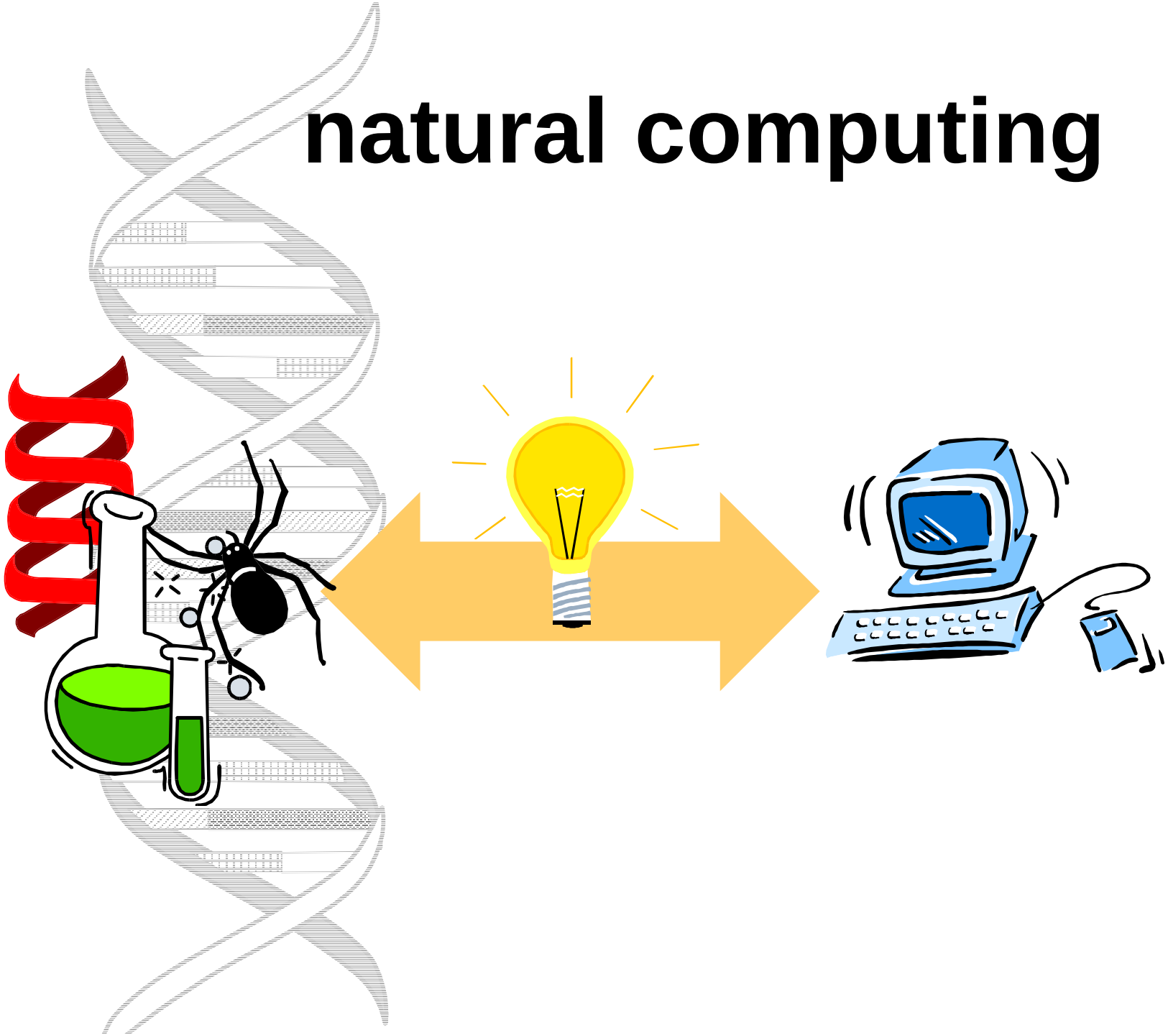
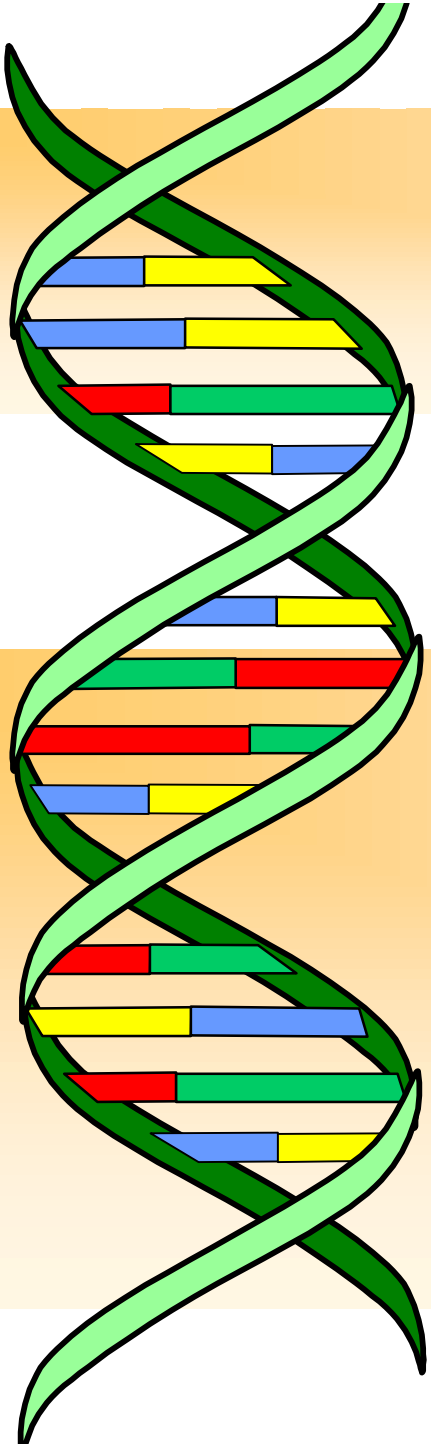
natural computing





Computational Molecular Biology

Hendrik Jan Hoogeboom
Fundamentele Informatica

FI1 : wiskundige begrippen
Datastructuren

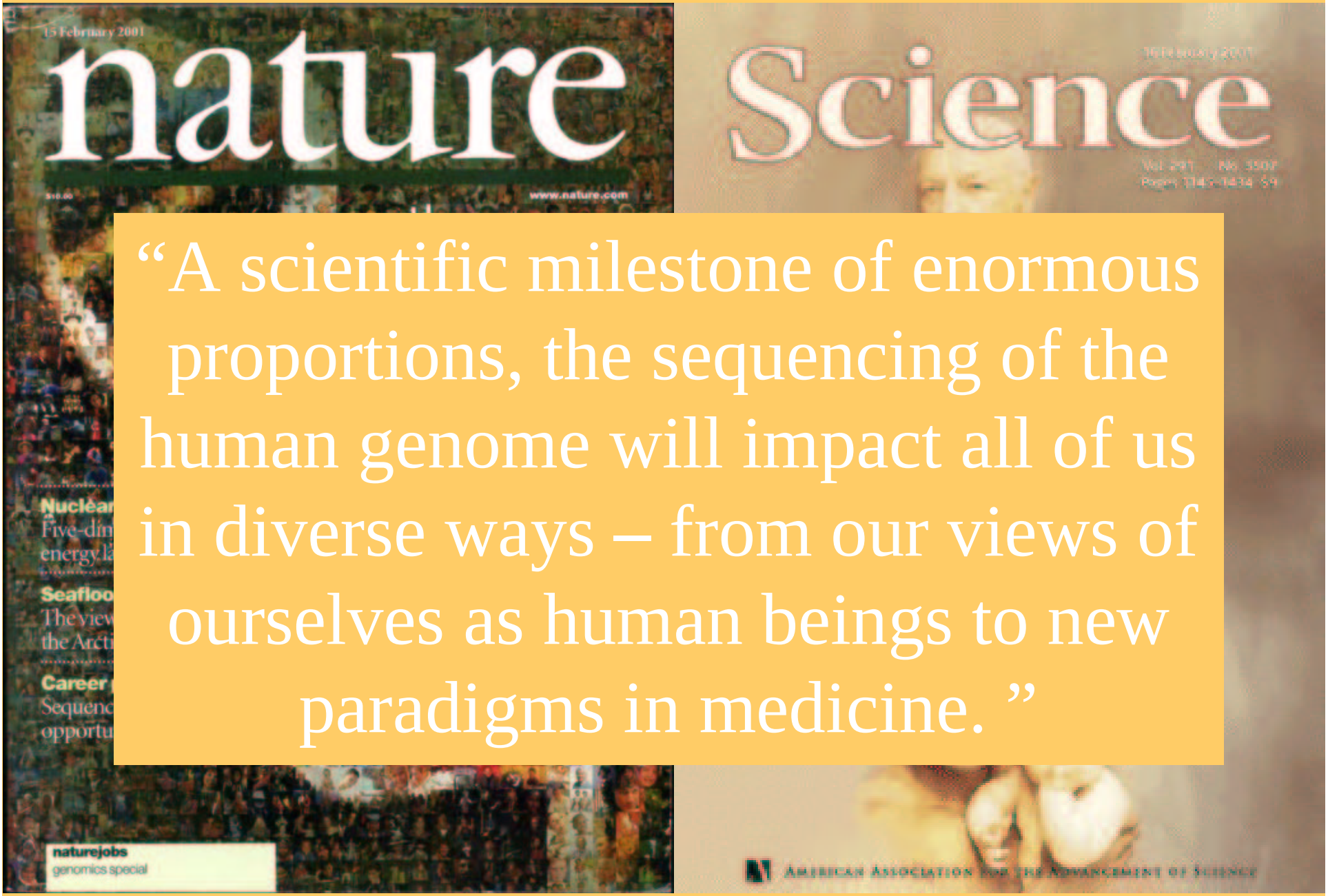
seminarium CMB

programmeerlijn

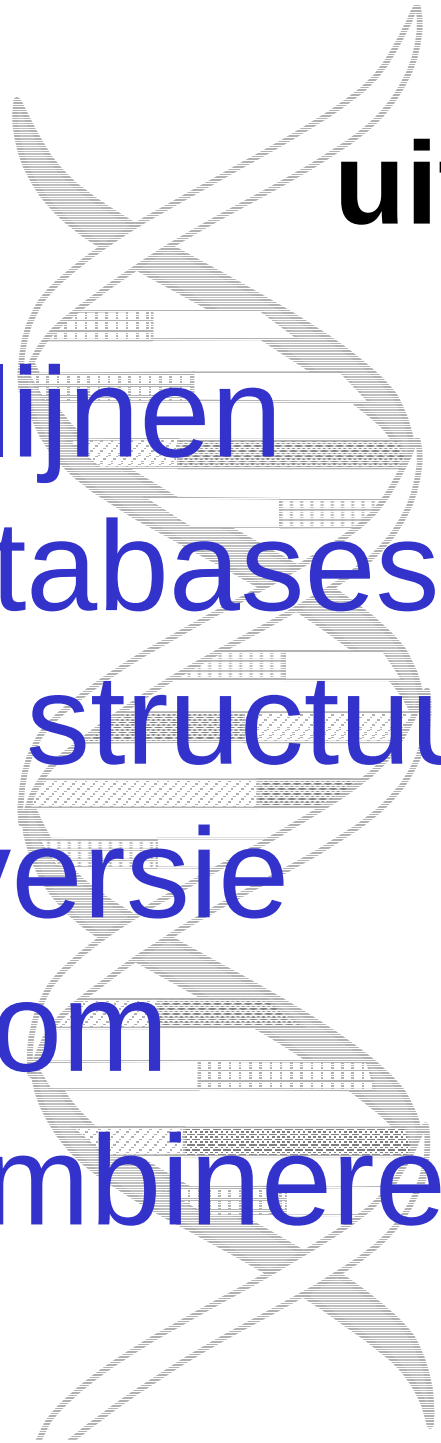
- programmeermethoden
- algoritmiek
- datastucturen

- concepten programmeertalen
- databases
- software engineering

- seminarium algoritmen:
  computational molecular biology



"A scientific milestone of enormous proportions, the sequencing of the human genome will impact all of us in diverse ways – from our views of ourselves as human beings to new paradigms in medicine. "



uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

uitdagingen

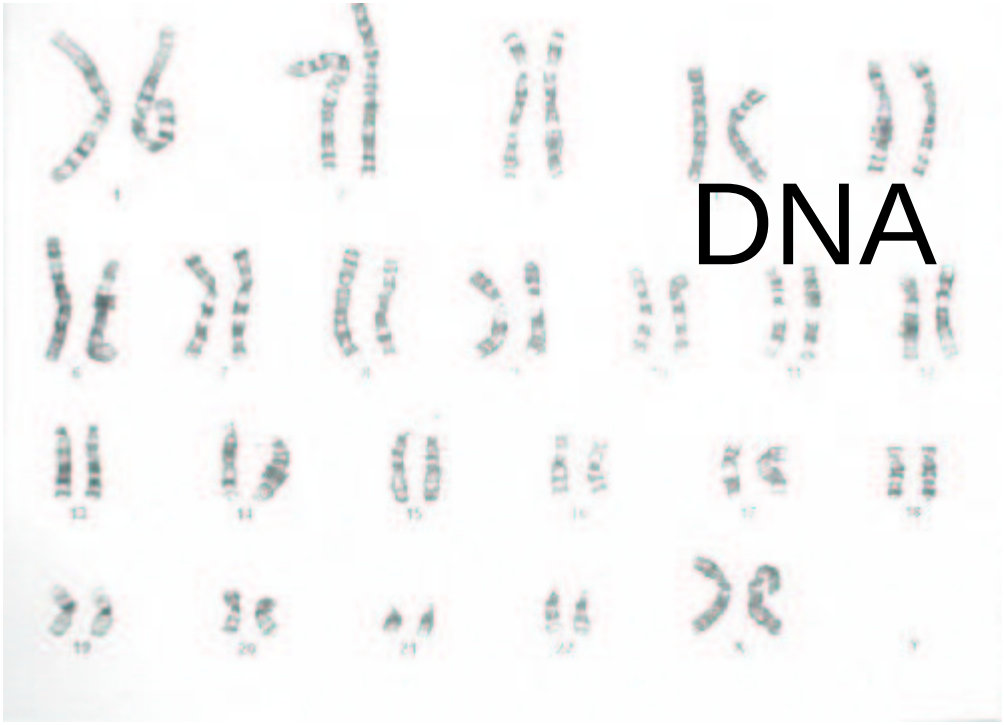
probleem $\Rightarrow$ model (bv. graaf)

