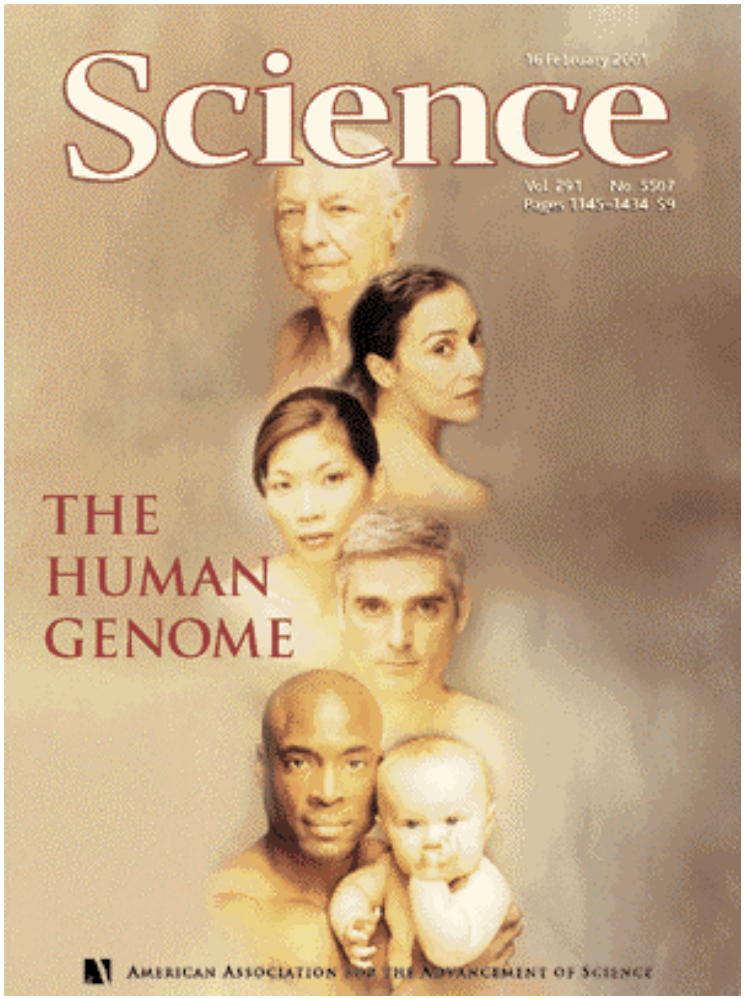
Applications of hidden Markov models to sequence analysis

Lior Pachter

Outline

- Why do we analyze sequences? What are we looking for?
- Annotation of DNA sequences I (and HMMs)
- Alignment
- Annotation of DNA sequences II
- Protein sequences

The Human genome



From the introduction to the Nature human genome paper:

"

- The genomic landscape shows marked variation in the distribution of a number of features…for example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome.
- There appear to be about 30,000-40,000 genes in the human genome- only about twice as many as in the worm or fly
- The full set of proteins encoded in the human is more complex than those of invertebrates….due in part to vertebrate specific protein domains and motifs.
- The pericentromeric and subtelomeric regions of the chromosomes are filled with large recent segmental duiplications of sequence….much more frequent than in yeast, fly or worm.
- More than 1.4 million single nucleotide polymorphisms have been identified.

"

Gene Structure I

DNA - - - - agacgagataaatcgattacagtca - - - -

Transcription

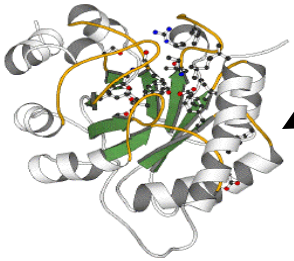
RNA - - - - agacgagauaaaaucgauuacaguca - - - -

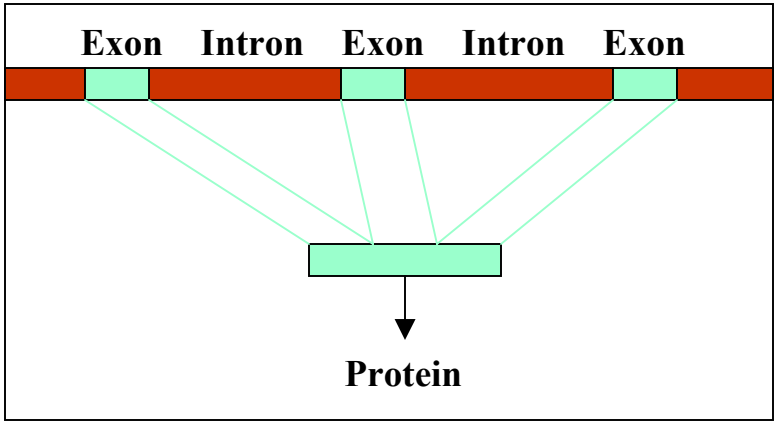
Translation

Protein - - - - - DEI - - - -

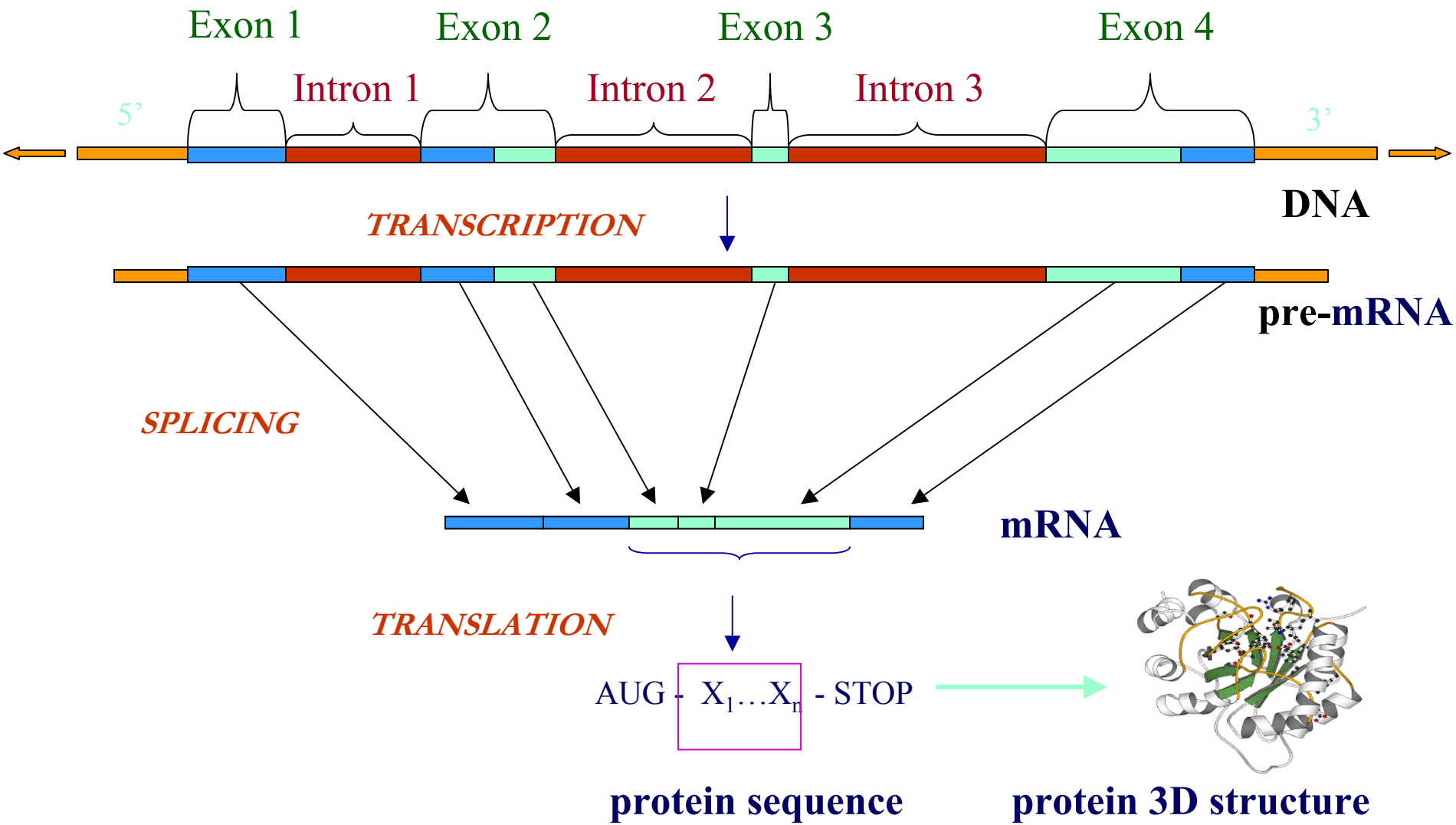
Splicing

Protein Folding Problem

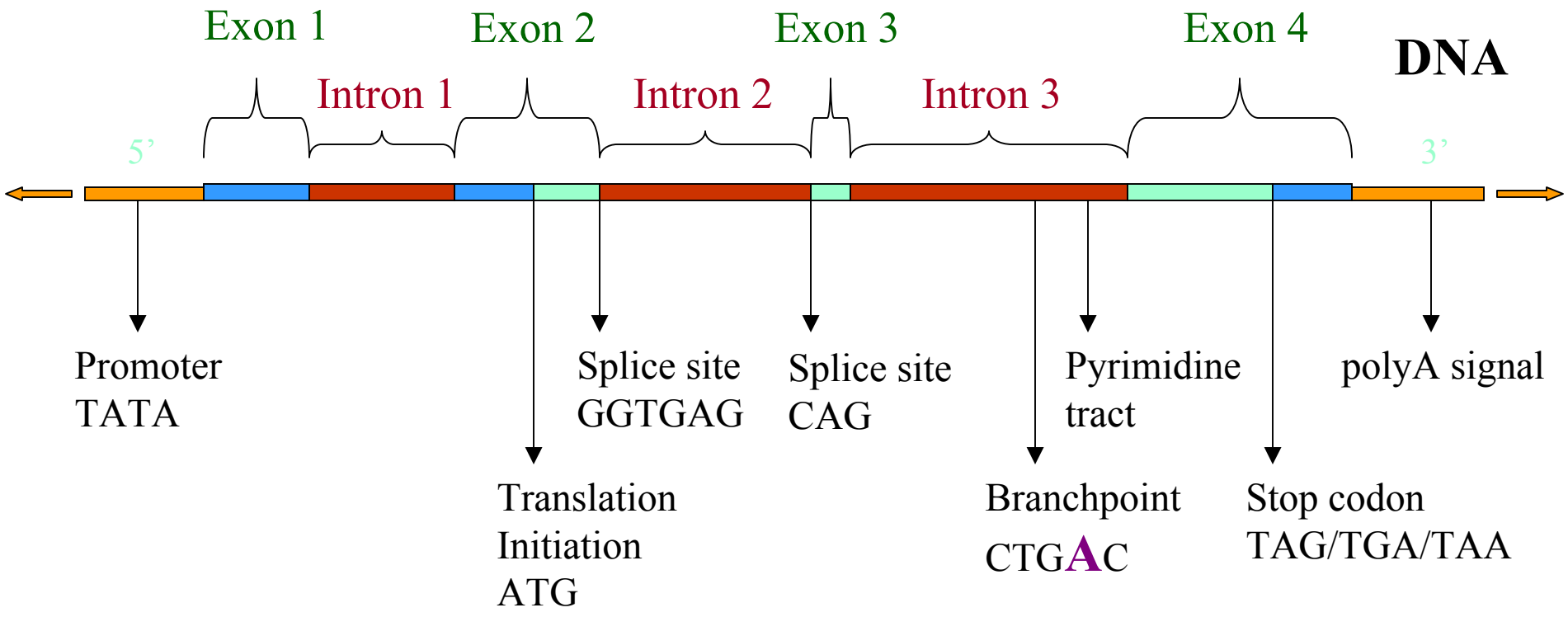




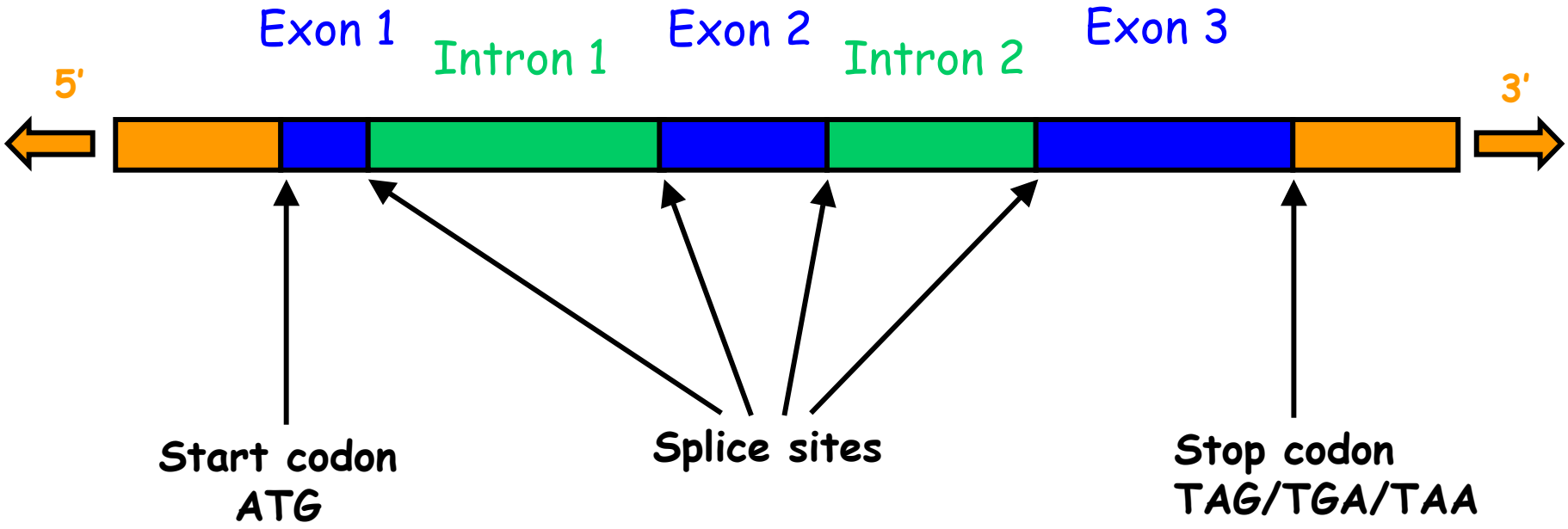
Gene Structure II



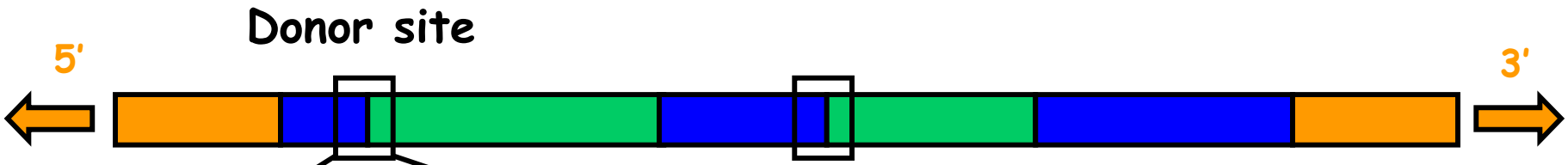
Gene Structure III



Finding genes

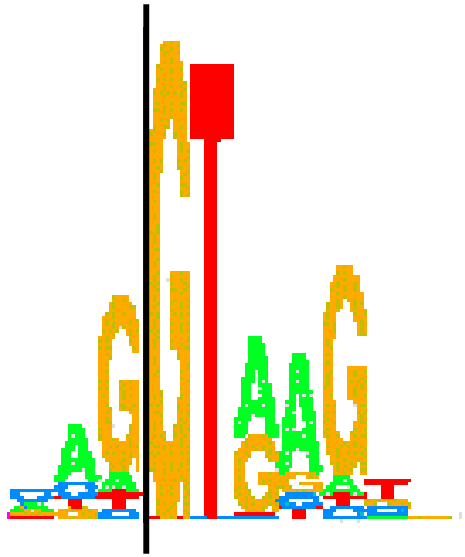


Splice site detection

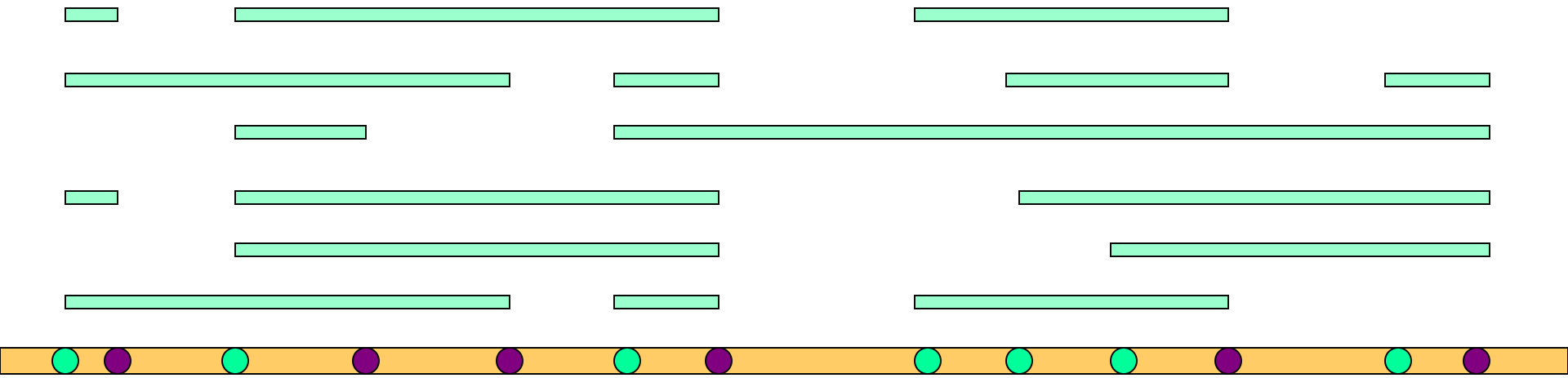


Position

%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	1	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	99	0	41	…	27
T	23	…	13	8	1	98	3	…	25



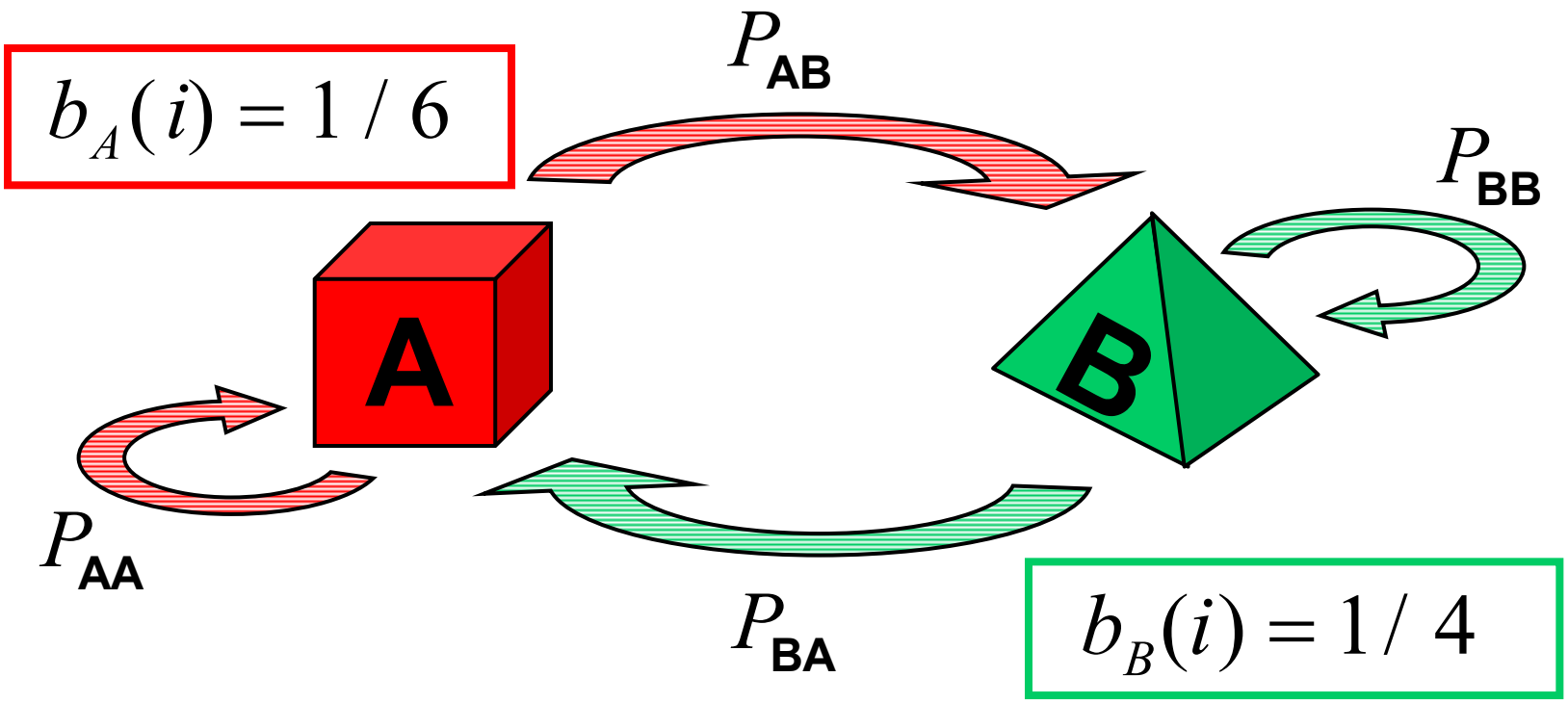
How Difficult is the Problem?



