

# Computational methods of systems biology

**Esko Ukkonen**

Department of Computer Science  
University of Helsinki

<http://www.cs.helsinki.fi/u/ukkonen/>

***Lectures in Fall 2009***

## Contents of the course (tentative)

- Lecture 0: Introduction
- Lecture 1: Hidden Markov Models
- Lecture 2: Applications of HMM – profile-HMMs for sequence families
- Lecture 3: Applications of HMM – modeling transcription binding sites with position weight matrices
- Lecture 4: Transcription networks – basics
- Lecture 5: Autoregulation – a network motif
- Lecture 6: The feed-forward loop network motif
- Lecture 7: Temporal programs and the global structure of transcription networks
- Lecture 8: Network motifs in developmental, signal transduction, and neuronal networks
- Lecture 9: Kinetic proofreading
- Lecture 10: Metabolic networks

## Comments on the contents

- The first part of the course (lectures 1-3) is an introduction to Hidden Markov models and their applications (profile models of sequence families, PWMs, cis-regulation) in biological sequence analysis. The text book by Durbin et al is the main source.
- The second part (lectures 4-10) is on biological networks, based on the recent text book by Uri Alon. The book has a novel approach to this very messy field. The idea is to understand various biological networks in terms of different regulatory 'motifs' that occur unexpectedly often in these networks (and hence should have been conserved in the evolution).
- One lecture (given by Esa Pitkänen) will be an introduction to metabolic networks and flux analysis

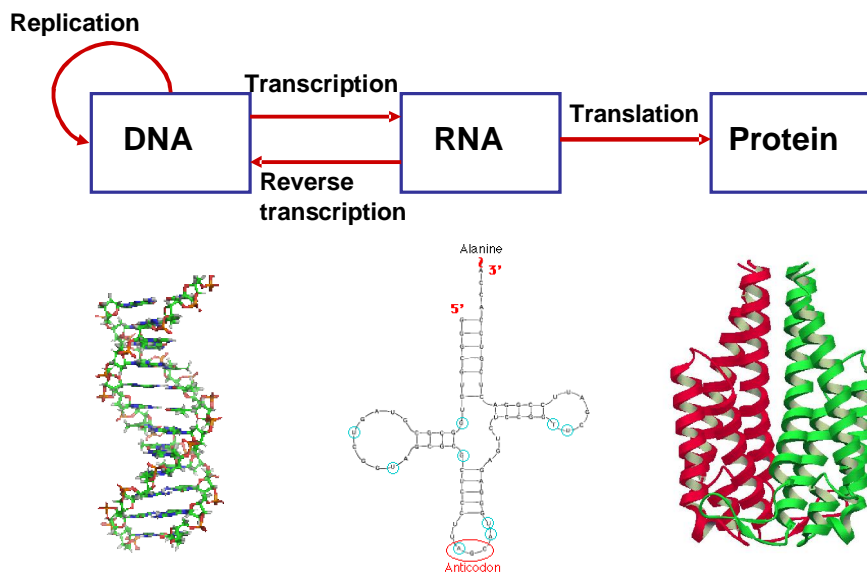
## Sources

- Text books
  - **R.Durbin, S.R.Eddy, A.Krogh & G.Mitchison: Biological sequence analysis. Cambridge University Press 1998**
    - Chapters 1.3, 3, 5 (lectures 1, 2, 3)
  - **U. Alon: An introduction to systems biology – Design principles of biological circuits. Chapman & Hall/CRC 2007**
    - Chapters 2, 3, 4, 5, 6, 9 (lectures 4-9)
  - **M. Zvelebil & J.O.Baum: Understanding Bioinformatics. Garland Science 2008**
    - Chapters 6.3 – 6.6 (lecture 3)
- Original articles

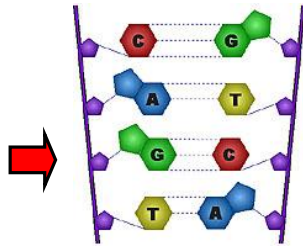
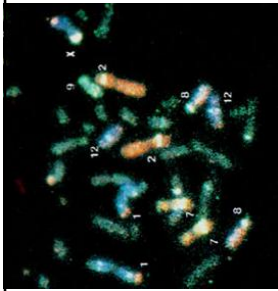
# Lecture 0: Introduction, background, summary of some topics

This lecture recalls some core bioinformatics problems and computational techniques that the participants are supposed to be familiar with. Some examples of the course content are also mentioned.