- bekende algoritmen
- karakterisatie

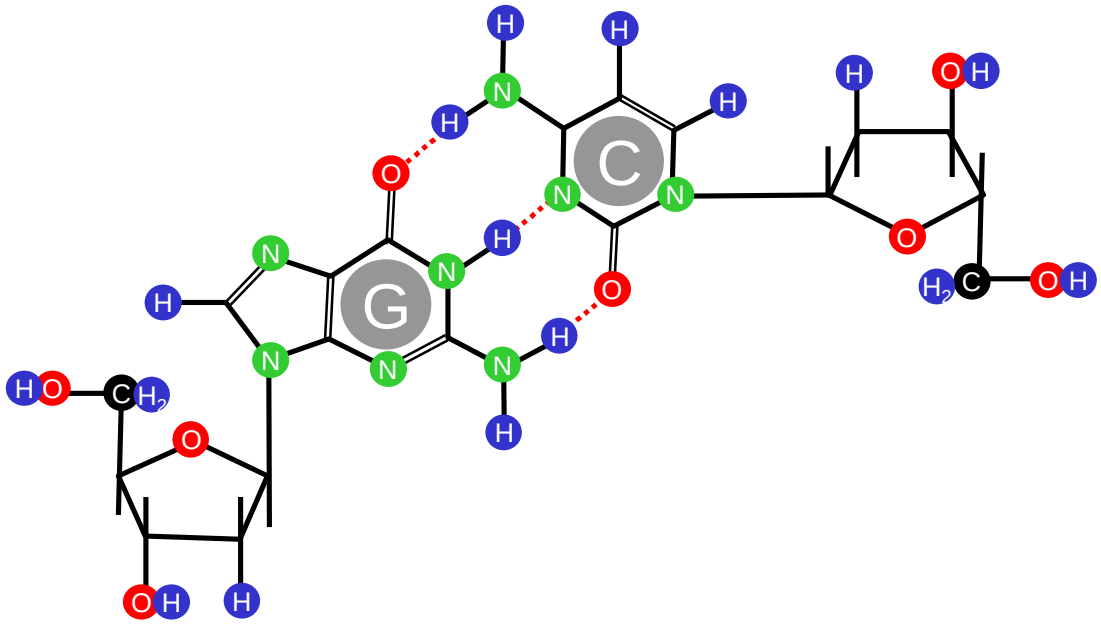
onnauwkeurigheid gegevens

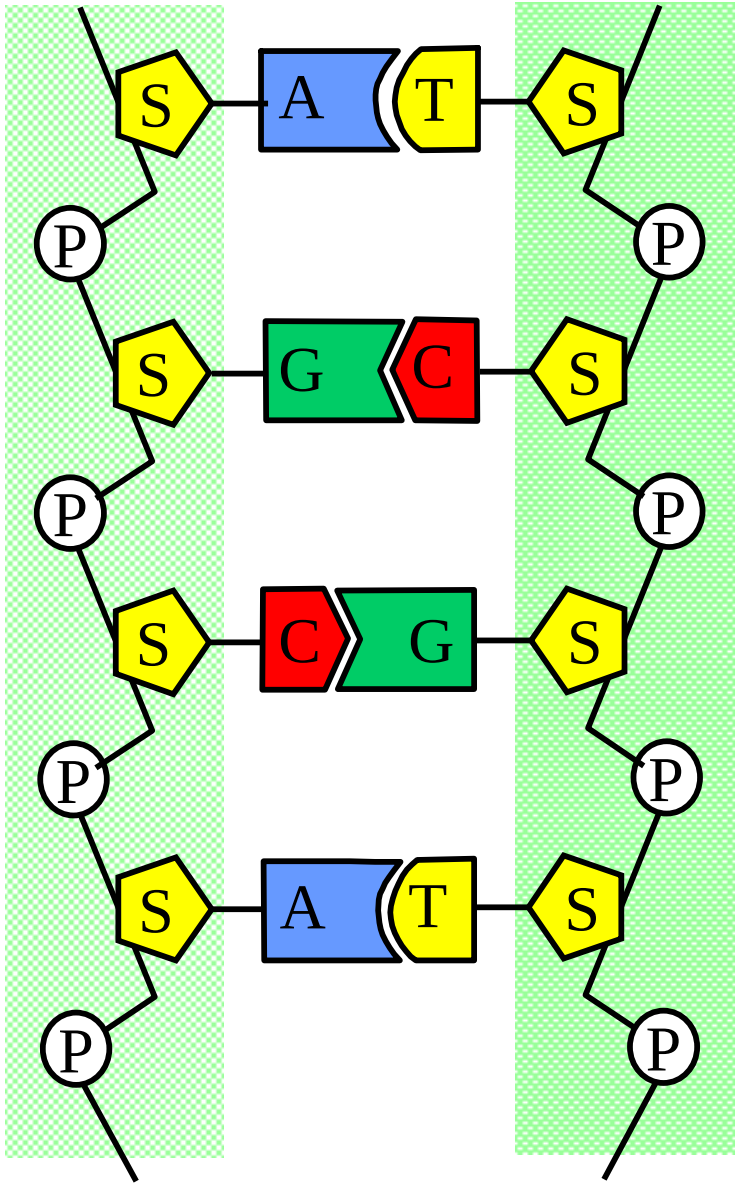
complexiteit

$\Rightarrow$ heuristieken



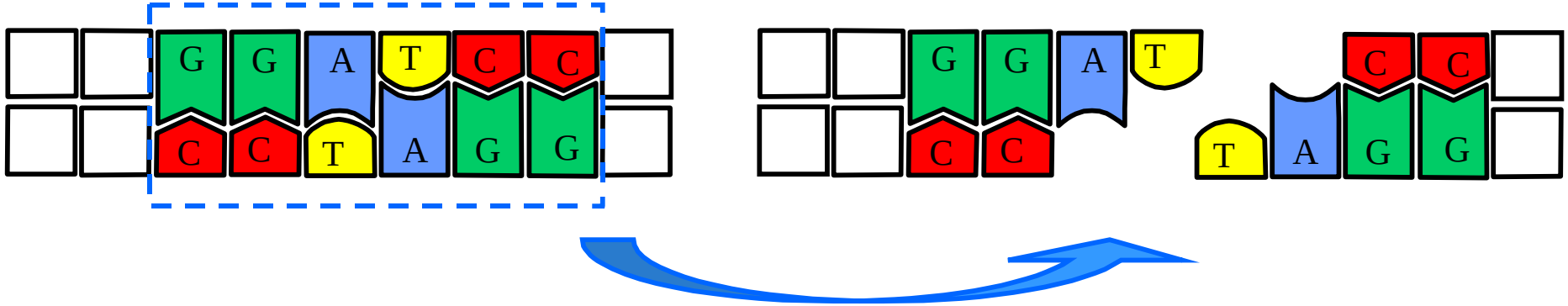
DNA





desoxyribonucleïnezuur

restrictie-enzymen



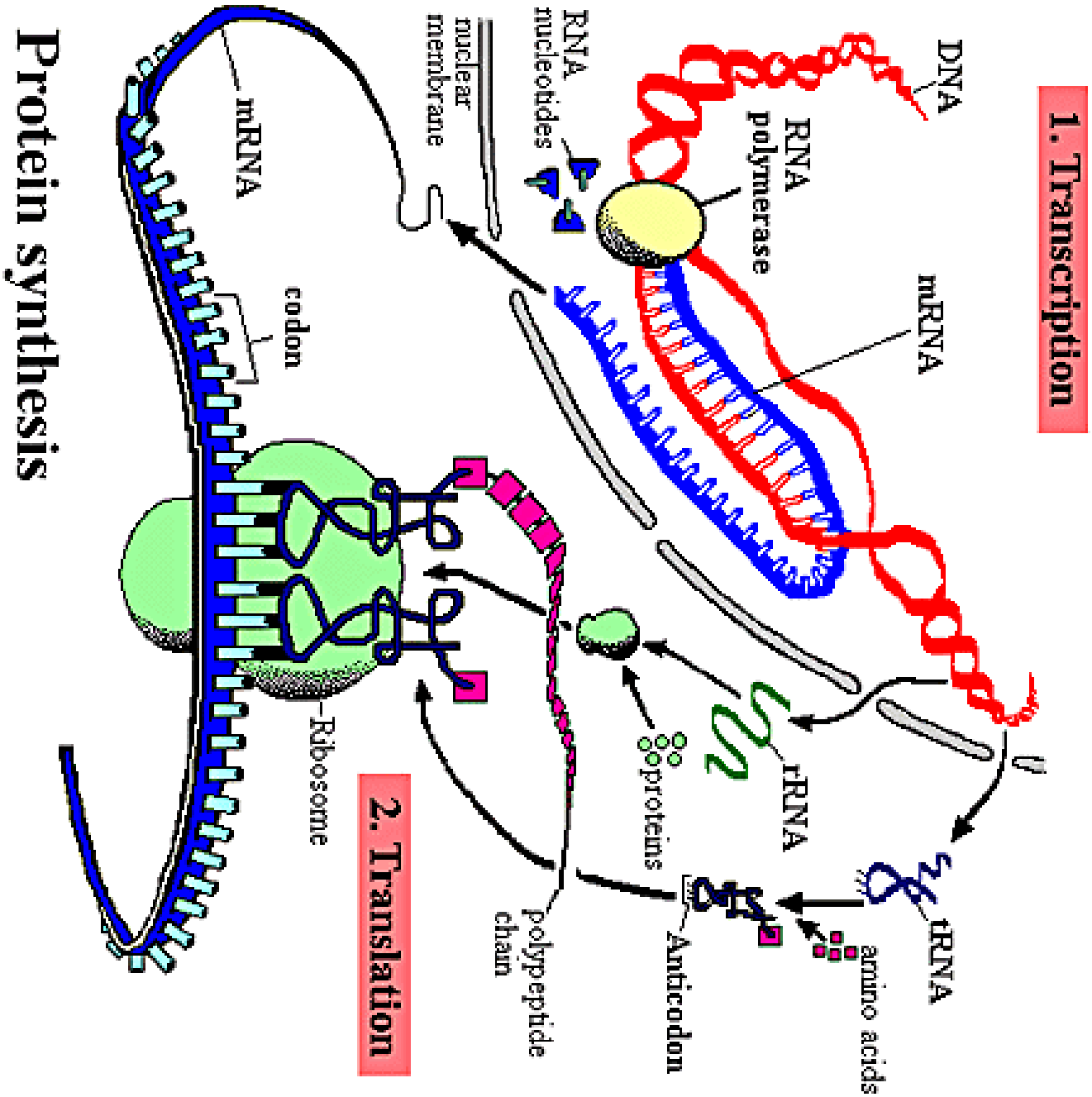
AGAGGATCCTTGCTGGATCCTGA
TCTCCTAGGAACGACCTAGGACT

AGAGGATCCTGA
TCTCCTAGGACT

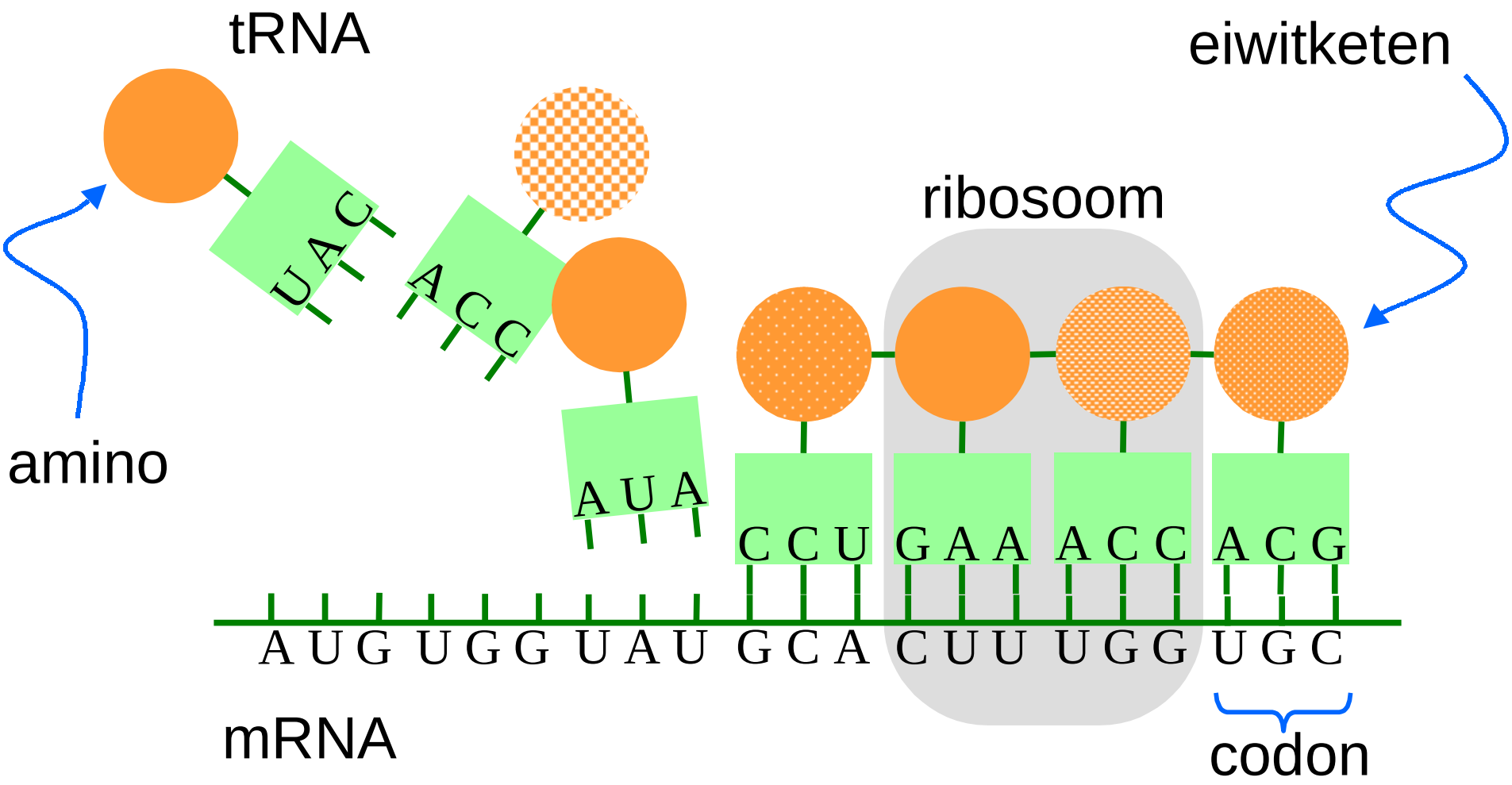
splicing

AGAGGATCCAGCAAGGATCCTGA
TCTCCTAGGTCGTTCCTAGGACT

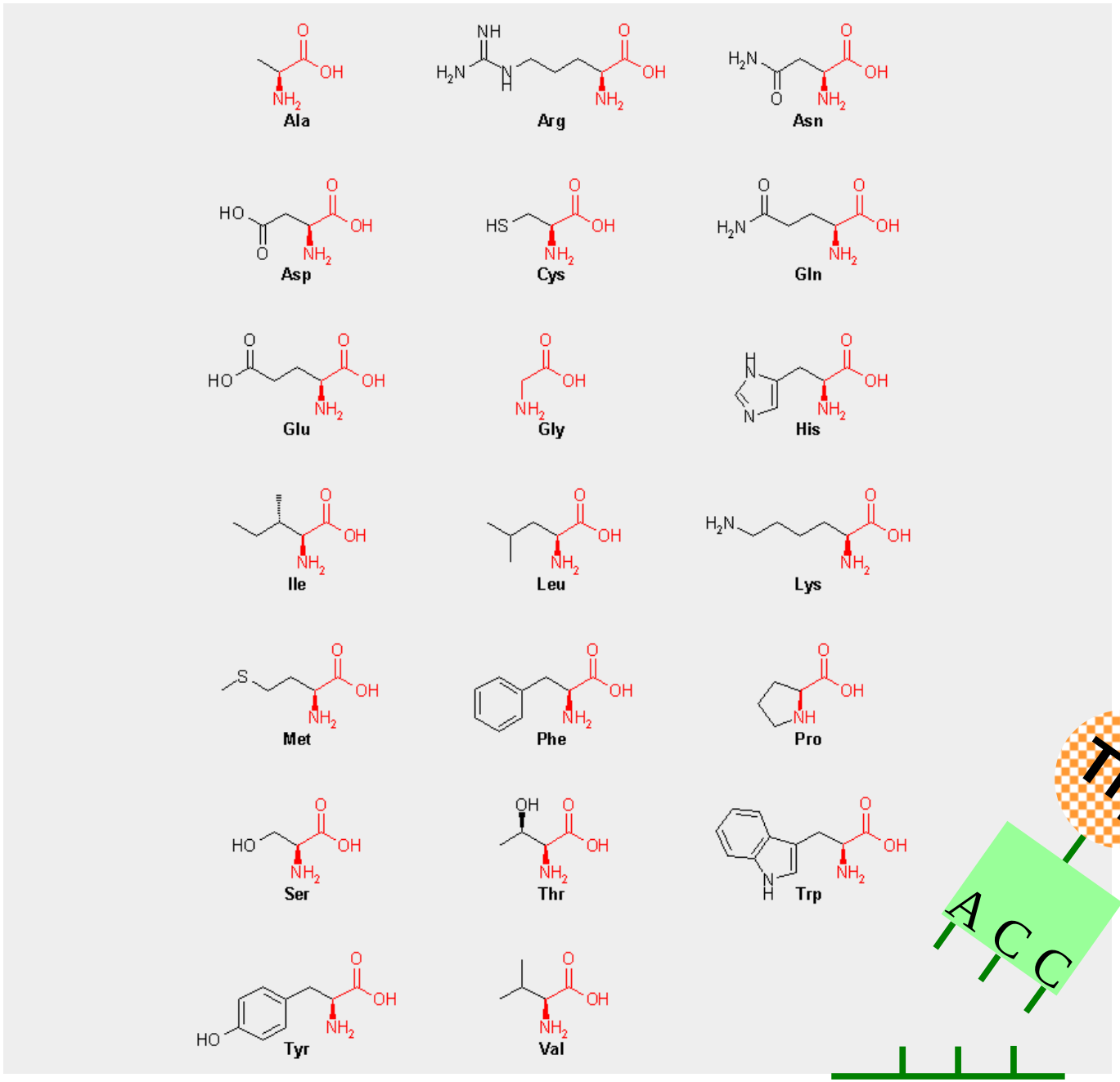
centraal dogma



translatie


tRNA
eiwitketen
ribosoom
amino
U A C
A C C
A U A
C C U
G A A
A C C
A C G
A U G U G G U A U G C A C U U U G G U G C
mRNA
codon

20 aminozuren



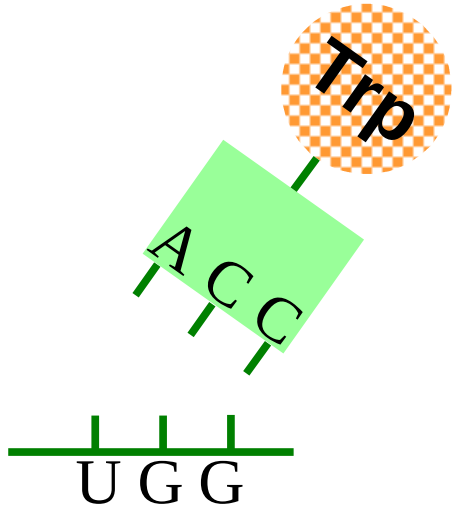
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

code

UGG

⇓

Trp



twee alfabetten

DNA

basen

4 symbolen

a c t g

eiwitten

aminozuren

20 symbolen

A R D N C
E Q G H I
L K M F P
S T W Y V

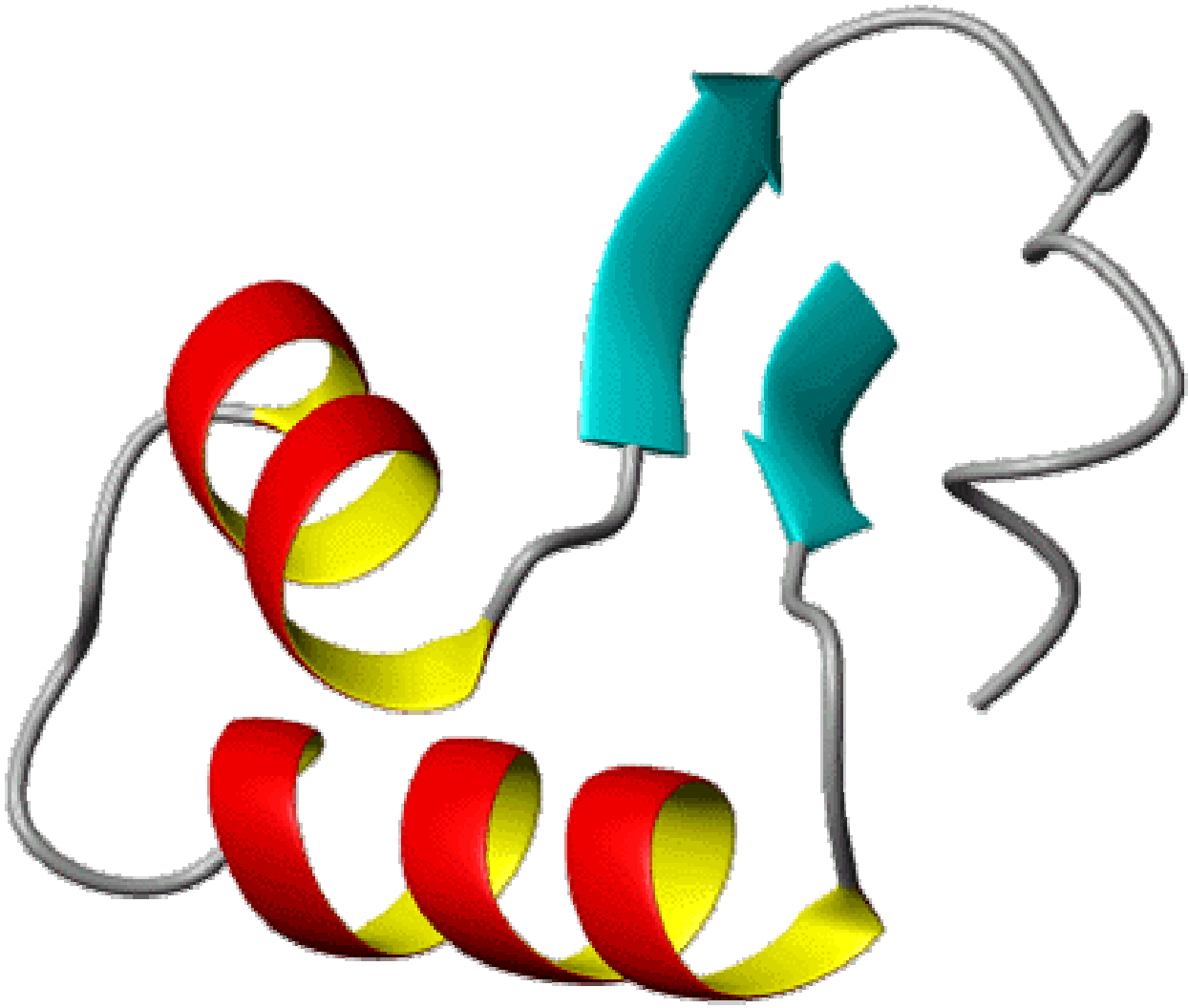
uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

alignment

```
MVVSGAPPAL GGGCLGTFTS LLLLASTAIL NAARIPVPPA CGKPQQLNRV VGGEDSTDSE
MMISRPPPAL GGDQFSILIL LVLLTSTAPI SAATIRVSPD CGKPQQLNRI VGGEDSMDAQ
*::* .**** **. :. :   *:**:*** : .** * *.*  *********: ****** *::

