Introduction to Bioinformatics

JTMS-19

Marc-Thorsten Hütt mhuett@constructor.university

Felix Jonas fjonas@constructor.university

session Wed, 6. Nov. 2024

Hidden Markov models I

What is this session about?

Hidden Markov models (HMMs) are introduced. Viterbi algorithm of HMMs is discussed. First ideas of posterior decoding are introduced.

How can you revise the material after the session?

Read Durbin et al. chapters 3.2, 3.3, 3.4
Read Baxevanis/Oullette pages 208 – 210
alternative reading: Hütt/Dehnert chapters 2.8.1 – 2.8.4

Markov chains as a tool for studying CpG islands

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

$$S(x) = \log\left(\frac{P(x \mid \text{model } +)}{P(x \mid \text{model } -)}\right) = \log\left(\frac{P(B)\prod_{i=1}^{L} a^{+}_{x_{i-1}x_i}}{P(B)\prod_{i=1}^{L} a^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \log\left(\frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

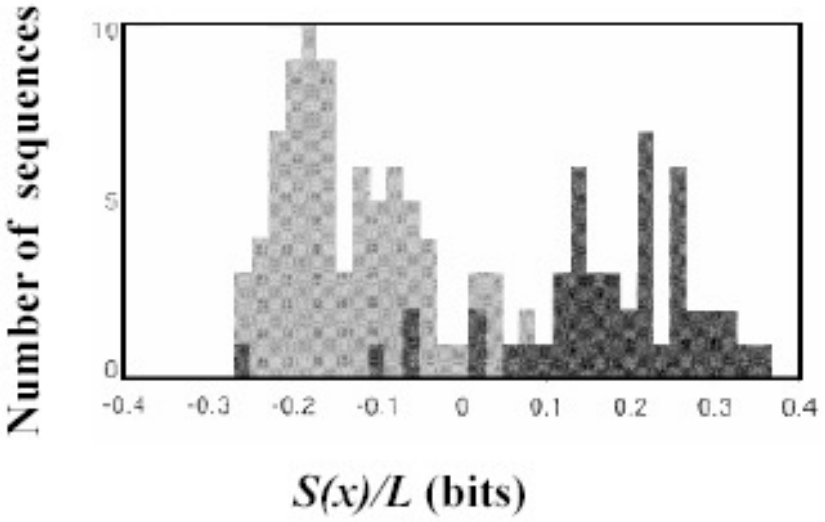
a number for each sequence x
→ histogram of score values S(x) for many sequences x

a number for each dinucleotide
→ table of "log-likelihoods"

Markov chains as a tool for studying CpG islands

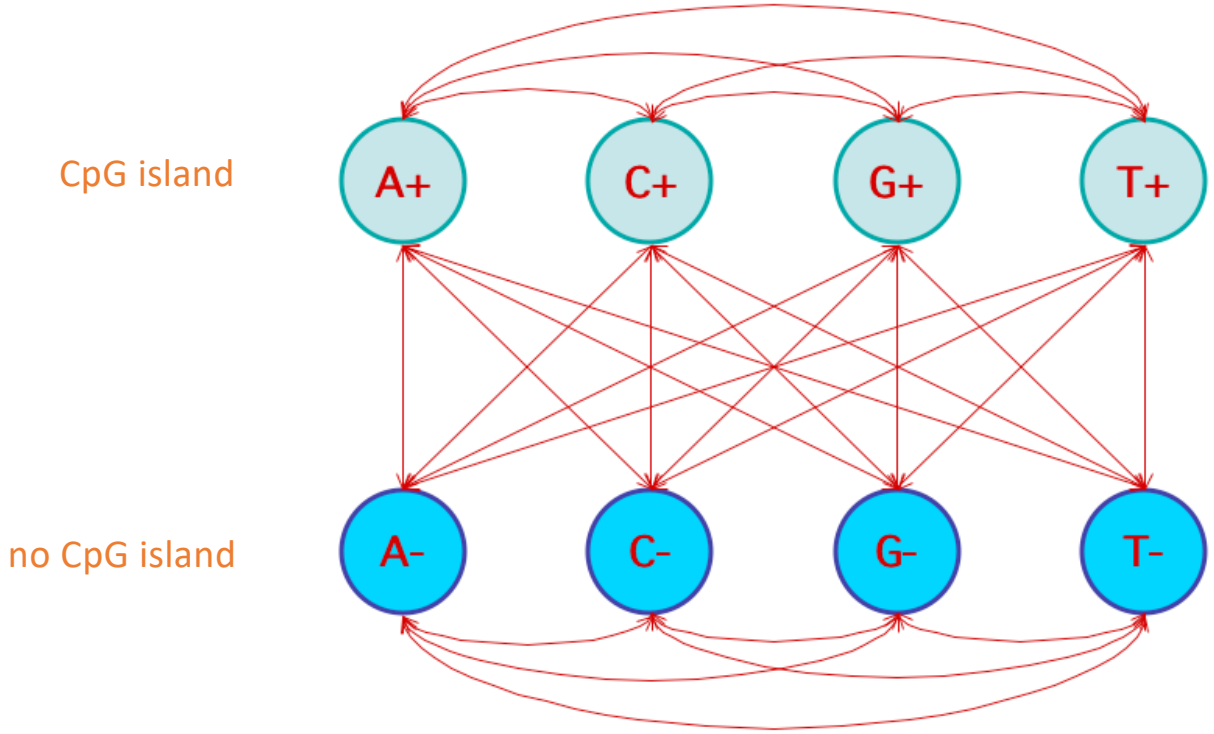
ß(log$_2$)	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.0685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

table of "log-likelihoods"



