KATHOLIEKE UNIVERSITEIT LEUVEN

Bioinformatics

Pairwise alignment

Revised 25/02/06

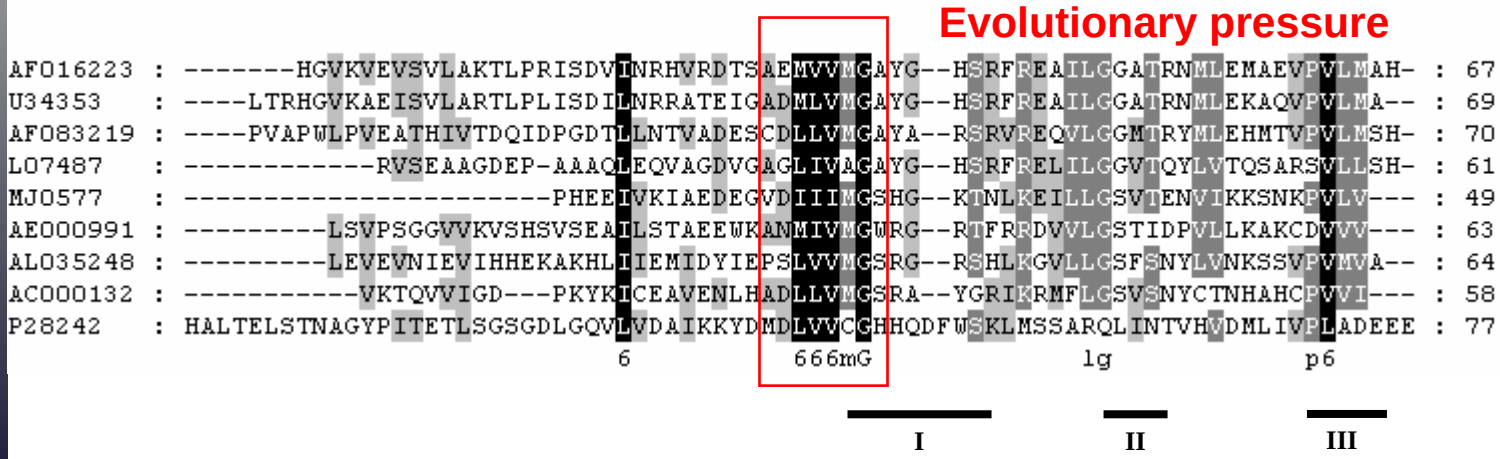


Bioi@Sista

Introduction

Why aligning sequences? Functional inference

- Clone and sequence gene with unknown function
- Aligning sequence with other sequence in databank
  - detect homologues with known function
    - Ortholog, paralog
    - detect conserved motifs characteristic for protein family
  - infer function from sequence alignment

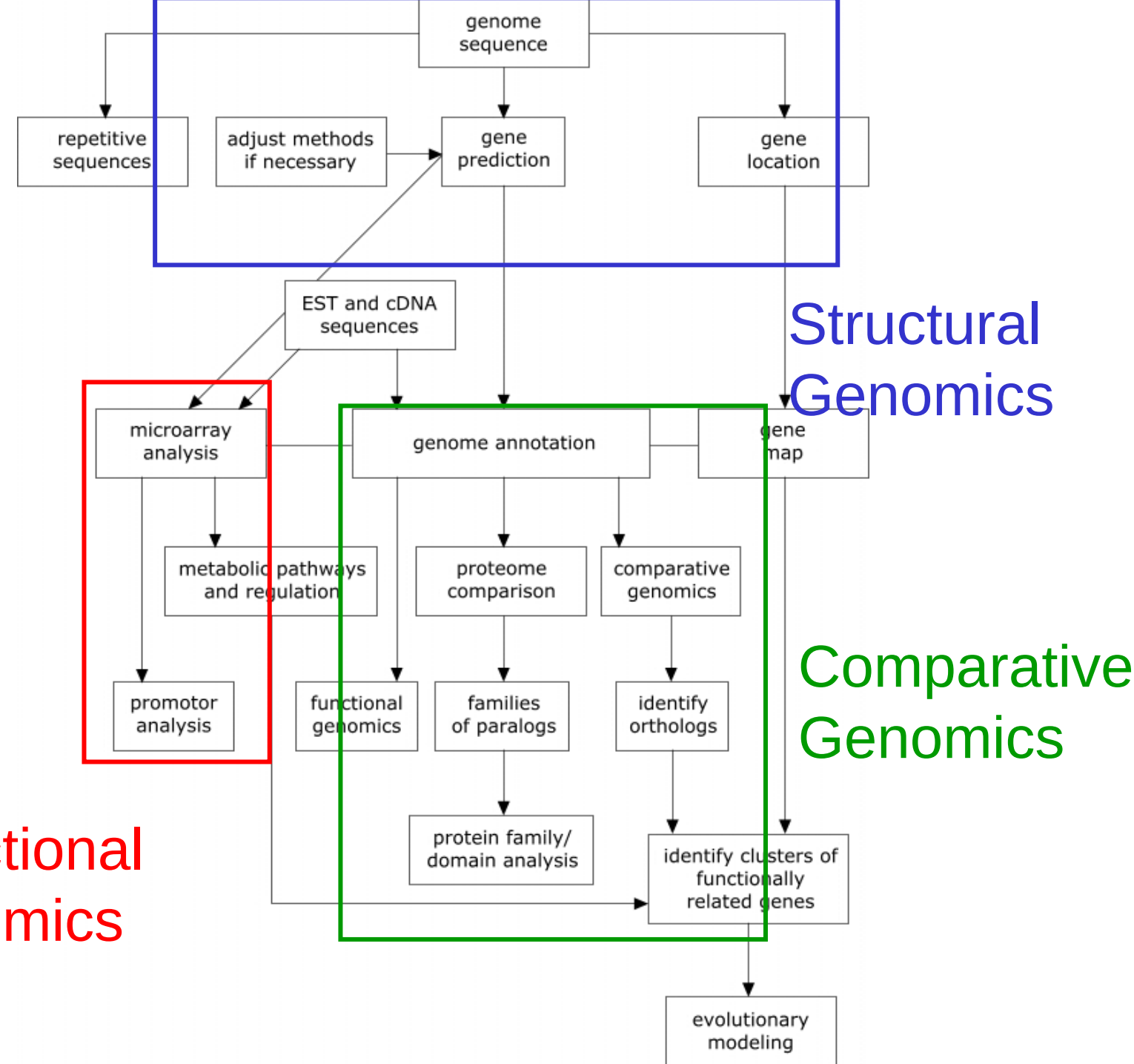


Introduction

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Homologous genes:**
  - **Exhibit sequence homology**
  - **Have similar ancestor**
    - **Orthologous genes**
    - **Paralogous genes**
- **Analogous genes:**
  - **convergent evolution**
  - **Similar function or structural protein fold**
  - **No common ancestor**

- **Alignment allows**
  - **functional inference**
  - **Reconstruction of phylogenetic relatedness**


genome sequence
repetitive sequences
adjust methods if necessary
gene prediction
gene location
EST and cDNA sequences
Structural Genomics
microarray analysis
genome annotation
gene map
metabolic pathways and regulation
proteome comparison
comparative genomics
promotor analysis
functional genomics
families of paralogs
identify orthologs
Comparative Genomics
protein family/ domain analysis
identify clusters of functionally related genes
Functional genomics
evolutionary modeling

Introduction

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Pairwise alignment:**
  1. aligning two sequences
  2. deciding whether the alignment is biologically relevant (two sequences are related) or whether the alignment occurred by chance

- **Key issues:**
  1. sorts of alignment (local versus global)
  2. the scoring system to rank the alignments
  3. algorithms to find alignments (versus heuristic)
  4. PAM and BLOSUM

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Pairwise alignment:**
  1. aligning two sequences
  2. deciding whether the alignment is biologically relevant (two sequences are related) or whether the alignment occurred by chance

- **Key issues:**
  1. sorts of alignment (local versus global)
  2. the scoring system to rank the alignments
  3. algorithms to find alignments (versus heuristic)
  4. PAM and BLOSUM