## Information flow in a cell



## DNA sequencing and genome projects



cgccgagtgacag  
agacgctaatacagg  
ctgtgttctcaggat  
gcgtaccgagtgga  
agacagcagcacg  
accag...

## DNA fragment assembly problem

cctcgagttaagtactgcccgcggcttcaacggatctgtcgggagtcg

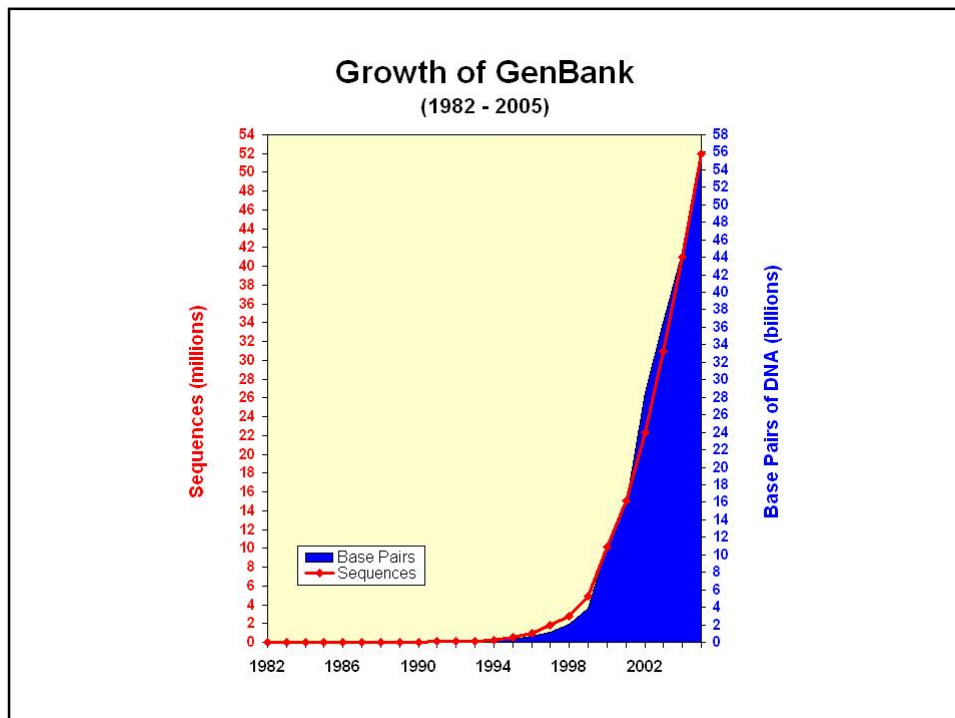


Re-assemble  
the puzzle?

cctcgagttaa  
tacttaactcgag  
cgggcagtacttaa  
aagtactgcccgcg  
gcccgcggcttcaacggat  
cccgcggcttcaacggatctgtg  
cccgacacagat  
tgtgtcgggagtcg

## Some numbers (Human Genome)

- Total length about 3 000 000 000 bp
- Celera's fragment data (Feb 2001):  
27 million fragments, each 150-800 bp
- at least 7-fold coverage by fragments  
needed => total data length 7 x 3 billion bp



# How does the DNA program work?

```
cgccgagtgacagagacgc
taatcaggctgtgttctca
ggatgcgtaccgagtggga
gacagcagcagaccagcg
gtggcagagacccttgcag
acatcaagctctttgggaa
caagtggagcaccgatgat
gtacagccgatcaatgaca
tttccctaatgcaggatta
cattgcagtgcccaaggag
aagtatg...
```