histogram of scores

Hidden Markov model (HMM) for CpG islands

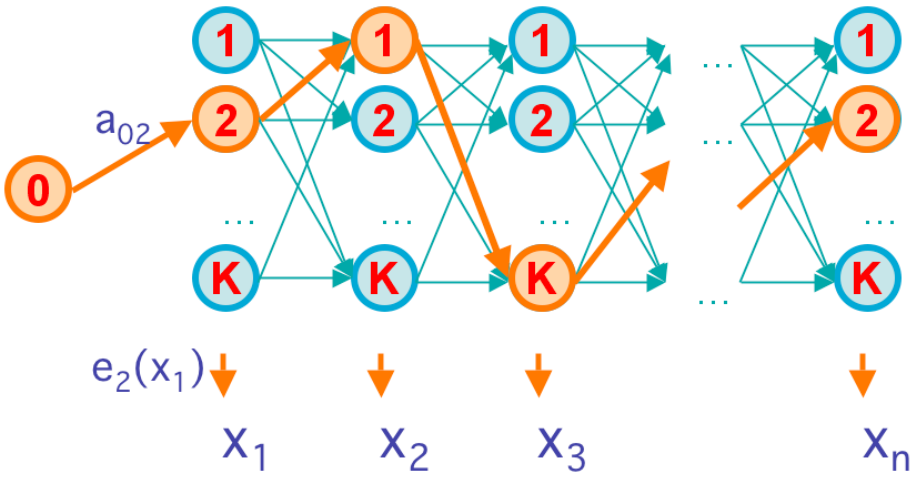


internal state: $A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-$

emitted state: A C G T A C G T

general properties of HMM

simulation of a sequence



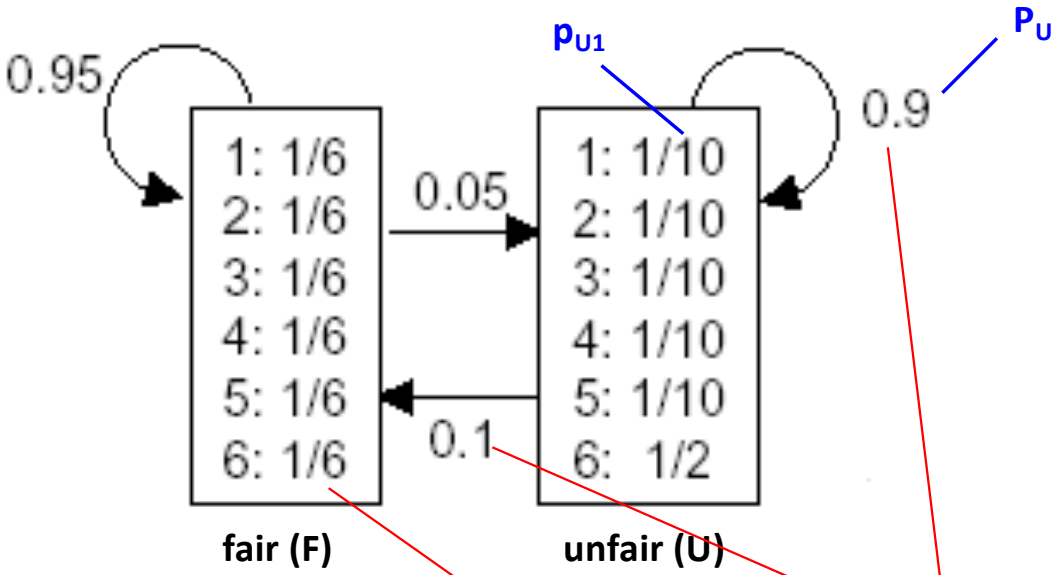
internal path

observed sequence x

joint probability of the path and the sequence

$$P(x, \pi) = a_{\pi_0 \pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i)\, a_{\pi_i \pi_{i+1}}$$

elementary example: casino with two dice



S_{HMM} = { F, U }

S_{HMM} = { 1F, 2F, ... , 6F, 1U, 2U, ... 6U }

- emission prob's are 0 or 1
- transition prob's are products (e.g. $a_{2U,1U} = P_U\, p_{U1}$)

Training:

parameter estimation from an ensemble of sequences with a given internal structure

