Annotation
Gene prediction
also with RNA-Seq
(Probabilistic models)

T. Schiex & E. Sallet
INRA – MIA & SPE
Toulouse

Overview

- Annotation
- RNA-Seq alone
- *Ab initio in silico* Gene Finding (proteins)
  - Probabilistic models, inference
  - Markovian models (generative)
  - Hidden Markov models for gene finding
  - Practice on EST cluster sequence analysis
- Using experimental evidence
  - Integration: conservation, dealing with dependencies
  - An annotation pipeline example
  - Bacterial annotation with oriented RNA-Seq

Sequencing Genomes

- Concerted effort to sequence model organisms
  - E. coli 4.5 Mb (1997)
  - S. cerevisiae 6.0 Mb (1997)
  - C. elegans 98 Mb (1998)
  - Human 3Gb (2000)
    - Estimated cost : $2.7 billion (1991 dollars)
    - Estimated time : 15 years

- 2003 : NHGRI commited to develop NGS.

  - 30x human genome : $100,000, then $1,000

New technologies

applied biosystems

SoLID (2007)

illumina

Genome Analyzer (2007)

Roche

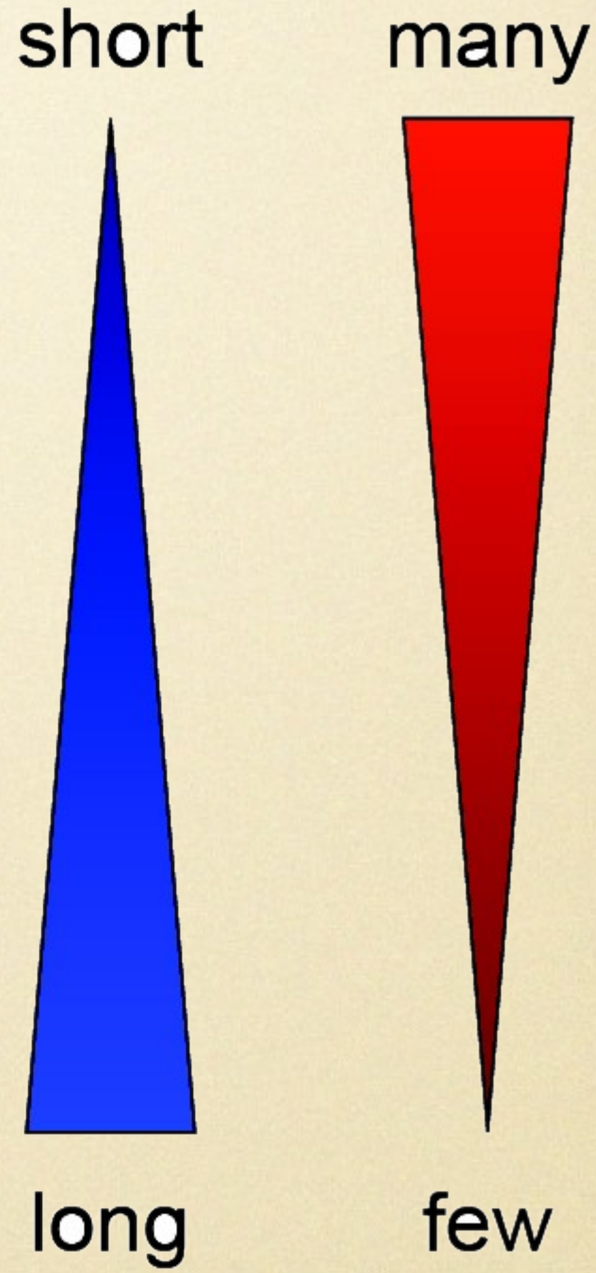
454 Sequencing (2005)

PACIFIC BIOSCIENCES

SMRT (2010)

Nanopore (2012)

life technologies | Semiconductor Sequencing for Life™ ion torrent



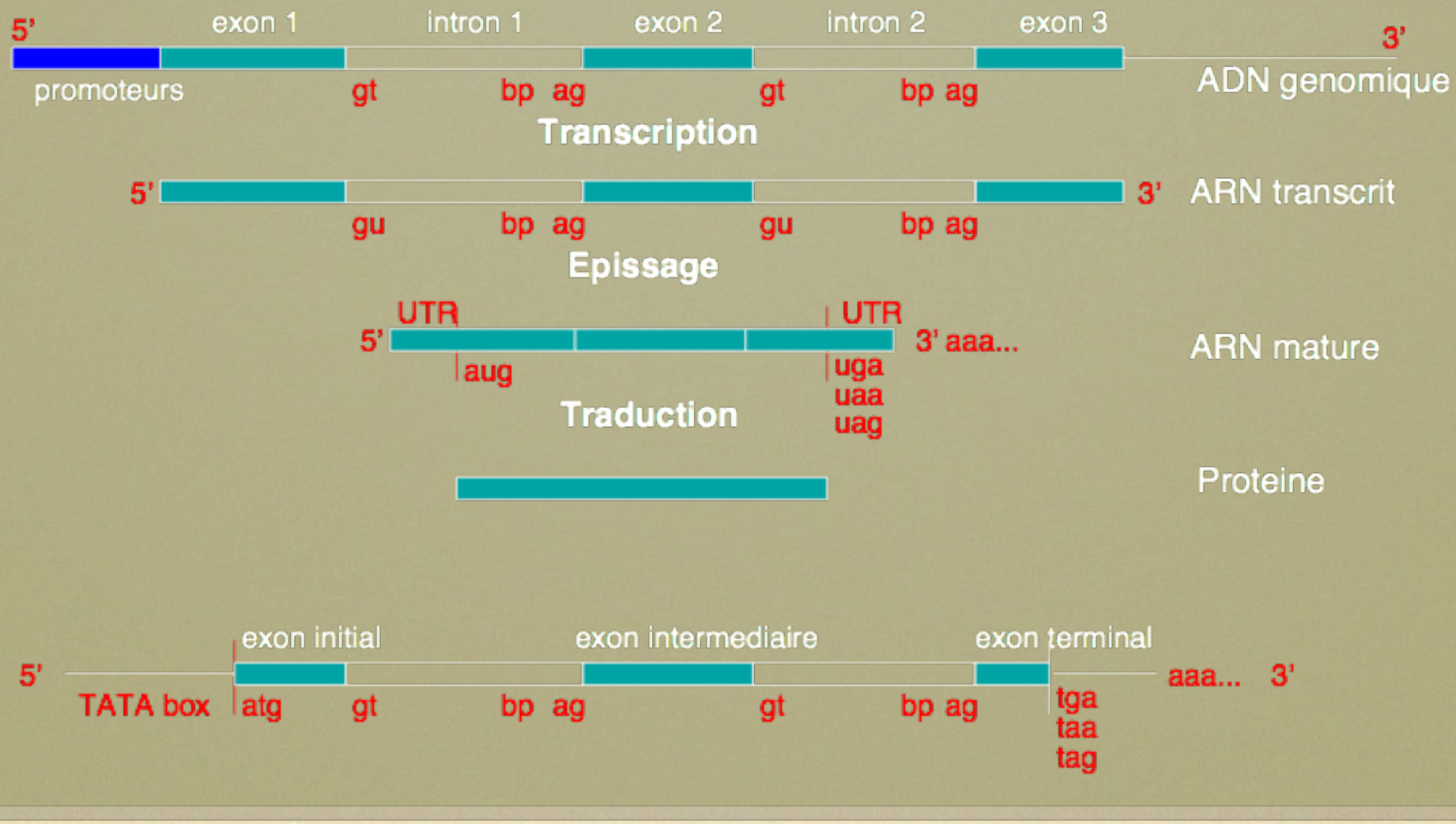
[Slide provided by Ali Mortazavi]

- DNA sequencing : genome
- RNA sequencing : transcriptome
- Chip-Seq, PET-Seq... potentially useful for annotation

Annotation

- From a raw genomic sequence
  - Repeated elements (TE) identification
  - RNA gene identification (tRNA, RFAM, miR)
  - **Protein gene identification**
  - Physical/Chemical/Functional annotations

Terminology



- Gene = several regions separated by "signals"

In silico extrinsic blunt approach

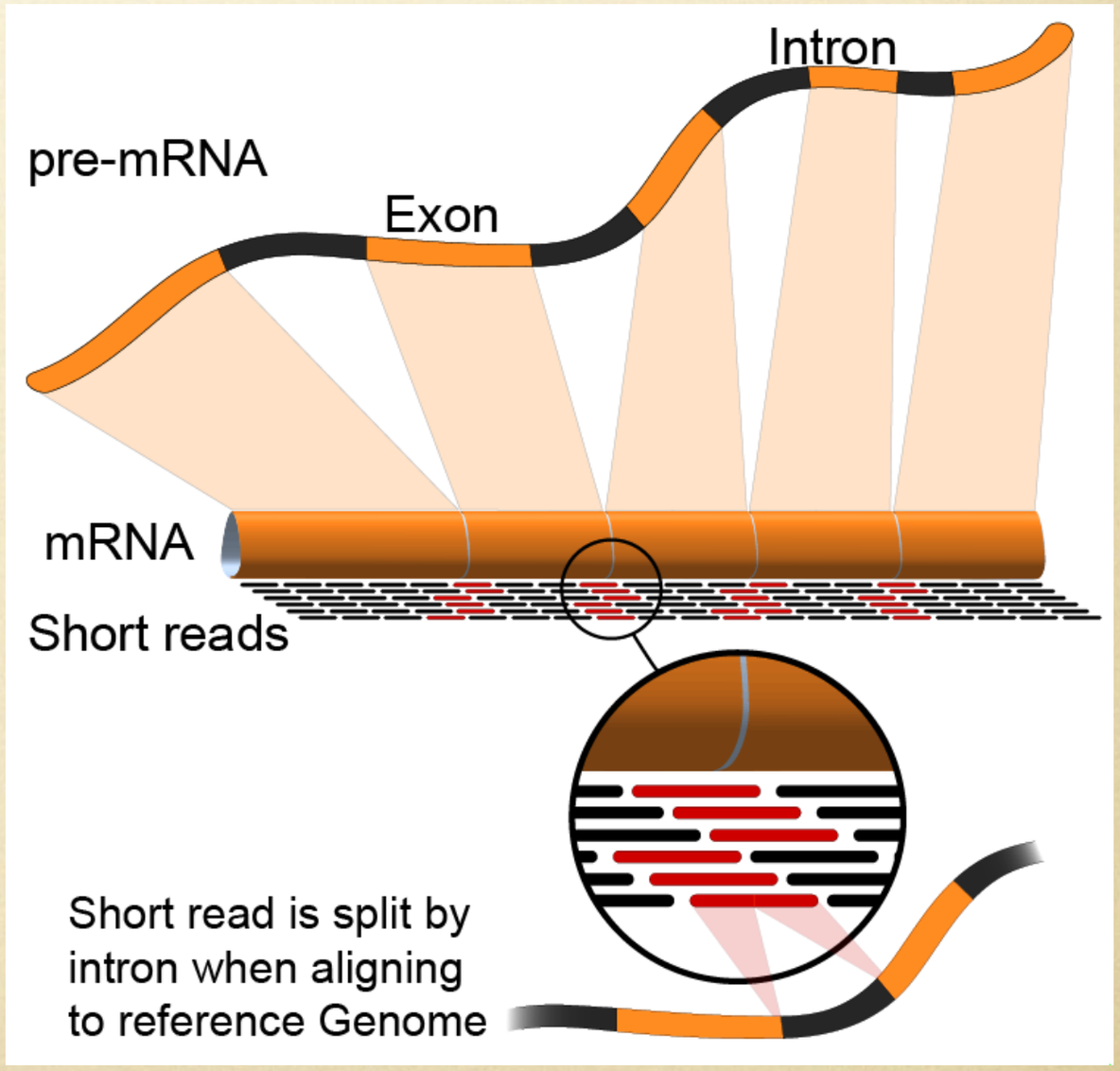
Alignment against know sequences

- DNA *vs* mRNA: BLASTN (transcribed)
- DNA vs peptides: BLASTX (translated)
- DNA vs DNA: TBLASTX (conserved)

Imprecise, blunt exon frontiers, conservative

Deep RNA-Seq

High throuput transcriptome measurement



Wikipedia

Spliced alignment

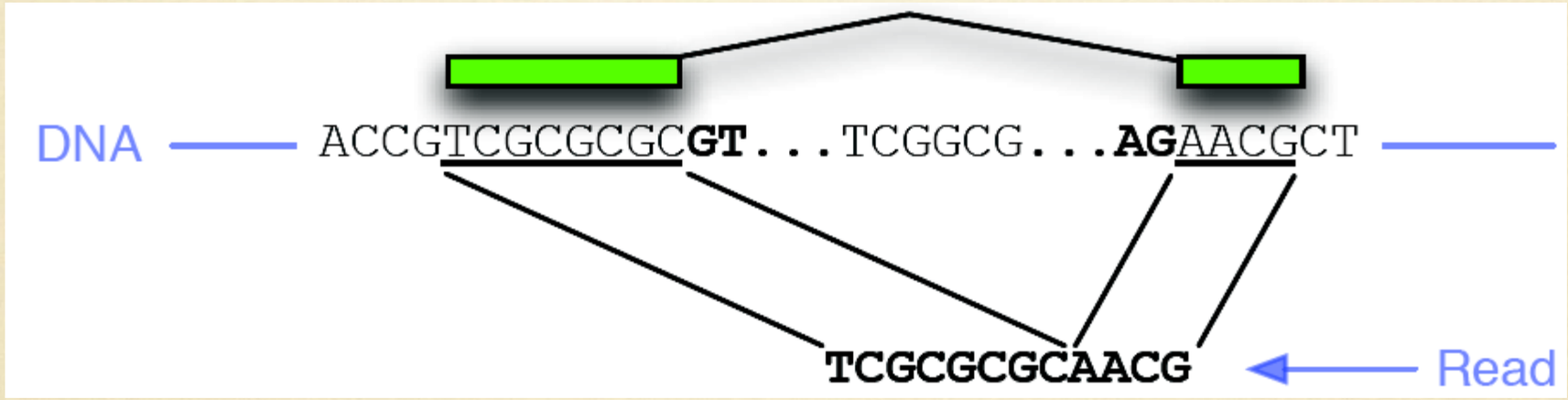


Figure by Gunnar Rätsch (StatSeq 2010)

- Can be done on long sequences (EST, mRNA)
- Or directly on « short » reads (less reliable for short overlaps)

RNA-Seq approaches

- Assembly : reconstructs (partial) mRNA. Sufficient depth needed. Oases, Trinity, Velvet..
- Spliced align : Genome Threader,...

