estimated size:
3 069.43 Mb
Total mapped:
2 843.41 Mb (92.64%)
No. of supercontigs:
350
Super contig N50 length:
29 104 799 bps
In super contigs > 10Mb
2 307.65 Mb (76 s'ctgs)
In super contigs > 1Mb
2 789.20 Mb (199 s'ctgs)
In super contigs > 100Kb
2 842.38 Mb (332 s'ctgs)



Chromosome 11

Length: 134,482,954 bps
Known Ensembl genes: 1,296
Novel Ensembl genes: 227
Snps: 245,108

Change Chromosome

Chromosome: 11 Go

Jump to Contigview

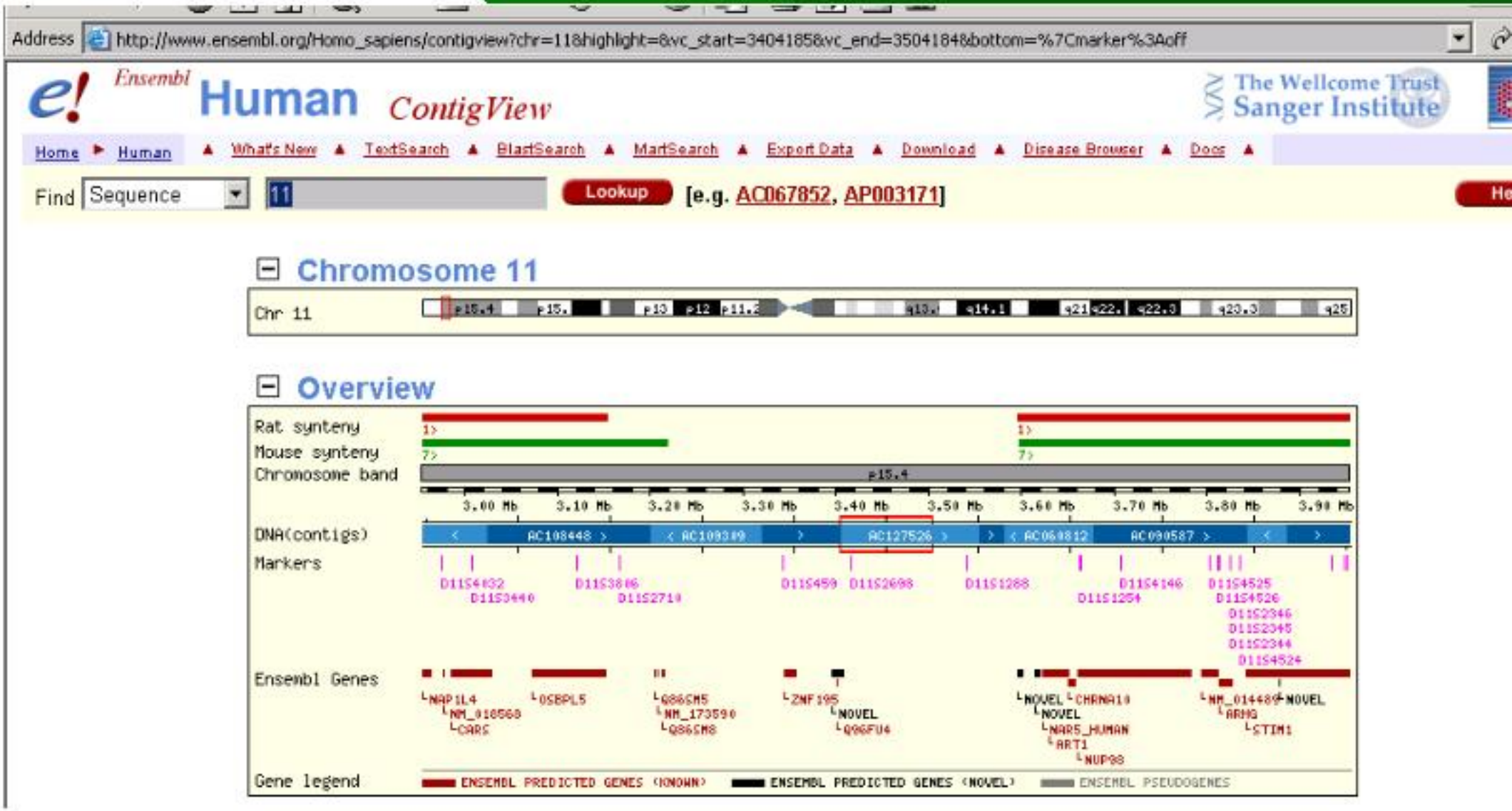
Click anywhere on the chromosome ideogram or one of the feature distribution p
level view of features at that point. Alternatively, you can jump to contigview bet
on this chromosome:

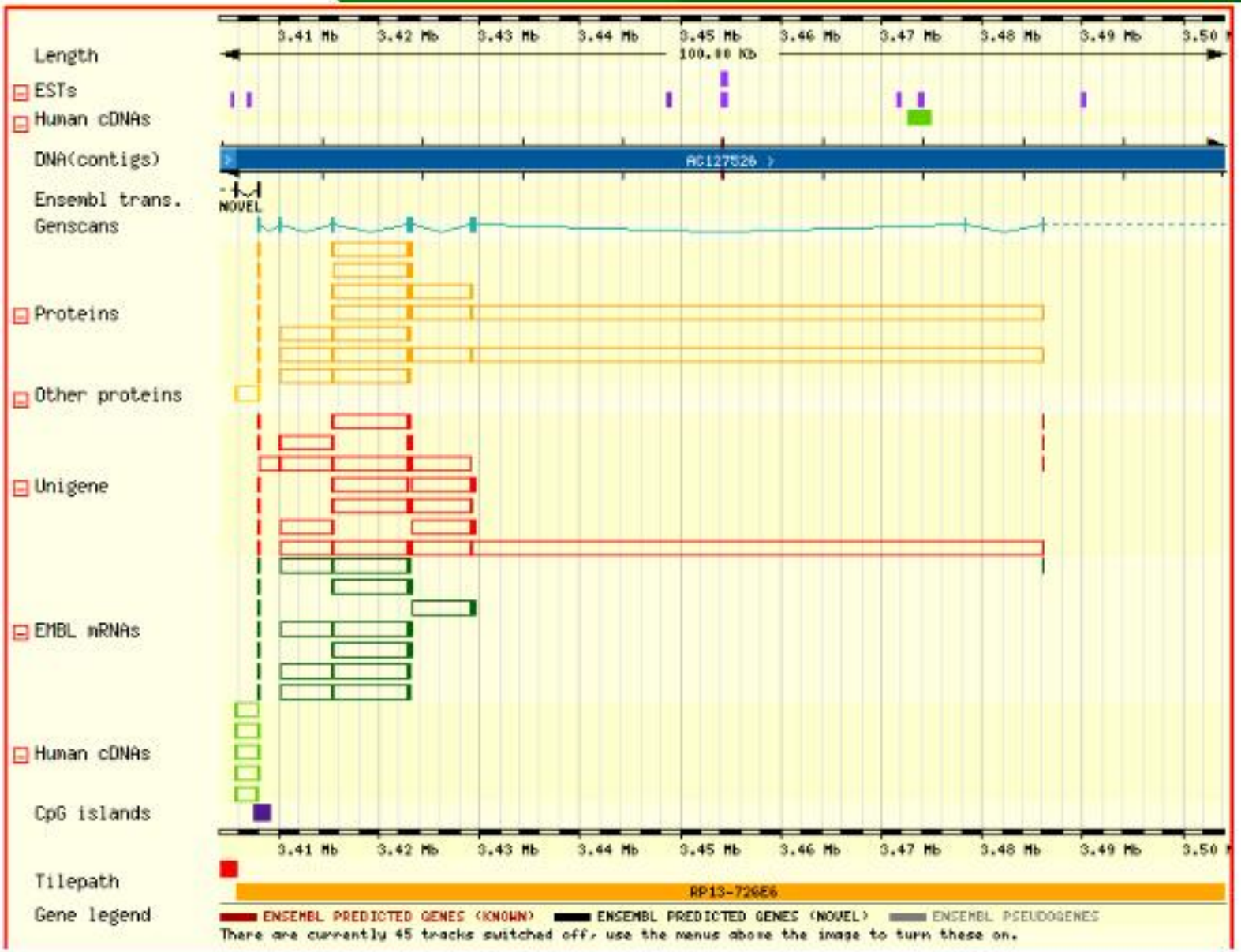
Between:
and:
Go

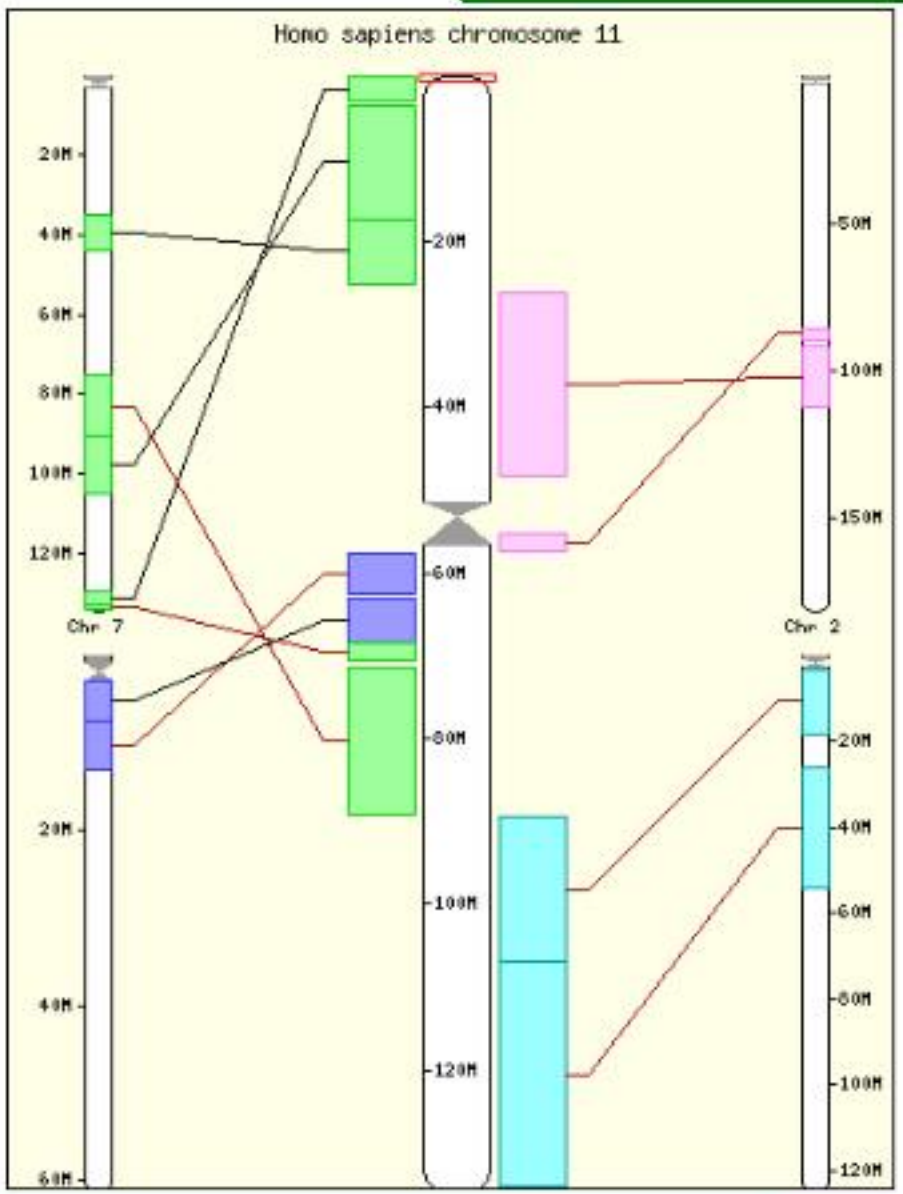
Display contig-level view between any two features.

Synteny

View Human Chr 11 vs Mouse Go


Address http://www.ensembl.org/Homo_sapiens/contigview?chr=11&highlight=&vc_start=3404185&vc_end=3504184&bottom=%7Cmarker%3Aoff
Ensembl Human ContigView
The Wellcome Trust Sanger Institute
Home Human What's New TextSearch BlastSearch MartSearch Export Data Download Disease Browser Docs
Find Sequence 11 Lookup [e.g. AC067852, AP003171]
Chromosome 11
Chr 11
Overview
Rat synteny
Mouse synteny
Chromosome band
DNA(contigs)
Markers
Ensembl Genes
Gene legend
ENSEMBL PREDICTED GENES (KNOWN)
ENSEMBL PREDICTED GENES (NOVEL)
ENSEMBL PSEUDOGENES


3.41 Mb
3.42 Mb
3.43 Mb
3.44 Mb
3.45 Mb
3.46 Mb
3.47 Mb
3.48 Mb
3.49 Mb
3.50 Mb
Length
100.00 Kb
ESTs
Human cDNAs
DNA(contigs)
AC127526 >
Ensembl trans.
Genscans
NOVEL
Proteins
Other proteins
Unigene
EMBL mRNAs
Human cDNAs
CpG islands
Tilepath
RP13-726E6
Gene legend
ENSEMBL PREDICTED GENES (KNOWN)
ENSEMBL PREDICTED GENES (NOVEL)
ENSEMBL PSEUDOGENES
There are currently 45 tracks switched off, use the menus above the image to turn these on.



Human Chromosome 11

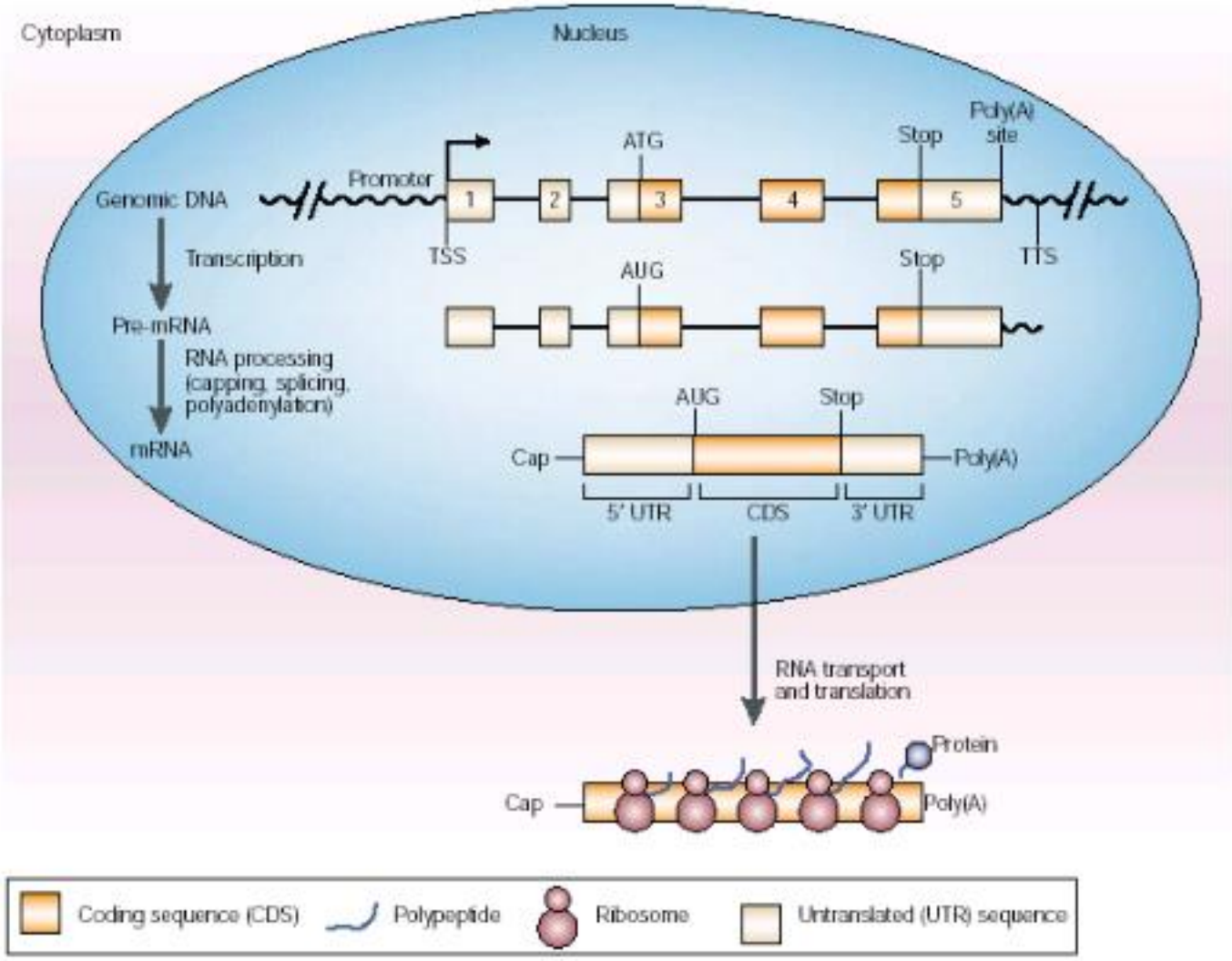
Jump to chromosome 11 Go

Jump to mapview for chromosome statistics.