# Pairwise alignment

(A) local

```
PI3-kinase DRHNSNIMVKDDGQLFHI DFG
cAMP PK DLKPENLLIDRRGGYIQVT DFG
```

(B) global

```
PI3-kinase HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELLFGNNEIIFKNGDDLRRQDMLT
cAMP PK GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDRFERIKTLGTGSEGRVML-
10 20 30 40 50
PI3-kinase 60 70 80 90 100 110
LqIIRIME--NIWQNRGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IRCKGGLKGA
cAMP PK ---VKHMETGNHYAMKILDKQKVVK-----EKQIEHTLNEKRILQAVNFPFLVKLEF
50 60 70 80 90 100
PI3-kinase 120 130 140 150 160
GFNSHT-LHQWLKDKNKGEIYDAA--IDLTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK SFKDNSNLYVMVEYVPGEMESHLLRRIGRFSEPHARFYAAQIVLTFEYLSLDTIYRDLK
110 120 130 140 150 160
PI3-kinase 170 180 190 200 210
GQLFHIDEGHFLDHHKKKFGYKRERVP-----EVLTDQFL--IVISKGAQECTKTREFE
cAMP PK PENLLIDGGYI--QVTDGFAK-RVKGRTWXLCGTPPEYLAPEIILSKGYNKAVDQWALG
170 180 190 200 210
PI3-kinase 230 240 250 260 270
RE-qEMC--YKAYLAIRRHANLFINEFSMMLGSGMPELQSFDDIAYIRKTLALDKTEGEA
cAMP PK VLIYEMAAGYPPPEFA-D&PPIQIYEKIVSGKVR--FISHFSSDLKDLRNLQVDLTKR--
230 240 250 260 270
PI3-kinase 280 290 300
LEYFMKQMNDAHHGGWTTKMDWI-----EHTIKQHALN---
cAMP PK EGNLKNQVNDIKNHKWFAITDWTIAYQRKVEAPFIPKFKGGDTSNEDDYEEEEIRVXIN
280 290 300 310 320 330 340
```

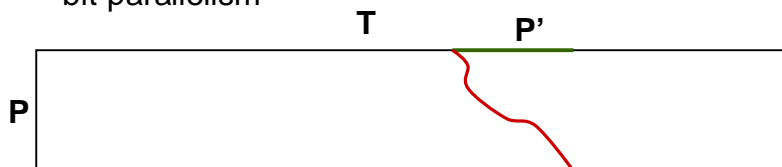
$$d_{i,j} = \min(\text{if } a_i=b_j \text{ then } d_{i-1,j-1} \text{ else } \infty, \\ d_{i-1,j} + 1, \\ d_{i,j-1} + 1)$$

A\B		S	T	O	C	K	H	O	L	M
	0	1	2	3	4	5	6	7	8	9
T	1	2	1	2	3	4	5	6	7	8
U	2	3	2	3	4	5	6	7	8	9
K	3	4	3	4	5	4	5	6	7	8
H	4	5	4	5	6	5	4	5	6	7
O	5	6	5	4	5	6	5	4	5	6
L	6	7	6	5	6	7	6	5	4	5
M	7	8	7	6	7	8	7	6	5	4
A	8	9	8	7	8	9	8	7	6	5

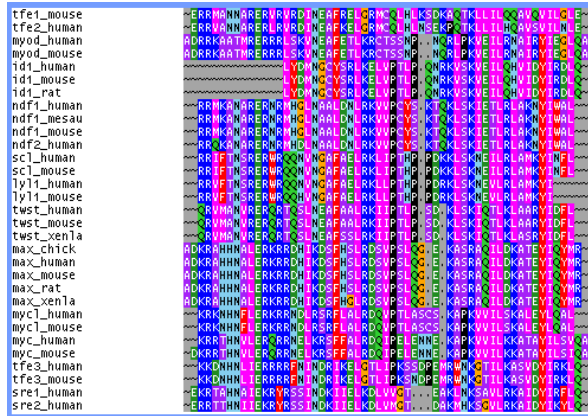
optimal alignment by trace-back = **Viterbi!**  $d_{ID}(A,B)$

## Search problem

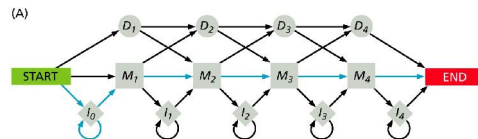
- find approximate occurrences  $P'$  of pattern  $P$  in text  $T$  such that  $d(P,P')$  small
- dyn progr with small modification:  $O(mn)$
- lots of (practical) improvements:
  - distance bound  $k \rightarrow O(kn)$  search
  - utilize regularities of the dp table
  - **filtration approach: BLAST (big success!)**
  - bit-parallelism



