Introduction to Bioinformatics

JTMS-19

Marc-Thorsten Hütt mhuett@constructor.university

Felix Jonas fjonas@constructor.university

session Wed, 20. Nov. 2024

Hidden Markov models III

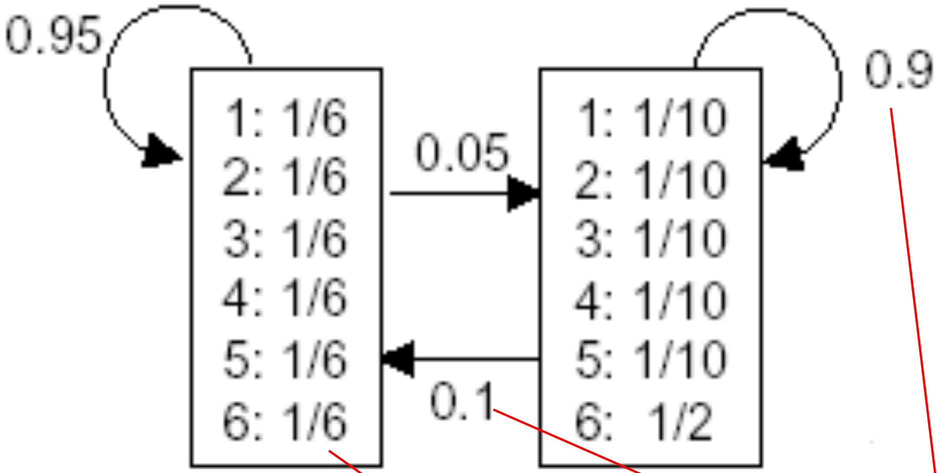
What is this session about?

Repetition of posterior decoding for HMMs. Additional technical aspects of HMMs are discussed. First applications of HMMs are explored.

How can you revise the material after the session?

Read Durbin et al. chapters 3.3, 3.4
Read Baxevanis/Oullette pages 208 – 210
Look at the HMM in Kundaje, et al. (2015). Nature, 518, 317-330.
alternative reading: Hütt/Dehnert chapters 2.8.3 – 2.8.4, 2.9

elementary example: casino with two dice



fair (F)

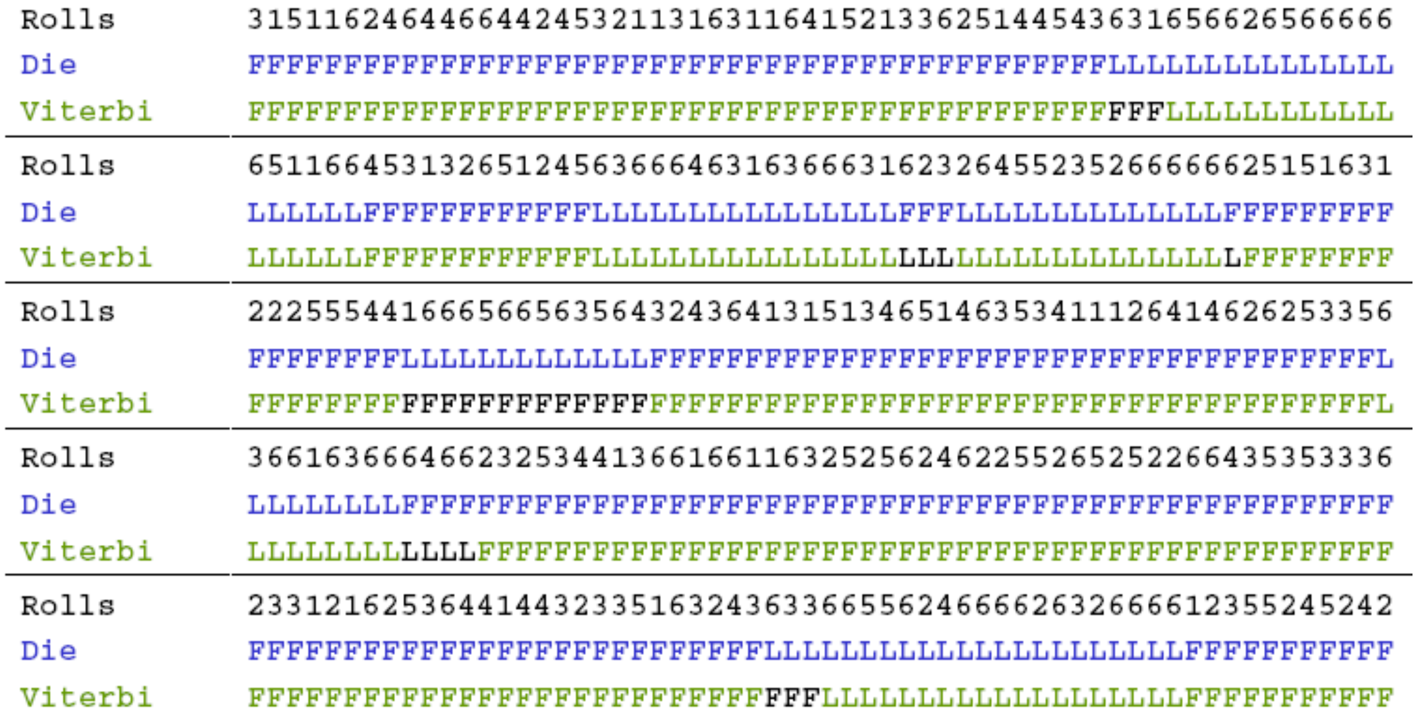
unfair (U)

emission probabilities

transition probabilities

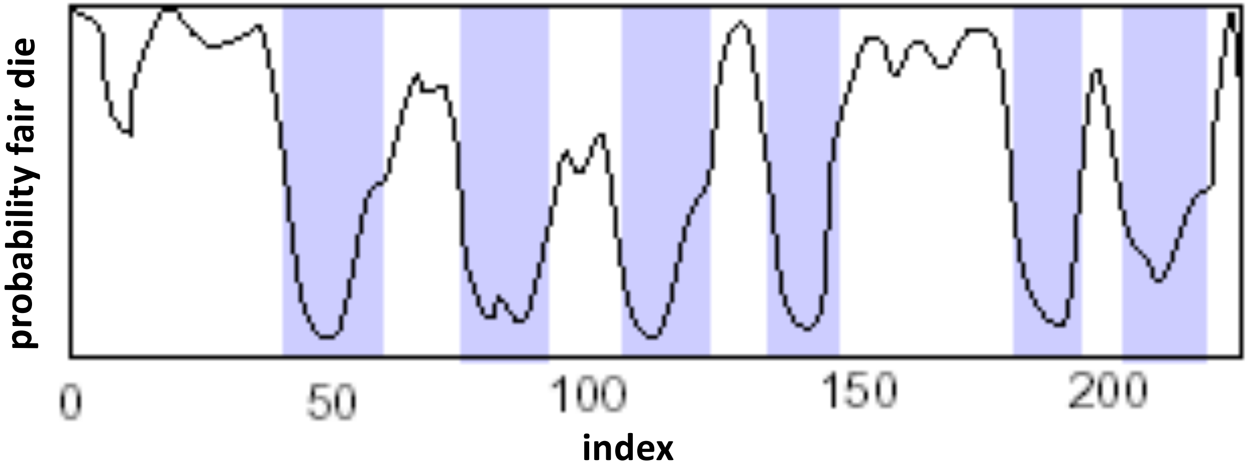
S_{HMM} = { F, U }

Casino: results



```
Rolls      315116246446644245321131631164152133625144543631656626566666
Die        FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLL
Rolls      651166453132651245636664631636663162326455235266666625151631
Die        LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLLFFFFFFFFF
Viterbi    LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
Rolls      222555441666566563564324364131513465146353411126414626253356
Die        FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL
Rolls      366163666466232534413661661163252562462255265252266435353336
Die        LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi    LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Rolls      233121625364414432335163243633665562466662632666612355245242
Die        FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
Viterbi    FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
```

Casino: results



Summary of the forward and backward algorithms

$$P(x,\pi_i=k) = \underbrace{P(x_1,\ldots,x_i,\pi_i=k)}_{\equiv f_k(i)}\ \underbrace{P(x_{i+1},\ldots,x_L\,|\,x_1,\ldots,x_i,\pi_i=k)}_{\substack{=P(x_{i+1},\ldots,x_L\,|\,\pi_i=k)\\ \equiv b_k(i)}}$$

$$f_0(0)=1\ ,\ f_k(0)=0\ \ \forall k\in\Sigma_{HMM}$$

initialization

$$f_l(i+1)=e_l(x_{i+1})\sum_k f_k(i)a_{kl}$$

recursion

$$P(x)=\sum_{k\in\Sigma_{HMM}} f_k(L)\,a_{k0}$$

termination

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

$$P(x)=\sum_\pi P(x,\pi)$$

marginal probability

Summary of the forward and backward algorithms

$$P(x, \pi_i = k) = \underbrace{P(x_1, \ldots, x_i, \pi_i = k)}_{\equiv f_k(i)} \; \underbrace{P(x_{i+1}, \ldots, x_L \,|\, x_1, \ldots, x_i, \pi_i = k)}_{\substack{= P(x_{i+1}, \ldots, x_L \,|\, \pi_i = k) \\ \equiv b_k(i)}}$$

$$b_k(L) = a_{k0} \quad k \in \Sigma_{HMM}$$ initialization

$$b_k(i) = \sum_{l \in \Sigma_{HMM}} a_{kl} e_l(x_{i+1}) \, b_l(i+1)$$ recursion

$$P(x) = \sum_{l \in \Sigma_{HMM}} a_{0l} e_l(x_1) \, b_l(1)$$ termination

Summary of the forward and backward algorithms

$$P(x, \pi_i = k) = \underbrace{P(x_1, \dots, x_i, \pi_i = k)}_{\equiv f_k(i)} \underbrace{P(x_{i+1}, \dots, x_L \,|\, x_1, \dots, x_i, \pi_i = k)}_{\substack{= P(x_{i+1}, \dots, x_L \,|\, \pi_i = k) \\ \equiv b_k(i)}}$$

$$P(x,\ \pi_i = k) = f_k(i)\, b_k(i)$$

intermediate result

$$P(x,\ \pi_i = k) = P(\pi_i = k \,|\, x) P(x)$$

definition of the conditional probability

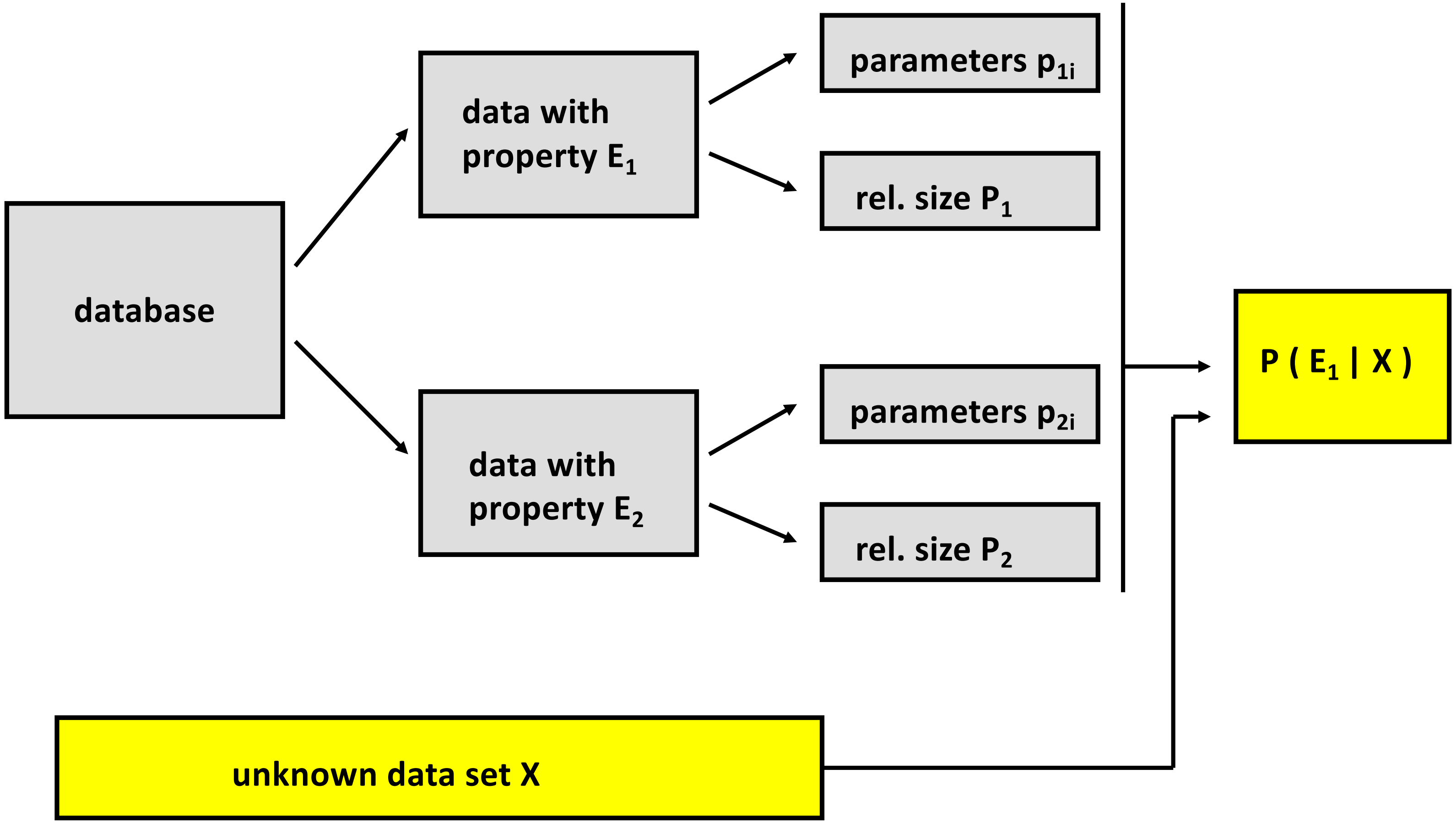
$$P(\pi_i = k \,|\, x) = \frac{f_k(i)\, b_k(i)}{P(x)}$$

final result:
posterior probability of a HMM state *k* at position i given the sequence *x.*

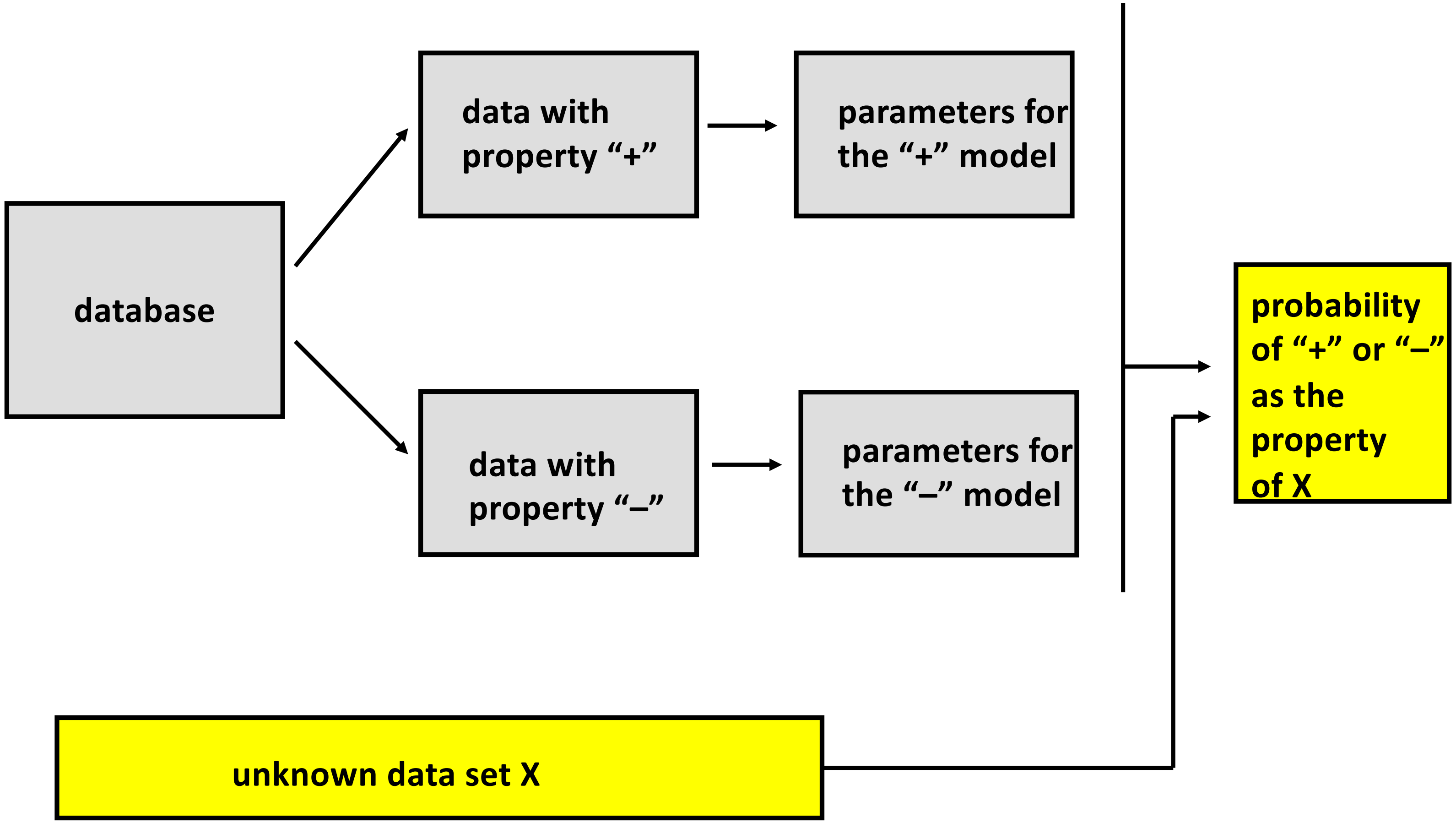
'posterior decoding'