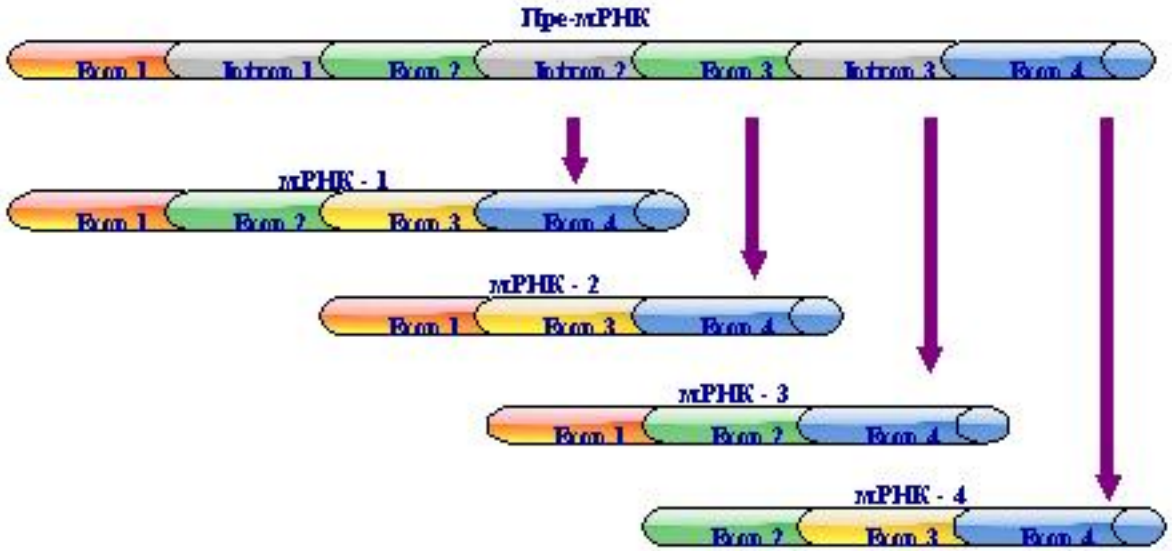
Homology Matches

Homo_sapiens Genes		*Mus_musculus* Homologues
ENSG00000170061 (149.64 Kb)		
ENSG00000177934 (152.41 Kb)		
NM_145651 (183.08 Kb)	->	ENSMUSG00000038801 (chr 7 : 129.87 Mb)
ODF3 (186.79 Kb)	->	1700011O04Rik (chr 7 : 129.87 Mb)
ENSG00000177951 (193.48 Kb)		
BET1L (195.30 Kb)	->	Bet1l (chr 7 : 129.87 Mb)
NM_021932 (198.85 Kb)	->	AI114950 (chr 7 : 129.88 Mb)
SIRT3 (206.14 Kb)	->	Sirt3 (chr 7 : 129.88 Mb)
PSMD13 (227.05 Kb)	->	Psmd13 (chr 7 : 129.90 Mb)
PYA5_HUMAN (268.57 Kb)	->	Pypaf5 (chr 7 : 129.94 Mb)
NM_025092 (279.13 Kb)	->	NM_145387 (chr 7 : 129.96 Mb)
ENSG00000142056 (288.60 Kb)	->	Hrmp1-pending (chr 7 : 129.97 Mb)

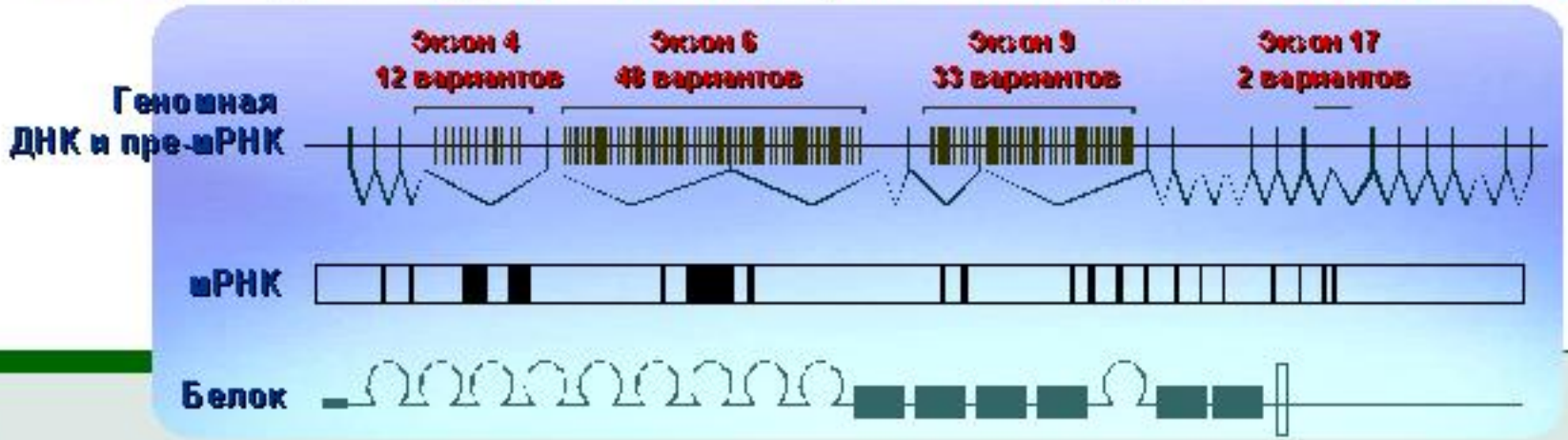
Модель эукариотического гена



Экзон-интронная структура и альтернативный сплайсинг обеспечивают огромную емкость кодирования генетической информации



Альтернативный сплайсинг обеспечивает продукцию огромного количества (более 30 000) вариантов белка DSCAM, участвующего в формировании тонкой нервной системы дрозофилы: $N = 12 \times 48 \times 33 \times 2$ ....



	Length of gene	Length of mRNA	Number of introns
β-globine	1,5	0,6	2
Insuline	1,7	0,4	2
Protéine kinase C	11	1,4	7
Albumine	25	2,1	14
Catalase	34	1,6	12
Récepteur des LDL	45	5,5	17
Facteur VIII	186	9	25
Thyroglobuline	300	8,7	36
Dystrophine*	plus de 2 000	17	plus de 50

* Une forme modifiée de ce gène provoque la dystrophie musculaire de Duchenne.
La taille des gènes indiquée ici comprend à la fois la partie transcrite du gène et les séquences d'ADN régulatrices voisines. (D'après des données fournies par Victor McKusick.)

Lengths are in kbp

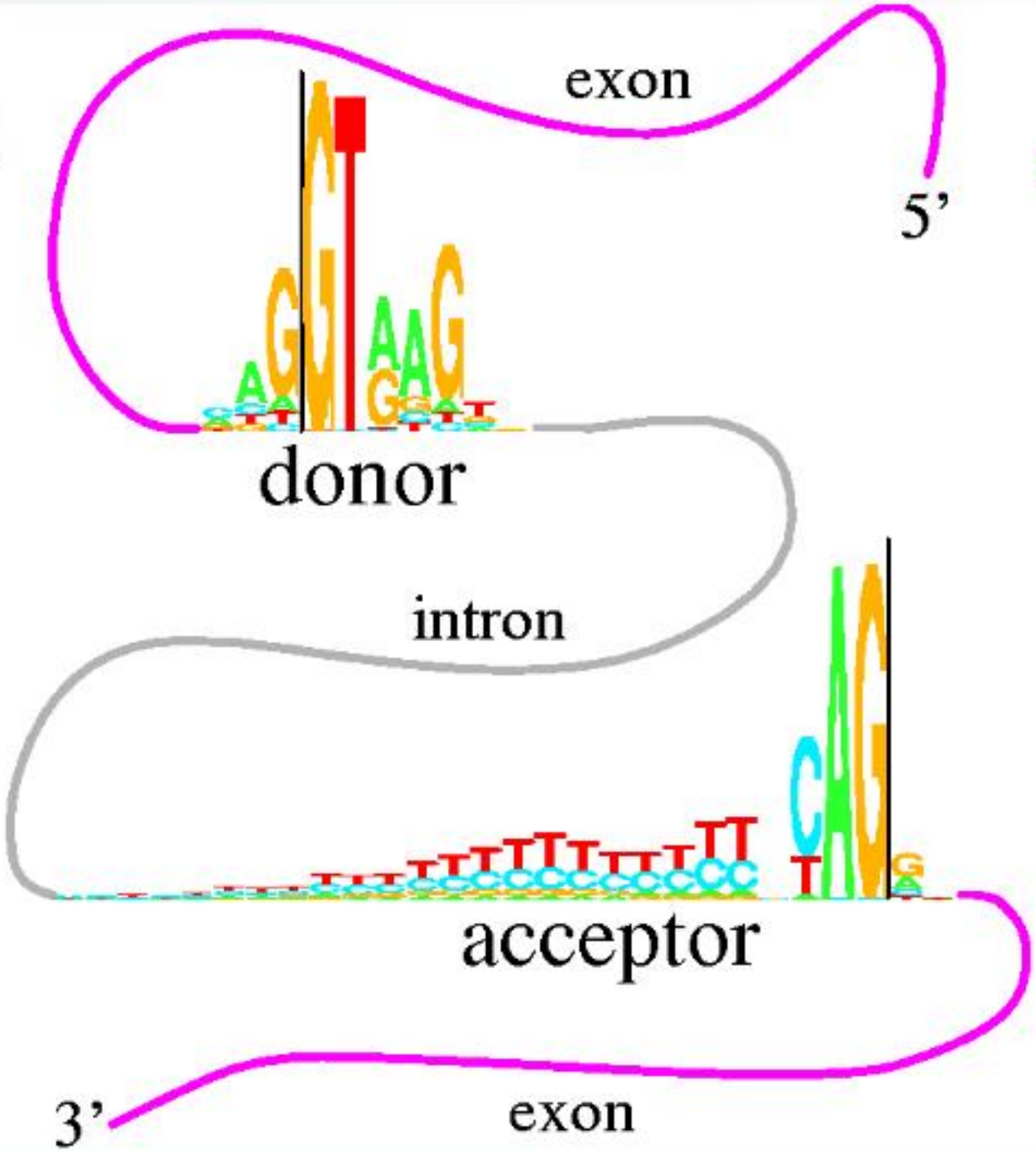
Типы эукариотических генов

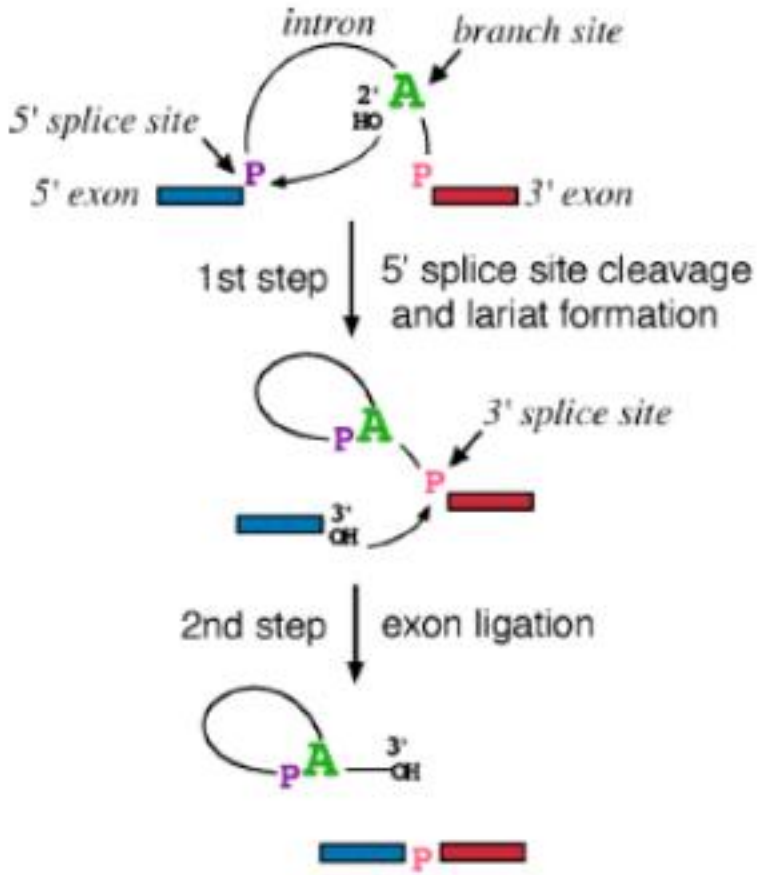
- Protein coding
  - most genes

- RNA genes
  - rRNA
  - tRNA
  - snRNA (small nuclear RNA)
  - snoRNA (small nucleolar RNA)

Основные типы информации, использующихся для распознавания генов

- Информационные свойства сигналов, входящих в структуру генов (донорные и акцепторные сайты, сайты связывания транскрипционных факторов и т.д.)
- Свойства контекста (смещения частот использования кодонов в кодирующих районах, длина открытой рамки считывания и т.д.)
- Сходство с известными гомологичными последовательностями представленными в базах данных.


exon
5’
donor
intron
acceptor
3’
exon



Chemical mechanism of intron excision by the spliceosome and group II introns.

- Consensus 'GT-AG'
  - Donor 5'
    - **(A,C)AG/**GT(A,G)AGT
  - Acceptor 3'
    - TTTTTNCAG**/GCCCCC**
  - Branch
    - CT(G,A)A(C,T)

Весовые матрицы, описывающие донорный и акцепторный сайты сплайсинга.

Весовая матрица донорного сайта сплайсинга

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

Весовая матрица акцепторного сайта сплайсинга

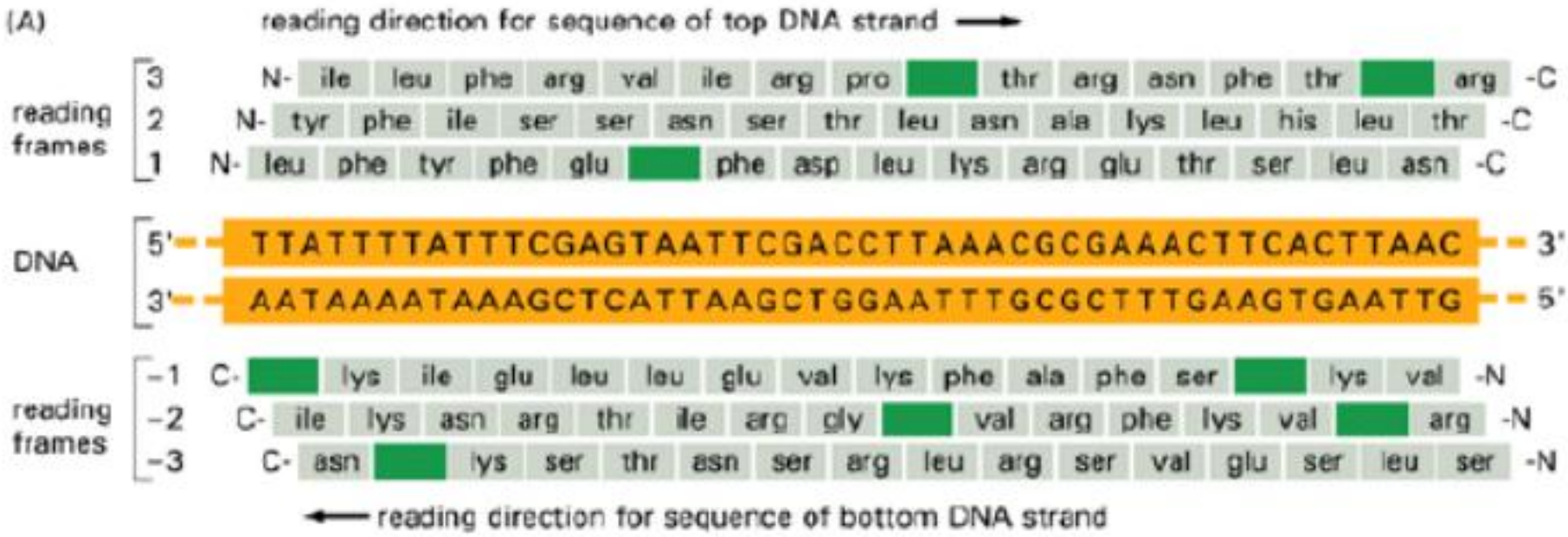
	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
A	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
C	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
G	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

