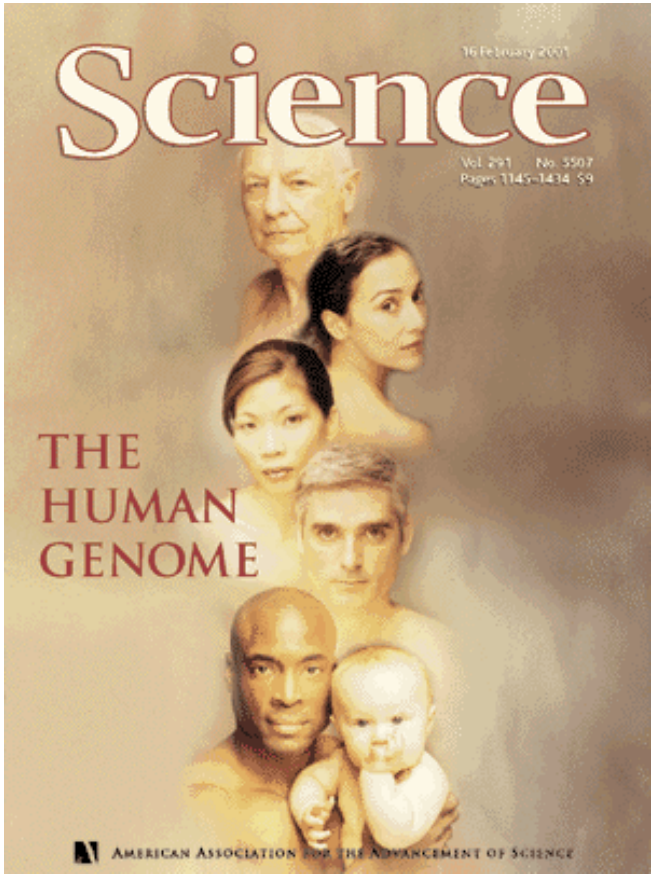
Bioinformatics

- Scientific Computing Lecture, UNICC

Marina Alexandersson

Göteborg, March 25, 2004

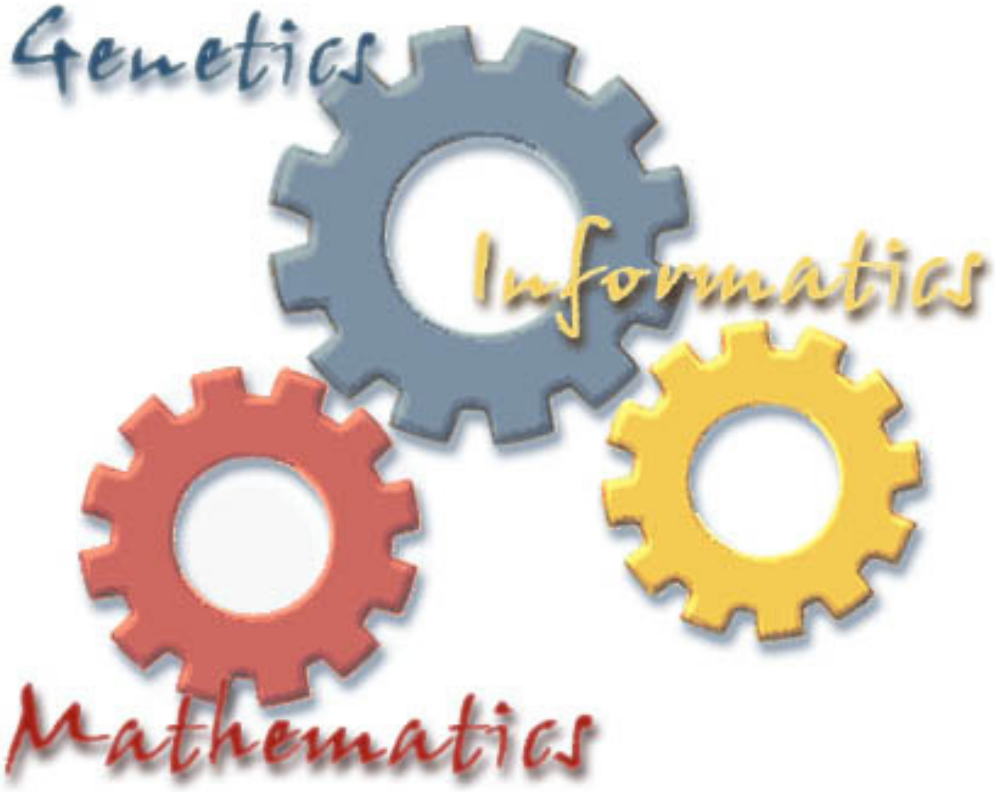
The human genome



Fraunhofer ITWM

CHALMERS
Research Centre
Industrial Mathematics

What is bioinformatics?



Where is bioinformatics used?

DNA analysis

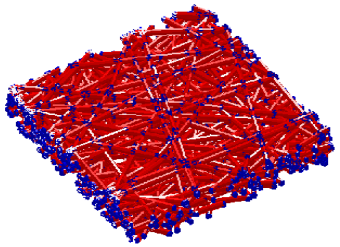
Drug

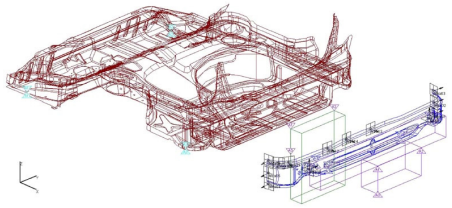
- DNA sequencing
- Sequence analysis
- Genes involved in diseases
- Protein structure and function
- Drug targets
- Lead compound
- Clinical trials

Technologies at the Fraunhofer-Chalmers centre

➢ Computational physics and fatigue

- **materials, mechanics, electromagnetics**
- **fatigue life and load analysis**





➢ Opitmization and quality engineering

- **design, geometry, qulity management**
- **logistics, scheduling, production**

➢ System theoretical methods

- **biostatistics, bioinformatics, systems biology**
- **telecommunciations, signal and image analysis**
- **finance and insurance**

Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

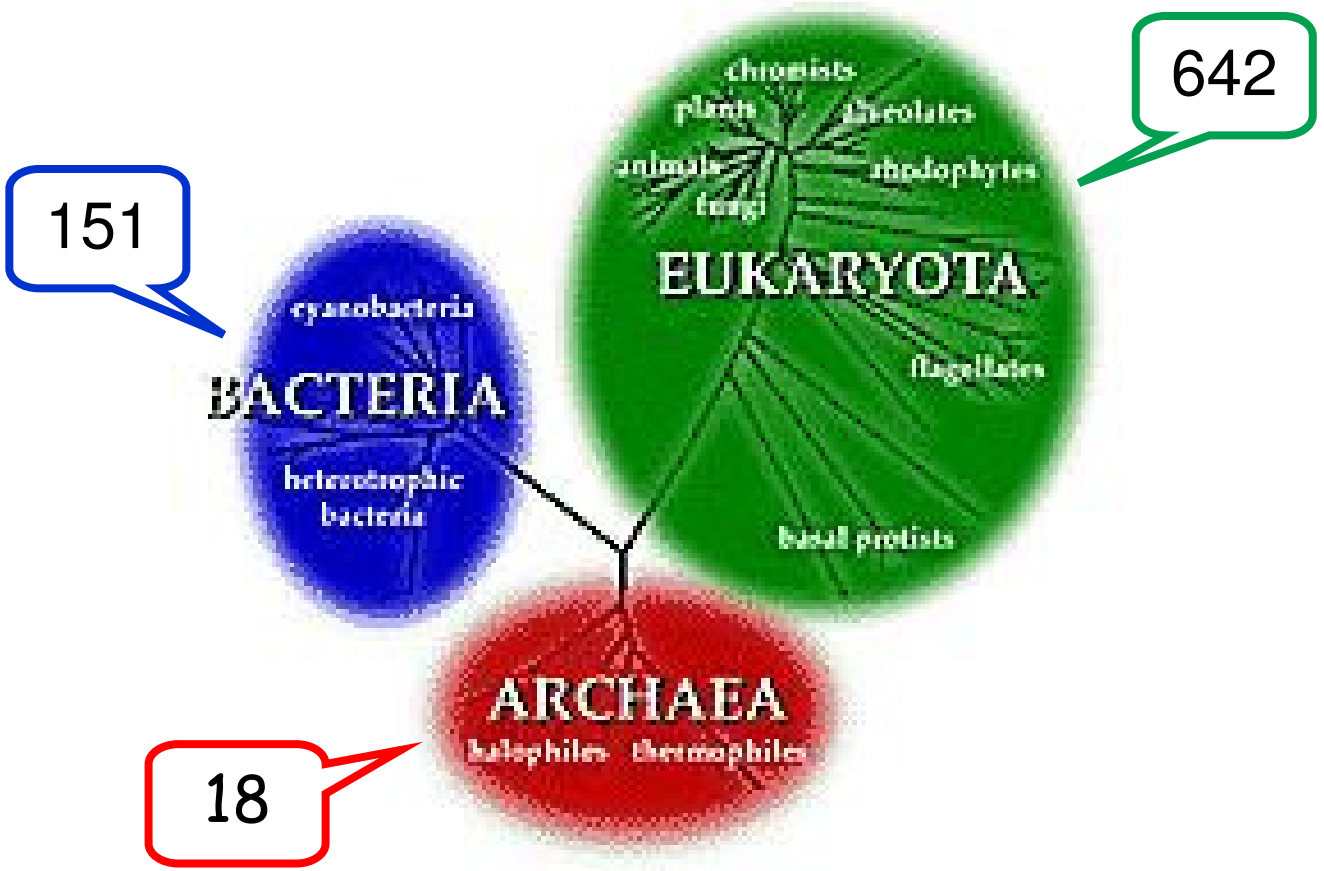
Bioinformatics

Research, development and application of computational methods and tools to expand the use of biological and medical data, including retrieving, storing, organizing, analyzing and visualizing such data.

- **Sequence analysis** to classify sequence data and identify functional elements
- **Comparative genomics** to detect similarities and differences in related organisms
- **Functional genomics** to infer biological structure and function from primary sequence
- **Proteomics** to identify, characterize and quantify proteins active in certain pathways, tissues, cells or organs
- **Biostatistics and statistical genetics** to analyze disease genotypes and phenotypes, detect environment interactions etc

ITWM
Fraunhofer
CHALMERS
Research Centre
Industrial Mathematics

Sequenced genomes



More to come...

ITWM
Fraunhofer
CHALMERS
Research Centre
Industrial Mathematics

Whole genome analysis - Why?

- Complete gene and protein sets
- Primary sequence of all genes
- Sequence relationships between genes and proteins
- Function of new proteins
- Transcriptional level of all genes
- Understanding methabolic pathways
- Trace disease genes
- ...

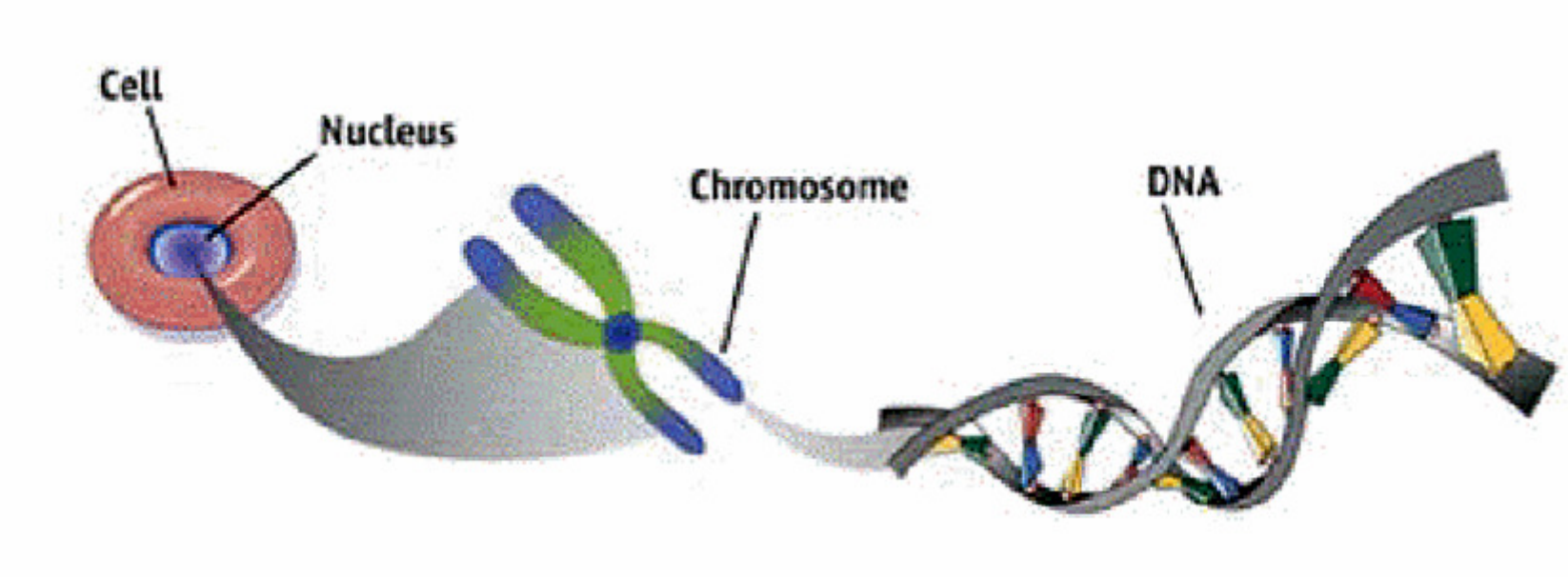
Whole genome analysis

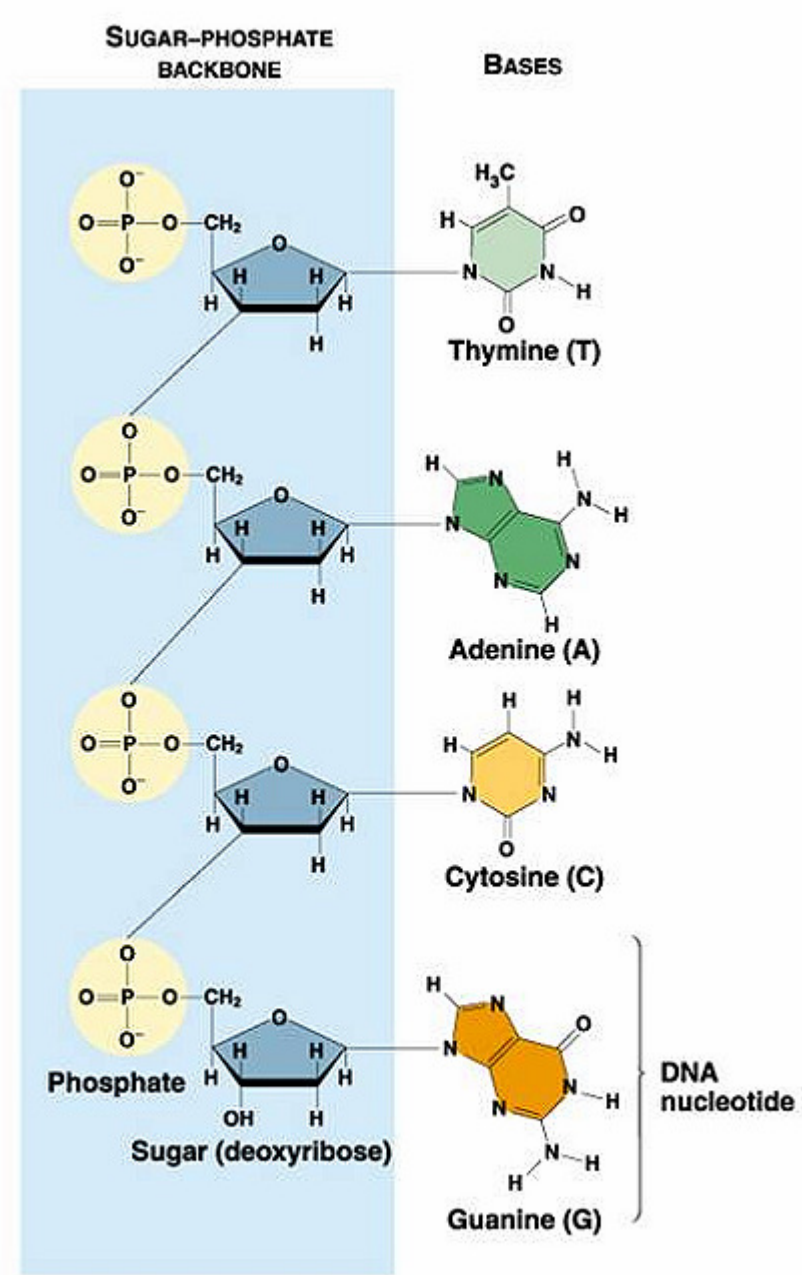
- Gene finding
- Sequence comparisons
- Regulatory region identification