# Multiple alignment



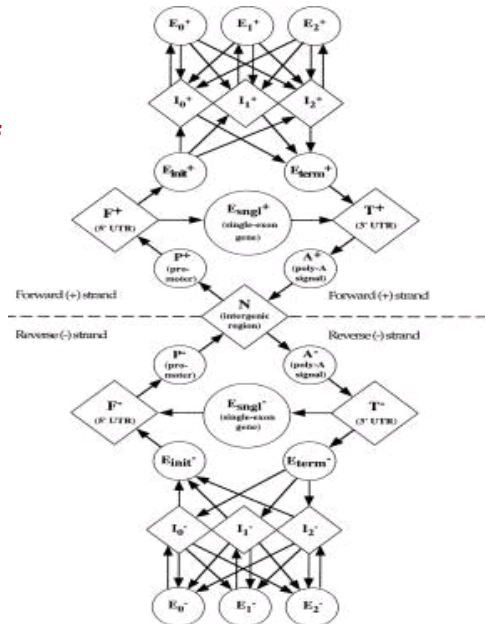
# Hidden Markov Models (HMM) for Sequence Families





## HMM architecture of GENESCAN

Prediction of genes



## Sequence motifs

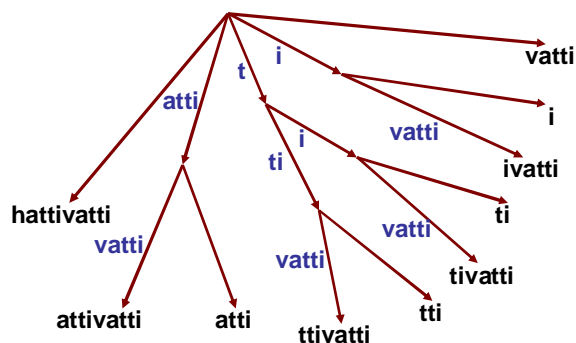
cgccgagtgacagagacgctaatacag  
gcgacccttgcagacatcaagctctt  
tgggaacaagtggagcaccgatgatg  
tacagccgatcaatgacatttccta  
atgcaggattacattgcagtgcccaa  
ggagaagtatgccaagtaatacctcc  
ctcacagtg...

## Sequence motifs

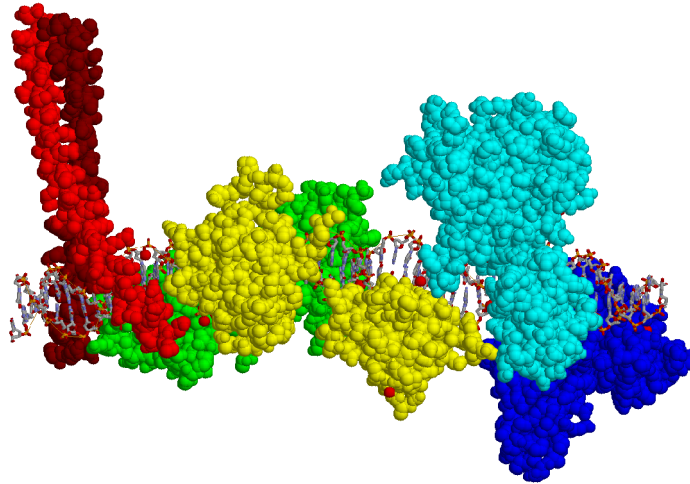
c**gcc**gagtg**a**cagagacgctaatacag  
gcgacccttgcagacatcaagctctt  
tgggaacaagtggagcaccgatgatg  
tacag**gcc**gatca**a**tgacatttccta  
atgcaggattacattgcagtg**ccca**a  
gg**a**agtat**cca**agta**a**tacctcc  
ctcacagtg...

## Suffix-trees

hattivatti  
attivatti  
ttivatti  
tivatti  
ivatti  
vatti  
atti  
tti  
ti  
i