HMM strategies

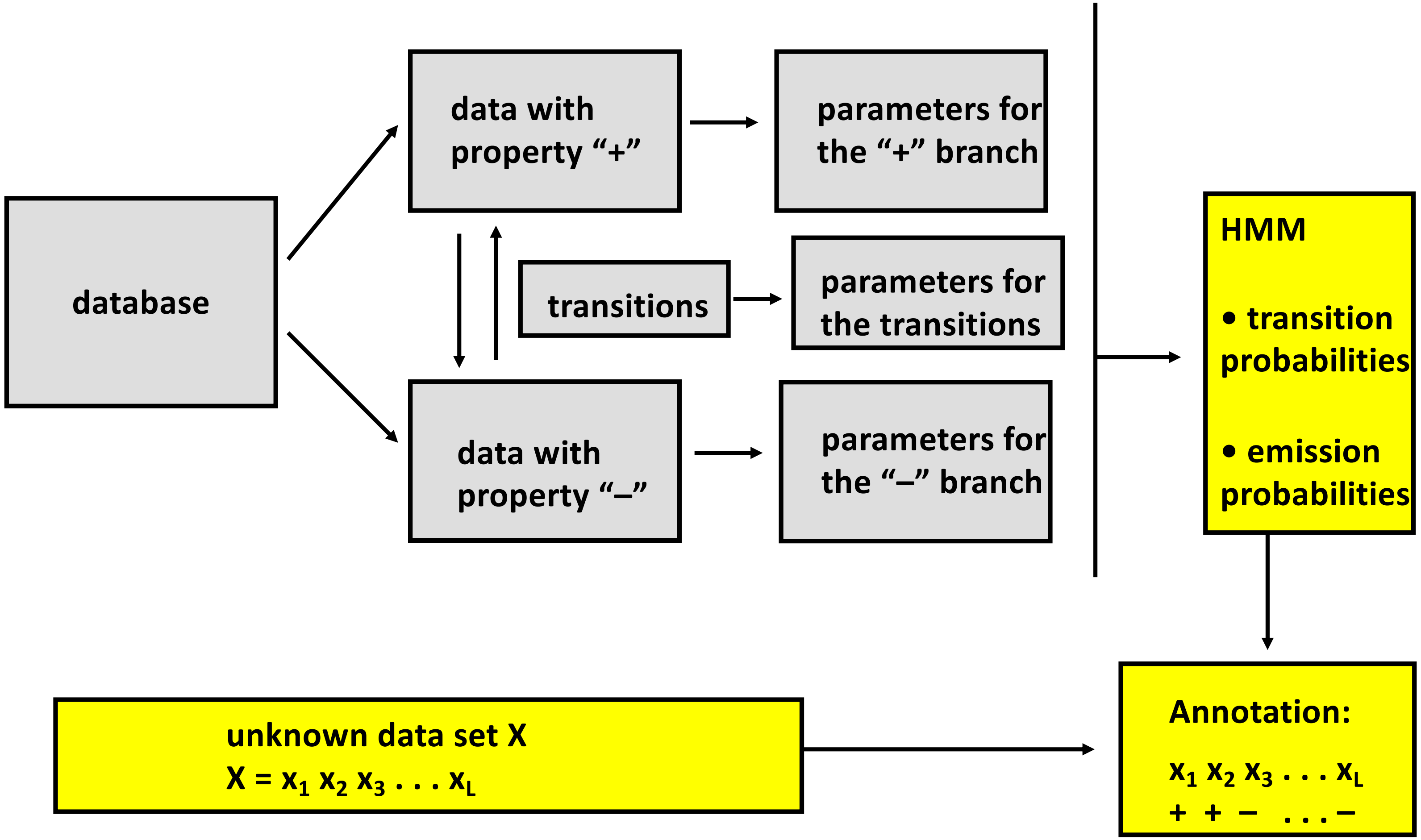
(1) Evaluation of the posterior probability (repeated)



(2) Predicting a property (Markov models)


database
data with property “+”
data with property “–”
parameters for the “+” model
parameters for the “–” model
probability of “+” or “–” as the property of X
unknown data set X

(3) Predicting an internal structure (Hidden Markov models)



database

data with property "+"

data with property "–"

transitions

parameters for the "+" branch

parameters for the transitions

parameters for the "–" branch

HMM

- transition probabilities
- emission probabilities

unknown data set X

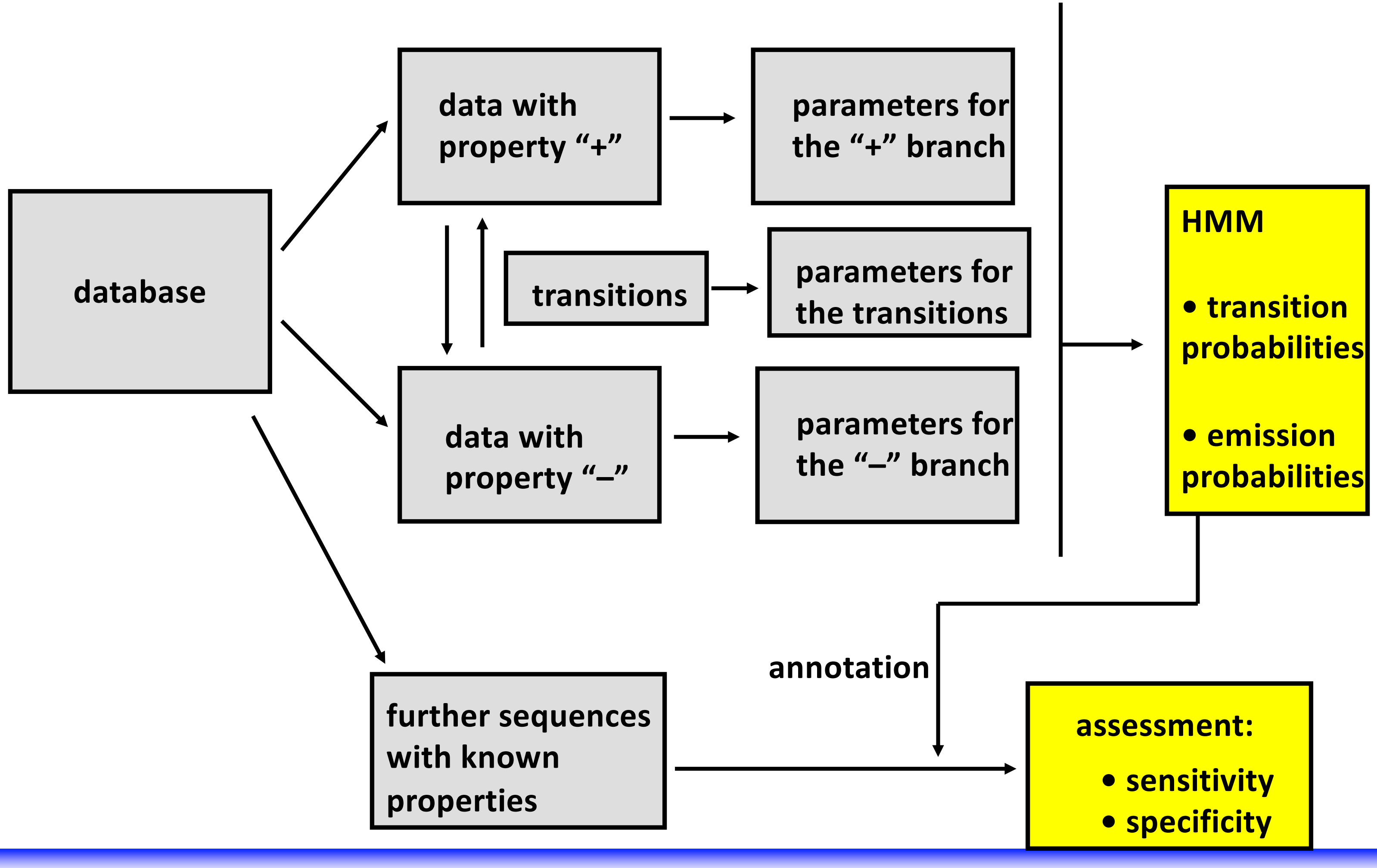
$X = x_1\ x_2\ x_3\ .\ .\ .\ x_L$

Annotation:

$x_1\ x_2\ x_3\ .\ .\ .\ x_L$

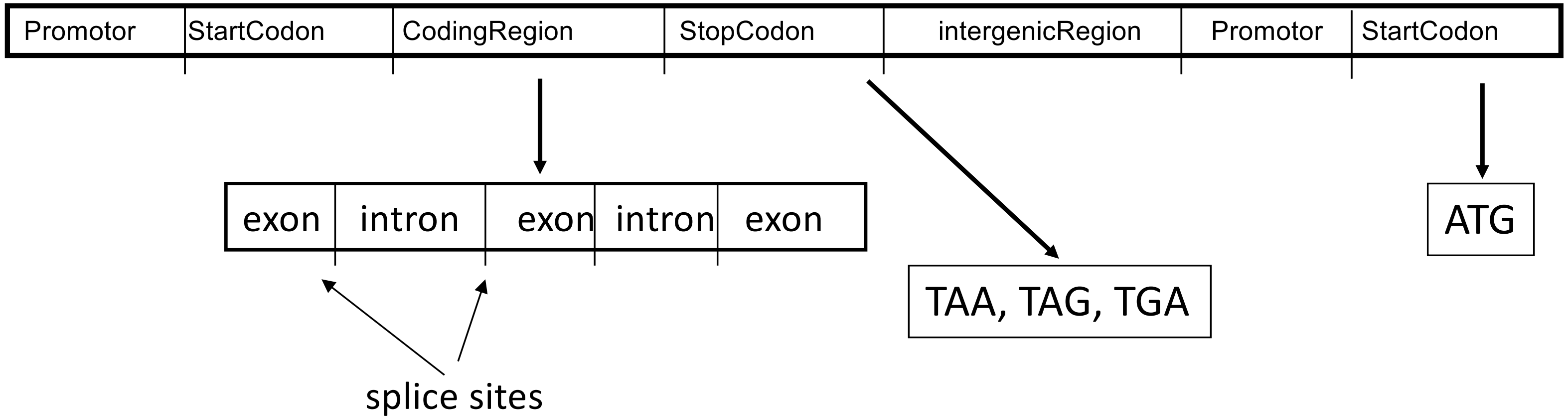
$+\ +\ -\ \ .\ .\ .\ -$

(4) Quality assessment for a Hidden Markov model


database
data with property "+"
data with property "–"
transitions
parameters for the "+" branch
parameters for the transitions
parameters for the "–" branch
HMM
• transition probabilities
• emission probabilities
further sequences with known properties
annotation
assessment:
• sensitivity
• specificity

Application 1:
Gene prediction and gene structure prediction

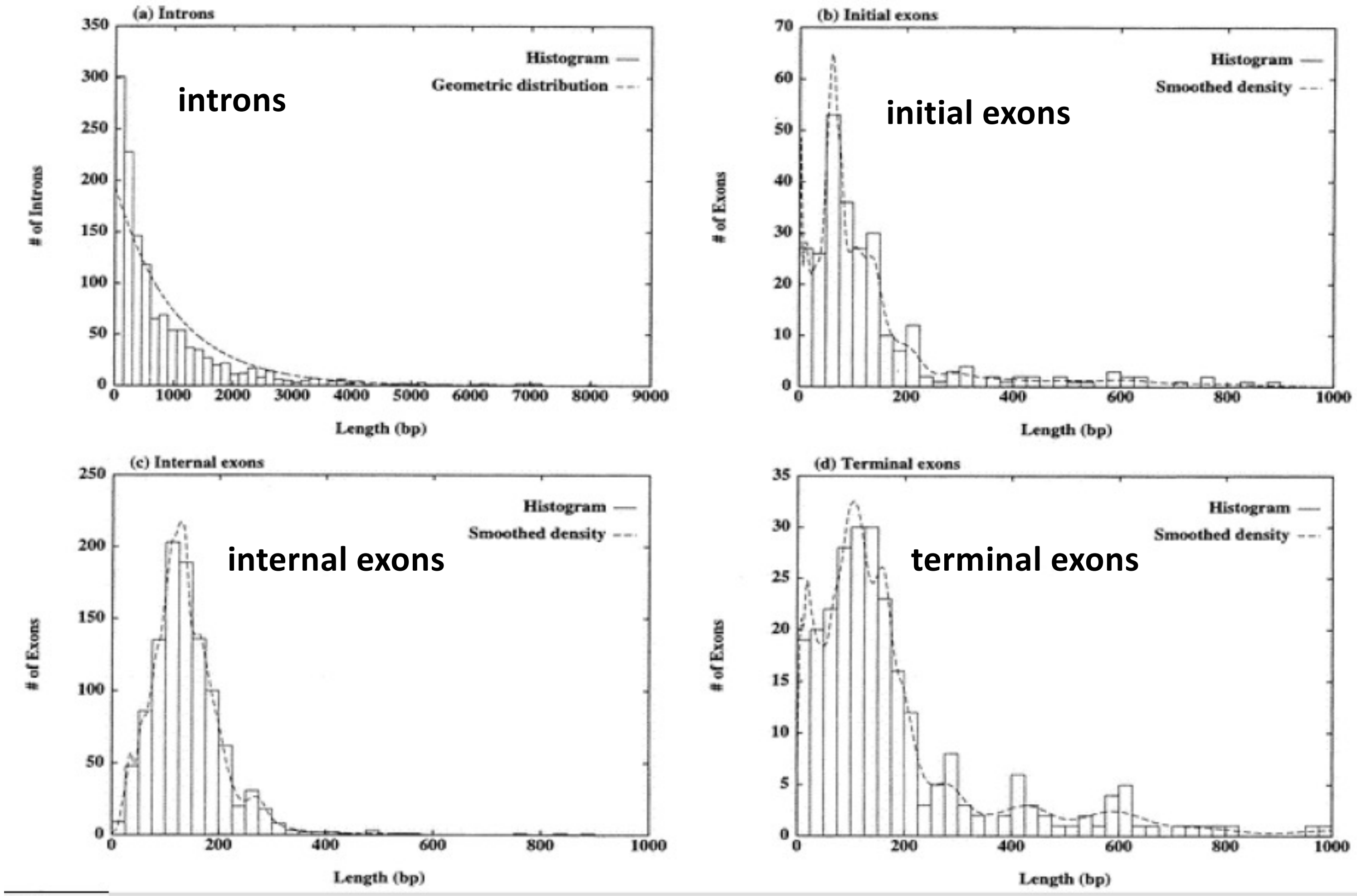
simplified layout of a DNA sequence



Basic idea:

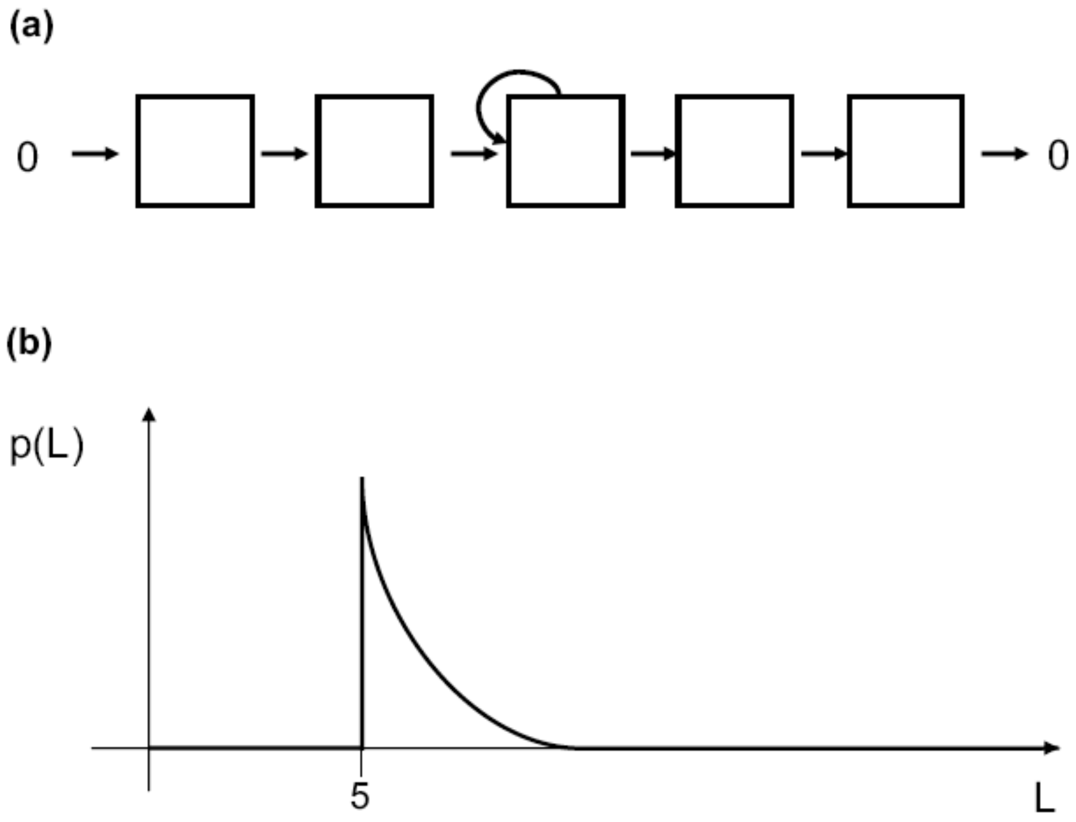
- set up a Markov model for each of these states
- link these models to form a HMM

Enpirical length distributions of different HMM states

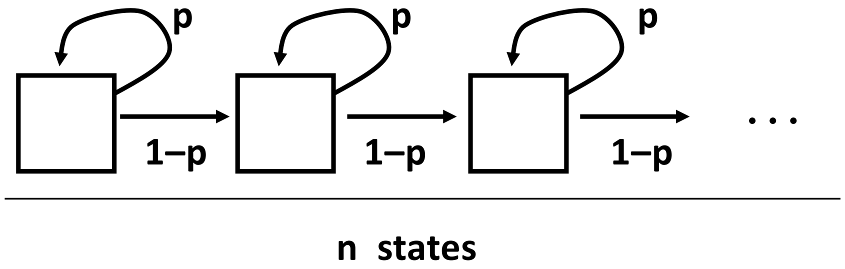


Digression:
length distributions in Hidden Markov models

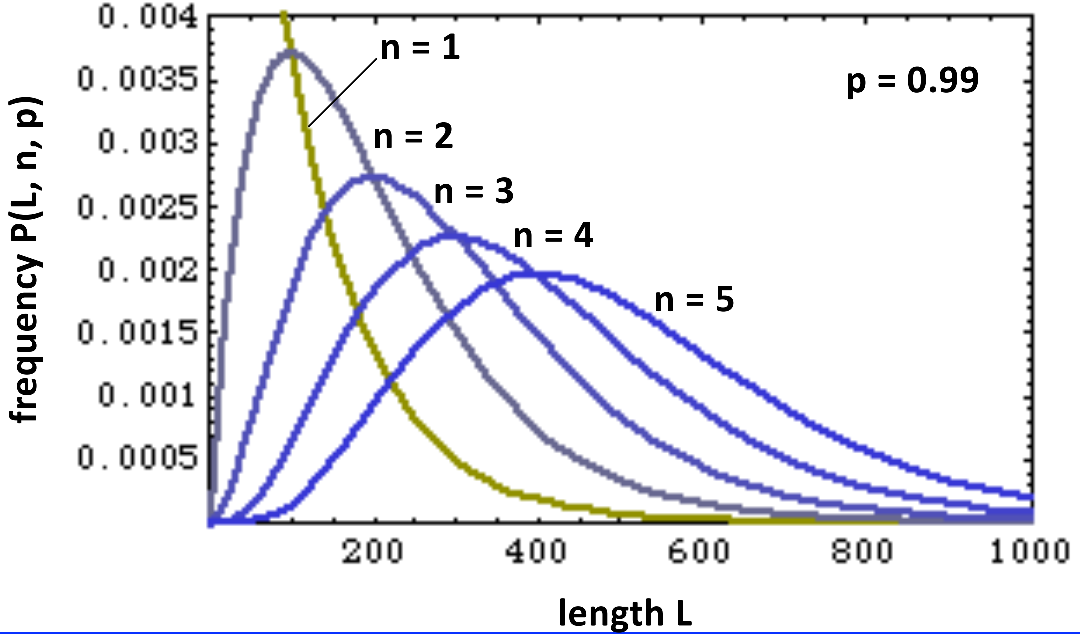
length distributions in Hidden Markov models



length distributions in Hidden Markov models



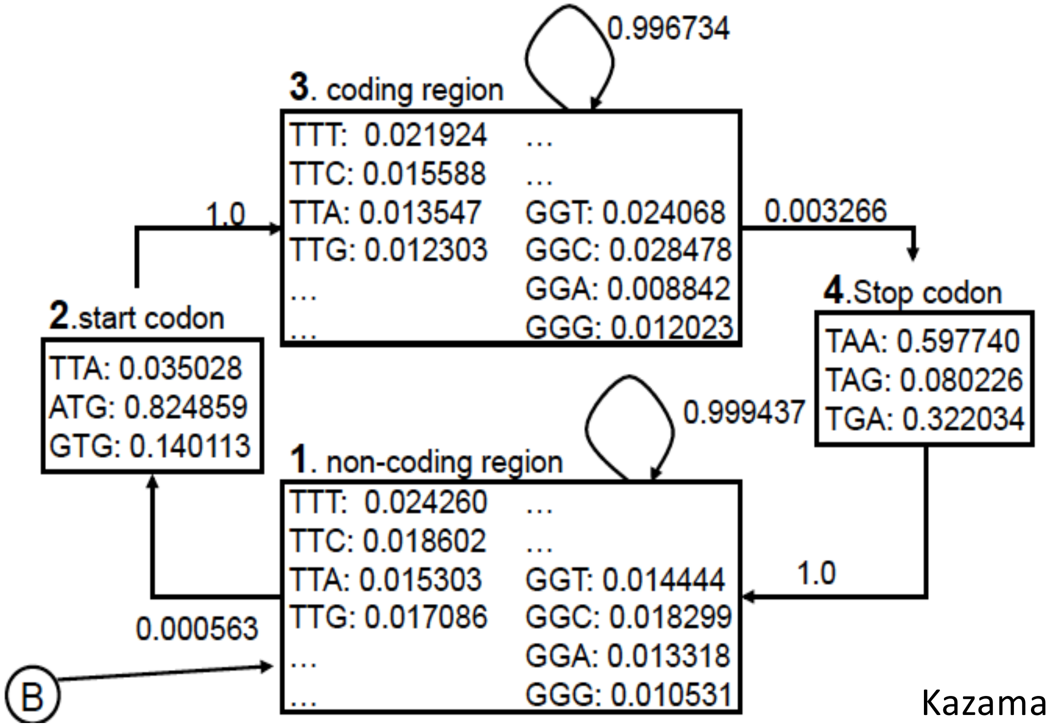
$$P(L) = \binom{L-1}{n-1} p^{L-n} (1-p)^n$$



A simple example: HMM at the codon level

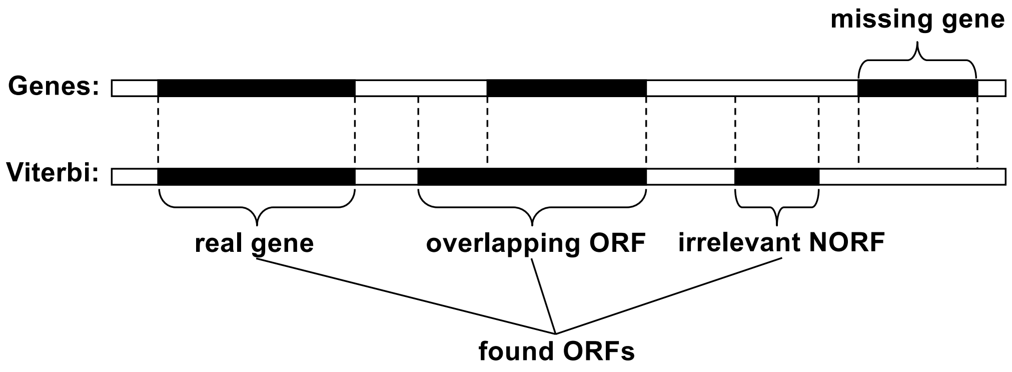
concept:

- 4 HMM states (coding, non-coding, start, stop)
- 64 codons
- different emission probabilities for the codons in the different HMM states



Kazama und Toyoizumi (2003)

A simple example: HMM at the codon level



	Exp. 1
Found ORFs	5760
Real genes	2907
Overlapping ORFs	1119
Irrelevant NORFs	1734
Missing Genes	262
Sn (Sensitivity)	93.88%
Sp (Specificity)	69.89%

Kazama und Toyoizumi (2003)

A less trivial example:

VEIL – Viterbi Exon-Intron Locator (Henderson et al., 2001)

HMM with exons, introns, intergenic regions, splice sites

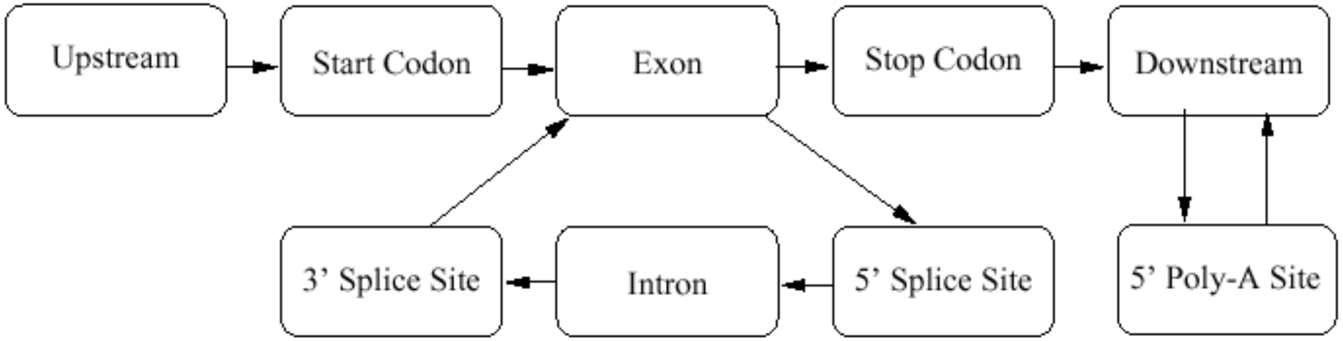


Figure 5: Combined Model Schematic

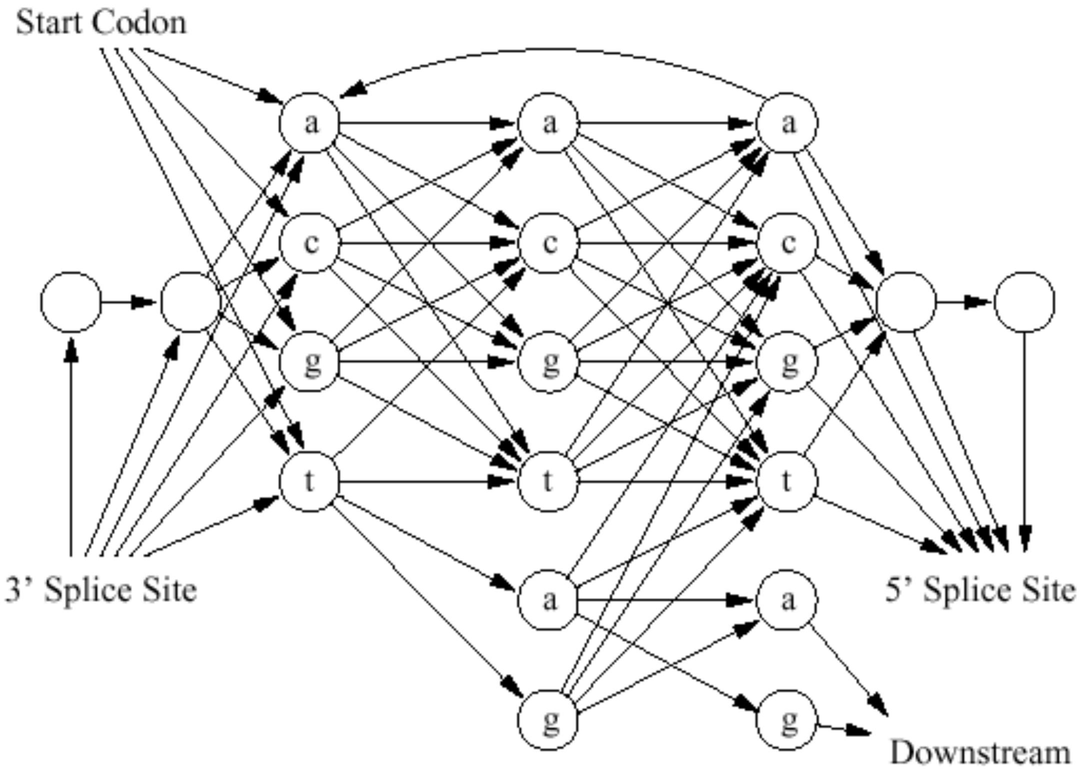
results:

(1) correct identification of both ends at 53 % of the known exons **(sensitivity)**
(2) 49% of all exons predicted by VEIL were correct **(specificity)**

A less trivial example:

VEIL – Viterbi Exon-Intron Locator (Henderson et al., 2001)

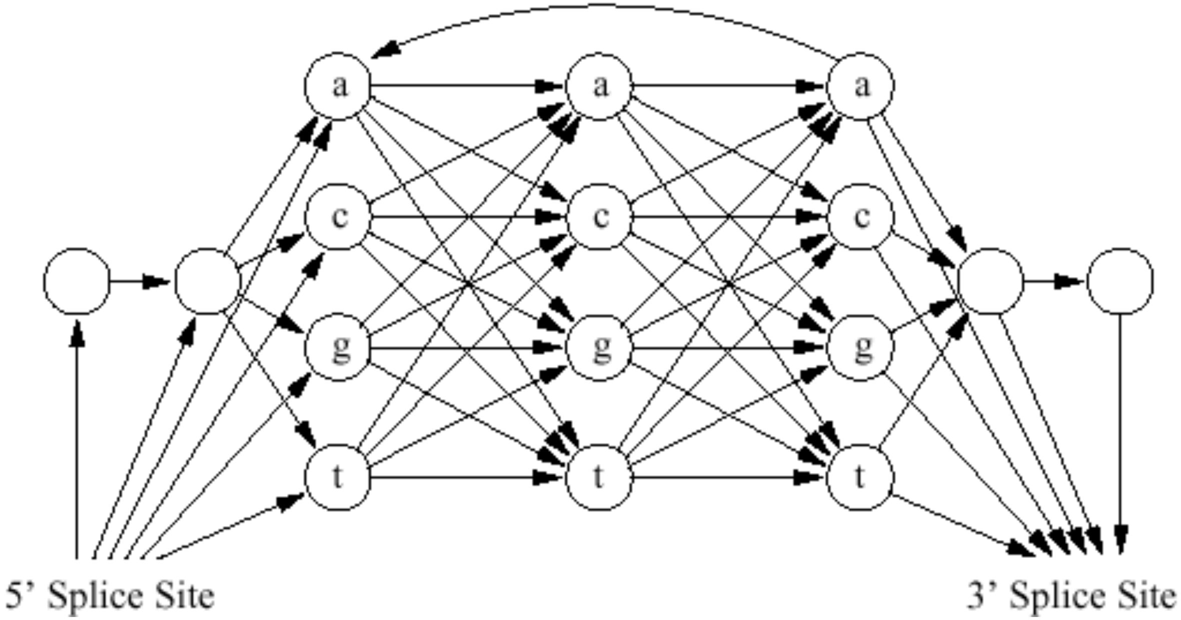
exon component

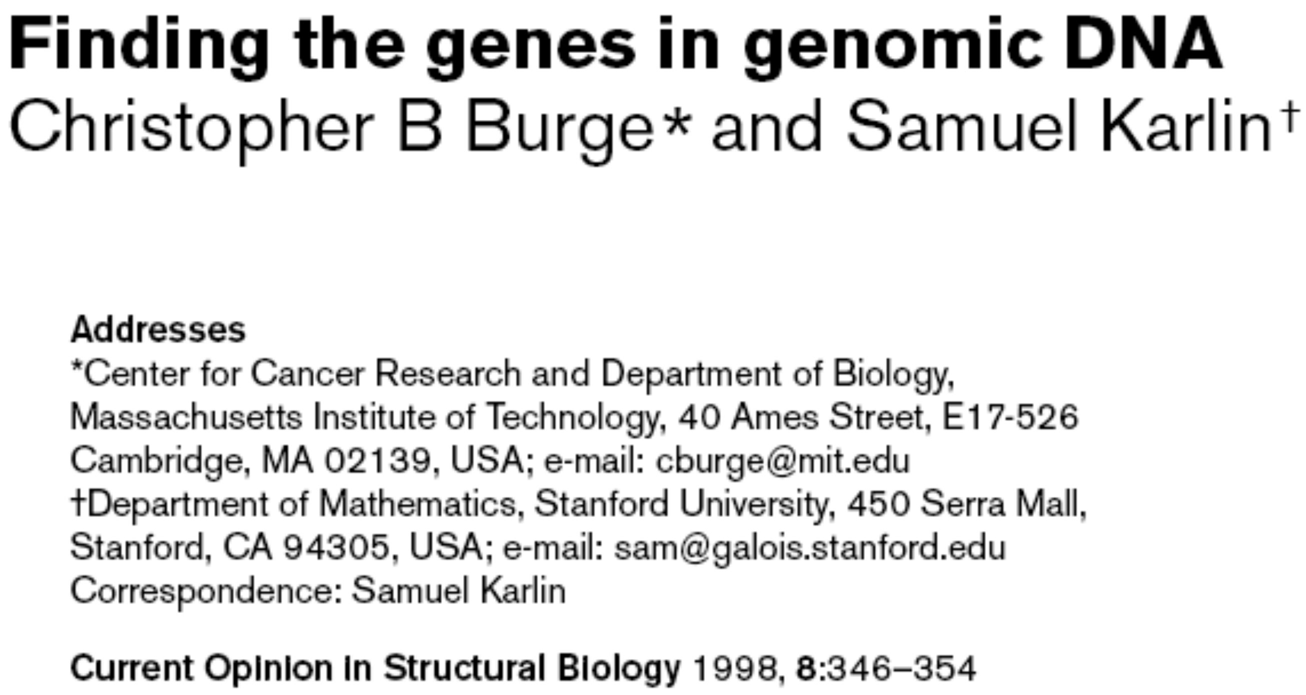


A less trivial example:

VEIL – Viterbi Exon-Intron Locator (Henderson et al., 2001)

intron component:





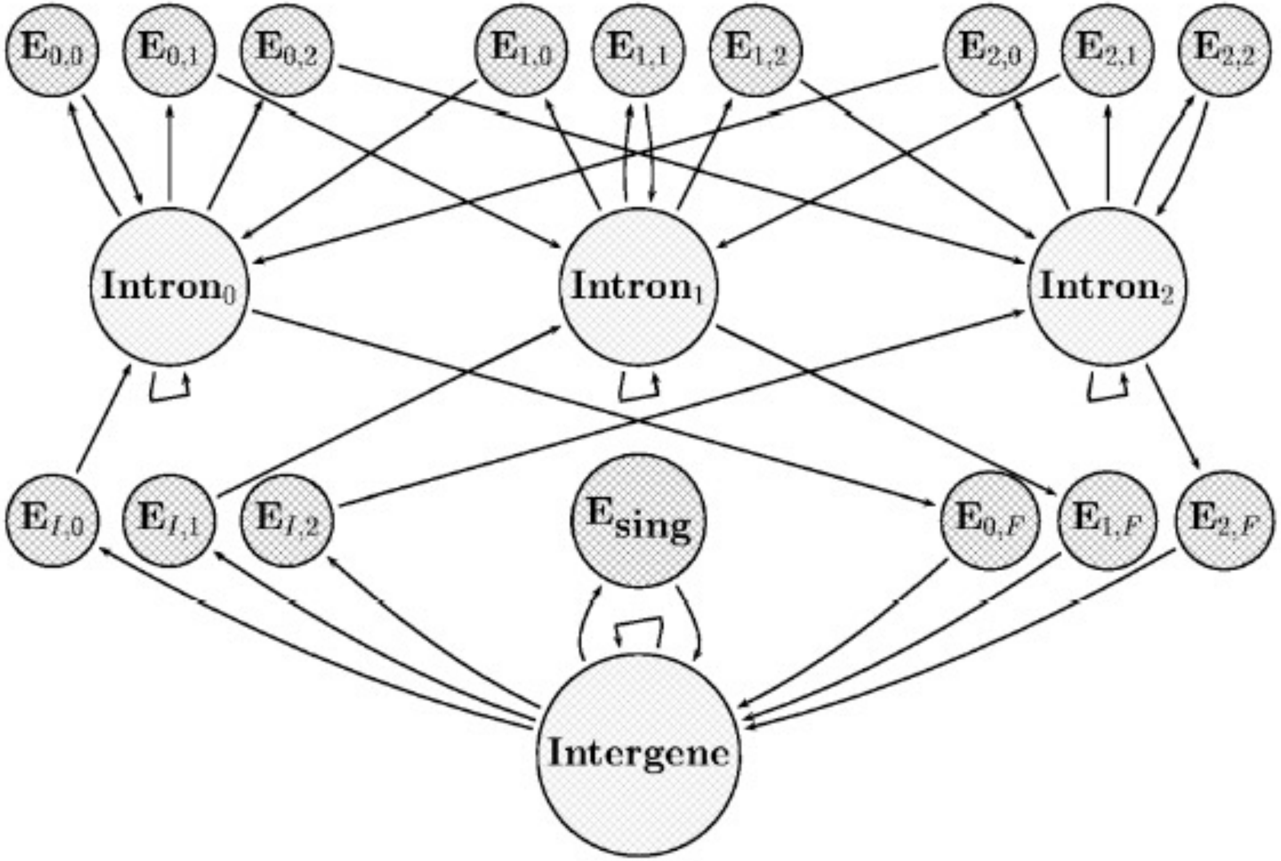
Finding the genes in genomic DNA

Christopher B Burge* and Samuel Karlin†

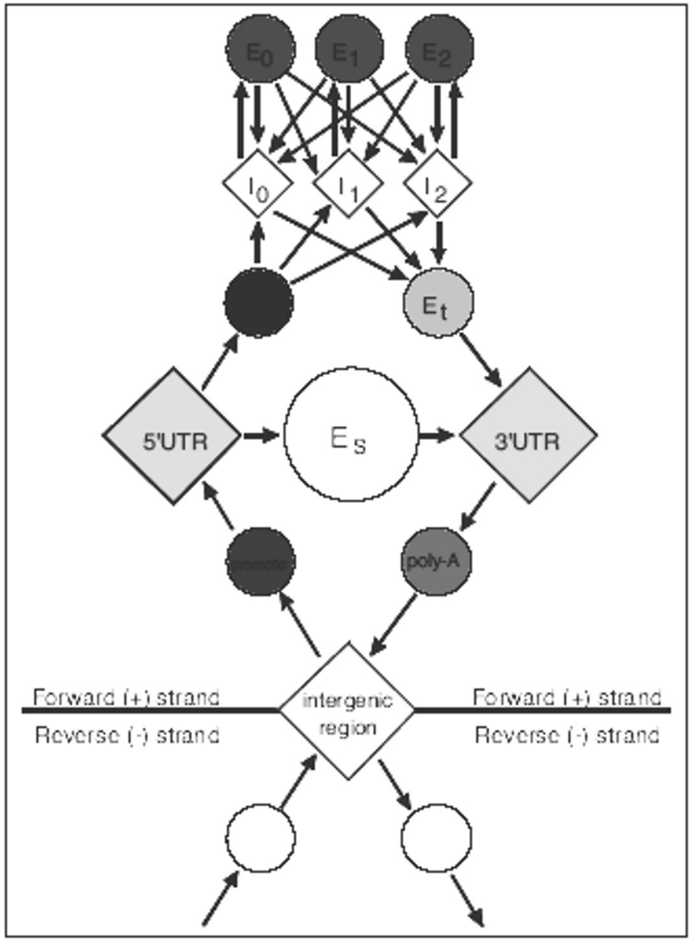
Addresses
*Center for Cancer Research and Department of Biology,
Massachusetts Institute of Technology, 40 Ames Street, E17-526
Cambridge, MA 02139, USA; e-mail: cburge@mit.edu
†Department of Mathematics, Stanford University, 450 Serra Mall,
Stanford, CA 94305, USA; e-mail: sam@galois.stanford.edu
Correspondence: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

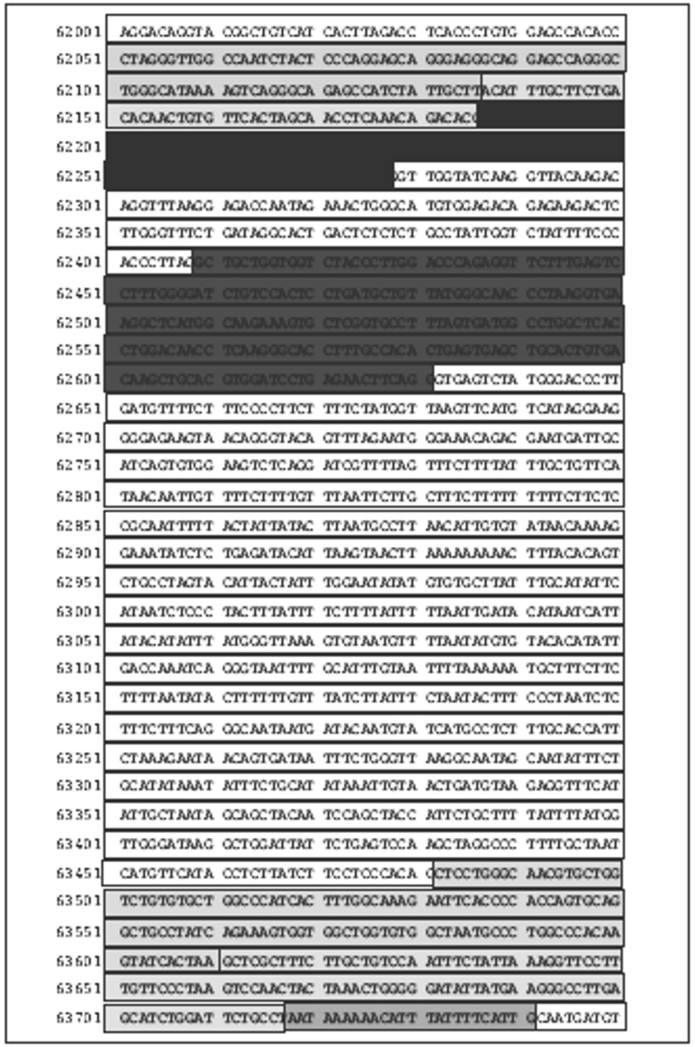
Current Opinion in Structural Biology 1998, **8**:346–354

The GenScan HMM

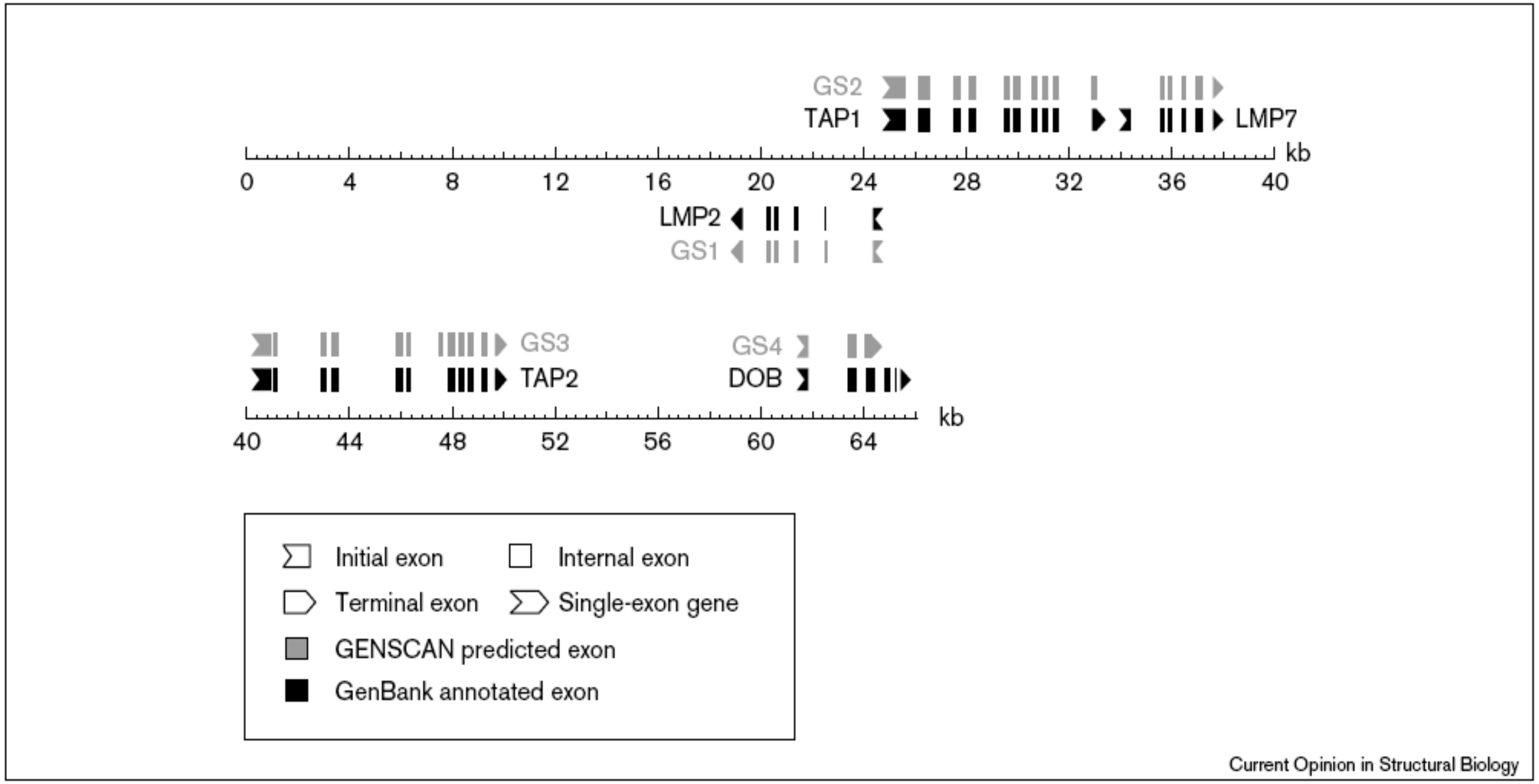


GENSCAN (Burge & Karlin)