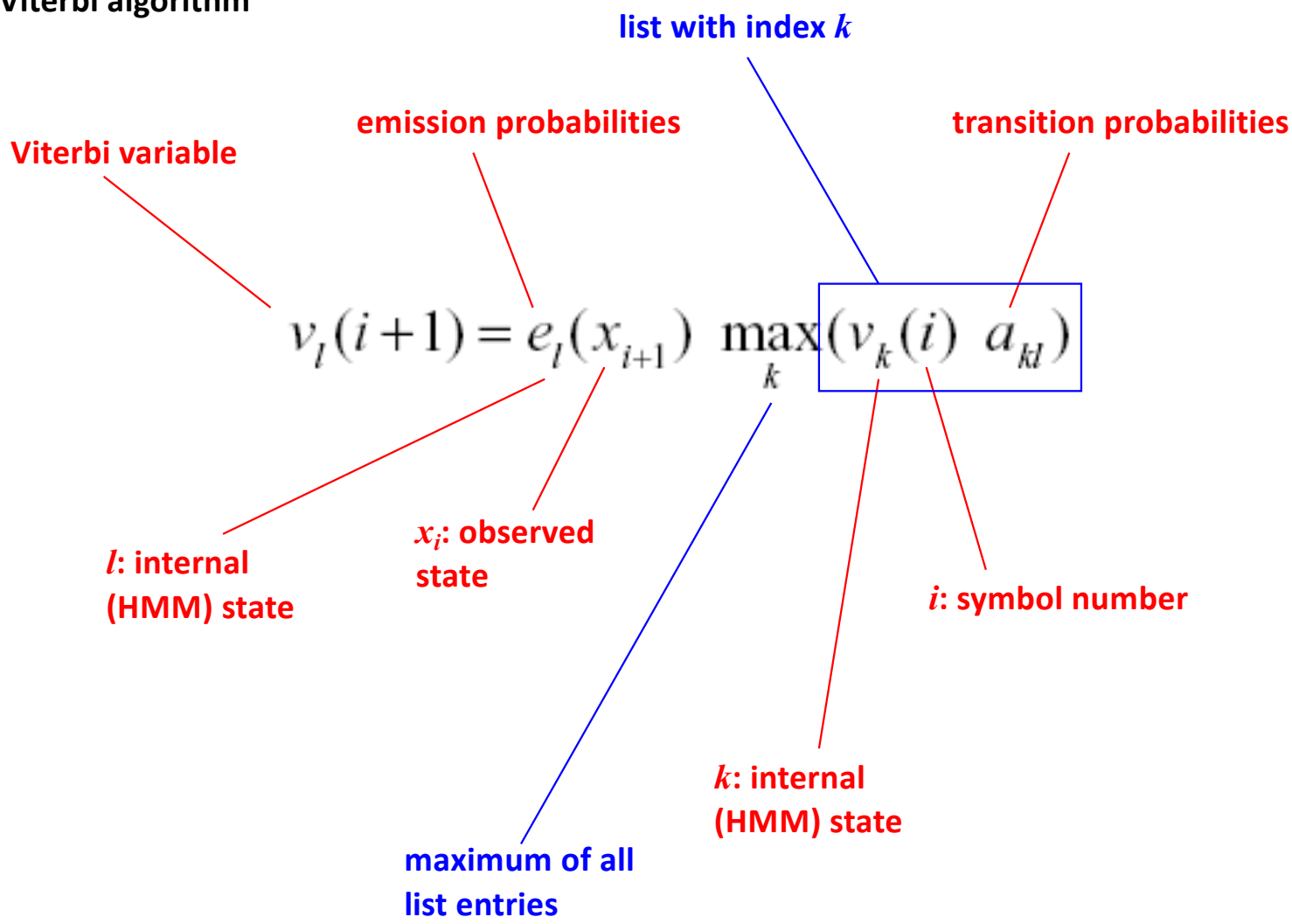
Evaluation:

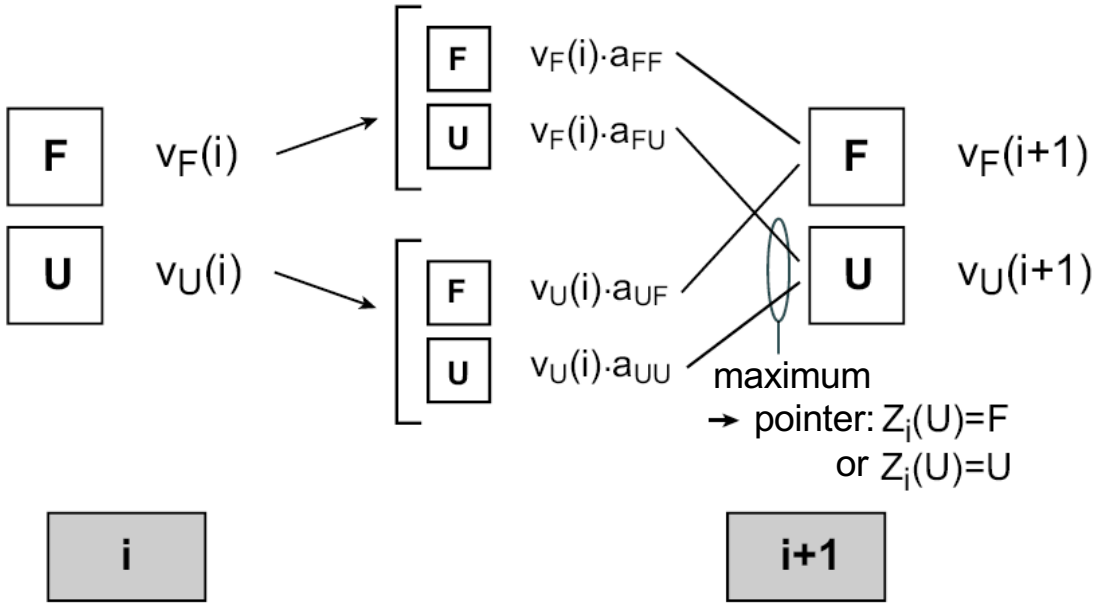
obtaining the sequence x out of a given internal sequence (path) p