Global alignment

KATHOLIEKE UNIVERSITEIT LEUVEN

Best global alignment

```
               10        20        30        40        50        60
304992 MALKDLLVVVDDTAAAAAANRCRRPTGRRRTDGHITGLYPVVPLTLPGYVEAELPDEVRH
       :.    .::                 :   :::  ...  :
_      MSAYKTVVV----------------G---TDGSDSSMRAV------------------
                                10          20

               70        80        90       100       110       120
304992 AARLHREPDRQGGGSLRRRGAPQRPDRPLGMAGPLRSPDGQRPALHGRYADVVVVGQADP
                               ::   .::     :          : .....   :
_      ------------------------DRAAQIAGA----D----------AKLIIASAYLP
                                     30                        40

              130       140       150       160       170       180
304992 HRDRDRPIAVPQDLVFECGRPLLVRALRPALSPTSGNRRVLVAWNGSREAARWPTRCPSS
       ...  :   . .:  ..               ..:.  .    . ..: :.
_      QHEDARAADILKDESYK---------------VTGTAPIYEILHDAKERAH--------
             50                        60        70

              190       200       210       220       230       240
304992 PPPKRVVVMAVNPKAGPADRRRAGRRHRQAPVAPWLPVEATHIVTDQIDPGDTLLNTVAD
                         :: ..          ::   ::     : :.:.: . .
_      --------------------NAGAKN----------VEERPIVG---APVDALVNLADE
                           80                  90          100

              250       260       270       280
304992 ESCDLLVMGAYARSRVREQVLGGMTRYMLEHMTVPVLMSH--
       :. ::::.:  . : .  ..::..   . ..  : ::. :
_      EKADLLVVGNVGLSTIAGRLLGSVPANVSRRAKVDVLIVHTT
          110       120       130       140
```

Sequences are aligned over their entire region:

- **High homology**
- **Similar length**

Local alignment

```
          33.3% identity in 51 aa overlap; score:   92

      230       240       250       260       270       280
_       PGDTLLNTVADESCDLLVMGAYARSRVREQVLGGMTRYMLEHMTVPVLMSH
        : :.:.: . .:. :::::.:  . : .  ..::..   . ..  : ::. :
_       PVDALVNLADEEKADLLVVGNVGLSTIAGRLLGSVPANVSRRAKVDVLIVH
            100       110       120       130       140


          18.2% identity in 44 aa overlap; score:   33

       90       100       110       120       130
_       GMAGPLRSPDGQRPALHGRYADVVVVGQADPHRDRDRPIAVPQD
        :   . .:. : .    . :  : .....   :...  :    . .:
_       GSDSSMRAVD-RAAQIAGADAKLIIASAYLPQHEDARAADILKD
              20        30        40        50
```

Islands of homology:

- low homology
- different length

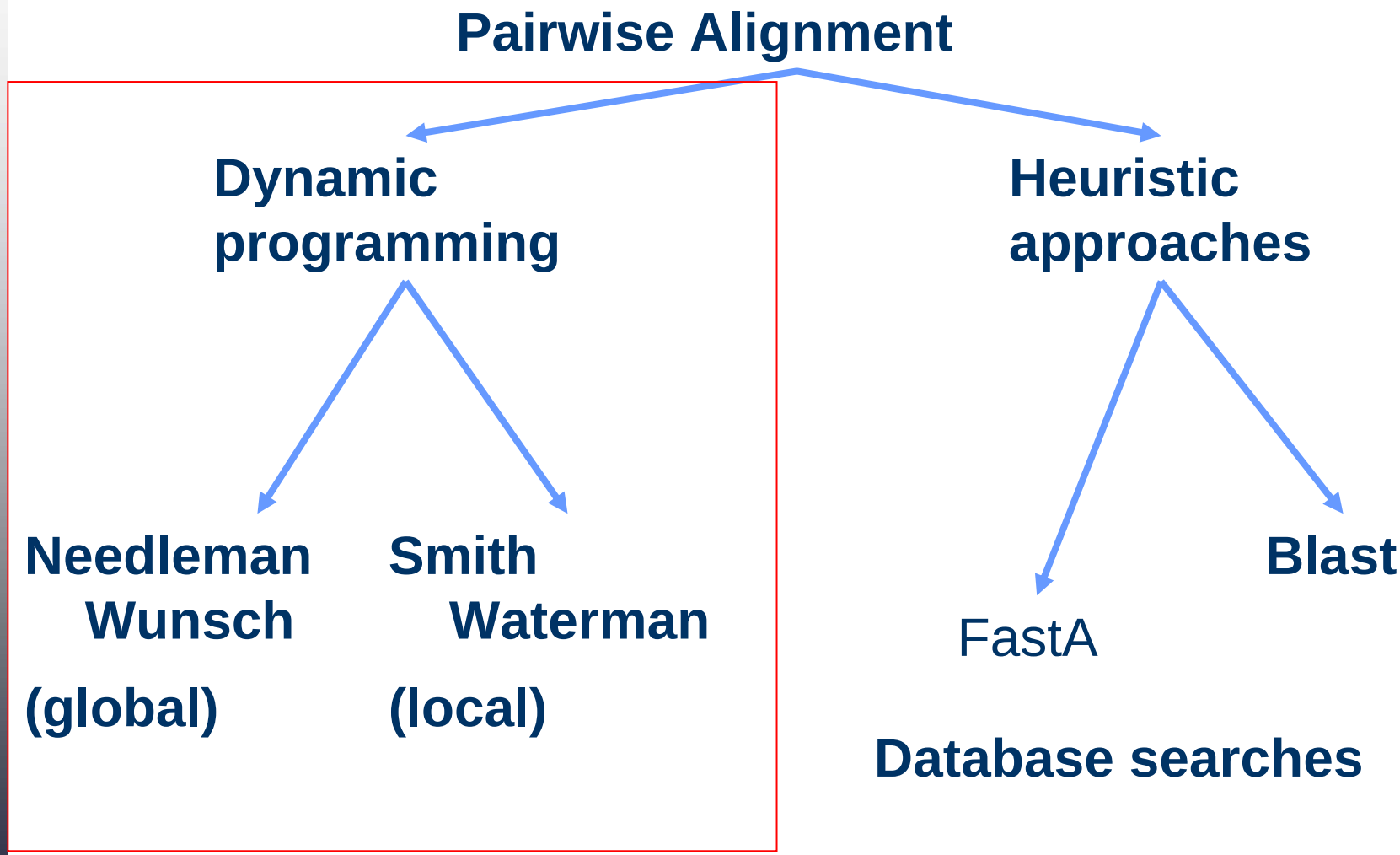
Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Pairwise alignment:**
  1. aligning two sequences
  2. deciding whether the alignment is biologically relevant (two sequences are related) or whether the alignment occurred by chance

- **Key issues:**
  1. sorts of alignment (local versus global)
  2. the scoring system to rank the alignments
  3. algorithms to find alignments (versus heuristic)
  4. PAM and BLOSUM

Algorithms



Chapter 1

Scoring Scheme

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Aligning = looking for evidence that sequences have diverged from a common ancestor by a process of natural selection.**
- **mutational processes:**
  1. **substitutions: change residues in a sequence,**
  2. **insertions: adding residues and**
  3. **deletions: removing residues.**

IGAx_i	IGALx	IGx--
LGVy_j	LGy--	LGVLy
substitution	insertion	deletion

- **total score of an alignment = the sum of terms**
  1. **For each aligned pair**
  2. **Plus terms for gaps**

Substitution Score

- **Ungapped global pairwise alignment:**
- **Assign a score to the alignment:**
  - **relative likelihood that sequences are related (MATCH MODEL)**
  - **to being unrelated (RANDOM MODEL)**

Random $p(x,y|R)=\prod_i q_{x_i}\prod_i q_{y_i}$

Match $p(x,y|M)=\prod_i p_{x_i y_i}$

- **ratio**

$$\prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

- **Log-odds ratio**

$$S=\sum_i s(x_i,y_i)$$

$$s(a,b)=\log\left(\frac{p_{ab}}{q_a q_b}\right)$$

IGAx

LGVy

Assumption of additivity! Independence between the aligned positions

Substitution Score

Substitution matrix (BLOSUM 50 matrix)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Log odds score can be positive (identities, conservative replacements) and negative

Gap Score

- **Gap penalties assign a negative score to the introduction of gaps (insertions, deletions)**

```
IGALx      IGx--
LGy--      LGVLy
```

- **Two types of gap scores have been defined:**
  - **linear score**
  - **affine score:**

$$\gamma(g) = -gd$$

$$\gamma(g) = -d - (g-1)e$$

with g gap length
with d gap open penalty
with e gap extension penalty

- **Gap penalties should be adapted to the substitution matrix**

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Pairwise alignment:**
  1. aligning two sequences
  2. deciding whether the alignment is biologically relevant (two sequences are related) or whether the alignment occurred by chance

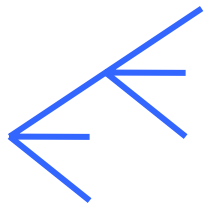
- **Key issues:**
  1. sorts of alignment (local versus global)
  2. the scoring system to rank the alignments
  3. algorithms to find alignments
  4. PAM and BLOSUM

Algorithms

- **an algorithm for finding an optimal alignment for a pair of sequences**
- **Suppose there are 2 sequences of length n that need to be aligned**

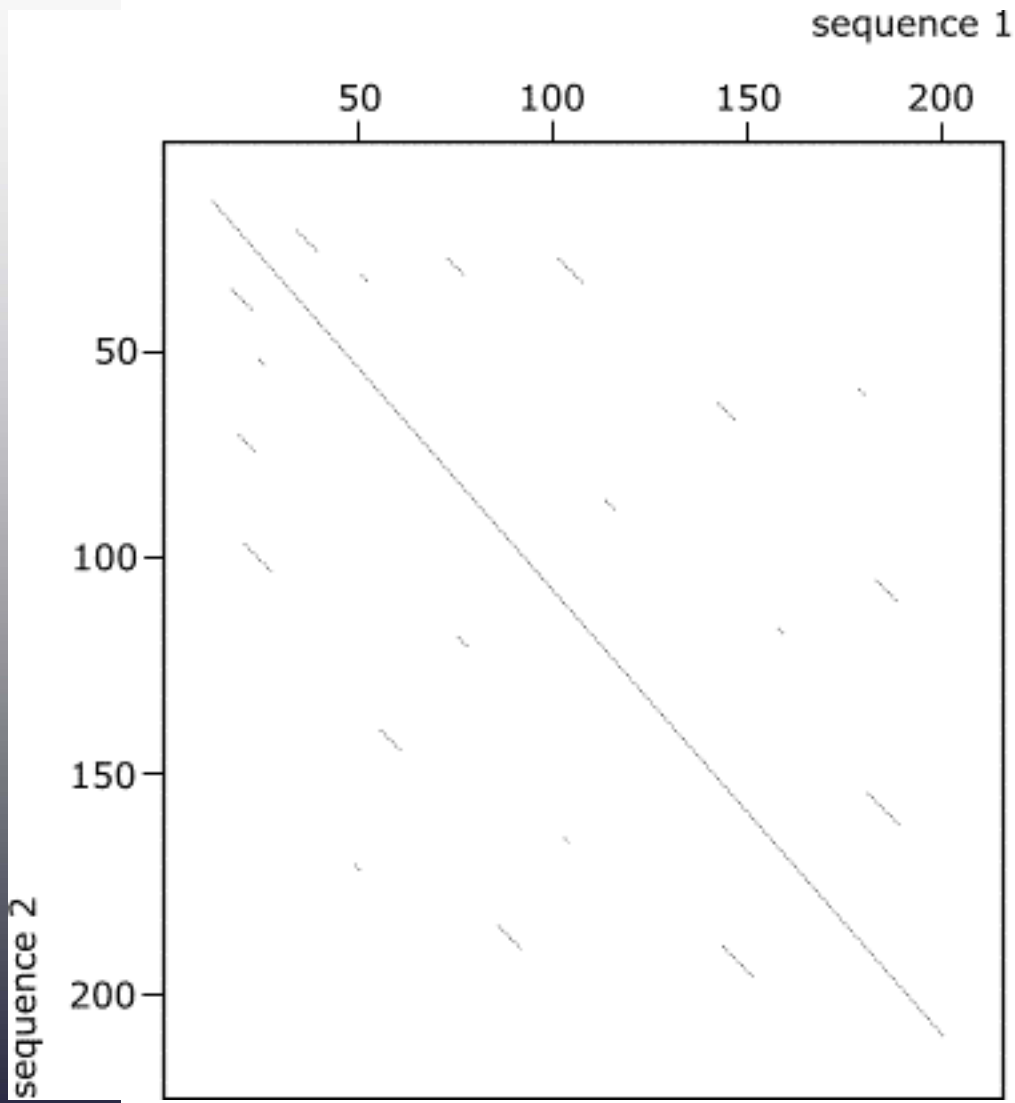
$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}}$$

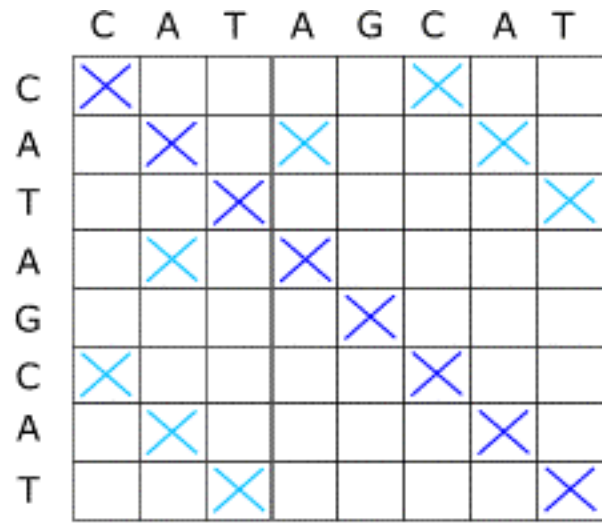
ATT and TTC	A	A	-
	T	-	T

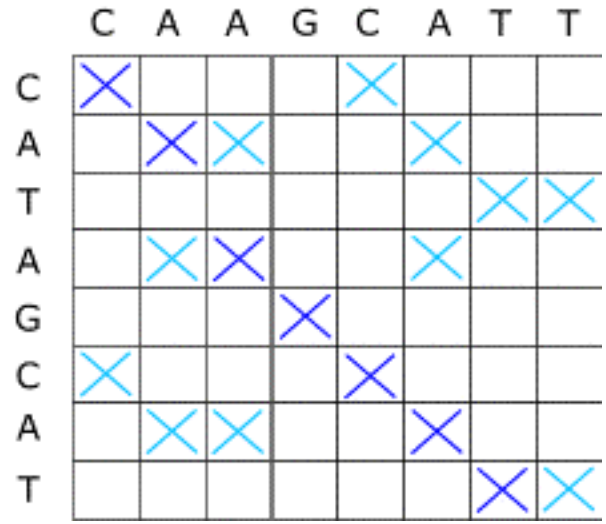


- Possible alignments between the 2 sequences
- Computationally infeasible to enumerate them all

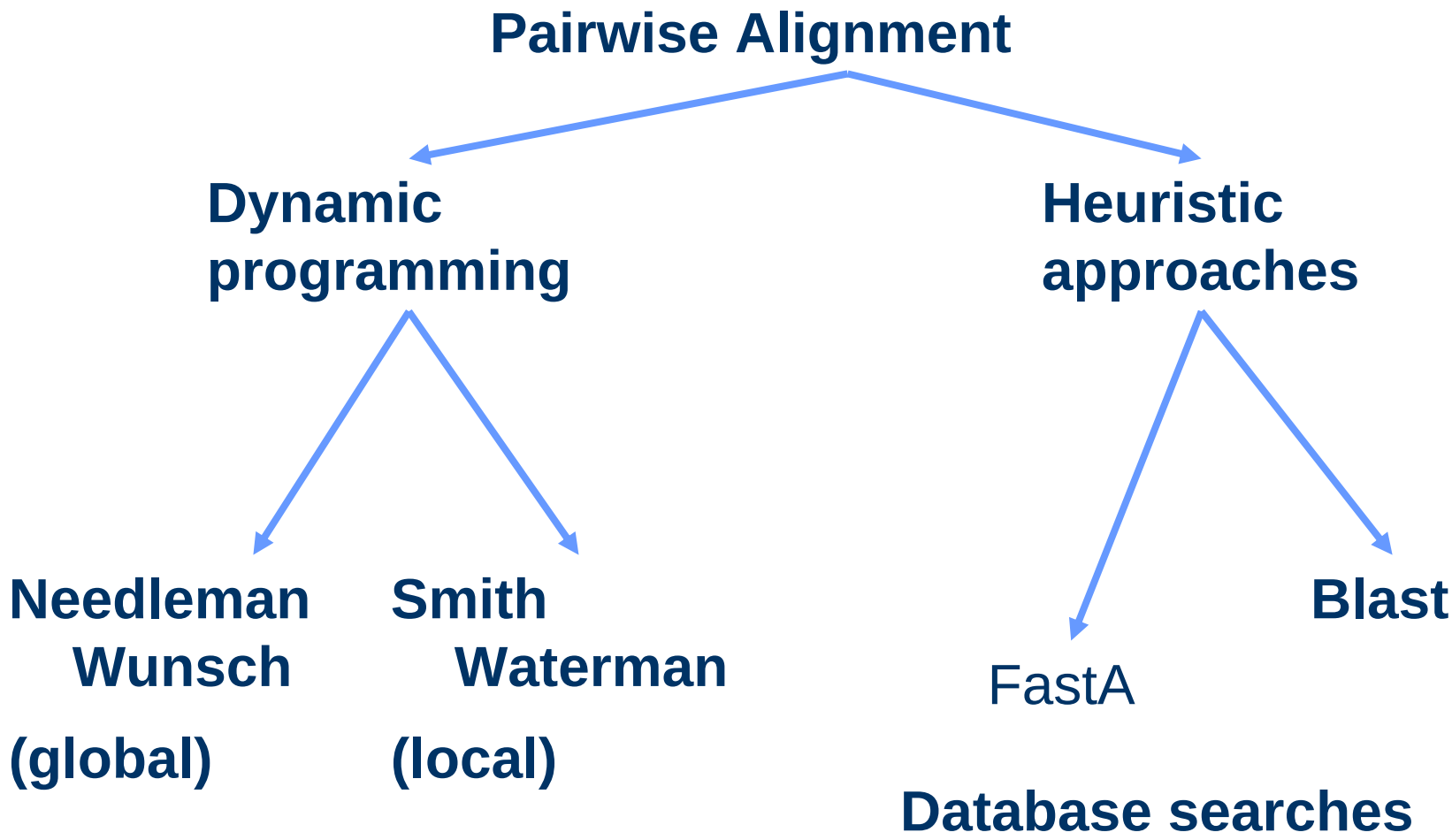
Visual Inspection







Algorithms


Pairwise Alignment
Dynamic programming
Heuristic approaches
Needleman Wunsch
(global)
Smith Waterman
(local)
FastA
Blast
Database searches

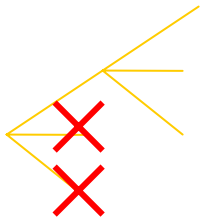
Dynamic Programming

KATHOLIEKE UNIVERSITEIT LEUVEN

Global Alignment: Needleman Wunsh

- Finding the optimal alignment = maximizing the score
- Construct matrix F, indexed by i and j
- F(i, j) is the score of the best alignment between the initial segment xi…1 of x up to xi and the initial segment y1…j up to yj
- Build F(i, j) recursively: start at F(0, 0) = 0 and proceed to fill the matrix from top left to bottom right

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \text{substitution} \\ F(i-1, j) - d & X_i \text{ aligned to a gap} \\ F(i, j-1) - d & Y_j \text{ aligned to a gap} \end{cases}$$



- Keep a pointer in each cell back to the cell from which it was derived
- Value of the final cell is the best score for the alignment

Dynamic Programming

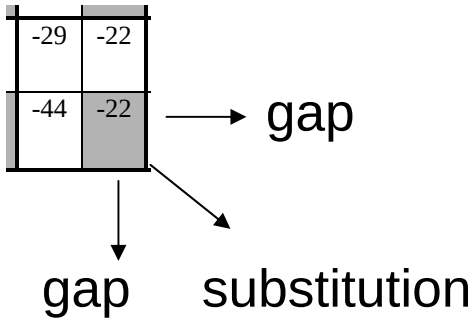
KATHOLIEKE UNIVERSITEIT LEUVEN

- **Alignment: path of choices which leads to the best score: traceback**
- **Build the alignment in reverse: move back to the cell from which F(i,j) was derived:**
  - » (i-1,j-1) depending on the pointer
  - » (i-1,j)
  - » (i, j-1)
- **Add a pair of symbols onto the current alignment**

- **Score is made of sum of independent pieces: score is the best score up to some point plus the incremental score**
- **Adaptations for local alignment, for more complex models (affine gap score)**

Dynamic programming

- **Any given point can only be reached from 3 possible positions**
- **Each new score is found by choosing the maximum of 3 possibilities**
- **For each square keep track of where the best score came from**



$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \text{substitution} \\ F(i-1, j) - d & X_i \text{ aligned to a gap} \\ F(i, j-1) - d & Y_j \text{ aligned to a gap} \end{cases}$$

Dynamic Programming

KATHOLIEKE UNIVERSITEIT LEUVEN

PAM250

Cys	C	12																		
Ser	S	0	2																	
Thr	T	-2	1	3																
Pro	P	-3	1	0	6															
Ala	A	-2	1	1	1	2														
Gly	G	-3	1	0	-1	1	5													
Asn	N	-4	1	0	-1	0	0	2												
Asp	D	-5	0	0	-1	0	1	2	4											
Glu	E	-5	0	0	-1	0	0	1	3	4										
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4									
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6								
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	8							
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
Leu	L	-8	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8			
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
Trp	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
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

Dynamic Programming

GAP		L		S	
	0	-12	-12	-16	-16
L	-12	8	-24	-15	-28
	-12	-24	8	-4	-4
S	-16	-15	-4	10	-16
	-16	-28	-4	-16	10

Dynamic Programming

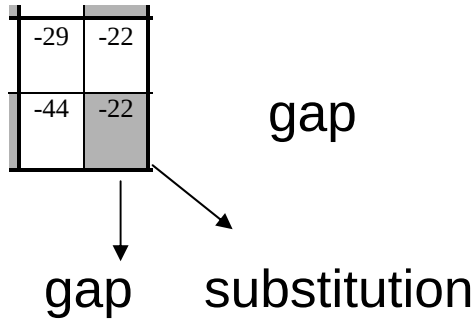
GAP		M		N		A		L		S		D		R		T	
	0	-12	-12	-16	-16	-20	-20	-24	-24	-28	-28	-32	-32	-36	-36	-40	-40
M	-12	6	-24	-14	-28	-19	-32	-16	-36	-26	-40	-31	-44	-29	-48	-37	-52
	-12	-24	6	-6	-6	-10	-10	-14	-14	-18	-18	-22	-22	-26	-26	-30	-30
G	-16		-6	6	-18	-5	-22	-14	-26	-13	-30	-19	-34	-25	-32	-26	-42
	-16	-28	-6	-18	6	-6	-5	-17	-14	-26	-13	-17	-17	-21	-21	-25	-25
S	-20	-18	-10	-5	-6	7	-17	-8	-26	-12	-25	-13	-29	-17	-33	-20	-37
	-20	-32	-10	-22	-5	-17	7	-5	-5	-9	-12	-24	-13	-25	-17	-29	-20
D	-24	-23	-14	-8	-17	-5	-5	3	-17	-5	-24	-8	-25	-14	-29	-17	-32
	-24	-36	-14	-36	-8	-20	-5	-17	3	-9	-5	-17	-8	-20	-14	-26	-17
R	-28	-24	-18	-14	-20	-10	-17	-8	-9	3	-17	-6	-20	0	-26	-15	-29
	-28	-40	-18	-30	-14	-26	-10	-22	-8	-20	3	-9	-6	-18	0	-12	-15
T	-32	-29	-22	-18	-26	-13	-22	-12	-20	-7	-9	3	-18	-7	-12	3	-27
	-32	-44	-22	-24	-18	-30	-13	-25	-12	-24	-7	-19	3	-9	-7	-19	3

Affine gap cost

•Gap open : -12

•Gap Extension: -4

Substitution cost: PAM250



MNALSDRT

M--GSDRT

Dynamic Programming

KATHOLIEKE UNIVERSITEIT LEUVEN

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6 (6)	-6 (-2)	-10	-14	-18	-22	-26	-30
G	-16	-6 (-3)	6 (0)	-5	-10	-13	-17	-22	-26
S	-20	-10	-5	7	-5	-8	-13	-17	-21
D	-24	-14	-8	-5	3	-5	-4	-14	-17
R	-28	-18	-14	-10	-8	3	-6	2	-10
T	-32	-22	-18	-13	-12	-7	3	-7	5
T	-36	-26	-22	-17	-15	-11	-7	2	-4
E	-40	-30	-25	-22	-20	-15	-8	-8	0
T	-44	-34	-30	-24	-24	-21	-15	-9	-5

MNALSDRT---

--MGSDRTTET

MNA-LSDRT

MGSDRTTET

Dynamic Programming

KATHOLIEKE UNIVERSITEIT LEUVEN

Local Alignment: Smith Waterman

- No negative scores are allowed
- Portions of each sequence that are in the high scoring regions are reported

	GAP	M	N	A	L	S	D	R	T
GAP	0	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	3	0	12	3
T	0	0	0	1	0	1	3	0	15
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

SDRT

SDRT

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Pairwise alignment:**
  1. aligning two sequences
  2. deciding whether the alignment is biologically relevant (two sequences are related) or whether the alignment occurred by chance

- **Key issues:**
  1. sorts of alignment (local versus global)
  2. the scoring system to rank the alignments
  3. algorithms to find alignments (versus heuristic)
  4. PAM and BLOSUM

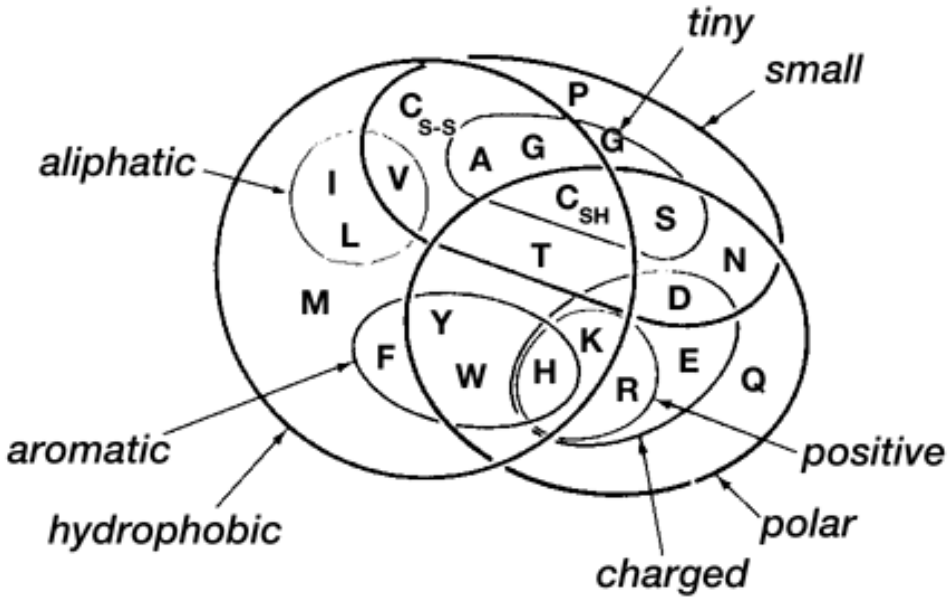
Substitution matrix

KATHOLIEKE UNIVERSITEIT LEUVEN

- a good random sample of confirmed alignments
- determine substitutions probabilities by counting the frequencies of the aligned residue pairs in the confirmed alignments and setting the probabilities to the normalized frequencies

The performance of the alignment programs depends to a large extent on how well the substitution matrices are adapted to the dataset to be aligned

Substitution matrix



BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

- BLOSUM: Henikoff and Henikoff

- Protein families from database
- Construct block = ungapped alignment
-

```
WWYIR   CASILRKIYIYGPV   GVSRLRTAYGGRK   NRG
WFYVR   CASILRHLYHRSPA   GVGSITKIYGGRK   RNG
WYYVR   AAAVARHIYLRKTV   GVGRLRKVHGSTK   NRG
WYFIR   AASICRHLYIRSPA   GIGSFEKIYGGRR   RRG
WYYTR   AASIARKIYLRQGI   GVHHFQKIYGGRQ   RNG
WFYKR   AASVARHIYMRKQV   GVGKLNKLYGGAK   SRG
WFYKR   AASVARHIYMRKQV   GVGKLNKLYGGSK   RRG
WYYVR   TASVARRLYIRSPT   GVGALRRVYGGNK   RRG
WFYTR   AASTARHLYLRGGA   GVGSMTKIYGGRQ   RNG
WWYVR   AAALLRRVYIDGPV   GVNSLRTHYGGKK   DRG
```

- counted the number of occurrences

BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

One block

R A R A

A A A C

A A C C

A A R A

A A C C

A A R C

Observed frequency

$$q_a = \frac{\sum_b A_{ab}}{\sum_{cd} A_{cd}}$$

q(A)	14/24
q(R)	4/24
q(C)	6/24

Proportion observed

$$p_{ab} = \frac{A_{ab}}{\sum_{cd} A_{cd}}$$

p(A to A)	26/60
p(A to R)	8/60
p(A to C)	10/60
p(R to R)	3/60
p(R to C)	6/60
p(C to C)	7/60

Proportion expected

$$e_{ab} = q_a q_b$$

e(A to A)	14/24 * 14/24
e(A to R)	(14/24 *4/24) *2
e(A to C)	(14/24 * 6/24) *2
e(R to R)	4/24*4/24
e(R to C)	(4/24 * 6/24) *2
e(C to C)	6/24*6/24

BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

aligned pair	proportion observed	proportion expected	2 $\log_2$(proportion observed/proportion expected)
A to A	26/60	196/576	0.70
A to R	8/60	112/576	-1.09
A to C	10/60	168/576	-1.61
R to R	3/60	16/576	1.70
R to C	6/60	48/576	0.53
C to C	7/60	36/576	1.80

BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

- p_{ab} i.e. the fraction of pairings between *a* and *b* out of all observed pairs.
- For each pair of amino acids *a* and *b*, the estimated e_{ab}
- s(a,b). This quantity is the ratio of the log likelihood that *a* and *b* are actually observed aligned in the same column in the blocks to the probability that they are aligned by chance, given their frequencies of occurrence in the blocks.
- The resulting log odds values are scaled and rounded to the nearest integer value. In this way, pairs that are more likely than chance will have positive scores, and those less likely will have negative scores.

BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

- The first four sequences possibly derive from closely related species and the last three from three more distant species. Since A occurs with high frequency in the first four sequences, the observed number of pairings of A with A will be higher than is appropriate if we are comparing more distantly related sequences.
- Ultimately, each block should have sequences such that any pair have roughly the same amount of 'evolutionary distance' between them
- those sequences in each block that are 'sufficiently close' to each are treated as a single sequence
- BLOSUM45, BLOSUM62, and BLOSUM80
- larger-numbered matrices correspond to recent divergence, smaller-numbered matrices correspond to distantly related sequences.
- BLOSUM62 standard for ungapped alignments, BLOSUM 50 alignments with gaps
-

```
A A A C
A A A C
A A C C
A A A C
C A C T
A R G C
```

BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

1 block

1 cluster

C	R	R
C	R	R
A	R	C
A	A	C

BLOSUM45: sequences that show a homology of at least 45% are treated as a single sequence

Observed frequency

$$q_a = \frac{\sum_b A_{ab}}{\sum_{cd} A_{cd}}$$

q(A)	3/9
q(R)	3/9
q(C)	3/9

Proportion observed

$$p_{ab} = \frac{A_{ab}}{\sum_{cd} A_{cd}}$$

p(A to A)	1/9
p(A to R)	2/ 9
p(A to C)	2/ 9
p(R to R)	1/ 9
p(R to C)	2/ 9
p(C to C)	1/ 9

BLOSUM62

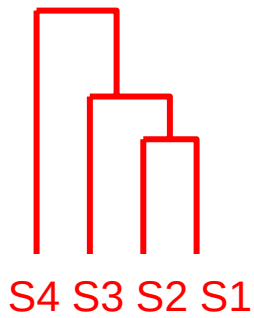
KATHOLIEKE UNIVERSITEIT LEUVEN

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

PAM

- The construction of PAM matrices starts with ungapped multiple alignments of proteins into blocks for which all pairs of sequences in any block are, as in the BLOSUM procedure, 'sufficiently close' to each other.
- This is important because the initial goal is to create a transition matrix for a short enough time period so that multiple mutations are unlikely.
- phylogenetic reconstruction (MP)

- In a maximum parsimony tree, the number of changes can be counted



PAM

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0	30	109	154	33	93	266	579	21	66	95	57	29	20	345	772	590	0	20	365
Arg	30	0	17	0	10	120	0	10	103	30	17	477	17	7	67	137	20	27	3	20
Asn	109	17	0	532	0	50	94	156	226	36	37	322	0	7	27	432	169	3	36	13
Asp	154	0	532	0	0	76	831	162	43	13	0	85	0	0	10	98	57	0	0	17
Cys	33	10	0	0	0	0	0	10	10	17	0	0	0	0	10	117	10	0	30	33
Gln	93	120	50	76	0	0	422	30	243	8	75	147	20	0	93	47	37	0	0	27
Glu	266	0	94	831	0	422	0	112	23	35	15	104	7	0	40	86	31	0	10	37
Gly	579	10	156	162	10	30	112	0	10	0	17	60	7	17	49	450	50	0	0	97
His	21	103	226	43	10	243	23	10	0	3	40	23	0	20	50	29	14	3	40	30
Ile	66	30	36	13	17	8	35	0	3	0	253	43	57	90	7	20	129	0	13	661
Leu	95	17	37	0	0	75	15	17	40	253	0	39	207	167	43	32	52	13	23	303
Lys	57	477	322	85	0	147	104	60	23	43	39	0	90	0	43	168	200	0	10	17
Met	29	17	0	0	0	20	7	7	0	57	207	90	0	17	4	20	28	0	0	77
Phe	20	7	7	0	0	0	0	17	20	90	167	0	17	0	7	40	10	10	260	10
Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7	0	269	73	0	0	50
Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	169	0	696	17	22	43
Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	695	0	0	23	186
Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0	0	6	0
Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	0	17
Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	0

Observed number of times Ala was replaced by Arg in a sequence and its immediate ancestor on the tree

$$A_{ab}$$

Convert the observed empirical observations into probabilities

$$P(j|k) = B_{j,k} = \frac{A_{jk}}{\sum_m Ajm}$$

PAM

- **Mutation probability matrix**

Values are multiplies by 10000

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1

One element in this matrix, [Mij], denotes the chance that an amino acid in column *j* will be replaced by an amino acid in row *i*, when these sequences have diverged over a 1 PAM distance.

PAM

- **To correct for longer evolutionary distances: multiply PAM1 eg PAM250**

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Values are multiplies by 100

PAM

For alignments PAM matrices are converted into log odds matrices

Cys	C	12																	
Ser	S	0	2																
Thr	T	-2	1	3															
Pro	P	-3	1	0	6														
Ala	A	-2	1	1	1	2													
Gly	G	-3	1	0	-1	1	5												
Asn	N	-4	1	0	-1	0	0	2											
Asp	D	-5	0	0	-1	0	1	2	4										
Glu	E	-5	0	0	-1	0	0	1	3	4									
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4								
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6							
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6						
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5					
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6				
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5			
Leu	L	-8	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8		
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	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
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

The odds score represents the likelihood that the two amino acids will be aligned in alignments of similar proteins divided by the likelihood that they will be aligned by chance in an alignment.

PAM vs BLOSUM

KATHOLIEKE UNIVERSITEIT LEUVEN

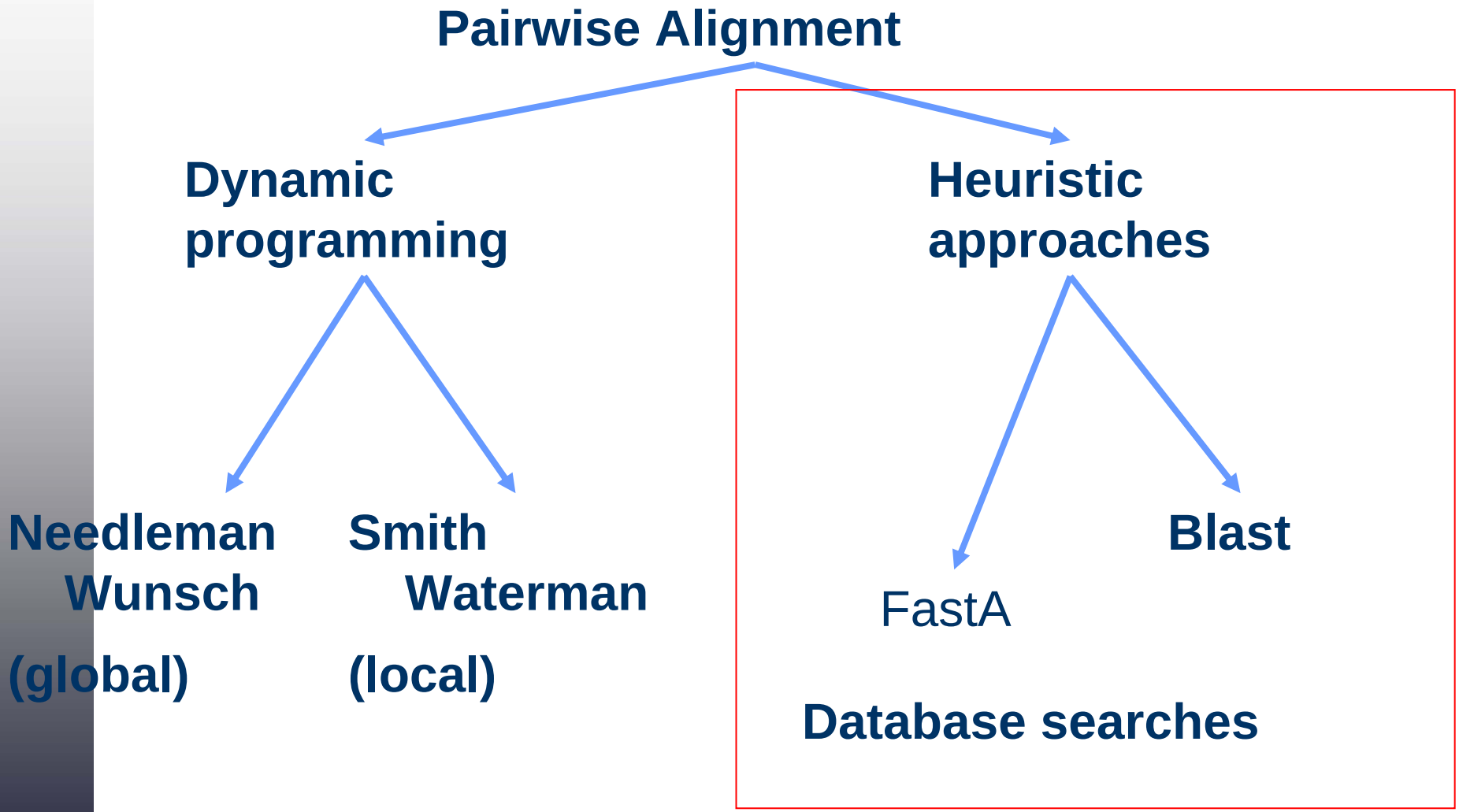
- PAM matrix based on an evolution model
  - All amino acids evolve at the same rate
  - The rate of evolution remains unaltered over long periods of time

PAM should be better than BLOSUM

More advanced scoring schemes for evolutionary modeling and phylogeny have been developed

- To detect sequence similarity:
  - The best alignment is obtained when an matrix adapted to the evolutionary distance between the 2 studied sequences is used

Algorithms



Chapter 1

Heuristic Pairwise: FASTA

KATHOLIEKE UNIVERSITEIT LEUVEN

Rather than comparing individual residues in two sequences, FASTA (Fast Alignment) searches for matching sequence patterns or words, or k-tuples.

sequence 1: ACNGTSCHQE
sequence 2: GCHCLSAGQD

Amino Acid	Position in Seq 1	Position in Seq 2	Offset Value
A	1	7	6
C	2	2	0
	2	4	2
	7	2	-5
	7	4	-3
D	-	10	-
E	10	-	-
G	4	1	-3
	4	8	-4
H	8	3	-5
L	-	5	-
N	3	-	-
Q	9	9	0
S	6	6	0
T	5	-	-

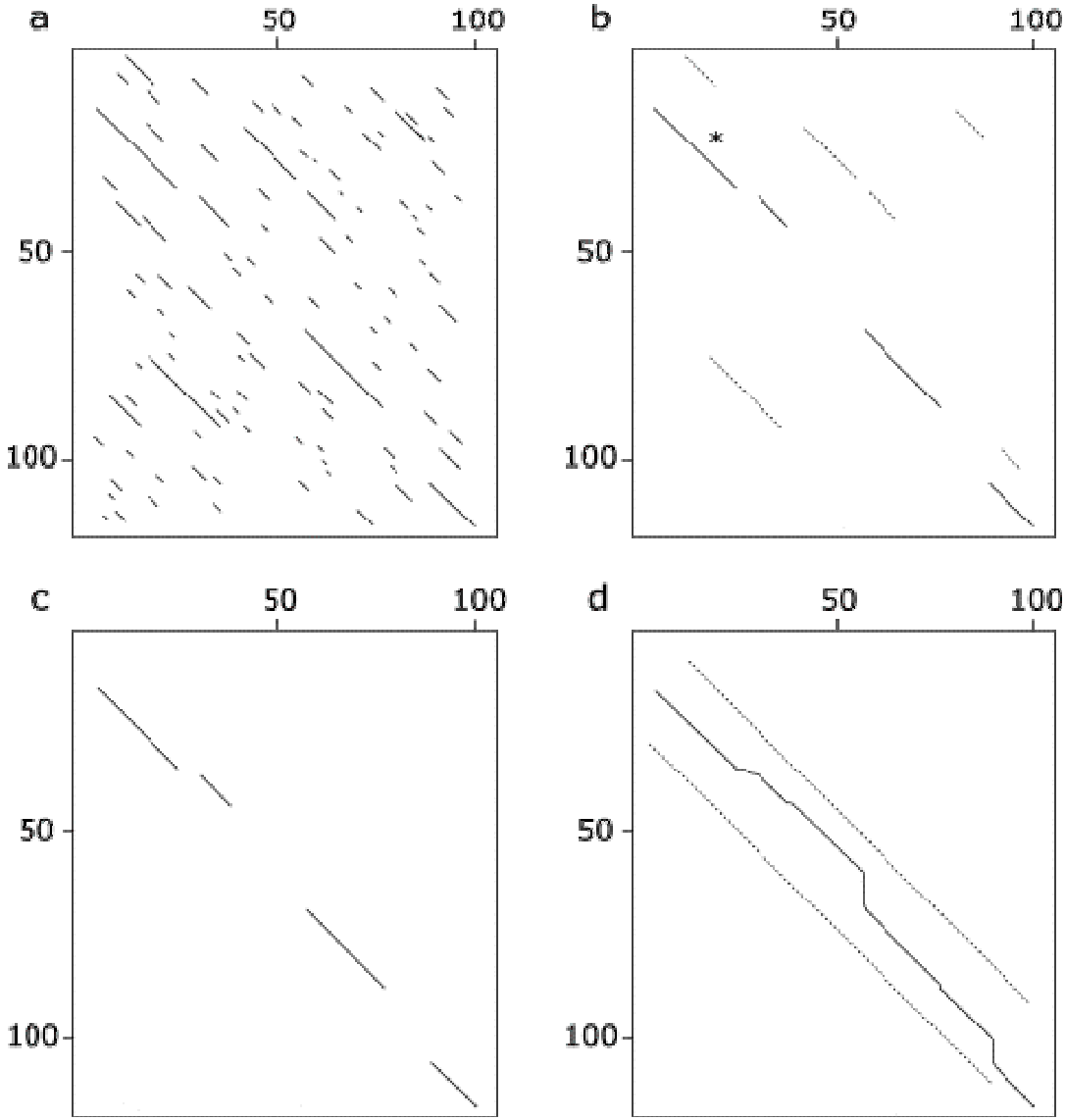
```
sequence 1: ACNGTSCHQE
              C   S  Q        << offset = 0
sequence 2: GCHCLSAGQD

sequence 1: ACNGTSCHQE---
                G  C          << offset = -3
sequence 2: ---GCHCLSAGQD

sequence 1: ACNGTSCHQE-----
                  CH          << offset = -5
sequence 2: -----GCHCLSAGQD
```

Heuristic Pairwise: FASTA

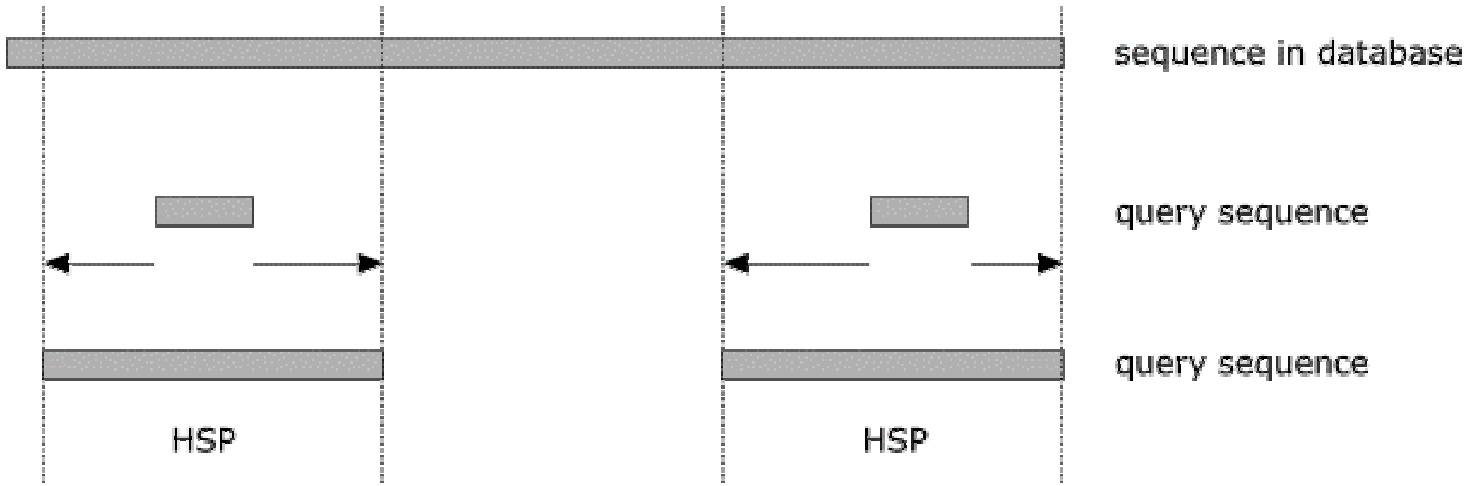
- all sets of k consecutive matches are detected (see dot plot).
- the 10 best-matching regions between the query sequence and the sequence in the database are identified.
- an optimal subset of regions is identified that can be combined into one initial, non-overlapping alignment.
- a full local alignment is performed using the Smith-Waterman dynamic programming algorithm.



Heuristic Pairwise: **Blast**

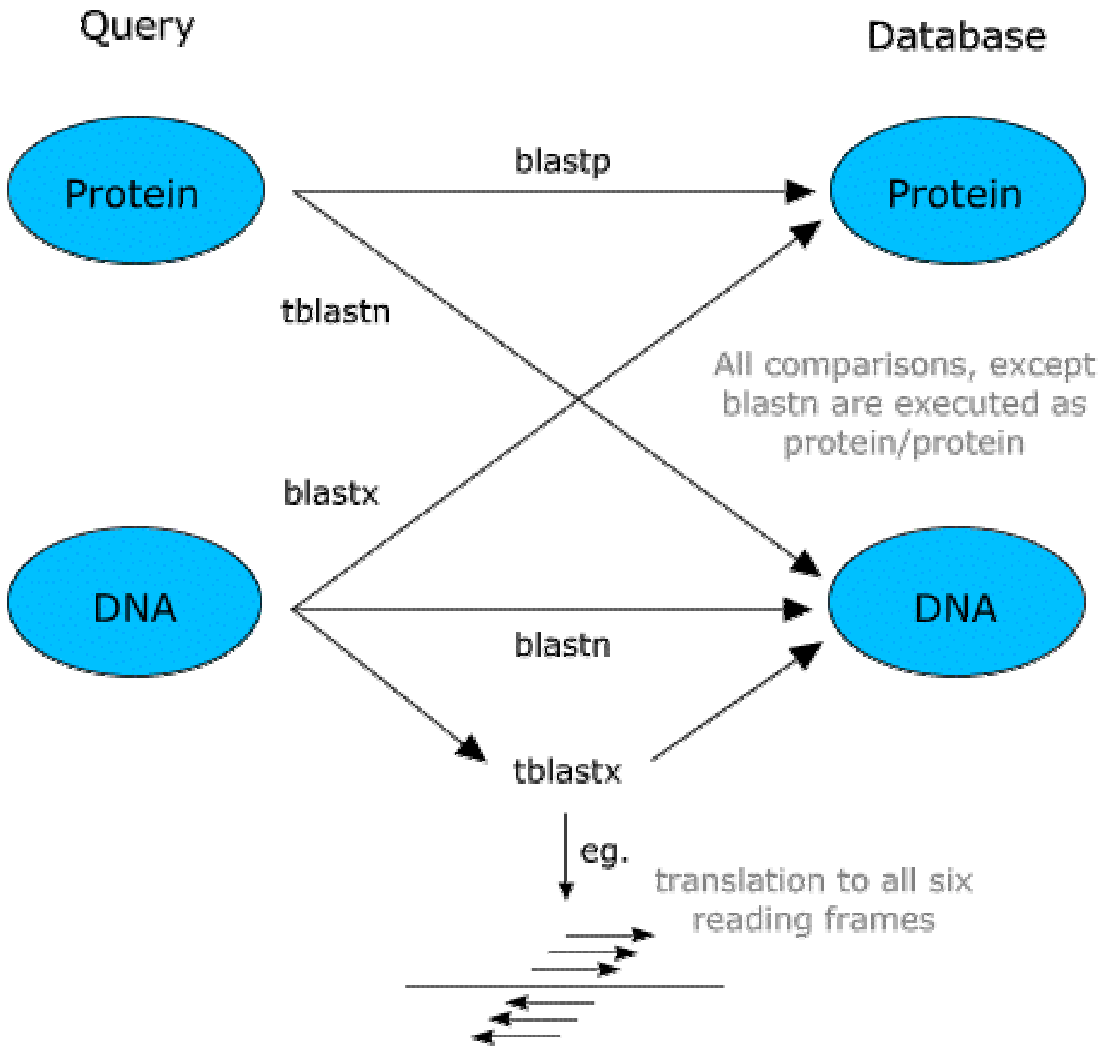
KATHOLIEKE UNIVERSITEIT LEUVEN

- **Heuristic approach**
  - **Starts with small words**
  - **Extends the words (HSP)**



Heuristic Pairwise: **Blast**

- **BlastP**
- **BlastN**
- **BlastX**
- **tBlastN**
- **tBlastX**



Heuristic Pairwise: **Blast**

```
gi|77565|pir||S08270   cytochrome-c oxidase (EC 1.9.3.1) chain I-beta - Paracoccus
          denitrificans
gi|2914141|pdb|1AR1|A  Chain A, Structure At 2.7 Angstrom Resolution Of The Paracoccus
          Denitrificans Two-Subunit Cytochrome C Oxidase Complexed
          With An Antibody Fv Fragment
gi|45479|emb|CAA68821.1|   cytochrome c oxidase beta subunit I; cytochrome-c oxidase
          [Paracoccus denitrificans]
         Length = 558

 Score =  963 bits (2489), Expect = 0.0
 Identities = 527/544 (96%), Positives = 527/544 (96%)

Query: 15  TRGFFTRWFMSTNHKDIGILYLFTAGIVGLISVCFTVYMRMELQHPGVQYMCLEGARLIA 74
           TRGFFTRWFMSTNHKDIGILYLFTAGIVGLISVCFTVYMRMELQHPGVQYMCLEGARLIA
Sbjct: 15  TRGFFTRWFMSTNHKDIGILYLFTAGIVGLISVCFTVYMRMELQHPGVQYMCLEGARLIA 74
```

- **Query:** sequence submitted
- **Sbjct:** homolog found
- **Bit score:** derived from log odds score Smith waterman
- **Expect:** Score derived from extreme value distribution: probability to observe such a homology (bitscore) by chance)
- **Identities:** percentage of identical residues on the length the sequence aligned
- **Positives**: percentage of similar residues on the length the sequence aligned

Absolute alignment scores

KATHOLIEKE UNIVERSITEIT LEUVEN

- Dependent on sequence length
- Not comparable between alignments
- Increase with the choice of a more appropriate scoring scheme

- To make scores comparable between alignments

=> statistical scores, p-values

Statistical significance

KATHOLIEKE UNIVERSITEIT LEUVEN

All tests of statistical significance involve a comparison between

1. observed values
2. that value that one would expect to find on average if only random variability was operating

P-value: probability of observing a score value by chance (assuming that the H0 is valid)

H0: sequences are unrelated = extreme value distribution

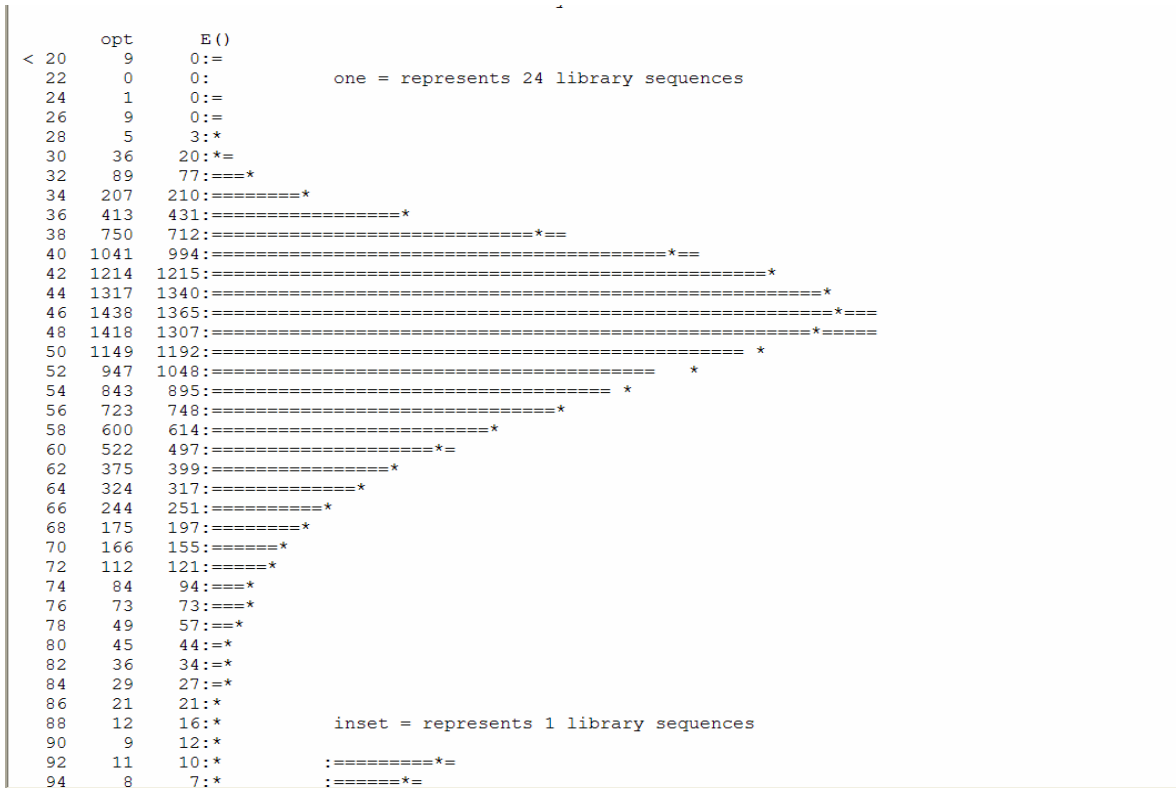
Calculate based on the p-value the E-value

The E-value takes into account the size of the database

E(score > S) = database size * p-value

Statistical significance

KATHOLIEKE UNIVERSITEIT LEUVEN



KATHOLIEKE UNIVERSITEIT LEUVEN

- **exercise**