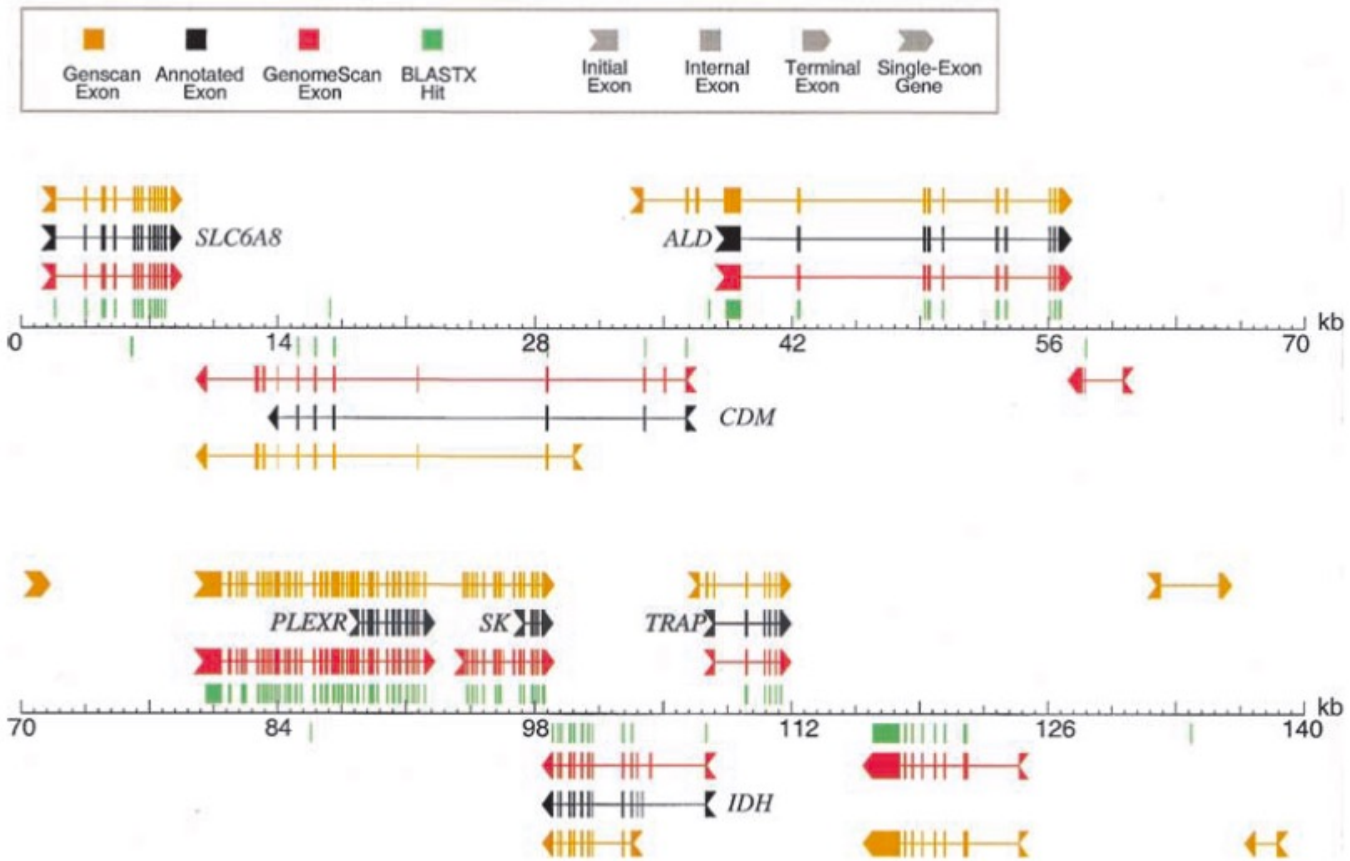




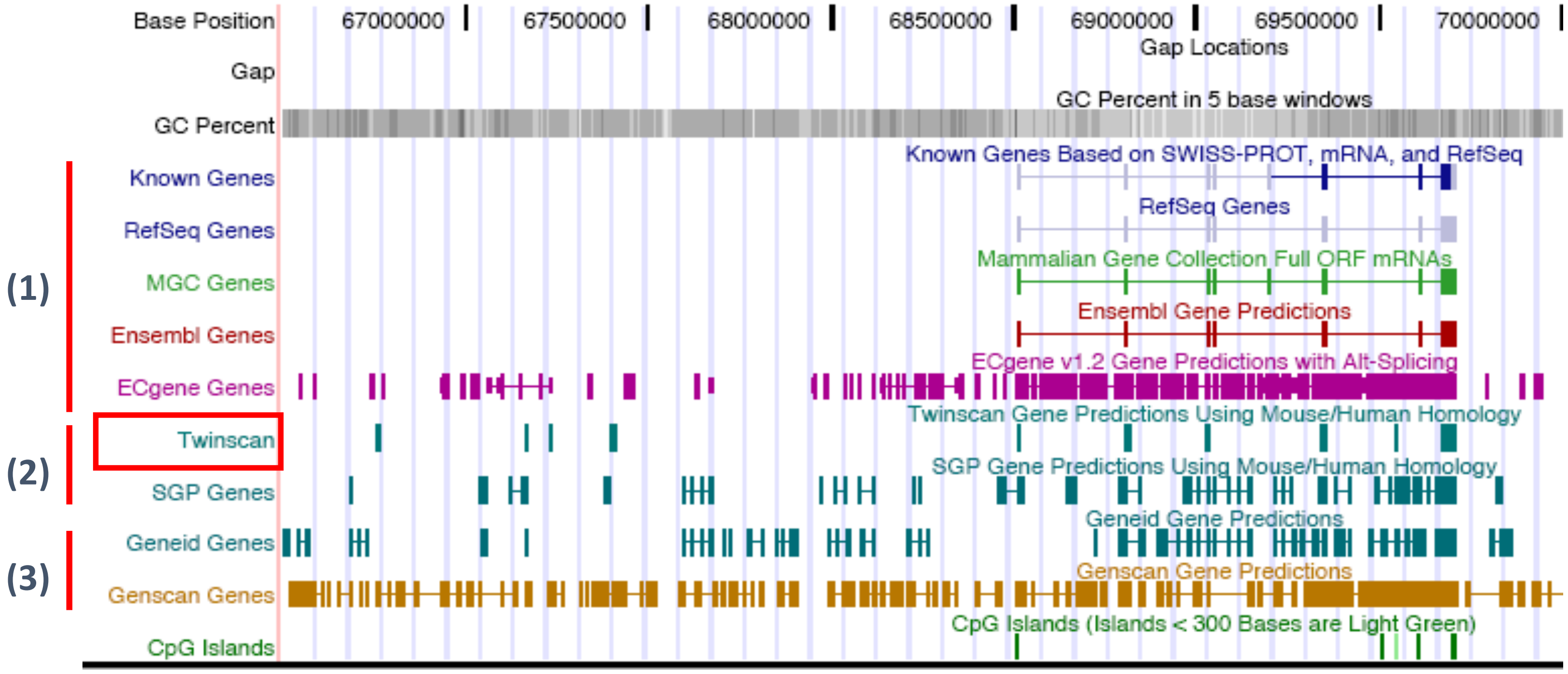
```
62001 AGGACAGGTA CGGCTGTCAT CACTTAGACC TCACCCTGTG GAGCCACACC
62051 CTAGGGTTGG CCAATCTACT CCCAGGAGCA GGGAGGGCAG GAGCCAGGGC
62101 TGGGCATAAA AGTCAGGGCA GAGCCATCTA TTGCTTACAT TTGCTTCTGA
62151 CACAACTGTG TTCACTAGCA ACCTCAAACA GACACC
62201
62251                                GT TGGTATCAAG GTTACAAGAC
62301 AGGTTTAAGG AGACCAATAG AAACTGGGCA TGTGGAGACA GAGAAGACTC
62351 TTGGGTTTCT GATAGGCACT GACTCTCTCT GCCTATTGGT CTATTTTCCC
62401 ACCCTTAGGC TGCTGGTGGT CTACCCTTGG ACCCAGAGGT TCTTTGAGTC
62451 CTTTGGGGAT CTGTCCACTC CTGATGCTGT TATGGGCAAC CCTAAGGTGA
62501 AGGCTCATGG CAAGAAAGTG CTCGGTGCCT TTAGTGATGG CCTGGCTCAC
62551 CTGGACAACC TCAAGGGCAC CTTTGCCACA CTGAGTGAGC TGCACTGTGA
62601 CAAGCTGCAC GTGGATCCTG AGAACTTCAG GGTGAGTCTA TGGGACCCTT
62651 GATGTTTTCT TTCCCCTTCT TTTCTATGGT TAAGTTCATG TCATAGGAAG
62701 GGGAGAAGTA ACAGGGTACA GTTTAGAATG GGAAACAGAC GAATGATTGC
62751 ATCAGTGTGG AAGTCTCAGG ATCGTTTTAG TTTCTTTTAT TTGCTGTTCA
62801 TAACAATTGT TTTCTTTTGT TTAATTCTTG CTTTCTTTTT TTTTCTTCTC
62851 CGCAATTTTT ACTATTATAC TTAATGCCTT AACATTGTGT ATAACAAAAG
62901 GAAATATCTC TGAGATACAT TAAGTAACTT AAAAAAAAAC TTTACACAGT
62951 CTGCCTAGTA CATTACTATT TGGAATATAT GTGTGCTTAT TTGCATATTC
63001 ATAATCTCCC TACTTTATTT TCTTTTATTT TTAATTGATA CATAATCATT
63051 ATACATATTT ATGGGTTAAA GTGTAATGTT TTAATATGTG TACACATATT
63101 GACCAAATCA GGGTAATTTT GCATTTGTAA TTTTAAAAAA TGCTTTCTTC
63151 TTTTAATATA CTTTTTTGTT TATCTTATTT CTAATACTTT CCCTAATCTC
63201 TTTCTTTCAG GGCAATAATG ATACAATGTA TCATGCCTCT TTGCACCATT
63251 CTAAAGAATA ACAGTGATAA TTTCTGGGTT AAGGCAATAG CAATATTTCT
63301 GCATATAAAT ATTTCTGCAT ATAAATTGTA ACTGATGTAA GAGGTTTCAT
63351 ATTGCTAATA GCAGCTACAA TCCAGCTACC ATTCTGCTTT TATTTTATGG
63401 TTGGGATAAG GCTGGATTAT TCTGAGTCCA AGCTAGGCCC TTTTGCTAAT
63451 CATGTTCATA CCTCTTATCT TCCTCCCACA GCTCCTGGGC AACGTGCTGG
63501 TCTGTGTGCT GGCCCATCAC TTTGGCAAAG AATTCACCCC ACCAGTGCAG
63551 GCTGCCTATC AGAAAGTGGT GGCTGGTGTG GCTAATGCCC TGGCCCACAA
63601 GTATCACTAA GCTCGCTTTC TTGCTGTCCA ATTTCTATTA AAGGTTCCTT
63651 TGTTCCCTAA GTCCAACTAC TAAACTGGGG GATATTATGA AGGGCCTTGA
63701 GCATCTGGAT TCTGCCTAAT AAAAAACATT TATTTTCATT GCAATGATGT
```



Current Opinion in Structural Biology



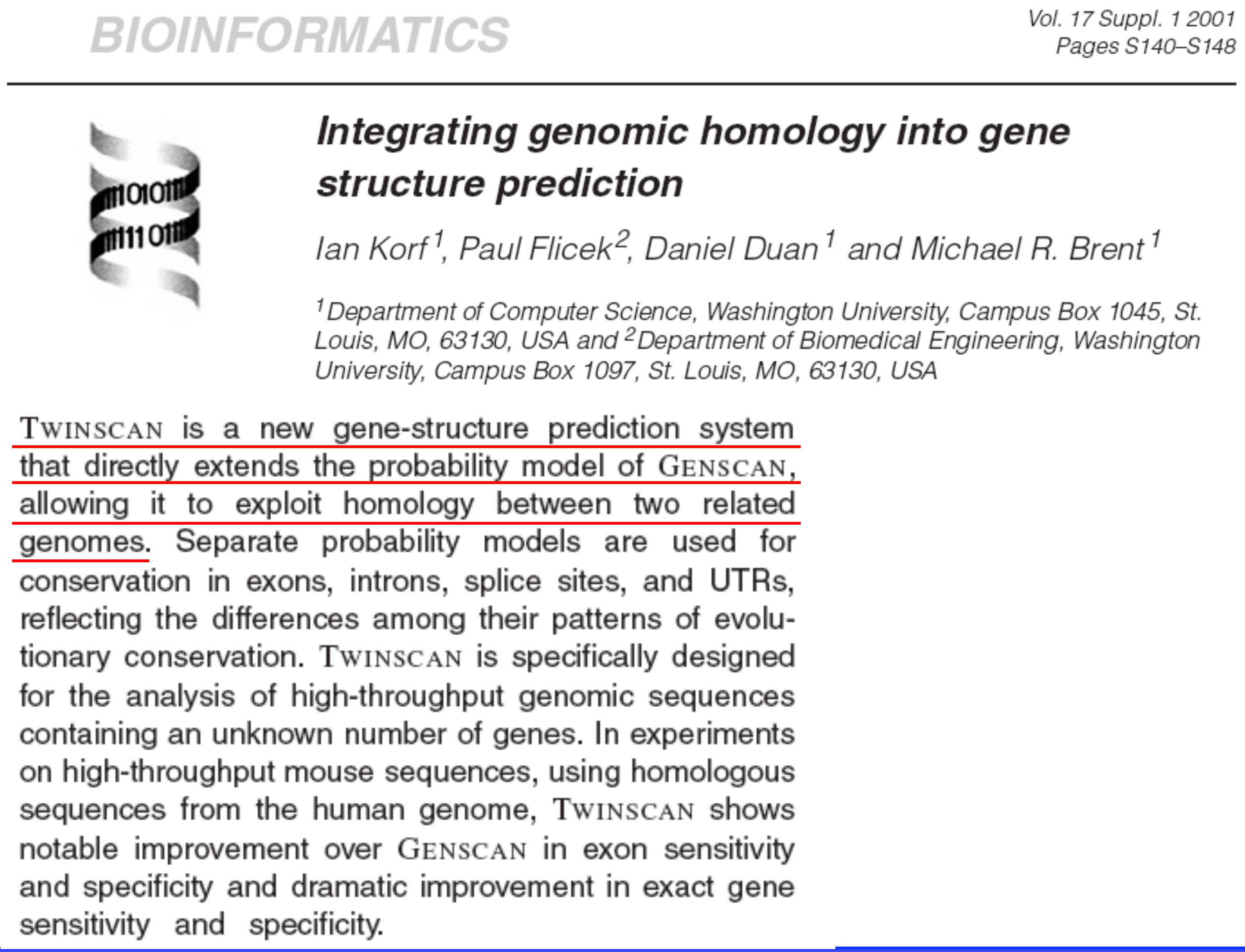
Yeh, Lim, and Burge, Genome Research 11:803-816, 2001.


Base Position
67000000
67500000
68000000
68500000
69000000
69500000
70000000
Gap Locations
Gap
GC Percent in 5 base windows
GC Percent
Known Genes Based on SWISS-PROT, mRNA, and RefSeq
Known Genes
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
MGC Genes
Ensembl Gene Predictions
Ensembl Genes
ECgene v1.2 Gene Predictions with Alt-Splicing
ECgene Genes
Twinscan Gene Predictions Using Mouse/Human Homology
Twinscan
SGP Gene Predictions Using Mouse/Human Homology
SGP Genes
Geneid Gene Predictions
Geneid Genes
Genscan Gene Predictions
Genscan Genes
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
(1)
(2)
(3)

(1) annotated genes and alignment-based gene models

(2) hybrid models

(3) ”ab initio” models



BIOINFORMATICS

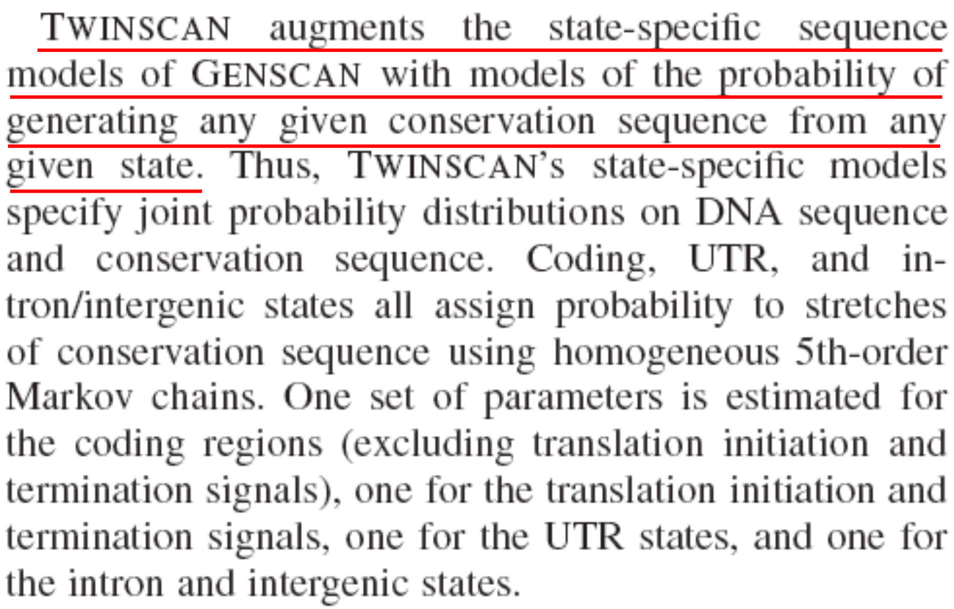
Vol. 17 Suppl. 1 2001
Pages S140–S148

Integrating genomic homology into gene structure prediction

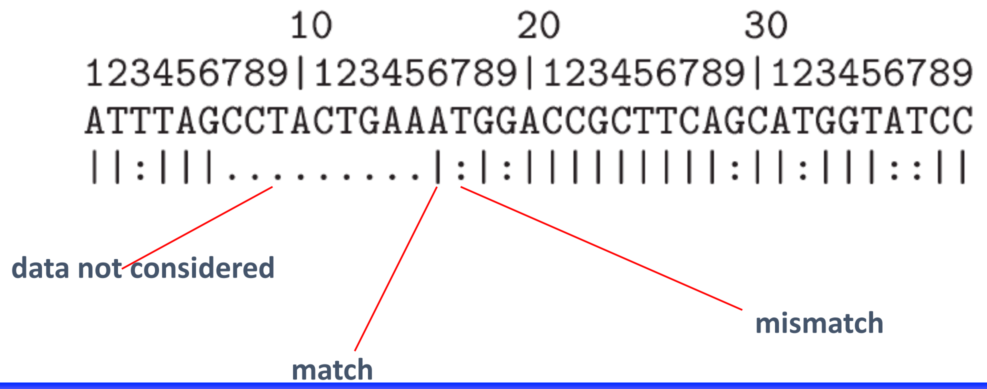
Ian Korf[1], Paul Flicek[2], Daniel Duan[1] and Michael R. Brent[1]

[1] Department of Computer Science, Washington University, Campus Box 1045, St. Louis, MO, 63130, USA and [2] Department of Biomedical Engineering, Washington University, Campus Box 1097, St. Louis, MO, 63130, USA

Twinscan is a new gene-structure prediction system that directly extends the probability model of Genscan, allowing it to exploit homology between two related genomes. Separate probability models are used for conservation in exons, introns, splice sites, and UTRs, reflecting the differences among their patterns of evolutionary conservation. Twinscan is specifically designed for the analysis of high-throughput genomic sequences containing an unknown number of genes. In experiments on high-throughput mouse sequences, using homologous sequences from the human genome, Twinscan shows notable improvement over Genscan in exon sensitivity and specificity and dramatic improvement in exact gene sensitivity and specificity.