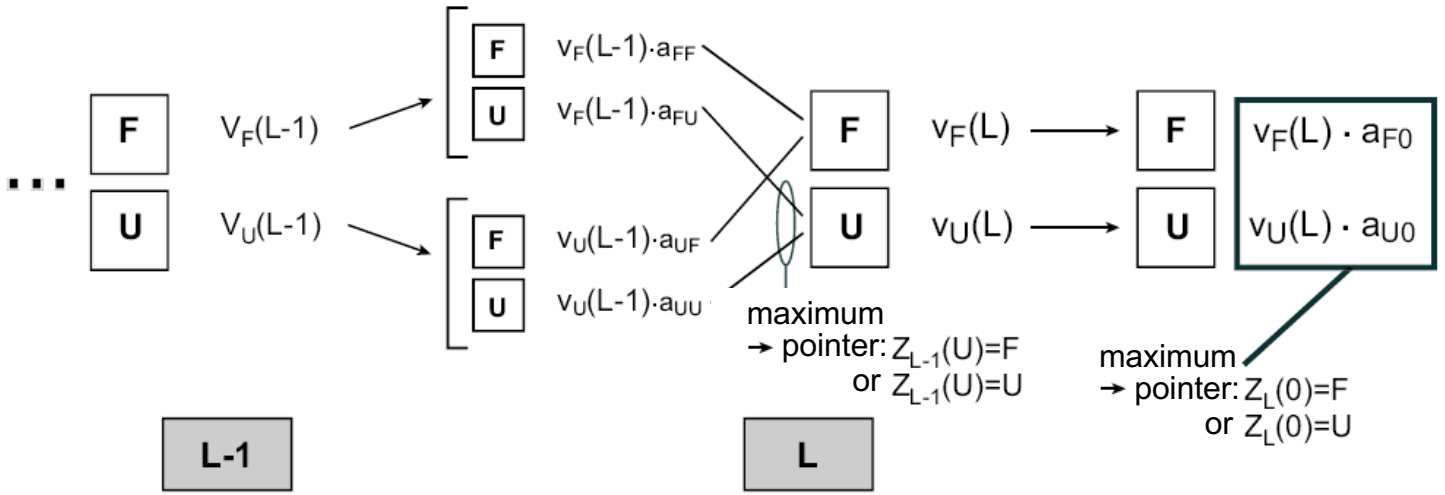
Decoding:

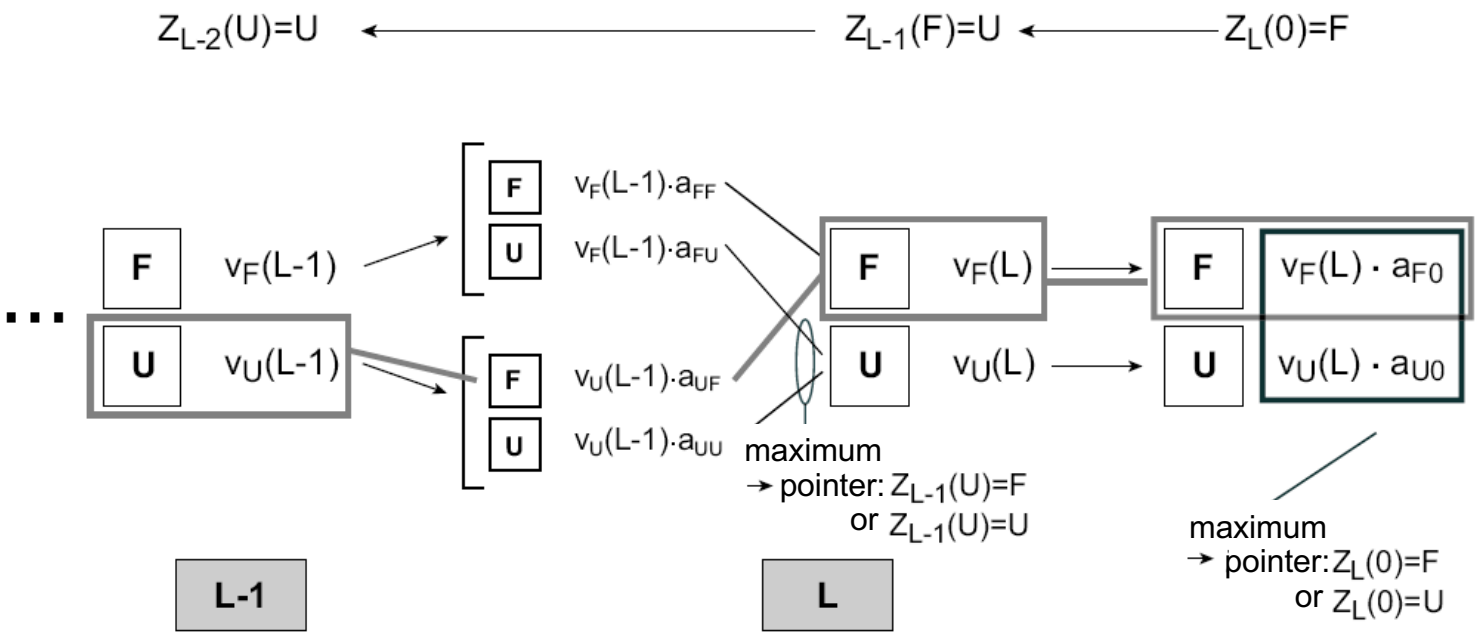
finding the internal path p behind an observed sequence x