WPWIVSIQKN GTHHCAGSLL TSRWVITAAH CFKDNLNKPY LFSVLLGAWQ LGNPGSRSQK
WPWIVSILKN GSHHCAGSLL TNRWVVTAAH CFKSNMDKPS LFSVLLGAWK LGSPGPRSQK
******* ** *:******** *.***:**** ***.*::**  *********: **.**.****

VGVAWVEPHP VYSWKEGACA DIALVRLERS IQFSERVLPI CLPDASIHLP PNTHCWISGW
VGIAWVLPHP RYSWKEGTHA DIALVRLEHS IQFSERILPI CLPDSSVRLP PKTDCWIAGW
**:*** ***  ******: * ********:* ******:*** ****:*::** *:*.***:**

GSIQDGVPLP HPQTLQKLKV PIIDSEVCSH LYWRGAGQGP ITEDMLCAGY LEGERDACLG
GSIQDGVPLP HPQTLQKLKV PIIDSELCKS LYWRGAGQEA ITEGMLCAGY LEGERDACLG
********** ********** ******:*.  ******** . ***.****** **********

DSGGPLMCQV DGAWLLAGII SWGEGCAERN RPGVYISLSA HRSWVEKIVQ GVQLRGRAQG
DSGGPLMCQV DDHWLLTGII SWGEGCAD-D RPGVYTSLLA HRSWVQRIVQ GVQLRG----
********** *. ***:*** *******: : ***** ** * *****:*** ******
```

ruimtelijke structuur

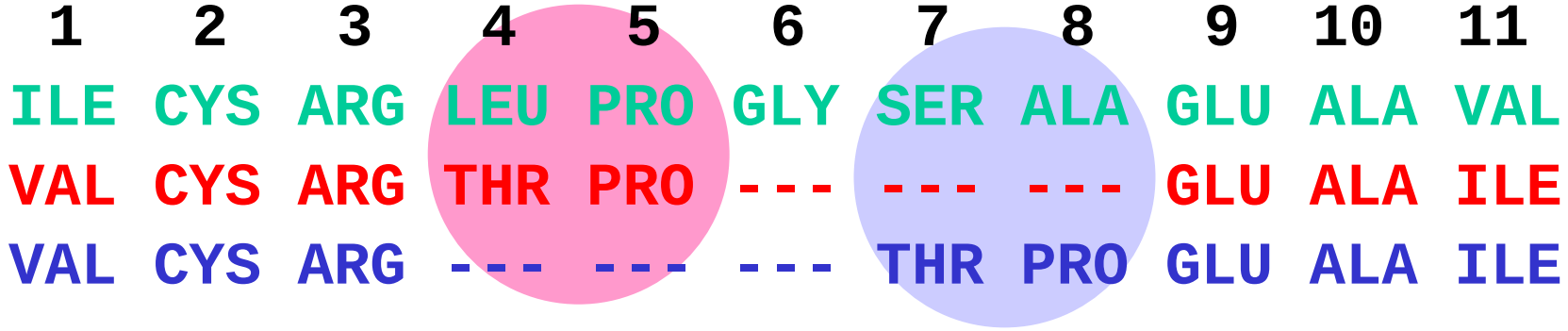


sequence alignment

Bekend vs. onbekend

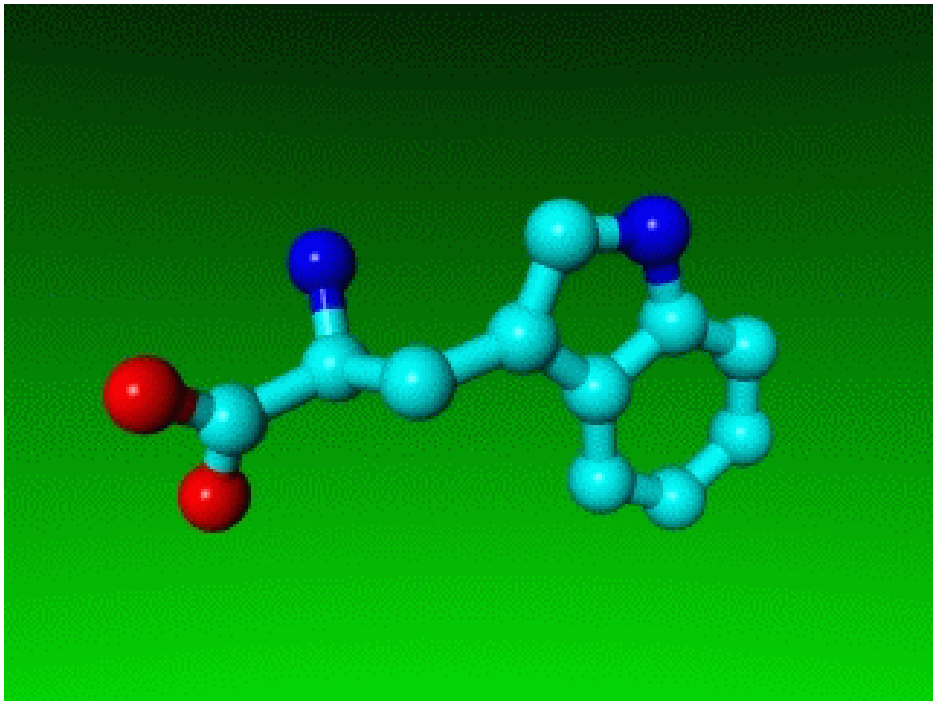
ILE CYS ARG LEU PRO GLY SER ALA GLU ALA VAL

VAL CYS ARG THR PRO GLU ALA ILE



W Trp Tryptophan

- Tryptophan is the biggest residue.
- It is aromatic.
- The nitrogen in the five-ring is donor for hydrogen bonds
- It is very hydrophobic.
- It doesn't care about helices or turns, but it loves strands.