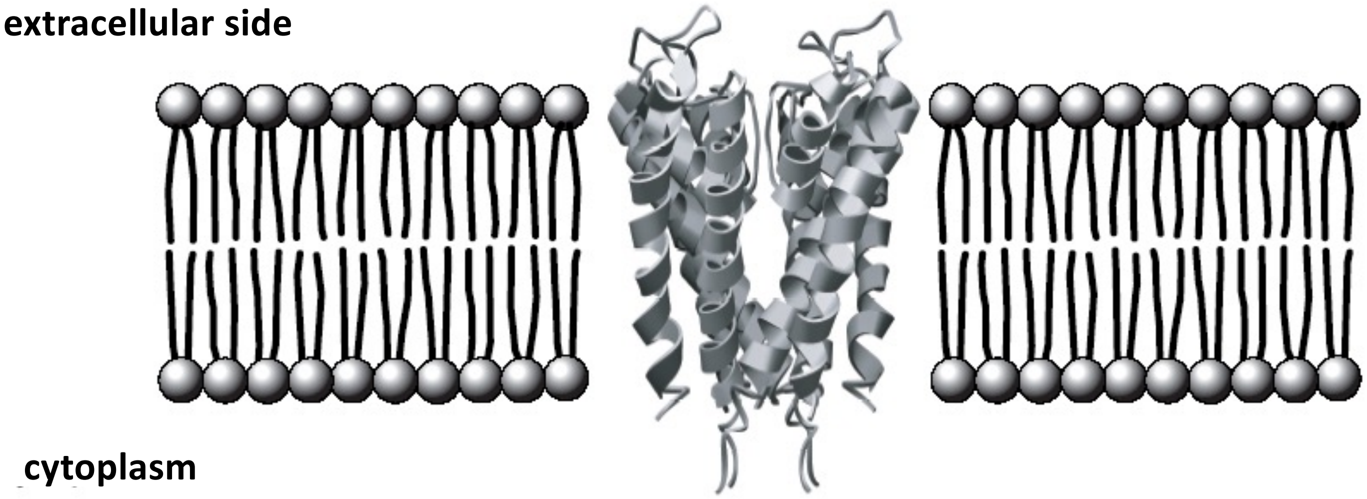


TWINSCAN augments the state-specific sequence models of GENSCAN with models of the probability of generating any given conservation sequence from any given state. Thus, TWINSCAN's state-specific models specify joint probability distributions on DNA sequence and conservation sequence. Coding, UTR, and intron/intergenic states all assign probability to stretches of conservation sequence using homogeneous 5th-order Markov chains. One set of parameters is estimated for the coding regions (excluding translation initiation and termination signals), one for the translation initiation and termination signals, one for the UTR states, and one for the intron and intergenic states.



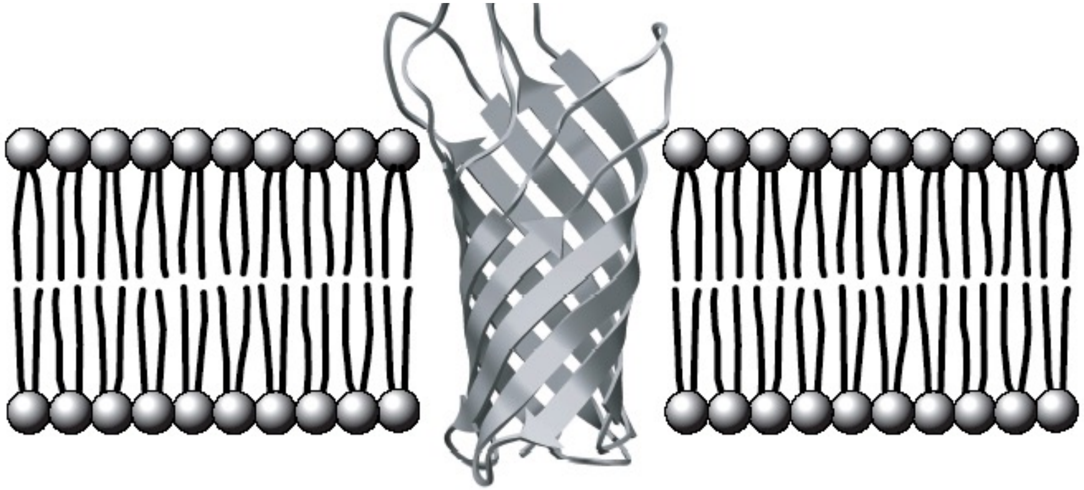
```
         10        20        30
123456789|123456789|123456789|123456789
ATTTAGCCTACTGAAATGGACCGCTTCAGCATGGTATCC
||:|||.........|:|:|||||||||:||:|||::||
```

Application 2:
membrane proteins

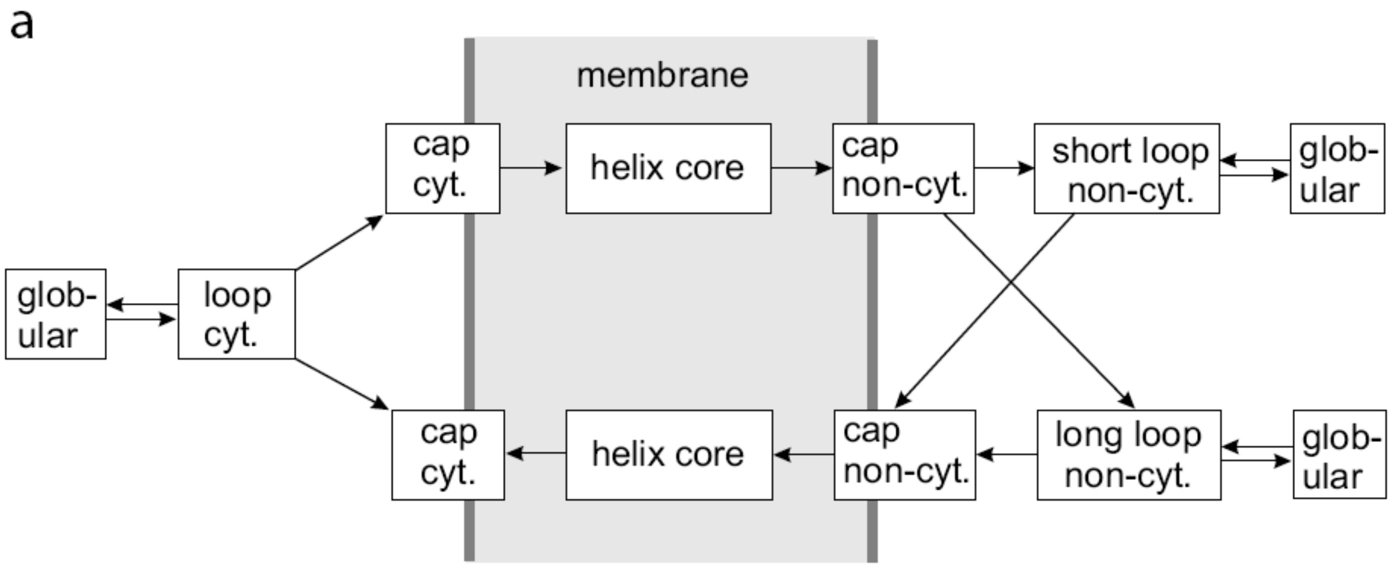


extracellular side

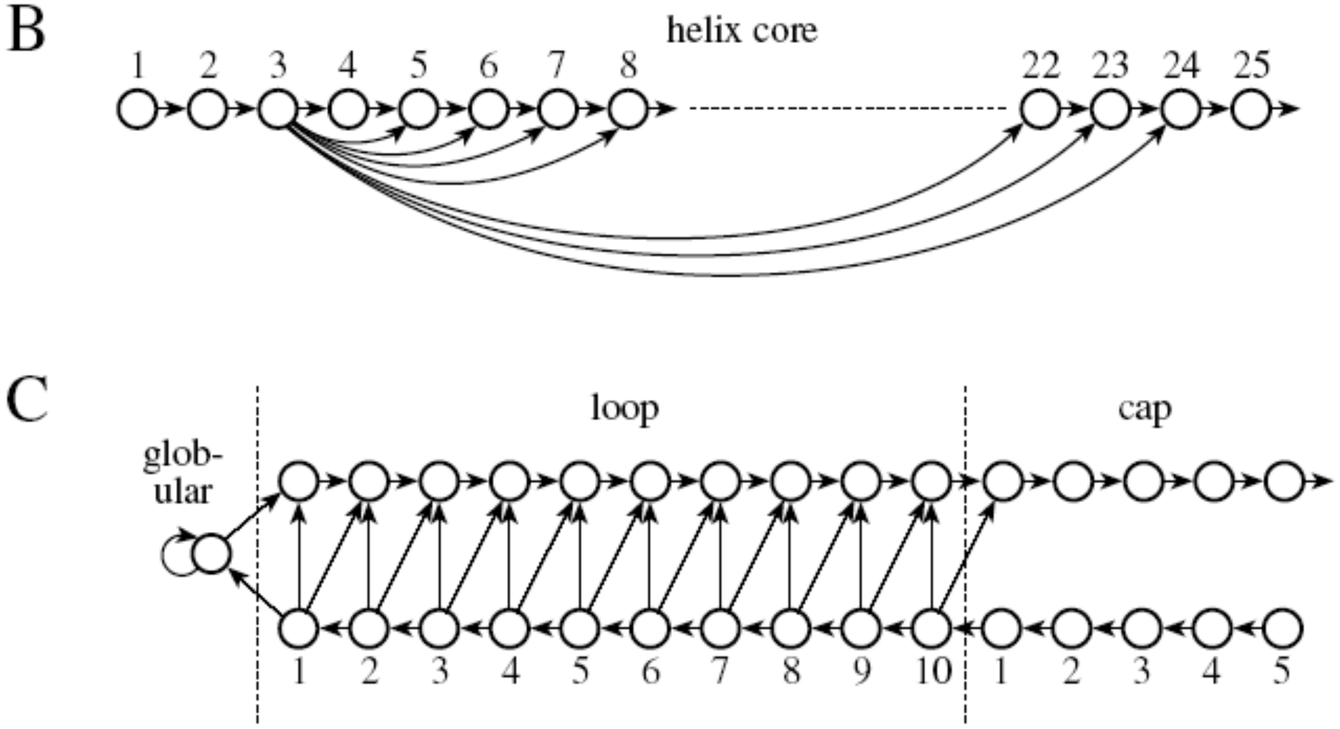
cytoplasm



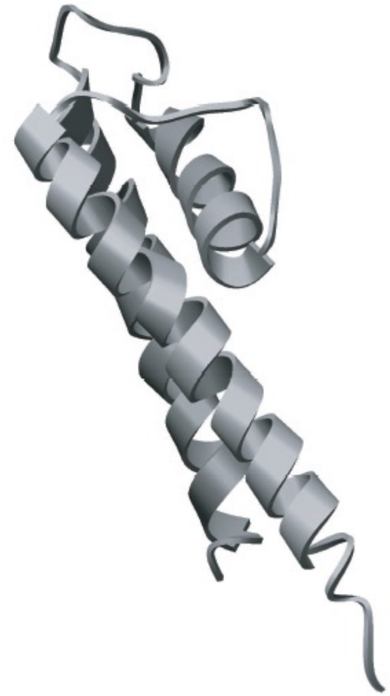
Transmembrane Hidden Markov Model (TMHMM)
(Krogh et al. 2001)



Transmembrane Hidden Markov Model (TMHMM)
(Krogh et al. 2001)

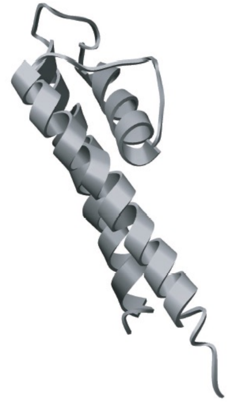


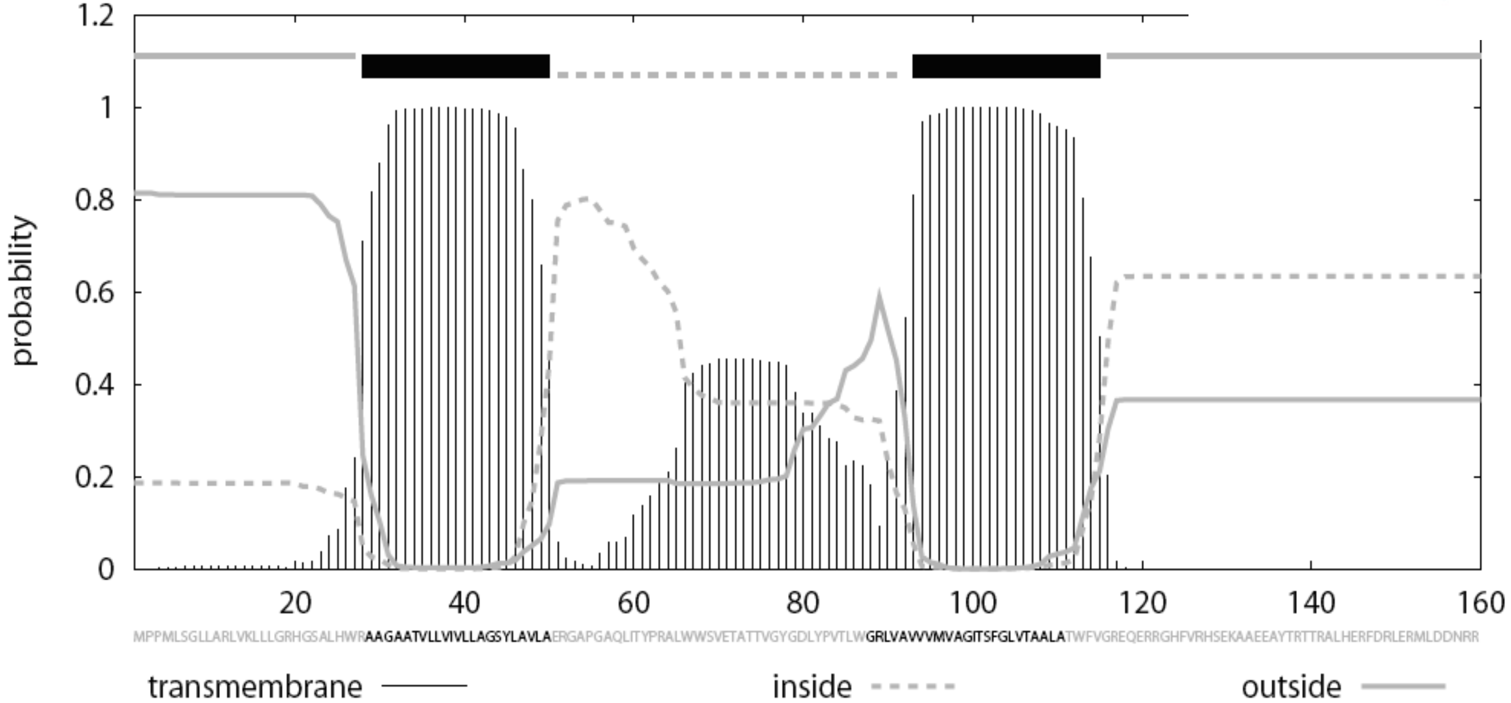
Transmembrane Hidden Markov Model (TMHMM)
(Krogh et al. 2001)



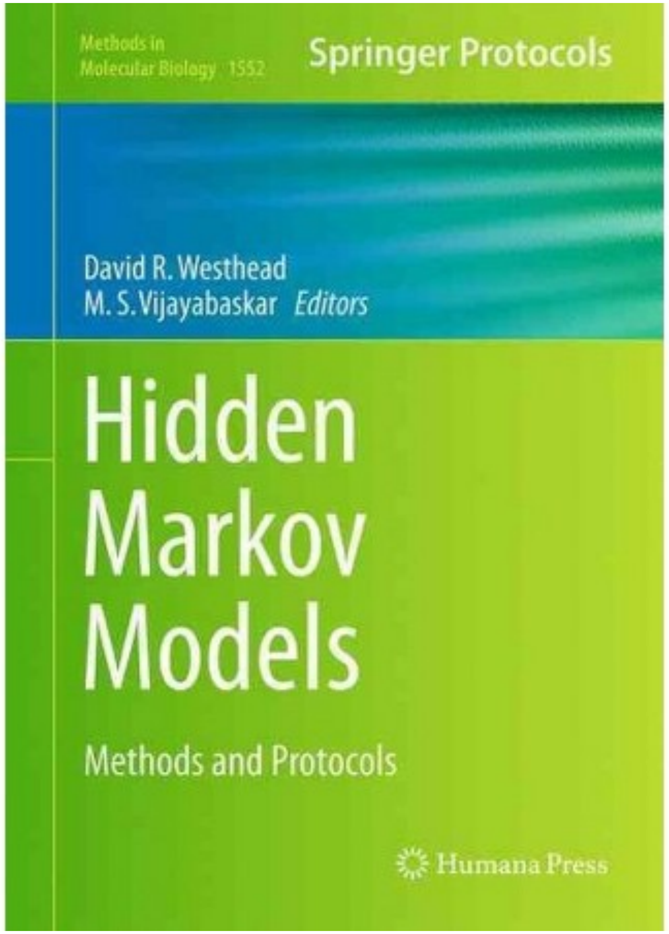
Known structure of one domain
of the **KirBac potassium channel**;
one sees the inner helix, the outer
helix and (the small helical segment)
the pore helix