- n = number of acceptor splice sites
- m = number of donor splice sites

Number of parses = F_{n+m+1} (Fibonacci)

A simple HMM

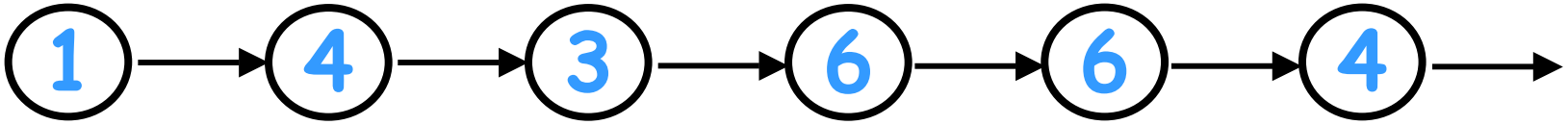


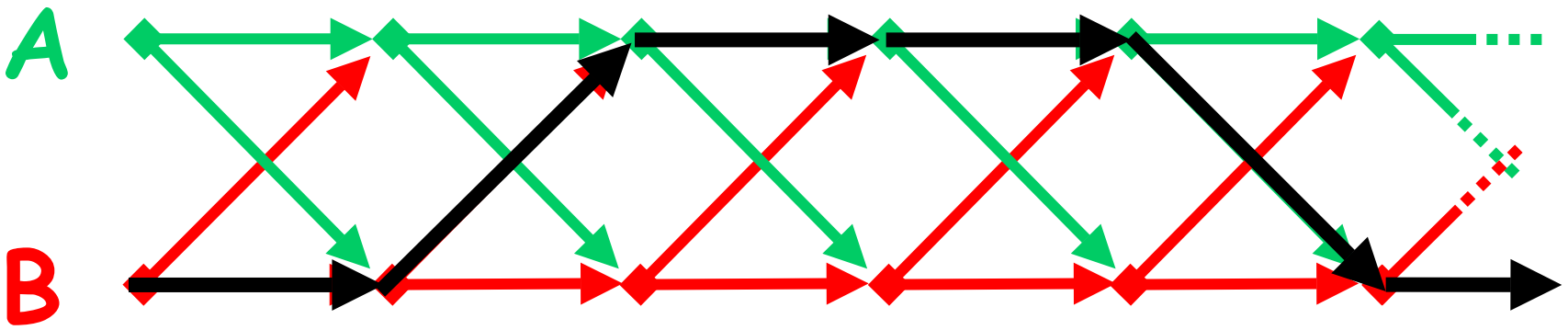
Initial distribution:

$$\pi = (\pi_A, \pi_B)$$

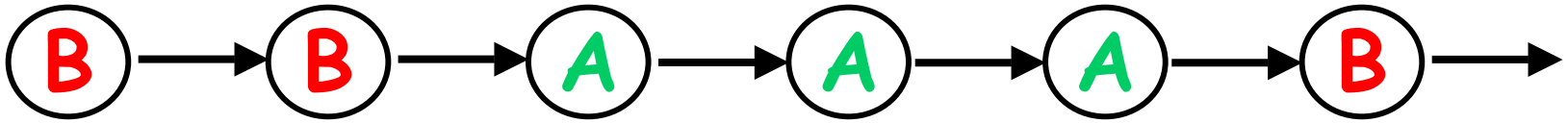
A lattice view

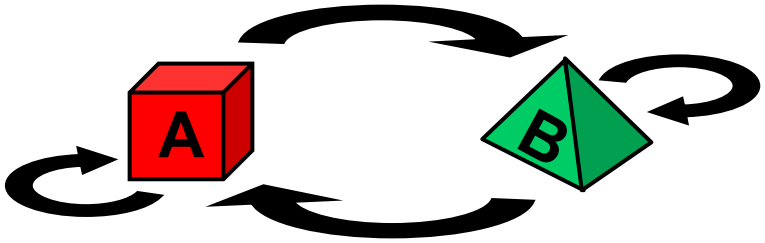
Observed sequence:





Hidden sequence:





Observed: 1,4,3,6,6,4…

Questions:

1. What is the most likely die sequence?
2. What is the probability of the observed sequence?
3. What is the probability that the 3rd state is B, given the observed sequence?

The HMM algorithms

Forward:

$\alpha_t(i)$ = P(observed sequence, ending in state i at base t)

Backward:

$\beta_t(i)$ = P(obs. after t | ending in state i at base t)

Viterbi:

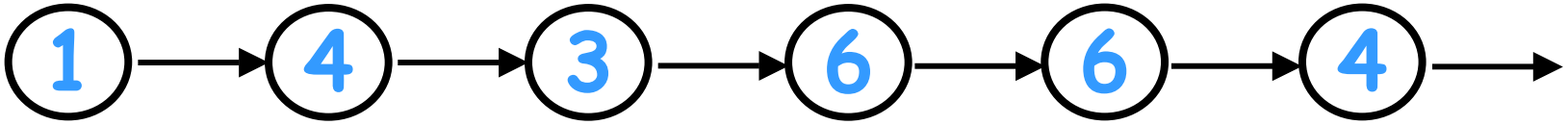
$\delta_t(i)$ = max P(obs., ending in state i at base t)

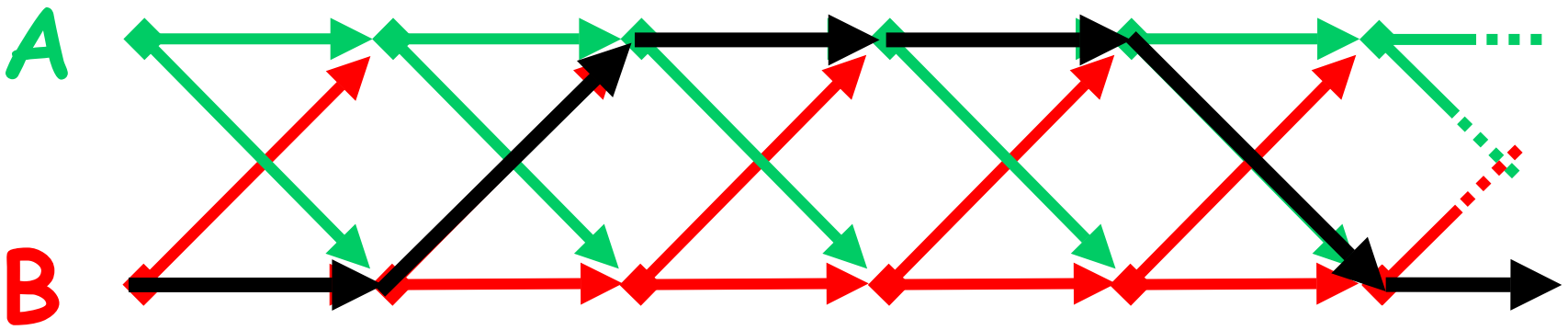
Questions:

1. What is the most likely die sequence? Viterbi
2. What is the probability of the observed sequence? Forward
3. What is the probability that the 3rd state is B, given the observed sequence? Backward

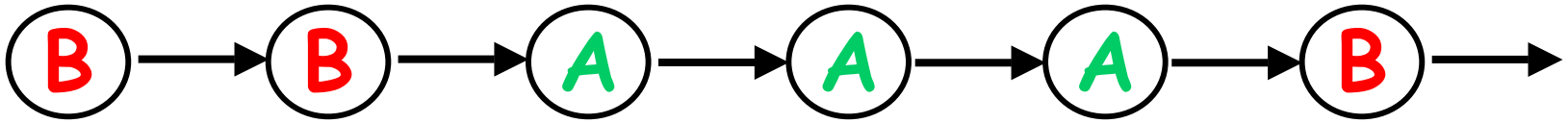
A lattice view

Observed sequence:





Hidden sequence:

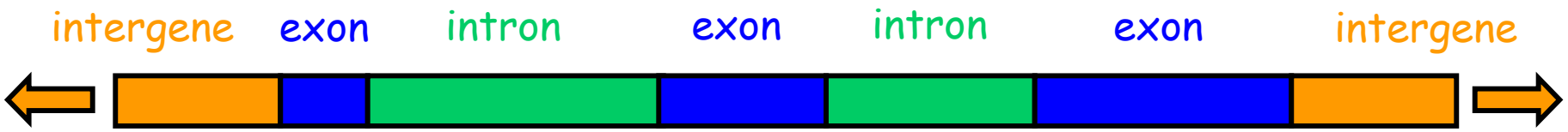


Hidden Markov Models (HMMs)

- Underlying generates a sequence of states.

  Markov chain = distribution of next state depends only on present

  Hidden = the state sequence



Observed = outputs from the states

GTCAGAGTAGCAAAGTAGACACTCCAGTAACGC

Approaches to Gene recognition

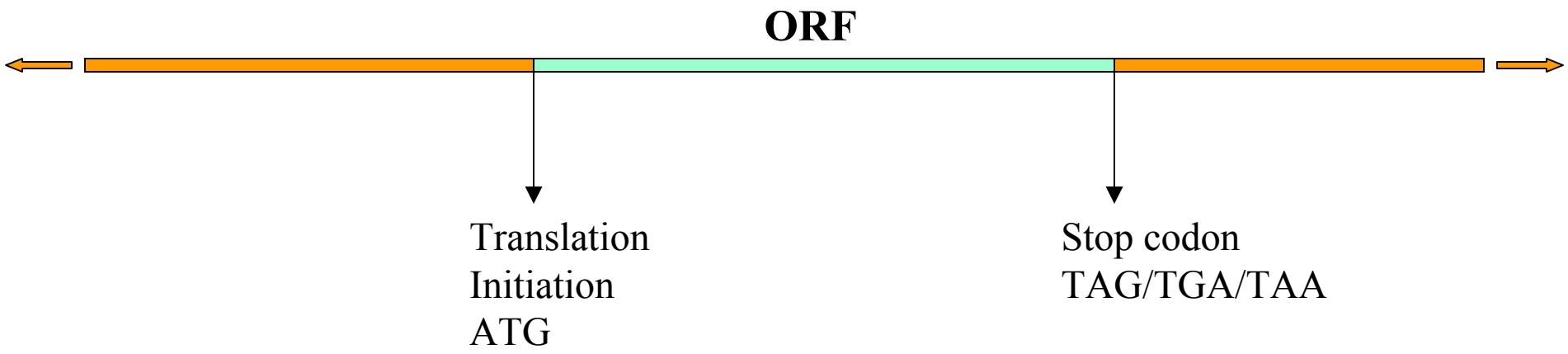
- Homology
  - BLAST, Procrustes
- De Novo
  - GRAIL, FGENEH, GENSCAN, Genie, Glimmer
- Hybrids
  - GenomeScan, Genie
- Comparative
  - Rosetta, Twinscan

Ab-initio gene finding: Generalized HMMs

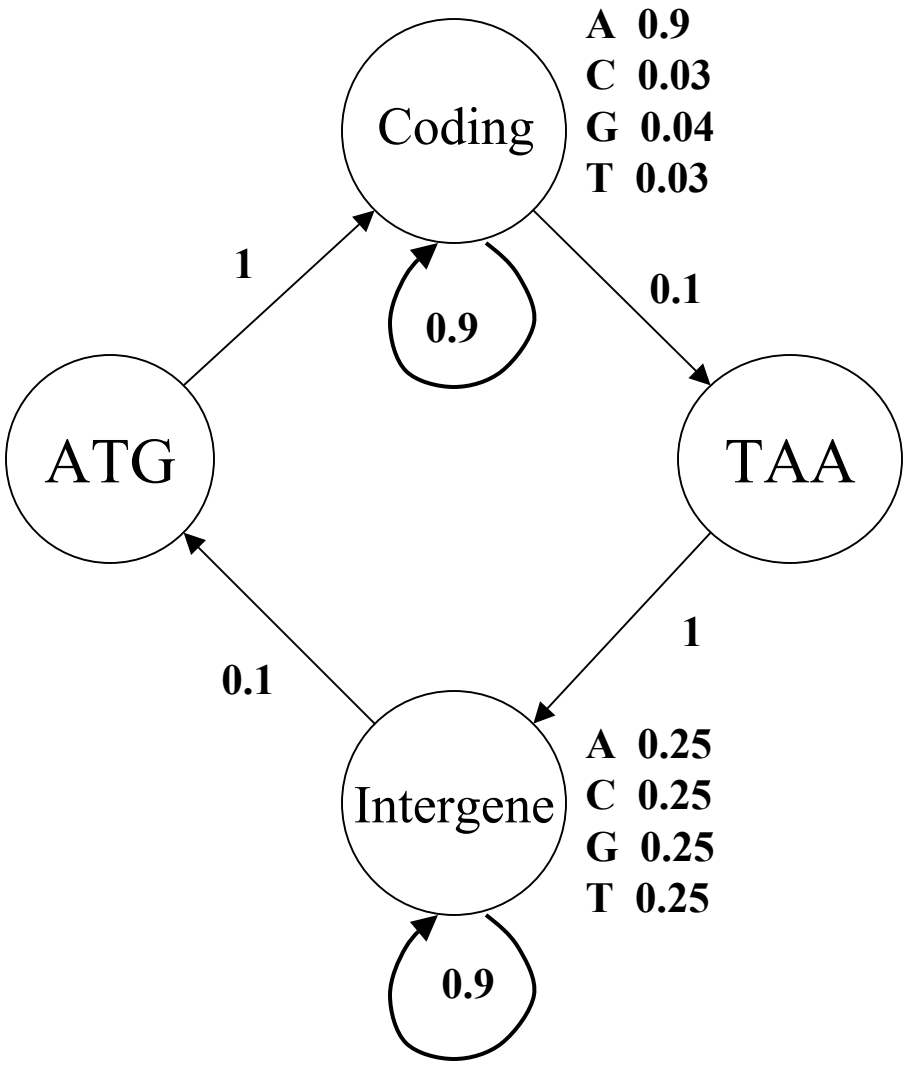
Example: Glimmer
Gene Finding in Microbial DNA

- No introns
- 90% coding
- Shorter genomes (less than 10 million bp)
- Lots of data

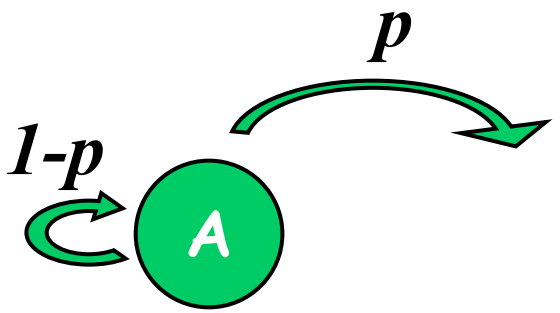
Gene Structure in Prokaryotes



Bacteriomaker (Walmart $3.95)

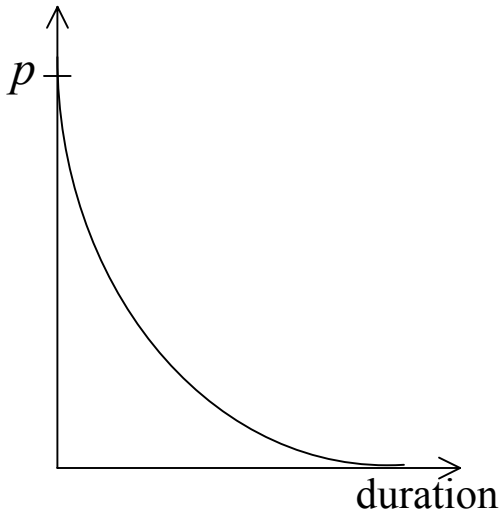


HMM state duration times

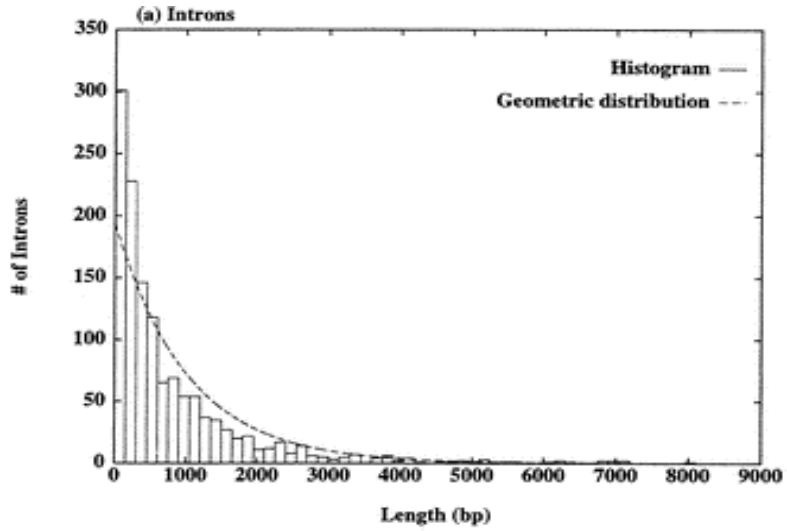


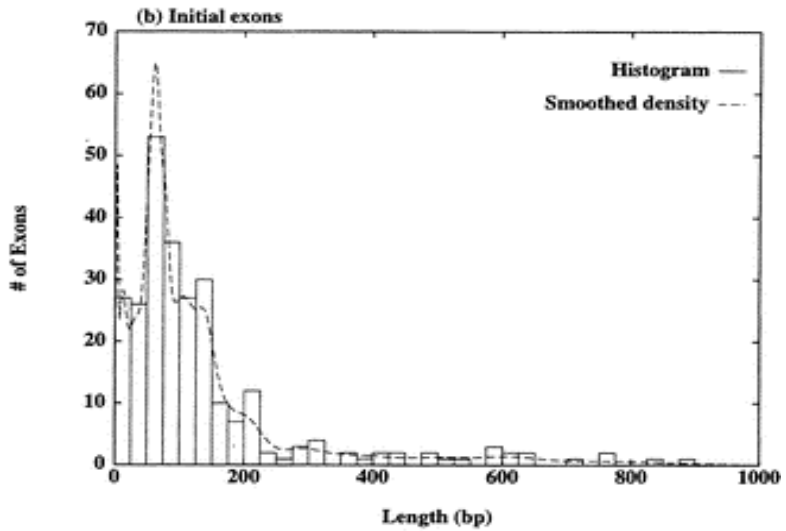
- Pr(leaving state) = p
- Pr(staying in state) = 1 - p
- Pr(output of exactly *r* in state) = $(1-p)^r p$