- Стартовый кодон

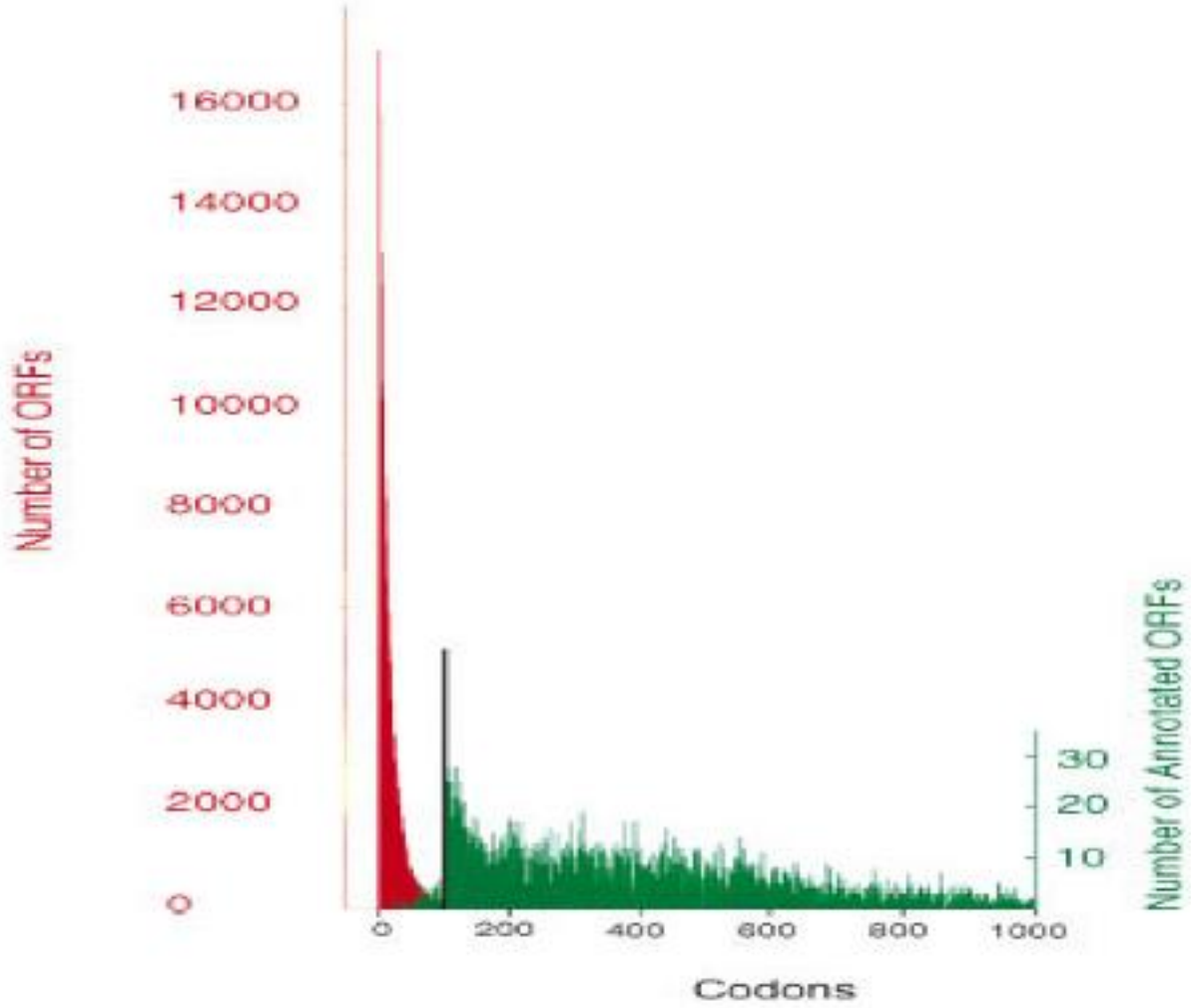
ATG

- Стоп - кодоны:

– TAA, TAG, TGA


(A)
reading direction for sequence of top DNA strand
reading frames
3
2
1
N- ile leu phe arg val ile arg pro thr arg asn phe thr arg -C
N- tyr phe ile ser ser asn ser thr leu asn ala lys leu his leu thr -C
N- leu phe tyr phe glu phe asp leu lys arg glu thr ser leu asn -C
DNA
5'
3'
TTATTTTATTTCGAGTAATTCGACCTTAAACGCGAAACTTCACTTAAC
AATAAAATAAAGCTCATTAAGCTGGAATTTGCGCTTTGAAGTGAATTG
3'
5'
reading frames
-1
-2
-3
C- lys ile glu leu leu glu val lys phe ala phe ser lys val -N
C- ile lys asn arg thr ile arg gly val arg phe lys val arg -N
C- asn lys ser thr asn ser arg leu arg ser val glu ser leu ser -N
reading direction for sequence of bottom DNA strand

Распределение длин ORF в геноме *S. cerevisiae*



Basrai MA, Hieter P, and Boeke J Genome Research 1997 7:768-771

Частоты использования кодонов кодирующих серин у различных организмов

Example: Codon frequencies (%) for serine (Ser or S) codons in different organisms

Codon	*E. coli*	Fruitfly	Man	Maise	Yeast
AGT	3	1	10	3	5
AGC	20	23	34	30	4
TCG	4	17	9	22	1
TCA	2	2	5	4	6
TCT	34	9	13	4	52
TCC	37	48	28	37	33

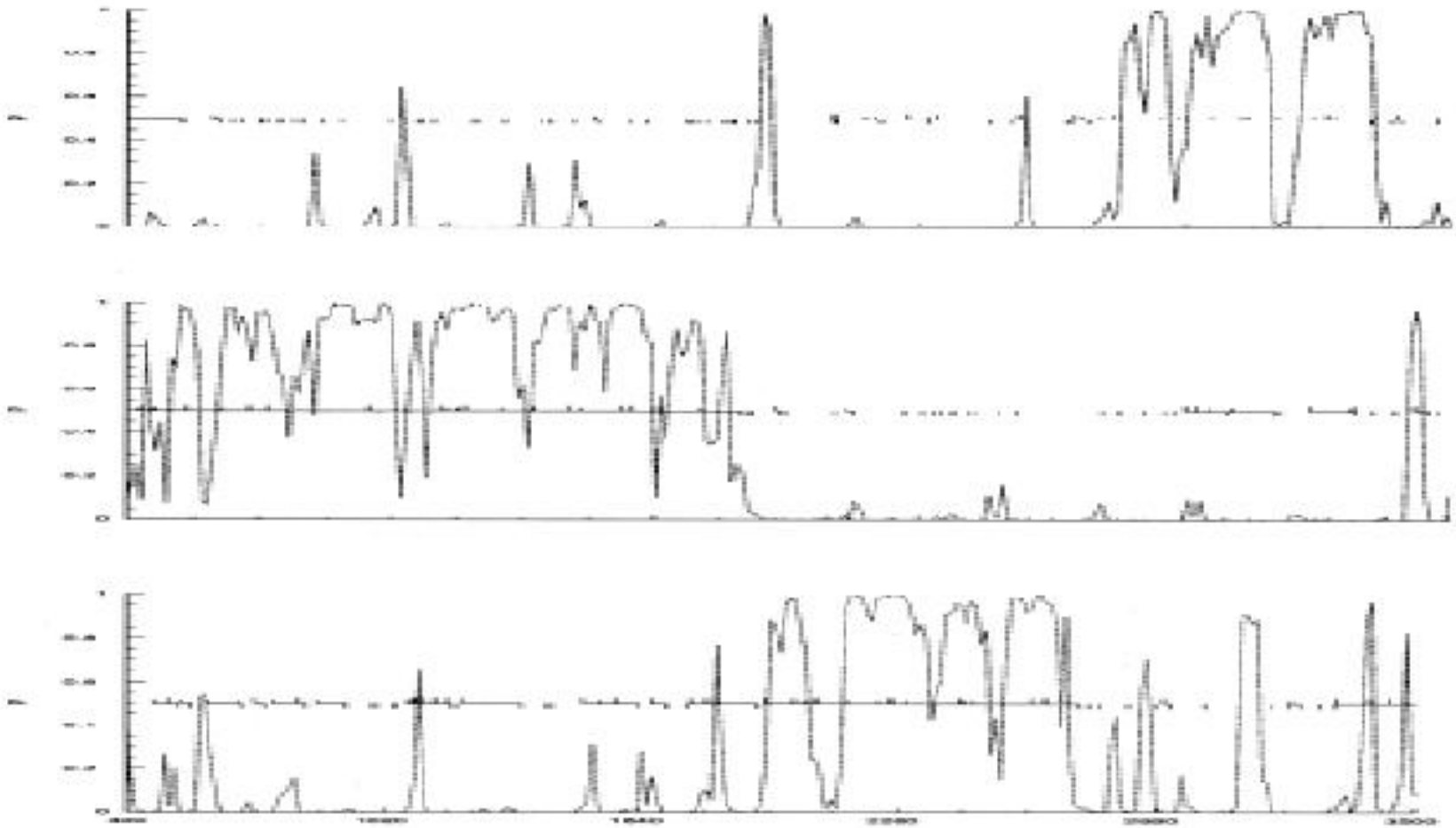
The most frequently used codons have the highest chance of being present in the CDS -> this is used in gene prediction

Индекс адаптации кодонов (CAI)

$$CAI = \prod_{i=codons} \left[\frac{f_{codon_i}}{f_{(codon_i)_{\max}}} \right]$$

Индексы адаптации кодонов, рассчитанные для человека и дрожжей

		hs	sc			hs	sc			hs	sc			hs	sc
UUU	Phe	16.6	26.0	UCU	Ser	14.5	23.6	UAU	Tyr	12.1	18.8	UGU	Cys	9.7	8.0
UUC	Phe	20.7	18.2	UCC	Ser	17.7	14.2	UAC	Tyr	16.3	14.7	UGC	Cys	12.4	4.7
UUA	Leu	7.0	26.3	UCA	Ser	11.4	18.8	UAA	stop	0.7	1.0	UGA	stop	1.3	0.6
UUG	Leu	12.0	27.1	UCG	Ser	4.5	8.6	UAG	stop	0.5	0.5	UGG	Trp	13.0	10.3
CUU	Leu	12.4	12.2	CCU	Pro	17.2	13.6	CAU	His	10.1	13.7	CGU	Arg	4.7	6.5
CUC	Leu	19.3	5.4	CCC	Pro	20.3	6.8	CAC	His	14.9	7.8	CGC	Arg	11.0	2.6
CUA	Leu	6.8	13.4	CCA	Pro	16.5	18.2	CAA	Gln	11.8	27.5	CGA	Arg	6.2	3.0
CUG	Leu	40.0	10.4	CCG	Pro	7.1	5.3	CAG	Gln	34.4	12.2	CGG	Arg	11.6	1.7
AUU	Ile	15.7	30.2	ACU	Thr	12.7	20.2	AAU	Asn	16.8	36.0	AGU	Ser	11.7	14.2
AUC	Ile	22.3	17.1	ACC	Thr	19.9	12.6	AAC	Asn	20.2	24.9	AGC	Ser	19.3	9.7
AUA	Ile	7.0	17.8	ACA	Thr	14.7	17.7	AAA	Lys	23.6	42.1	AGA	Arg	11.2	21.3
AUG	Met	22.2	20.9	ACG	Thr	6.4	8.0	AAG	Lys	33.2	30.8	AGG	Arg	11.1	9.3
GUU	Val	10.7	22.0	GCU	Ala	18.4	21.1	GAU	Asp	22.2	37.8	GGU	Gly	10.9	23.9
GUC	Val	14.8	11.6	GCC	Ala	28.6	12.6	GAC	Asp	26.5	20.4	GGC	Gly	23.1	9.7
GUA	Val	6.8	11.7	GCA	Ala	15.6	16.2	GAA	Glu	28.6	45.9	GGA	Gly	16.4	10.9
GUG	Val	29.3	10.7	GCG	Ala	7.7	6.1	GAG	Glu	40.6	19.1	GGG	Gly	16.5	6.0



Марковские цепи, как модель генерации последовательности.

$$s = s_1 s_2 s_3 s_4 \cdots$$

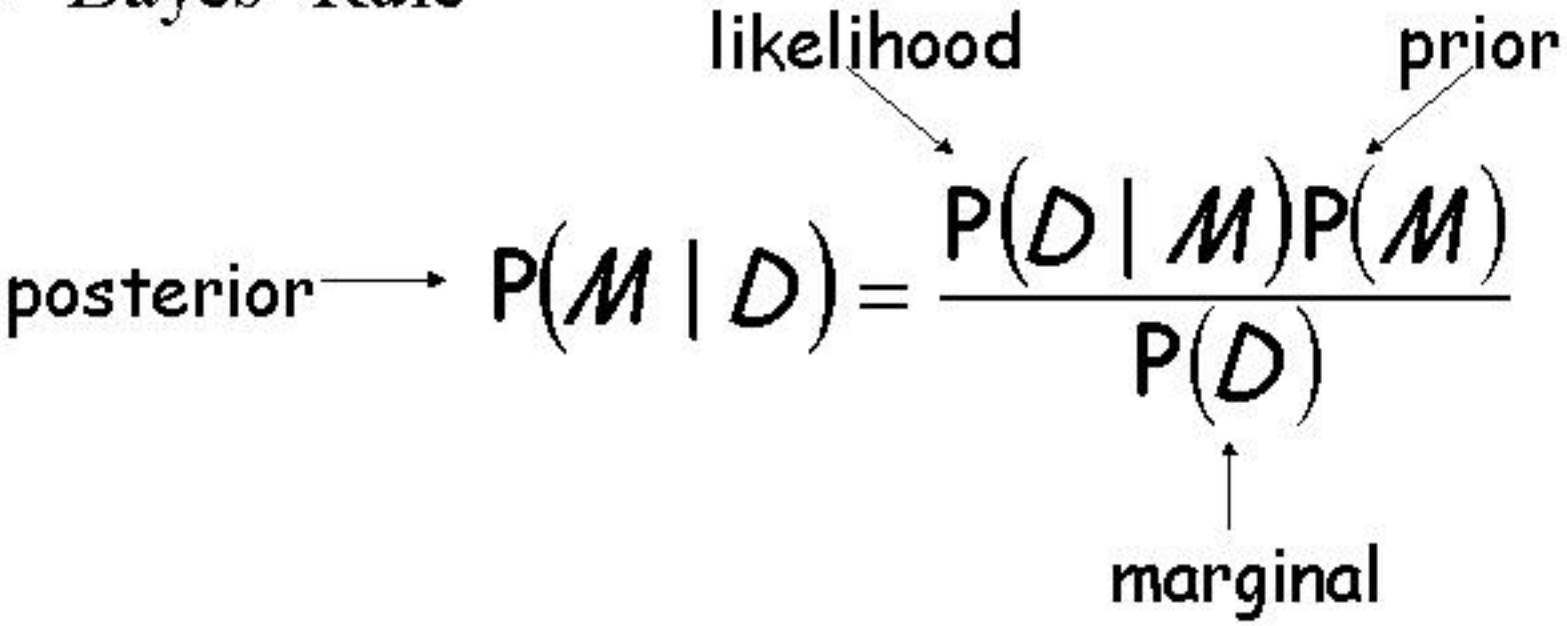
$$P_0(s) = \mathrm{p}(s_1) \cdot \mathrm{p}(s_2) \cdot \mathrm{p}(s_3) \cdots = \prod_{i=1}^{N} \mathrm{p}(s_i)$$

$$P_1(s) = \mathrm{p}(s_1) \cdot \mathrm{p}(s_2 \mid s_1) \cdot \mathrm{p}(s_3 \mid s_2) \cdots = \mathrm{p}(s_1) \cdot \prod_{i=2}^{N} \mathrm{p}(s_i \mid s_{i-1})$$

$$P_2(s) = \mathrm{p}(s_1 s_2) \cdot \mathrm{p}(s_3 \mid s_1 s_2) \cdot \mathrm{p}(s_4 \mid s_2 s_3) \cdots = \mathrm{p}(s_1 s_2) \cdot \prod_{i=3}^{N} \mathrm{p}(s_i \mid s_{i-2} s_{i-1})$$

Формула Байеса

- Bayes' Rule



- M: the model, D: data or evidence

$$P(D) = \sum P(D \mid M)P(M) \cdot \text{discrete}$$

$$= \int P(D \mid M)P(M)dM \cdot \text{continuous}$$

LOG-ODDS Ratio

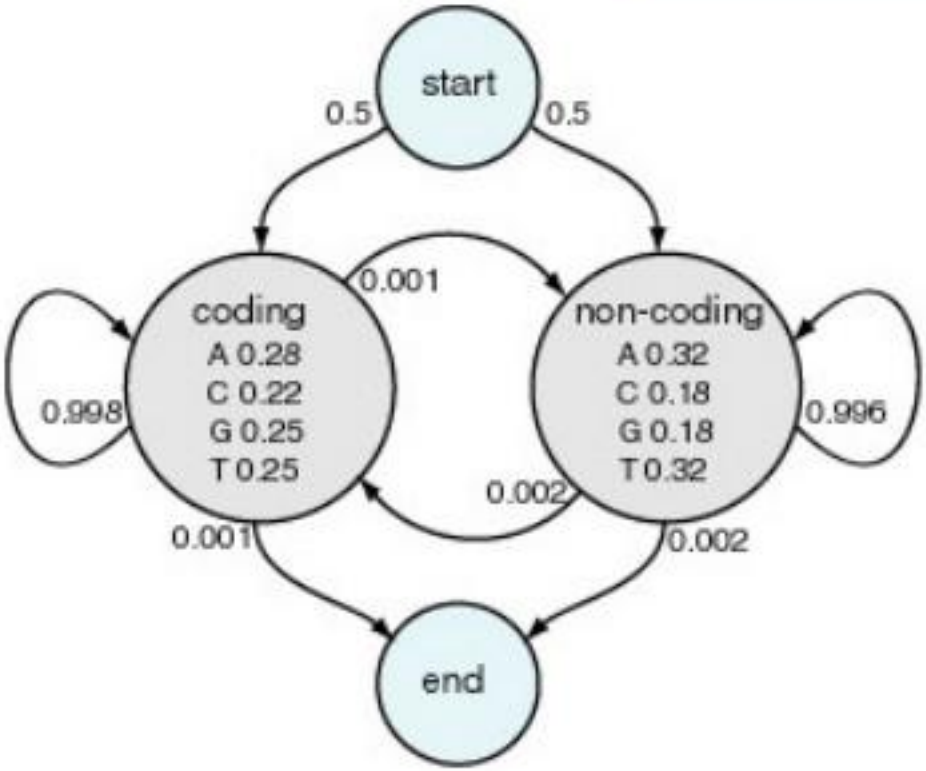
$$\log \frac{P(\mathrm{c}|data)}{P(\mathrm{nc}|data)} = \log \frac{P(data|\mathrm{c})}{P(data|\mathrm{nc})} + \log \frac{P(\mathrm{c})}{P(\mathrm{nc})}$$

- GeneMark uses a 5th-order Markov chain:

$$x = \text{GCTACGCGCTAGGAT...}$$

$$\Pr(x) = \Pr(GCTAC) \times \Pr(G \mid GCTAC) \times \Pr(C \mid CTACG) \ \ldots$$

Модель скрытой Марковской цепи (HMM)



$$\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0.998 & 0.002 & 0 \\ 0.5 & 0.001 & 0.996 & 0 \\ 0 & 0.001 & 0.002 & 0 \end{bmatrix}$$

$$H = \begin{bmatrix} 0.28 & 0.32 \\ 0.22 & 0.18 \\ 0.25 & 0.18 \\ 0.25 & 0.32 \end{bmatrix}$$

$H(m,y_i)$ = probability of emitting character y_i in state m;

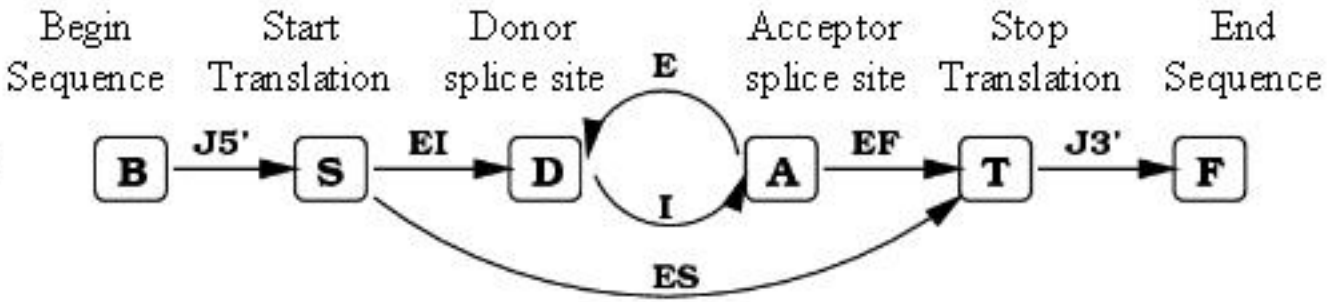
Φ_{mk} = probability of transition from state k to m.