Transmembrane Hidden Markov Model (TMHMM)
(Krogh et al. 2001)




TMHMM posterior probabilities for sp_P0A334_KCSA_STRI
probability
1.2
1
0.8
0.6
0.4
0.2
0
20
40
60
80
100
120
140
160
MPPMLSGLLARLVKLLLGRHGSALHWRAAGAATVLLVIVLLAGSYLAVLAERGAPGAQLITYPRALWWSVETATTVGYGDLYPVTLWGRLVAVVVMVAGITSFGLVTAALATWFVGREQERRGHFVRHSEKAAEEAYTRTTRALHERFDRLERMLDDNRR
transmembrane
inside
outside

Current status of HMMs in Bioinformatics

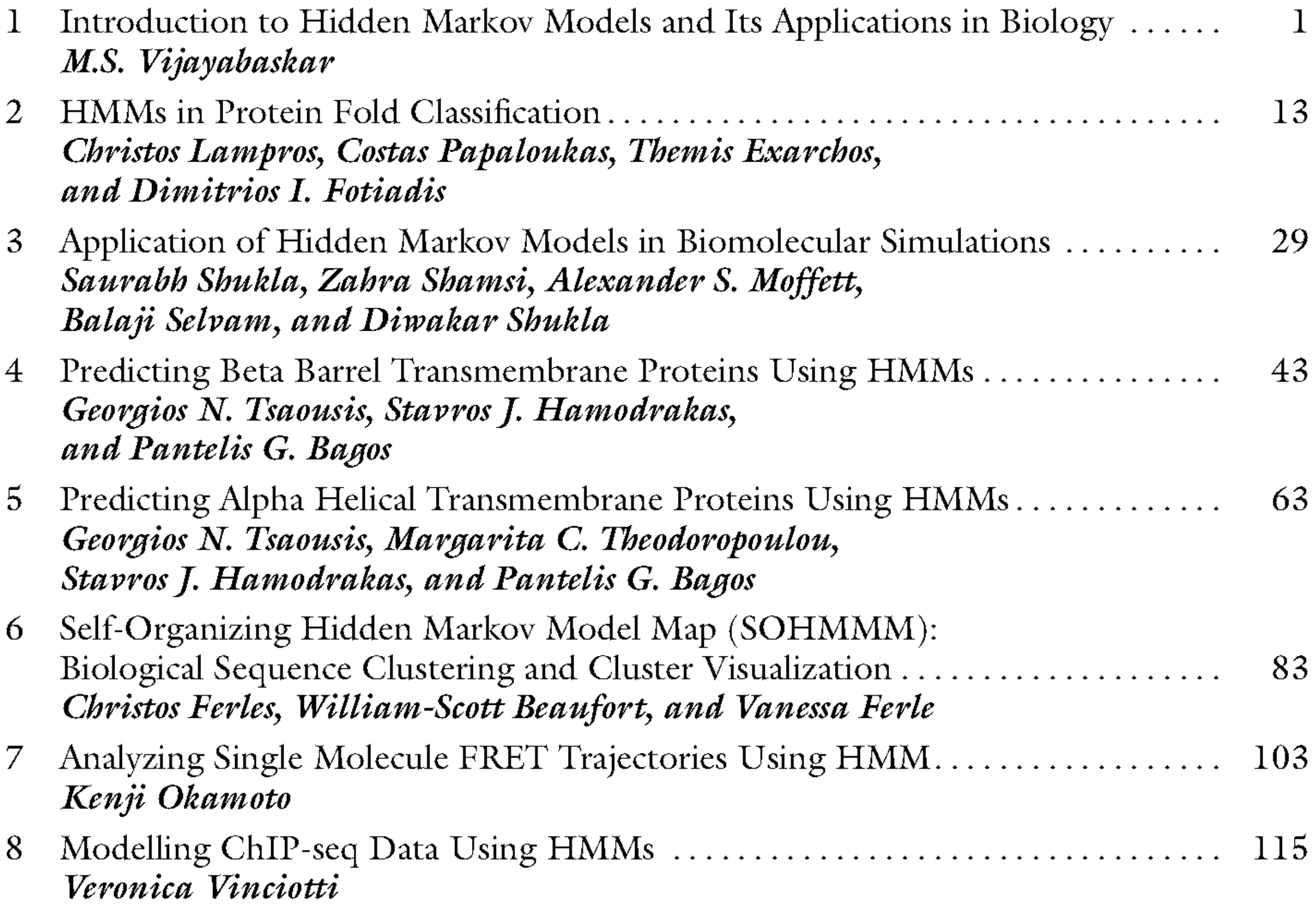


"With the increasing influence of computer-based algorithms and statistics in biology, we have successfully come up with methods for modeling such complex and noisy biological systems. Hidden Markov Model (HMM) is one such statistical model widely used in modeling complex systems and in the identification of “hidden” patterns.

The beauty of HMM is that it is simple to understand and easy to apply to real-world scenarios."

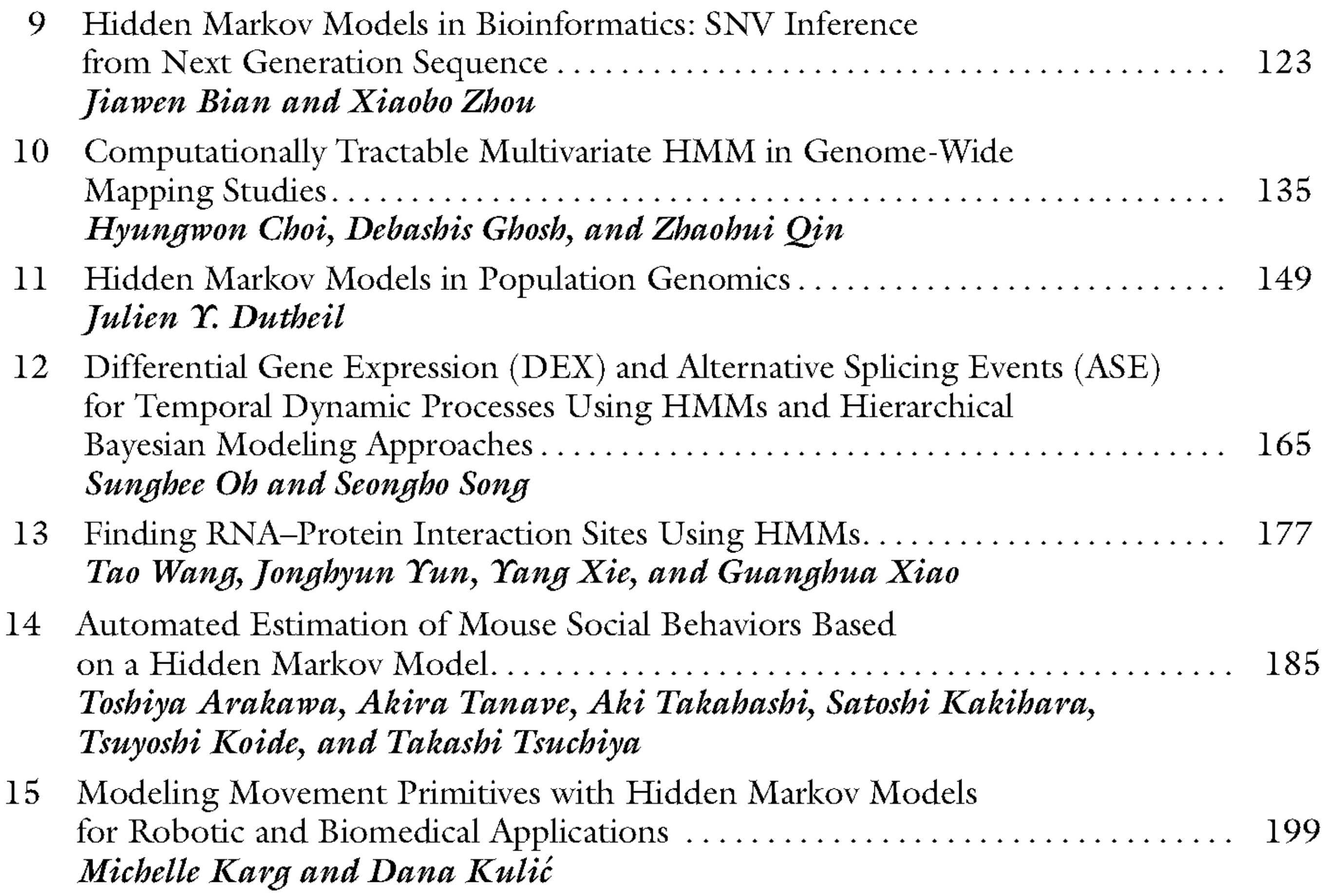
Westhead, Vijayabaskar (Eds.) Hidden Markov Models, Springer Protocols 2017

Current status of HMMs in Bioinformatics



Westhead, Vijayabaskar (Eds.) Hidden Markov Models, Springer Protocols 2017

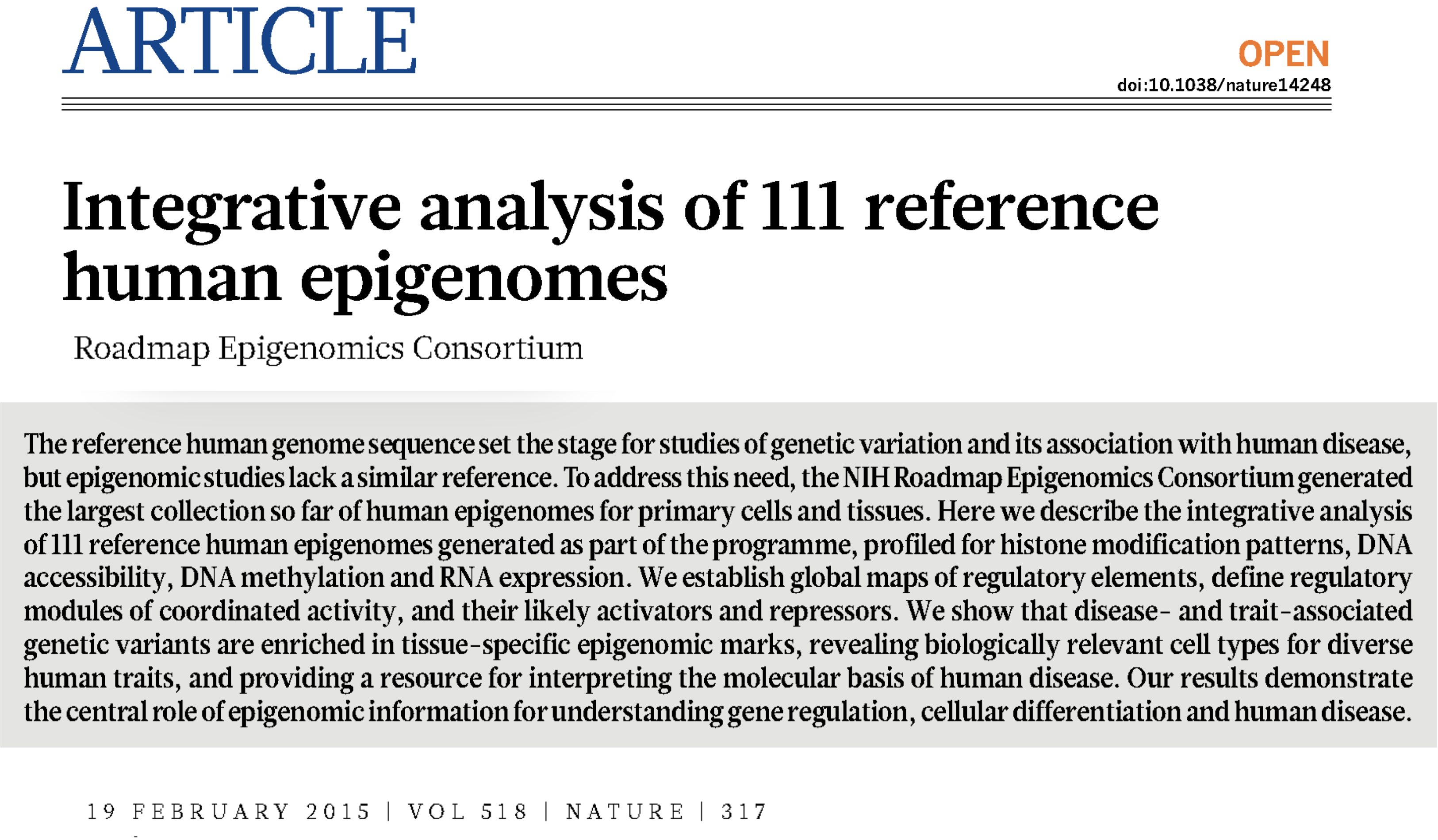
Current status of HMMs in Bioinformatics



Westhead, Vijayabaskar (Eds.) Hidden Markov Models, Springer Protocols 2017

Current status of HMMs in Bioinformatics

Digression: diversity of bioinformatics methods in this publication



ARTICLE

OPEN

doi:10.1038/nature14248

Integrative analysis of 111 reference human epigenomes

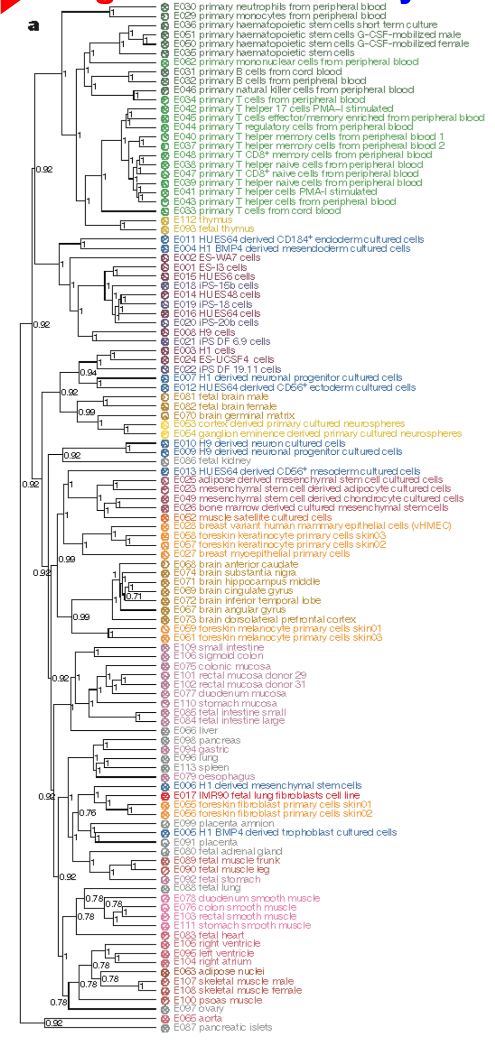
Roadmap Epigenomics Consortium

The reference human genome sequence set the stage for studies of genetic variation and its association with human disease, but epigenomic studies lack a similar reference. To address this need, the NIH Roadmap Epigenomics Consortium generated the largest collection so far of human epigenomes for primary cells and tissues. Here we describe the integrative analysis of 111 reference human epigenomes generated as part of the programme, profiled for histone modification patterns, DNA accessibility, DNA methylation and RNA expression. We establish global maps of regulatory elements, define regulatory modules of coordinated activity, and their likely activators and repressors. We show that disease- and trait-associated genetic variants are enriched in tissue-specific epigenomic marks, revealing biologically relevant cell types for diverse human traits, and providing a resource for interpreting the molecular basis of human disease. Our results demonstrate the central role of epigenomic information for understanding gene regulation, cellular differentiation and human disease.