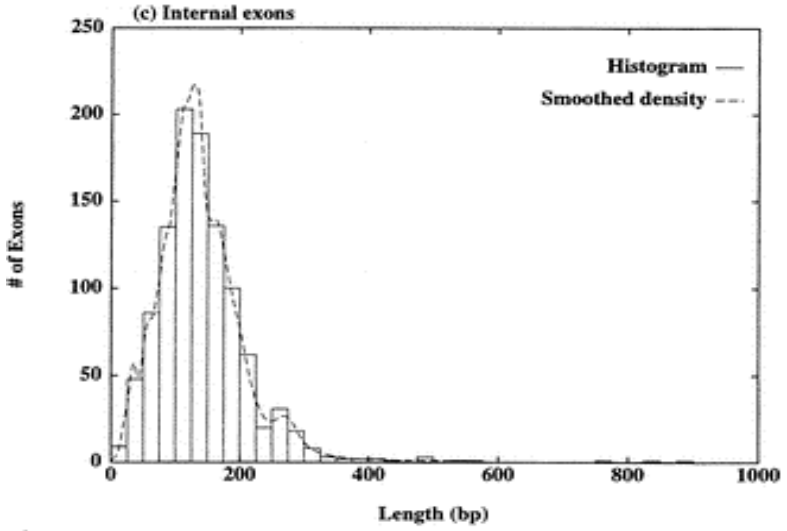
- Geometric distribution

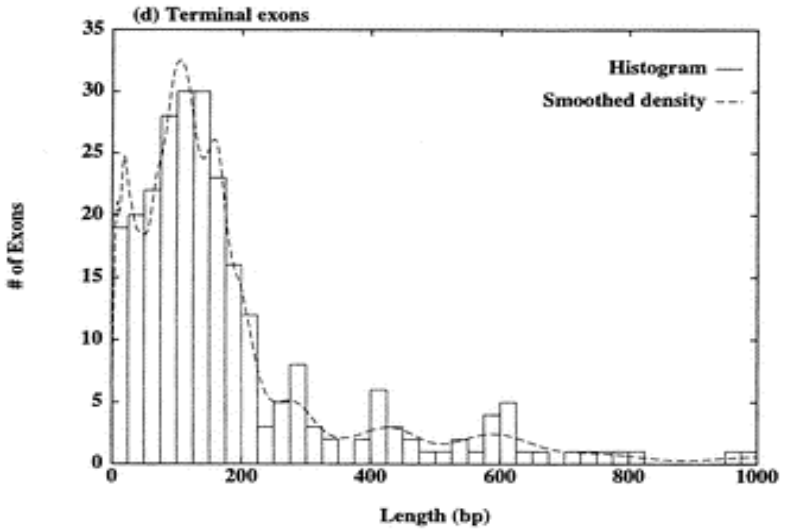


Observed duration times

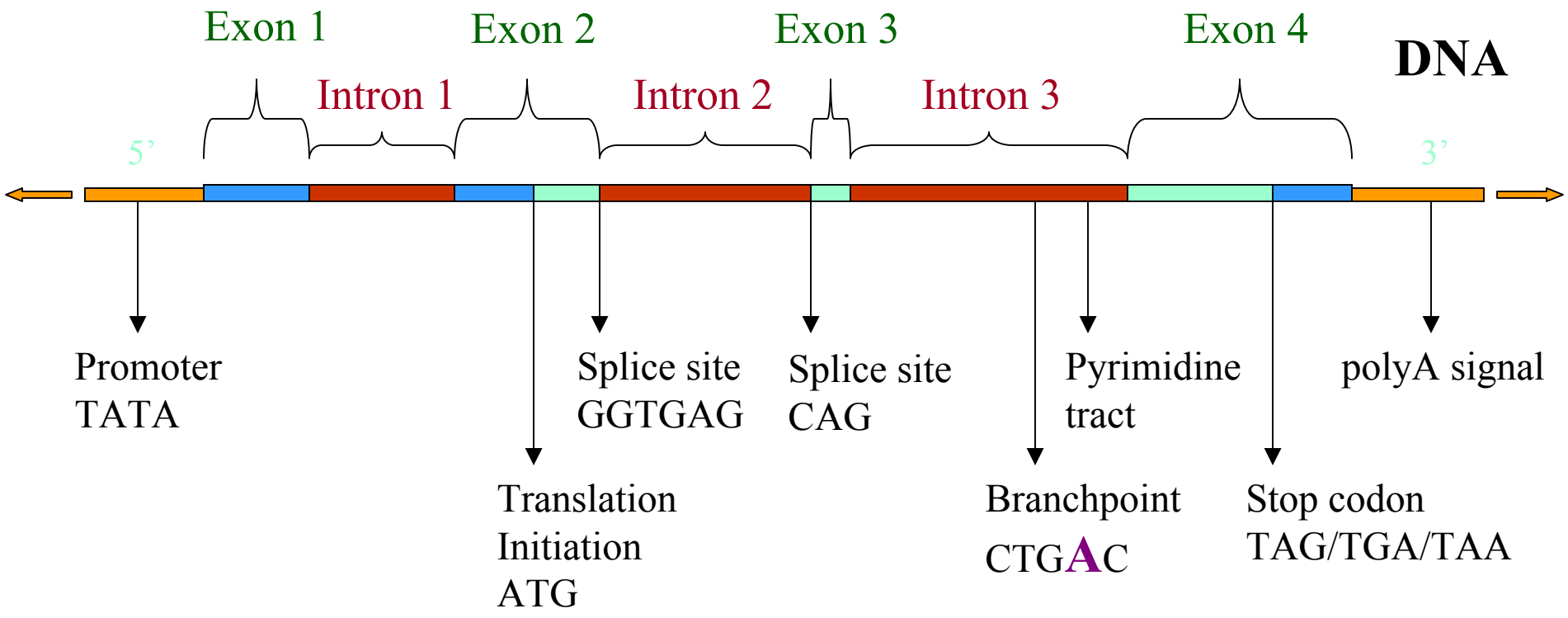




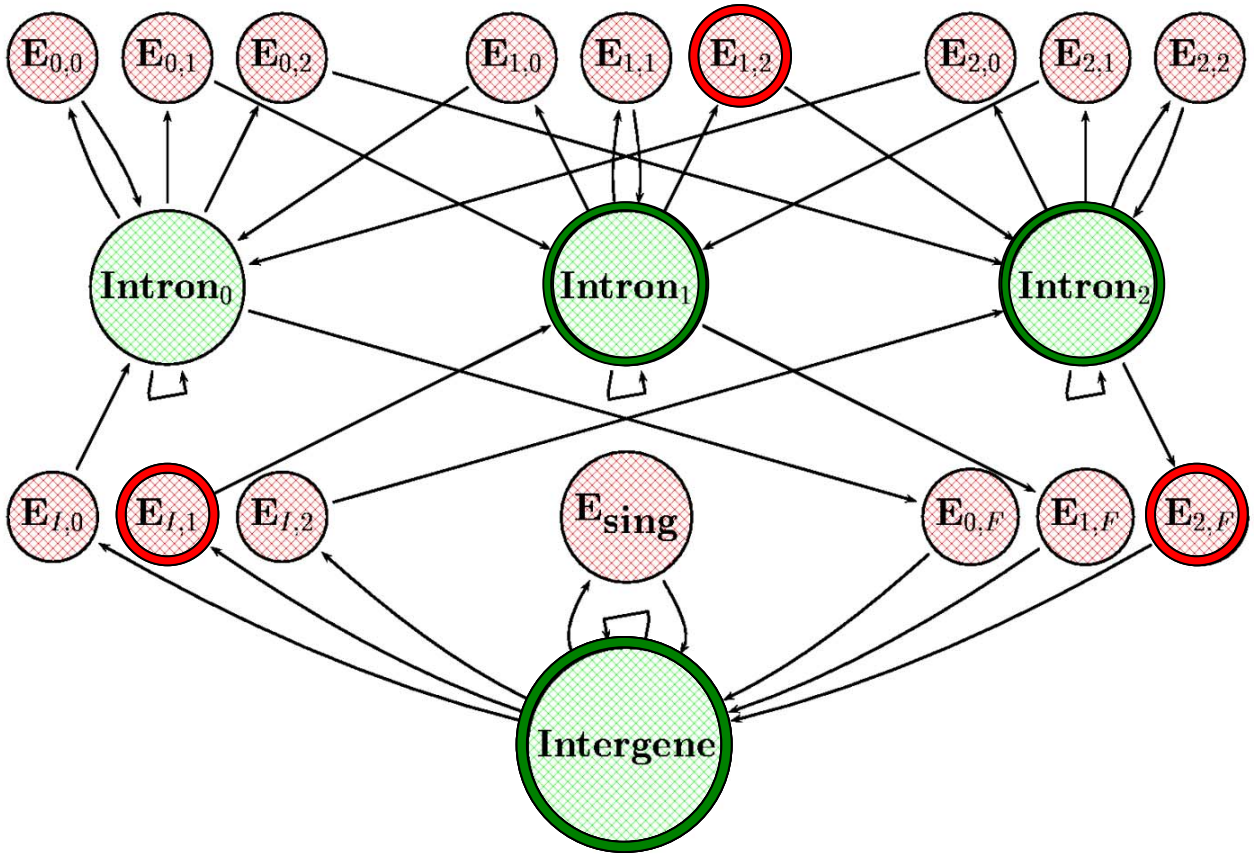




The Gene Finding Problem




TAAT
ATGTCCACGG
GTATTGAG
CATTGTACACGGG
GTATTGAG
CATGTAA
TGAA


$E_{0,0}$
$E_{0,1}$
$E_{0,2}$
$E_{1,0}$
$E_{1,1}$
$E_{1,2}$
$E_{2,0}$
$E_{2,1}$
$E_{2,2}$
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

Using GHMMs for ab-initio gene finding

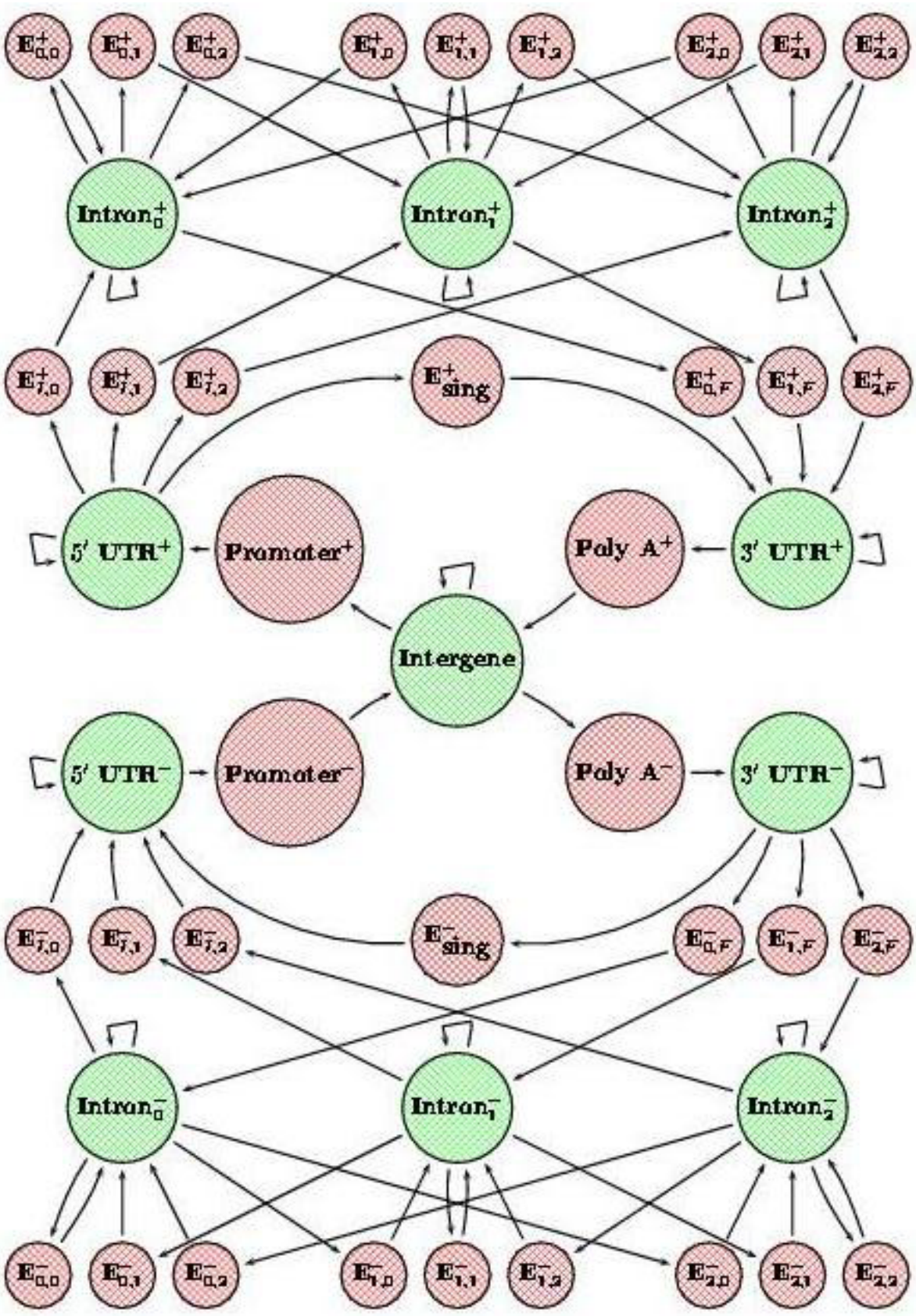
In practice, have observed sequence

TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

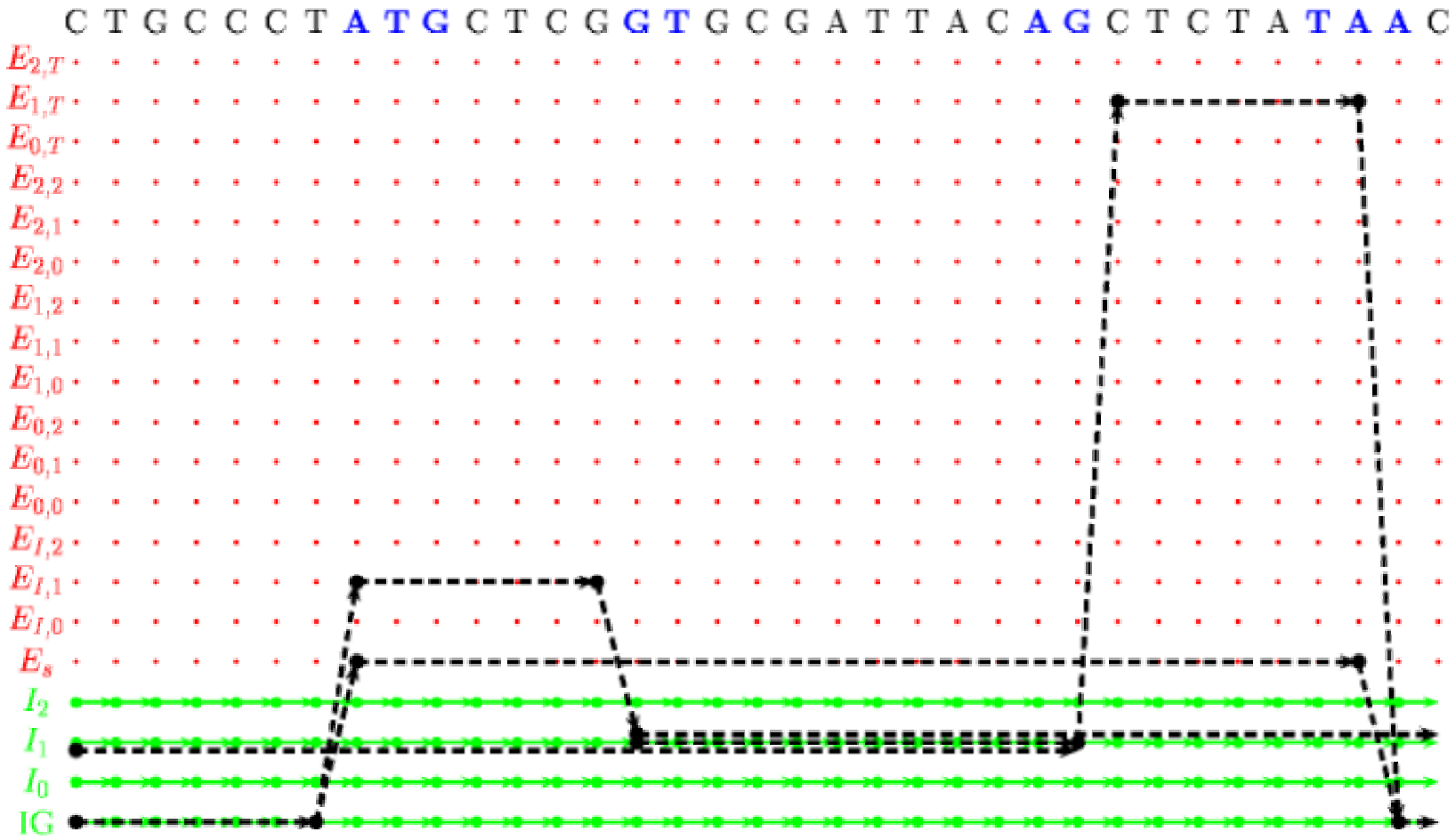
Predict genes by estimating hidden state sequence

Usual solution: single most likely sequence of hidden states (Viterbi).

The Genscan HMM

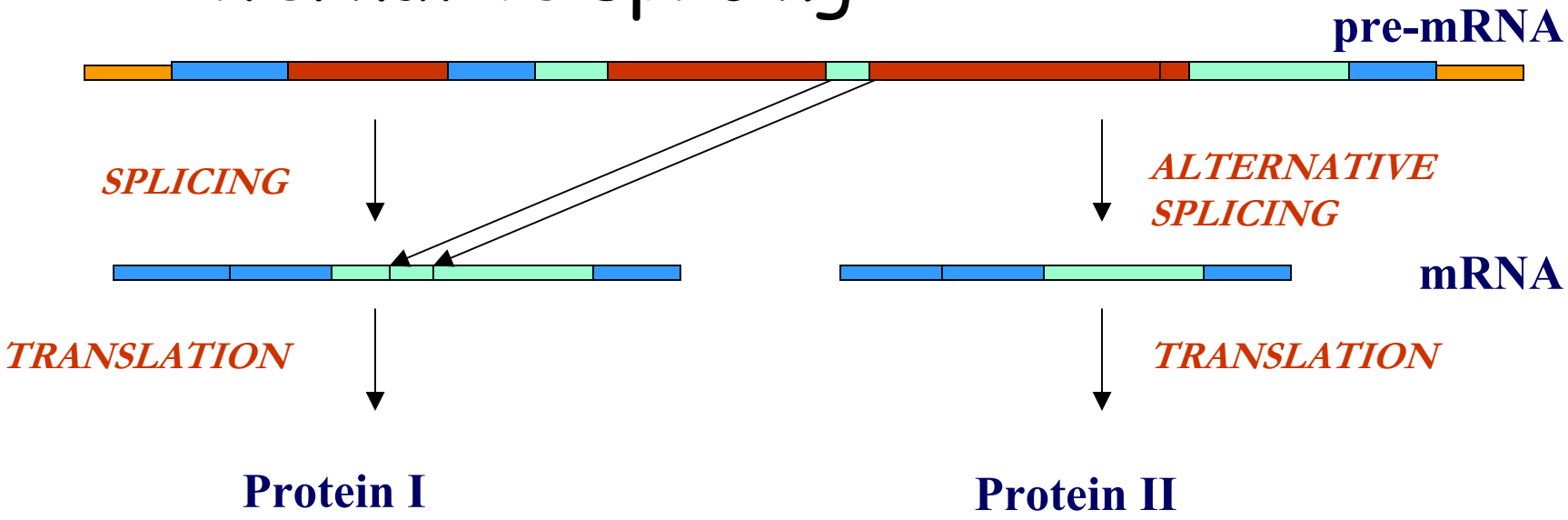


Lattice view

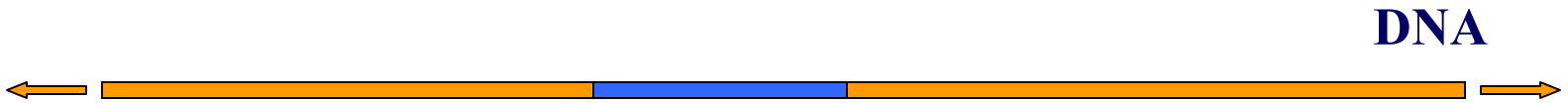


Life is complicated

- Alternative splicing



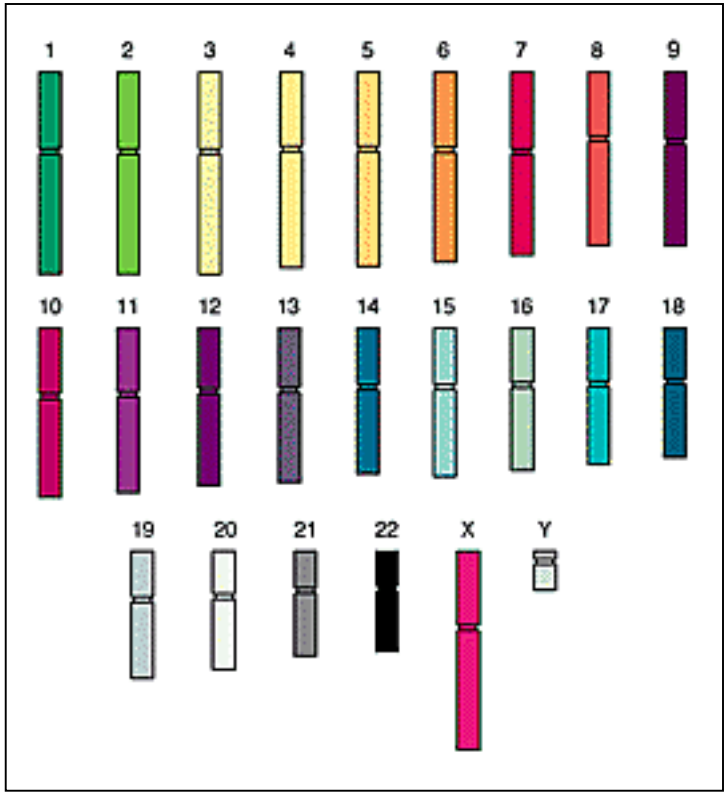
- Pseudo genes



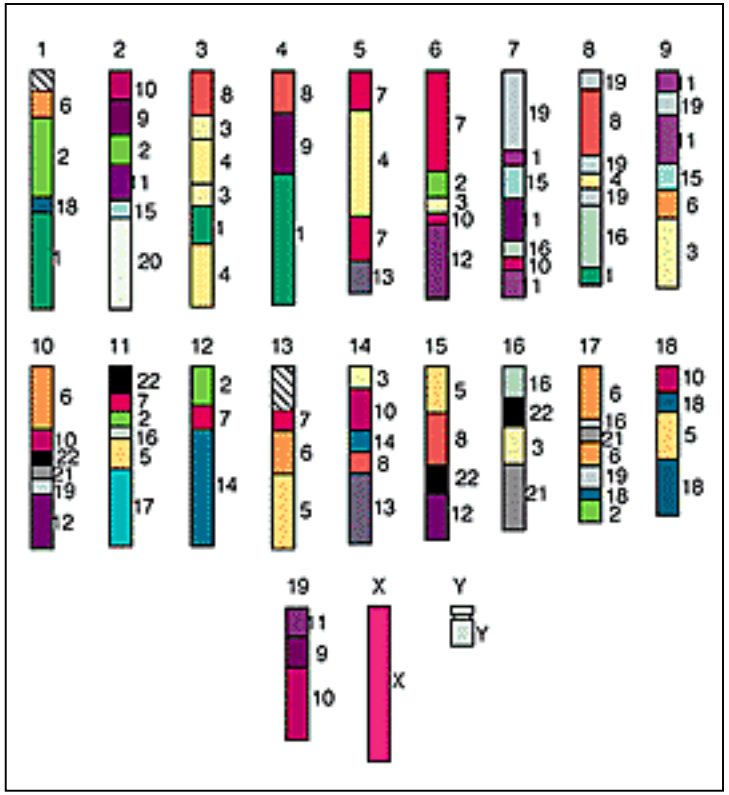
Alignment

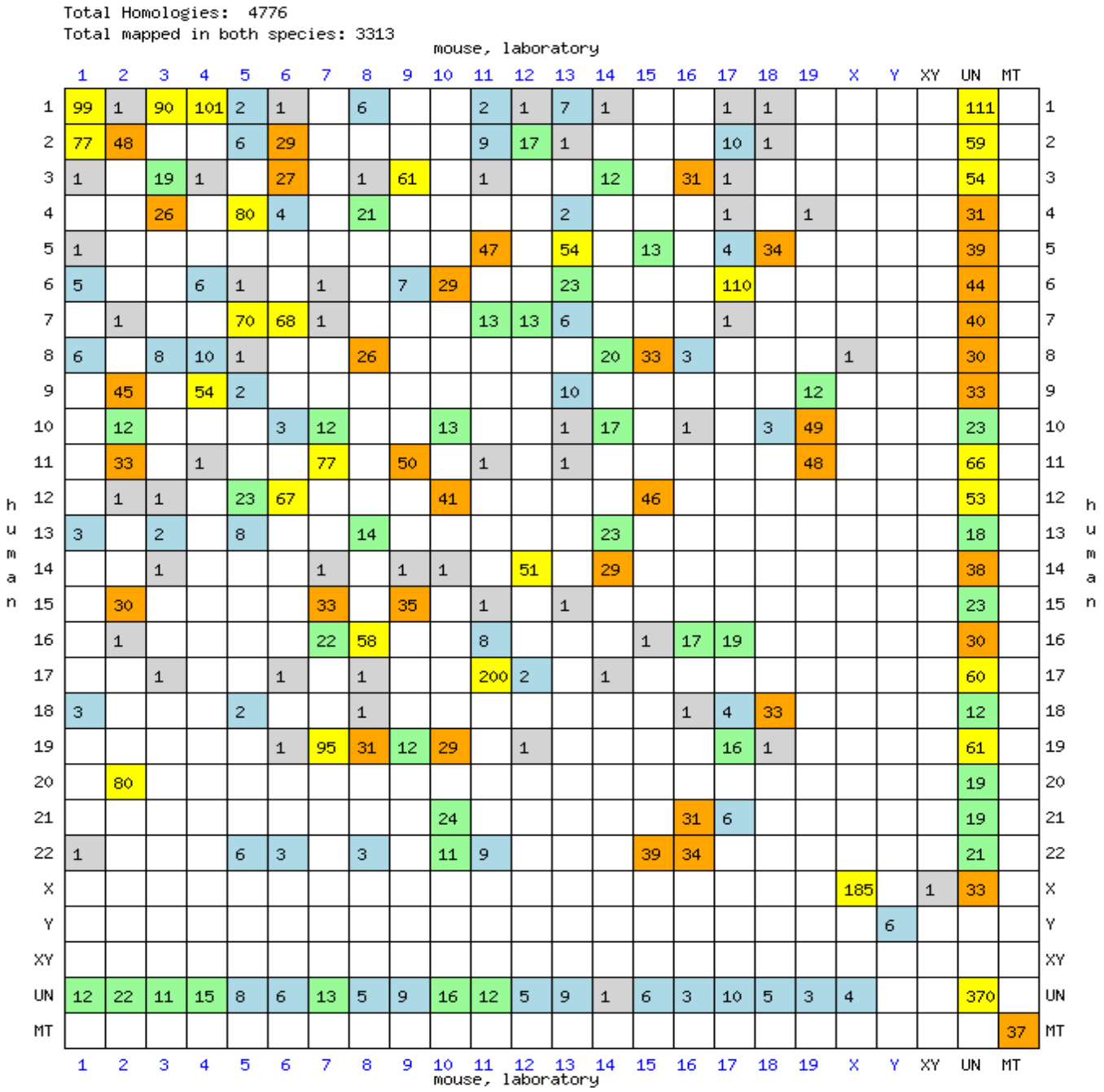
Chromosome Comparison

Human



Mouse




Total Homologies: 4776
Total mapped in both species: 3313
mouse, laboratory
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y XY UN MT
1 99 1 90 101 2 1 6 2 1 7 1 1 1 111
2 77 48 6 29 9 17 1 10 1 59
3 1 19 1 27 1 61 1 12 31 1 54
4 26 80 4 21 2 1 1 31
5 1 47 54 13 4 34 39
6 5 6 1 1 7 29 23 110 44
7 1 70 68 1 13 13 6 1 40
8 6 8 10 1 26 20 33 3 1 30
9 45 54 2 10 12 33
10 12 3 12 13 1 17 1 3 49 23
11 33 1 77 50 1 1 48 66
12 1 1 23 67 41 46 53
13 3 2 8 14 23 18
14 1 1 1 1 51 29 38
15 30 33 35 1 1 23
16 1 22 58 8 1 17 19 30
17 1 1 200 2 1 60
18 3 2 1 1 4 33 12
19 1 95 31 12 29 1 16 1 61
20 80 19
21 24 31 6 19
22 1 6 3 3 11 9 39 34 21
X 185 1 33
Y 6
XY
UN 12 22 11 15 8 6 13 5 9 16 12 5 9 1 6 3 10 5 3 4 370
MT 37
human
mouse, laboratory

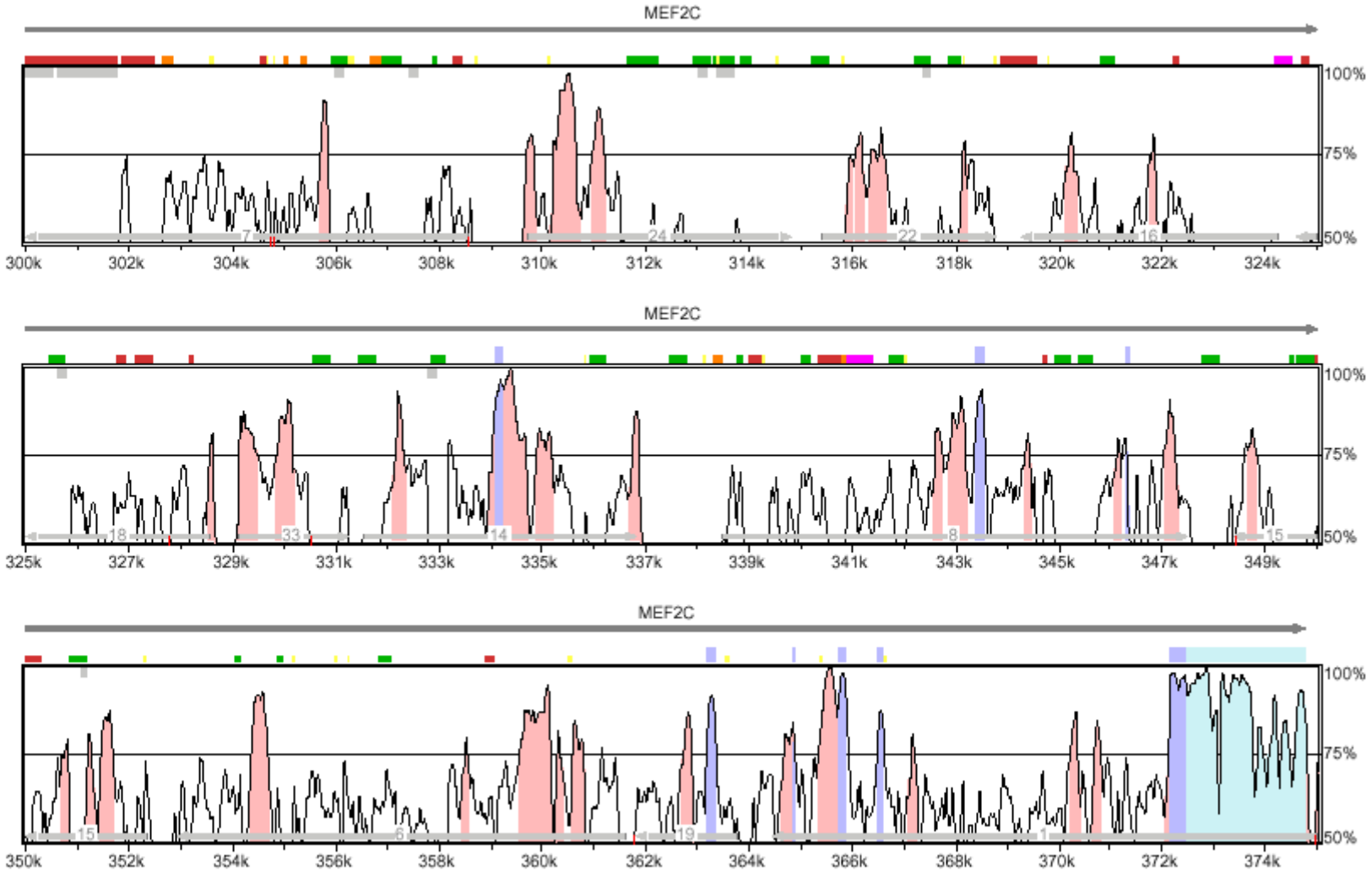
MEF2C

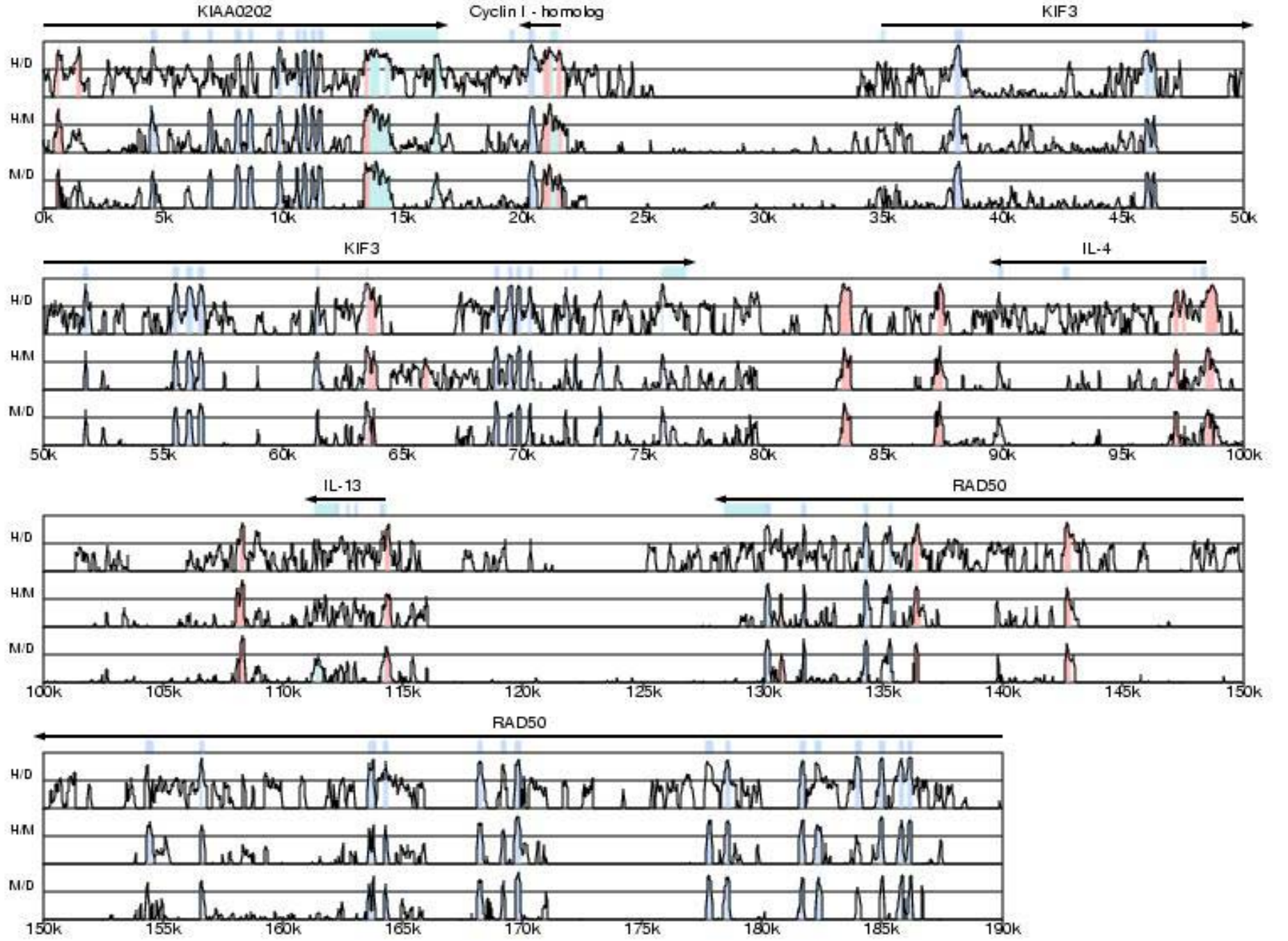
```
Alignment 1
 Seq1: human
 Seq2: mouse
 Reg id: 80
 Reg length: 100
 Plot min: 50
 Regions: 103

X-axis: human
Resolution: 39
Window size: 100
Min gap: 100
```

- Contig
- Gene
- Exon
- UTR
- CNS
- Gap in seq1
- Gap in seq2

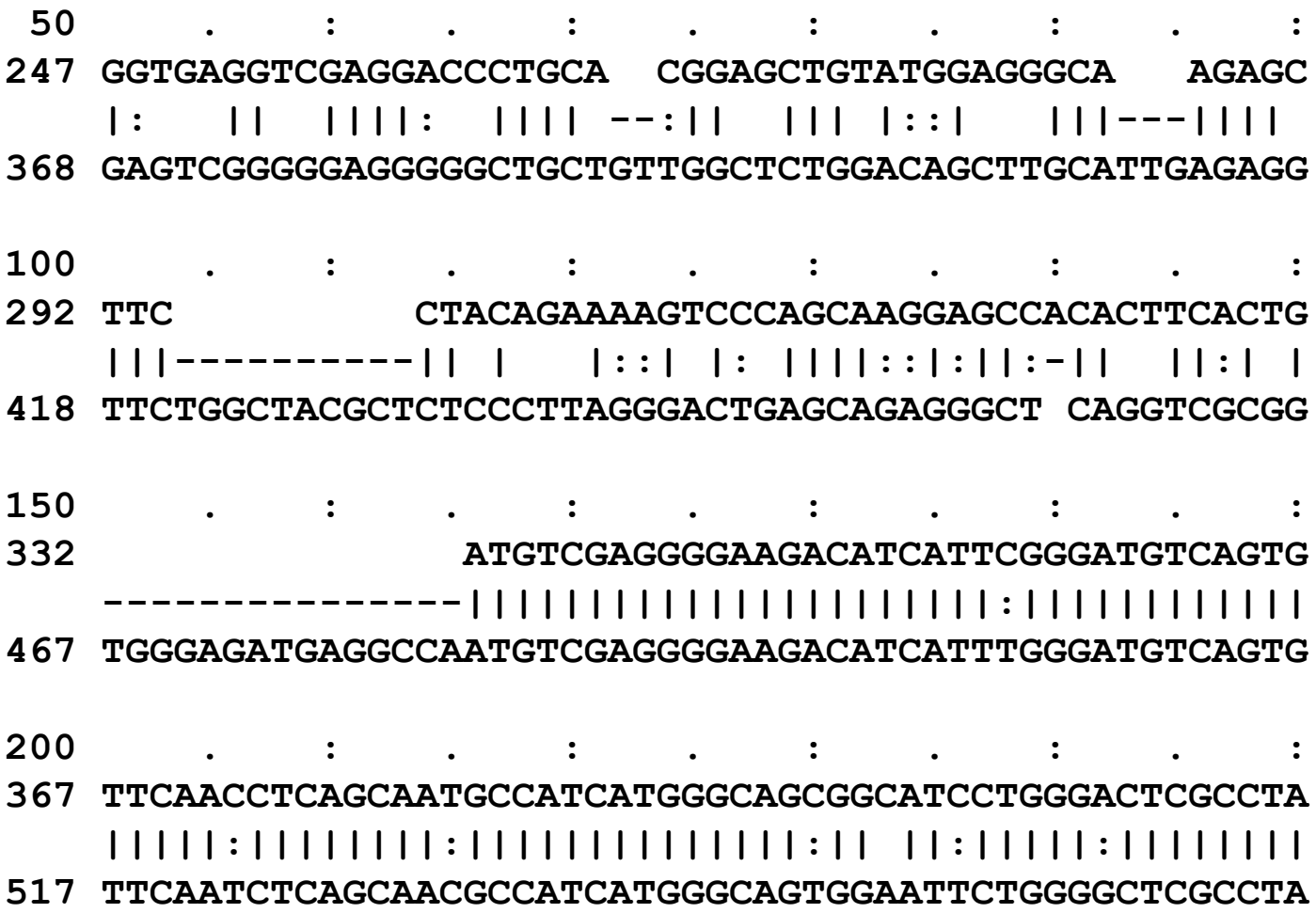
Repeats:

- LINE
- LTR
- SINE
- RNA
- DNA
- Other




KIAA0202
Cyclin I - homolog
KIF3
H/D
H/M
M/D
0k
5k
10k
15k
20k
25k
30k
35k
40k
45k
50k
KIF3
IL-4
50k
55k
60k
65k
70k
75k
80k
85k
90k
95k
100k
IL-13
RAD50
100k
105k
110k
115k
120k
125k
130k
135k
140k
145k
150k
RAD50
150k
155k
160k
165k
170k
175k
180k
185k
190k

Pair HMMs


50 . : . : . : . : . :
247 GGTGAGGTCGAGGACCCTGCA CGGAGCTGTATGGAGGGCA AGAGC
|: || |||| : |||| --:|| ||| |::| |||---||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
100 . : . : . : . : . :
292 TTC CTACAGAAAAGTCCCAGCAAGGAGCCACACTTCACTG
|||---------|| | |::| |: ||||::|:||:-|| ||:| |
418 TTCTGGCTACGCTCTCCCTTAGGGACTGAGCAGAGGGCT CAGGTCGCGG
150 . : . : . : . : . :
332 ATGTCGAGGGGAAGACATCATTCGGGATGTCAGTG
----------|||||||||||||||||||:||||||||||
467 TGGGAGATGAGGCCAATGTCGAGGGGAAGACATCATTTGGGATGTCAGTG
200 . : . : . : . : . :
367 TTCAACCTCAGCAATGCCATCATGGGCAGCGGCATCCTGGGACTCGCCTA
|||||:||||||||:||||||||||||||:|| ||:|||||:||||||||
517 TTCAATCTCAGCAACGCCATCATGGGCAGTGGAATTCTGGGGCTCGCCTA

Alignment Formalization

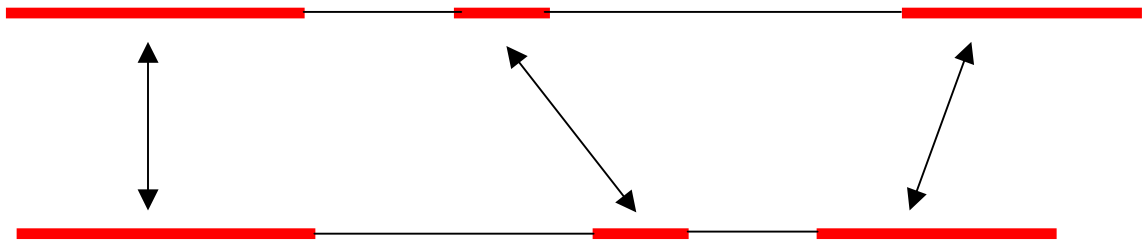
"…consider a pair of strings on a finite alphabet…

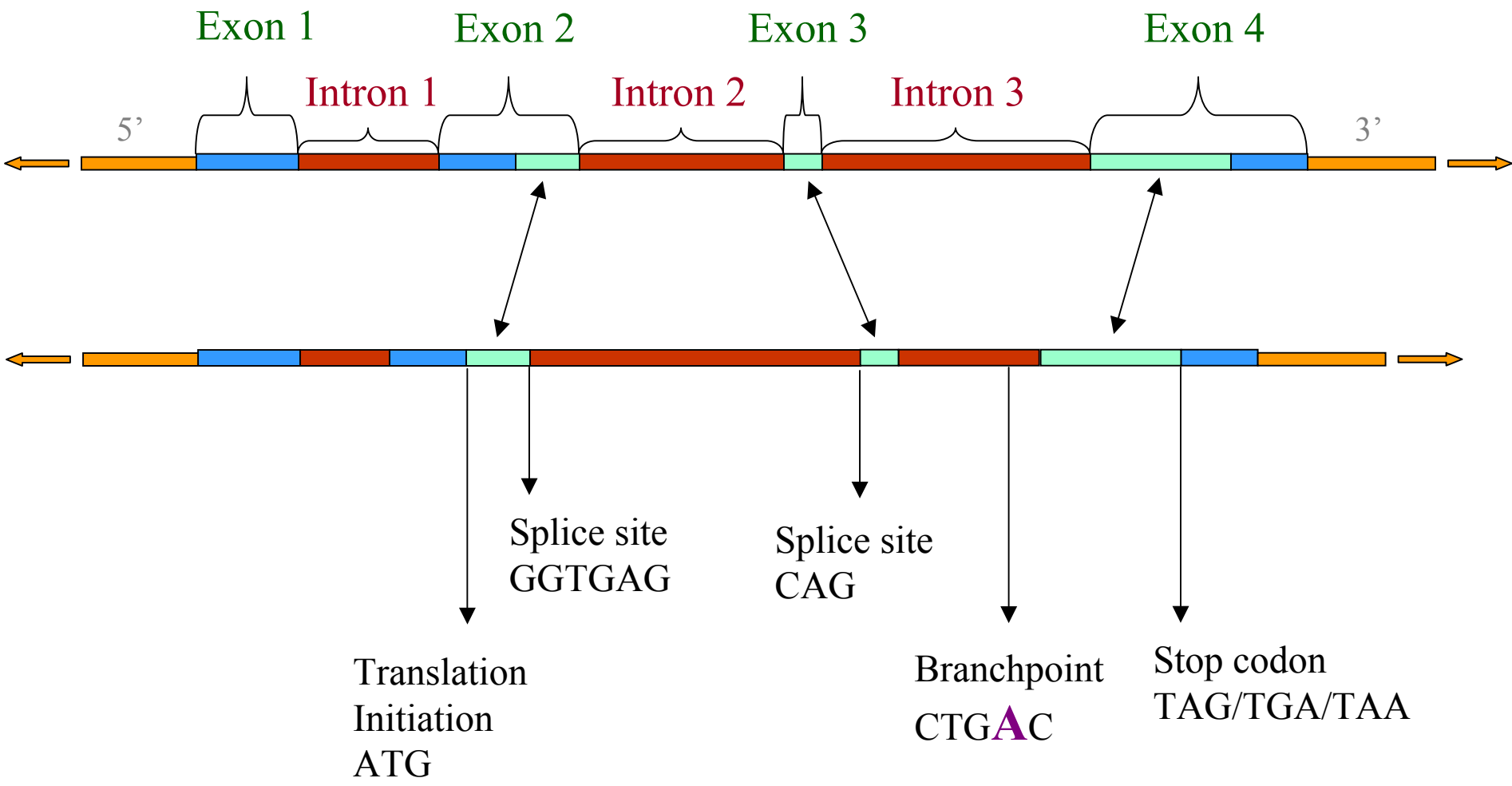
"…an alignment is a string of match/mismatch/indel symbols…"

"…we show how to find the optimal alignment where the scoring function is given by…"

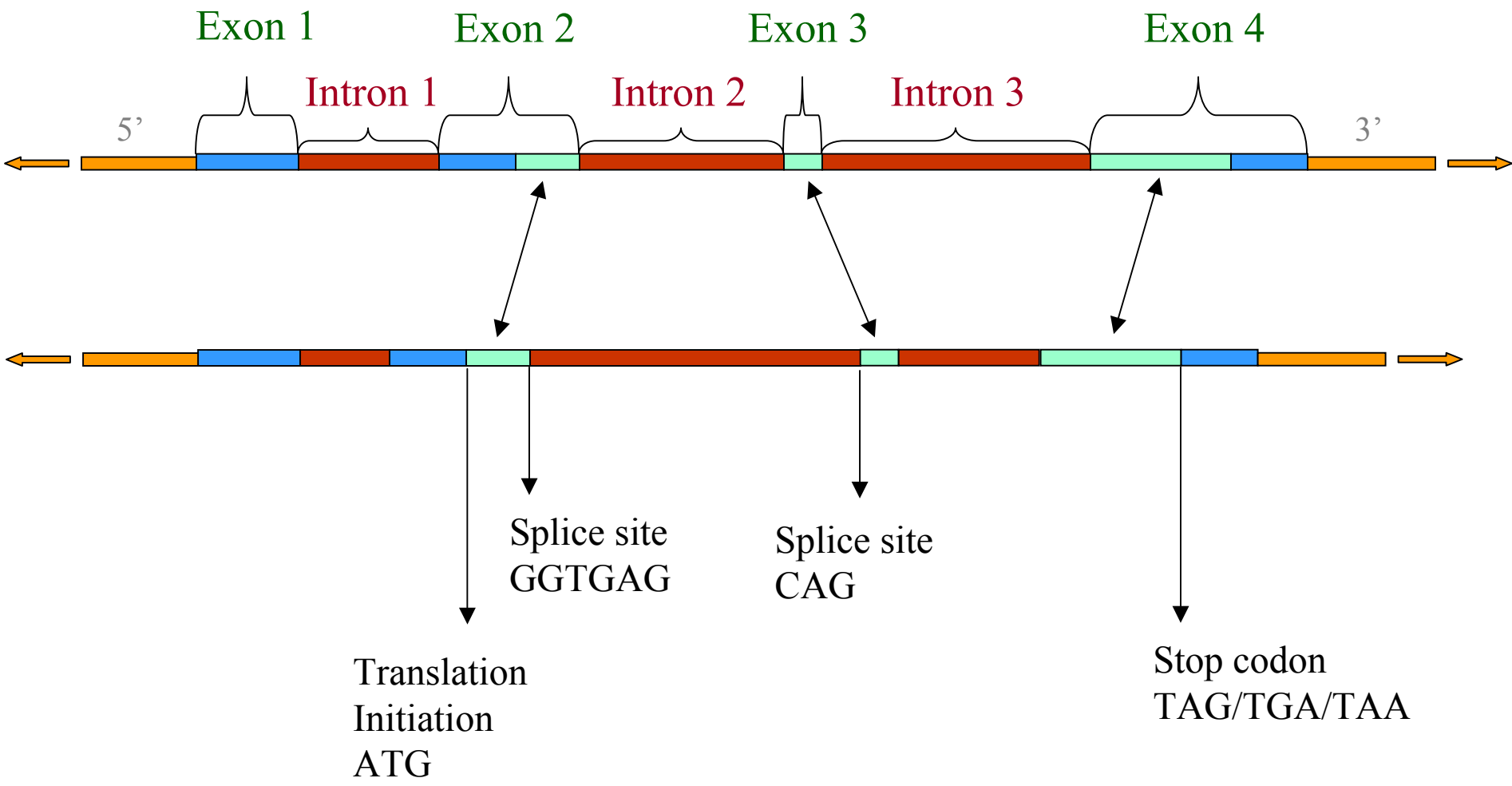
Want to take into account that the sequences are **genome** sequences:

Example: a pair of syntenic genomic regions

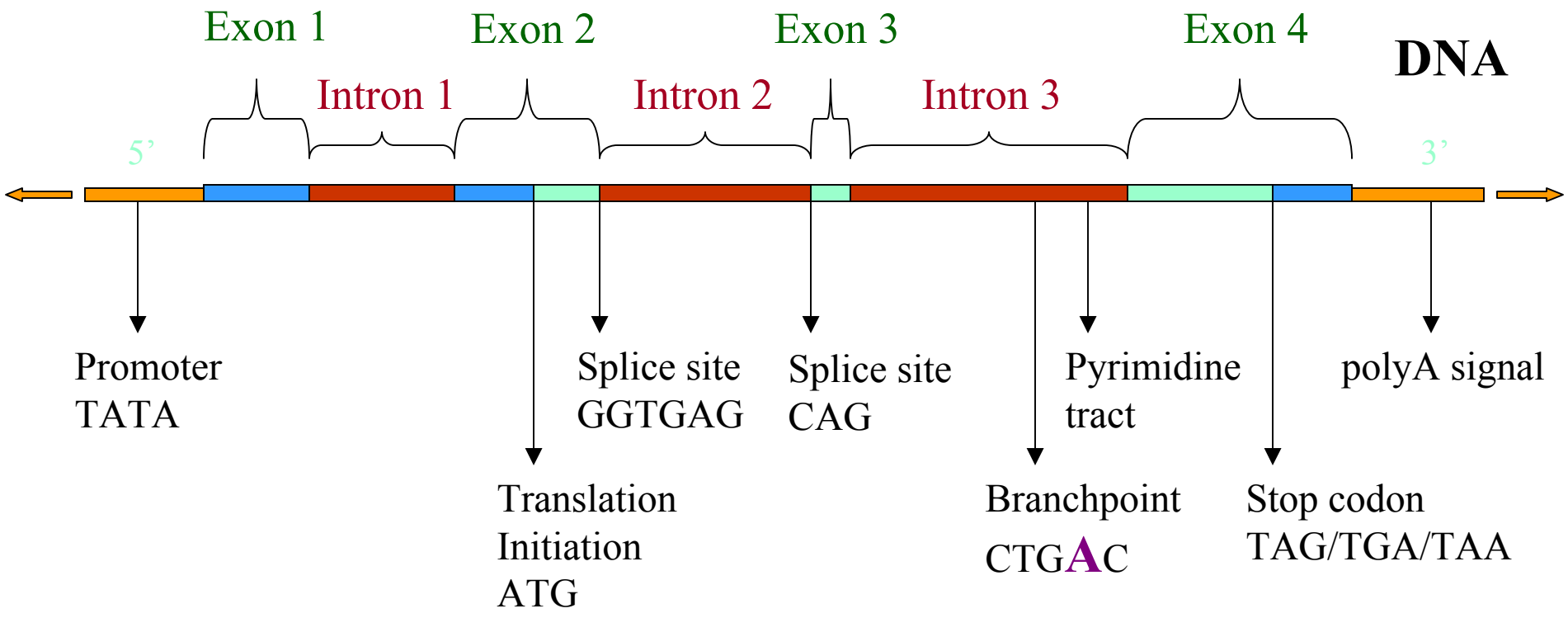



Exon 1
Exon 2
Exon 3
Exon 4
Intron 1
Intron 2
Intron 3
5'
3'
Splice site
GGTGAG
Splice site
CAG
Translation
Initiation
ATG
Branchpoint
CTGAC
Stop codon
TAG/TGA/TAA

Question: How do we align sequences so that the alignments are biologically meaningful?

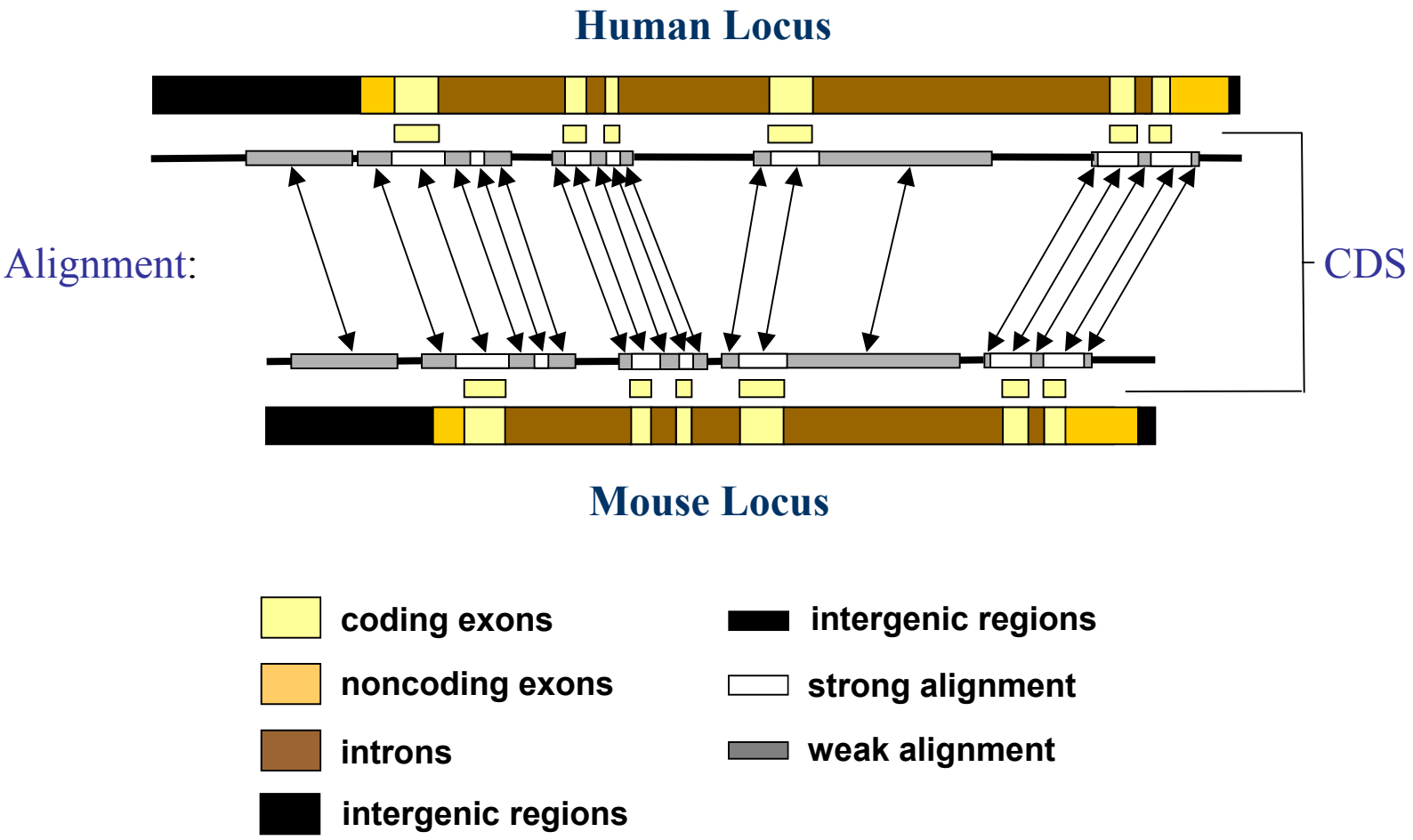

Exon 1
Exon 2
Exon 3
Exon 4
Intron 1
Intron 2
Intron 3
5’
3’
Splice site
GGTGAG
Splice site
CAG
Translation
Initiation
ATG
Stop codon
TAG/TGA/TAA

The Gene Finding Problem



Example: a human/mouse ortholog

Proliferating cell nuclear antigen (PCNA)



Suggestion: In order to find genes in two syntenic regions, first align them and then use the alignment to assist in the gene finding.

Alignment 1
Seq1: human
Seq2: macaque
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 7

Alignment 2
Seq1: human
Seq2: pig
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 6

Alignment 3
Seq1: human
Seq2: rabbit
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 4

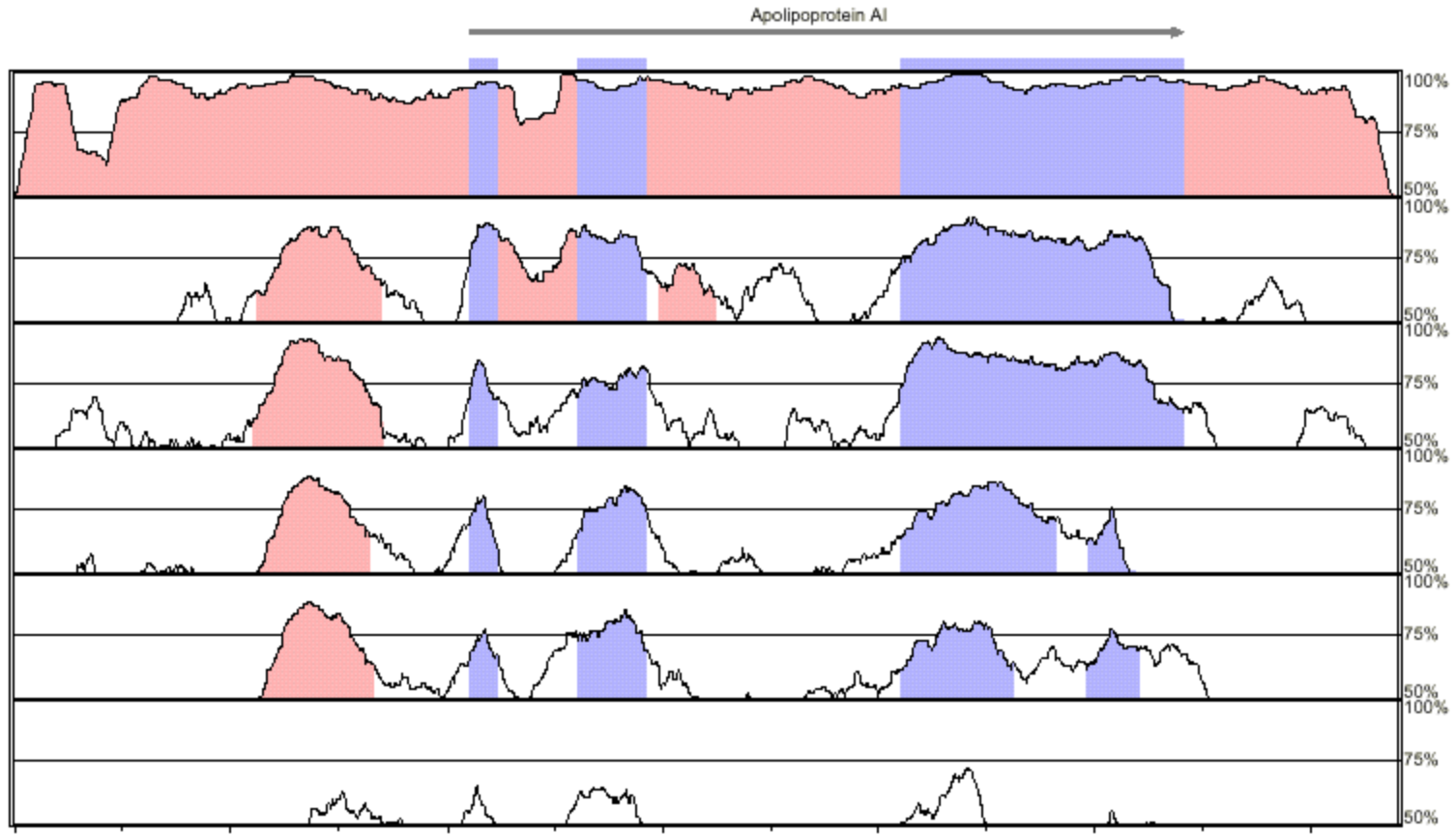
Alignment 4
Seq1: human
Seq2: mouse
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 5

Alignment 5
Seq1: human
Seq2: rat
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 5

Alignment 6
Seq1: human
Seq2: chicken
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 0

Resolution: 4
Window size: 100
Start: 1

Exon
UTR
CNS



Comparison of 1196 orthologous genes (Makalowski et al., 1996)

- Sequence identity:
  - exons: 84.6%
  - protein: 85.4%
  - introns: 35%
  - 5′ UTRs: 67%
  - 3′ UTRs: 69%
- 27 proteins were 100% identical.

Observation:
- Finding the genes will help to find biologically meaningful alignments.
-Finding a good alignment will help in finding the genes.

Which came first, the chicken or the egg?

They were both generated by a generalized pair hidden Markov model

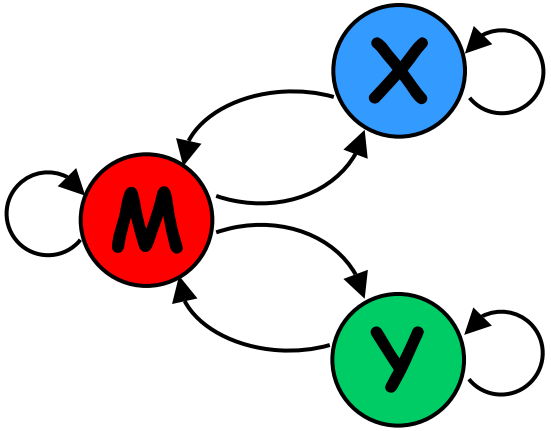
Hidden Markov models

- Sequence alignment with Pair HMMs
- Gene Prediction with Generalized HMMs
- Both simultaneously with GPHMMs

HMMs for sequence alignment: Pair HMMs

Pair HMMs

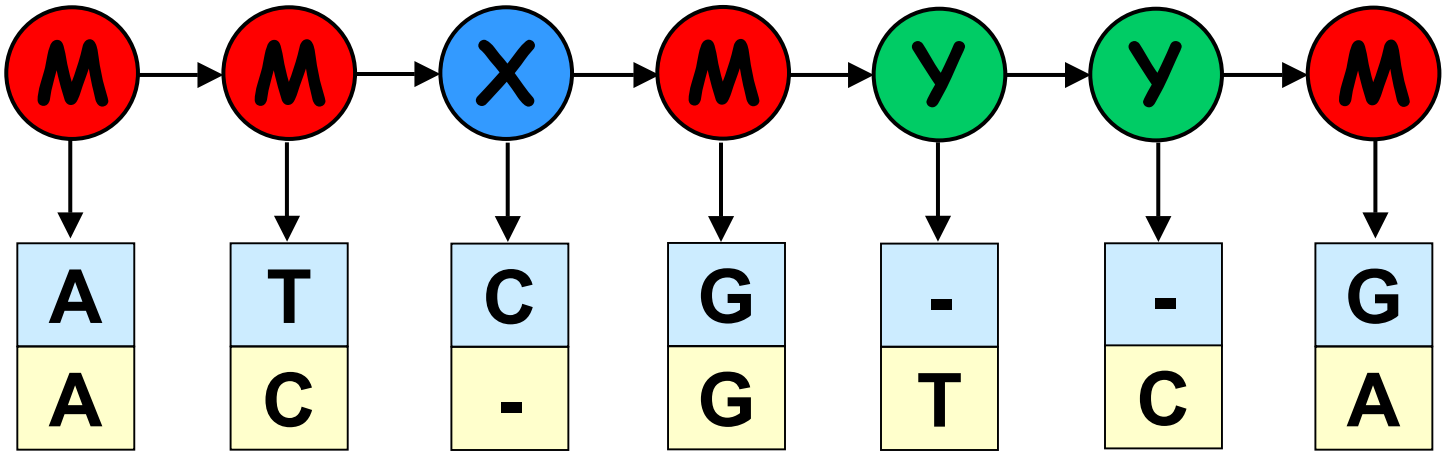
Simple sequence-alignment PHMM



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

Pair HMMs

Hidden sequence:



Hidden alignment:

ATCG--G
AC-GTCA

Observed sequence:

ATCGG
ACGTCA

Using the Pair HMM

In practice, we have observed sequence

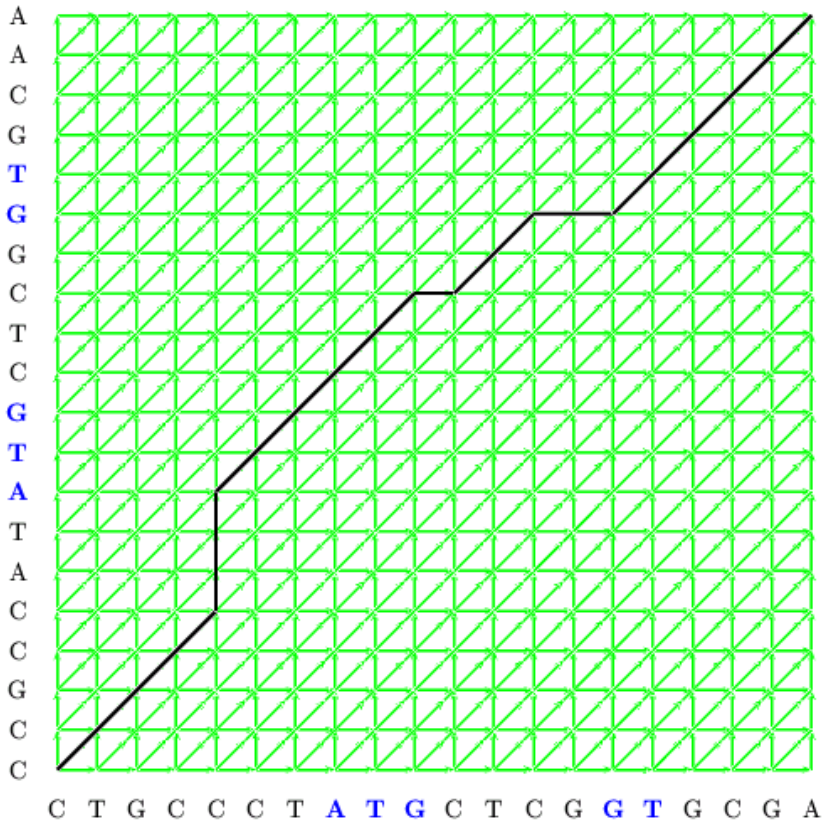
ATCGG
ACGTCA

for which we wish to infer the underlying hidden states

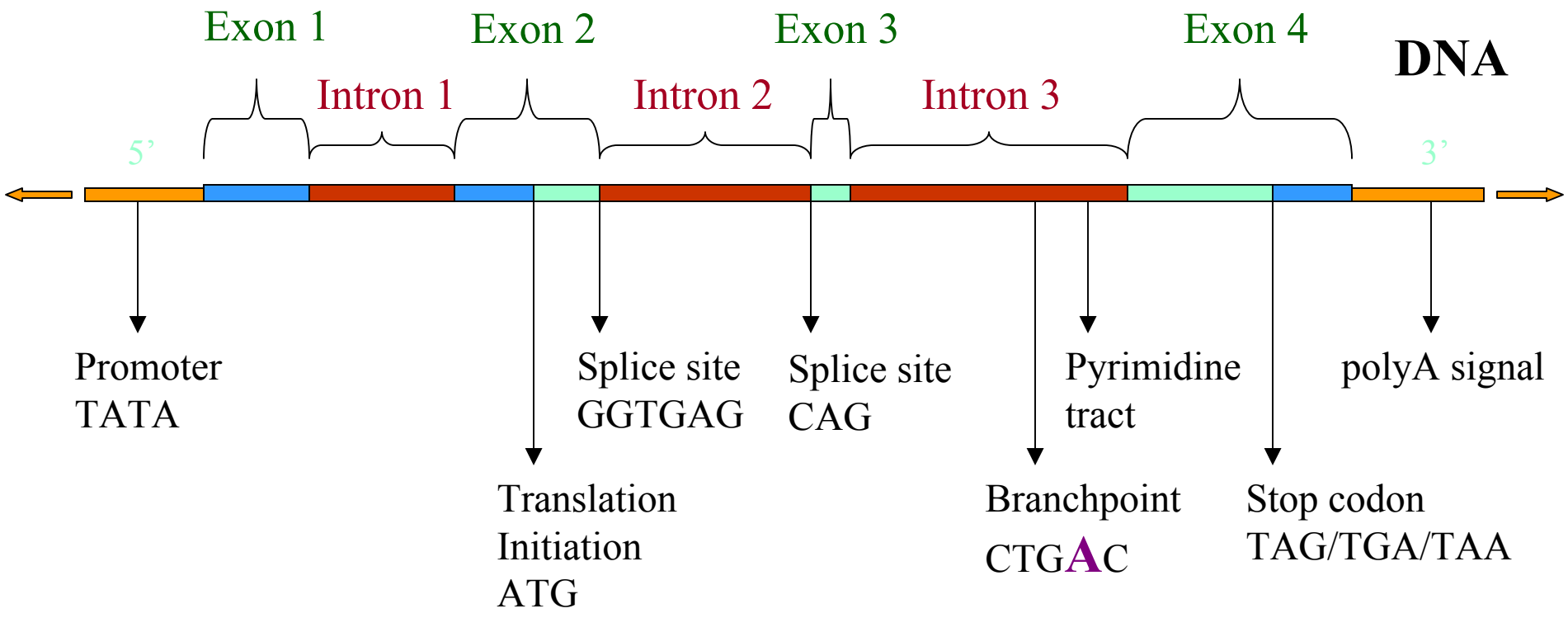
MMXMYYM
ATCG--G
AC-GTCA

One solution: among all possible sequences of hidden states, determine the most likely (Viterbi algorithm).

Viterbi in PHMM = Needleman Wunsch



The Gene Finding Problem



Using GHMMs for ab-initio gene finding

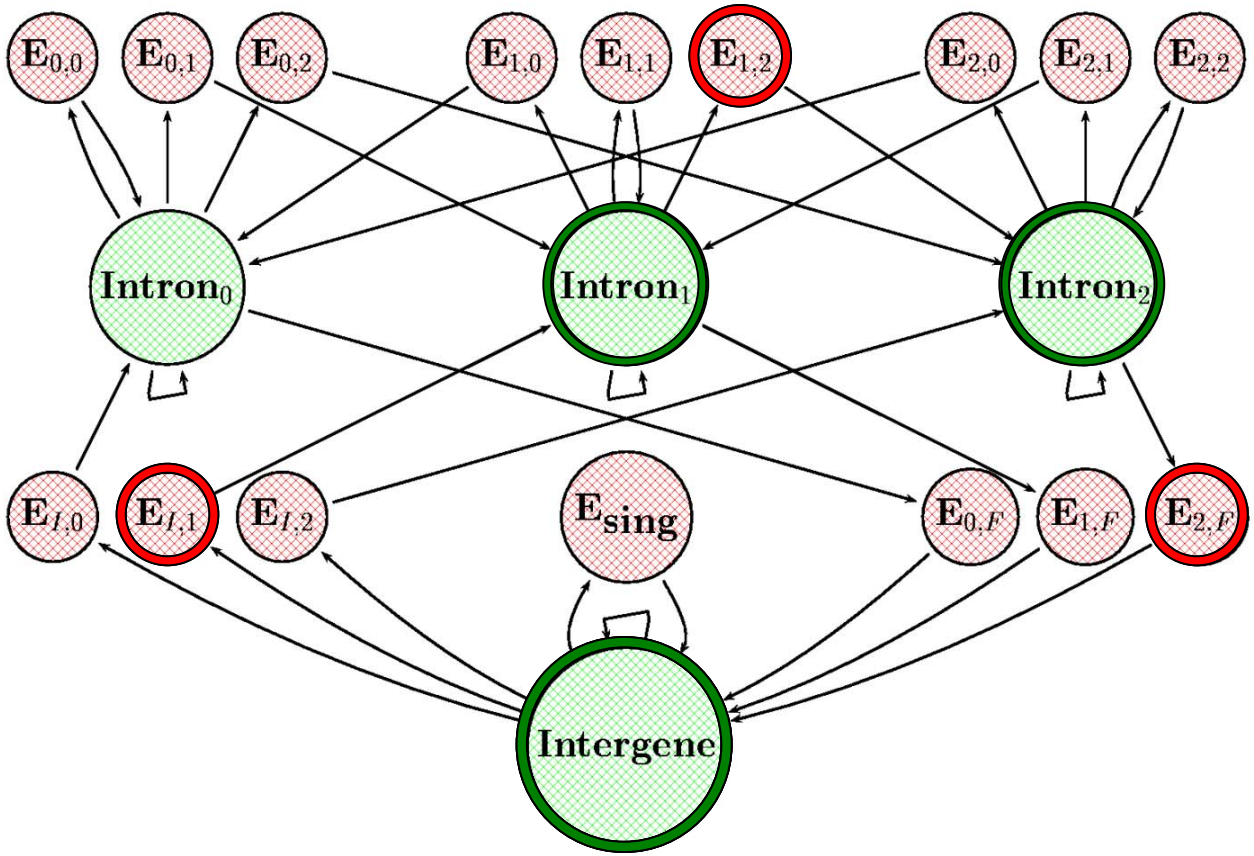
In practice, have observed sequence

TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

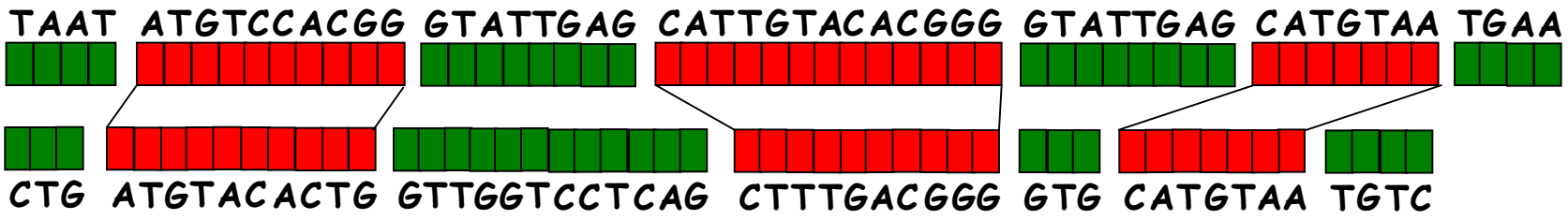
Predict genes by estimating hidden state sequence

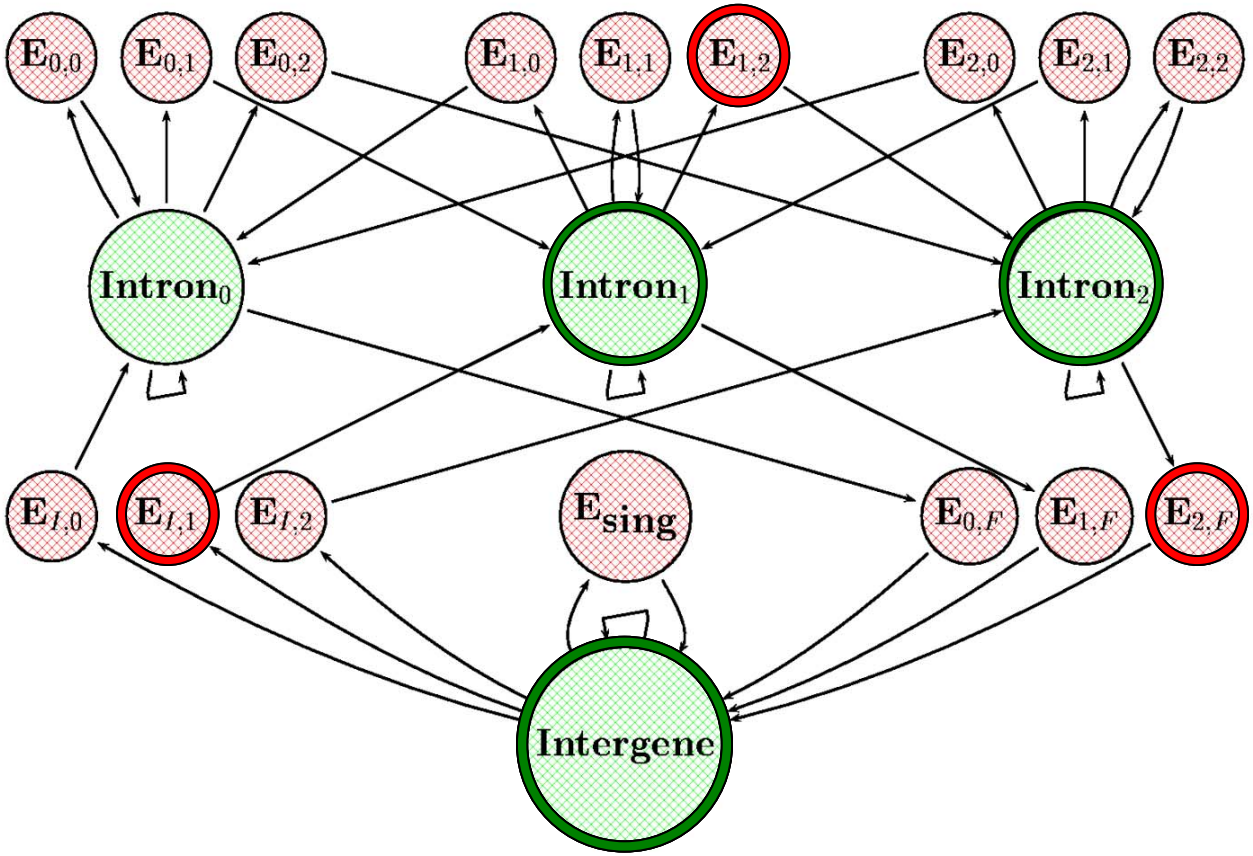
Usual solution: single most likely sequence of hidden states (Viterbi).


TAAT
ATGTCCACGG
GTATTGAG
CATTGTACACGGG
GTATTGAG
CATGTAA
TGAA


$E_{0,0}$
$E_{0,1}$
$E_{0,2}$
$E_{1,0}$
$E_{1,1}$
$E_{1,2}$
$E_{2,0}$
$E_{2,1}$
$E_{2,2}$
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

HMMs for simultaneous alignment and gene finding: Generalized Pair HMMs


TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA
CTG ATGTACACTG GTTGGTCCTCAG CTTTGACGGG GTG CATGTAA TGTC


$E_{0,0}$
$E_{0,1}$
$E_{0,2}$
$E_{1,0}$
$E_{1,1}$
$E_{1,2}$
$E_{2,0}$
$E_{2,1}$
$E_{2,2}$
$Intron_0$
$Intron_1$
$Intron_2$
$E_{I,0}$
$E_{I,1}$
$E_{I,2}$
E_{sing}
$E_{0,F}$
$E_{1,F}$
$E_{2,F}$
Intergene

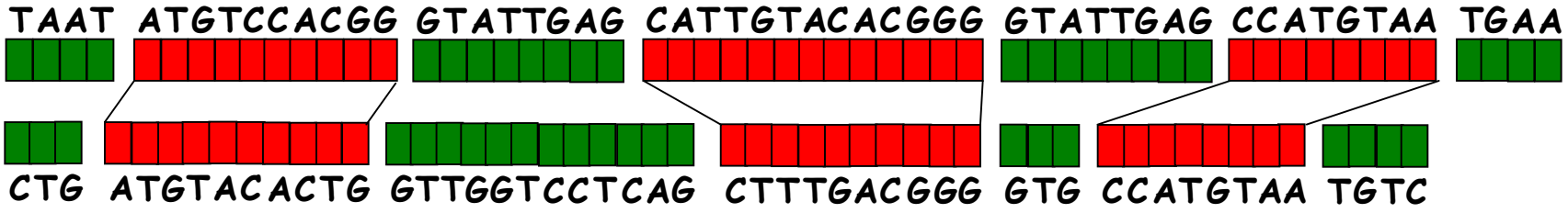
Using GPHMMs for cross-species gene finding

given a pair of syntenic sequences

TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCCATGTAATGAA

CTGATGTACACTGGTTGGTCCTCAGCTTTGACGGGGTGCCATGTAATGTC

predict genes by estimating hidden state sequence



Predict exon-pairs using single most likely sequence of hidden states (Viterbi).

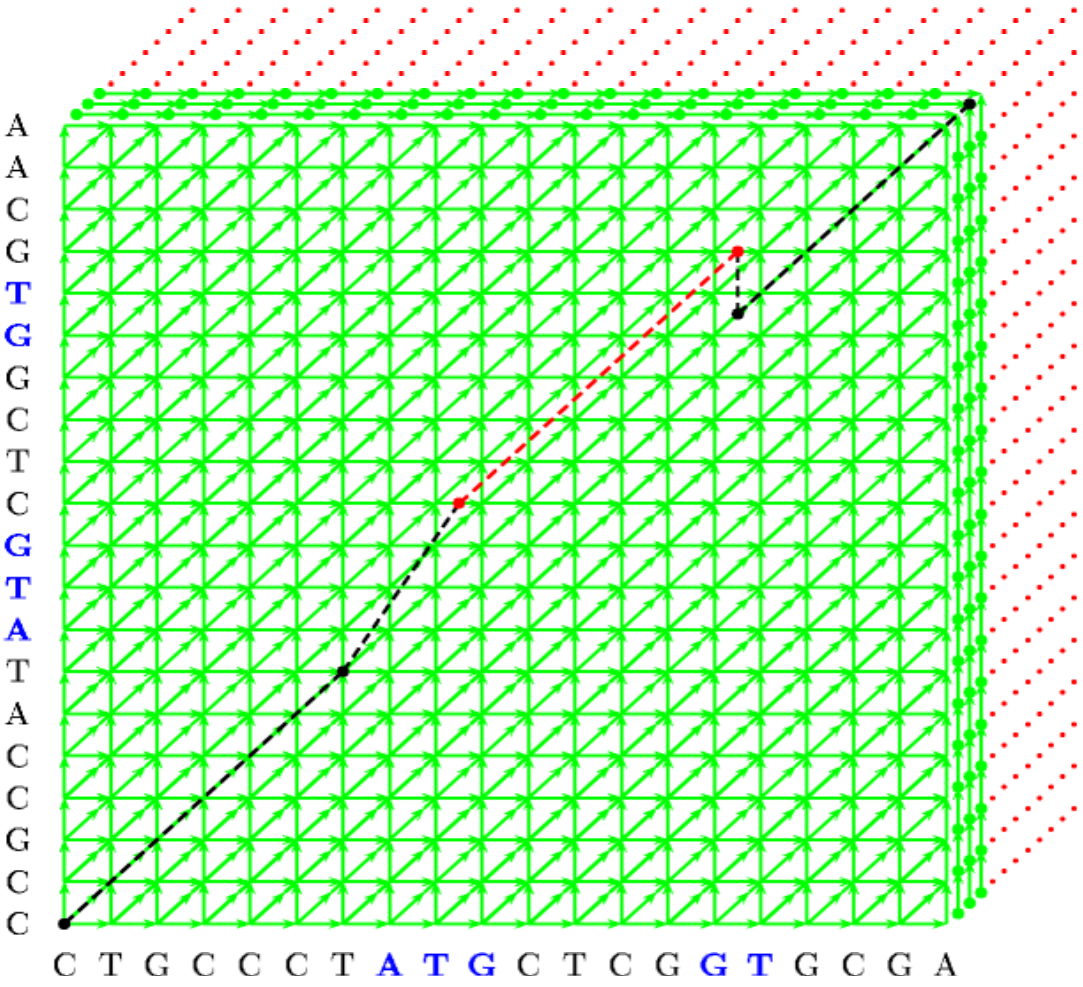
Computational Complexity

$N =$ # HMM states $T =$ length seq1

$D =$ max duration $U =$ length seq2

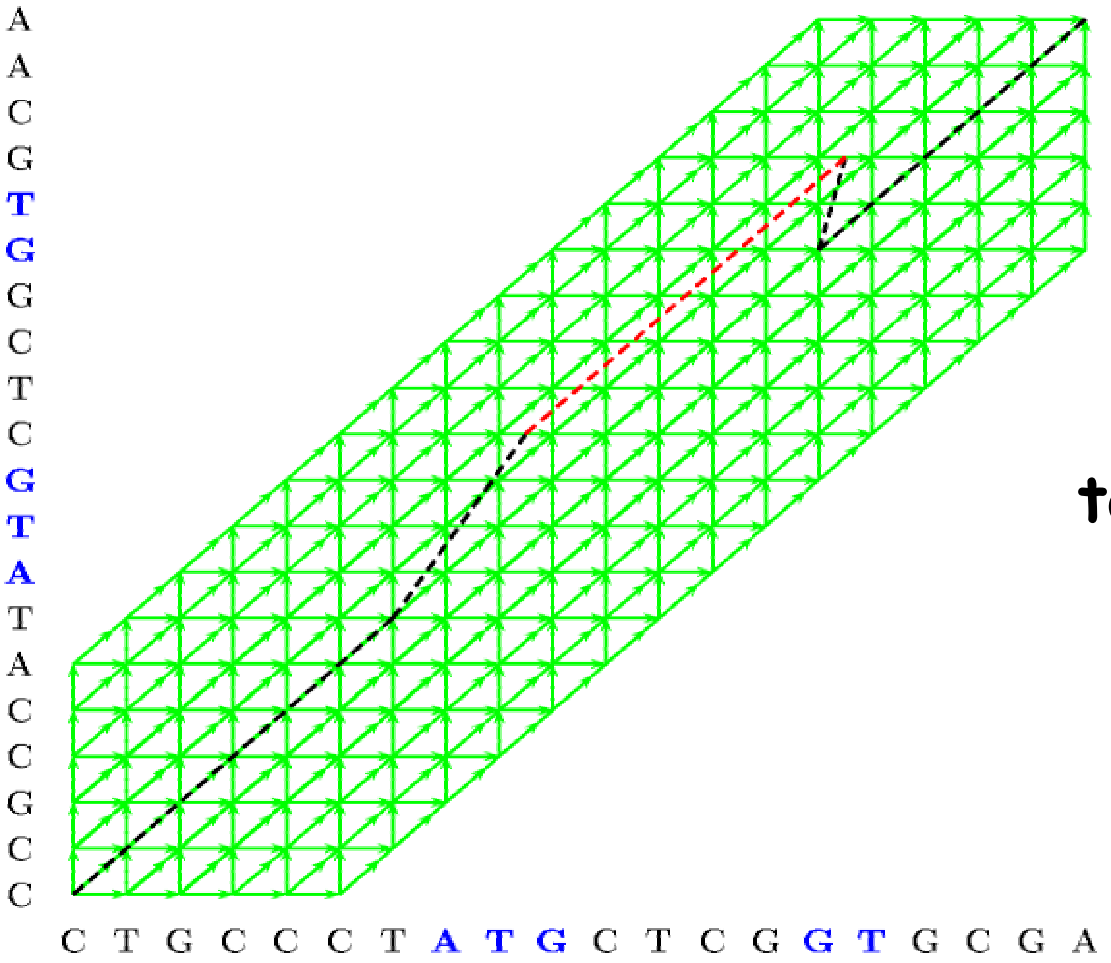
Model	Time	Space
HMM	N^2T	NT
PHMM	N^2TU	NTU
GHMM	D^2N^2T	NT
GPHMM	D^4N^2TU	NTU

lattice view



Introns
Exons

Approximate alignment



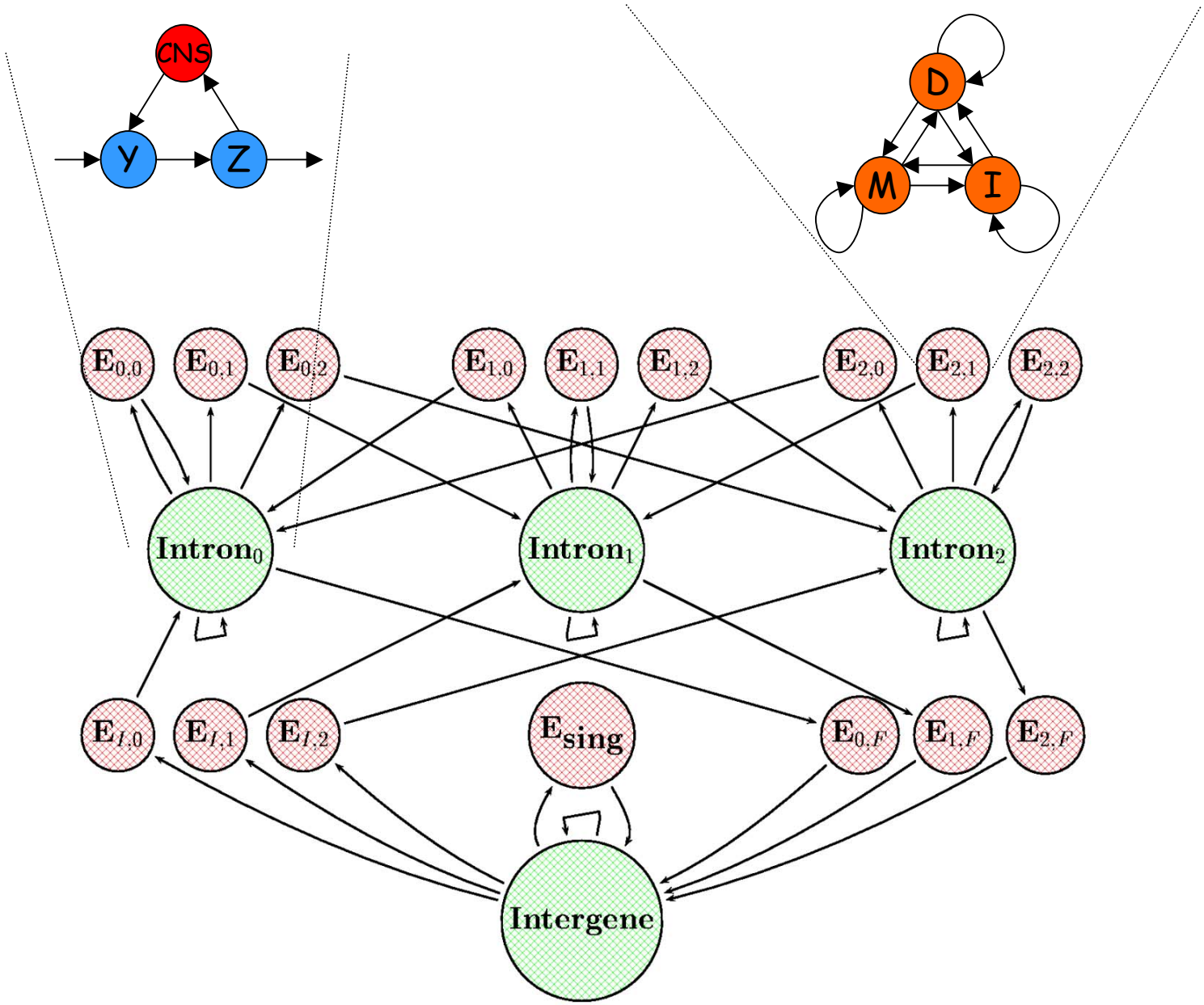
Reduces

from O(TU)

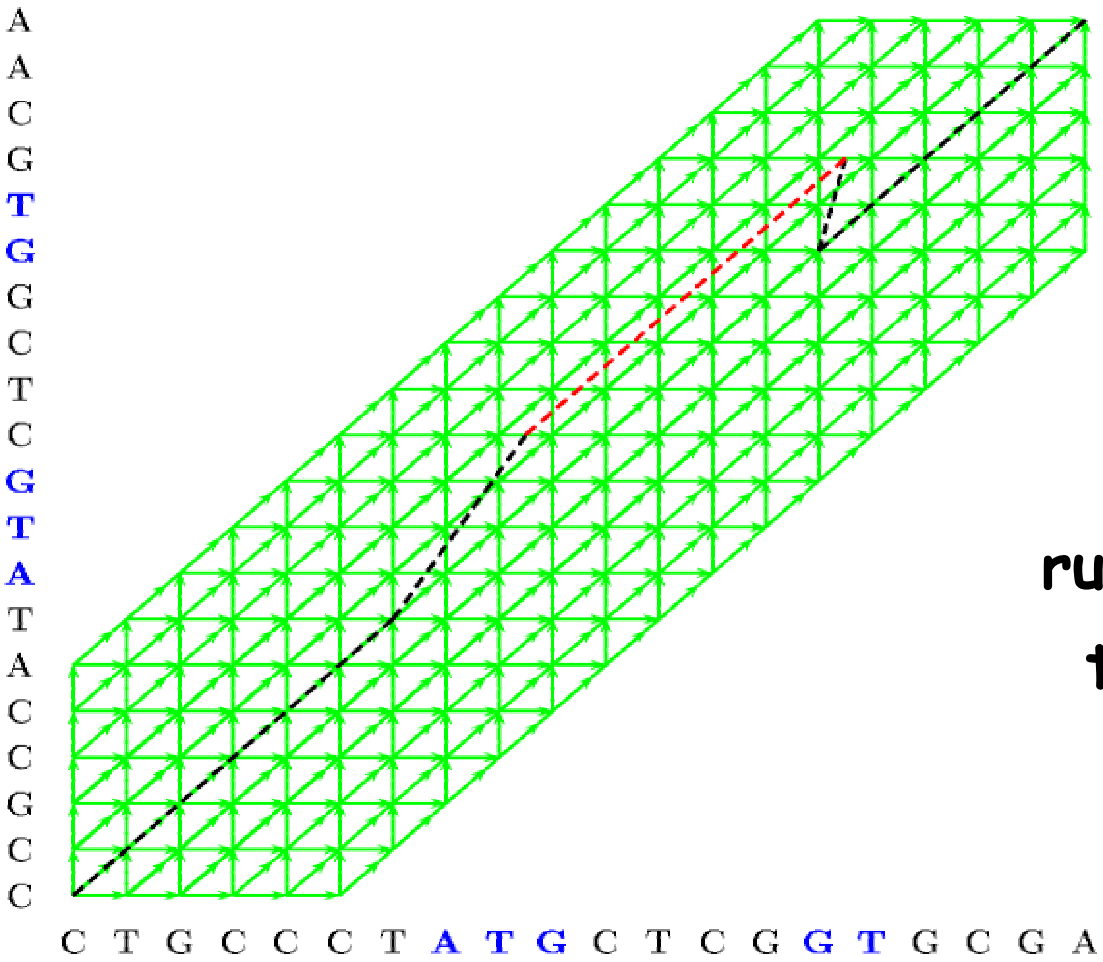
to O(max(T,U))

A GPHMM implementation SLAM

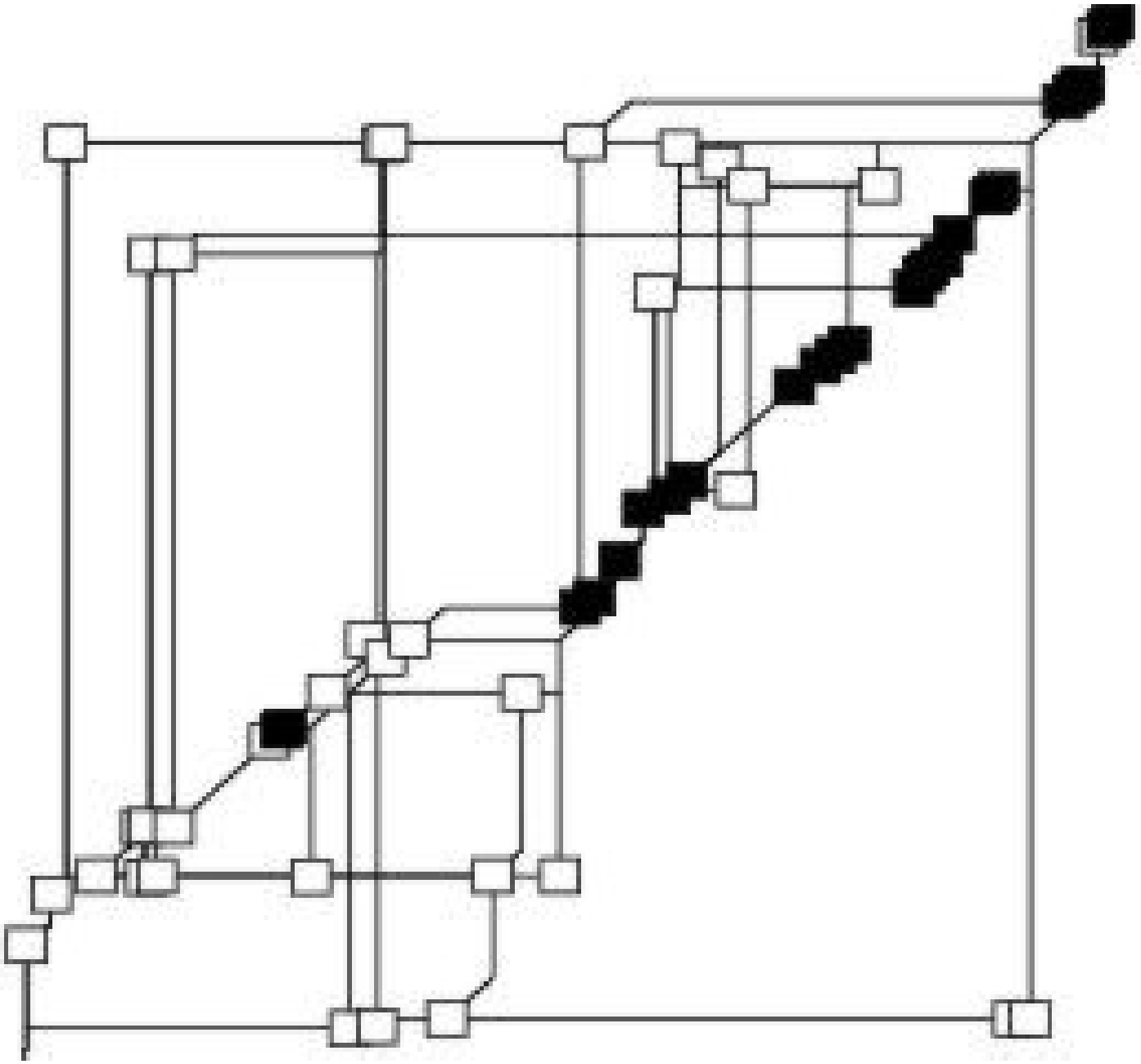
- SLAM components
  - Splice sites (Variable length Markov models).
  - Introns and Intergenic regions (2nd order Markov models, independent geometric lengths, CNS states).
  - Coding sequences (3-periodic Markov models, generalized length distributions, protein-based pairHMM.)
- Input
  - Pair of syntenic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

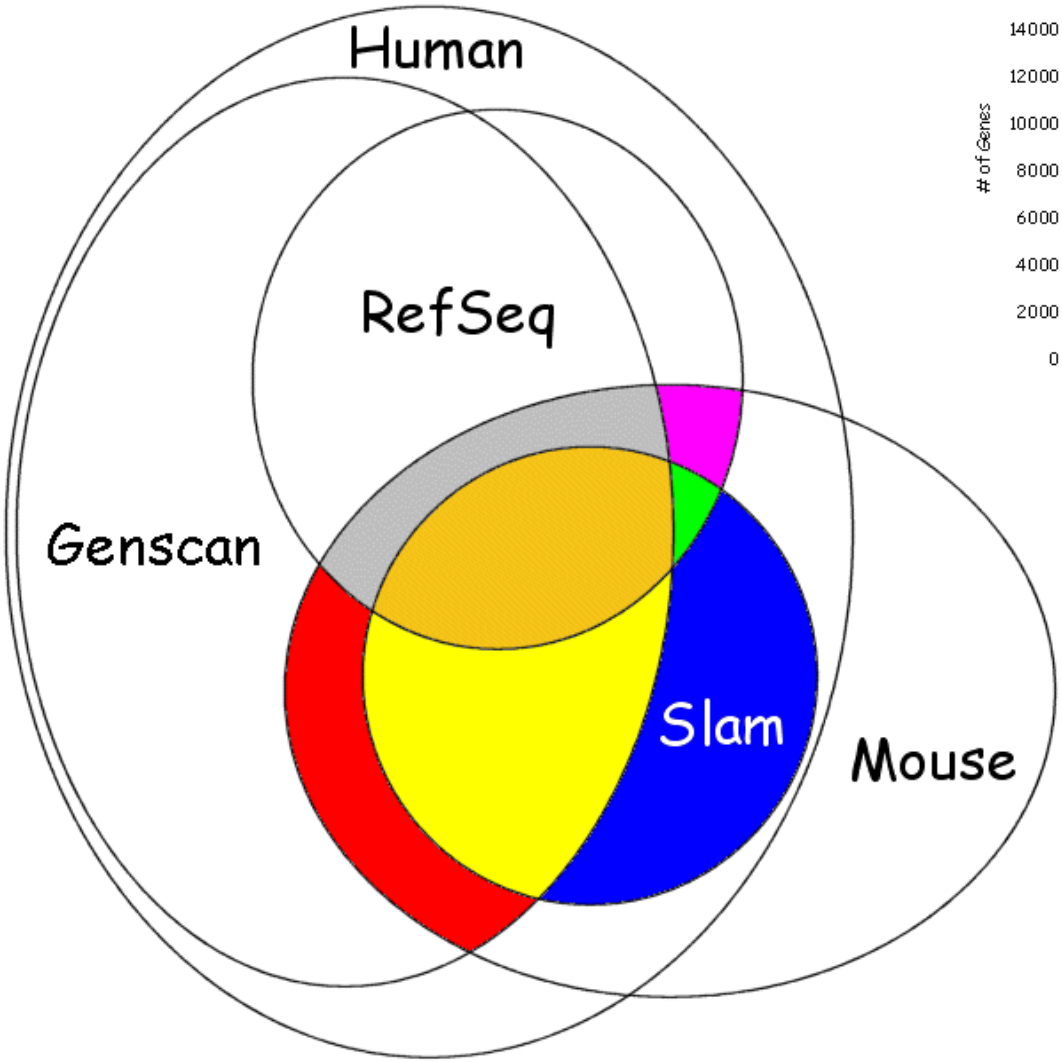

CNS
Y
Z
D
M
I
$E_{0,0}$
$E_{0,1}$
$E_{0,2}$
$E_{1,0}$
$E_{1,1}$
$E_{1,2}$
$E_{2,0}$
$E_{2,1}$
$E_{2,2}$
$Intron_0$
$Intron_1$
$Intron_2$
$E_{I,0}$
$E_{I,1}$
$E_{I,2}$
E_{sing}
$E_{0,F}$
$E_{1,F}$
$E_{2,F}$
Intergene

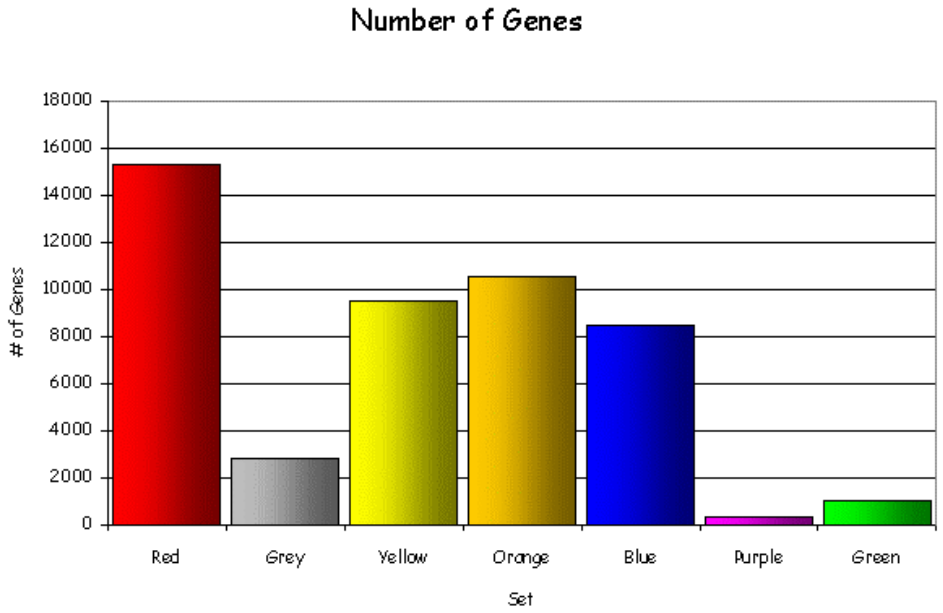
Approximate alignment



Currently generated by running AVID and then "relaxing"




Human
RefSeq
Genscan
Slam
Mouse


Number of Genes
# of Genes
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Red
Grey
Yellow
Orange
Blue
Purple
Green
Set

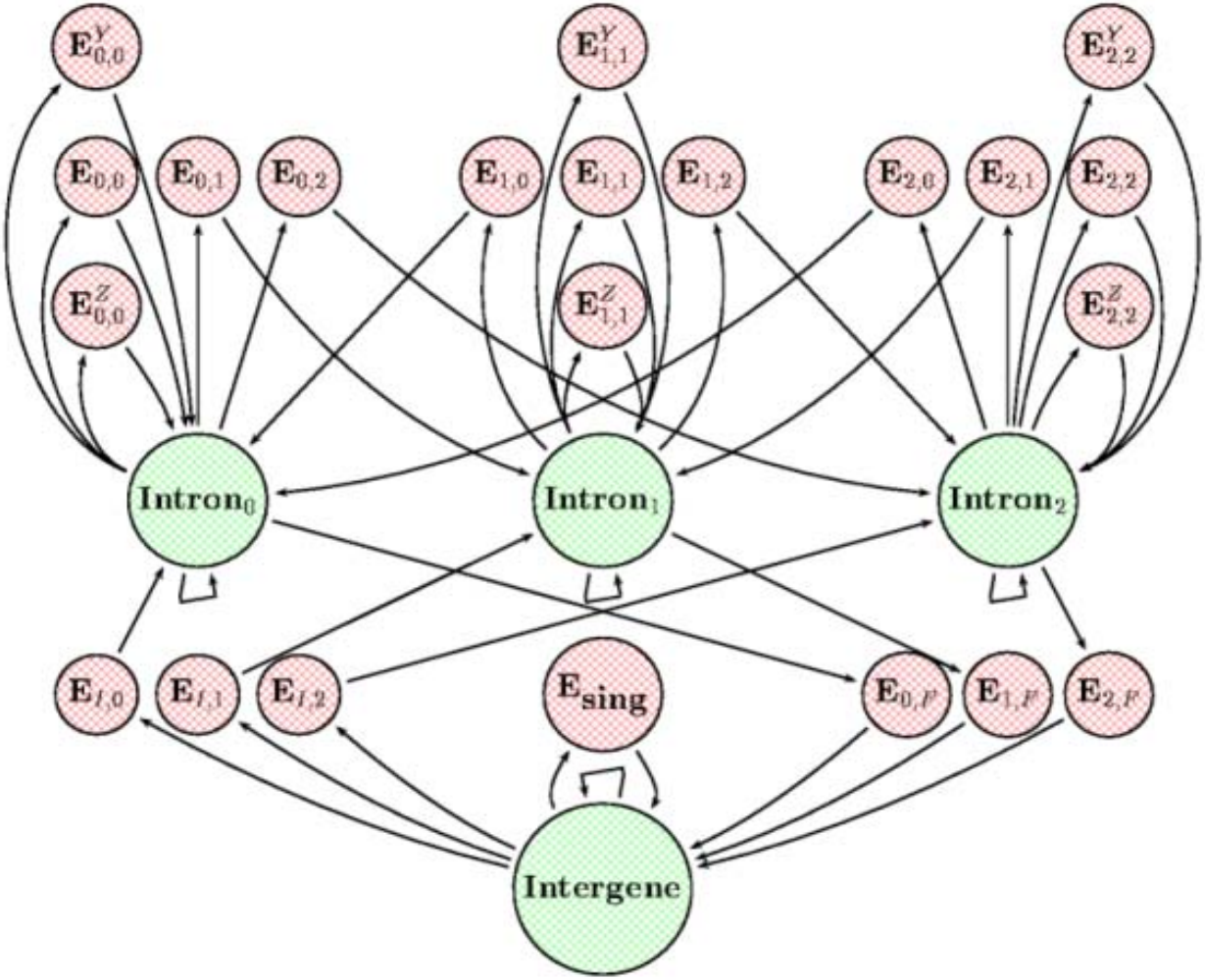
GPHMM applications

- Ideally suited for alignment/feature finding problems
  - DNA/DNA
  - DNA/cDNA
  - DNA/protein
- Extension to more than 2 sequences computationally challenging.

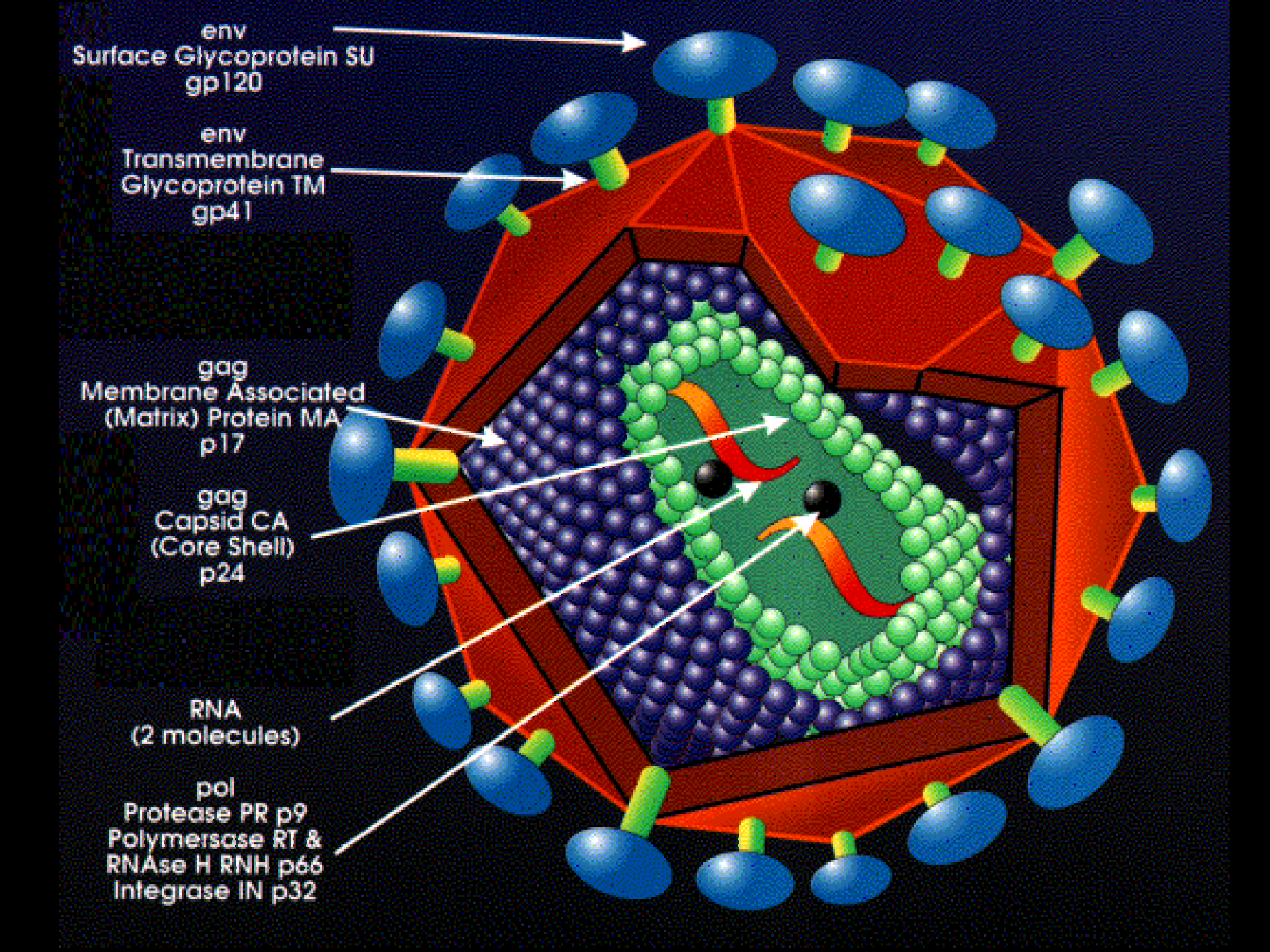
"Its difficult to predict, in particular the future"- GB Shaw

- SLAM improvements
  - modeling more features in pairs
  - states for untranslated regions
  - frameshifts
- Limitations
  - genomic rearrangements
  - overlapping genes

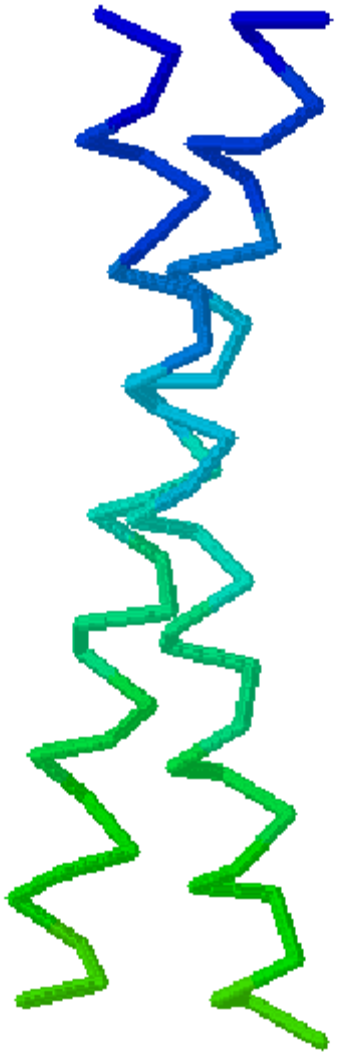
Allowing for inserted exons

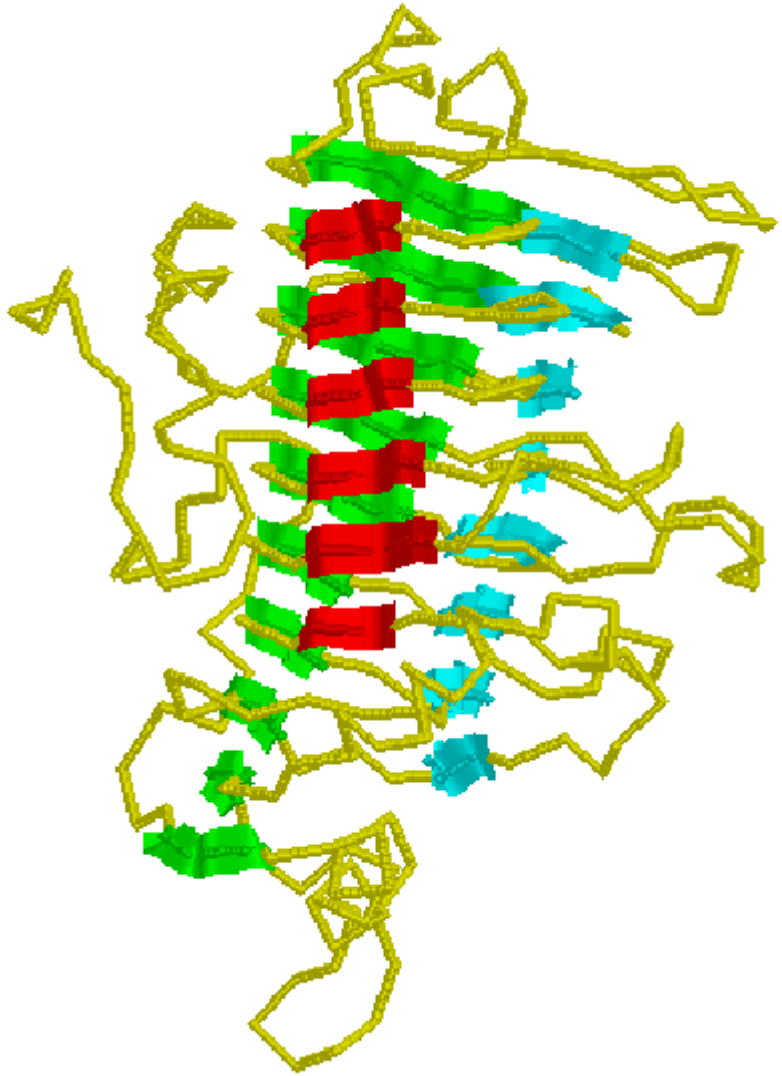


Analysis of Protein Sequences


env
Surface Glycoprotein SU
gp120
env
Transmembrane
Glycoprotein TM
gp41
gag
Membrane Associated
(Matrix) Protein MA
p17
gag
Capsid CA
(Core Shell)
p24
RNA
(2 molecules)
pol
Protease PR p9
Polymersase RT &
RNAse H RNH p66
Integrase IN p32

Examples of Super-secondary Structure

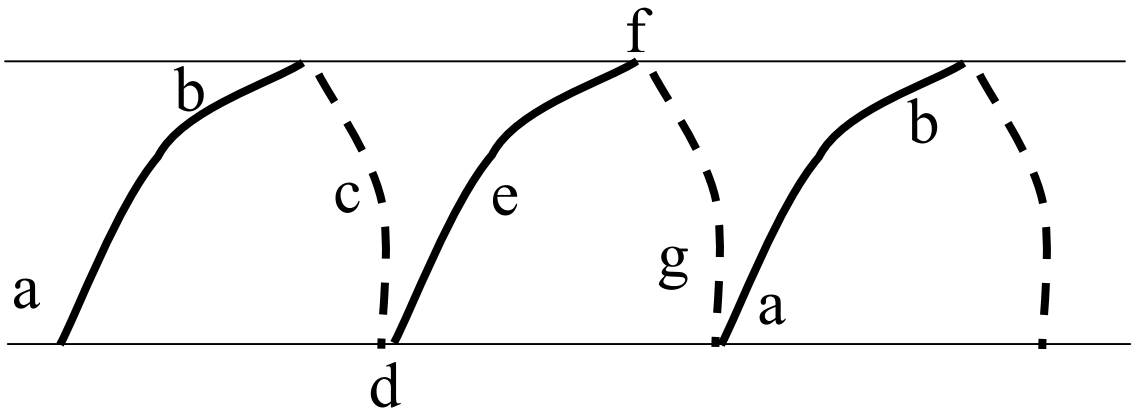




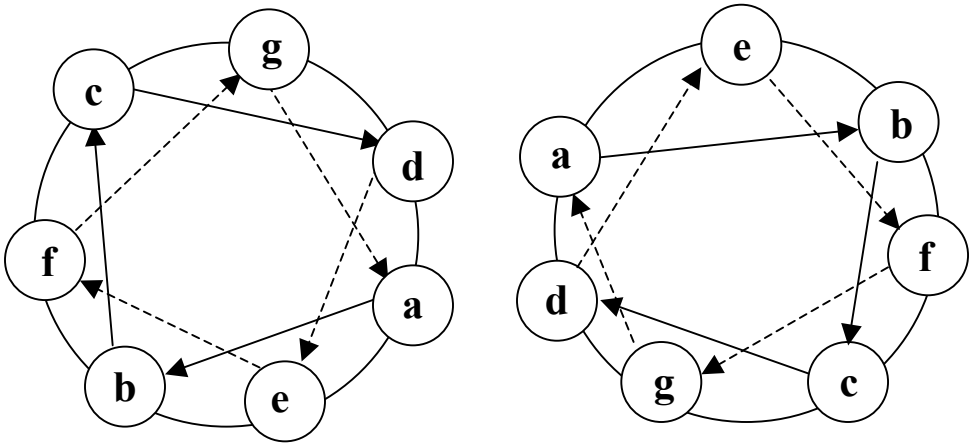
Geometry of Coiled Coil

7 repeating positions (a--g) in a coiled coil:

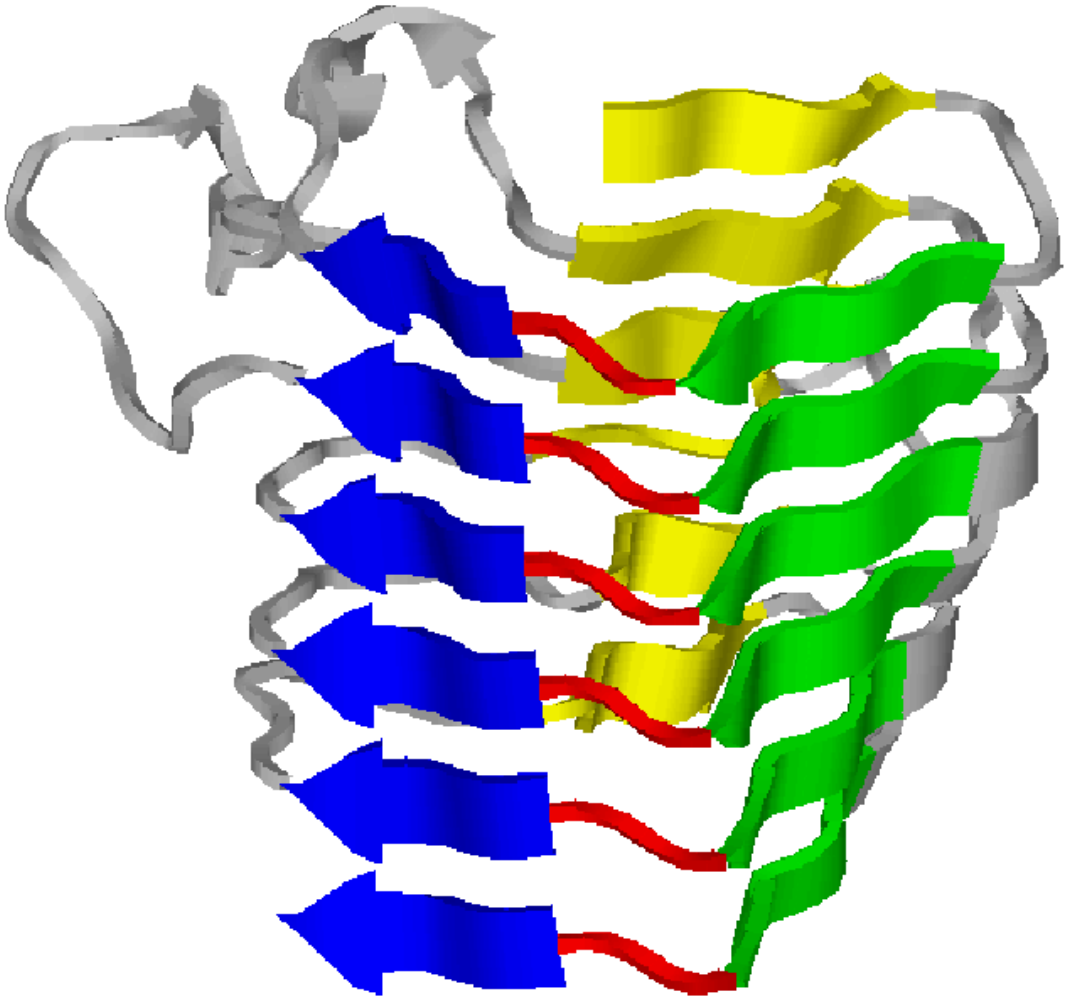
side view:



top view:



Beta Helices



Thanks

- Marina Alexandersson
- Simon Cawley
- CCSF HIV page
- Inna Dubchak
- Eddy Rubin
- Bonnie Berger
- Ethan Wolf
- Serafim Batzoglou
- Robert Gentleman
- Sandrine Dutoit