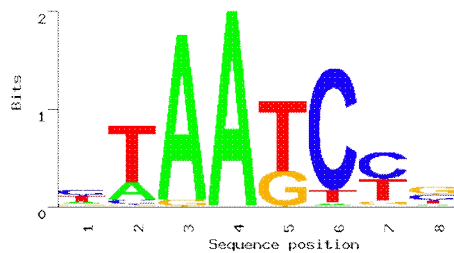


## Transcription factor binding sites

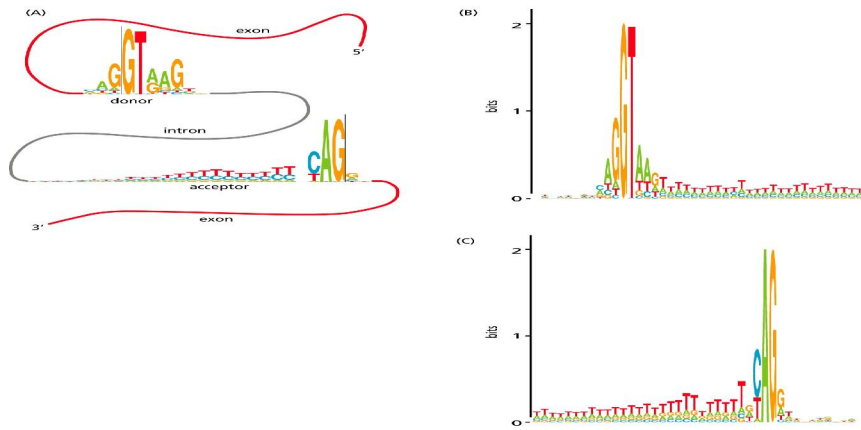


## Binding affinity matrices (PWMs)

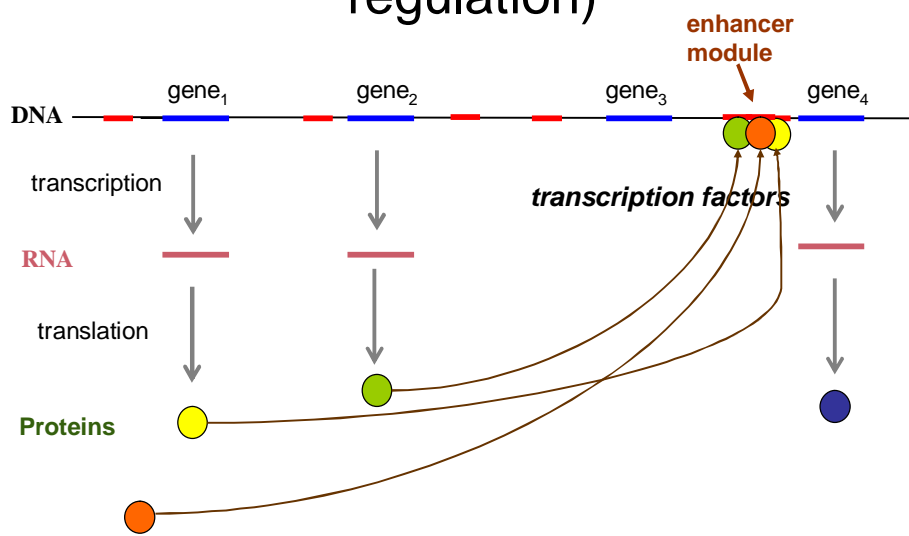
<b>A</b>	9	11	49	51	0	1	1	4
<b>C</b>	19	3	0	0	0	45	25	16
<b>G</b>	5	1	2	0	17	0	4	21
<b>T</b>	18	36	0	0	34	5	21	10