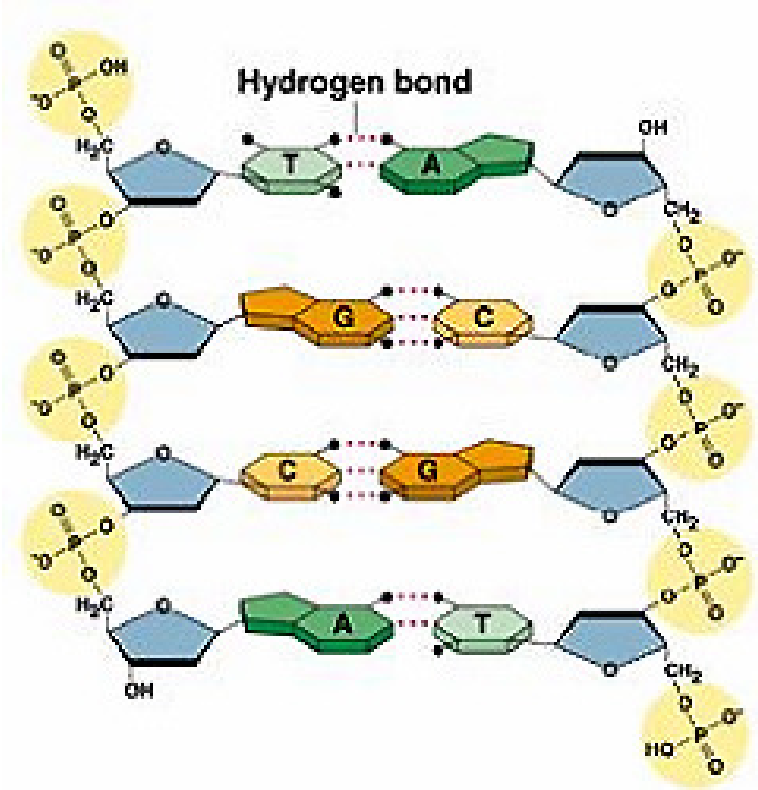
Gene finding
- *Generalized HMMs*

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

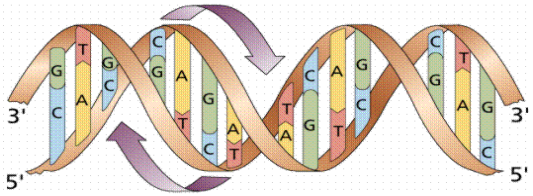
Basic genetics

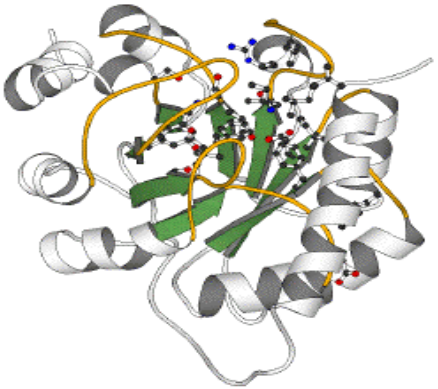



SUGAR–PHOSPHATE BACKBONE
BASES
Thymine (T)
Adenine (A)
Cytosine (C)
Guanine (G)
Phosphate
Sugar (deoxyribose)
DNA nucleotide


Hydrogen bond
T
A
G
C
C
G
A
T

Gene expression





DNA

transcription

RNA

translation

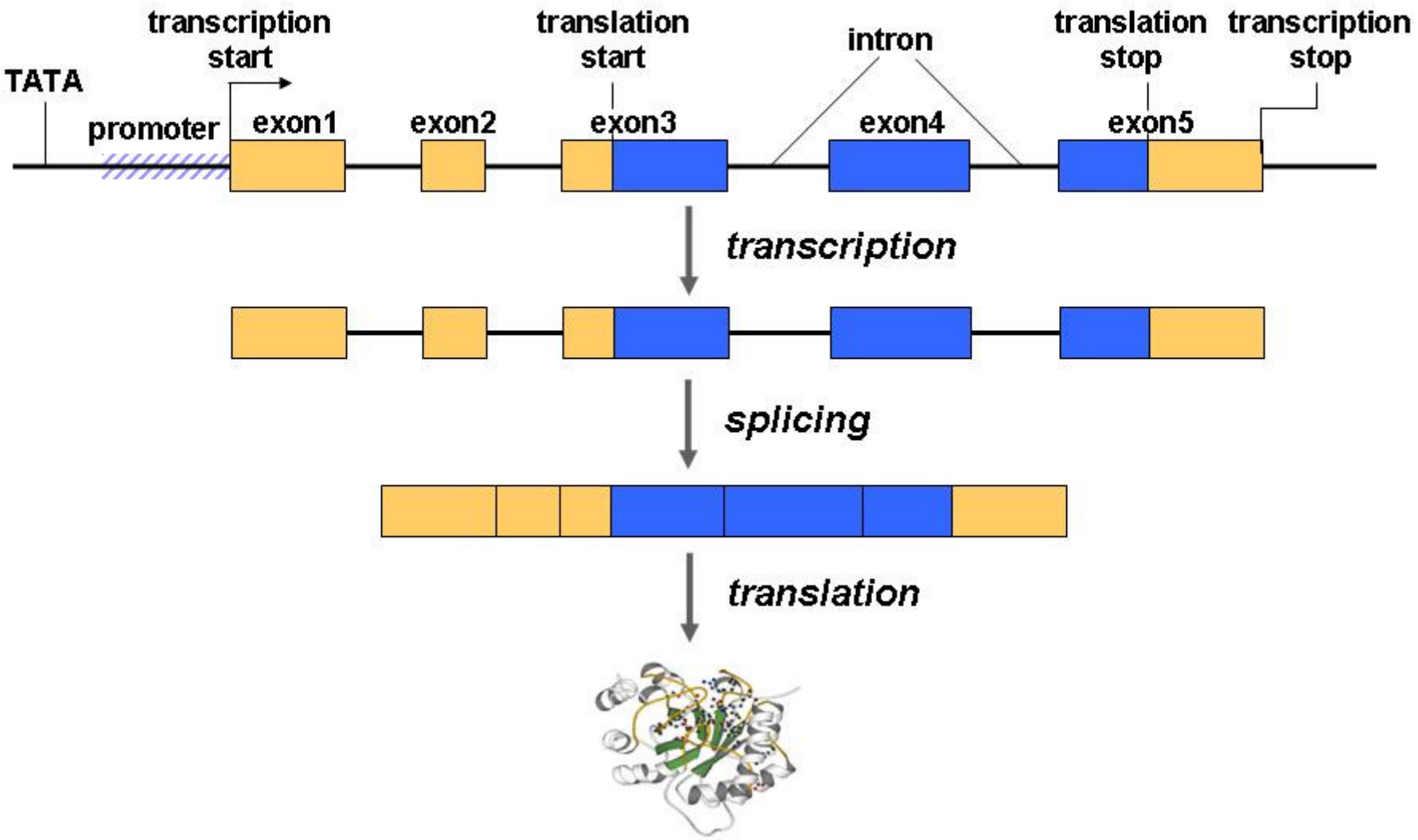
Protein

CCTGAGCCAACTATTGATGAA

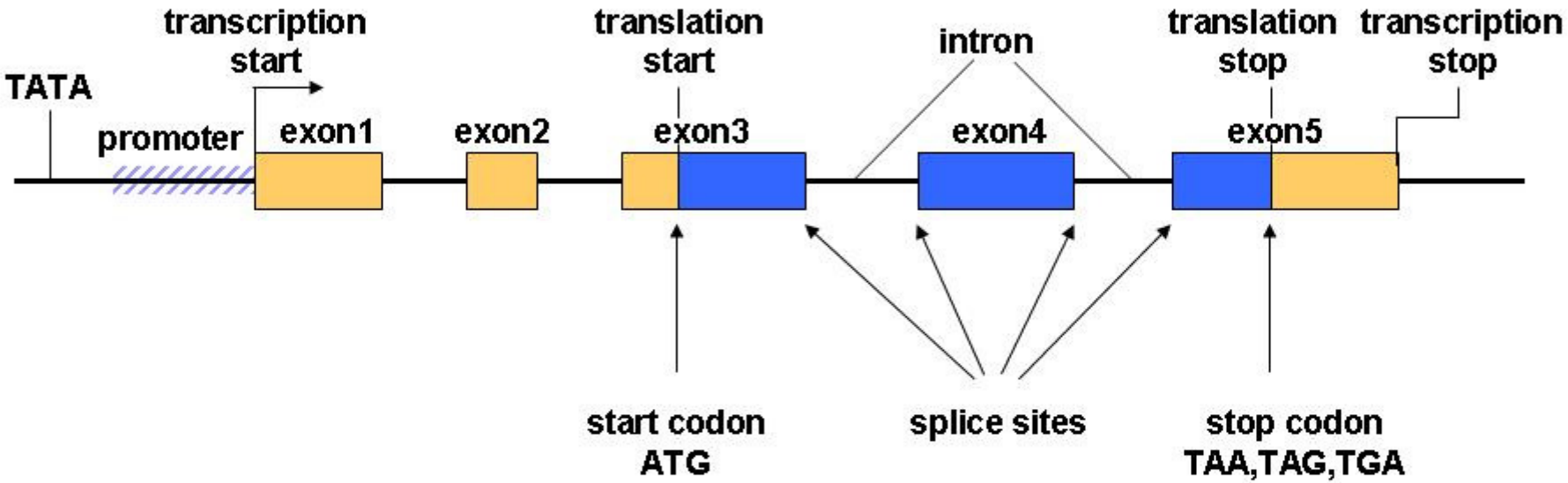
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Gene structure



Finding genes



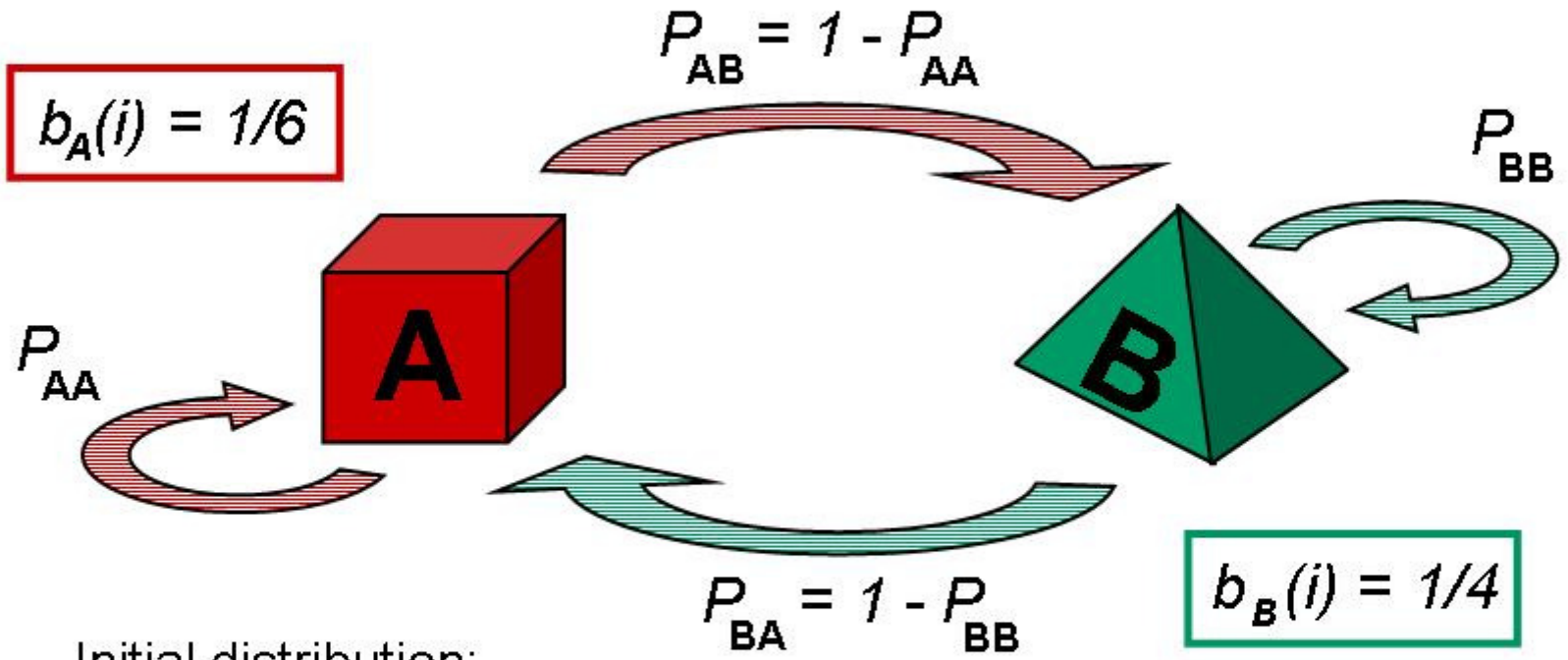
> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatccttaccaaccgtggcatc
gcccagatggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaacccccttccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaagggtca
aaggaaaggcccaagatcccccagagaggggaggacagggcatggccctttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctcccccacctgtcctgcagtc
tttgtcttttccttctttttttgccacccttcaggaaccctgtatggtttttagtttaaattaaaggga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
tggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg

> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatccttaccaaccgtggcatc
gcccagatggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaacccctccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaagggtca
aaggaaaggcccaagatcccccagagaggggaggacagggcatggccctttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctcccccacctgtcctgcagtc
tttgtcttttccttctttttttgccacccttcaggaaccctgtatggtttttagtttaaattaaaggga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
tggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg

Approaches to gene finding

- Homology searches
  - BLAST, Procrustes
- Ab initio
  - Genscan, Genie, GeneID
- Hybrids
  - GenomeScan, GenieEST, Twinscan, SGP, ROSETTA, CEM, TBLASTX, SLAM

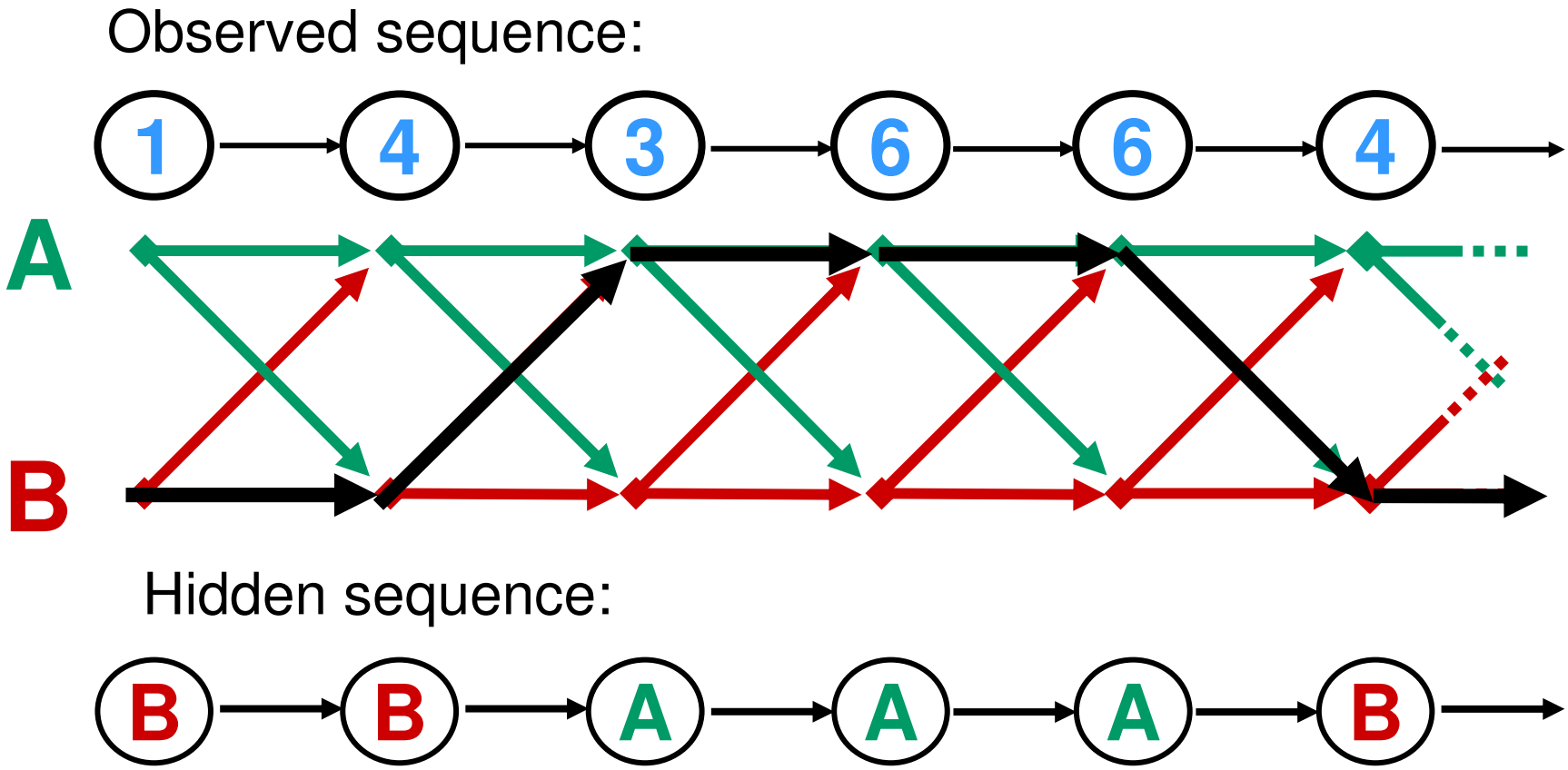
A simple Hidden Markov Model (HMM)



Initial distribution:

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

A lattice view



Two fundamental problems

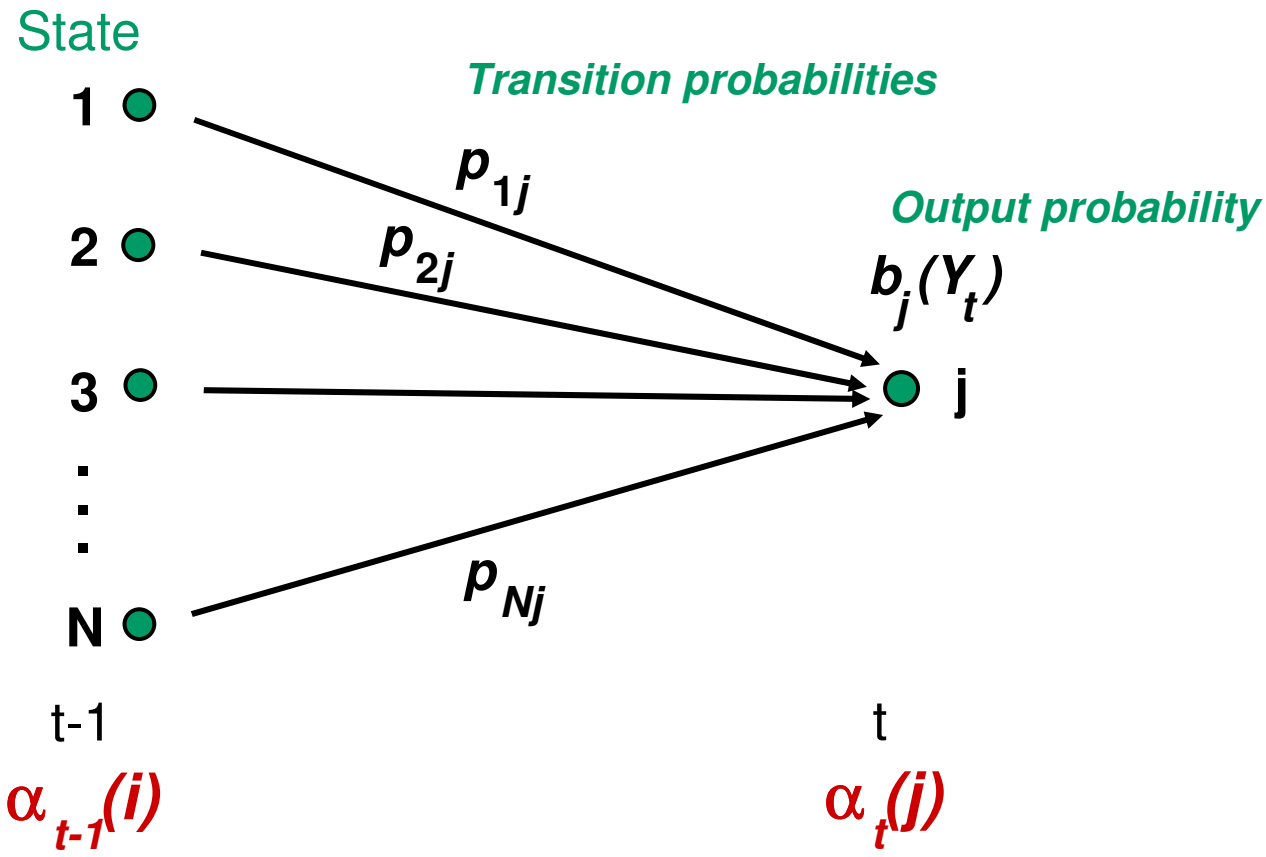
- The probability of the observed data given the model.

  The forward algorithm

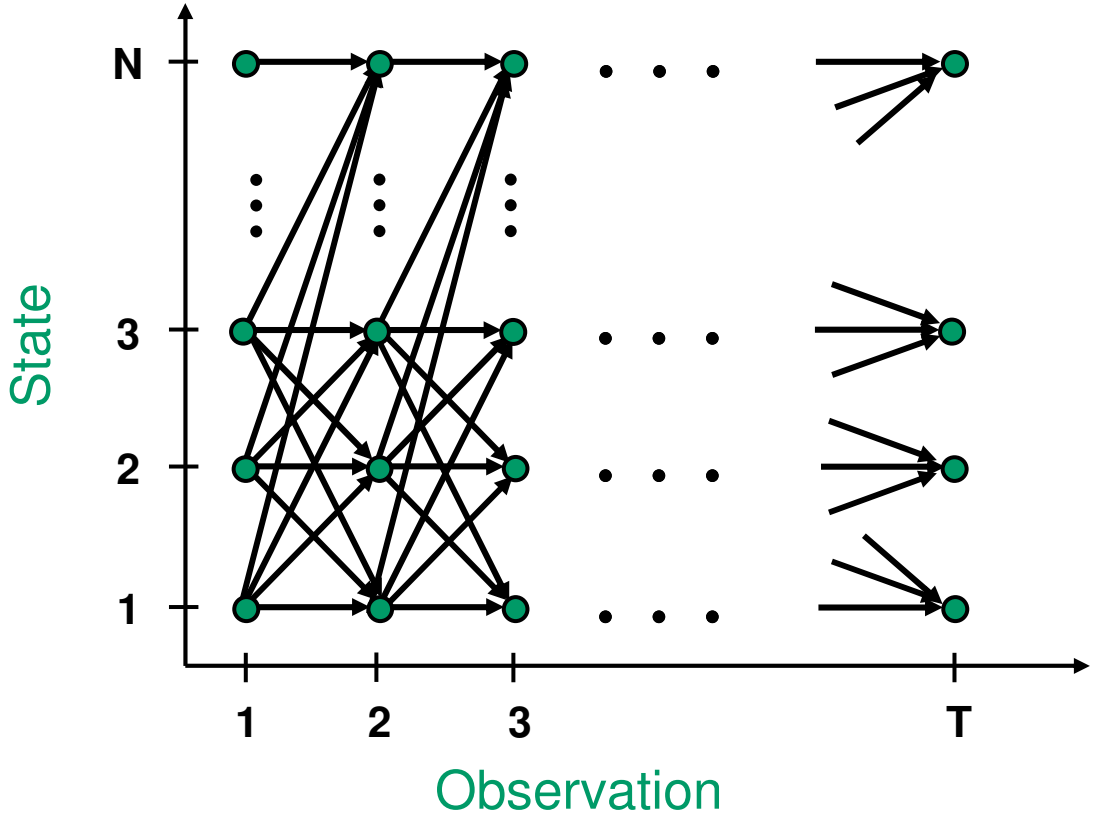
- The best hidden state sequence given the data.