PAM250 Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17
	C	**S**	**T**	**P**	**A**	**G**	**N**	**D**	**E**	**Q**	**H**	**R**	**K**	**M**	**I**	**L**	**V**	**F**	**Y**	**W**

- mutatiekans (evolutie)
- biochemische eigenschappen

questions

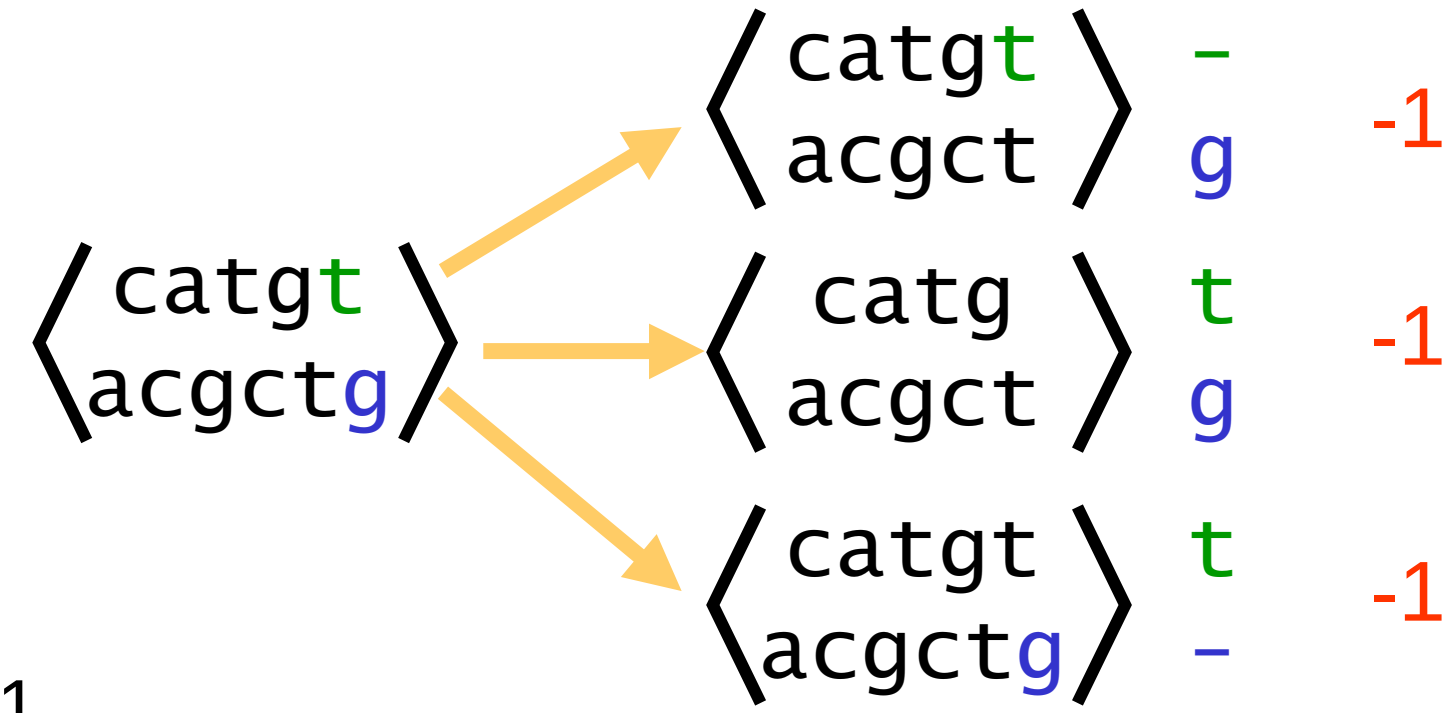
- Lookup
  - Is the gene known for my protein (or vice versa)?
  - **On which chromosome is the gene located?**
  - What sequence patterns are present in my protein?
  - Are the mutations known which cause this disease?
  - **To what class or family does my protein belong? What is known?**
- Compare
  - Are there sequences in the database resembling my protein?
  - How can I optimally align the members of this protein family?
  - Are these two sequences similar?
- Predict
  - Can I predict the active site residues of this enzyme?
  - Why are these patients ill?
  - **Can I make a 3D model for my protein?**
  - Can I predict a (better) drug for this target?
  - How can I improve the thermostability? (protein engineering)
  - How can I predict the genes located on this genome?

een algoritme

alignment

- recursief
- dynamisch programmeren

alignment: recursief



σ(-,x) = -1
σ(x,-) = -1
σ(x,y) = -1
σ(x,x) = 2

‘straf en beloning’

‘dynamisch programmeren’

⟨ cat
acgct ⟩

	-	c	a	t	g	t
-						
a						
c						
g						
c						
t				*		
g						

⟨ catgt
acgctg ⟩

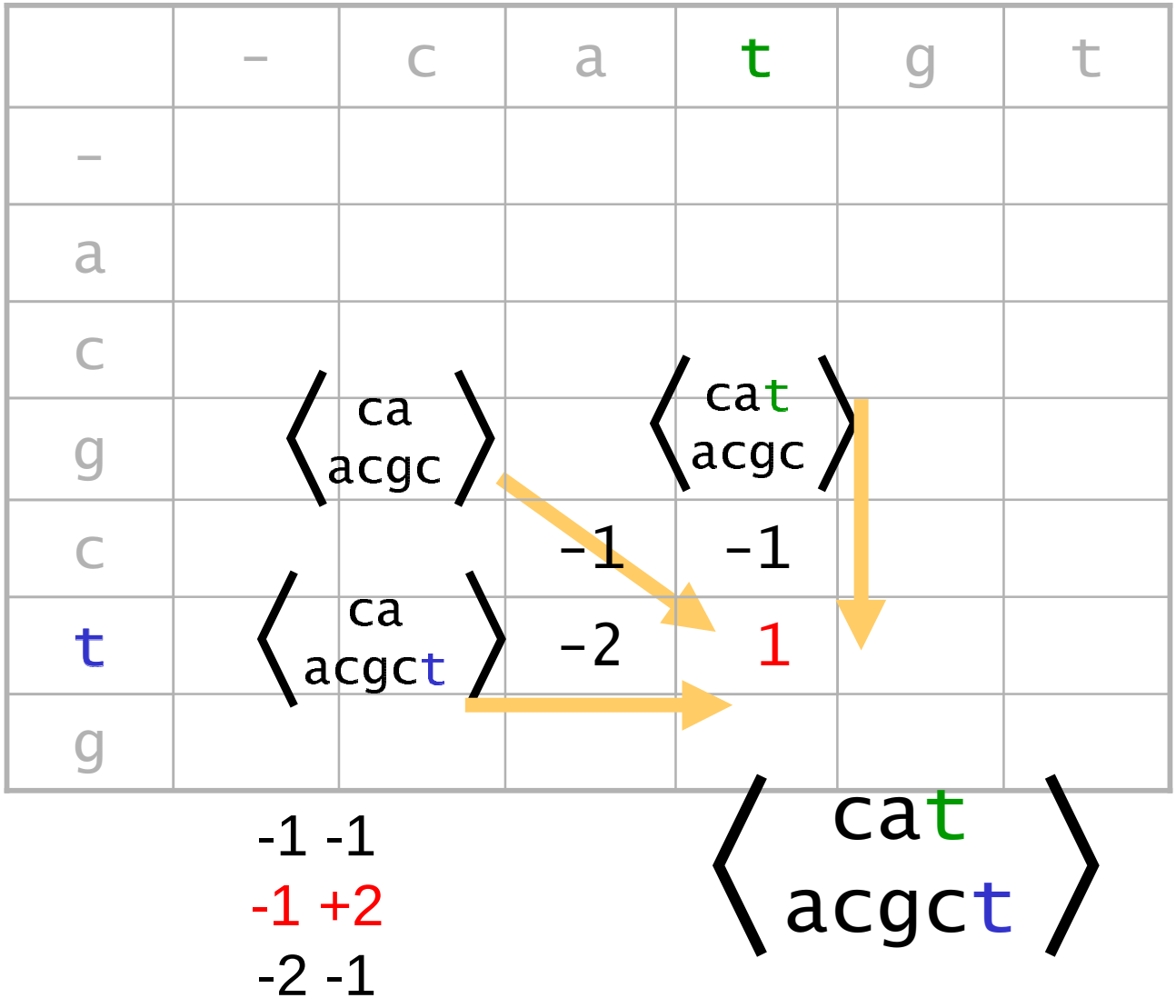
‘dynamisch programmeren’

⟨ cat
acgct ⟩

	-	c	a	t	g	t
-	0	-1	-2	-3	-4	-5
a	-1	-1	1	0	-1	-2
c	-2	1	0	0	-1	-2
g	-3	0	0	-1	2	1
c	-4	-1	-1	-1	1	1
t	-5	-2	-2	1	0	3
g	-6	-3	-3	0	3	2

⟨ catgt
acgctg ⟩

‘dynamisch programmeren’



alignment

	-	c	a	t	g	t
-	0	-1	-2	-3	-4	-5
a	-1	-1	1	0	-1	-2
c	-2	1	0	0	-1	-2
g	-3	0	0	-1	2	1
c	-4	-1	-1	-1	1	1
t	-5	-2	-2	1	0	3
g	-6	-3	-3	0	3	2

⟨ catgt
acgctg ⟩

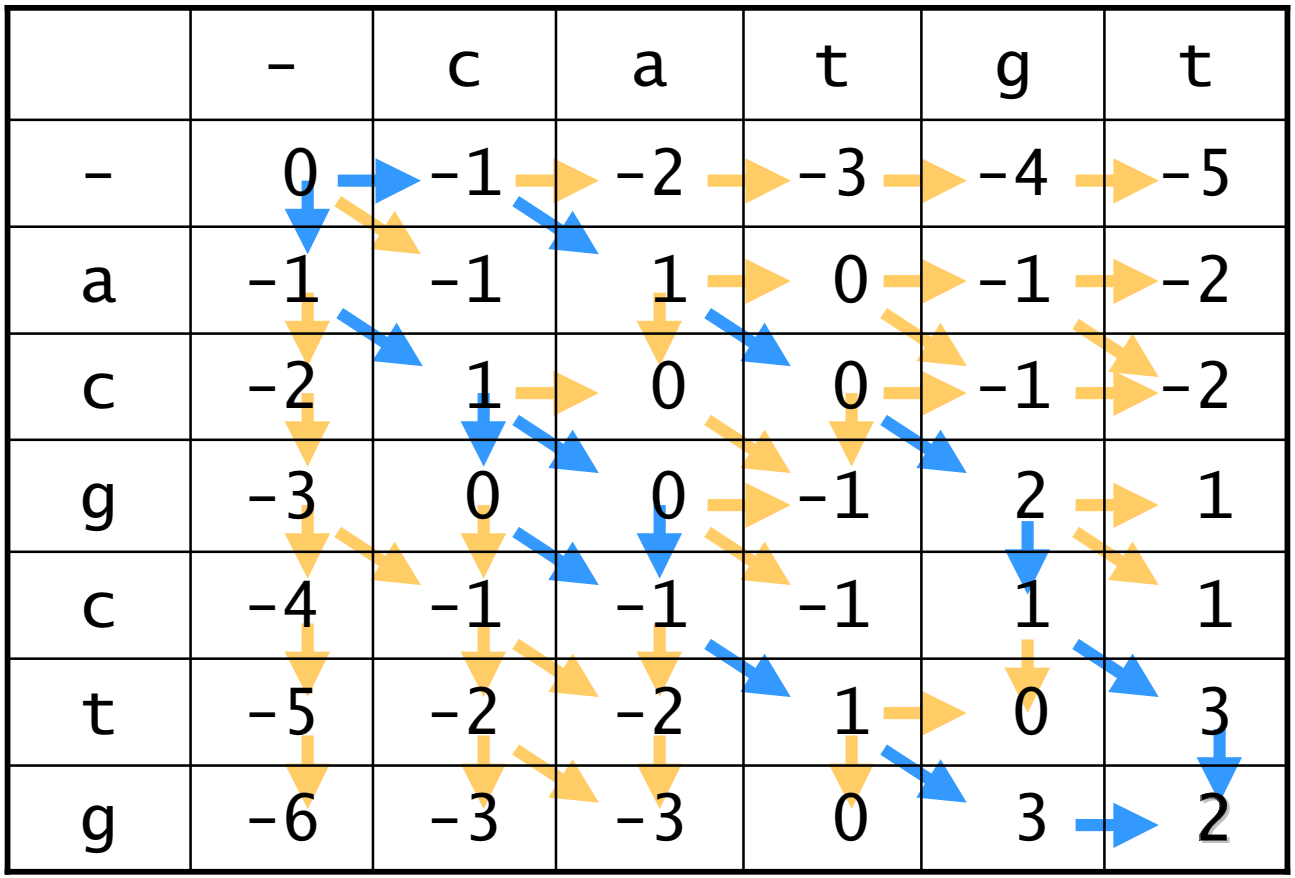
alignment

catg-t-
-acgctg

-ca-tgt
acgctg-

-c-atgt
acgctg-

⟨ catgt
⟨ acgctg ⟩



probleem opgelost !?