- Spliced read mapping : Qpalmapper,TopHat...
  - Sequence_alignment_software#Short-Read_Sequence_Alignment (wikipedia)
  - seqanswers.com
  - BWA : extended suffix array & BW transform

BIOINFORMATICS REVIEW

Vol. 28 no. 24 2012, pages 3169–3177
doi:10.1093/bioinformatics/bts605

Sequence analysis

Advance Access publication October 11, 2012

Tools for mapping high-throughput sequencing data

Nuno A. Fonseca*, Johan Rung, Alvis Brazma and John C. Marioni

EMBL Outstation, European Bioinformatics Institute (EBI), Hinxton, Cambridge CB10 ISD, UK

Associate Editor: Jonathan Wren

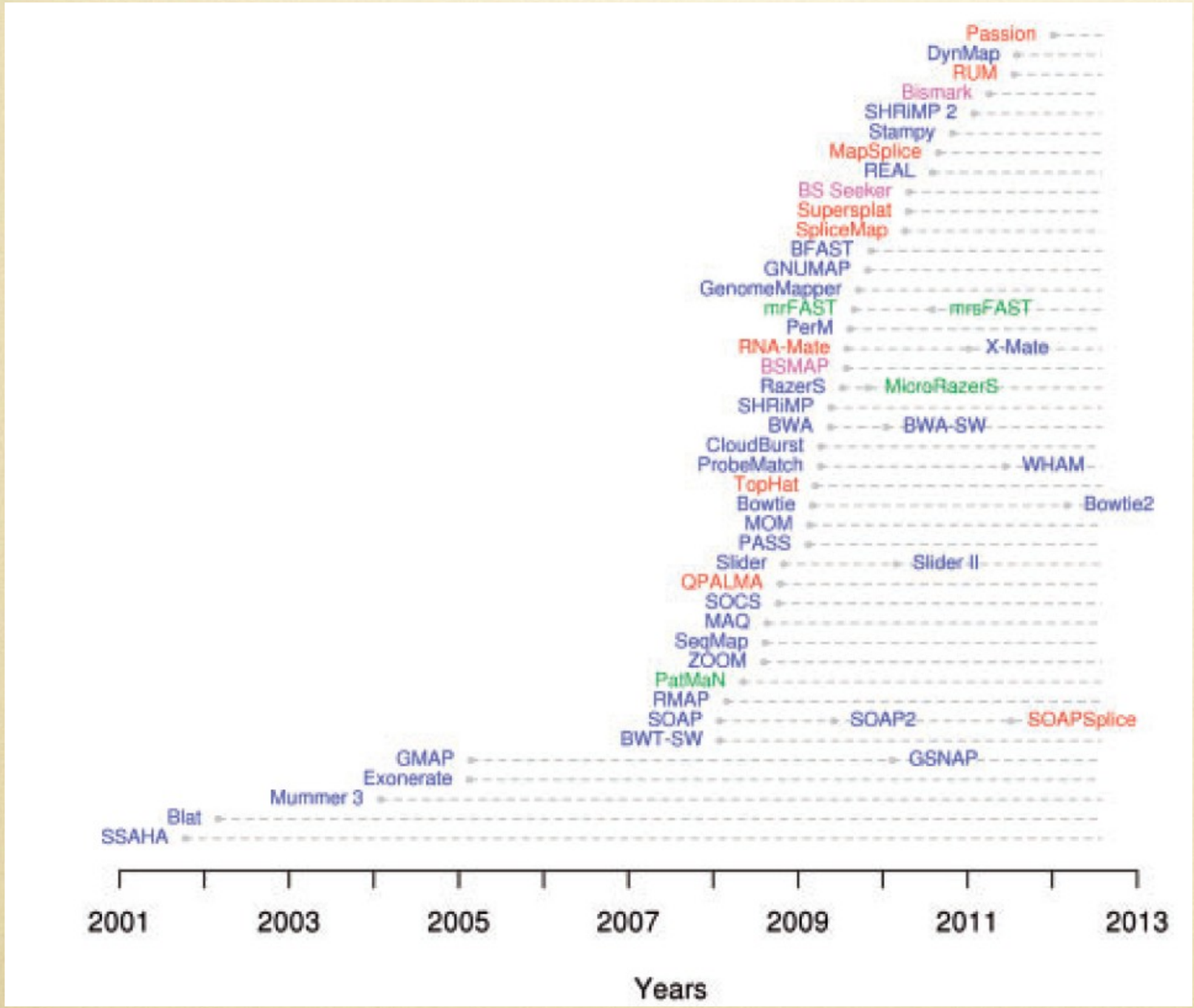
Motivation: A ubiquitous and fundamental step in high-throughput sequencing analysis is the alignment (mapping) of the generated reads to a reference sequence. To accomplish this task, numerous software tools have been proposed. Determining the mappers that are most suitable for a specific application is not trivial.

Results: This survey focuses on classifying mappers through a wide number of characteristics. The goal is to allow practitioners to compare the mappers more easily and find those that are most suitable for their specific problem.

Availability: A regularly updated compendium of mappers can be found at http://wwwdev.ebi.ac.uk/fg/hts_mappers/.

Contact: nf@ebi.ac.uk

Mappers



Blue : DNA

Red : RNA

Green: miRNA

Purple : bisulphite

In silico ab initio approach

Capture the « intrinsic » properties of genes in a mathematical model built from *known data*

- Ability to generalize
- Can only capture what is in the learning set

Textures (CDS), **Signals** (Splice sites...)

Ubiquitous: probabilistic (Markovian) models

Probabilistic models

- Model the universe of all possibilities using variables ($X_1,\ldots,X_n$ for sequences of length n)
- A probabilistic model defines a probability distribution over all possible cases (sequences)
- Extra variables can be included (other, possibly unknown, informations)

Probabilities & Coins

- Positive, $\Sigma(p) = 1$
- Coins: $P(X=h)= q$, $P(X=t) = 1-q$.
- note $h \equiv 1$, $t \equiv 0$ then $P(X=x) = P(x) = q^x.(1-q)^{(1-x)}$
- iid trials: $P(x_1,x_2...,x_n) = \prod P(x_i) = q^{\#h}.(1-q)^{\#t}$
- $P(\#h=k) = C_k^N.q^k.(1-q)^{N-k} \Rightarrow$ max. in $q=k/N$

Conditional probabilities

- $P(X \mid Y)$ = probability of X knowing Y
  - X must take one value: $\sum P(X=x \mid Y) = 1$
  - coupled coins: $P(X_2 \mid X_1 = h) \neq P(X_2 \mid X_1 = t)$
  - P(X,Y): needs to take all interactions in account

Marginal & Conditionals

Coin 2 \ Coin 1	h	t	+
h	71	27	98
t	35	67	102
+	106	94	200

$P_e(h,h)=71/200$

$P_e(h \mid h) = 71/98$

$P_e(X_2=h) = 98/200$

$P(X_1,X_2) = P(X_1 \mid X_2).P(X_2)$

$P(X_1,X_2) = P(X_2 \mid X_1).P(X_1)$

Probabilistic models bioinformatics

- Many variables (sequences of length *n*)
- Often their values is know (data, *D*)
- Or missing (*M*: class, segmentation)
- May depend on parameters (θ)
- $P_\theta(D,M)$ is usually easy to express

What can we do ?

- **Generation**: sample from $P_\theta(D,M)$
- **Evaluation**: compute $P_\theta(D,M)$ or

$$P_\theta(D) = \sum_M P_\theta(D,M) \qquad \text{(missing data)}$$

- **Classification**: most probable class. We « just » need:

$$P_\theta(M \mid D) = P_\theta(M,D)/P_\theta(D)$$

More problems ?

- **Estimate** θ (or M) from known data D

max. likelihood (**ML**)

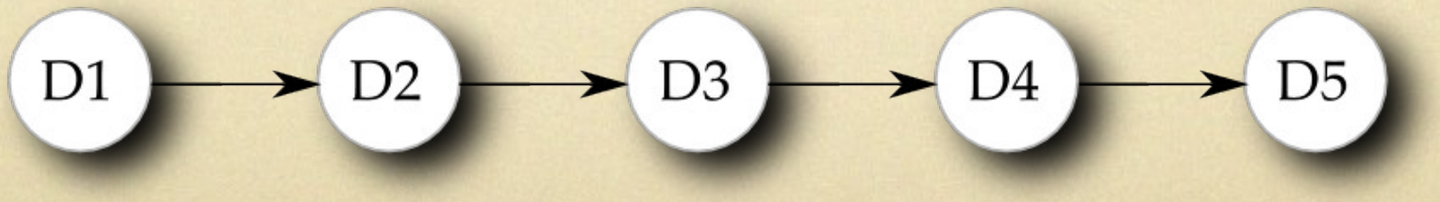
$$\max P_\theta(D) = \sum_M P_\theta(D,M)$$

Coin example

- 2 coins thrown N times overall, 1 coin change
- HHHTHHHTHHTHHHHHTHTTHHTTTTT
- D = sequence, M = coin change, $\theta = q_1, q_2$
- $P_\theta(D,M) = q_1^{\#h<M}.\ q_2^{\#h\geq M}.(1-q_1)^{\#t<M}.(1-q_2)^{\#t\geq M}$
- could do ML estim. of θ, explore $P_\theta(M \mid D)$

Markov chain models

- A sequence of random variables $D_1D_2D_3$... (discrete time process). $D_i \in \{\texttt{a,t,g,c}\}$
- D_i are not necessarily independent. D_i is an order 1 Markov chain when:
- $P(D_i=b \mid D_1...D_{i-1}) = P(D_i \mid D_{i-1})$
- Future depends on past only through present



Markov chain models

Dependencies given by $\pi(a,b) = P(D_i=b \mid D_{i-1}=a)$

	a	c	g	t
a	0.1	0.4	0.2	0.3
c	0.5	0.0	0.25	0.25
g	0.3	0.2	0.2	0.3
t	0.25	0.2	0.35	0.2

$P(D_i=a \mid D_{i-1}=t) = 0.25$
$P(D_i=c \mid D_{i-1}=c) = 0$
$P(D_i=g \mid D_{i-1}=a) = 0.2$

Conditional probabilities (each line sum to 1)

The initial law $\nu(x)=P(D_1=x)$.

A simple case: Bernouilli model

	a	c	g	t
a	0.1	0.4	0.2	0.3
c	0.1	0.4	0.2	0.3
g	0.1	0.4	0.2	0.3
t	0.1	0.4	0.2	0.3

When the probabilities do not depend on the previous nucleotide

Each line is equal to ν

Models the `a,t,g,c` composition

Sampling the distribution

- Draw a character according to the initial law
- Given the previous character, draw a new character from the conditional law
- Repeat...

$$P(D_1=s_1,D_2=s_2,\ldots,D_n=s_n) = \nu(s_1).\pi(s_1,s_2)\ldots\pi(s_{n-1},s_n)$$

$$= \nu(s_1).\prod_{ab}\pi(a,b)^{\#ab}$$

Estimating the θ...