  The Viterbi algorithm

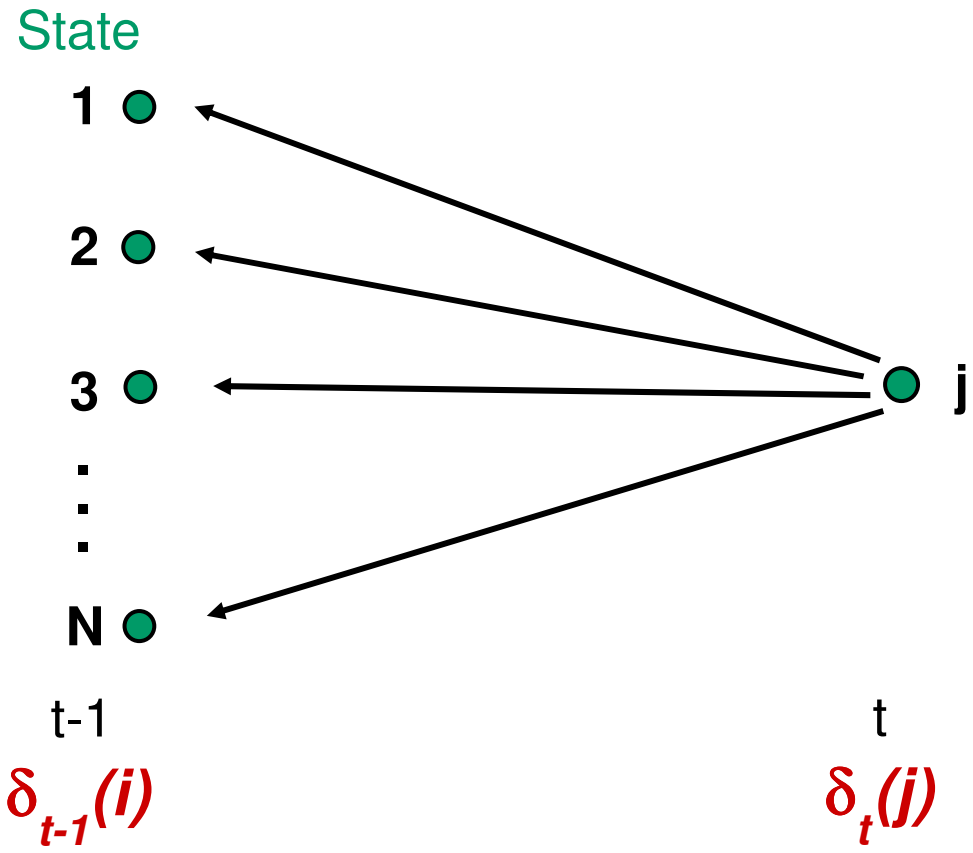
The forward algorithm - α



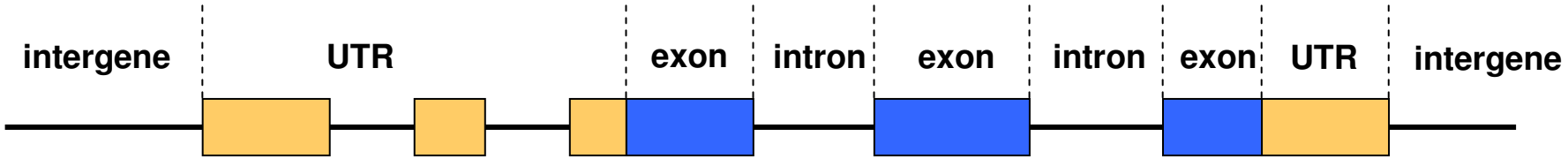
The forward algorithm



The Viterbi algorithm - δ



HMMs for gene finding



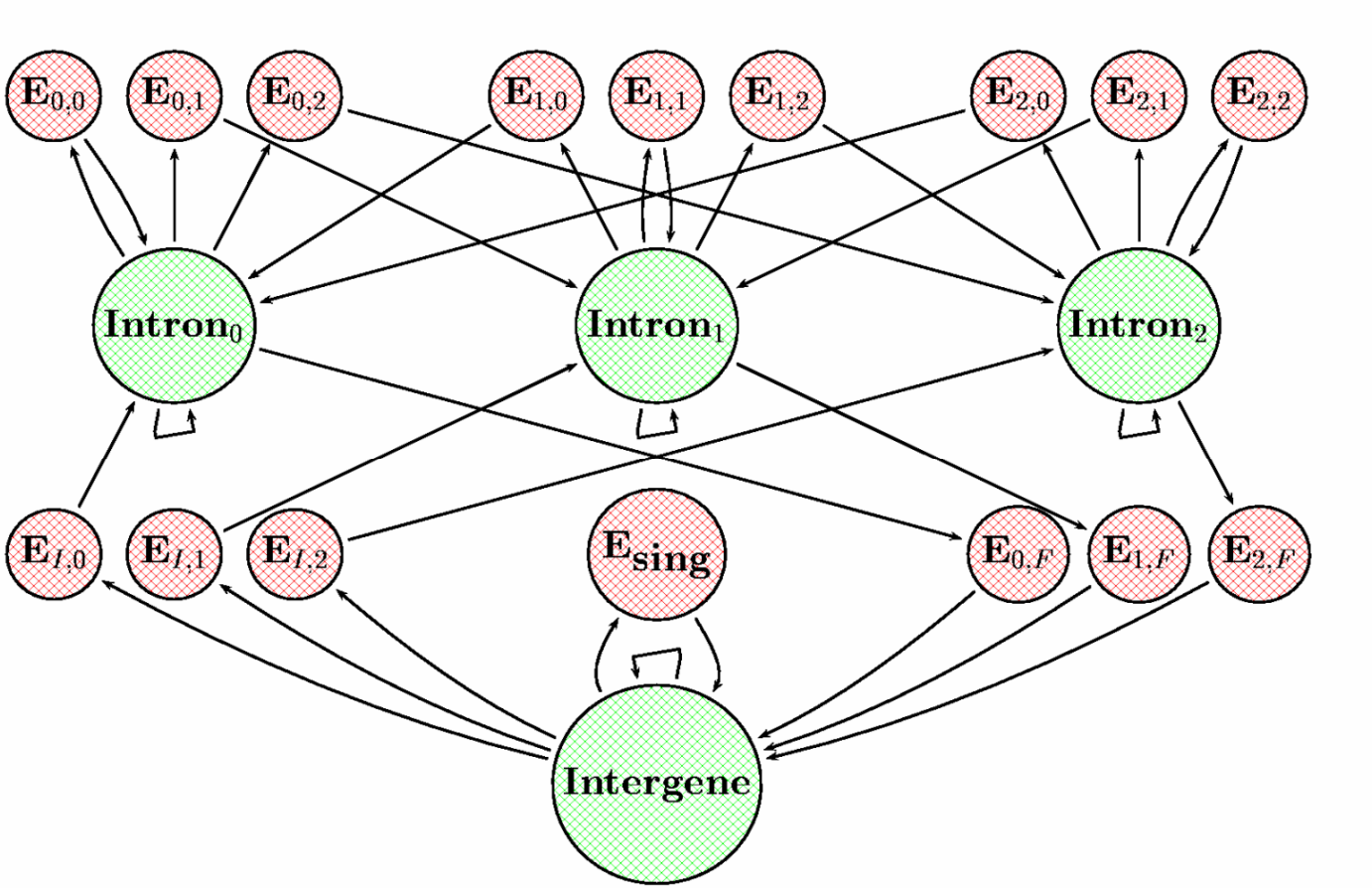
Observed:

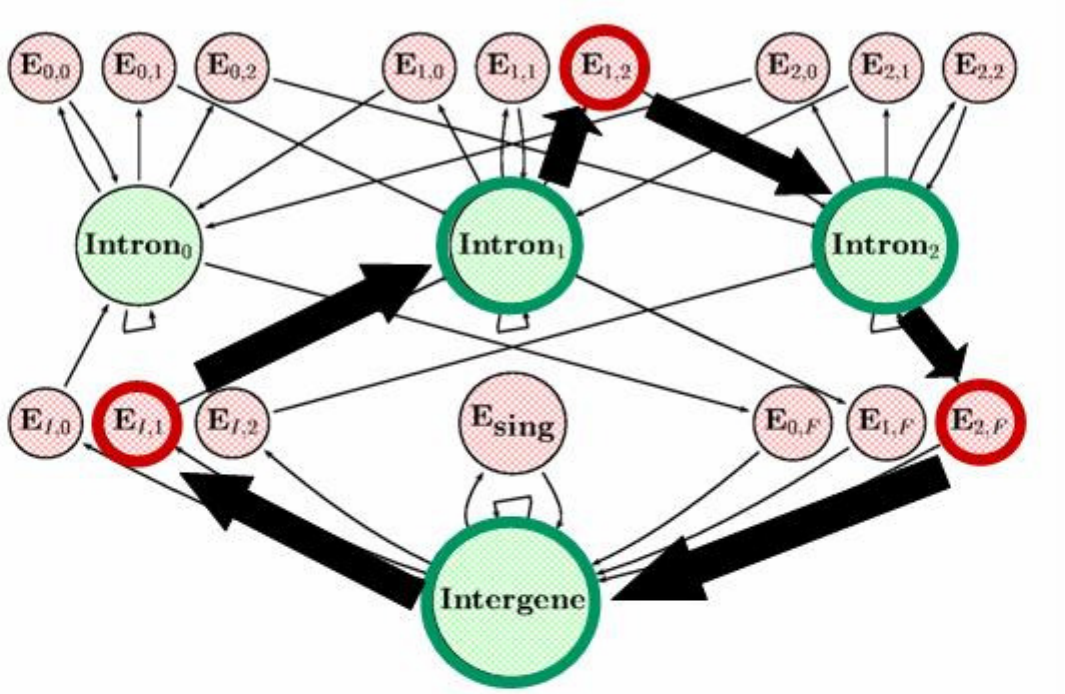
CTTGATGCTGGCACGTTCTGCTTCATCGGAGACAAATTACGGCTTTCCGGAGCA

Hidden:

CTTGATGCTGGCACGTTCTGCTTCATCGGAGACAAATTACGGCTTTCCGGAGCA

Generalized HMM





Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

Sequence alignment

- Pair HMMs

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Sequence alignment

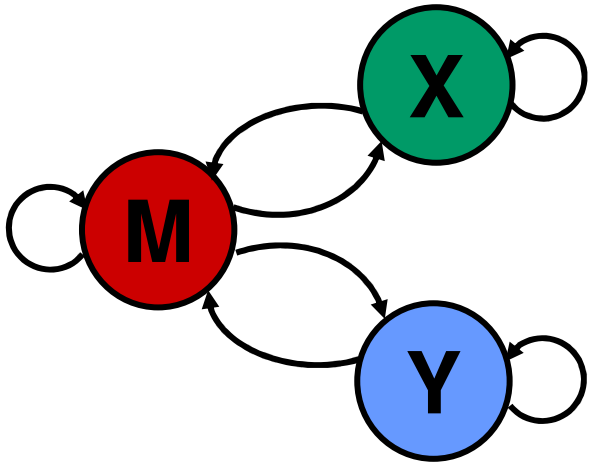
```
 50     .    :    .    :    .    :    .    :    .    :
247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA   AGAGC
    |:   ||  ||||:  |||| --:||  ||| |::|   |||---||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG

100     .    :    .    :    .    :    .    :    .    :
292 TTC          CTACAGAAAAGTCCCAGCAAGGAGCCACACTTCACTG
    |||----------|| |   |::| |: ||||::|:||:-||  ||:| |
418 TTCTGGCTACGCTCTCCCTTAGGGACTGAGCAGAGGGCT CAGGTCGCGG

150     .    :    .    :    .    :    .    :    .    :
332               ATGTCGAGGGGAAGACATCATTCGGGATGTCAGTG
    ---------------|||||||||||||||||||:||||||||||||
467 TGGGAGATGAGGCCAATGTCGAGGGGAAGACATCATTTGGGATGTCAGTG

200     .    :    .    :    .    :    .    :    .    :
367 TTCAACCTCAGCAATGCCATCATGGGCAGCGGCATCCTGGGACTCGCCTA
    |||||:||||||||:||||||||||||||:|| ||:||||||:||||||||
517 TTCAATCTCAGCAACGCCATCATGGGCAGTGGAATTCTGGGGCTCGCCTA
```

Pair HMMs for alignment



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

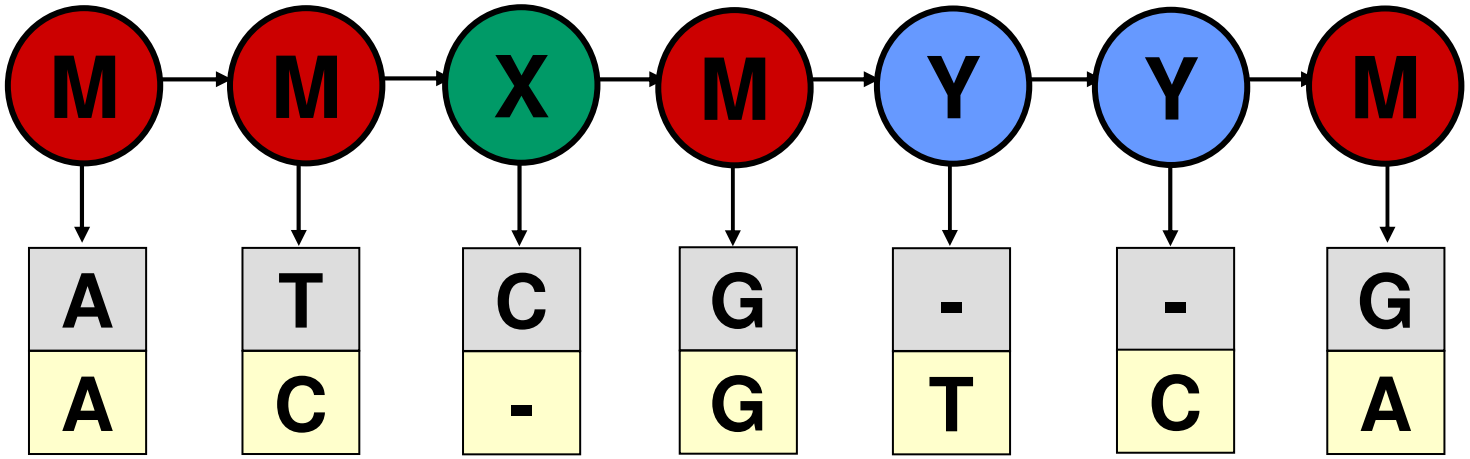
Output sequence:

ATCG- - G
AC- GTCA

Observed sequences:

ATCGG
ACGTCA

Pair HMMs



Output sequence:

ATCG- - G

AC- GTCA

Observed sequences:

ATCGG

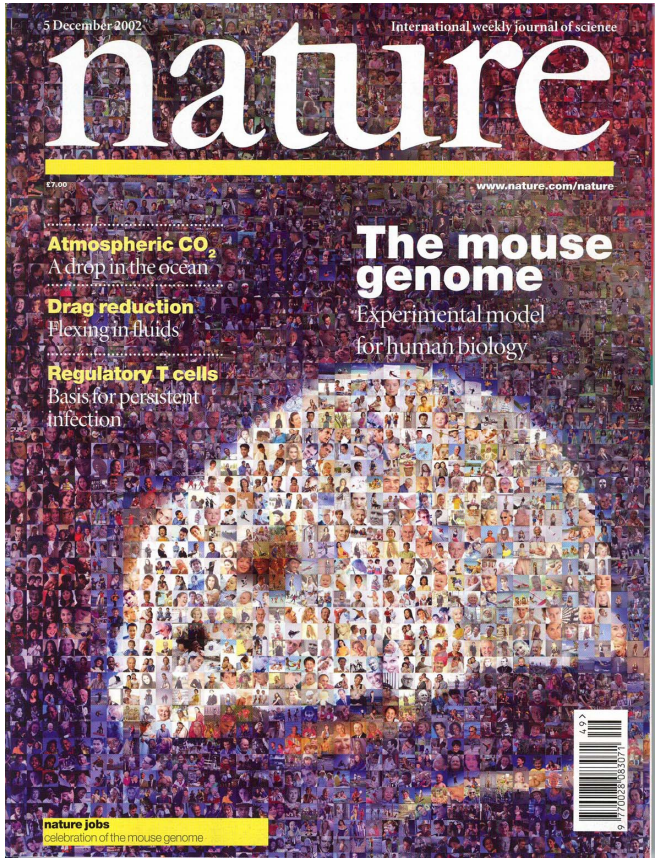
ACGTCA

Comparative gene finding

- Generalized Pair HMMs

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Comparing human and mouse

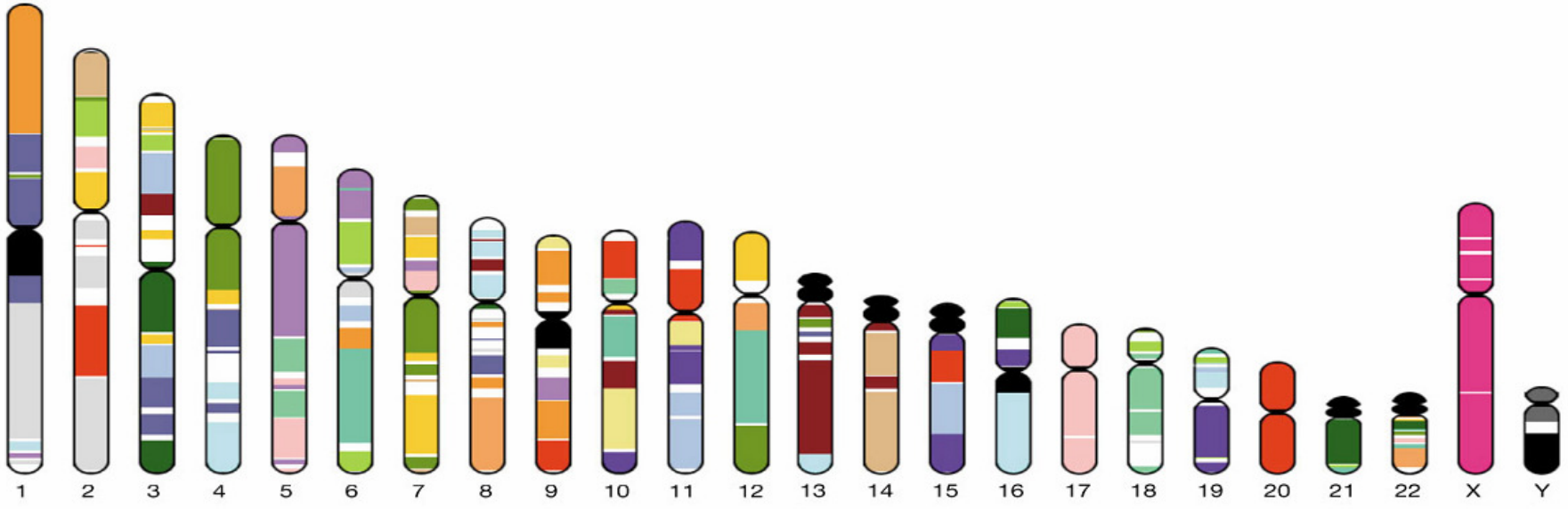


ITWM

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

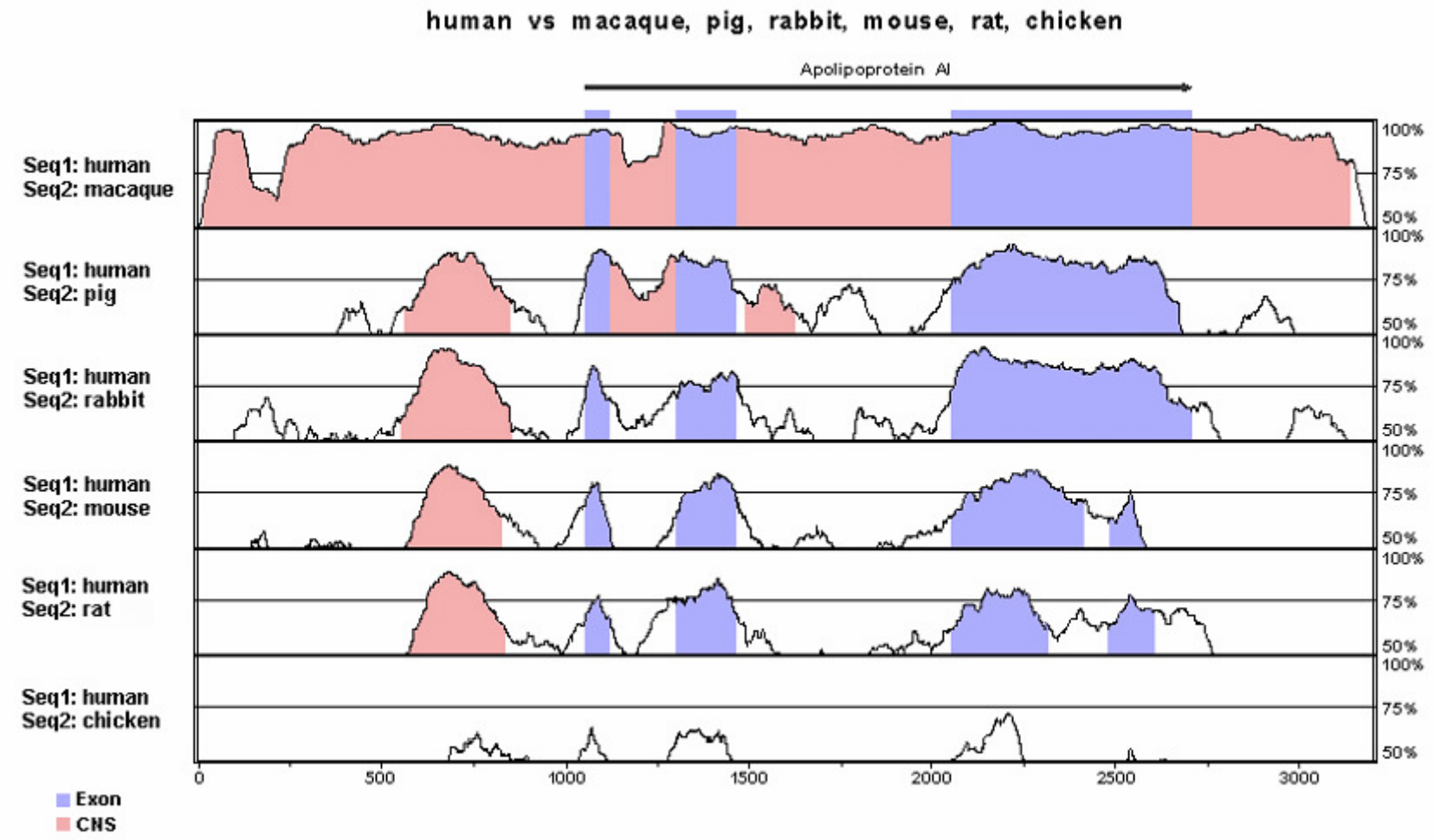
Why mouse?

Human

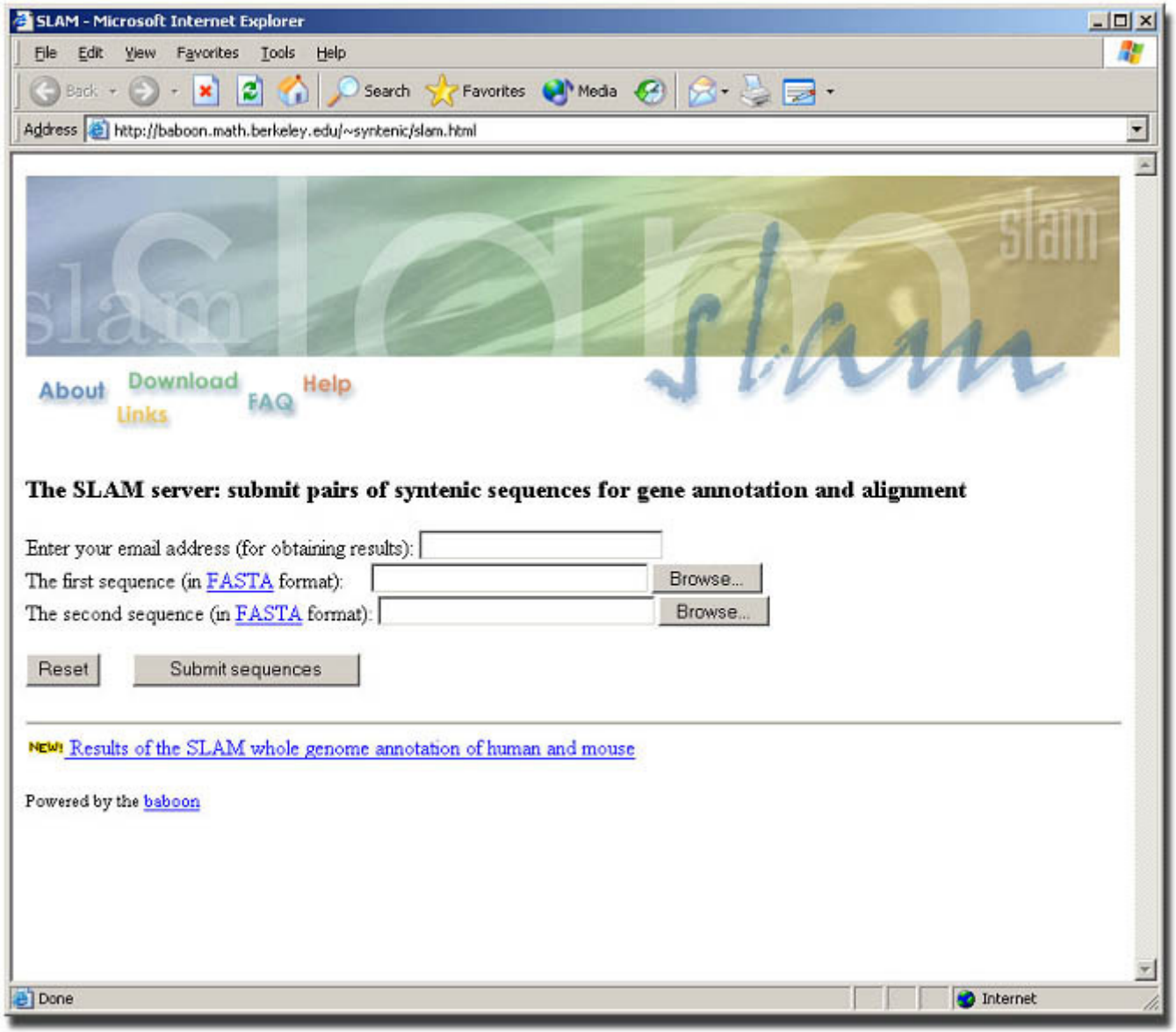

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Mouse


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Y

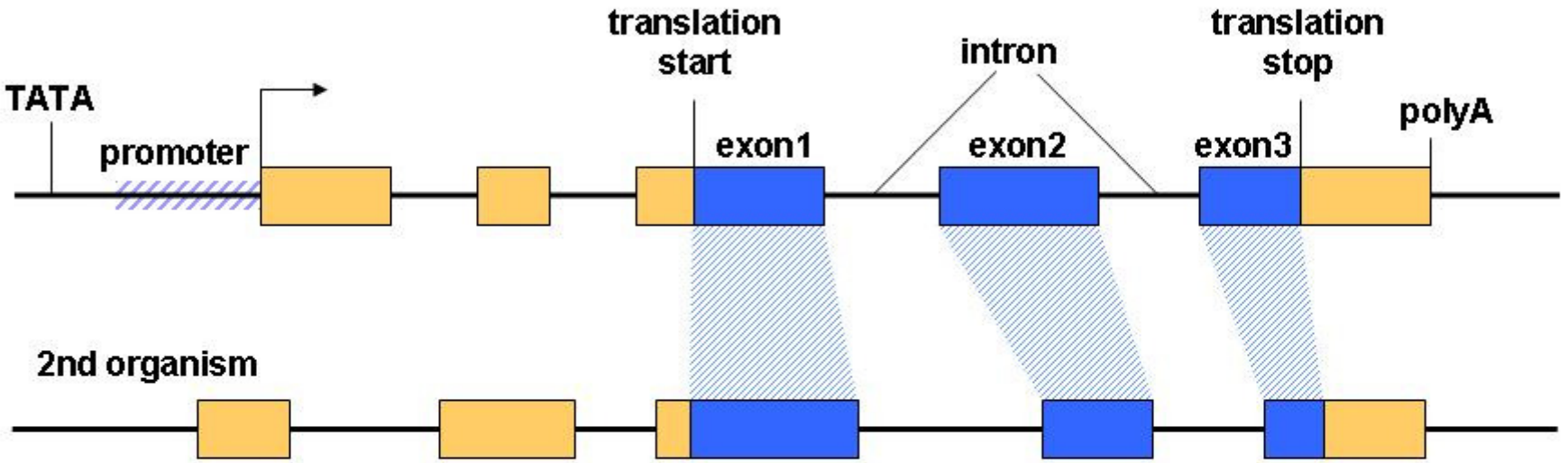


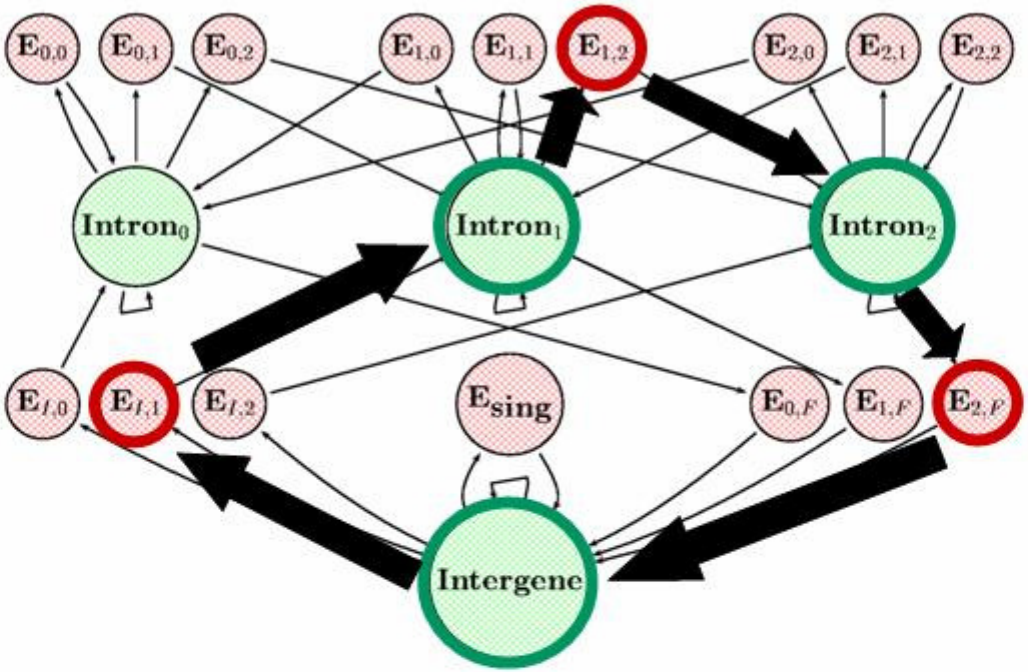
Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

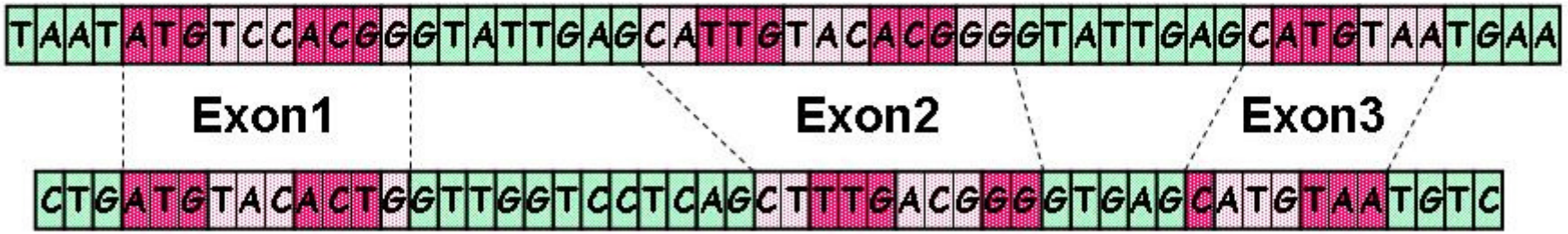


http://bio.math.berkeley.edu/slam/

Comparative gene finding in SLAM







Fraunhofer ITWM

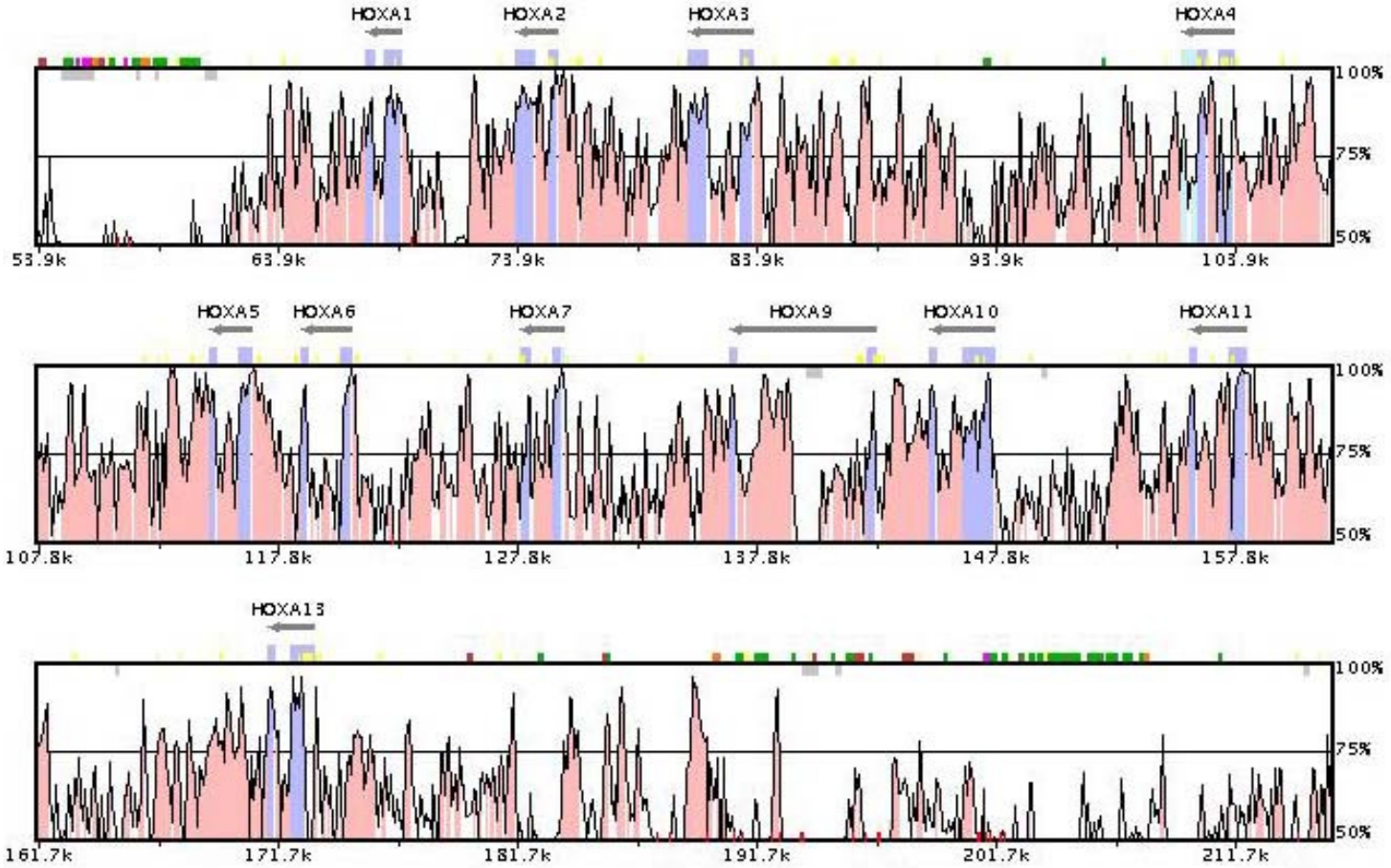
CHALMERS
Research Centre
Industrial Mathematics