te langzaam

- lange strings
- grote databases

heuristieken

- langs diagonaal
- exacte overeenkomst

multiple alignment

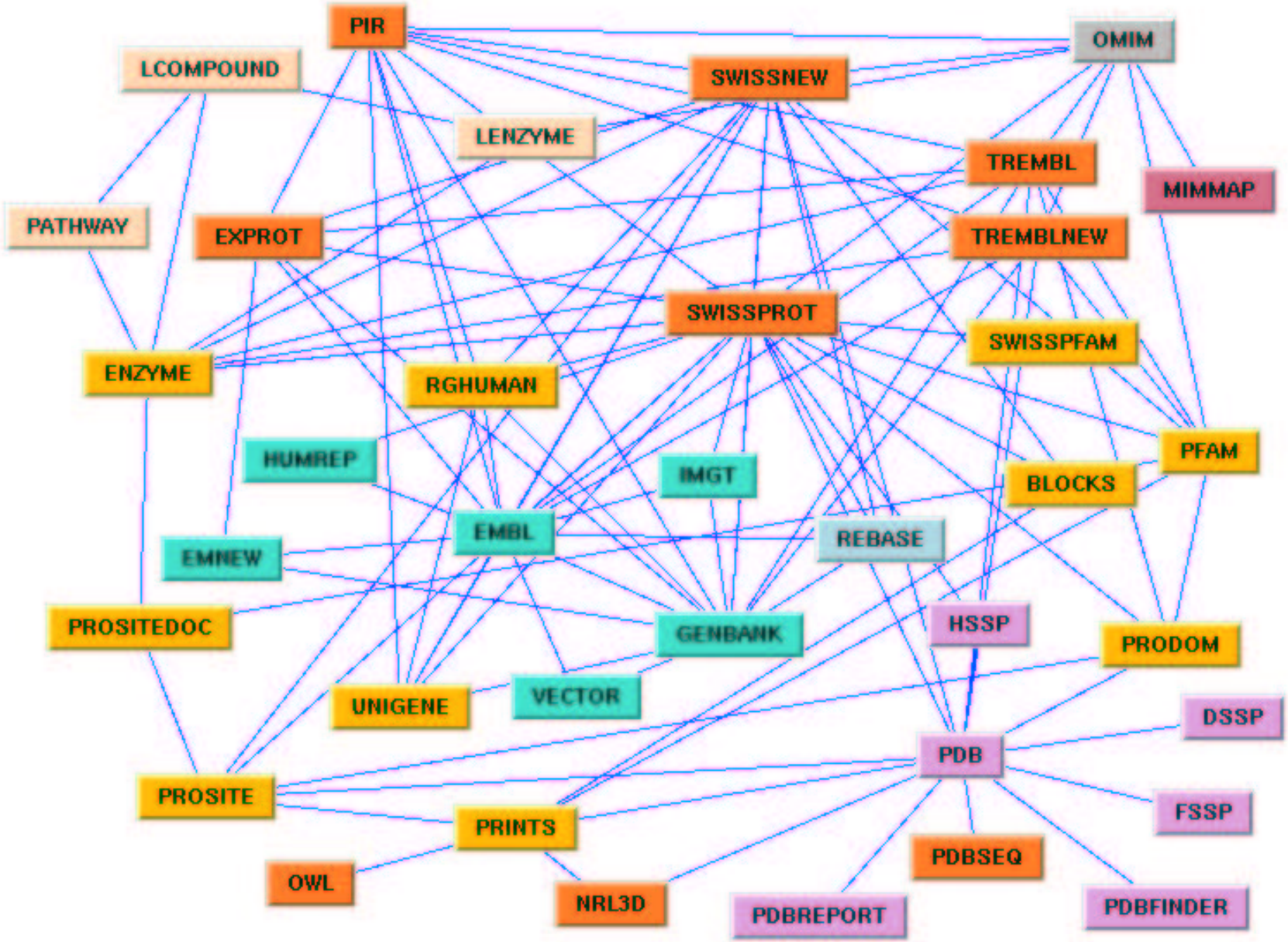
(meerdere strings)

NP compleet … exponentieel

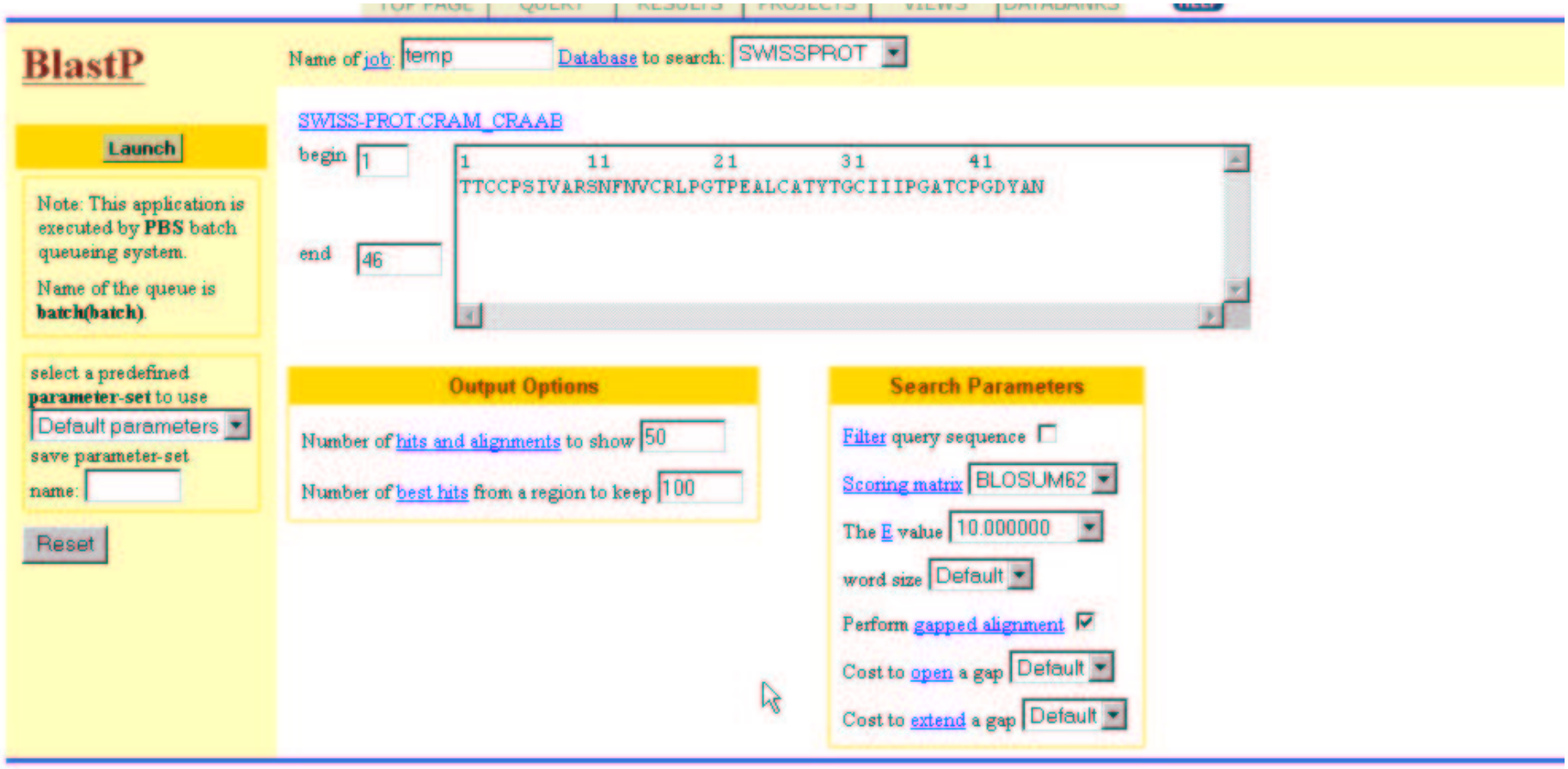
uitdagingen

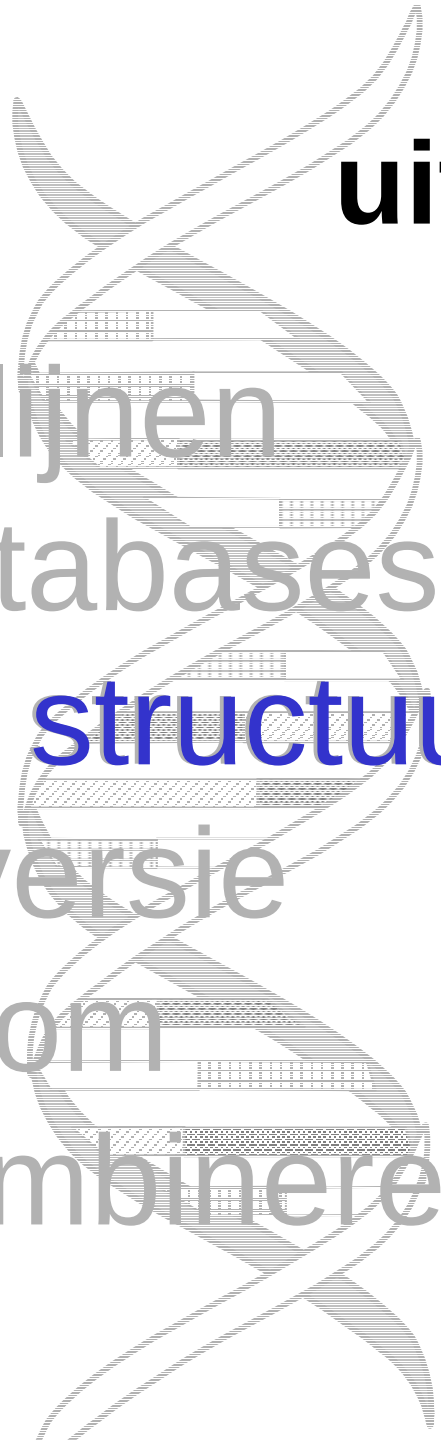
- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

databases


PIR
OMIM
LCOMPOUND
SWISSNEW
LENZYME
TREMBL
MIMMAP
PATHWAY
EXPROT
TREMBLNEW
SWISSPROT
SWISSPFAM
ENZYME
RGHUMAN
PFAM
HUMREP
IMGT
BLOCKS
EMBL
REBASE
EMNEW
PROSITEDOC
HSSP
GENBANK
PRODOM
UNIGENE
VECTOR
DSSP
PDB
PROSITE
FSSP
PRINTS
PDBSEQ
OWL
NRL3D
PDBREPORT
PDBFINDER

‘*launch*’

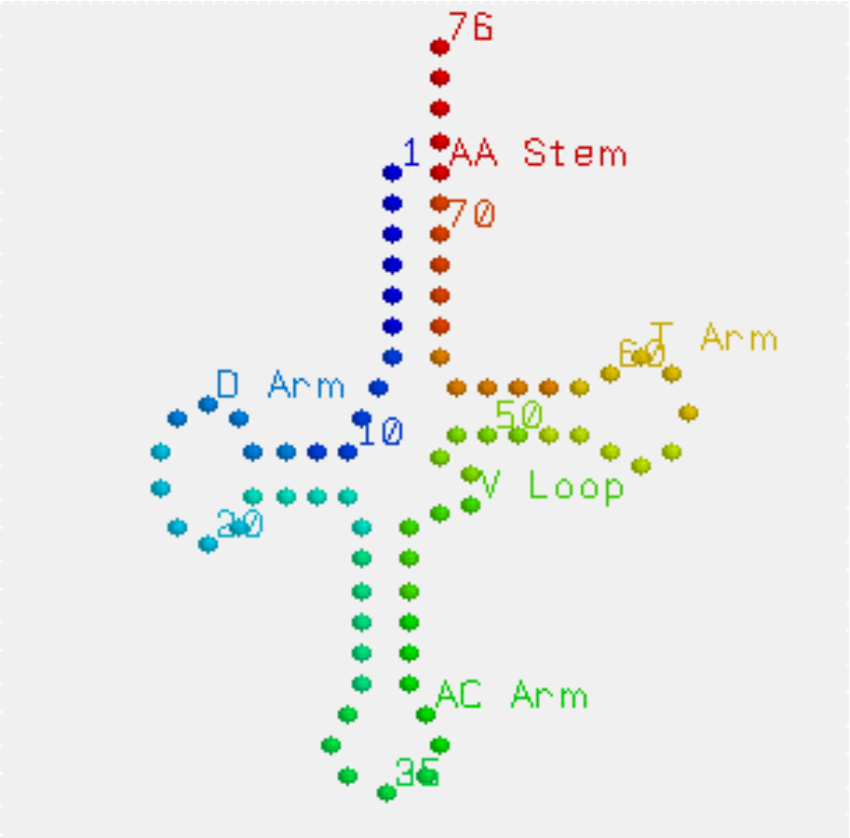

BlastP
Name of job: temp
Database to search: SWISSPROT
SWISS-PROT:CRAM_CRAAB
Launch
Note: This application is executed by PBS batch queueing system.
Name of the queue is batch(batch).
begin 1
end 46
1 11 21 31 41
TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
select a predefined parameter-set to use
Default parameters
save parameter-set name:
Reset
Output Options
Number of hits and alignments to show 50
Number of best hits from a region to keep 100
Search Parameters
Filter query sequence
Scoring matrix BLOSUM62
The E value 10.000000
word size Default
Perform gapped alignment
Cost to open a gap Default
Cost to extend a gap Default



uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

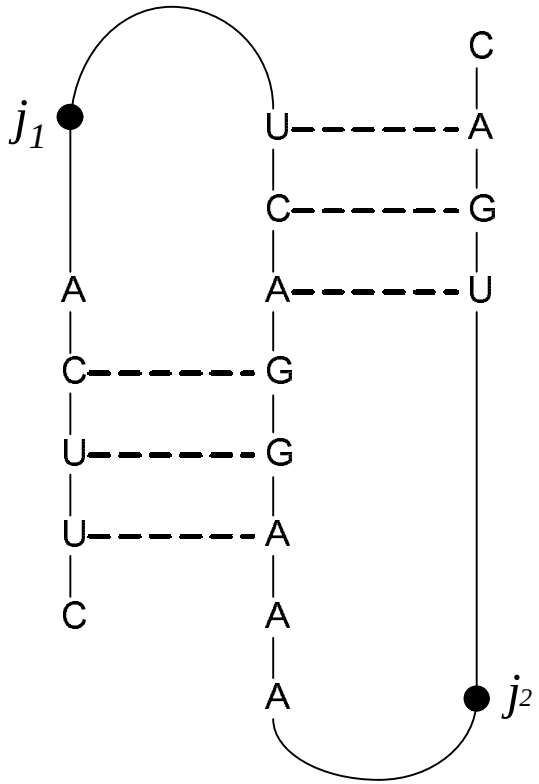
2D & 3D Structures of Yeast Phenylalanyl-Transfer RNA