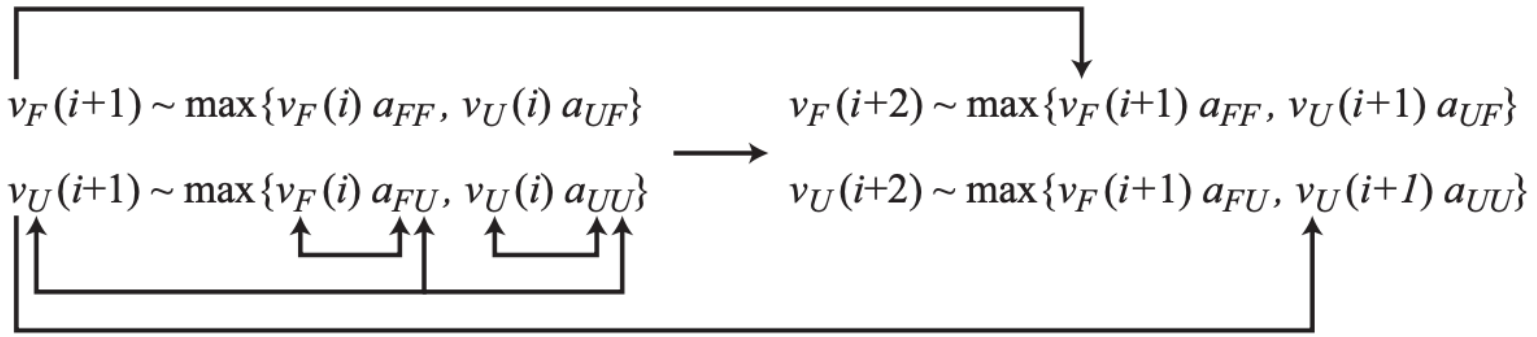
Viterbi algorithm




F
U
$v_F(i)$
$v_U(i)$
F
U
F
U
$v_F(i)\cdot a_{FF}$
$v_F(i)\cdot a_{FU}$
$v_U(i)\cdot a_{UF}$
$v_U(i)\cdot a_{UU}$
F
U
$v_F(i+1)$
$v_U(i+1)$
maximum
→ pointer: $Z_i(U)=F$
or $Z_i(U)=U$
i
i+1


F
U
$v_F(L-1)$
$v_U(L-1)$
F
U
F
U
$v_F(L-1)\cdot a_{FF}$
$v_F(L-1)\cdot a_{FU}$
$v_U(L-1)\cdot a_{UF}$
$v_U(L-1)\cdot a_{UU}$
maximum
→ pointer: $Z_{L-1}(U)=F$
or $Z_{L-1}(U)=U$
F
U
$v_F(L)$
$v_U(L)$
F
U
$v_F(L)\cdot a_{F0}$
$v_U(L)\cdot a_{U0}$
maximum
→ pointer: $Z_L(0)=F$
or $Z_L(0)=U$
L-1
L


$Z_{L-2}(U)=U$
$Z_{L-1}(F)=U$
$Z_L(0)=F$
F
U
$v_F(L-1)$
$v_U(L-1)$
$v_F(L-1)\cdot a_{FF}$
$v_F(L-1)\cdot a_{FU}$
$v_U(L-1)\cdot a_{UF}$
$v_U(L-1)\cdot a_{UU}$
$v_F(L)$
$v_U(L)$
$v_F(L)\cdot a_{F0}$
$v_U(L)\cdot a_{U0}$
maximum
→ pointer: $Z_{L-1}(U)=F$
or $Z_{L-1}(U)=U$
maximum
→ pointer: $Z_L(0)=F$
or $Z_L(0)=U$
L-1
L


$v_F(i+1) \sim \max\{v_F(i)\, a_{FF},\, v_U(i)\, a_{UF}\}$
$v_U(i+1) \sim \max\{v_F(i)\, a_{FU},\, v_U(i)\, a_{UU}\}$
$v_F(i+2) \sim \max\{v_F(i+1)\, a_{FF},\, v_U(i+1)\, a_{UF}\}$
$v_U(i+2) \sim \max\{v_F(i+1)\, a_{FU},\, v_U(i+1)\, a_{UU}\}$

Casino: results

```
Rolls    3151162464466442453211316311641521336251445436316566265666666
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLL
---------------------------------------------------------------------
Rolls    6511664531326512456366646316366631623264552352666666251516 31
Die      LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi  LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
---------------------------------------------------------------------
Rolls    2225554416665665635643243641315134651463534111264146262533 56
Die      FFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
---------------------------------------------------------------------
Rolls    3661636664662325344136616611632525624622552652522664353533 36
Die      LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi  LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
---------------------------------------------------------------------
Rolls    2331216253644144323351632436336655624666626326666123552452 42
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFF
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLFFFFFFFF
```