The Rosetta set

Local	SN	SP	AC
Genscan	0.96	0.91	0.93
Twinscan	0.96	0.94	0.94
SGP-1	0.94	0.96	0.94
SLAM	0.95	0.98	0.96

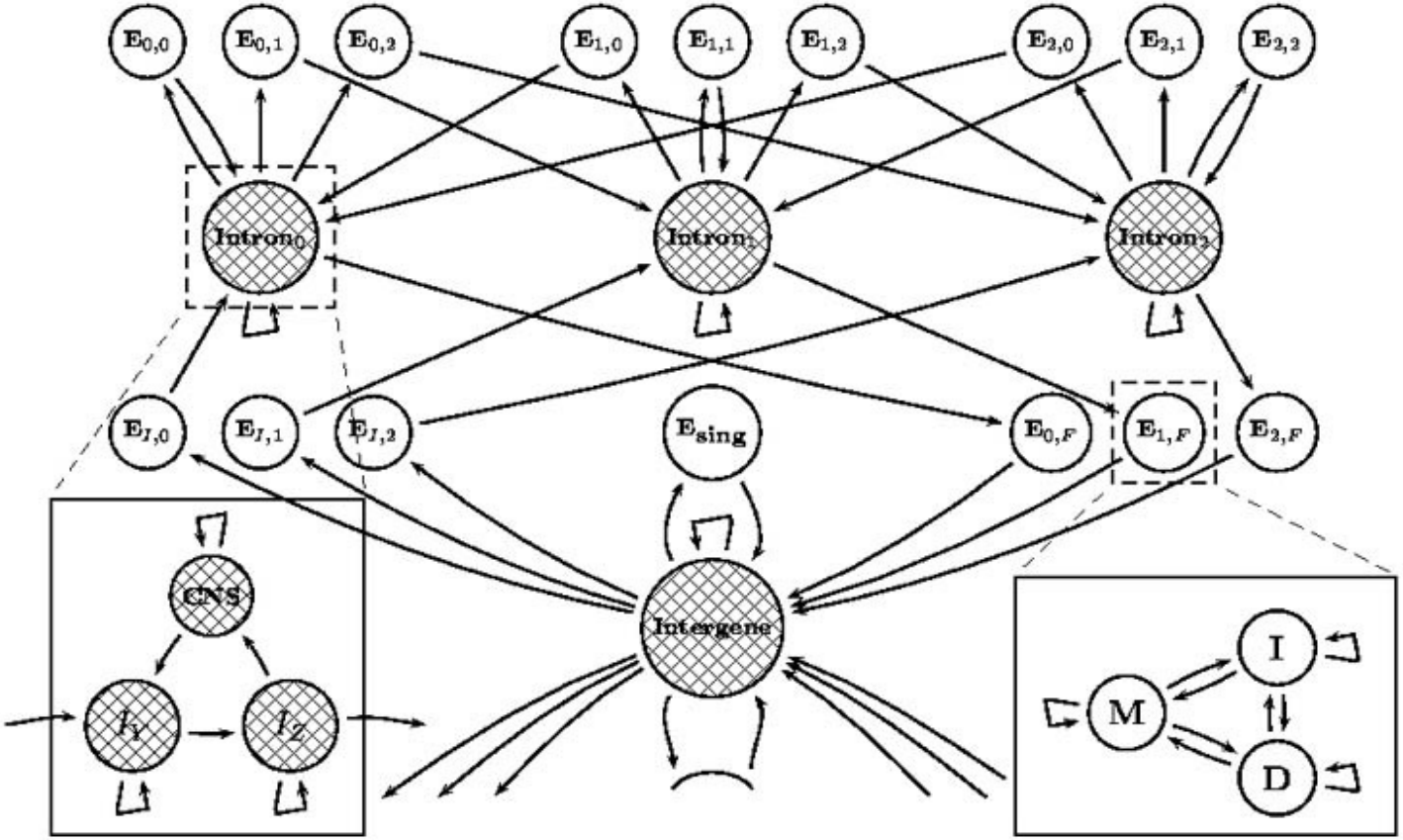
Global	SN	SP	AC
Genscan	0.87	0.87	0.84
Twinscan	0.89	0.92	0.89
SGP-2	0.88	0.94	0.89
SLAM	0.87	0.96	0.90