Некоторые программы распознавания генов использующие скрытые Марковские цепи

- GENSCAN (Burge 1997)
- FGENESH (Solovyev 1997)
- HMMgene (Krogh 1997)
- GENIE (Kulp 1996)
- GENMARK (Borodovsky & McIninch 1993)
- VEIL (Henderson, Salzberg, & Fasman 1997)

Genie

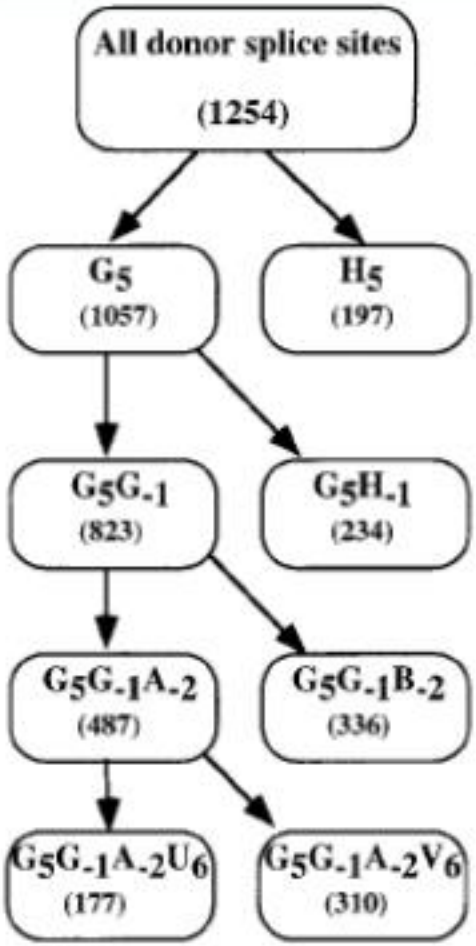
- J5' – 5' UTR
- EI – Initial Exon
- E – Exon, Internal Exon
- I – Intron
- EF – Final Exon
- ES – Single Exon
- J3' – 3'UTR



GenScan

- Учитывает распределение длин интронов и экзонов
- Весовая позиционная матрица (WMM) для описания TATA box, PolyA signal, CAP end and Transcription Initiation End (TIE) of 5'UTR.
- Weight Array model (WAM) для описания акцепторного сайта связывания.
- Дерево решений (maximal dependence decomposition) для моделирования донорных сайтов сплайсинга

Maximal Dependence Decomposition (MDD)



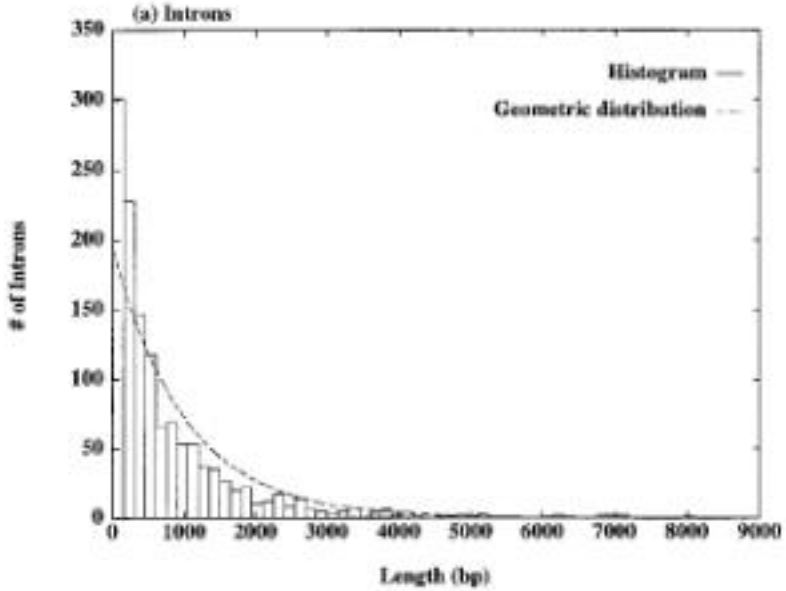
Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19

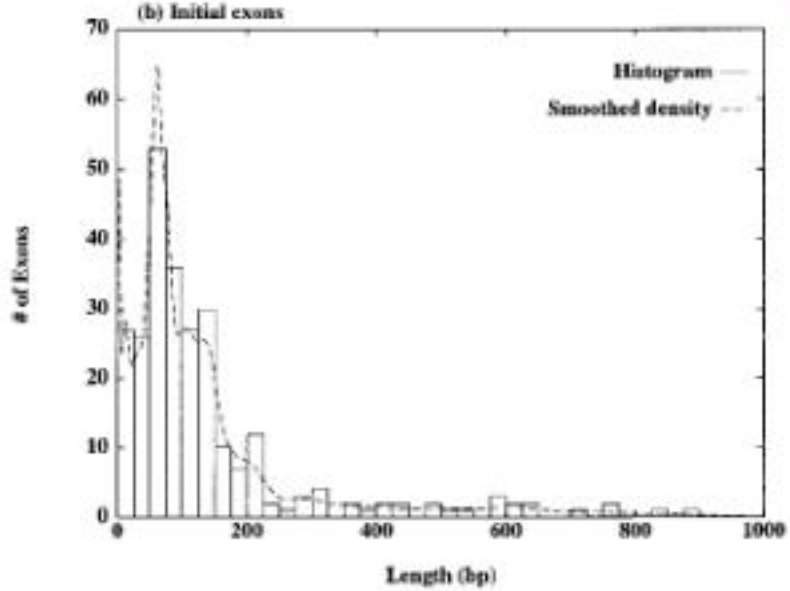
Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

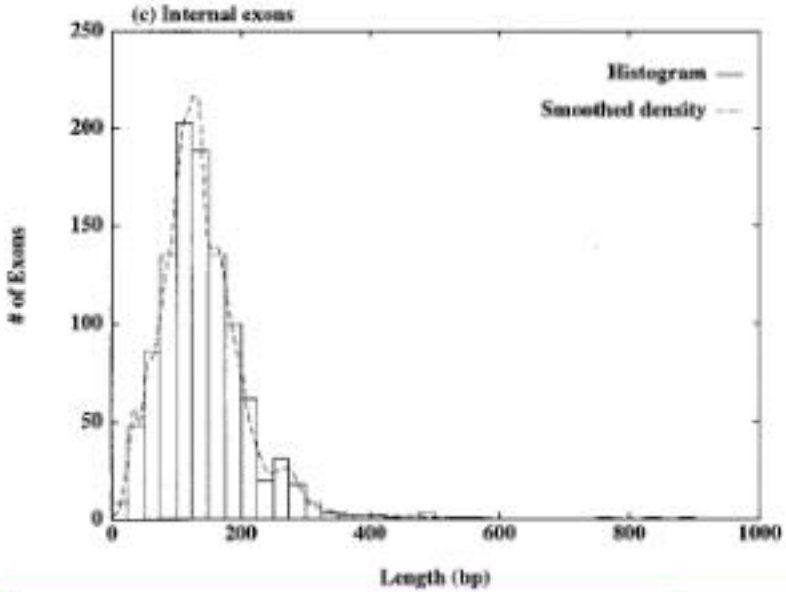
All sites: ---------- **Position** ----------

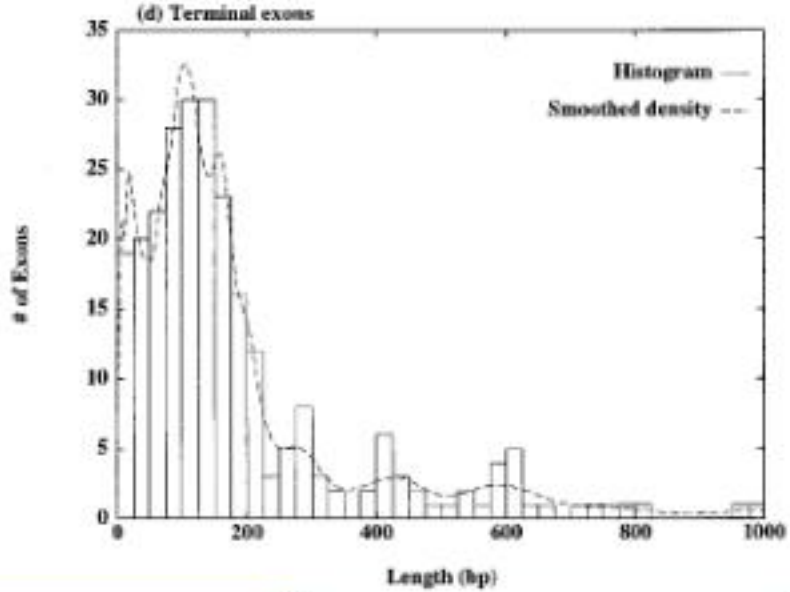
Base	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	**33**	**60**	8	0	0	**49**	**71**	6	15
C%	**37**	13	4	0	0	3	7	5	19
G%	18	14	**81**	**100**	0	**45**	12	**84**	20
U%	12	13	7	0	**100**	3	9	5	**46**
U1 snRNA:	**3'**	**G**	**U**	**C**	**C**	**A**	**U**	**U**	**C**	**A**	**5'**