19 FEBRUARY 2015 | VOL 518 | NATURE | 317

Current status of HMMs in Bioinformatics

Digression: diversity of bioinformatics methods in this publication



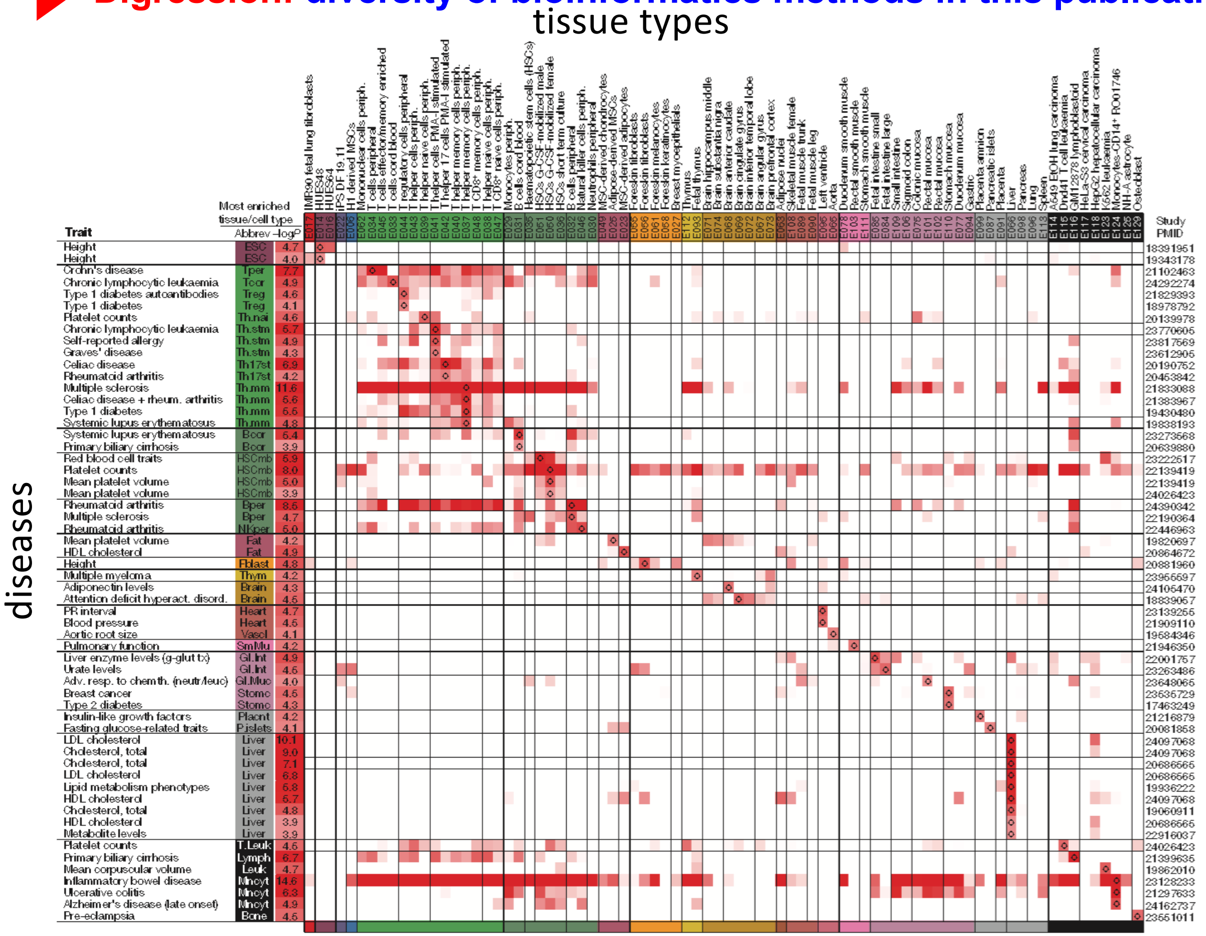
Taken from: Roadmap Epigenomics Consortium (2015) Nature 518, 317

Hierarchical clustering of tissue types according to some epigenetic marker

Current status of HMMs in Bioinformatics

Digression: diversity of bioinformatics methods in this publication

Taken from: Roadmap Epigenomics Consortium (2015) Nature 518, 317

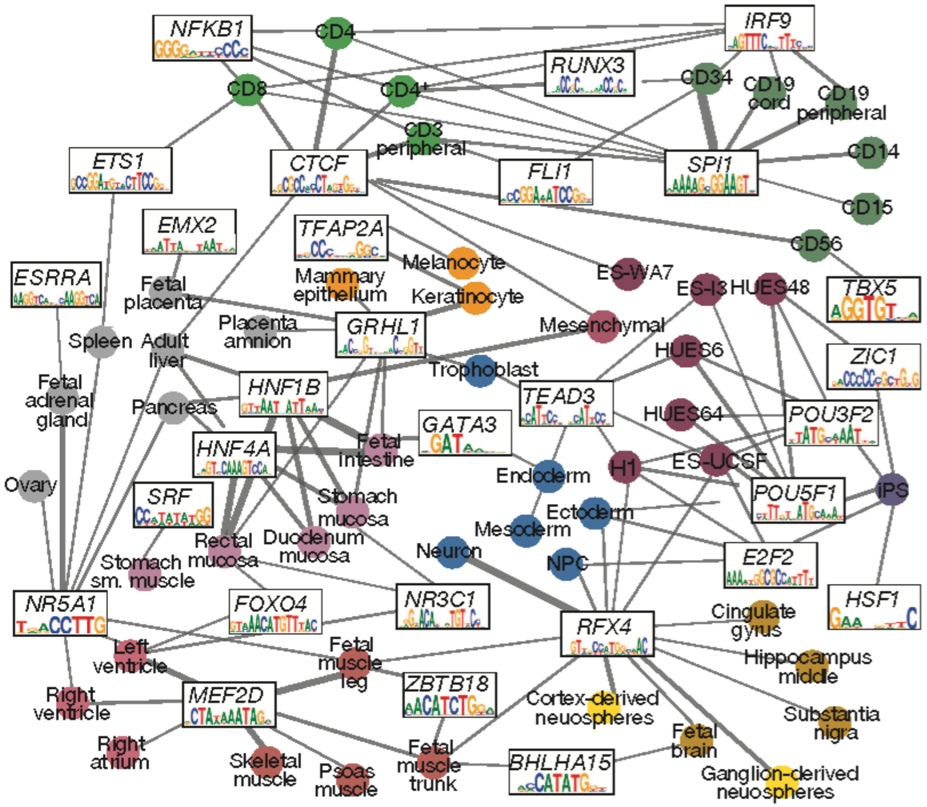


statistical enrichment of epigenetic markers in disease-associated genetic variants

Current status of HMMs in Bioinformatics

Digression: diversity of bioinformatics methods in this publication

Taken from: Roadmap Epigenomics Consortium (2015) Nature 518, 317

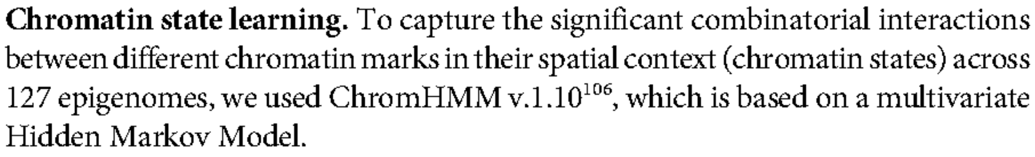


networks of associations of regulators with tissues

Current status of HMMs in Bioinformatics

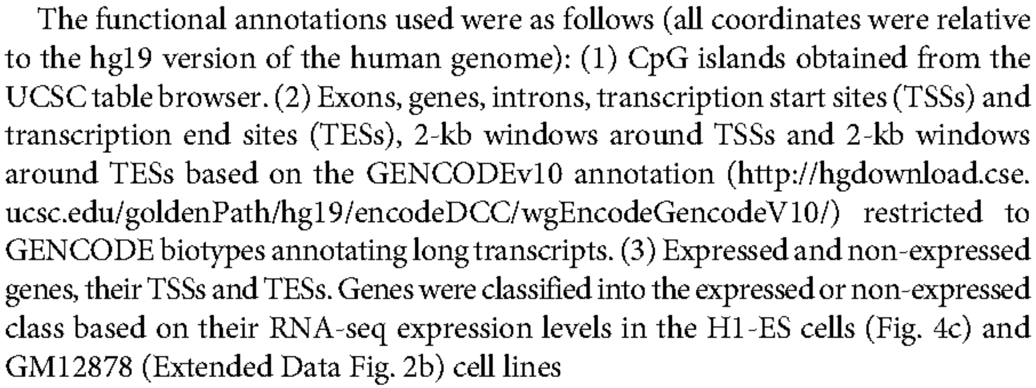
Digression: diversity of bioinformatics methods in this publication

Taken from: Roadmap Epigenomics Consortium (2015) Nature 518, 317



Chromatin state learning. To capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states) across 127 epigenomes, we used ChromHMM v.1.10[106], which is based on a multivariate Hidden Markov Model.

Hidden Markov models to annotate chromosomal states



The functional annotations used were as follows (all coordinates were relative to the hg19 version of the human genome): (1) CpG islands obtained from the UCSC table browser. (2) Exons, genes, introns, transcription start sites (TSSs) and transcription end sites (TESs), 2-kb windows around TSSs and 2-kb windows around TESs based on the GENCODEv10 annotation (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV10/) restricted to GENCODE biotypes annotating long transcripts. (3) Expressed and non-expressed genes, their TSSs and TESs. Genes were classified into the expressed or non-expressed class based on their RNA-seq expression levels in the H1-ES cells (Fig. 4c) and GM12878 (Extended Data Fig. 2b) cell lines

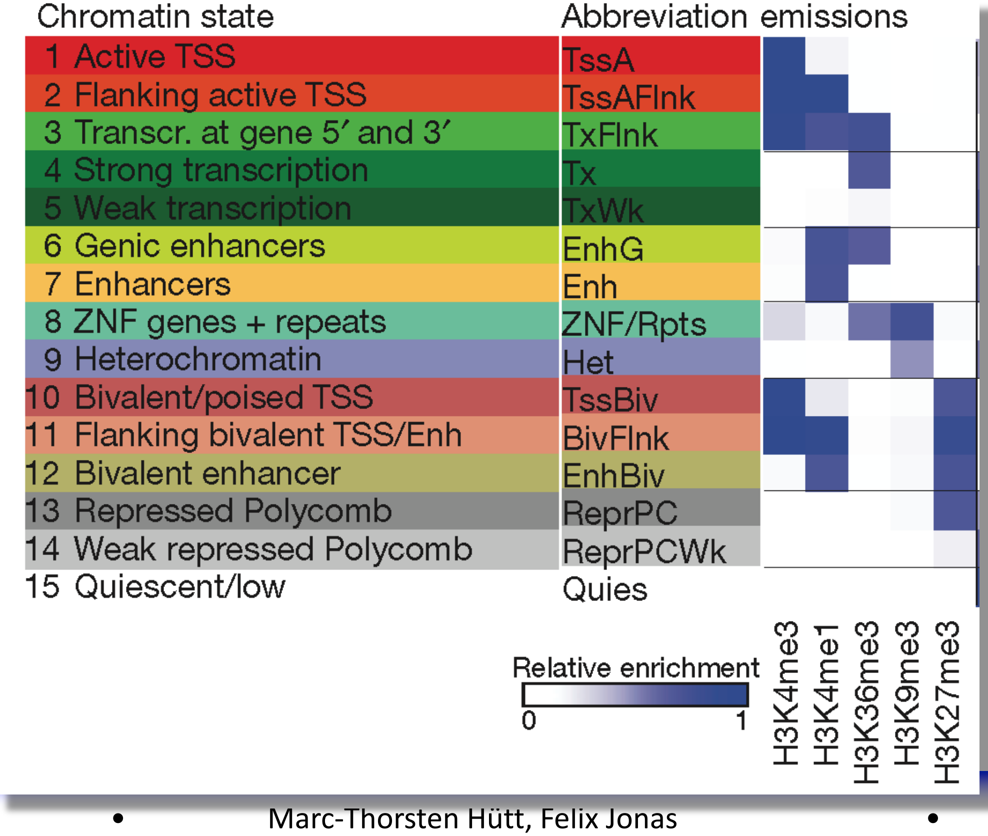
Use of **databases** and multi-'omics' **data integration**

Current status of HMMs in Bioinformatics

Taken from: Roadmap Epigenomics Consortium (2015) Nature 518, 317

Chromatin state learning. To capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states) across 127 epigenomes, we used ChromHMM v.1.10[106], which is based on a multivariate Hidden Markov Model.

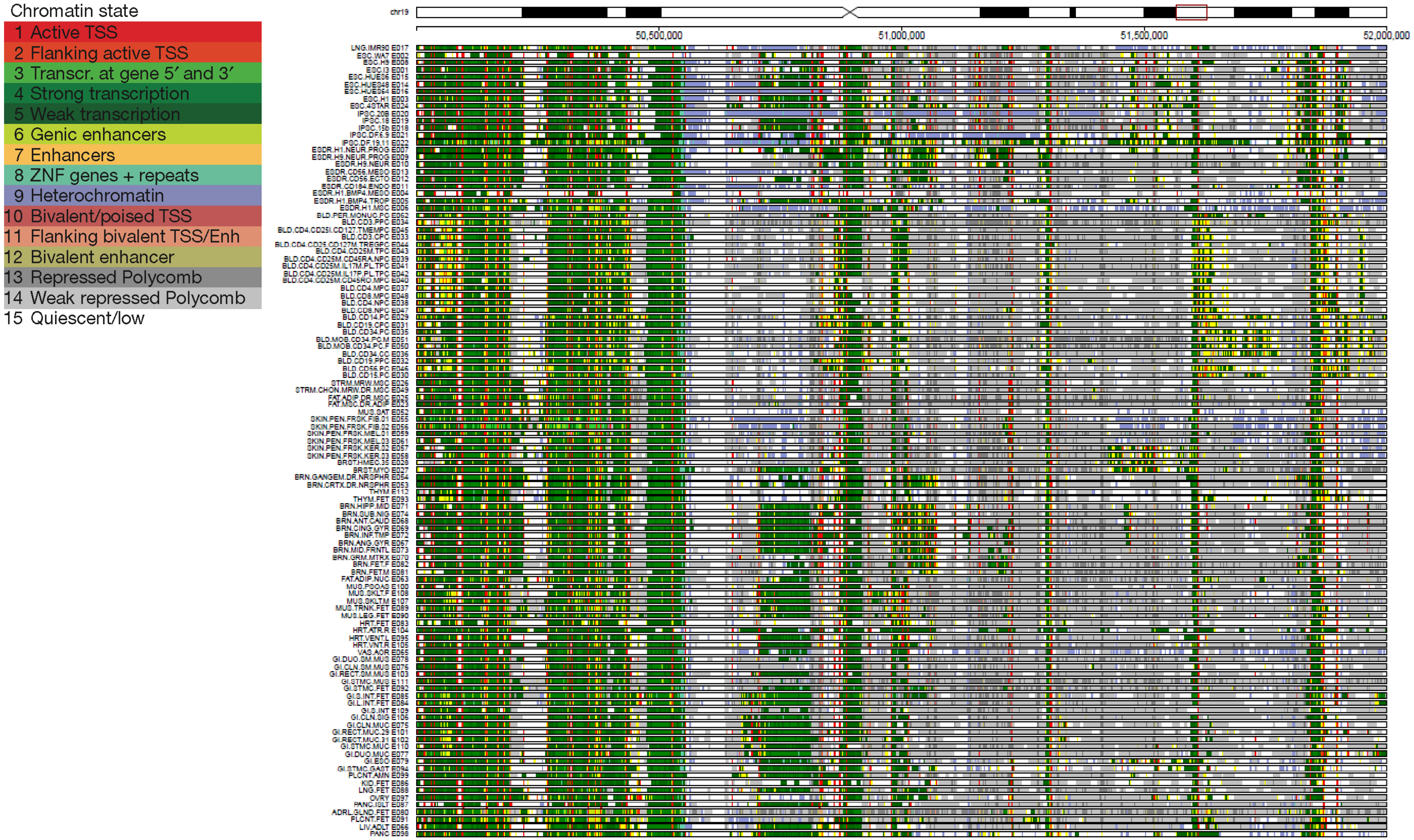
Hidden Markov models to annotate chromosomal states



Chromatin state	Abbreviation
1 Active TSS	TssA
2 Flanking active TSS	TssAFlnk
3 Transcr. at gene 5′ and 3′	TxFlnk
4 Strong transcription	Tx
5 Weak transcription	TxWk
6 Genic enhancers	EnhG
7 Enhancers	Enh
8 ZNF genes + repeats	ZNF/Rpts
9 Heterochromatin	Het
10 Bivalent/poised TSS	TssBiv
11 Flanking bivalent TSS/Enh	BivFlnk
12 Bivalent enhancer	EnhBiv
13 Repressed Polycomb	ReprPC
14 Weak repressed Polycomb	ReprPCWk
15 Quiescent/low	Quies

Current status of HMMs in Bioinformatics

The complexity of genome organization and genomic annotations


Chromatin state
1 Active TSS
2 Flanking active TSS
3 Transcr. at gene 5′ and 3′
4 Strong transcription
5 Weak transcription
6 Genic enhancers
7 Enhancers
8 ZNF genes + repeats
9 Heterochromatin
10 Bivalent/poised TSS
11 Flanking bivalent TSS/Enh
12 Bivalent enhancer
13 Repressed Polycomb
14 Weak repressed Polycomb
15 Quiescent/low
chr19