- to build the Markov model of a sequence
- In order to:
  - recognize similar sequences (classify)
  - use it as a background model for tests

Estimating the θ

- The probability of a sequence is

$$P(X_1 \ldots X_n = s_1 \ldots s_n) = \nu(s_1).\pi(s_1,s_2) \ldots \pi(s_{n-1},s_n)$$

$$= \nu(s_1).\prod_{ab} \pi(a,b)^{\#ab}$$

- Direct estimation is ML estimation

$$\nu(a) = \#a/n \quad \text{and} \quad \pi_e(a,b) = \#ab/\#a$$

- Sequences generated under π have the same composition in words of length 2 as $s_1,\ldots,s_n$

Classification Durbin et al. 1998

- CpG islands: CG often suppressed except for regions in front of genes (promoters, starts)

M_1: estimated on CpG islands

	a	c	g	t
a	0.18	0.27	0.43	0.12
c	0.17	0.37	0.27	0.19
g	0.16	0.34	0.38	0.12
t	0.07	0.36	0.39	0.18

M_2: estimated on the rest of the sequence

	a	c	g	t
a	0.3	0.2	0.28	0.22
c	0.32	0.30	0.08	0.30
g	0.25	0.25	0.3	0.2
t	0.18	0.24	0.29	0.29

Loglikelihood ratios

$$\text{Score}(s_1,\ldots,s_n) = \log P_{\theta 1}(s_1,\ldots,s_n)/P_{\theta 2}(s_1,\ldots,s_n)$$

$$= \sum (\#ab)\log \pi(a,b)$$

	a	c	g	t
a	-0.74	0.42	0.58	-0.8
c	-0.91	-0.30	1.81	-0.69
g	-0.62	0.46	0.33	-0.73
t	-1.17	0.57	0.39	-0.68

Use a sliding window

Possible tests

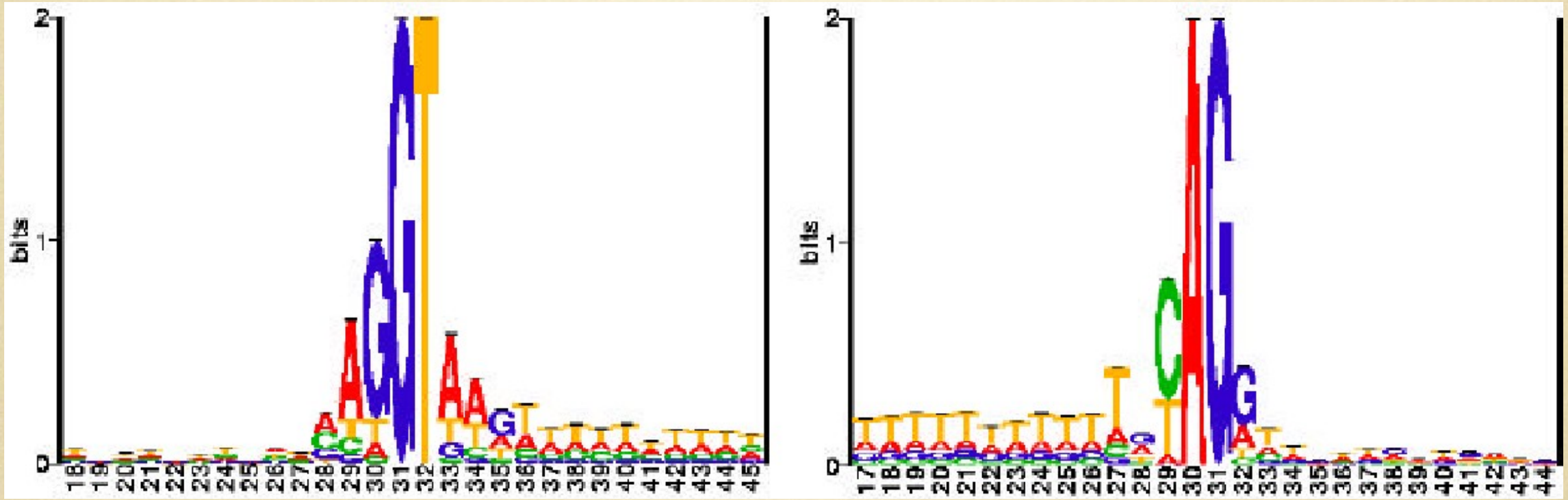
Frontiers ?

Position Weight Matrices

- Non homogeneous 0 order Markov chains.
- One model per position on aligned patterns.
  - Estimate $P_i(x)$ by frequency of x in column *i*
  - One positive (P) and one negative (P') model.
  - $Score(s_1,\ldots,s_n) = \sum_i -\log(P_i(s_i)/P'_i(s_i))$

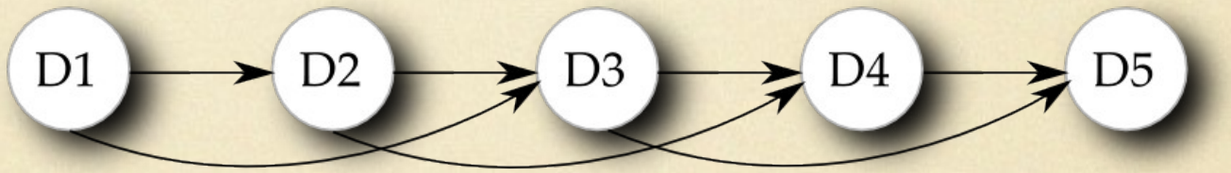
PWM = logos

(with background)



Generalization to higher orders

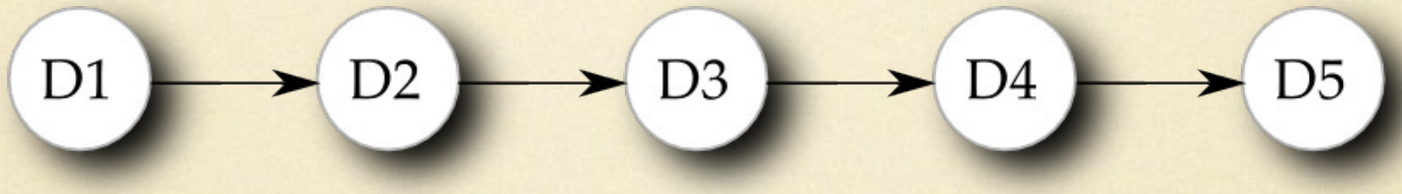
- In an order *k* Markov chain, the futur depends on the recent past (bounded memory)
- $P(D_i \mid D_1,\ldots,D_{i-1}) = P(D_i \mid D_{i-m},\ldots,D_{i-1})$
- order 2



- The initial law is defined on *k*-uplets
- The size of the transition matrix is 4^{k+1}

Other signal models

- WAM: Inhomogeneous order 1 Markov chain



- Neural networks
- Support vector machines (LDA+)
- Other classifiers...

Order 2 example

	a	c	g	t
aa	0.1	0.4	0.2	0.3
..	..	..	..	.
ta	0.0	0.3	0.0	0.7
tc	0.25	0.2	0.35	0.2
tg	0.0	0.25	0.05	0.7
tt	0.5	0.1	0.4	0.3

$P(a \mid tg) = 0.0$
$P(g \mid tg) = 0.05$
$P(t \mid tg) = 0.7$
$P(c \mid tg) = 0.25$

captures triplet comp.
(and more).

Direct estimation MLE

Overfitting

- The higher the order, the more flexible the model is. It can learn more (capacity)
- More parameters: smaller cardinalities to estimate parameters... poor estimation
- Characterizes the learning set well but learns details: does not generalize to other sequences.

Size of the training set !

3-périodic Markov chains

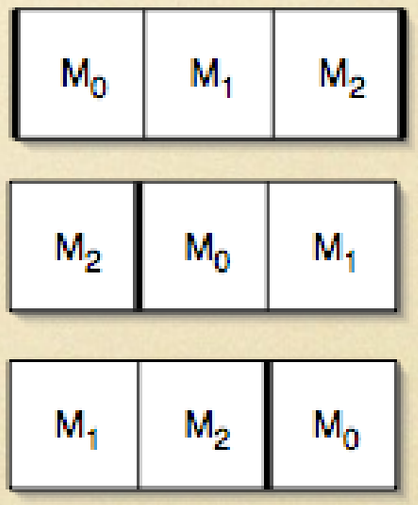
- For a CDS, the previous model is fine only if X_i is in the 3rd position of a codon.

- 3 models M_0,M_1,M_2 one for each position.

- phase 0: $\nu_0(X_1X_2)\prod P_{i \bmod 3}(X_i \mid X_{i-2},X_{i-1})$

- phase 1: $\nu_1(X_1X_2)\prod P_{i+2 \bmod 3}(X_i \mid X_{i-2},X_{i-1})$

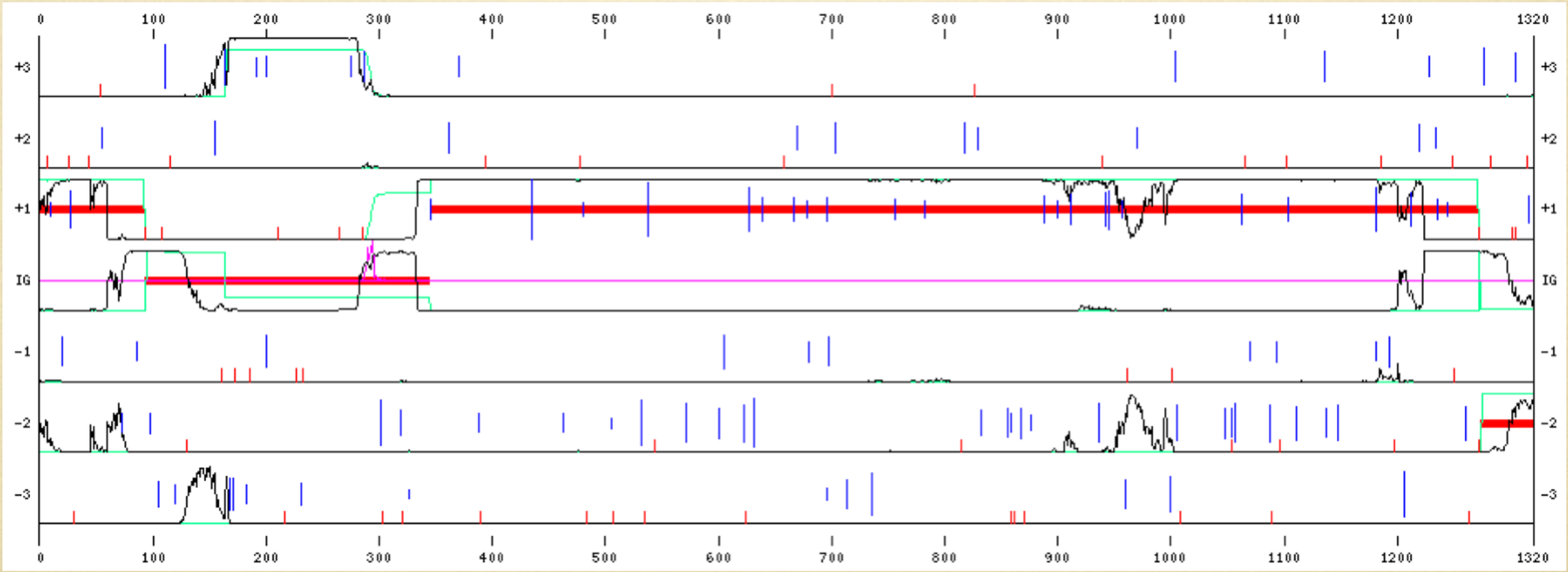
- phase 2: $\nu_2(X_1X_2)\prod P_{i+1 \bmod 3}(X_i \mid X_{i-2},X_{i-1})$



Looking for CDS (procaryotes/mRNA)

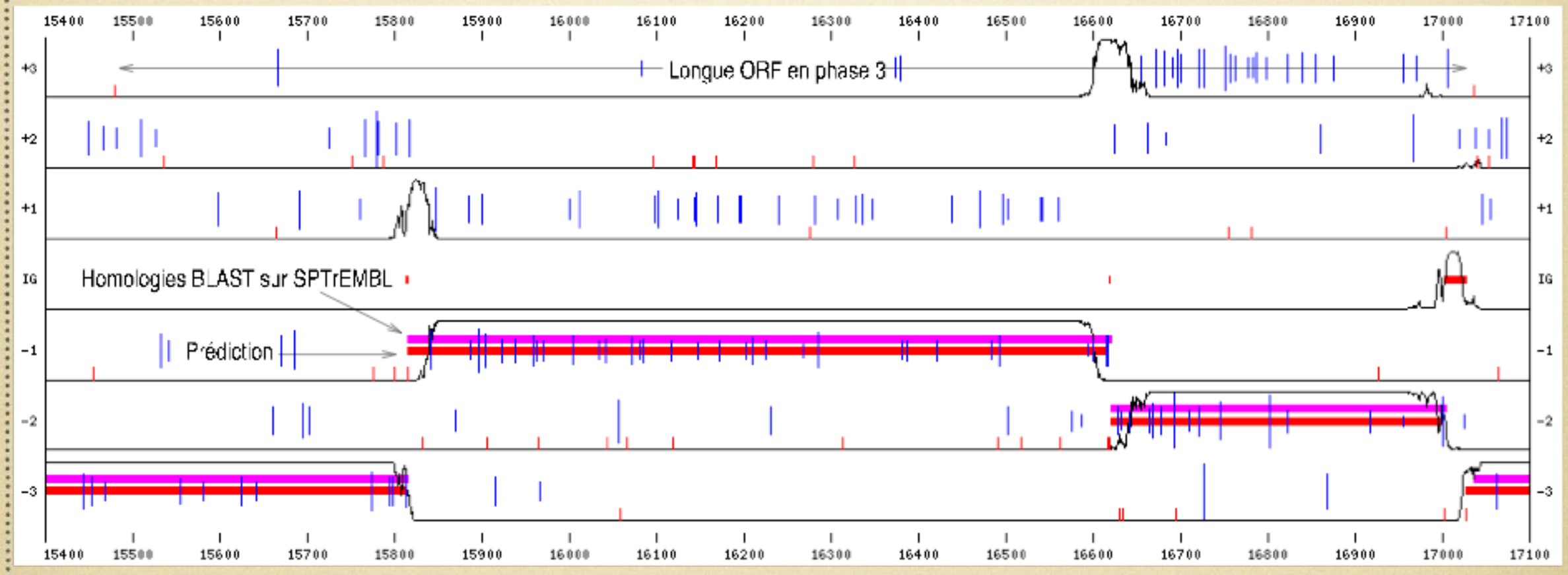
FrameD (NAR 2003)