Распределение длин экзонов и интронов, входящих в структуру человеческих генов



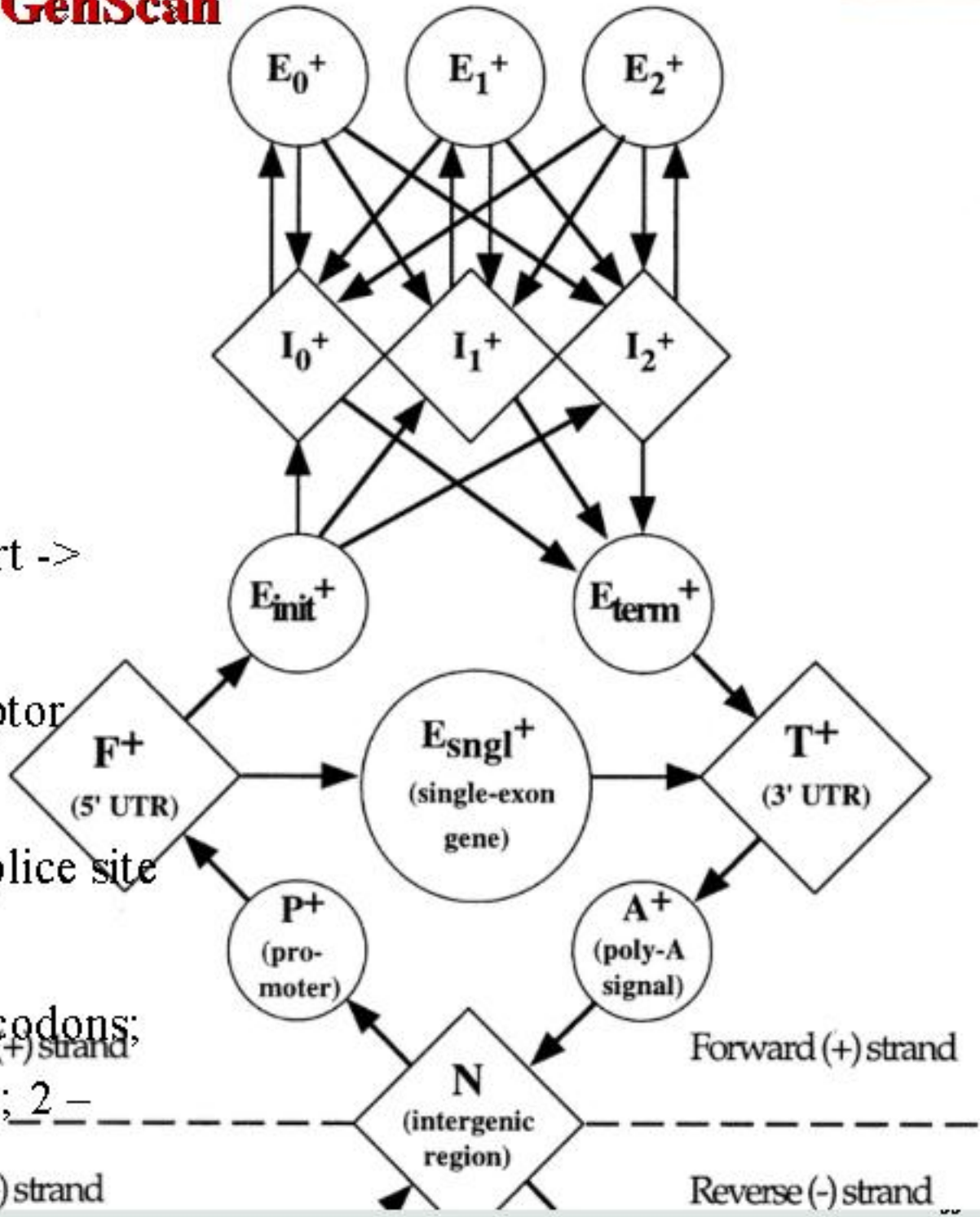


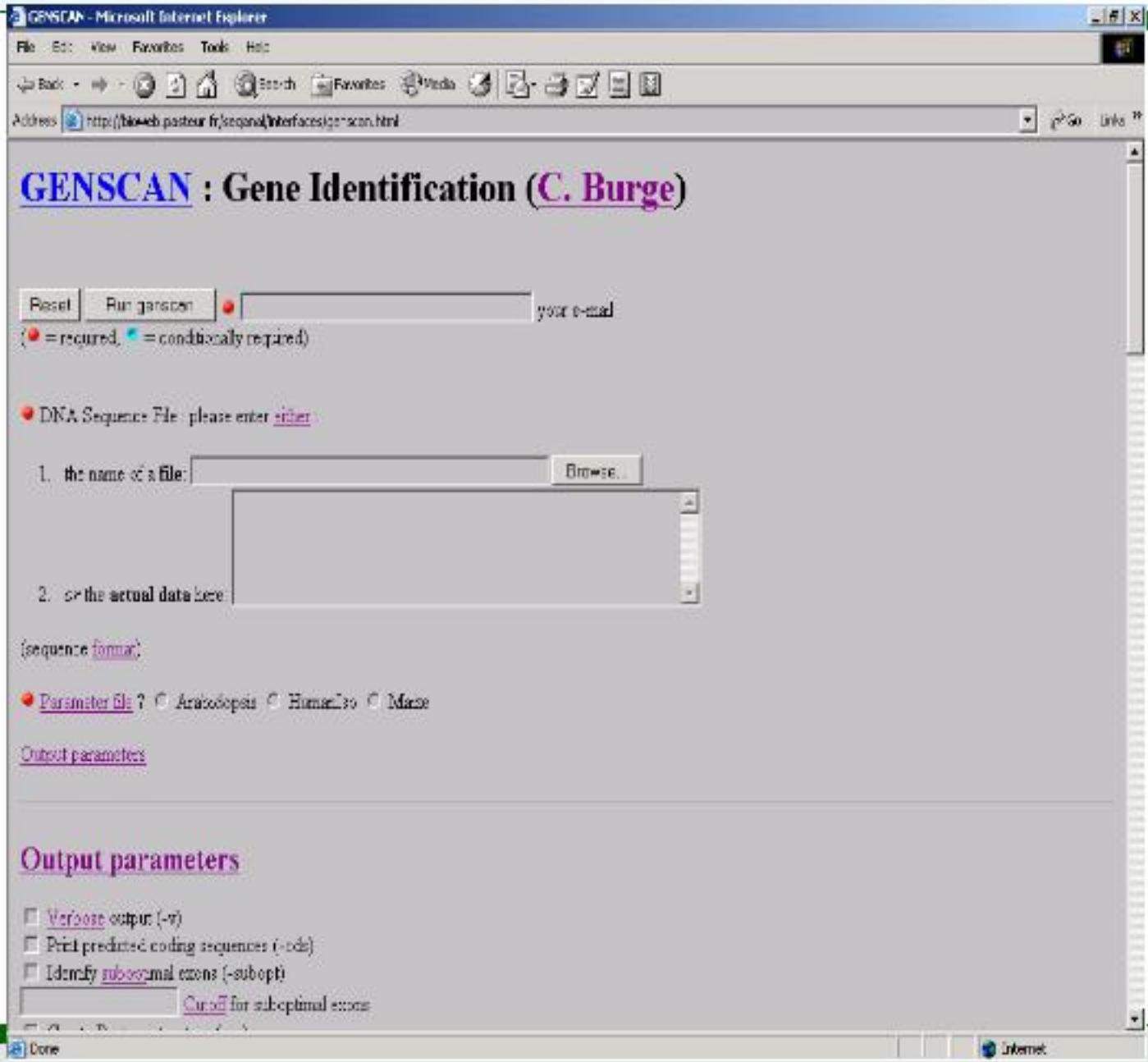




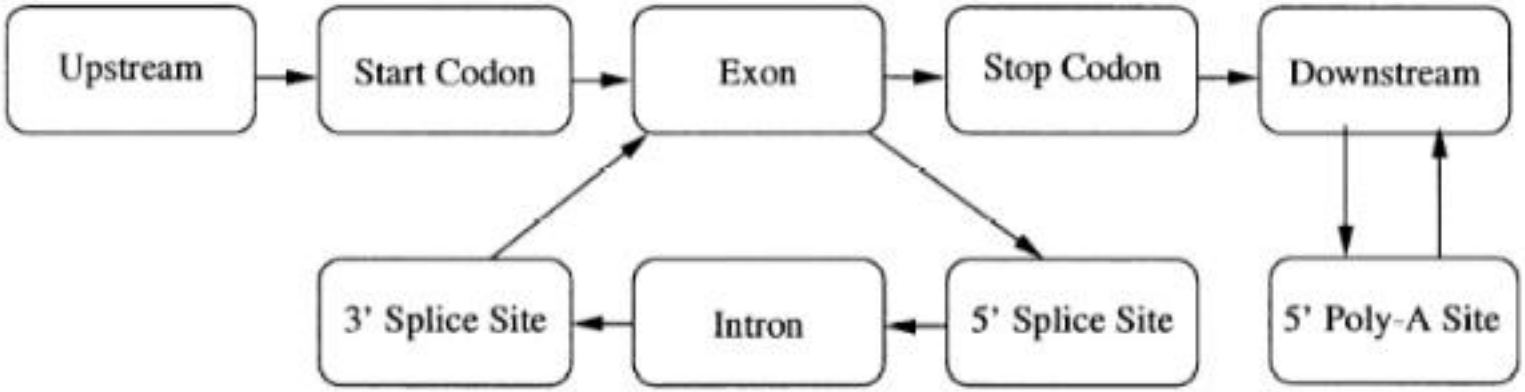
GenScan

- N - intergenic region
- P - promoter
- F - 5' untranslated region
- E_{sngl} – single exon (intronless) (translation start -> stop codon)
- E_{init} – initial exon (translation start -> donor splice site)
- E_k – phase k internal exon (acceptor splice site -> donor splice site)
- E_{term} – terminal exon (acceptor splice site -> stop codon)
- I_k – phase k intron: 0 – between codons; 1 – after the first base of a codon; 2 – after the second base of a codon




GENSCAN - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://(biowb.pasteur.fr/seqanal/interfaces/genscan.html
GENSCAN : Gene Identification (C. Burge)
Reset
Run genscan
your e-mail
(= required, = conditionally required)
DNA Sequence File : please enter either
1. the name of a file:
Browse...
2. or the actual data here
(sequence format)
Parameter file ? Arabidopsis HumanIso Maize
Output parameters
Output parameters
Verbose output (-v)
Print predicted coding sequences (-cds)
Identify suboptimal exons (-subopt)
Cutoff for suboptimal exons
Done
Internet

VEIL (Viterbi Exon-Intron Locator)

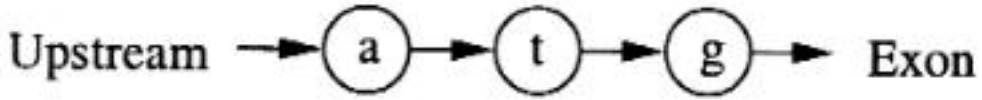


- Each box represents a separate HMM
- Edges represent multiple edges between HMMs
- Total model has 241 states and 1003 edges

Upstream/Downstream and Start Codons

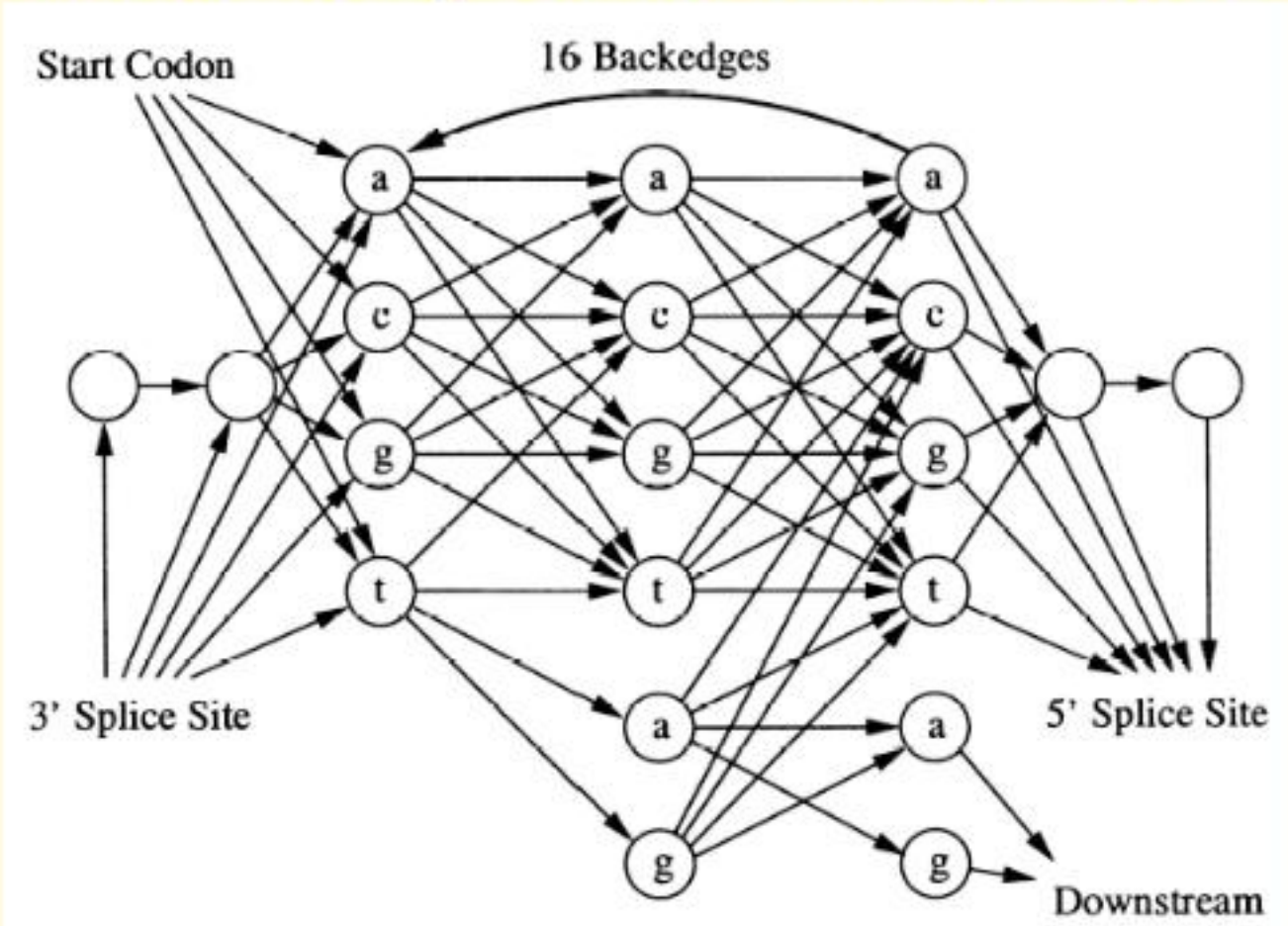
- **Upstream model:**
  - **15-stage chain**
  - **Loops at end to absorb extra bases**

- **Start Codon Model:**

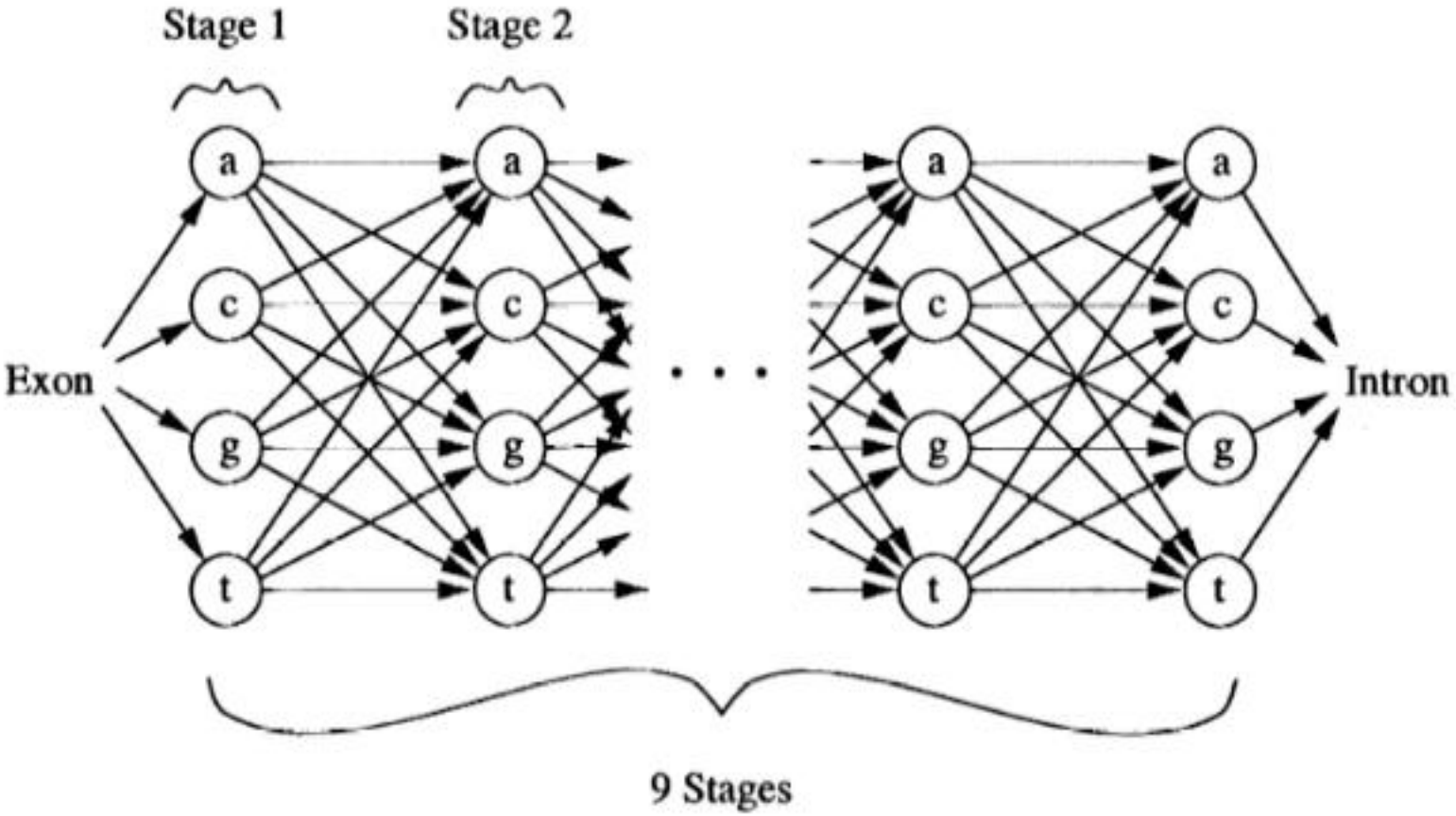


- **Downstream model:**
  - **10-stage chain**
  - **Loops to absorb extra bases**

Exon and Stop Codon Models in VEIL

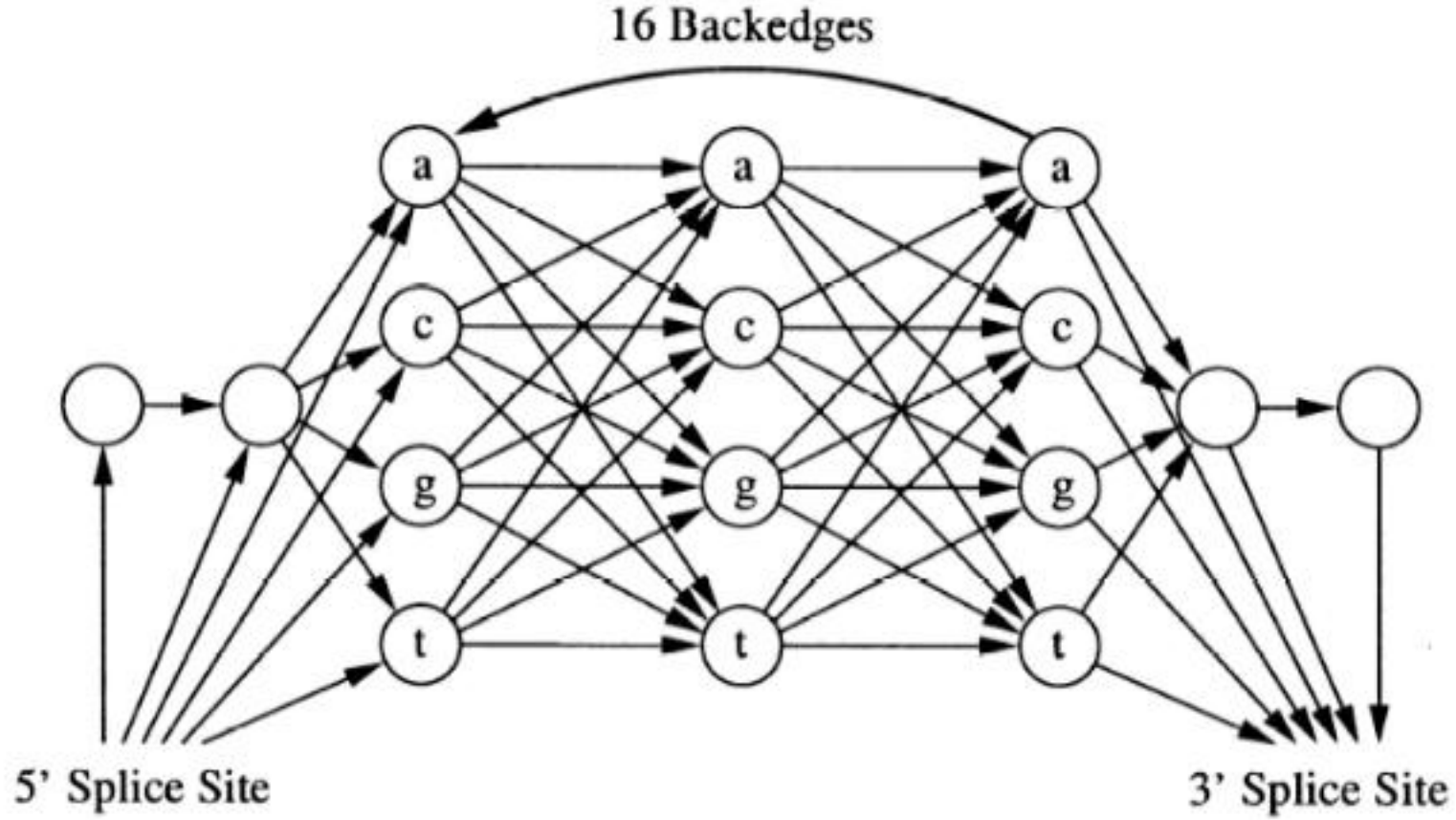


Donor Site (5' splice site) Model

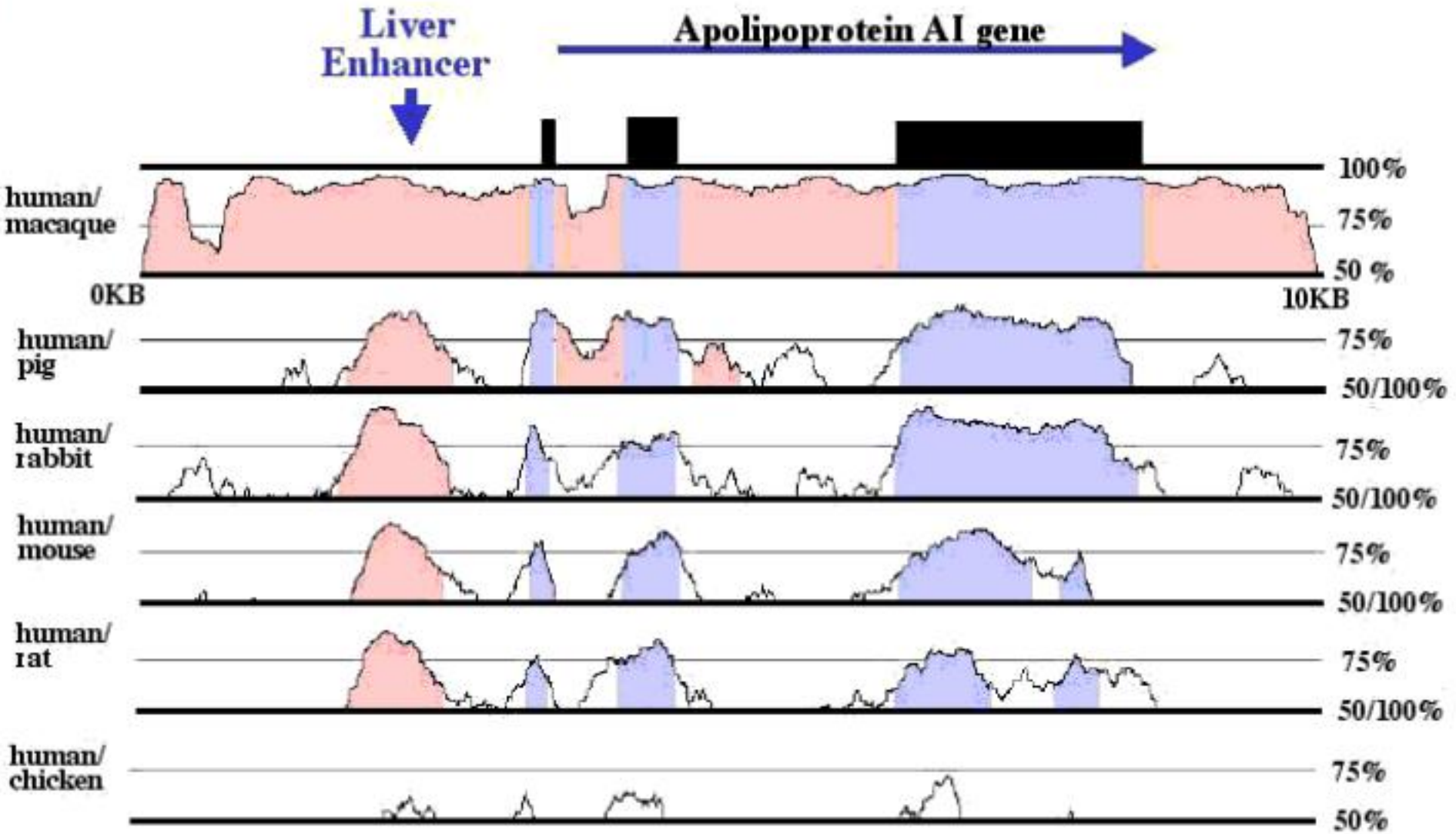


- Acceptor site (3' splice site) has 15 stages
- Length based on consensus sequences (Mount et. al, 1995)