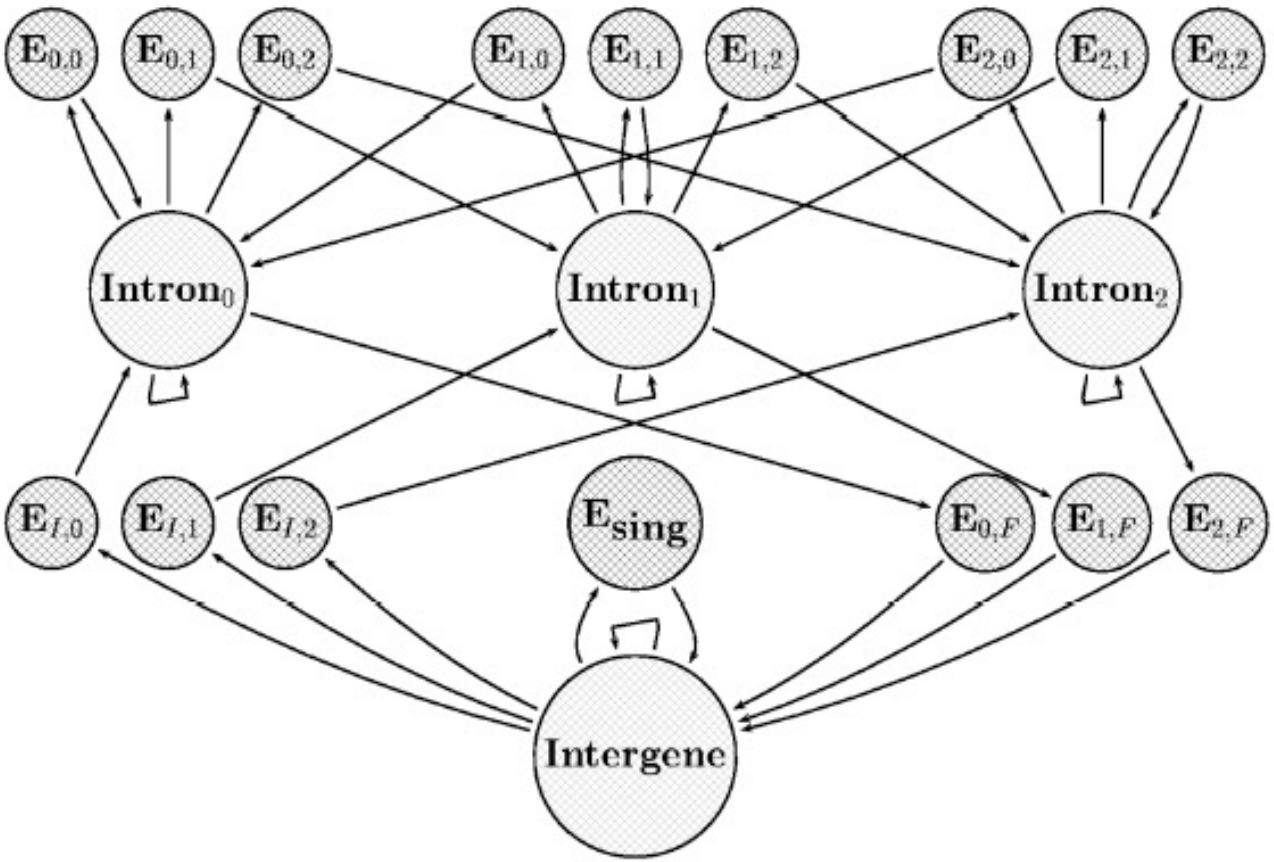
Finding the genes in genomic DNA

Christopher B Burge* and Samuel Karlin†

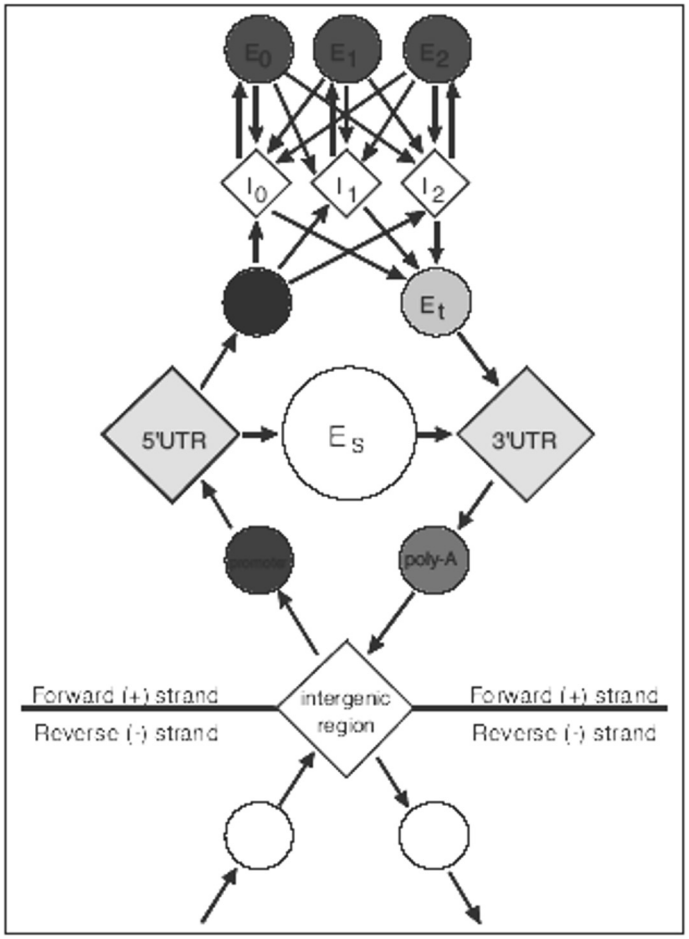
Addresses
*Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, 40 Ames Street, E17-526 Cambridge, MA 02139, USA; e-mail: cburge@mit.edu
†Department of Mathematics, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA; e-mail: sam@galois.stanford.edu
Correspondence: Samuel Karlin

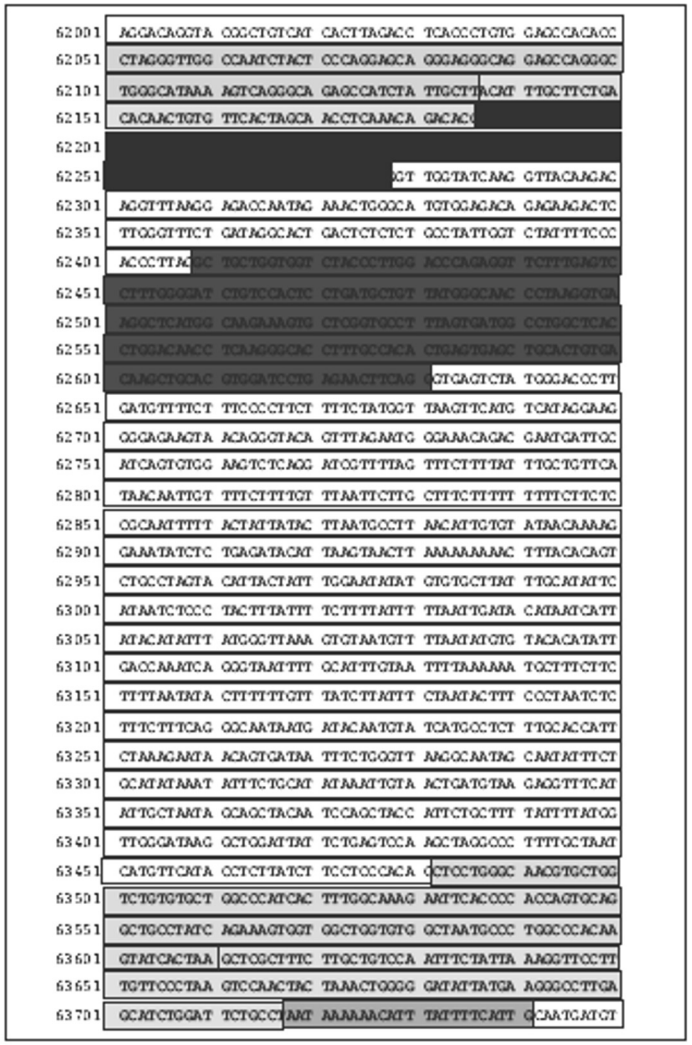
Current Opinion in Structural Biology 1998, **8**:346–354