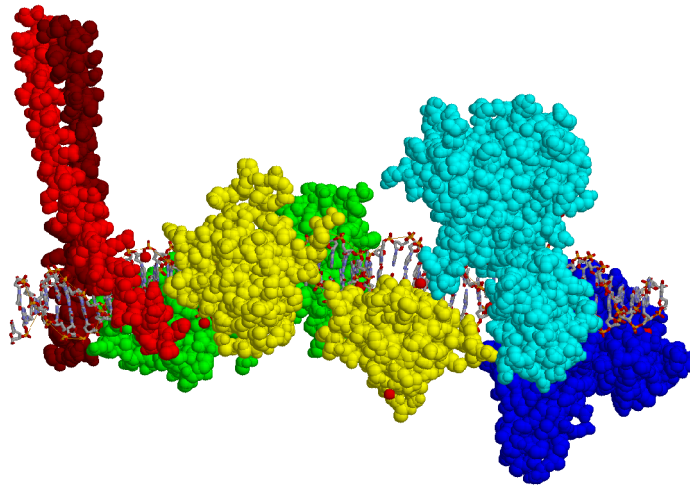
# Search for PWM occurrences



# Gene enhancer modules (cis-regulation)



# Enhancer module



## Drosophila enhancer

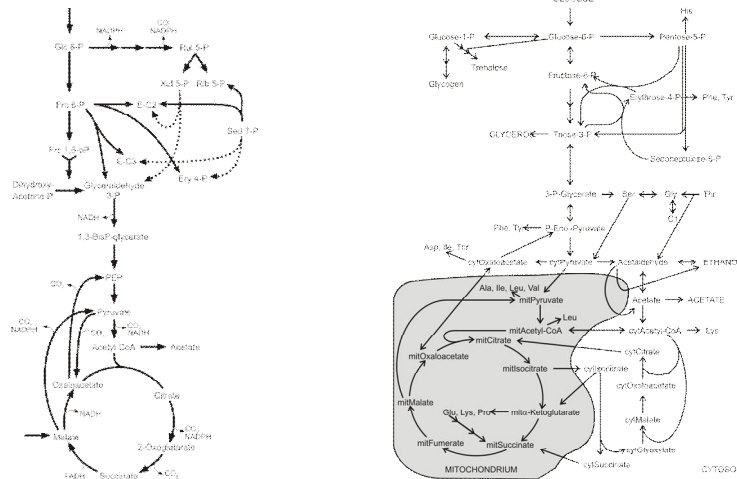
- Drosophila even-skipped gene stripe 2 enhancer
- Score = 487.05

```

10460 : cccgaggatgcatcctggcccggcag-gacgacctcgctgcattagaaAACTAGATCAg Kni1
127636 : ---aggatcc-tc--gaaa-cg-agag-cgacctcgctgcattagaaAACTAGATCAg Kni1
10519 : TTTT-TGTTCI-----a---ATT-TTGTGCCc-gcccTGCCTCCTTatgcattattgtt Hb14 Hb14 Kni2
127686 : TTTT-TGTTCIggccgaccgATT-TTGTGCCcgg---TGCCTCCTTtagc-----gtt Hb14 Kni2
10571 : taaggtc-cattccatttcc-sttttcatttccacctccattgtttggccgcaaaaCAA
127736 : taaggcccggtcccatctccagctc-tttgt--tcc---g---ggct-cagaa-at
10629 : accccgacgggaattatggatggtatatgcagATTTTATGGG-cactcgggtgatct Hb12
127784 : -c---gtatgg-aattatggtat---at--gcagATTTTATGGGtc-c-cggc-gatcc Hb12
10688 : agctccgggaatggccgcta-cctqtagccc-gggacctcgaaccggccctcggaggat
127832 : gg_lccgggaacgggagLl--cclgcccagagcgl-cclcg-ccggcga-cc-----l Kni4 Hb11
10747 : atctgtatgtctatattaggaAACTAGATCAgTTTTCCTCCcatttgcccttttt Kni4 Hb11
127983 : -t--gtc-gcccgattaggaAACTAGATCAgTTTTCCTCCcatttgcccttttt Kni4 Hb10
10807 : cgdTGGCTAGT-ttttcccgaaacgcagcaaacctgctctaaTTTTTAATTctcaag Kni5 Hb8
127939 : cgdTGGCTAGT-ttttcccccgaacccagcaaacctgctctaaTTTTTAATTctcaag Kni5 Hb8
10867 : gg-ttttattgtgctcctggaaaaactacggtctccacaaggttagagcgtcttagtta Hb8
127999 : gctttcattgggctcctggaaacaacg-cgg-----acaaggtataacgctctactta Hb8
10927 : ccgtaattgtggccataaacacacattcaag-cgcattcagtgctctcATT-TTAAGA Hb8
128052 : cccgc-aattgtggccataaac-----g-c---a--c--tgctctcATT-TTAAGA Hb8
10987 : Taagttctttctctgtgt-ttctgttctgtctgttcaatcaatTTTTATGACgct- Hb8
128095 : Tcegtt-tgt--tgtgttgtt--gtcc-gcgatgg-cattcaagTTTTATGACg-ctc Hb8

```

# Regulatory networks



## Computational techniques – Do you already know all these?

- Dynamic programming
- Needleman-Wunsch algorithm, Smith-Waterman algorithm, Viterbi algorithm
- Probabilistic modeling
- Hidden Markov Models
- Maximum likelihood estimation
- Expectation Maximization (EM) algorithm
- Combinatorial pattern matching
- Exact and approximate string matching
- Aho-Corasick & Boyer-Moore algorithms
- Index structures for sequential data
- Suffix-trees, suffix-arrays, BW transform