- From known sequences: a 3-periodic & an intergenic model.
- Compare likelihoods $P_\theta(D)$ on a sliding window



Better than the long ORF approach

- GC rich genomes: ORFs make ORFs



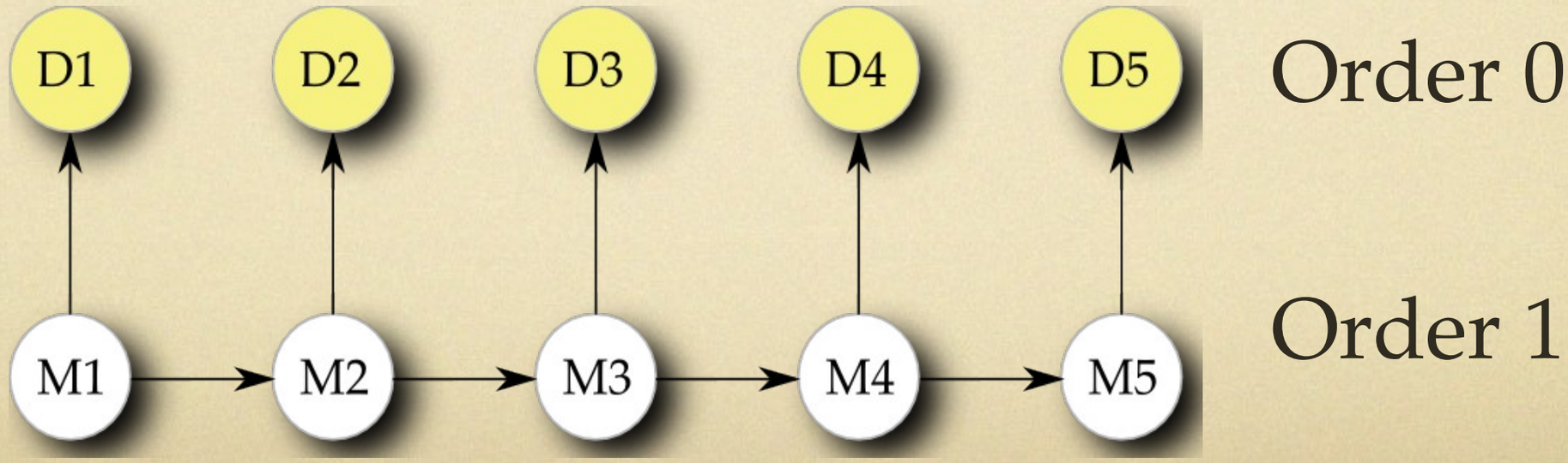
The segmentation problem

- We are given a sequence $s_1,\ldots,s_n$
- We want to split it in regions of "homogeneous" style
  - **supervised**: we know the styles
  - **unsupervised**: we don't know the styles and possibly not even their number

HMM: prob. model

Style, change of style = Markov

- Two type of variables:
  - M_i, hidden states = style used (missing data)
  - D_i, sequence = observed data
- For a style: sequence follows a Markov model
- The change of style follows a Markov model



Back to coins

- Two biased coins

h	t
1/2	1/2

h	t
1/4	3/4

emission probabilities $e_1(x)$, $e_2(x)$

- She draws a sequence of `h/t` using these coins. We just see this.

- She changes coin with probability 0.05

Transition probability $t(1,2)=t(2,1)$

- How likely it is she uses coins this way? Where did she change? What is the probability of using this very coin at some point?

Complete likelihood

- Note M the sequence of hidden states (coins)
- Note $D=(s_1,\ldots,s_n)$ the sequence (of h/t), known

$$P_\theta(M,D) = t(M_1).\prod[e_{Mi}(s_i).t(M_{i-1},M_i)]$$

- but... we don't know M !

MPE

(Most Probable Explanation)

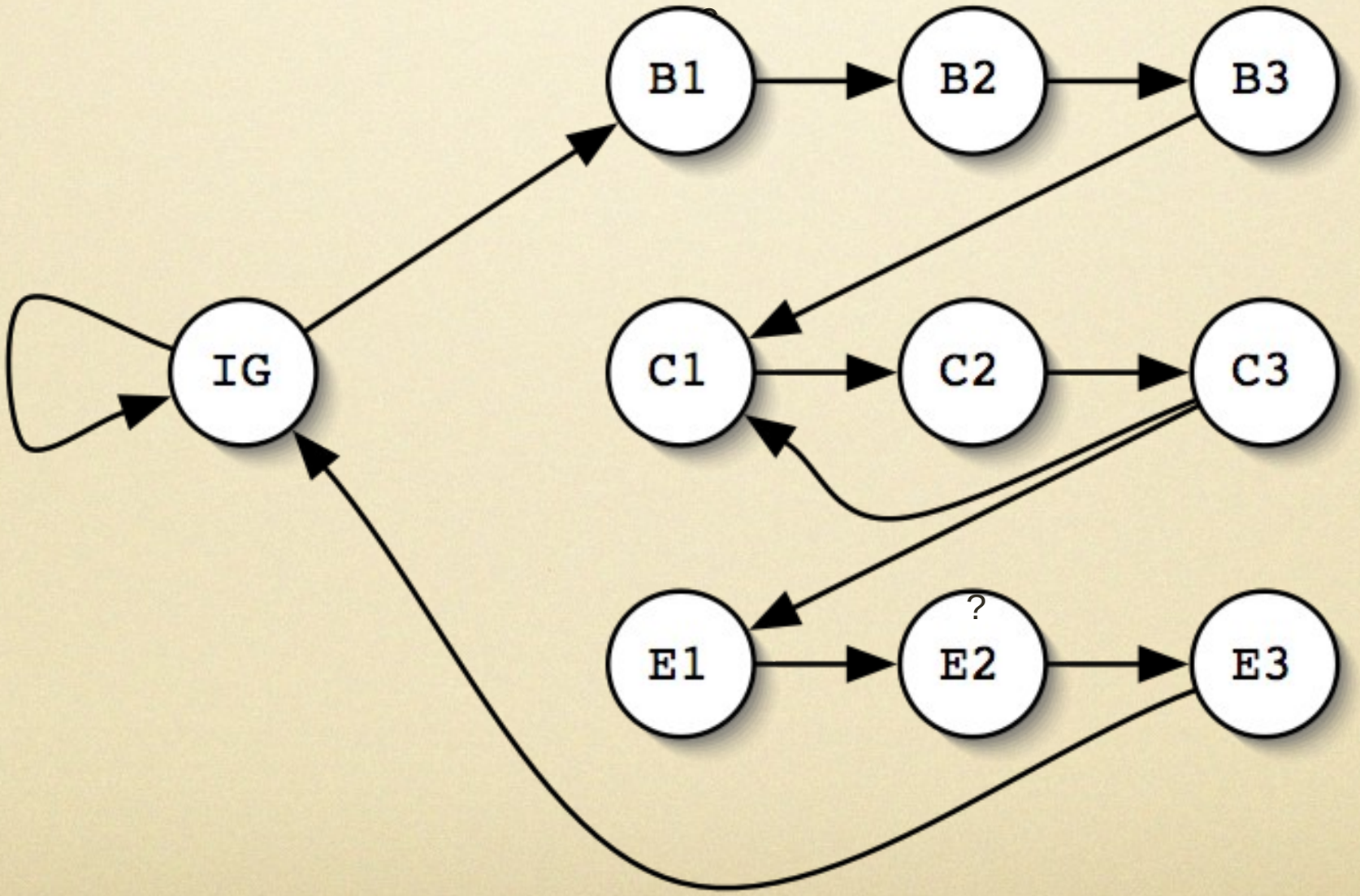
- If we don't know M we choose M* that maximizes

$$P_\theta(M,D)$$

- $M^* = \text{argmax}\ t(M_1).\prod[e_{Hi}(s_i).t(M_{i-1},M_i)]$
- Worst case exponential number of possible M but not cpu-intensive (Dynamic Programming)

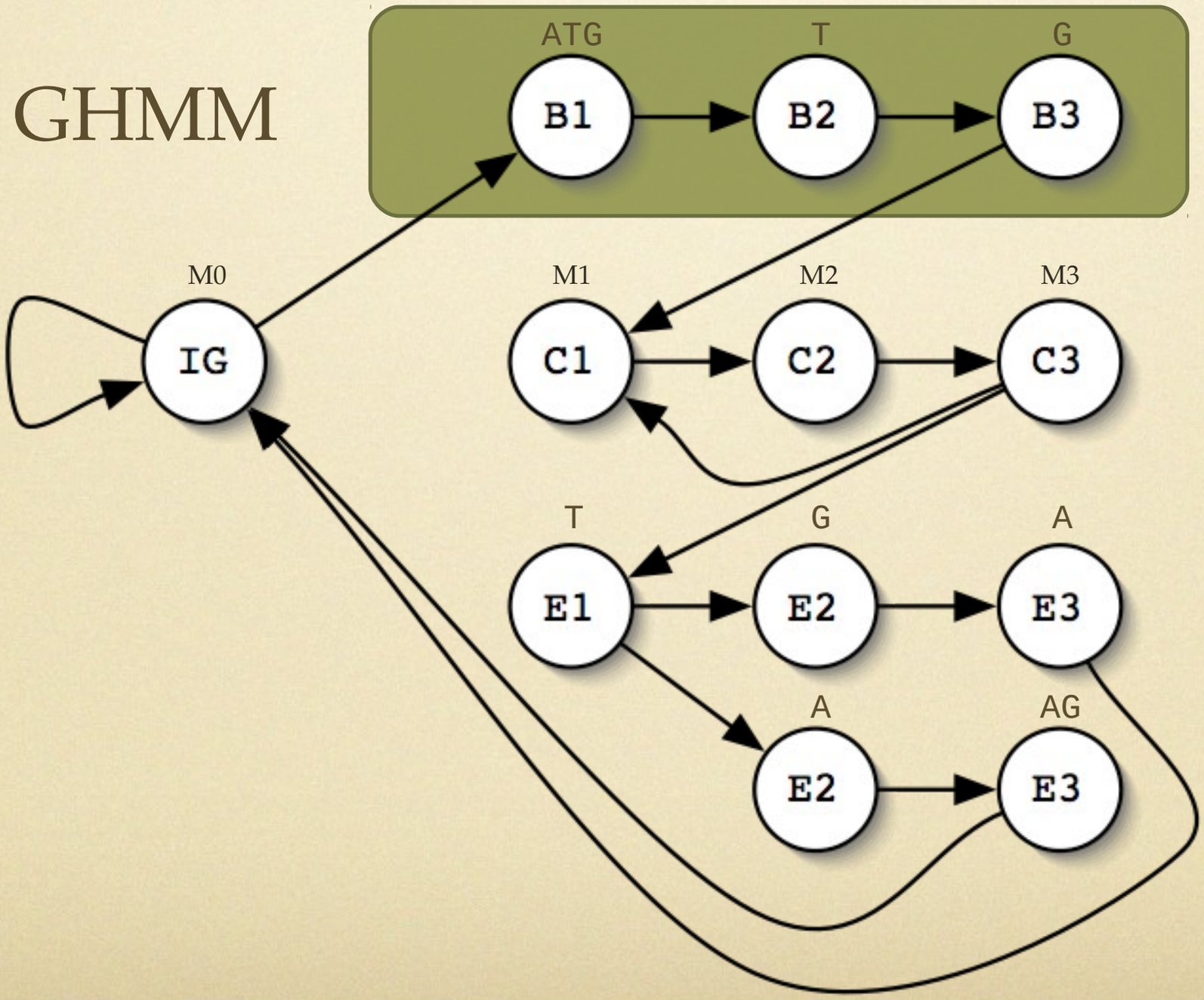
Gene finding

- Hidden states: intergenic, Start, CDS, Stop



Gene finding

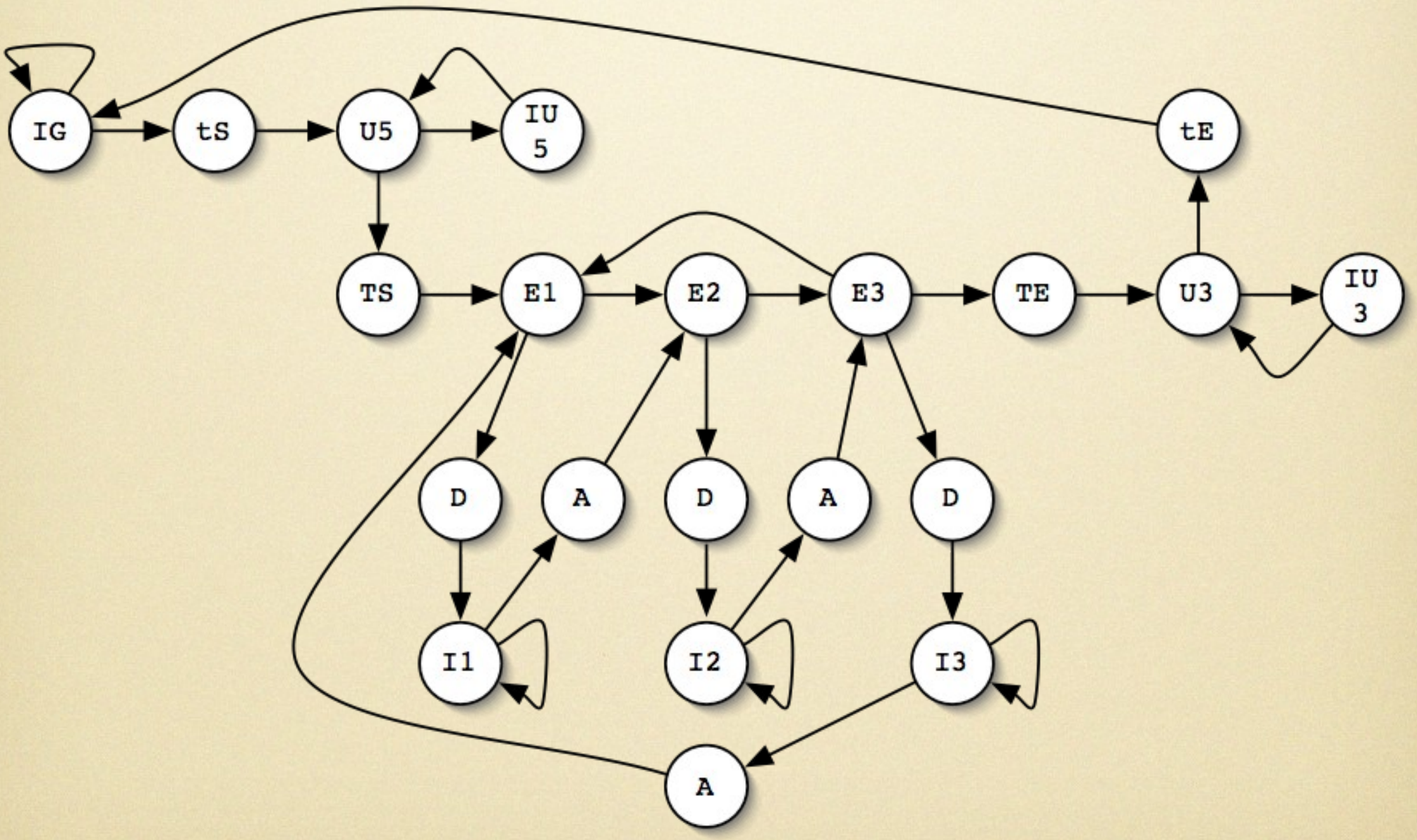
GHMM



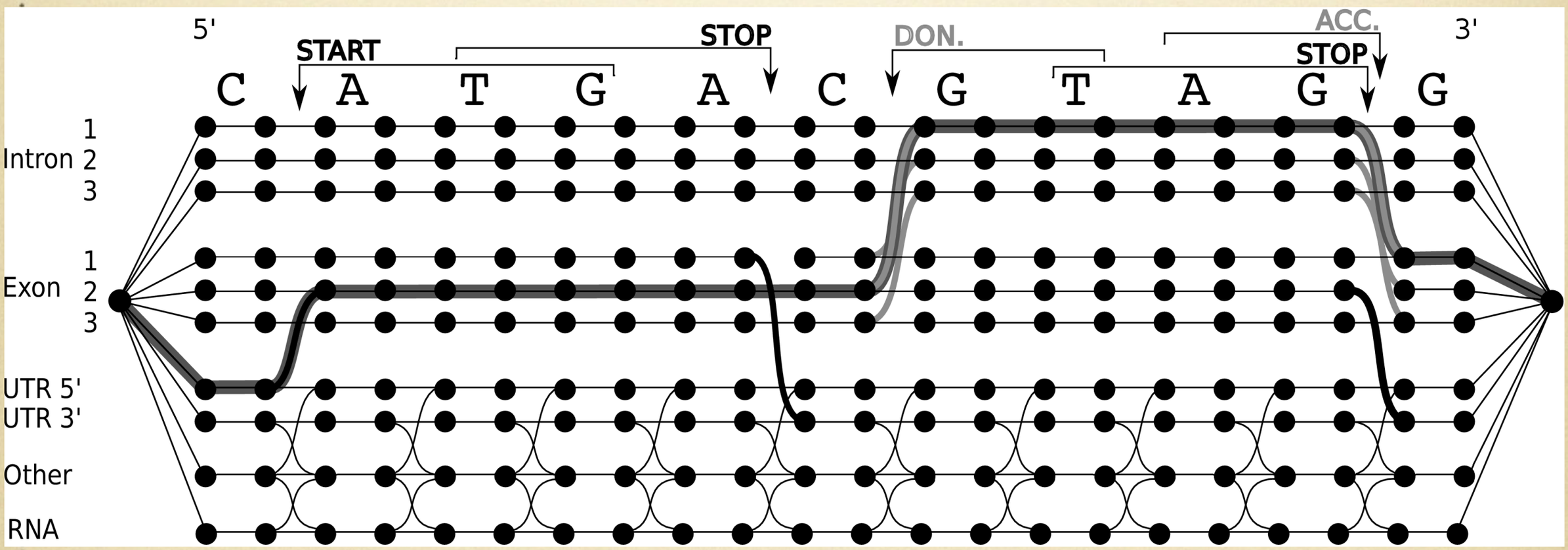
PWM

Anything that maps a triplet to a prob.

Eukaryota



Unwrapped over DNA

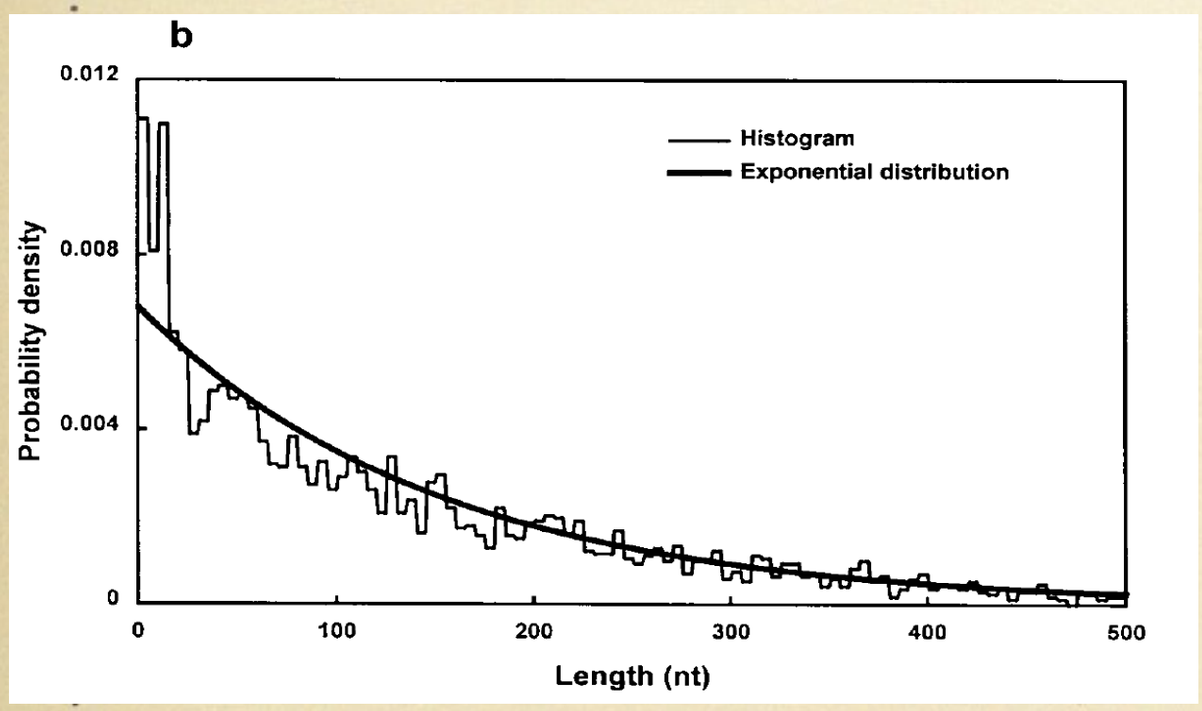


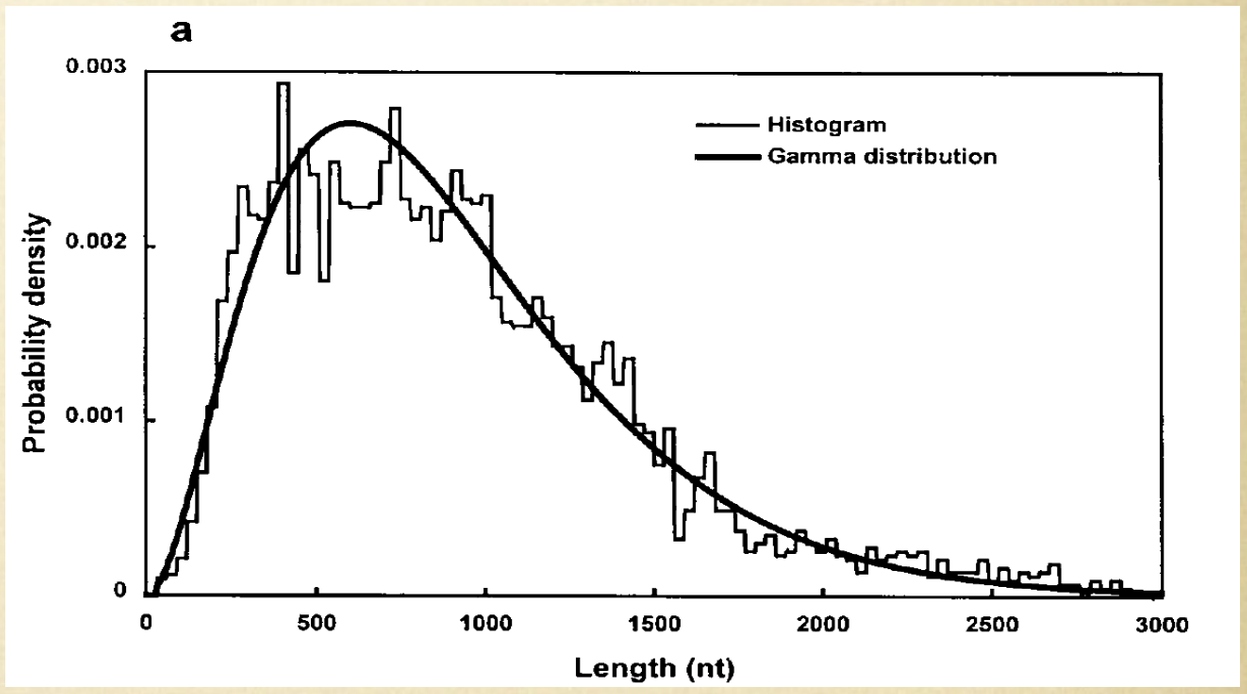
As in Smith & Waterman : with a proper weighting, a shortest path in the graph is an optimal segmentation

Lengths (1998)

- length *k*, state *i*: $t(i,i)^k.(1-t(i,i))$ probability

Exponential distribution





- Semi-HMM introduce explicit lengths (Viterbi becomes quadratic unless exponential tail).

Existing ab initio gene finders

- 3-periodic MM (order 5 typ.) for exons
- simple MM for introns...
- signals (ATG, Splice sites...): ad hoc models
- length modeling
- many hidden states (first, internal, terminal exons... more than 40 in EuGène)

AUGUSTUS, GeneMark.HMM, Glimmer, GenScan....

Training

- **supervised** training: we estimate all models on sequences with known hidden states.
- Direct estimation.
- Training set **quality**/composition crucial
  - specific to each organism (codon bias...)
  - representative of a "general case" ?
  - Auto-trainable heuristic (iterative) *ab initio* gene finders: GeneMarkES, AGENE.
  - Non supervised learning: EVIGAN (requires good existing gene predictors)

Weaknesses

- Little biology inside (no spliceosome, ribosome)
- Models are learned. Atypical behaviors are counter-selected:
  - signals (AG/GC or AT/AC sites)
  - codon usage/GC % - TE/Repeats
  - lengths (short exons, long introns)
- Alternatives ignored (splicing, trans.). In progress
- Unusual structures (genes in introns)
- Merge/split adjacent genes
- Long introns/intergenic difference ?

In silico extrinsic smarter approach

Spliced alignment against know sequences

- We can model splice sites (PWM, WAM, SVM)
- We use an alignment algorithm which opens gap easily on *good* splice site pairs.

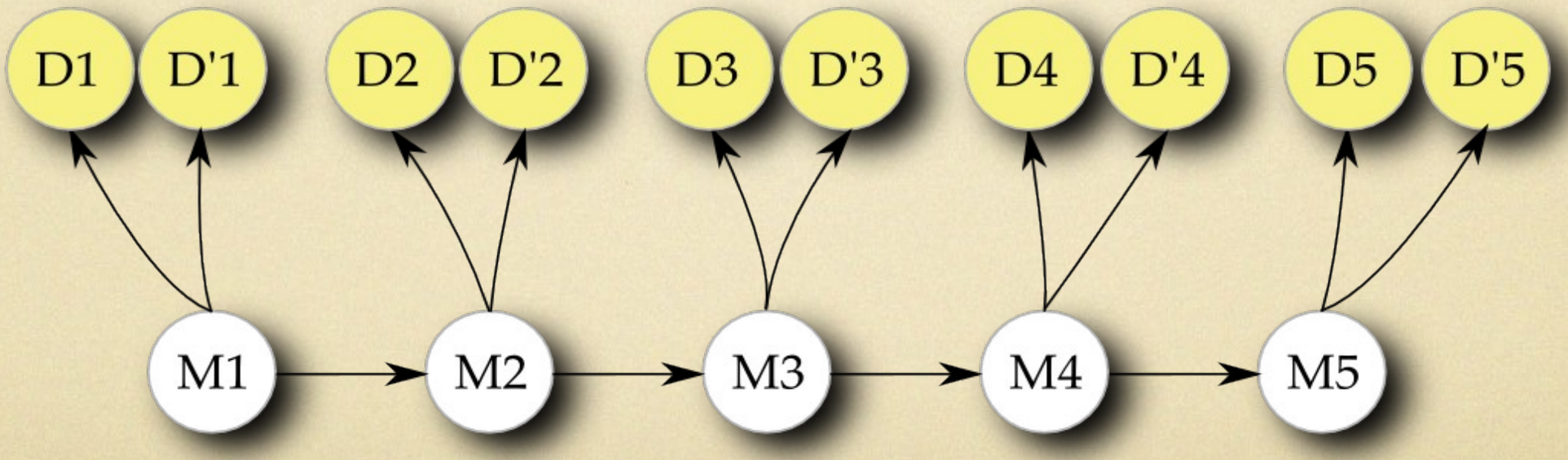
`SIM4``(92)` **`GeneWise`**`(04)` **`GenomeThreader`**`(05)`:
DNA/Proteins. Efficient. Free for public research.

Close match required.

Integrative approach

- Integrate *ab initio* with extrinsic data (mRNA, EST, protein matches, repeats, conservation...), other predictions, human expertise...

- An HMM could independently « emit » all these ?



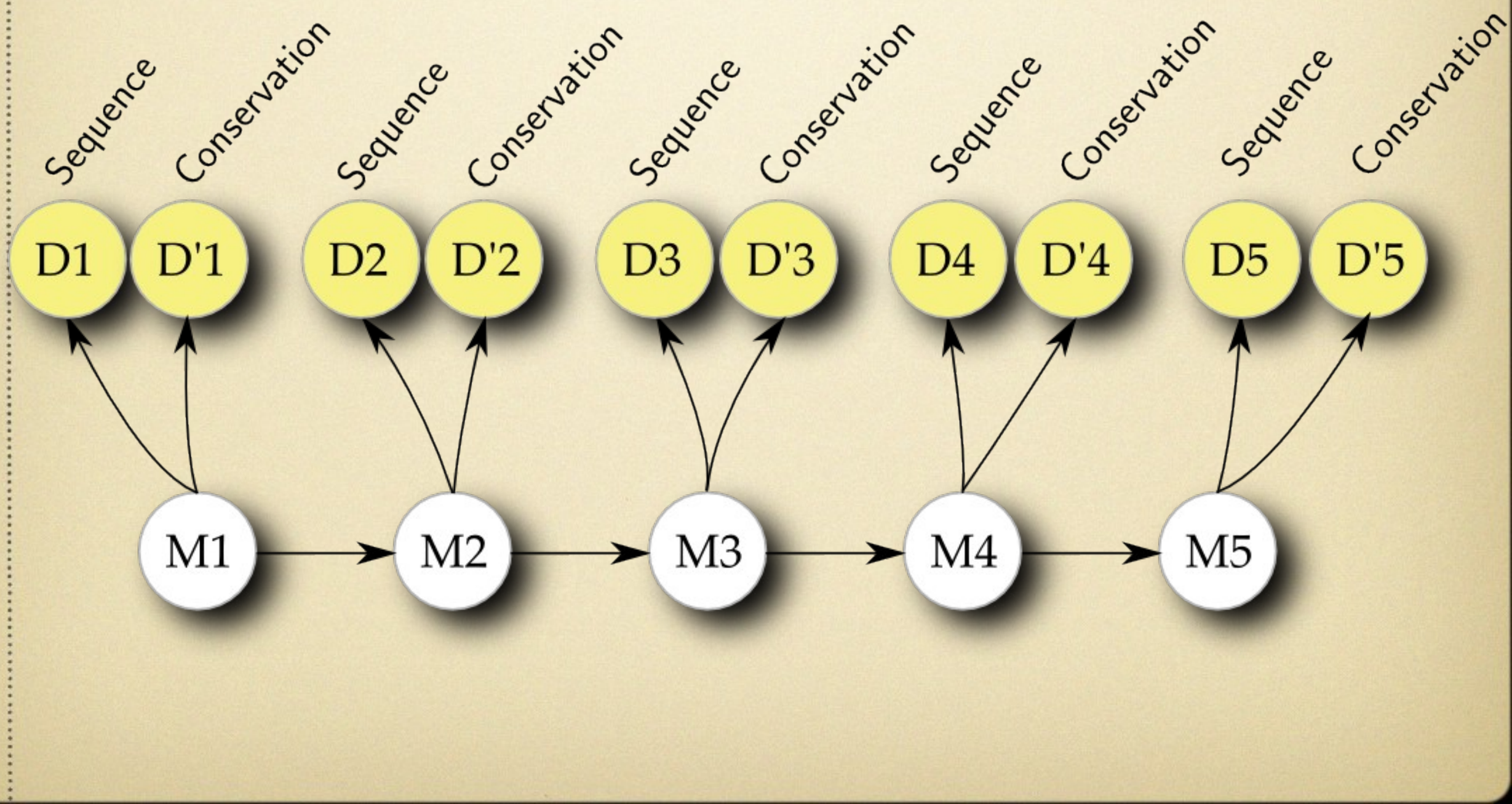
Conservations

- **TwinScan (2001)**: one informant genome.
  - DNA whole genome alignement produces alignment strings: | : - (match mism. gap).

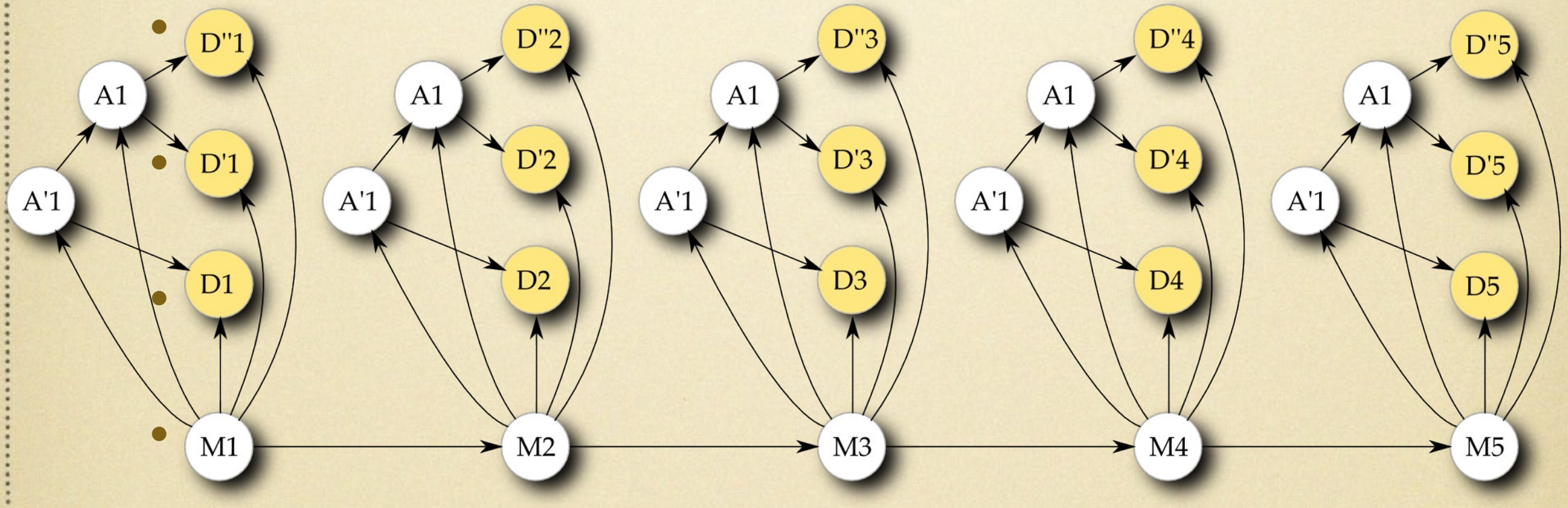
```
12345 6789 target position
 GAATT-CCGT target alignment
   ||  || | Blast Match symbols
 --ATCACC-T informant alignment
 ..||: ||:| conservation sequence
```

  - Different alignment strings 'styles' on exons, introns,... => estimation of Markov chains

TwinScan


Sequence
Conservation
Sequence
Conservation
Sequence
Conservation
Sequence
Conservation
Sequence
Conservation
D1
D'1
D2
D'2
D3
D'3
D4
D'4
D5
D'5
M1
M2
M3
M4
M5

Phylogenetic HMMs

- NSCAN (2006): several orthologous genome
- One multiple alignment column/position
- 

- Models the correlations in the data, the missing, and between them

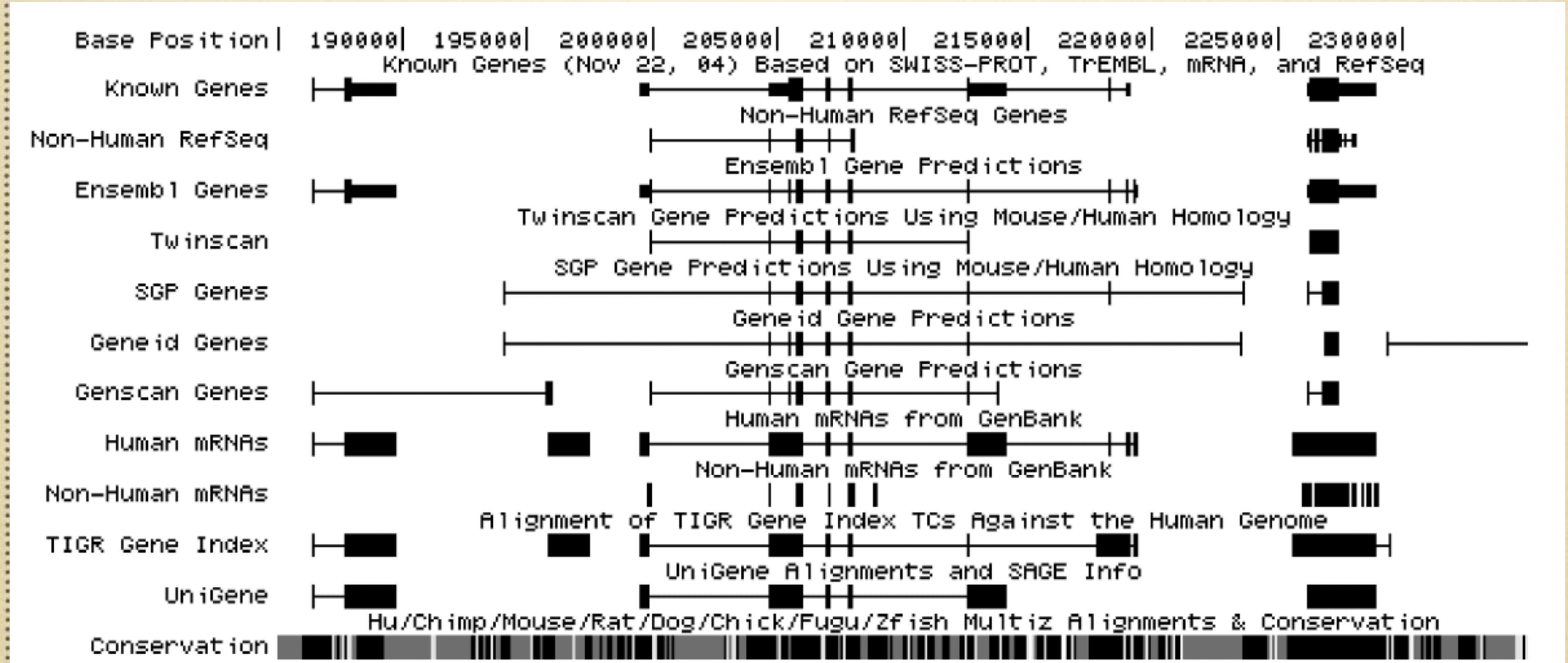
General situation

- Gene with limited expression, no or few EST
- No EST but protein matches, or both.
- Exon predicted but looking like repeats
- Ensembl (Curwen et al. 2004): strict rules for reliable annotation…

Likely missed genes.

Lot of dependencies in the data.

Human UCSC Genome Browser Chromosome 20



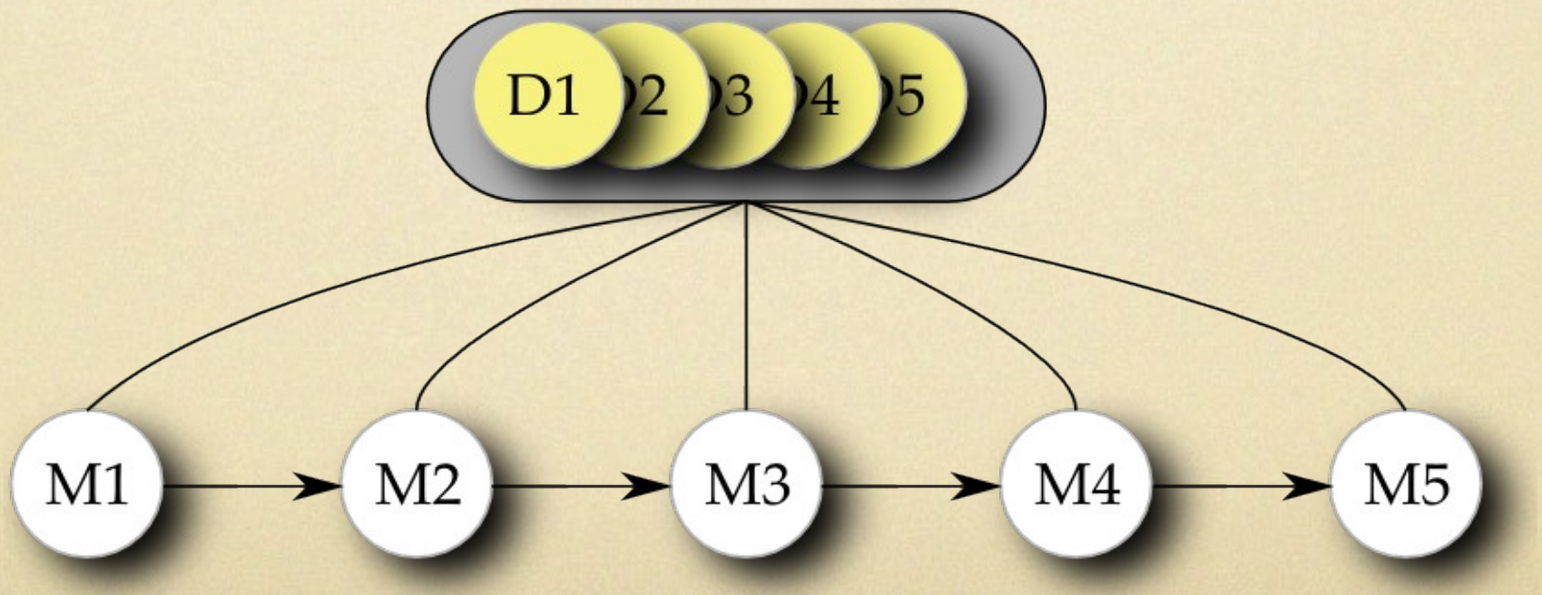
Discriminative GF

- Instead of modeling $P_\theta(D,M)$, we just model

$$P_\theta(M \mid D)$$

Linear CRF: Conditional Random Field (2001)

- Not generative, just discriminative



Data - Prediction

- Related by « **features** » $\phi_k(M_{i-1},M_i,D)$
  - If (any condition on Data) then a given « vote » distribution applies to M_i,M_j
  - $$P(M|D) \propto \exp\left(\sum_i \sum_k \lambda_k \cdot \phi_k(M_{i-1}, M_i, D)\right)$$
  - λ_k = strength of feature *k. Unknown parameter*
  - If (base *i* is part of a repeat) then vote for intron, UTR or intergenic at M_i

Parameter estimation

- Can maximize the « conditional likelihood » $P_\theta(M \mid D)$ on a learning set.
- Maximize the « actual or expected predictive quality » on a learning set.
  - Less sensitive to model weaknesses
  - Very sensitive to the data set
  - Hard (actual predictive qual. no derivative)

EuGène
A discriminative GF

Initial motivation (1999)

- To be able to use statistical information for region contents characterization: *ab initio features*
- But also the output of arbitrary signal prediction software: *additional features*
- And (almost) any other information at hand

Ab initio « features »

- **Region texture**: use Markov Models as in HMM . Votes with $-log(P)$, λ fixed to 1.

- **Signals**: use any classifier giving score v_i. Corresponding feature $\lambda_1+\lambda_2.v_i$ votes.

- All λ optimized by "maximum of success" on a set of known sequences by a genetic algorithm.

More « features » Similarities

- **Proteins**
  - Vote for “matches” (BlastX) for in-frame coding tracks using BlastX score.
  - also vote for intron on gap borders
  - One λ per databank (SP, PIR, TrEMBL)

- **EST/ADNc/Assembled RNA-Seq**
  - Vote for exon/UTR on hits, for introns on gaps (spliced alig.). One λ.

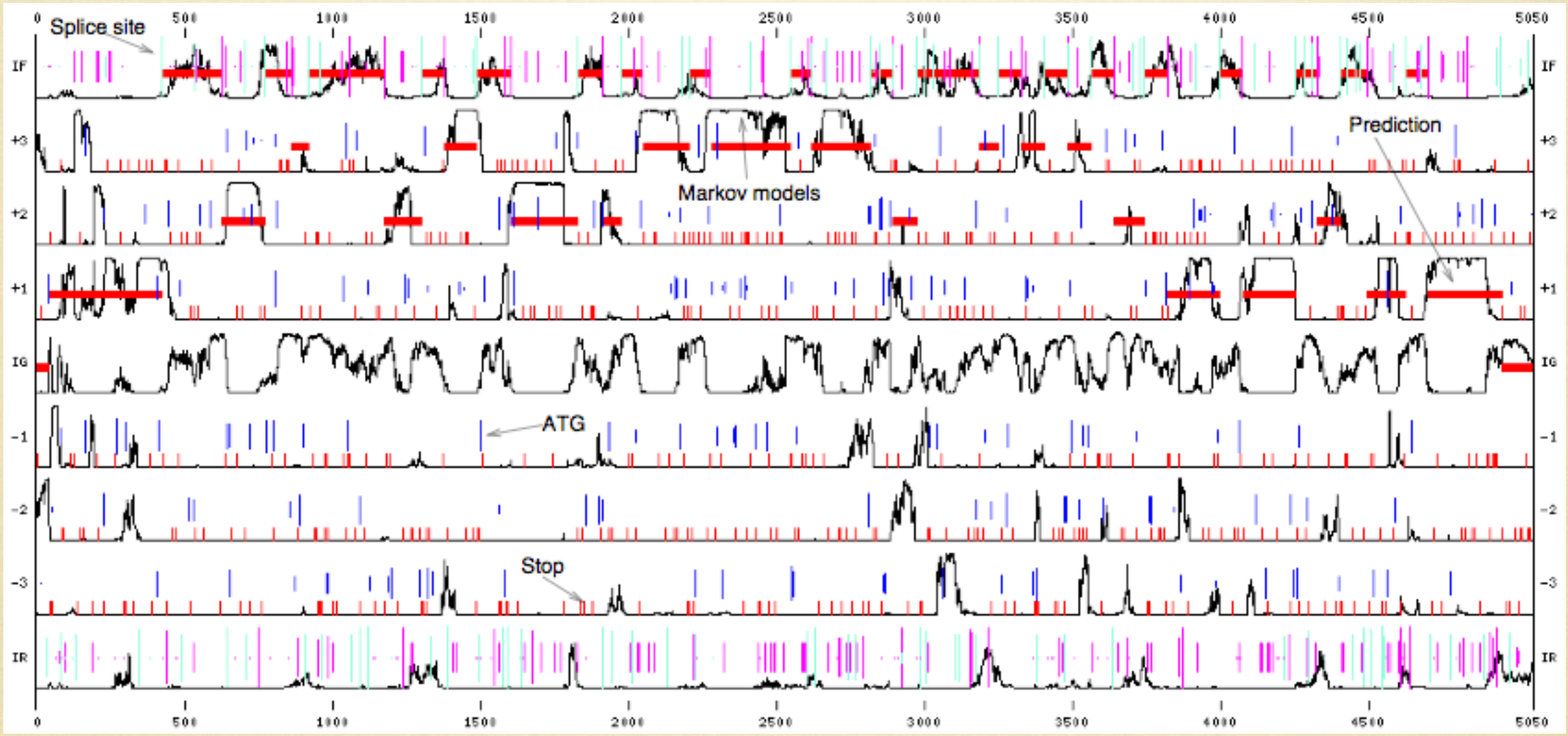
Other features

- **Conservation (TBlastX)**
  - Vote for "matches" on in-frame coding tracks using similarity score.
- **Repeats (RepeatMasker)**
  - Vote against coding exons on repeats
- **Peptide signals (**Predotar**)**
  - Vote for nearby Translation Start.
- **Other predictions** (FGENESH...)
  - May vote for ATG, Stop, SS, CDS...

A general plugin

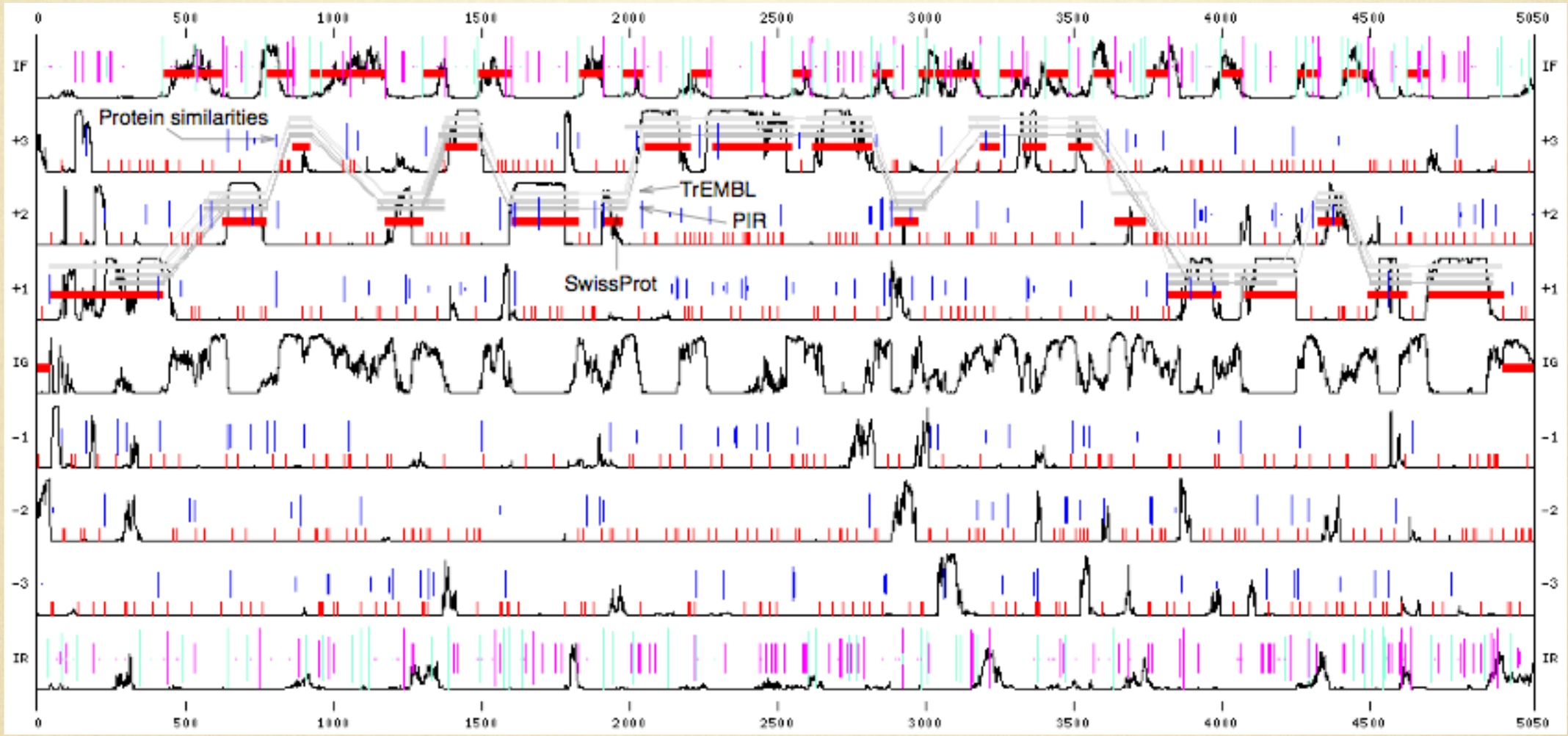
- Annotastruct (GFF3 based)
  - Votes for the desired signals
  - Votes for desired regions or high-level regions (transcribed, translated)
  - Example : Spliced RNA-Seq reads can be used for splice site prediction.

Example

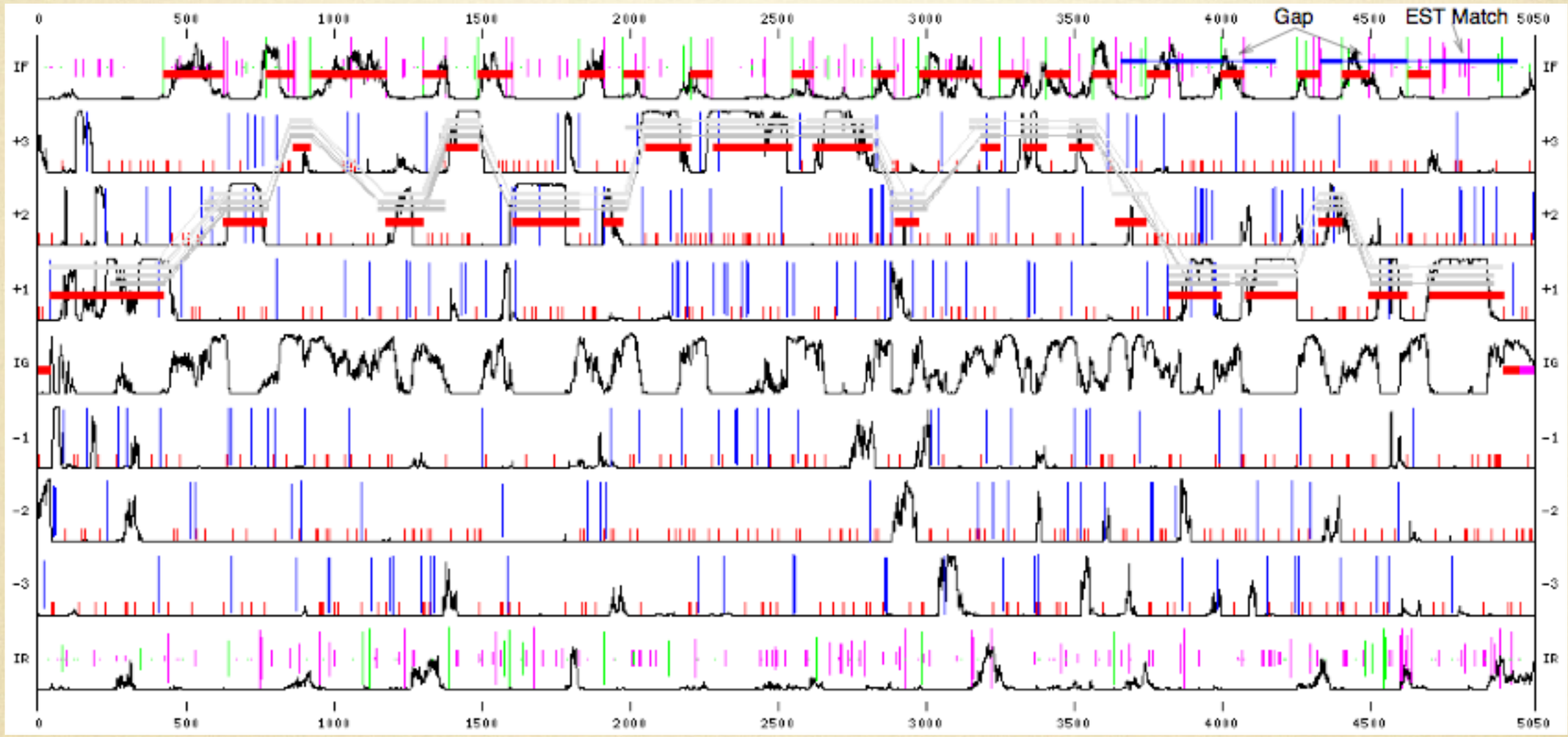

Splice site
Markov models
Prediction
ATG
Stop
IF
+3
+2
+1
IG
-1
-2
-3
IR
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5050

Example

(proteins)



Example (prot.+EST)



S. Foissac et al. NAR 2003

S. Foissac, T. Schiex, BMC Bioinformatics 2005
Foissac et al 2008 Cur Bioinfo.

http://www.inra.fr/bia/T/EuGene/

LETTER

doi:10.1038/nature11119

The tomato genome sequence provides insights into fleshy fruit evolution

ARTICLES

nature biotechnology

Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*

ARTICLES

nature genetics

The genome of *Theobroma cacao*

LETTER

doi:10.1038/nature10625

The *Medicago* genome provides insight into the evolution of rhizobial symbioses

Sequencing centers

Univ. Oklahoma | **TIGR** | **Genoscope** | **Sanger**

Sequence submission to public db

Genbank / EMBL

Once a month or once 50 BACs sequences have been added

Collect sequences and identify additions/modifications

MIPS

First run of analyses

TIGR

+ gene prediction with **FGenesH**

+ spliced alignements with **PASA**

INRA - Toulouse

Second run of analyses, integration and gene prediction

- **Blastx** vs MENS peptide database, ProDom, SP, At
- **SpliceMachine** (Start, Acceptor, Donor sites)
- **Predotar** (Start coding for proteins with peptide signal)
- **SIM4** vs MENS Mt EST Clusters and TIGR Tentative consensus (Mt, Lj, Gm)

Integration and gene prediction with **EuGene**

Protein analyses & description

TIGR

MIPS

Public release of the IMGAG Medicago annotation

A pipeline example

Info Source	Elements	Plugin	
Statitics contents	Exon/UTR/Intron/IG	Int. Markov models	Ab initio 1
Statistics signals	ATG/Acc/Don	Splice Machine	Ab initio 2
Predotar	ATG	PepSignal	
FGenesH Mt	All	Annotastruct	
Mt MENS peptides	Exons	BlastX	Ab initio + BlastX
ProDom	Exons	BlastX	
PASA	All	Annotastruct	Ab initio + proteins +transcripts
Transcripts (legumes)	All	Est	

PASA: Haas et al., NAR 2003

Measuring quality

- Positive: True Positive, False Negative
- Negative: True Negative, False Positive
- Sensitivity = TP/(TP+FN)
- Specificity = TP/(TP+FP)
- Can be applied at the nucleotide level (coding), exon level (precise borders) or gene (all coding exon borders)

Evaluation

Gene predictor	Sens Gene	Spec Gene	Sens Exon	Spec Exon
GenScan	25.8	29	69.6	78
GeneMark.Hmm	32.4	31.6	73.1	76.6
FGenesH At	47	46.5	85.3	81.4
FGenesH Mt	52.8	47.8	85.1	80.7
EGN Ab initio 1	55.5	50.5	84.7	85.4
+FgenesH	63.2	56.4	90	86.9
+AA	69.2	61.8	92.4	88
+AA+EST+PASA	80.2	79.4	94.4	94.6

RNA Seq alone ?

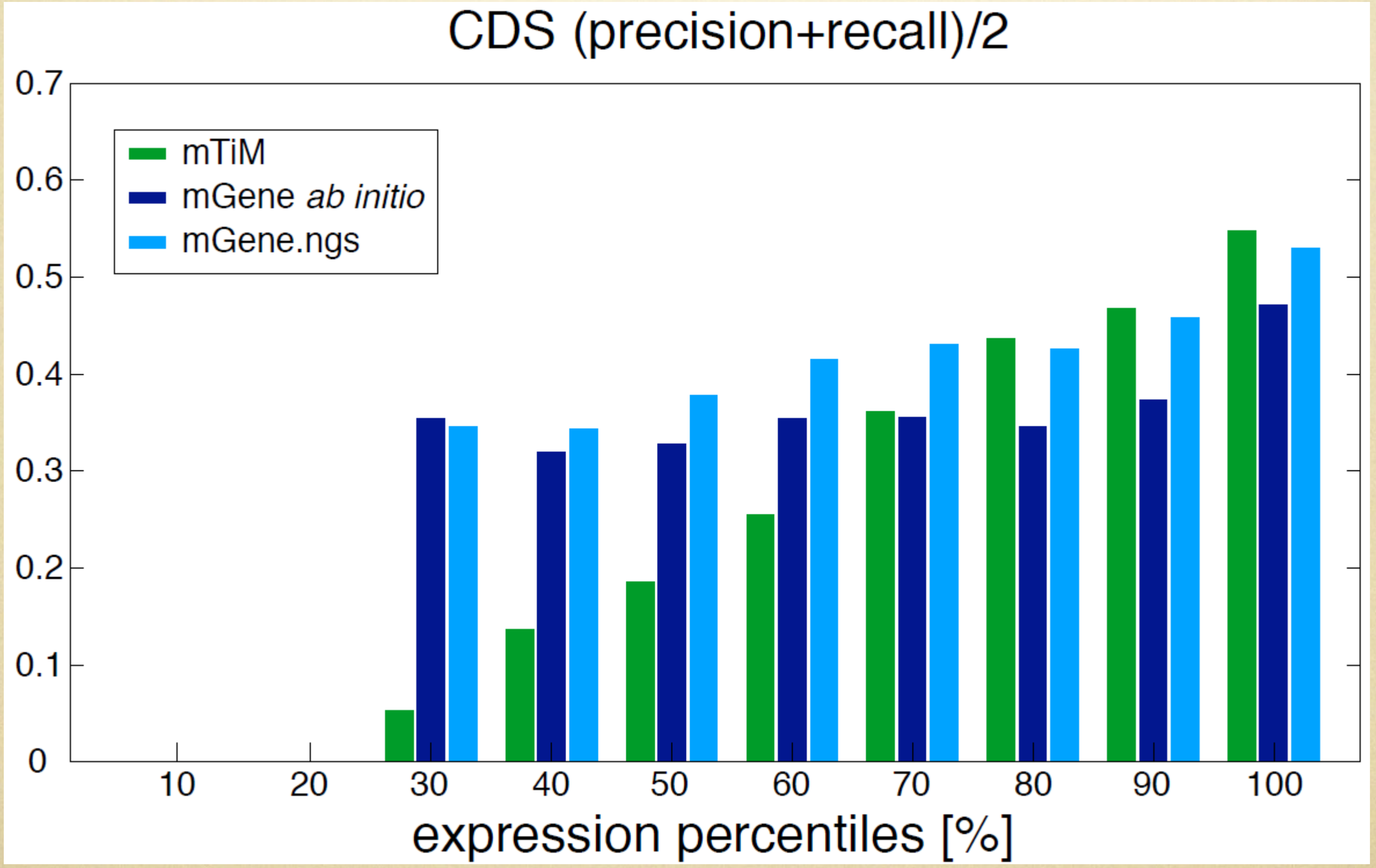


Figure by Gunnar Rätsch (StatSeq 2010)

Conclusion

- Automated gene prediction far from a closed problem:
  - alternative splicing, TSS, regulation, overlapping or imbricated genes,
  - New sequences: draft genomes, sequencing errors (improving quickly)
  - More information: protein tags, PET-Seq...
  - Need : full automatization

For more information

- HMM: Rabiner tutorial (http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf)
- HMM & biology:
  - ADN, mots et modèles (S. Robin, F. Rodolphe et S. Schbath)
  - Biological Sequence Analysis (Durbin, Eddy, Krogh et Mitchison)
- Gene finding:
  - http://linkage.rockfeller.edu/wli/gene
  - Reviews: Mathé et al. NAR 2002, Zhang Nature 2002, Brent Genome research 2005, Brent Nature Biotechnology 2007, Foissac et al 2008 Cur Bioinfo.