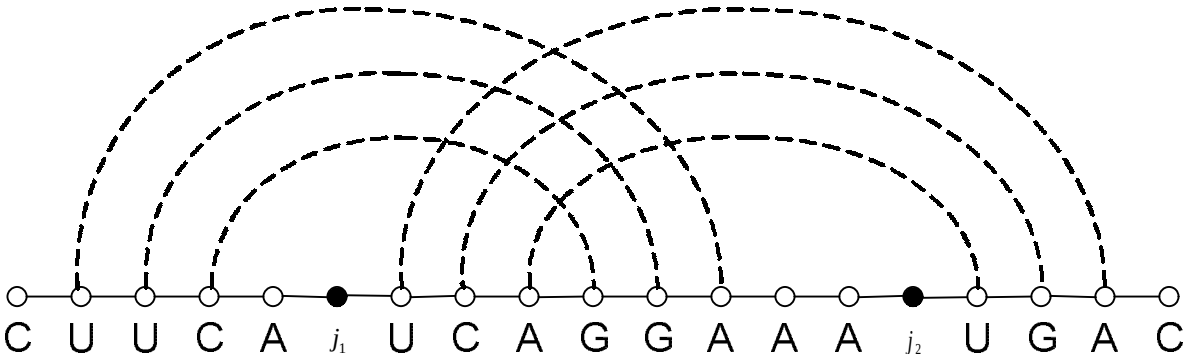


2D Structure

3D Structure

RNA Secondary Structure with Simple Pseudoknots



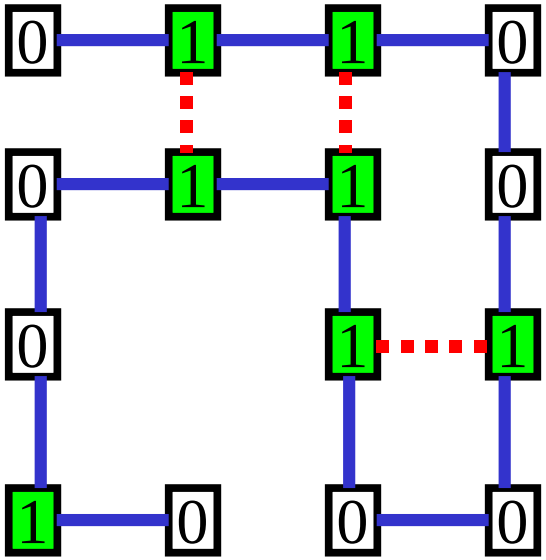


(ACTU: dit is RNA)

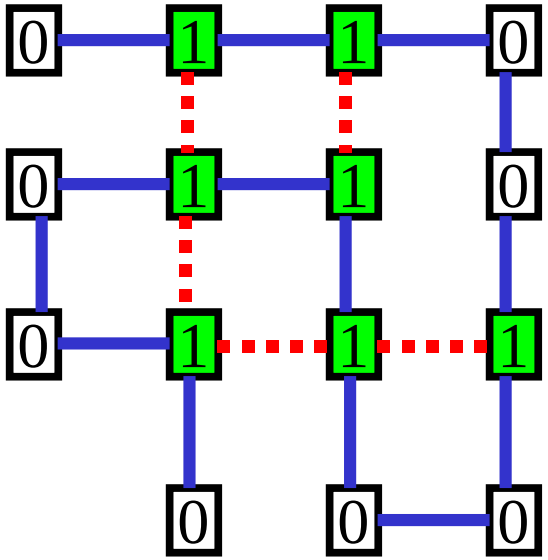
vereenvoudigd model

1 = *H* (***hydrophobic***, non-polar) (hating water)
0 = *P* (***hydrophilic***, polar) (loving water)

- Instance: 011001001110010

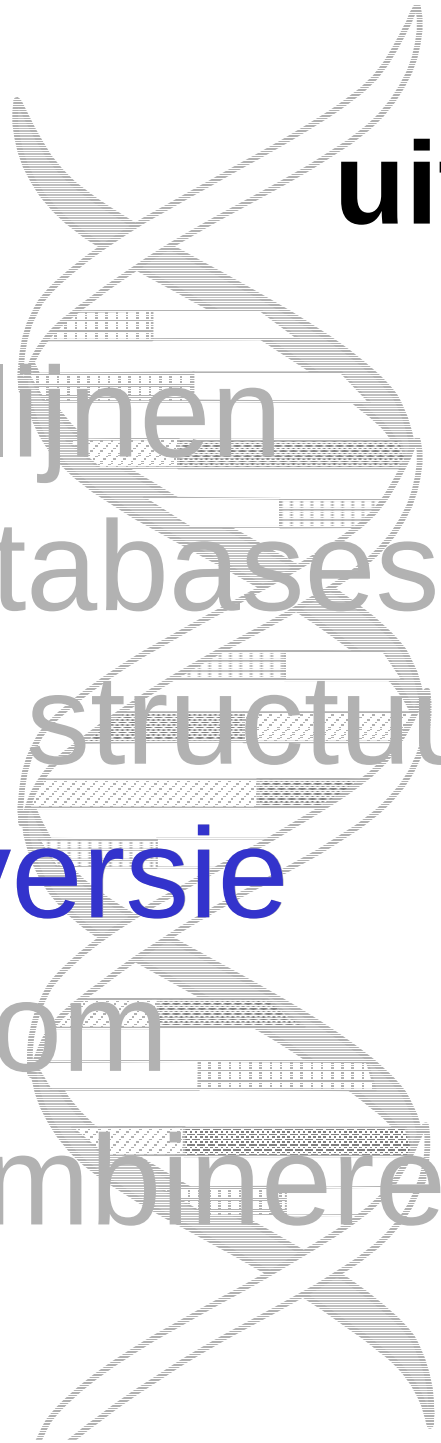


Score = 3



Score = 5

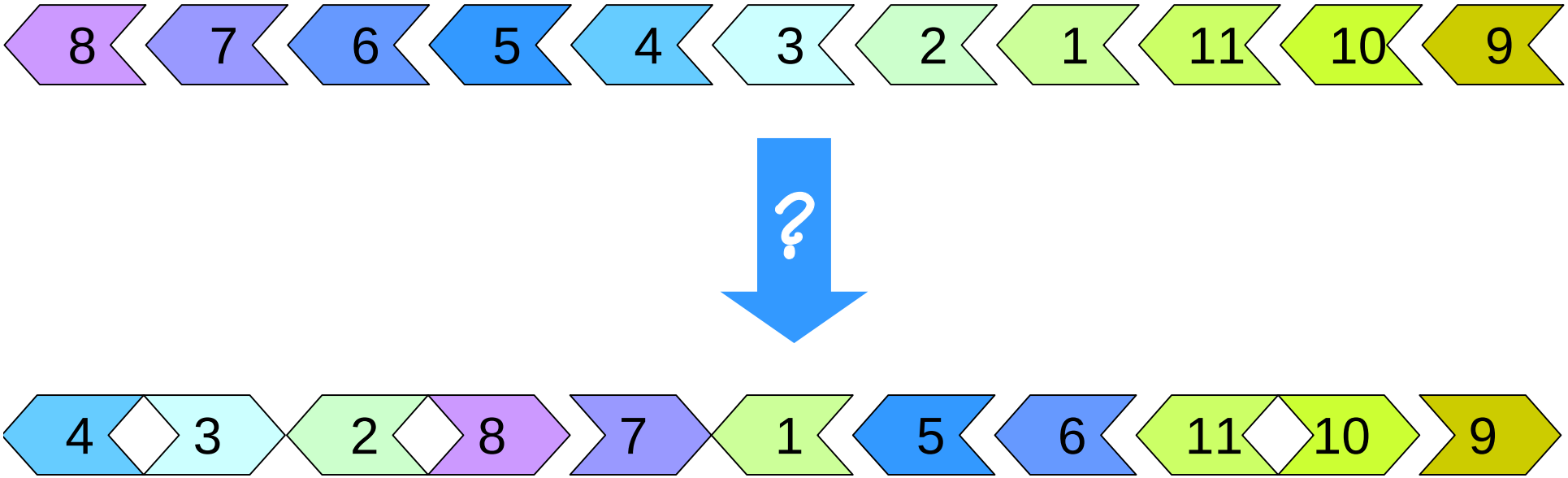
(dit is eiwit)



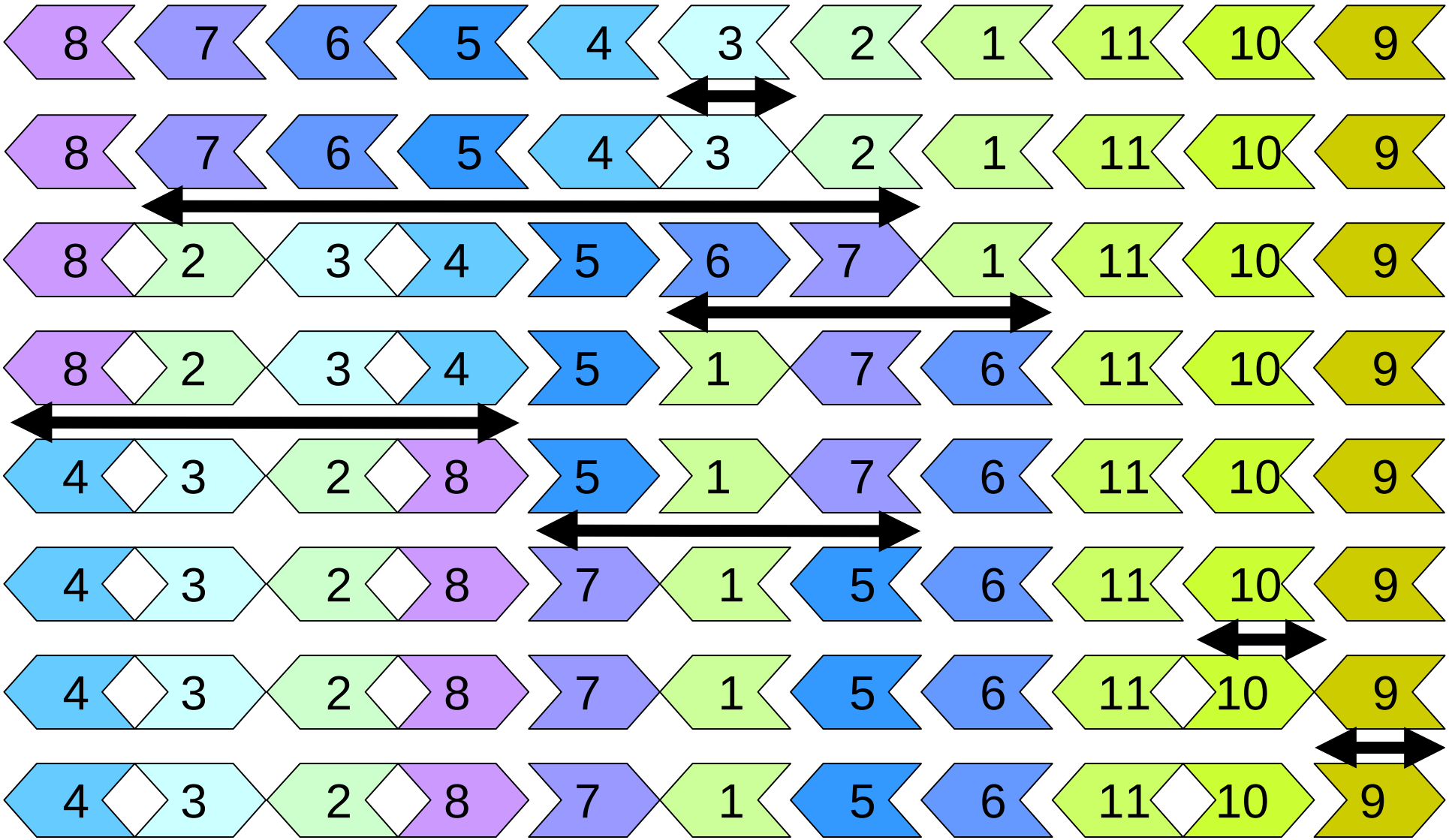
uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