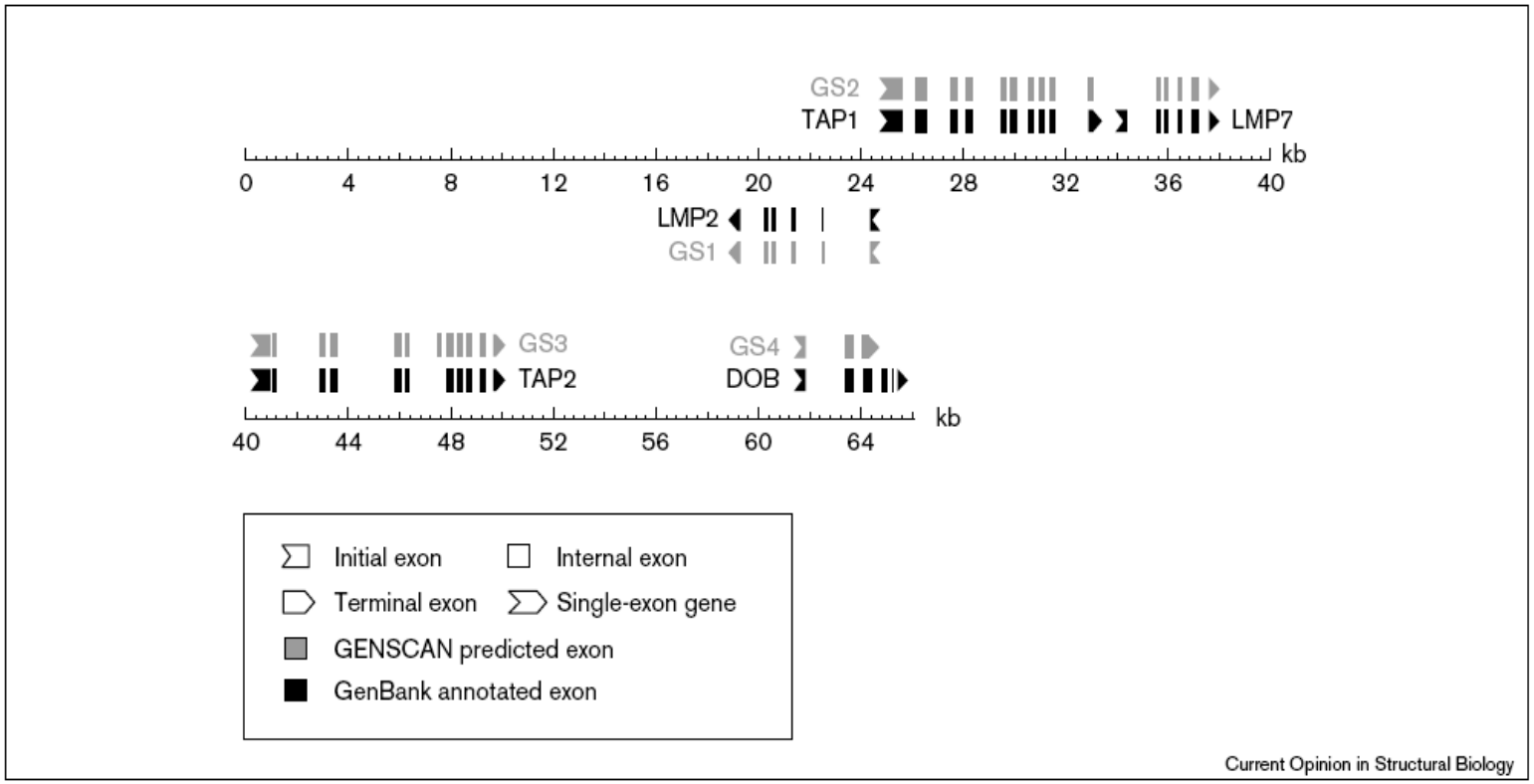
The GenScan HMM

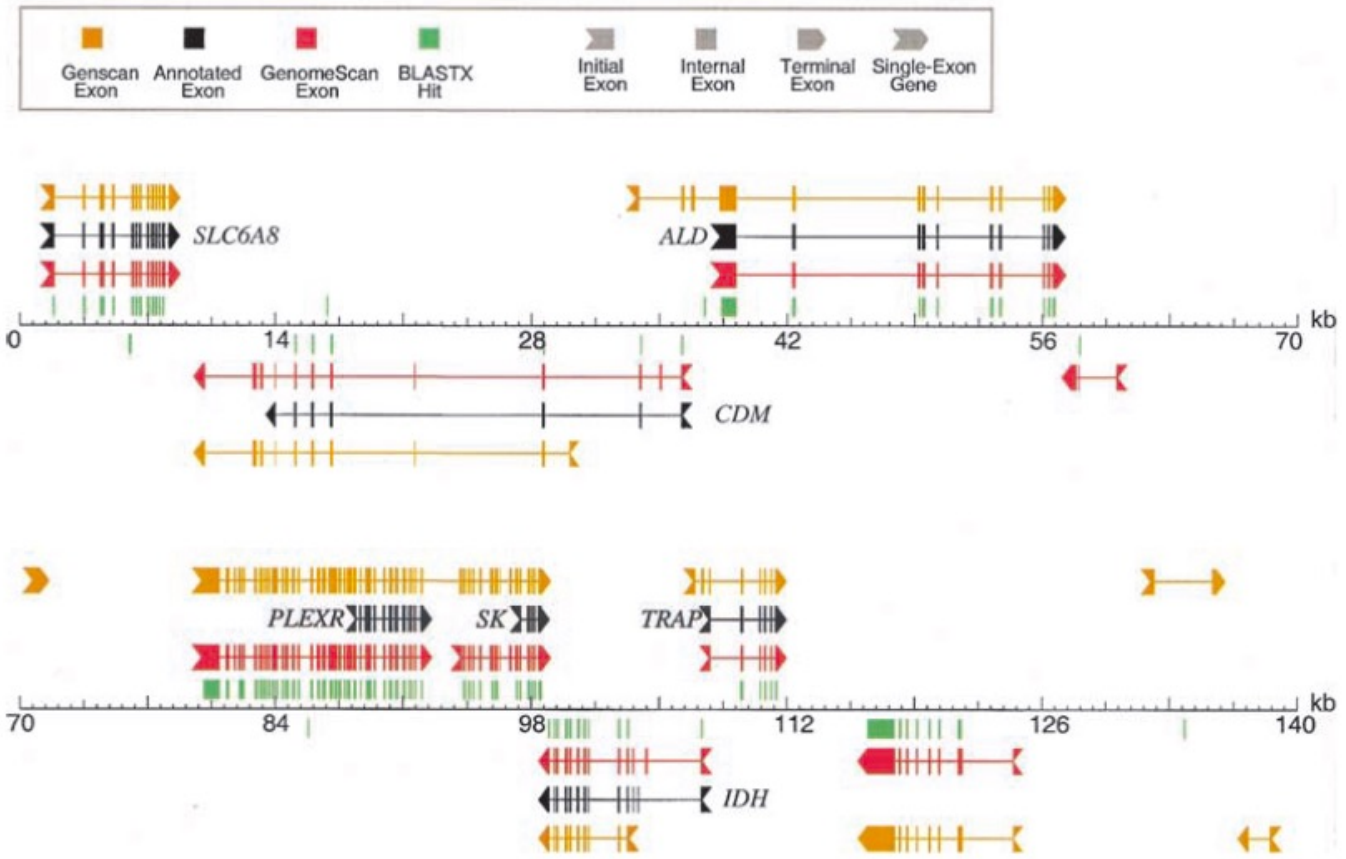


GENSCAN (Burge & Karlin)



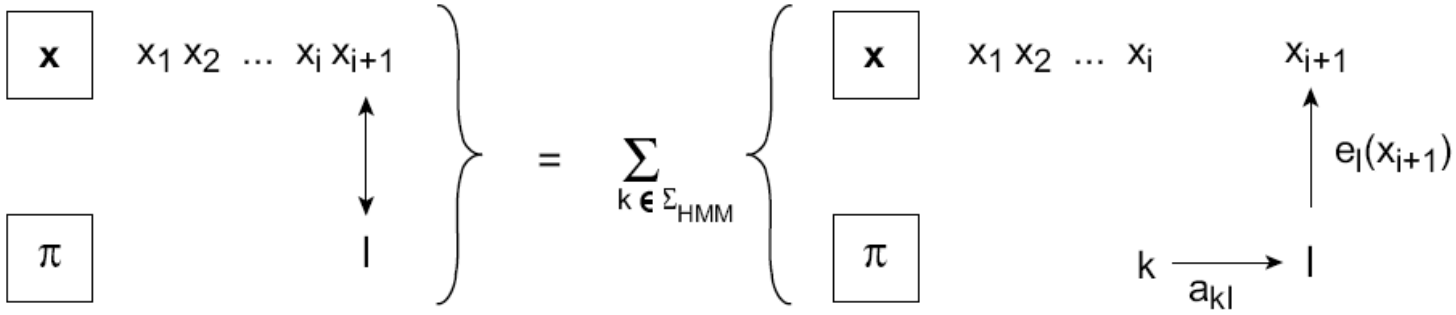



GS2
TAP1
LMP7
kb
0
4
8
12
16
20
24
28
32
36
40
LMP2
GS1
GS3
TAP2
GS4
DOB
kb
40
44
48
52
56
60
64
Initial exon
Internal exon
Terminal exon
Single-exon gene
GENSCAN predicted exon
GenBank annotated exon
Current Opinion in Structural Biology



Yeh, Lim, and Burge, Genome Research 11:803-816, 2001.

Casino: results



Summary of the Viterbi algorithm

initialization → recursion → termination → traceback

$$v_0(0) = 1\ ,\ v_k(0) = 0 \quad \forall\, k \in \Sigma_{HMM}$$ initialization

$$v_l(i+1) = e_l(x_{i+1}) \max_k \{v_k(i)\, a_{kl}\}$$ recursion

$$\max_k \{v_k(L)\, a_{k0}\} = P(x, \pi^*)$$ termination

$$a_{k0} = \frac{1}{|\Sigma_{HMM}|} \quad \forall\, k \in \Sigma_{HMM}$$

$$\pi^*_L$$

$$\pi^*_{L-1} = Z_{L-1}(\pi^*_L)$$

$$\vdots$$

$$\pi^*_{i-1} = Z_{i-1}(\pi^*_i)$$

$$\vdots$$

traceback