A problem: HoxA human-mouse



Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

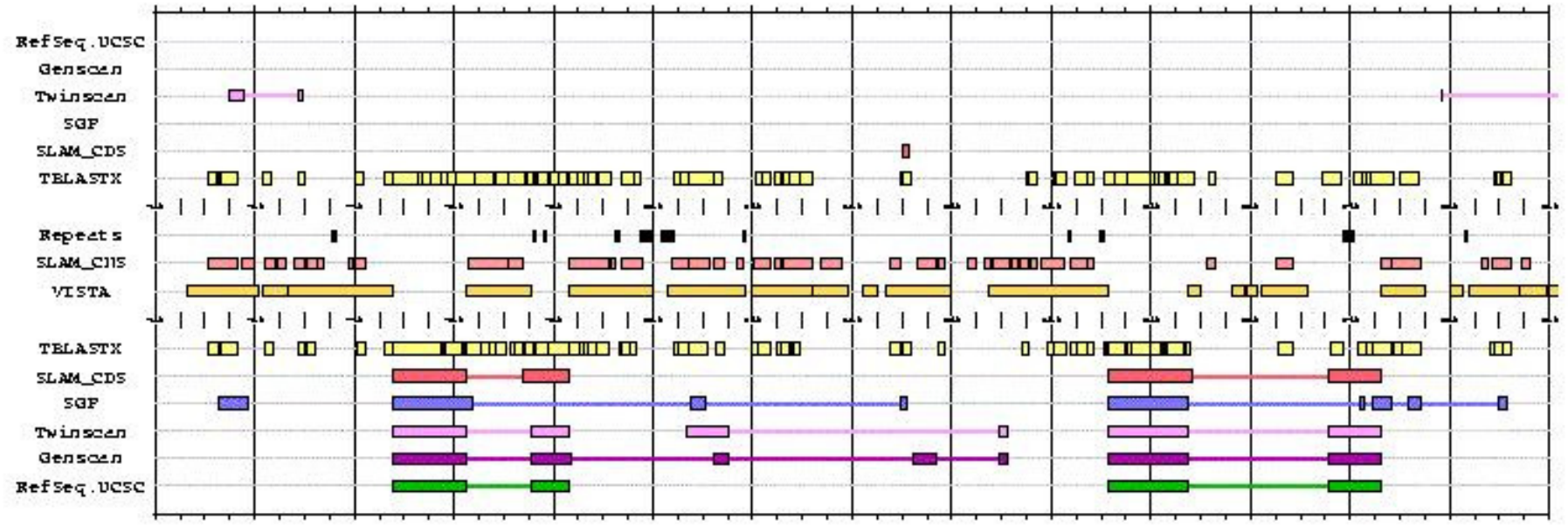
The SLAM model



Fraunhofer ITWM

CHALMERS
Research Centre
Industrial Mathematics

Example: HoxA2 and HoxA3

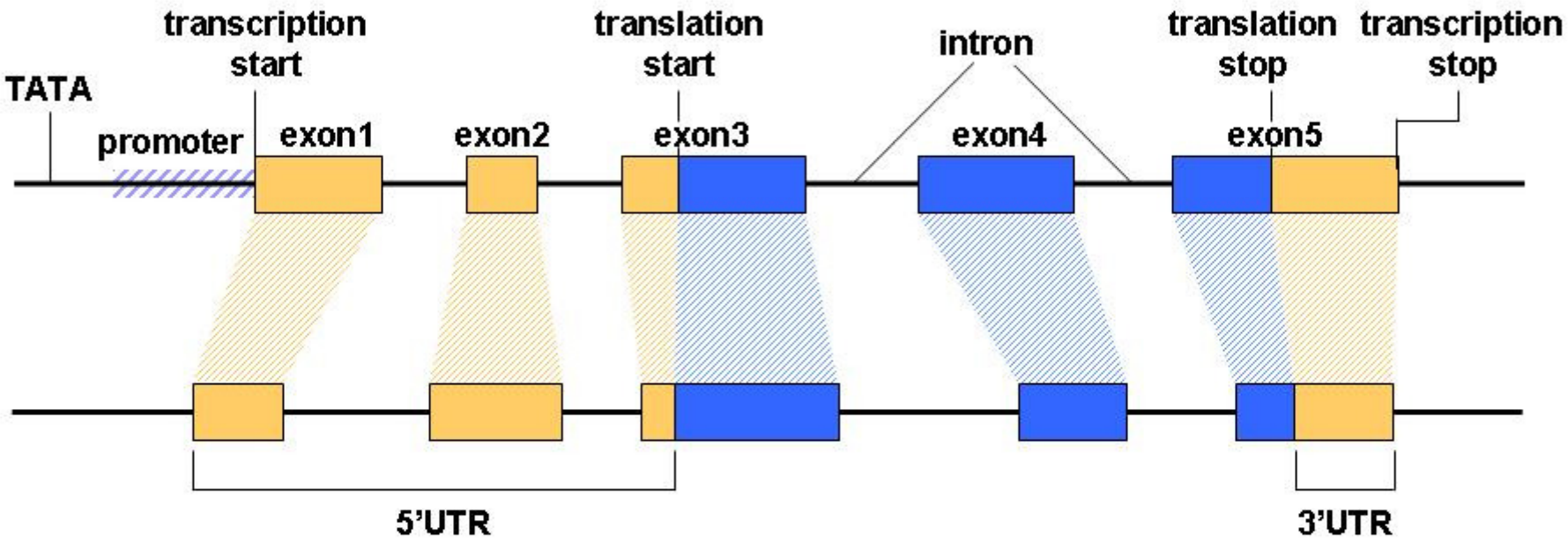


VISTA TBLASTX SLAM Twinscan

RefSeq SLAM CNS SGP-2 Genscan

Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

Comparative UTR prediction



Reducing computational complexity

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

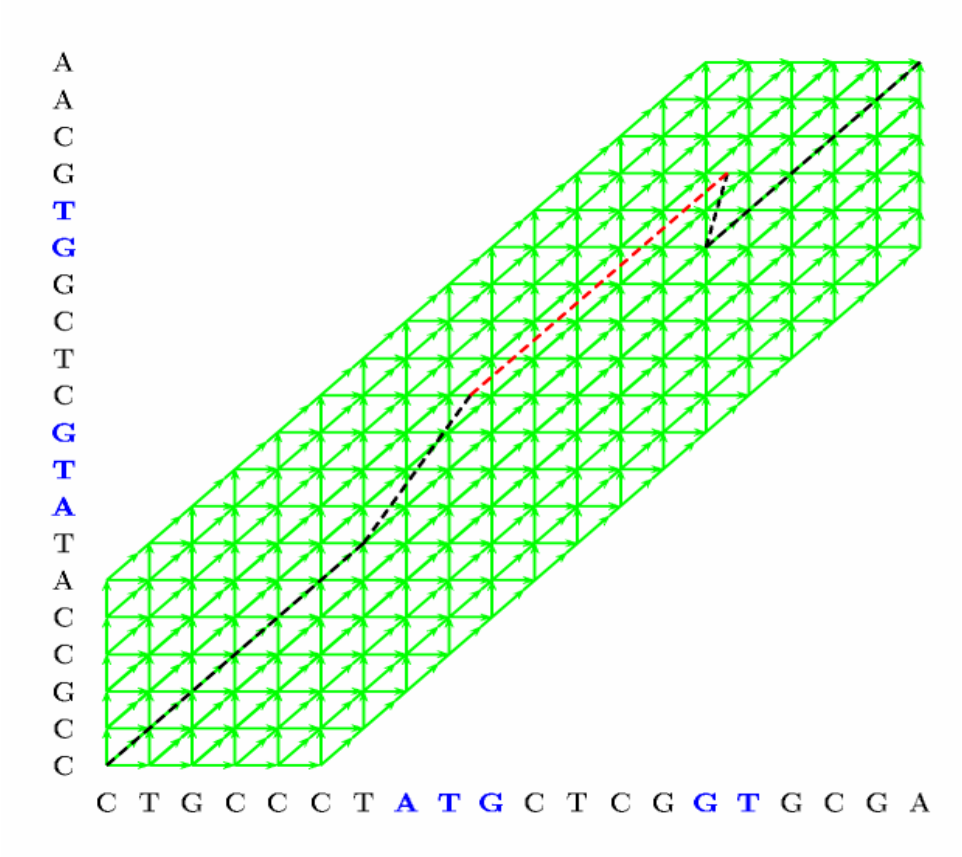
Computational complexity

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

N = no. states T = length seq1

D = max duration U = length seq2

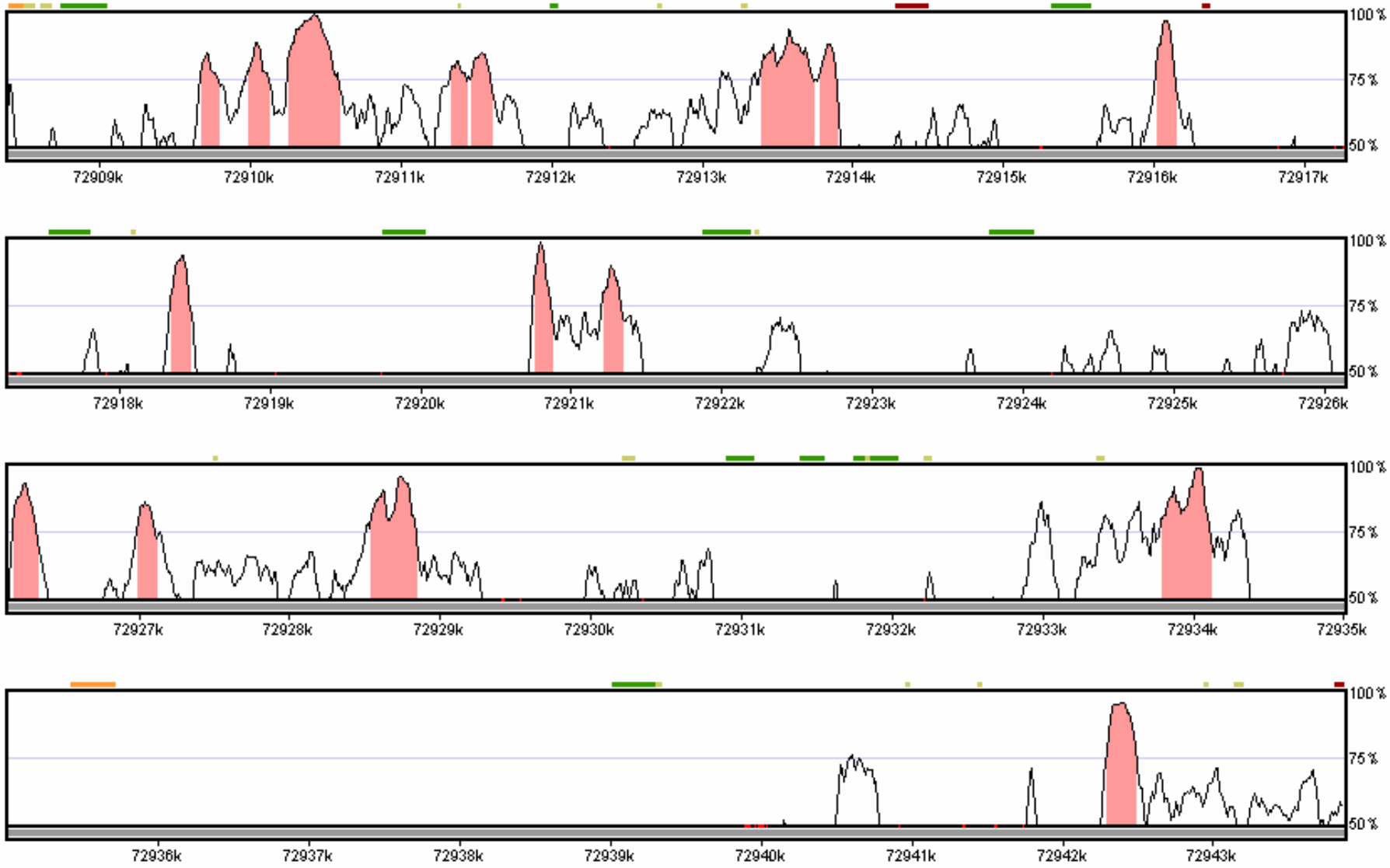
Approximate alignment

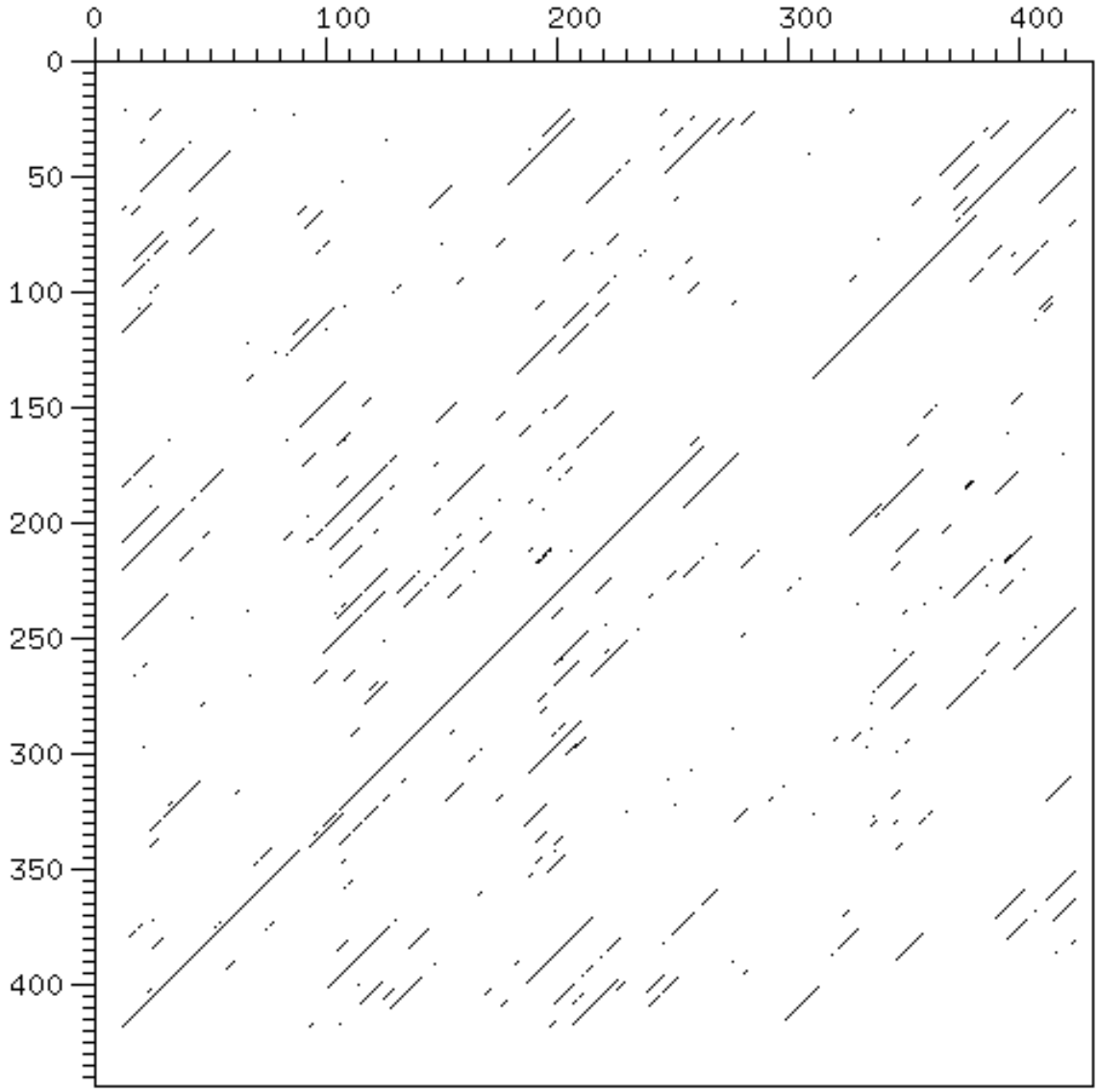


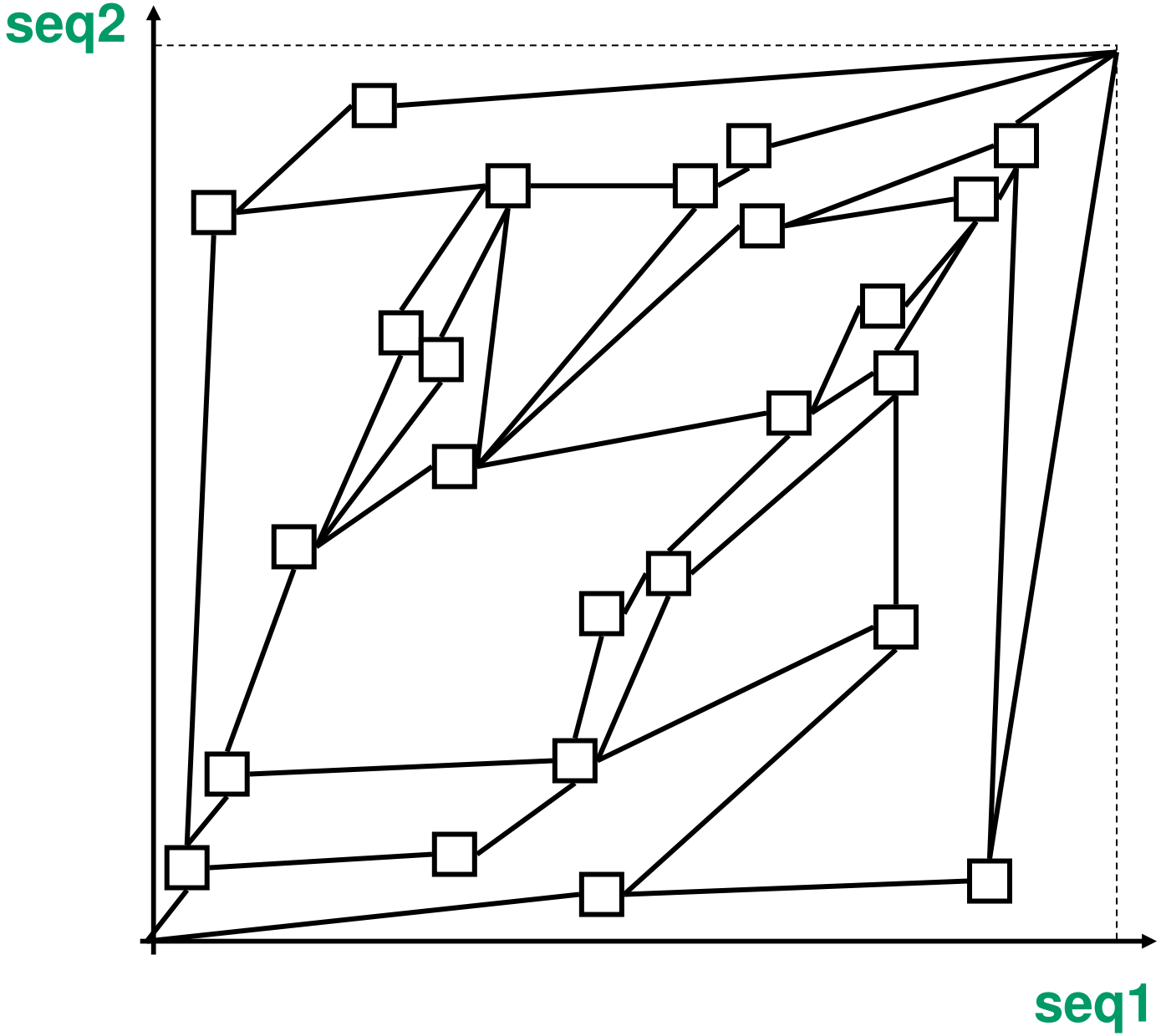
Reduces

TU*-factor to *hT

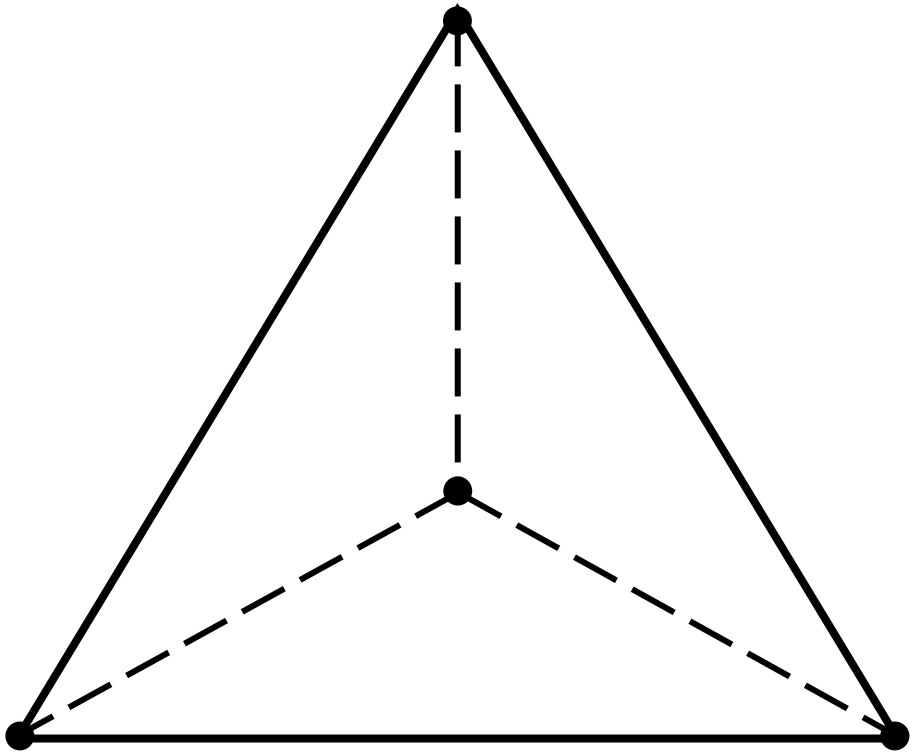
Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics


100 %
75 %
50 %
72909k
72910k
72911k
72912k
72913k
72914k
72915k
72916k
72917k
72918k
72919k
72920k
72921k
72922k
72923k
72924k
72925k
72926k
72927k
72928k
72929k
72930k
72931k
72932k
72933k
72934k
72935k
72936k
72937k
72938k
72939k
72940k
72941k
72942k
72943k