genoom: van kool naar raap



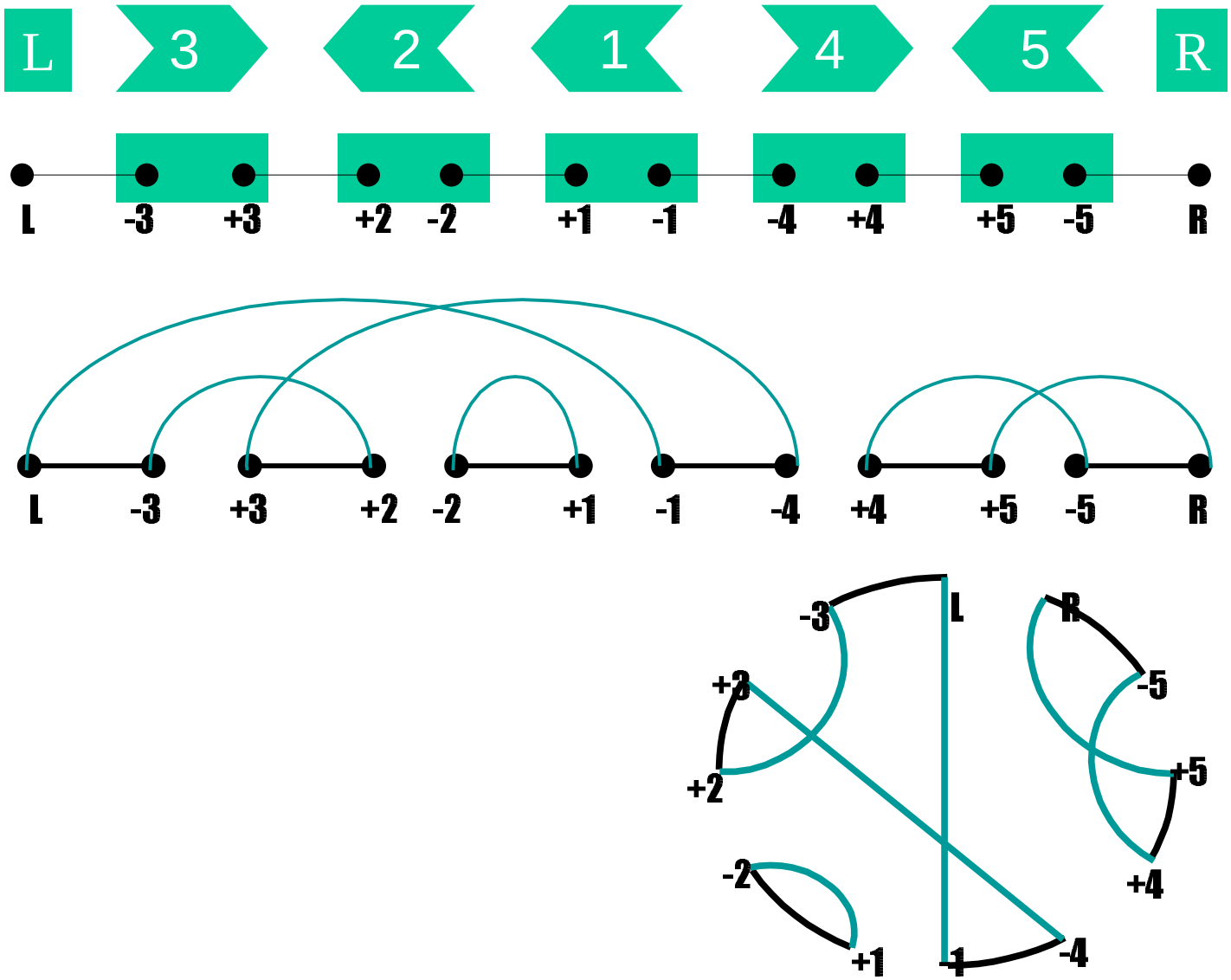
genoom: van kool naar raap



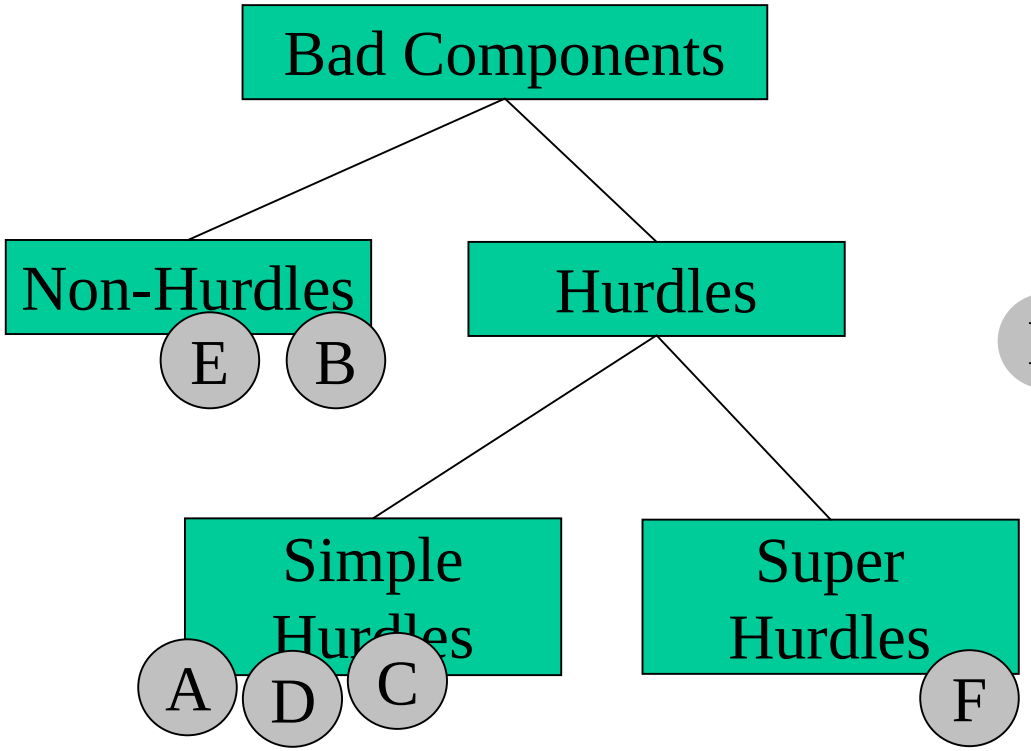
Transformation of mitochondrial DNA: worm Ascaris Suum into human

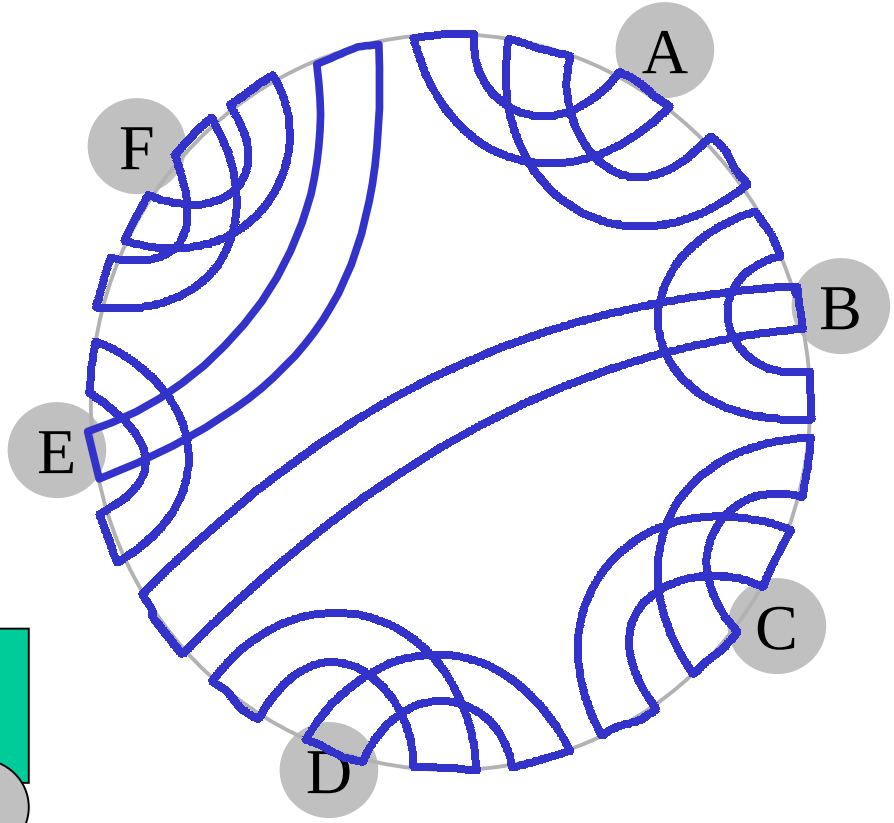
12 31 34 28 26 17 29 4 9 36 18 35 19 1 16 14 32 33 22 15 11 27 5 20 13 30 23 10 6 3 24 21 8 25 2 7
12 31 34 28 26 17 29 4 9 36 18 35 19 1 16 14 33 32 22 15 11 27 5 20 13 30 23 10 6 3 24 21 8 25 2 7
12 31 32 33 14 16 1 19 35 18 36 9 4 29 17 26 28 34 22 15 11 27 5 20 13 30 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 13 20 5 27 11 15 22 34 28 26 17 29 4 9 36 18 35 19 1 16 14 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 29 17 26 28 34 22 15 11 27 5 20 13 4 9 36 18 35 19 1 16 14 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 29 28 26 17 34 22 15 11 27 5 20 13 4 9 36 18 35 19 1 16 14 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 29 28 27 11 15 22 34 17 26 5 20 13 4 9 36 18 35 19 1 16 14 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 29 28 27 26 17 34 22 15 11 5 20 13 4 9 36 18 35 19 1 16 14 23 10 6 3 24 21 8 25 2 7
12 33 32 31 30 29 28 27 26 25 8 21 24 3 6 10 23 14 16 1 19 35 18 36 9 4 13 20 5 11 15 22 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 21 8 3 6 10 23 14 16 1 19 35 18 36 9 4 13 20 5 11 15 22 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 10 6 3 8 21 14 16 1 19 35 18 36 9 4 13 20 5 11 15 22 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 15 11 5 20 13 4 9 36 18 35 19 1 16 14 21 8 3 6 10 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 14 16 1 19 35 18 36 9 4 13 20 5 11 15 8 3 6 10 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 13 4 9 36 18 35 19 1 16 14 5 11 15 8 3 6 10 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 35 18 36 9 4 13 1 16 14 5 11 15 8 3 6 10 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 35 36 9 4 13 1 16 14 5 11 15 8 3 6 10 34 17 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 34 10 6 3 8 15 11 5 14 16 1 13 4 9 36 35 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14 5 11 15 8 3 6 10 34 1 13 4 9 36 35 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 11 5 14 8 3 6 10 34 1 13 4 9 36 35 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 5 11 8 3 6 10 34 1 13 4 9 36 35 2 7
12 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 1 34 10 6 3 8 11 5 4 9 36 35 2 7
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 34 10 6 3 4 5 11 8 9 36 35 2 7
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 34 35 36 9 8 11 5 4 3 6 10 2 7
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 34 35 36 9 8 7 2 10 6 3 4 5 11
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 34 35 36 9 8 7 6 10 2 3 4 5 11
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 34 35 36 9 8 7 6 5 4 3 2 10 11
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 1 2 3 4 5 6 7 8 9 36 35 34 10 11
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 9 8 7 6 5 4 3 2 1 10 11
1 2 3 4 5 6 7 8 9 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36

model: *reality and desire*



hurdle & fortress





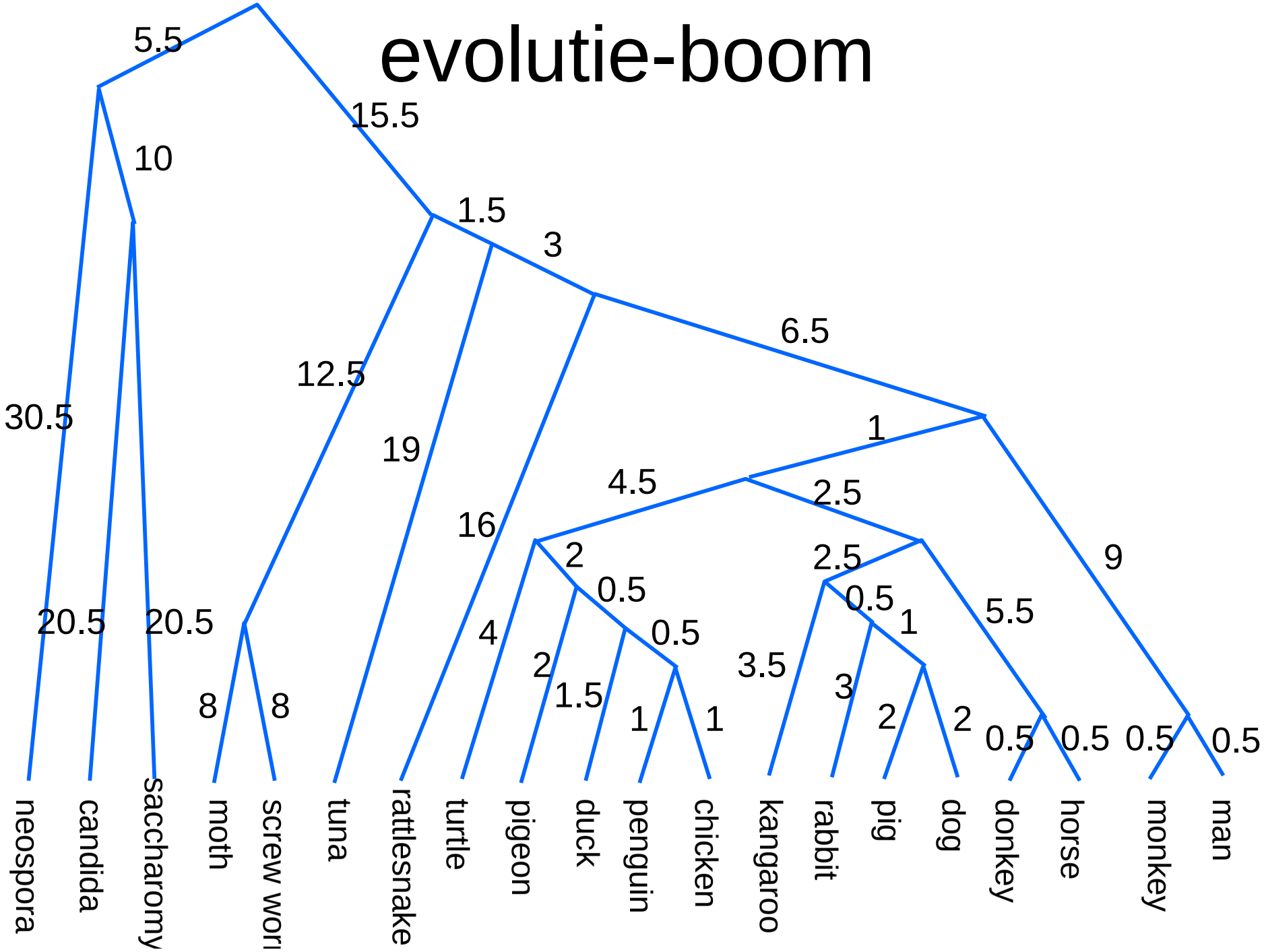
$$d(\pi) \geq b(\pi) - c(\pi) + h(\pi) + f(\pi)$$

uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

evolutie-boom

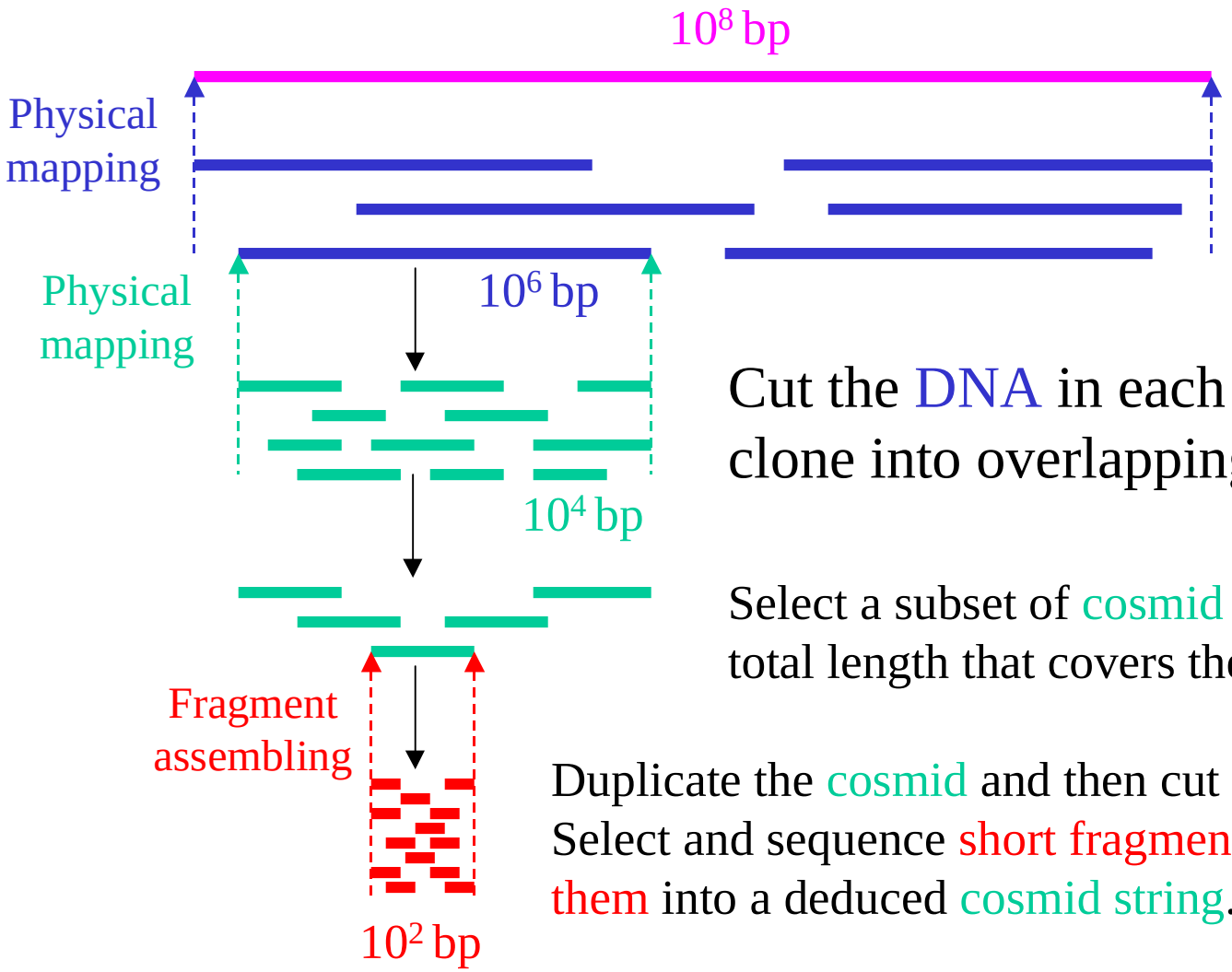
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	0	1	13	17	16	13	12	12	17	16	18	18	19	20	31	33	36	63	56	66	Man
2		0	12	16	15	12	11	13	16	15	17	17	18	21	32	32	35	62	57	65	Monkey
3			0	10	8	4	6	7	12	12	14	14	13	30	29	24	28	64	61	66	Dog
4				0	1	5	11	11	16	16	16	17	16	32	27	24	33	64	60	68	Horse
5					0	4	10	12	15	15	15	16	15	31	26	25	32	64	59	67	Donkey
6						0	6	7	13	13	13	14	13	30	25	26	31	64	59	67	Pig
7							0	7	10	8	11	11	11	25	26	23	29	62	59	67	Rabbit
8								0	14	14	15	13	14	30	27	26	31	66	58	68	Kangaroo
9									0	3	3	3	7	24	26	25	29	61	62	66	Pekin duck
10										0	4	4	8	24	27	26	30	59	62	66	Pigeon
11											0	2	8	28	26	26	31	61	62	66	Chicken
12												0	8	28	27	28	30	62	61	65	King penguin
13													0	30	27	30	33	65	64	67	Snapping turtle
14														0	38	40	41	61	61	69	Rattlesnake
15															0	34	41	72	66	69	Tuna
16																0	16	58	63	65	Screwworm fly
17																	0	59	60	61	Moth
18																		0	57	61	Neurospora
19																			0	41	Saccharomyces
20																				0	Candida


evolutie-boom
5.5
15.5
10
1.5
3
6.5
12.5
30.5
19
1
16
4.5
2.5
2
2.5
9
0.5
0.5
20.5
20.5
4
0.5
5.5
1
3.5
2
1.5
3
1
1
2
2
0.5
0.5
0.5
0.5
8
8
neospora
candida
saccharomy
moth
screw wor
tuna
rattlesnake
turtle
pigeon
duck
penguin
chicken
kangaroo
rabbit
pig
dog
donkey
horse
monkey
man

uitdagingen

- uitlijnen *alignment*
- databases
- 3d structuur
- inversie *sorting by reversal*
- boom *phylogenetic tree*
- combineren *physical mapping*

physical mapping



C: Full DNA

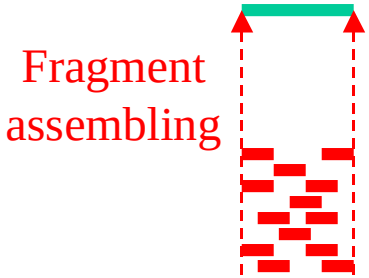
Cut *C* and clone into overlapping YAC clones.

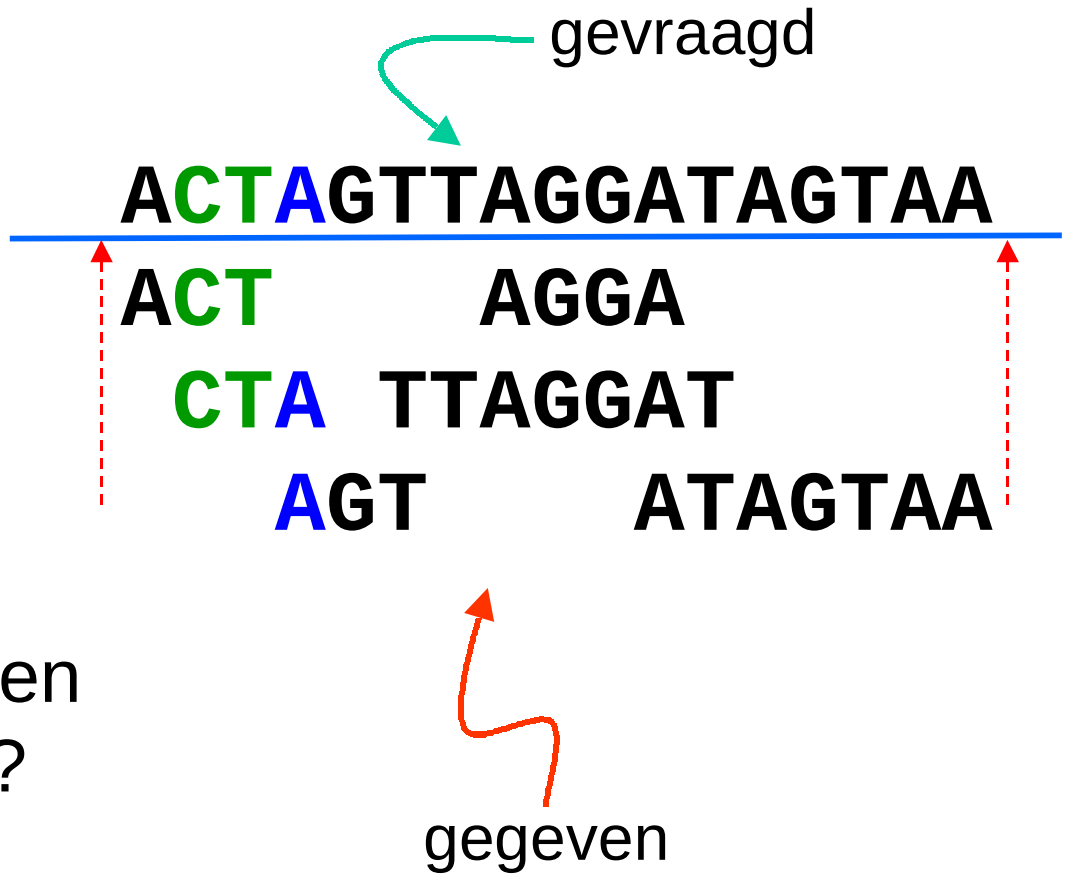
Cut the DNA in each YAC clone and clone into overlapping cosmid clones.

Select a subset of cosmid clones of minimum total length that covers the YAC DNA.

Duplicate the cosmid and then cut the copies randomly. Select and sequence short fragments and then reassemble them into a deduced cosmid string.

shortest common superstring





onnauwkeurigheden
unieke oplossing ?
NP-compleet :(

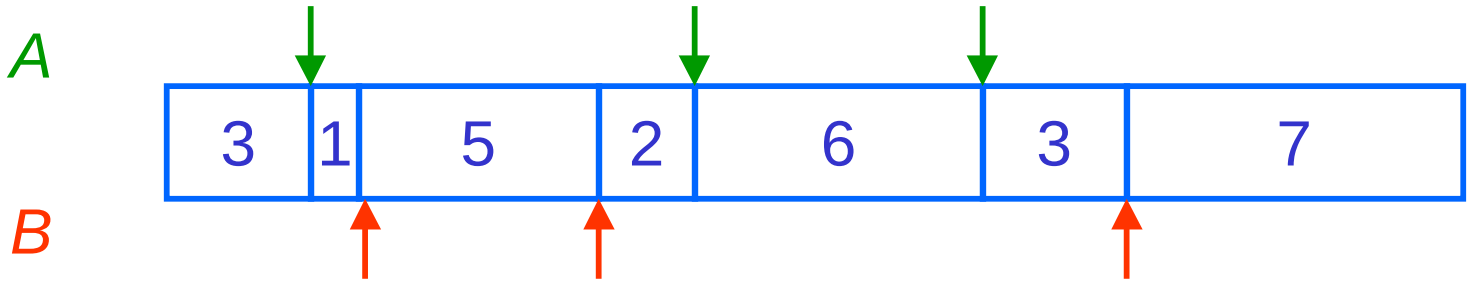
‘gretig’ algoritme

bepaal overlap tussen paren strings

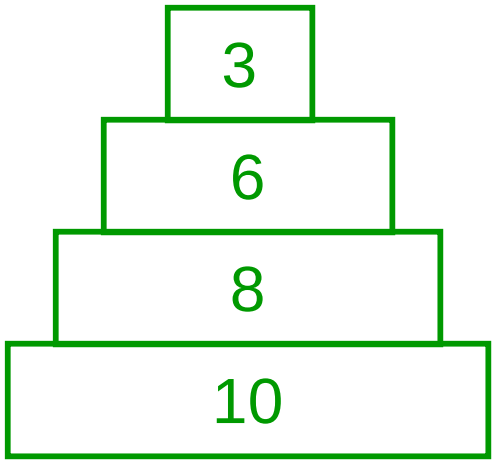
herhaal:
voeg paar met grootste overlap samen
bereken nieuwe overlaps
grootste overlap met zichzelf !?
apart leggen

tenslotte:
voeg apart gelegde strings samen

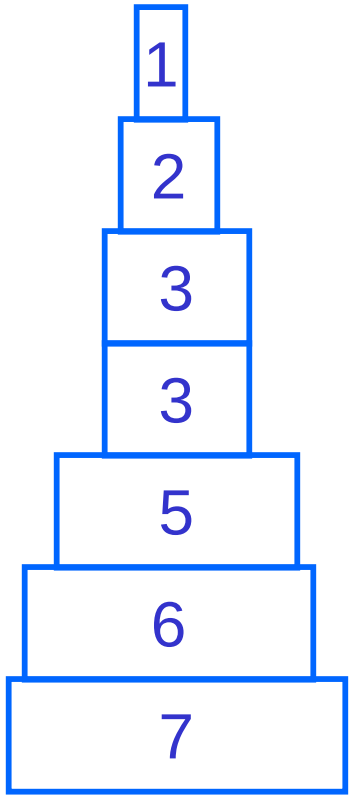
digest problem

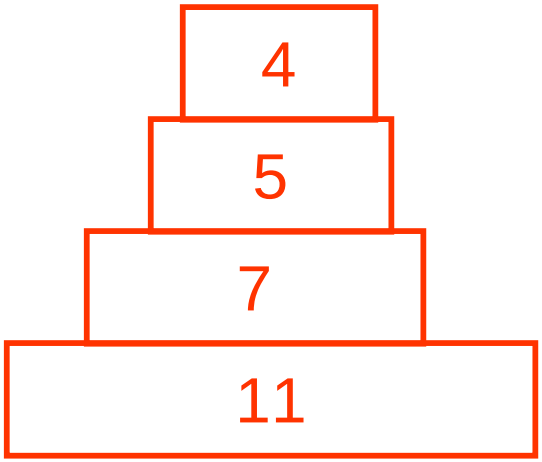


lange segmenten:
onbekende sequenties



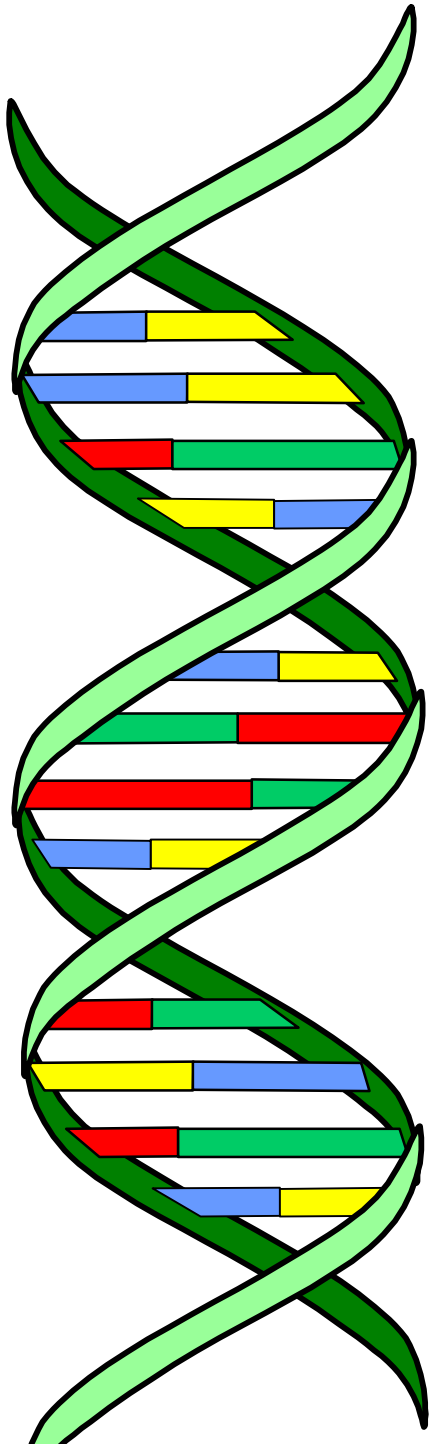
enzym A {3,6,8,10}





enzym B {4,5,7,11}

A+B {1,2,3,3,6,7}



Hendrik Jan Hoogeboom
hoogeboom@liacs.nl
voorjaar 2003
Universiteit Leiden
proefstuderen /
studievaardigheden

Een aantal plaatjes is
op internet gevonden, of
uit presentaties gehaald.
Bedankt google, CMBI Nijmegen,
R.C.T. Lee @ Chinan Univ.
(en al die anderen)


Leiden Institute of Advanced Computer Science
Research & Education

www.liacs.nl