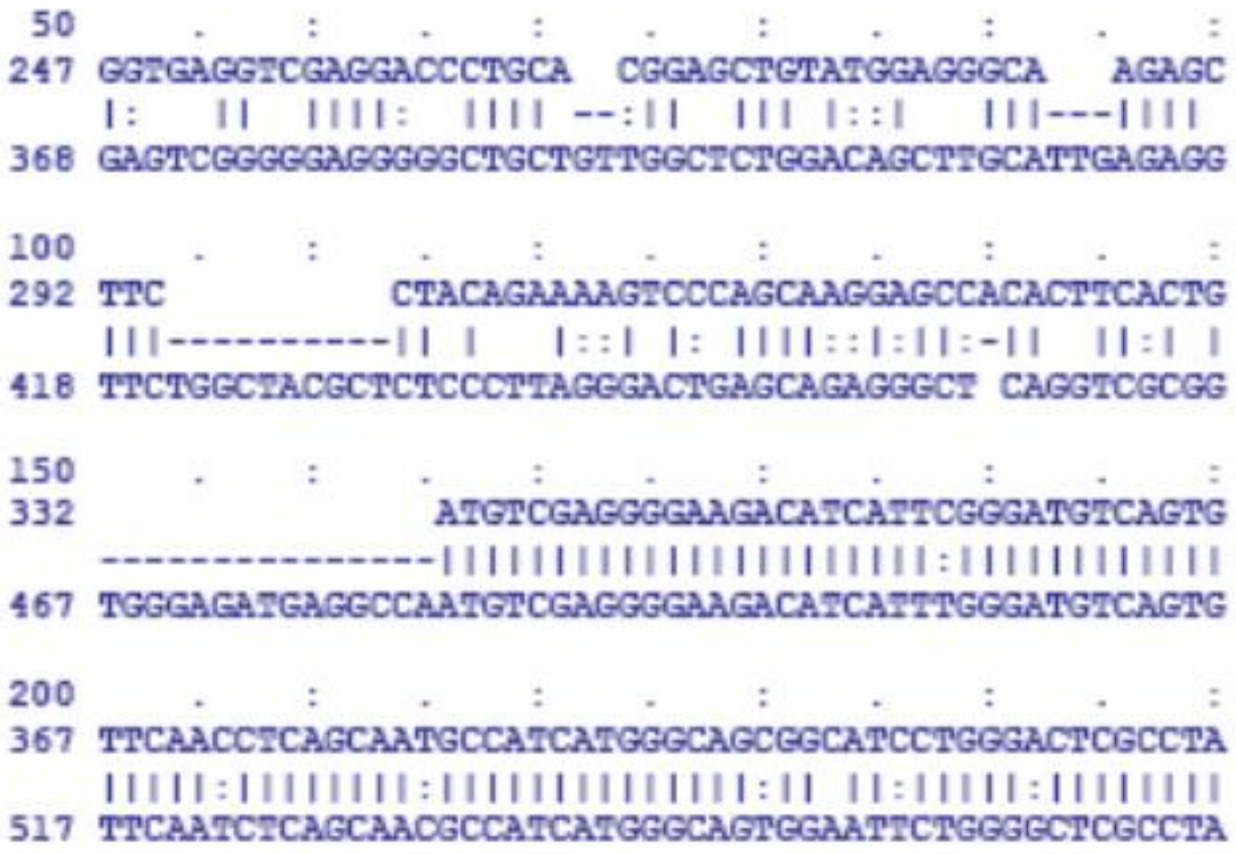
Intron Model in VEIL



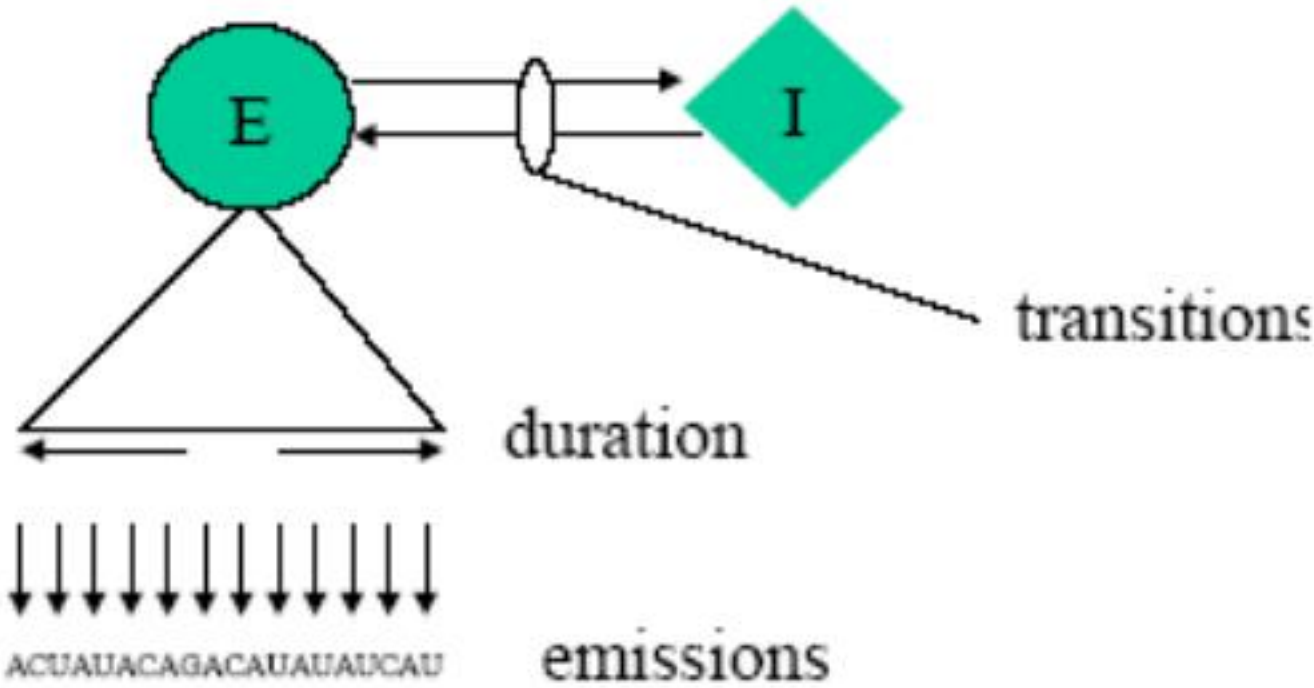
Multi-Species Comparative Analysis



Выравнивание ортологичных генов



Twinscan


E
I
transitions
duration
ACUAUACAGACAUAUAUCAU
emissions

1. Align the two sequences (eg. from human and mouse)

2. Mark each human base as gap (-), mismatch (:), match (|)

New “alphabet”: 4 x 3 = 12 letters

Σ = { A-, A:, A|, C-, C:, C|, G-, G:, G|, U-, U:, U| }

3. Run Viterbi using emissions $e_k(b)$
   where $b \in \{A-, A:, A|, \ldots, T| \}$

Note:

Emission distributions $e_k(b)$ estimated from real genes from human/mouse

$e_I(x|) < e_E(x|)$: matches favored in exons

$e_I(x-) > e_E(x-)$: gaps (and mismatches) favored in introns

Human: ACGGCGACUGUGCACGU

Mouse: ACUGUGAC GUGCACUU

Align: ||:|:|||-|||||||:|

Input to Twinscan HMM:

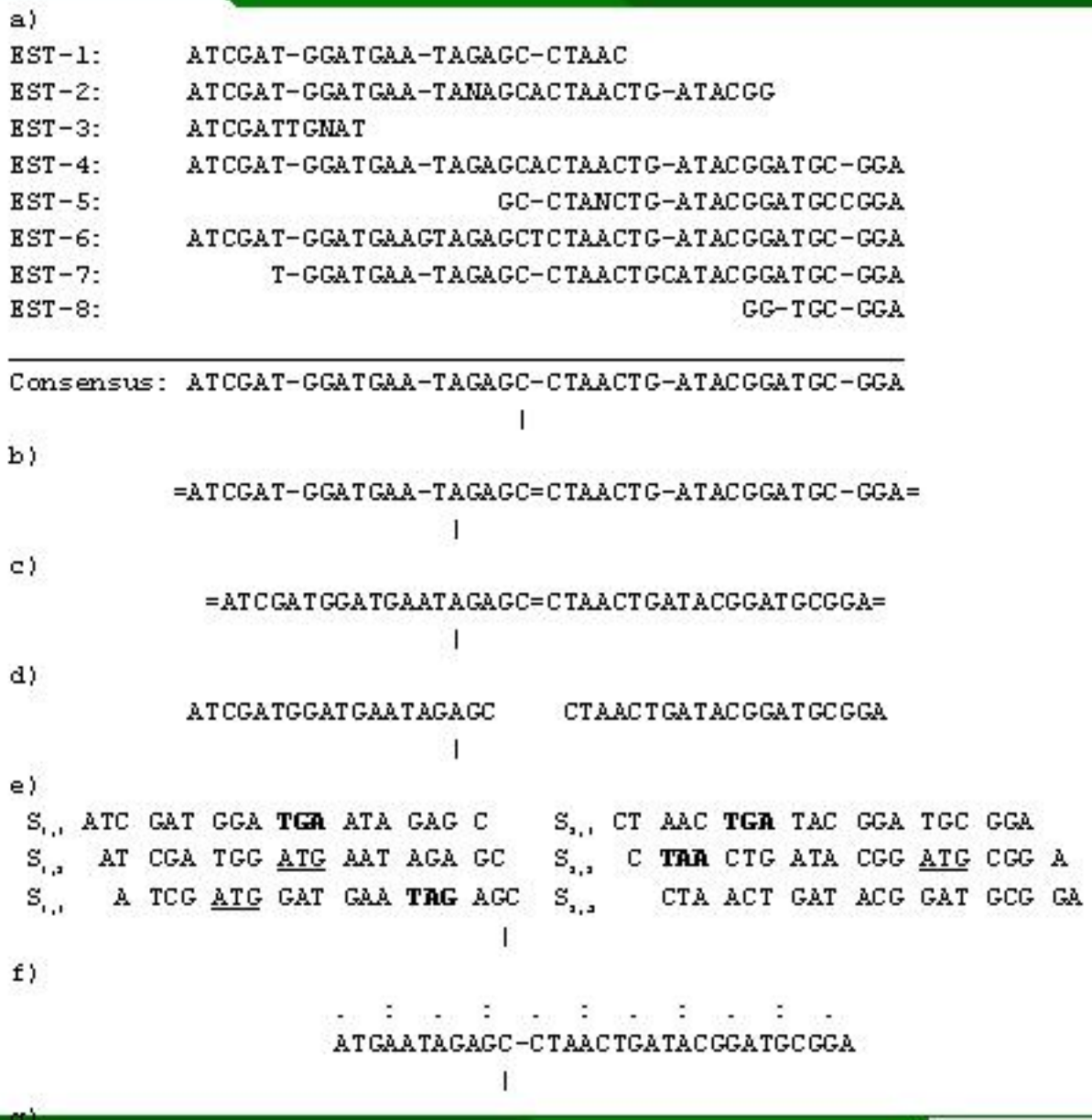
A| C| G: G| C: G| A| C| U- G| U| G| C| A| C| G: U|

Recall, $e_E(A|) > e_I(A|)$

$e_E(A-) < e_I(A-)$

Likely exon

ORFScan

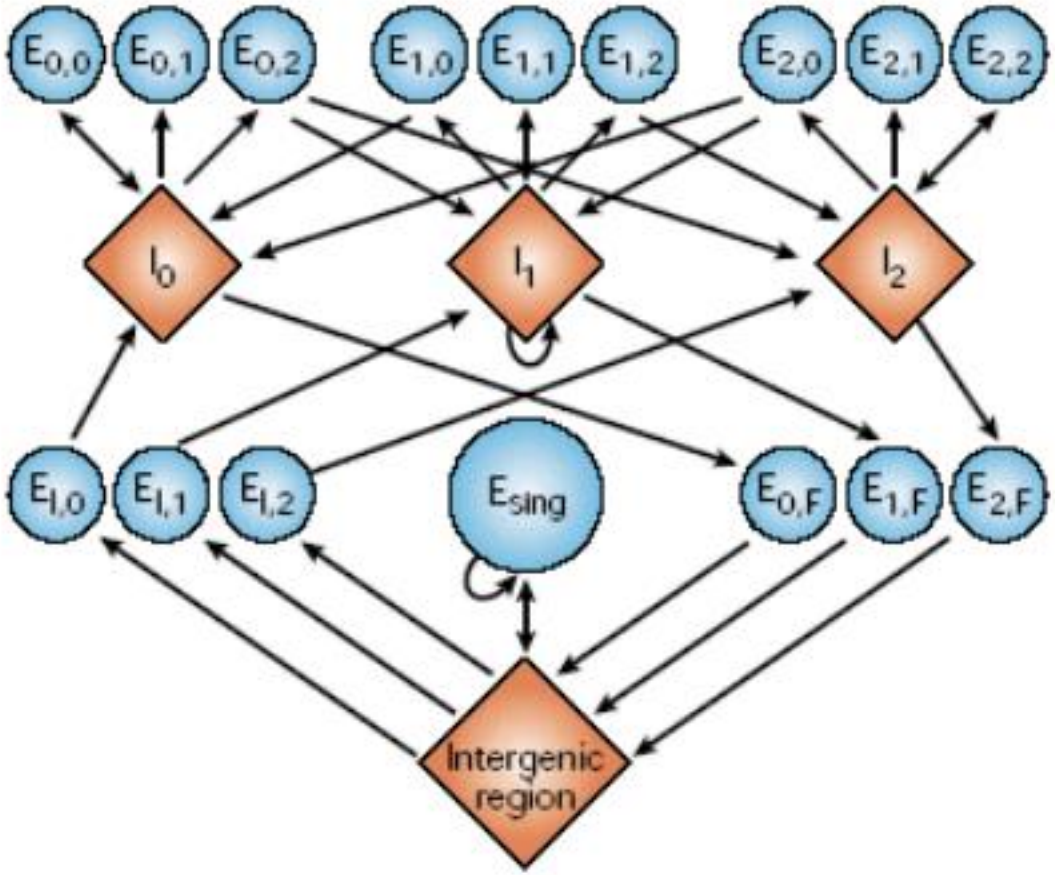

a)
EST-1: ATCGAT-GGATGAA-TAGAGC-CTAAC
EST-2: ATCGAT-GGATGAA-TANAGCACTAACTG-ATACGG
EST-3: ATCGATTGNAT
EST-4: ATCGAT-GGATGAA-TAGAGCACTAACTG-ATACGGATGC-GGA
EST-5: GC-CTANCTG-ATACGGATGCCGGA
EST-6: ATCGAT-GGATGAAGTAGAGCTCTAACTG-ATACGGATGC-GGA
EST-7: T-GGATGAA-TAGAGC-CTAACTGCATACGGATGC-GGA
EST-8: GG-TGC-GGA
Consensus: ATCGAT-GGATGAA-TAGAGC-CTAACTG-ATACGGATGC-GGA
b)
=ATCGAT-GGATGAA-TAGAGC=CTAACTG-ATACGGATGC-GGA=
c)
=ATCGATGGATGAATAGAGC=CTAACTGATACGGATGCGGA=
d)
ATCGATGGATGAATAGAGC CTAACTGATACGGATGCGGA
e)
S1,1 ATC GAT GGA TGA ATA GAG C S2,1 CT AAC TGA TAC GGA TGC GGA
S1,2 AT CGA TGG ATG AAT AGA GC S2,2 C TAA CTG ATA CGG ATG CGG A
S1,3 A TCG ATG GAT GAA TAG AGC S2,3 CTA ACT GAT ACG GAT GCG GA
f)
ATGAATAGAGC-CTAACTGATACGGATGCGGA
g)
M N R ? L T D T D A

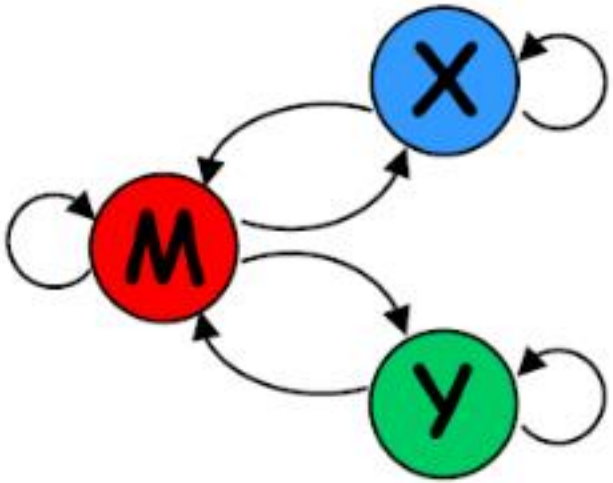
Предсказание кодирующих частей генов с помощью программ ORFScan, ESTScan, GENSCAN, GRAIL, GENFIND на примере DT_101886.

Ген human ferrochelatase. Ошибочно определенные позиции маркированы цветом

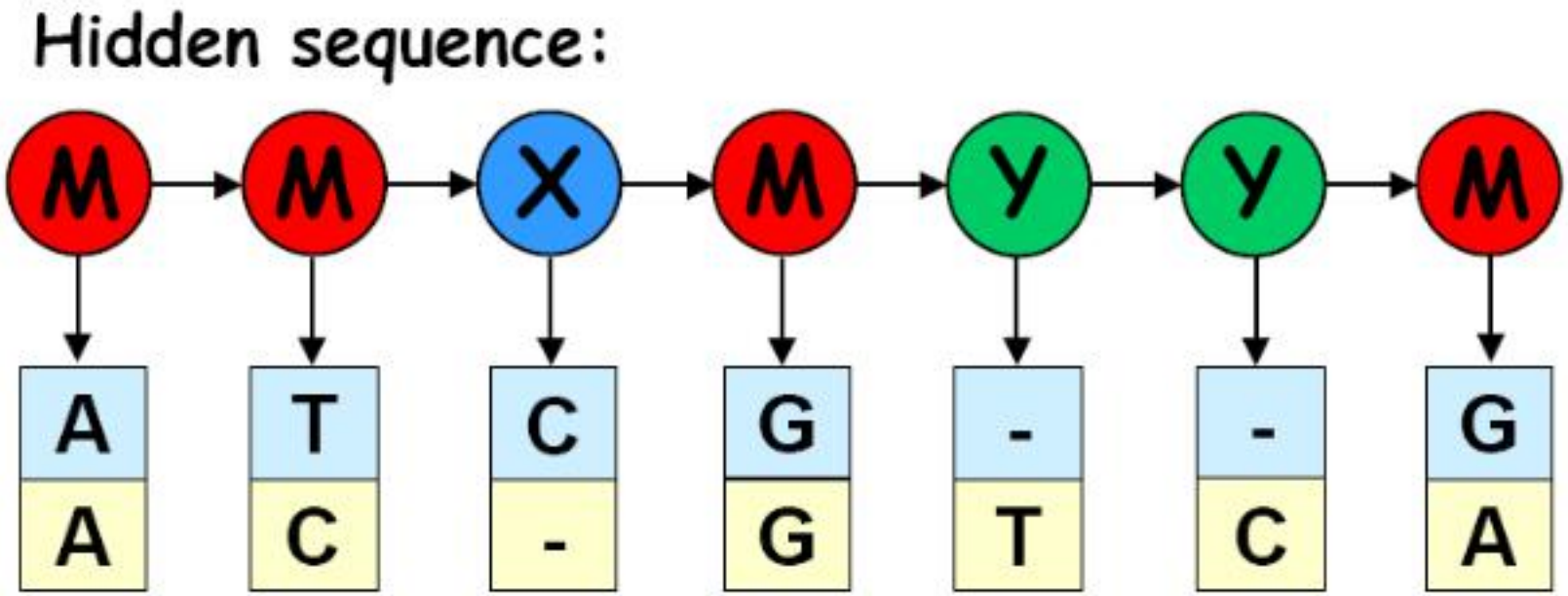
```
true     : mrslgarmaaalraagvllrdplassswrvcqpwrwksgaaaaavttetaqhaqgakpqvqpqkrkpktgilmlrmggpetlgdvhdfllrlfldqdlmtlpiqrklapf
ORFScan  : MRSLGAMMAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
ESTScan  : NAFIGAMMAAALRAAGVLLRXSAGIQQLEGLSAMEVEVGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GENSCAN  : _______MAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GRAIL    : MRSLGAMMAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GenFind  : _______MAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF

true     : iakrrtpkiqeqyrrigggspikiwtskqgegmvklldelspntaphkyyigfryvhplteeaieemerdgleraiaftqypqyscsttgsslnaiyryynqvgrkptmk
ORFScan  : IAKRRTPKIQEQYRRLEADPPSRY??SKQGEGMVKLLDELSPNTAPHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
ESTScan  : IAKRRTPKIQEQYRRXWRRIPHQIMDFQQGEGMVKLLDELSPNTAPHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
GENSCAN  : IAKRRTPKIQEQYRRXWRRIPHQIMDFQAGRGH__________APHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
GRAIL    : IAKRRTPKIQEQYRRTWRRIPHQIMDFQAGRGH_____________________________________________________________________
GenFind  : IAKRRTPKIQEQYRRTWRRIPHQIMDFQAGRGH_____________________EEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK

true     : wstidrwpthhlliqcfadhilkeldhfplekrsevvilfsahslpmsvvnrgdpypqevsatvqkvmerleycnpyrlvwqskvgpmpwlgpqtdesikglcergrkni
ORFScan  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
ESTScan  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
GENSCAN  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
GRAIL    : _____________________________________________GEAAGQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGP_____________________
GenFind  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI

true     : llvpiaftsdhietlyeldieysqvlakecgvenirraeslngnplfskaladlvhshiqsnelcskqltlscplcvnpvcretksfftsqql
ORFScan  : LLVPIAFTSDHIETLYELDIEYSQVLAKECGVENIRRAESLNGIHCS?KALADLV?FTHPVKRAVFQAADPDCPLCVNPVCRETKSFFT?PAAVTPAGGPRGVSKCPTSRYLRCGEGVI
ESTScan  : LLVPIAFTSDHIETLYELDIEYSQVLAKDVGVENIXKELSLFNGKFHCSLKAPGRLGAFTHPSQTSCVPKQLTPELVNPVCRETKSFXHPASSCEPPPVDPVALANAQPPDTSDVER
GENSCAN  : LLVPIAFTSDHIETLYELDIEYSQVLAKECGS_____________________________________________________________________
GRAIL    :                                                                                  CPLCVNPVCRETKSFFTQPAAVNPRRVTPVR
```

Pair HMM


E0,0
E0,1
E0,2
E1,0
E1,1
E1,2
E2,0
E2,1
E2,2
I0
I1
I2
E1,0
E1,1
E1,2
Esing
E0,F
E1,F
E2,F
Intergenic region



M = (mis)match
X = insert seq1
Y = insert seq2



Hidden alignment:

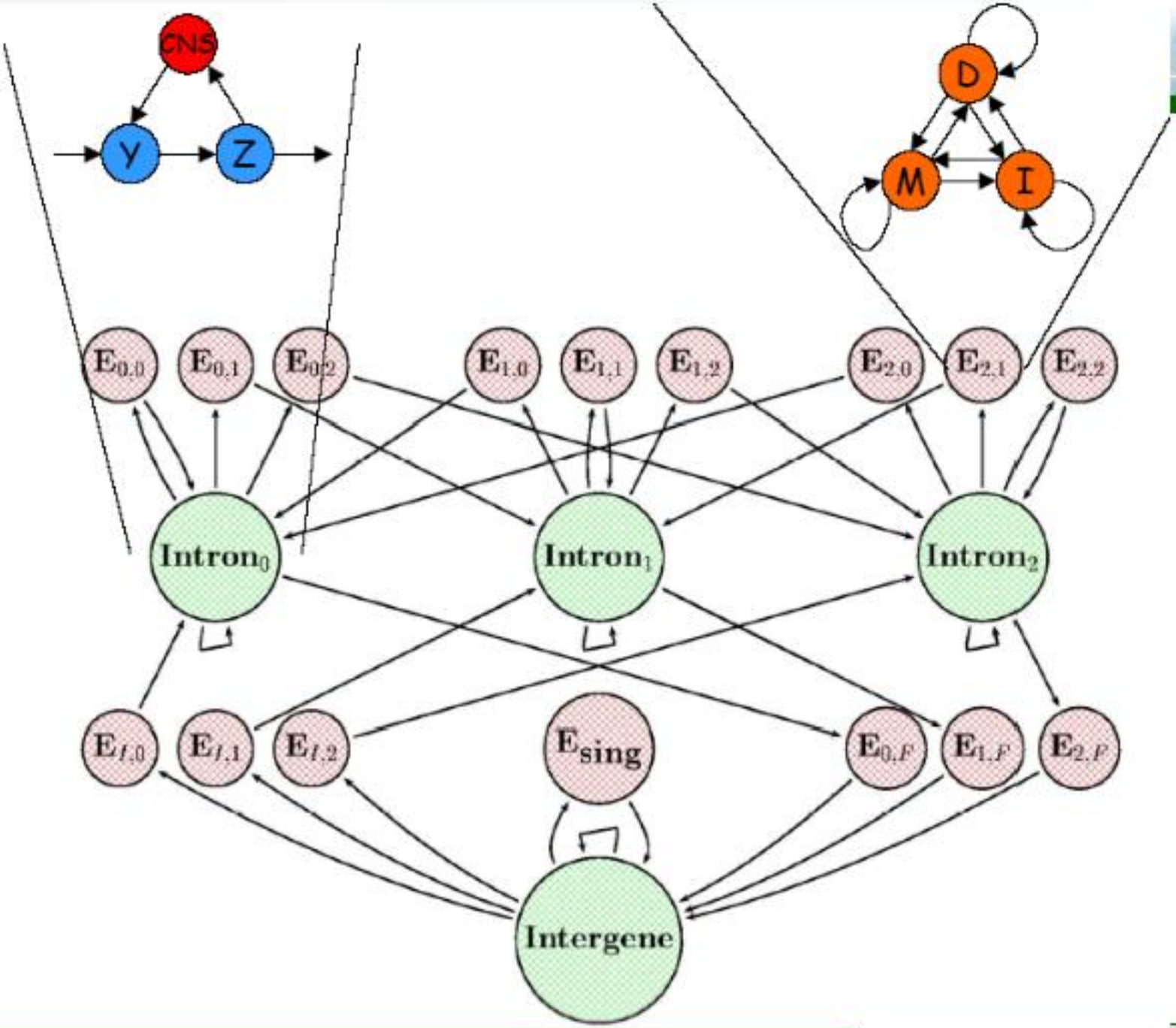
ATCG--G
AC-GTCA

Observed sequence:

ATCGG
ACGTCA

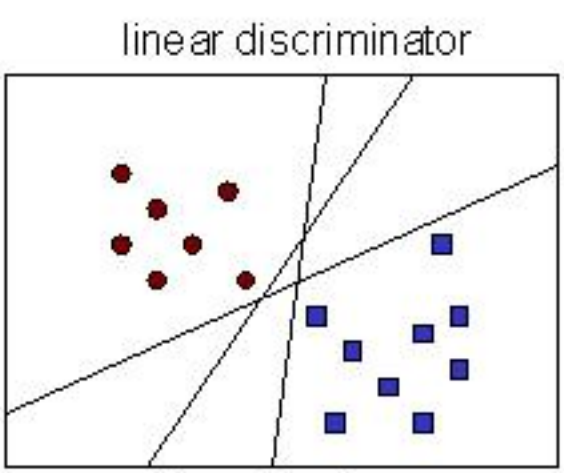
SLAM

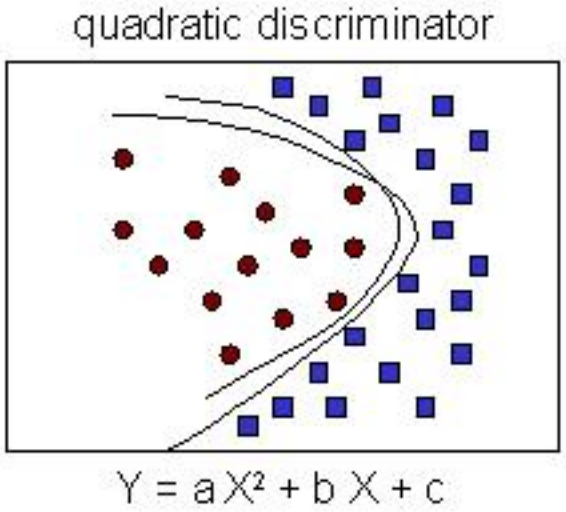
- SLAM components
  - Splice sites (Variable length Markov models).
  - Introns and Intergenic regions (2nd order Markov models, independent geometric lengths, CNS states).
  - Coding sequences (3-periodic Markov models, generalized length distributions, protein-based pairHMM.)
- Input
  - Pair of syntenic genomic sequences.
  - Approximate alignment.
- Output
  - CDS predictions in *both* sequences.


CNS
Y
Z
D
M
I
$E_{0,0}$
$E_{0,1}$
$E_{0,2}$
$E_{1,0}$
$E_{1,1}$
$E_{1,2}$
$E_{2,0}$
$E_{2,1}$
$E_{2,2}$
$Intron_0$
$Intron_1$
$Intron_2$
$E_{I,0}$
$E_{I,1}$
$E_{I,2}$
E_{sing}
$E_{0,F}$
$E_{1,F}$
$E_{2,F}$
Intergene

Дискриминантный анализ.

Программа MZEF.



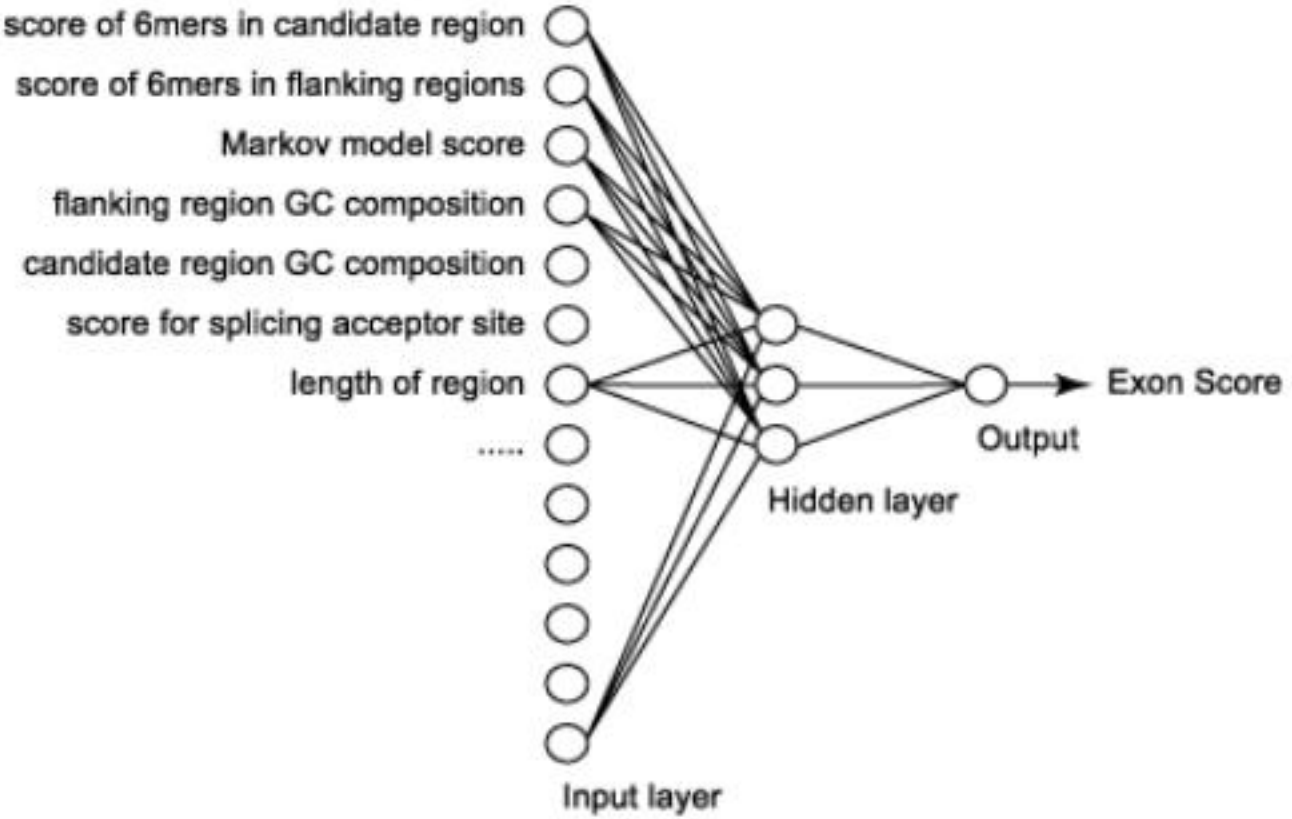


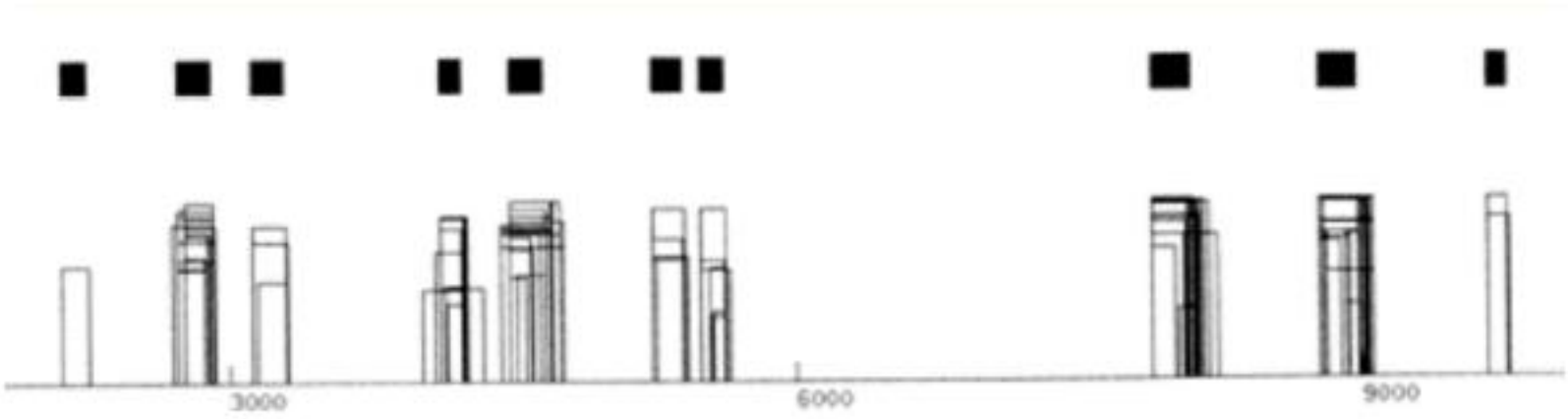
GRAIL

∀• Частоты олигонуклеотидов длины 6.

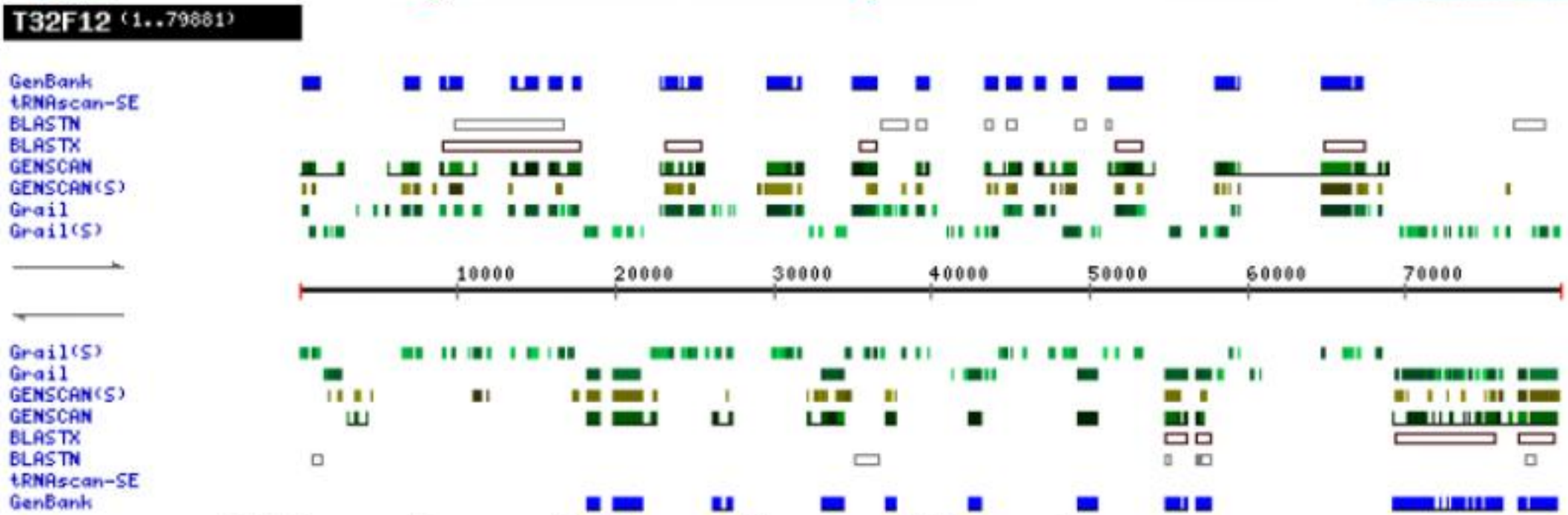
- Использование марковских цепей пятого порядка для трех рамок считывания, подобно GenMark.
- Учет распределения длин экзонов.
- Учет G+C состава 2kb прилежащих районов.
- Нейронные сети для распознавания сайтов сплайсинга.
- и т.д.

Feature vectors	Scores
$<f_{11}, f_{12}, f_{13}, \ldots f_{1n}>$	$score_1$
$<f_{21}, f_{22}, f_{23}, \ldots f_{2n}>$	$score_2$
...	
$<f_{m1}, f_{m2}, f_{m3}, \ldots f_{mn}>$	$score_m$


score of 6mers in candidate region
score of 6mers in flanking regions
Markov model score
flanking region GC composition
candidate region GC composition
score for splicing acceptor site
length of region
.....
Input layer
Hidden layer
Output
Exon Score



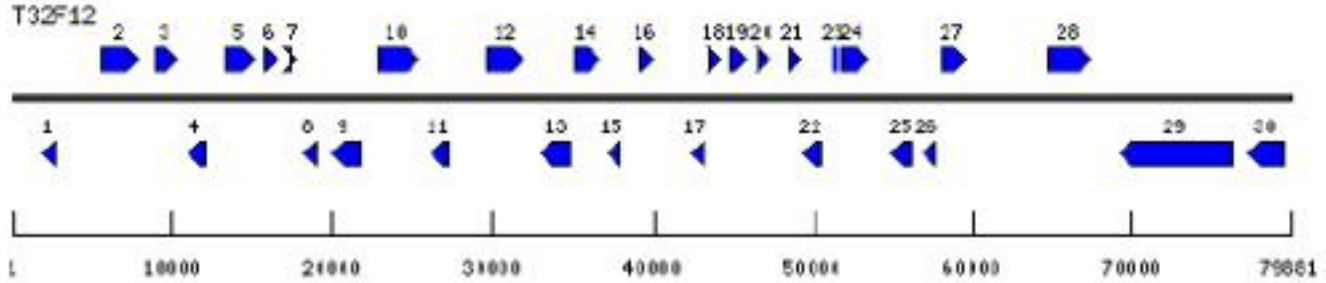
Комбинированные методы распознавания генов



Different predictions from different programs.

TIGR "Combiner"

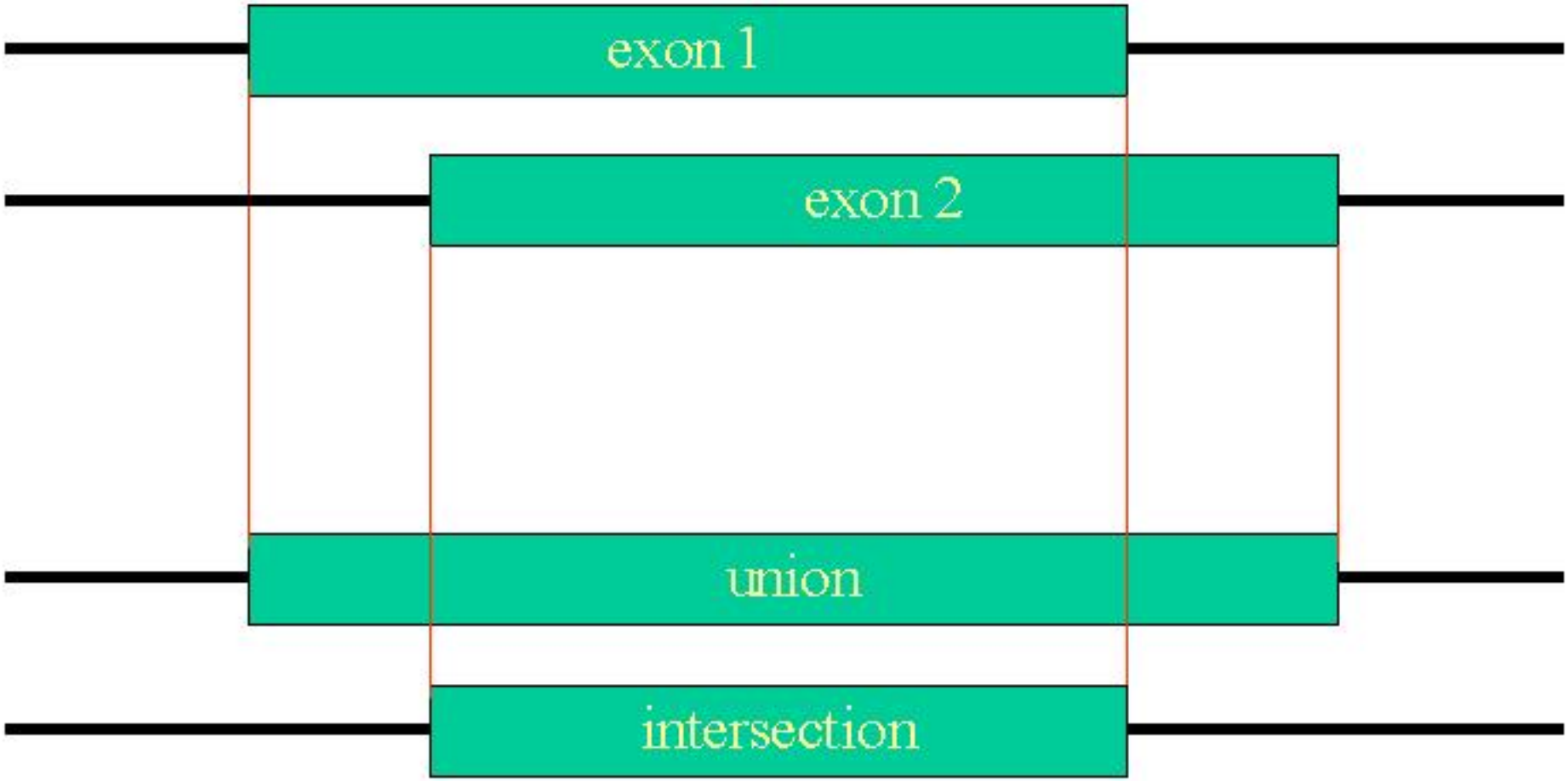
Final prediction is manually determined



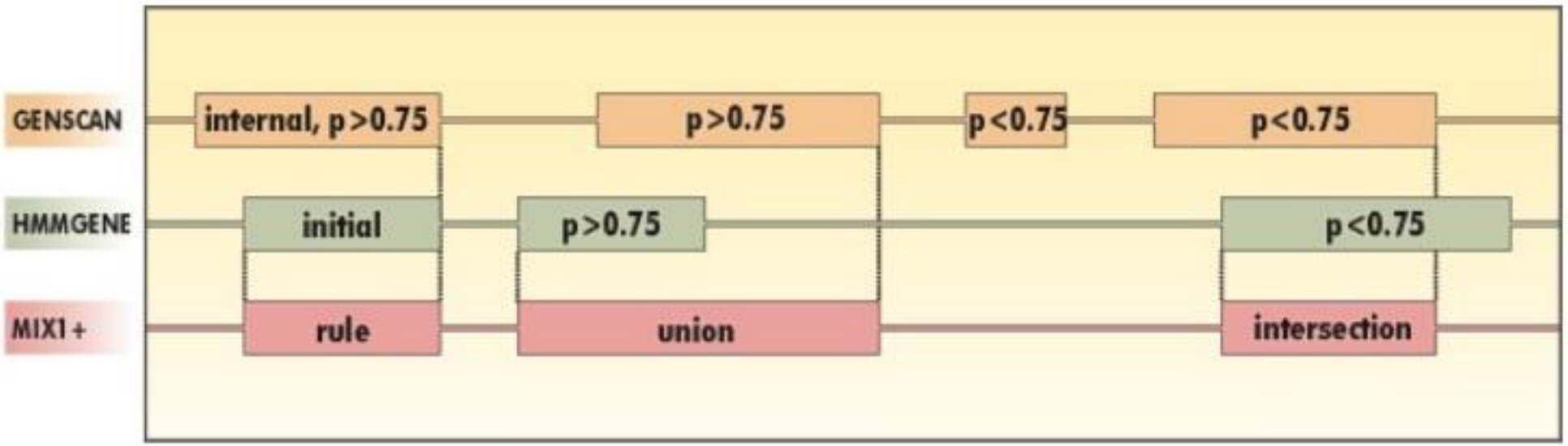
Genes on both strands

(this is one BAC clone drawn with the genes = 80,000 base pairs)

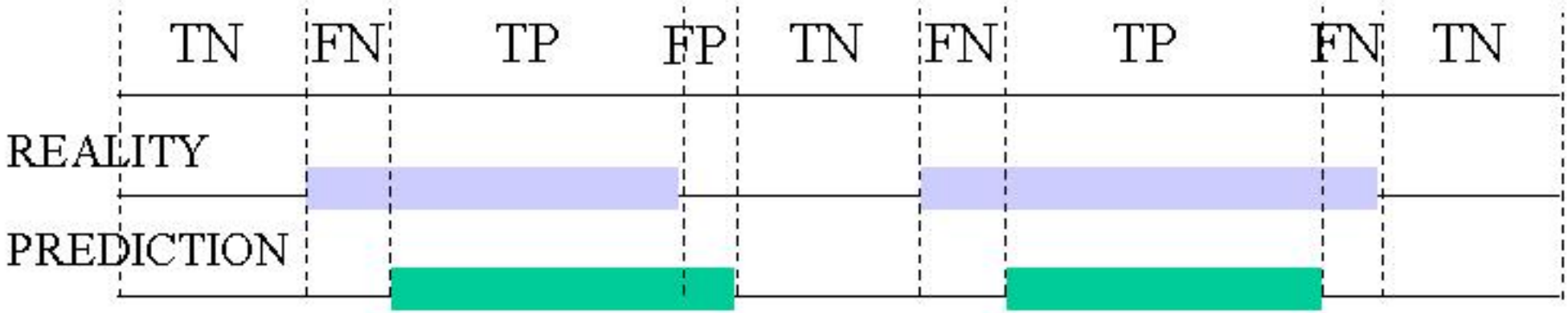
AND и OR методы



EUI Метод


GENSCAN
HMMGENE
MIX1+
internal, p>0.75
p>0.75
p<0.75
p<0.75
initial
p>0.75
p<0.75
rule
union
intersection

Оценка качества работы программ предсказания на уровне нуклеотидов.



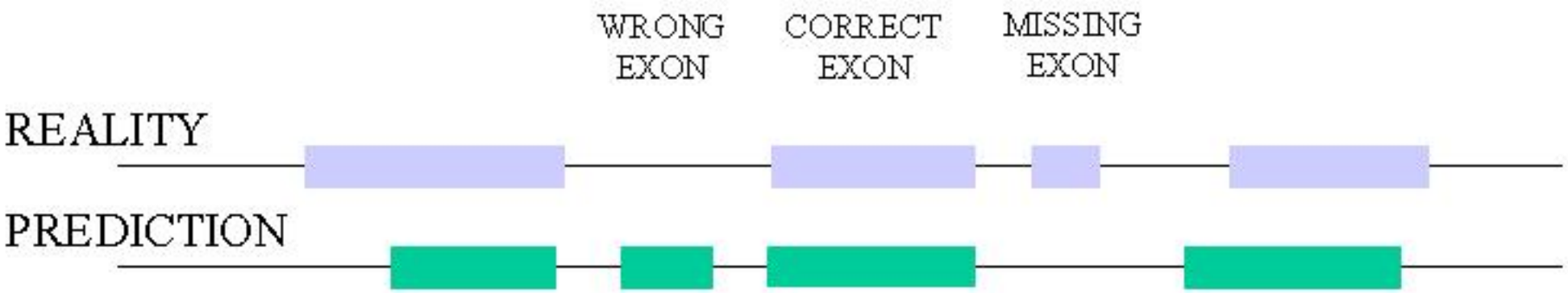
Sensitivity $Sn = \frac{TP}{TP + FN}$

$$\frac{\text{number of correct exons}}{\text{number of actual exons}}$$

Specificity $Sp = \frac{TP}{TP + FP}$

$$\frac{\text{number of correct exons}}{\text{number of predicted exons}}$$

Оценка качества работы программ предсказания на уровне целых экзонов.



$$ESn = \frac{TE}{AE} \qquad ESp = \frac{TE}{PE}$$

$$AC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right) - 1$$

$$CC = \frac{(TP * TN) - (FN * FP)}{((TP+FN) * (TN+FP) * (TP+FP) * (TN+FN))^{\frac{1}{2}}}$$

Table 1. Nucleotide and Exon Level Accuracy

Programs	No. of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	Sp	AC	CC	ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark.hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
Genscan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.38	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69	0.46	0.41	0.43 ± 0.26	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

For each sequence in the HMR195 dataset, the exons predicted on the forward (+) strand were compared to the annotated exons. The standard measures of predictive accuracy on nucleotide and exon level were calculated for each sequence and averaged over all sequences for which they were defined. This was done separately for each of the programs tested.

(No. of sequences) number of sequences effectively analyzed by each program; in parentheses is the number of sequences where the absence of gene was predicted; (Sn) nucleotide level sensitivity; (Sp) nucleotide level specificity; (AC) approximate correlation; (CC) correlation coefficient; (ESn) exon level sensitivity; (ESp) exon level specificity; (ME) missed exons; (WE) wrong exons; (PCa) proportion of real exons that were partially predicted (only one exon boundary correct); (PCp) proportion of predicted exons that were only partially correct; (OL) proportion of predicted exons that overlap an actual exon. AC and (ESn+ESp)/2 are given with standard deviation.

Спасибо за внимание!