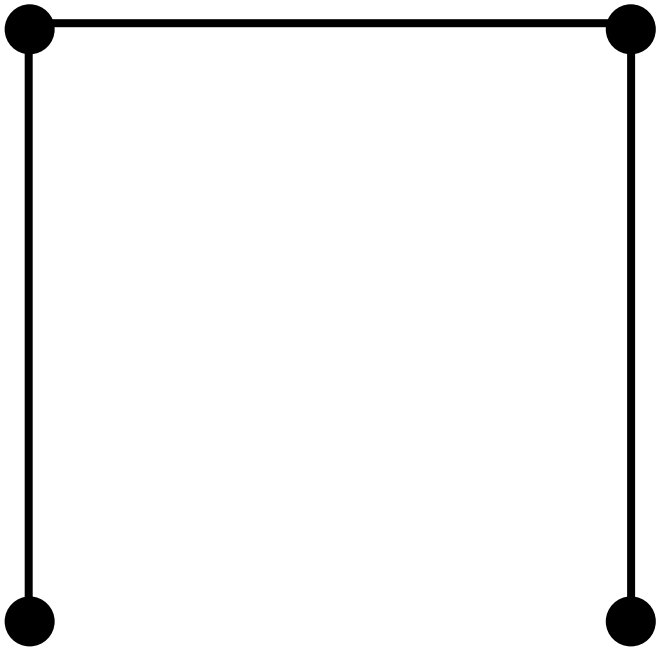

0
100
200
300
400
0
50
100
150
200
250
300
350
400


seq2
seq1

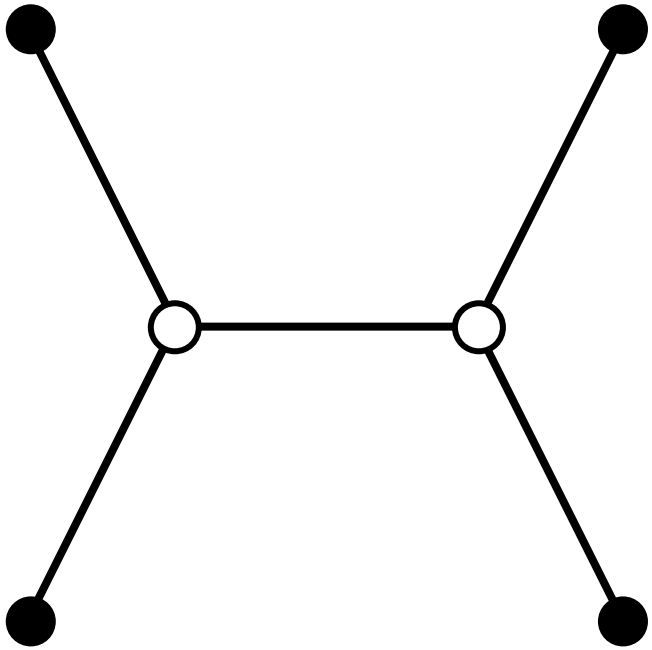
Steiner trees



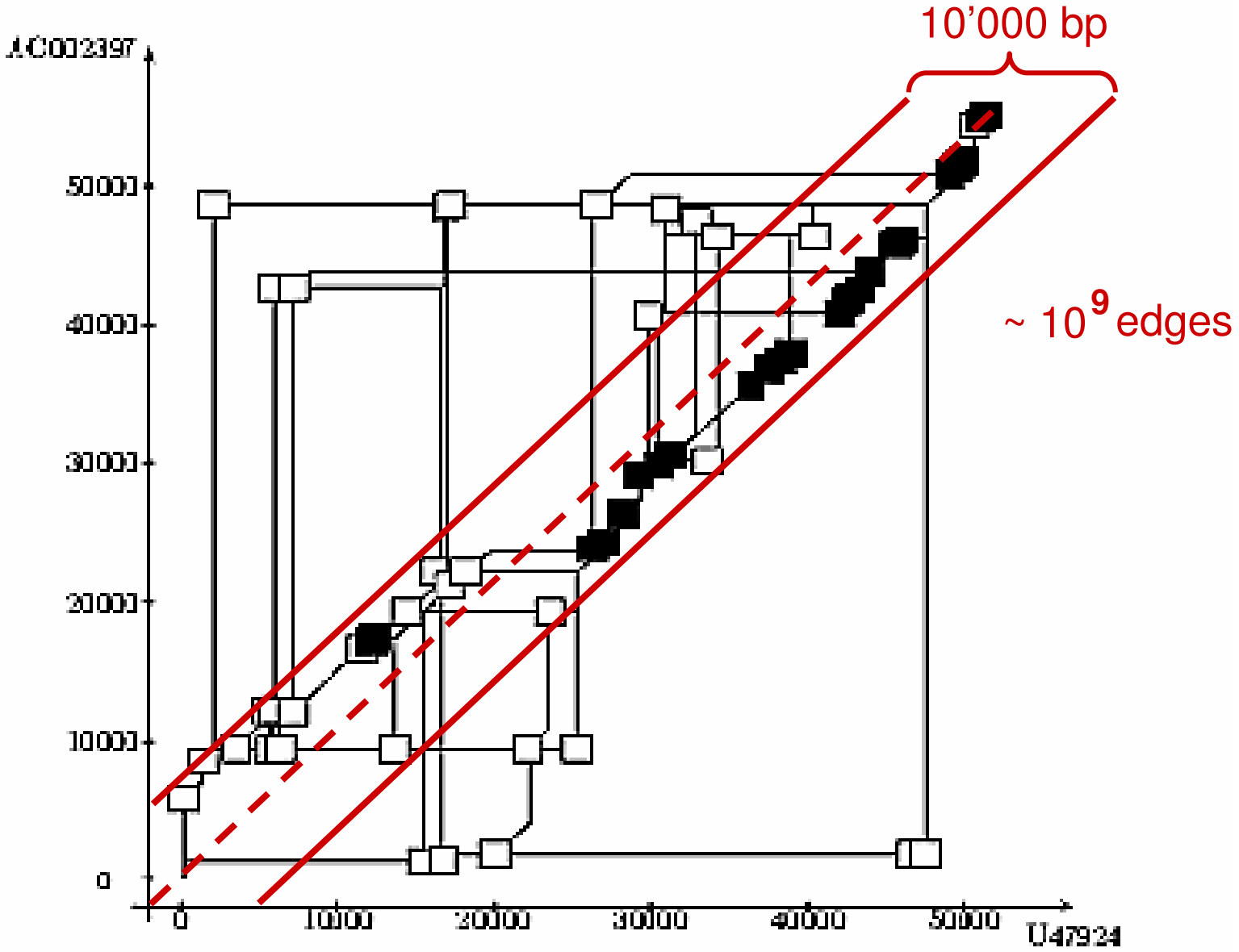
Steiner trees



Minimum spanning tree



Steiner tree

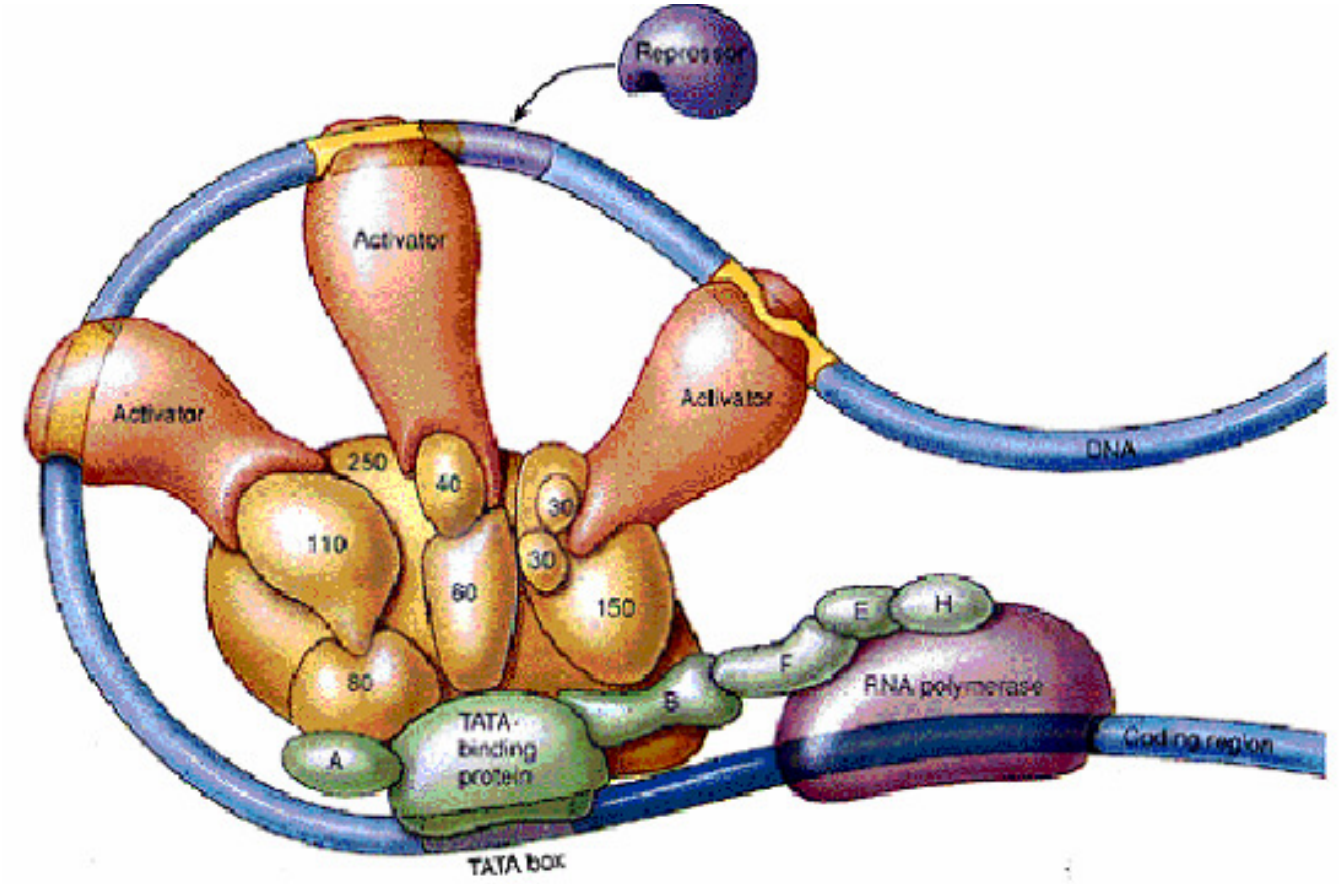


2-approximation in $O(n^3)$

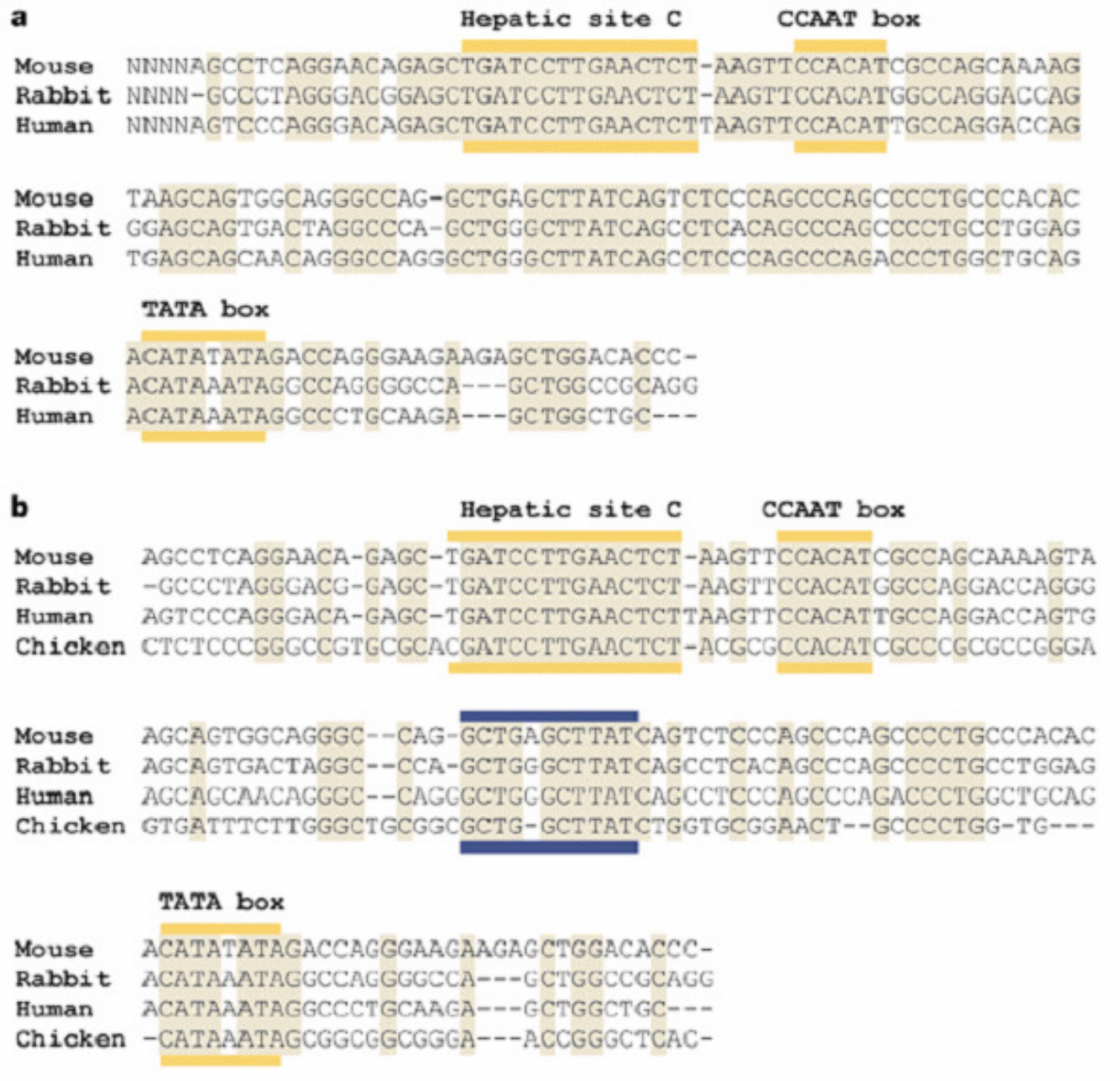
Regulatory regions

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Regulatory complex



Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics


a
Hepatic site C
CCAAT box
Mouse NNNNAGCCTCAGGAACAGAGCTGATCCTTGAACTCT-AAGTTCCACATCGCCAGCAAAAG
Rabbit NNNN-GCCCTAGGGACGGAGCTGATCCTTGAACTCT-AAGTTCCACATGGCCAGGACCAG
Human NNNNAGTCCCAGGGACAGAGCTGATCCTTGAACTCTTAAGTTCCACATTGCCAGGACCAG
Mouse TAAGCAGTGGCAGGGCCAG-GCTGAGCTTATCAGTCTCCCAGCCCAGCCCCTGCCCACAC
Rabbit GGAGCAGTGACTAGGCCCA-GCTGGGCTTATCAGCCTCACAGCCCAGCCCCTGCCTGGAG
Human TGAGCAGCAACAGGGCCAGGGCTGGGCTTATCAGCCTCCCAGCCCAGACCCTGGCTGCAG
TATA box
Mouse ACATATATAGACCAGGGAAGAAGAGCTGGACACCC-
Rabbit ACATAAATAGGCCAGGGGCCA---GCTGGCCGCAGG
Human ACATAAATAGGCCCTGCAAGA---GCTGGCTGC---
b
Hepatic site C
CCAAT box
Mouse AGCCTCAGGAACA-GAGC-TGATCCTTGAACTCT-AAGTTCCACATCGCCAGCAAAAGTA
Rabbit -GCCCTAGGGACG-GAGC-TGATCCTTGAACTCT-AAGTTCCACATGGCCAGGACCAGGG
Human AGTCCCAGGGACA-GAGC-TGATCCTTGAACTCTTAAGTTCCACATTGCCAGGACCAGTG
Chicken CTCTCCCGGGCCGTGCGCACGATCCTTGAACTCT-ACGCGCCACATCGCCCGCGCCGGGA
Mouse AGCAGTGGCAGGGC--CAG-GCTGAGCTTATCAGTCTCCCAGCCCAGCCCCTGCCCACAC
Rabbit AGCAGTGACTAGGC--CCA-GCTGGGCTTATCAGCCTCACAGCCCAGCCCCTGCCTGGAG
Human AGCAGCAACAGGGC--CAGGGCTGGGCTTATCAGCCTCCCAGCCCAGACCCTGGCTGCAG
Chicken GTGATTTCTTGGGCTGCGGCGCTG-GCTTATCTGGTGCGGAACT--GCCCCTGG-TG---
TATA box
Mouse ACATATATAGACCAGGGAAGAAGAGCTGGACACCC-
Rabbit ACATAAATAGGCCAGGGGCCA---GCTGGCCGCAGG
Human ACATAAATAGGCCCTGCAAGA---GCTGGCTGC---
Chicken -CATAAATAGCGGCGGCGGGA---ACCGGGCTCAC-

How many genes do we have???

27’462

Acknowledgements

- Lior Pachter, UC Berkeley
- Simon Cawley, Affymetrix

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics