Bioinformatik

Stefanie Hartmann

Wintersemester 2019 / 2020, Universität Potsdam

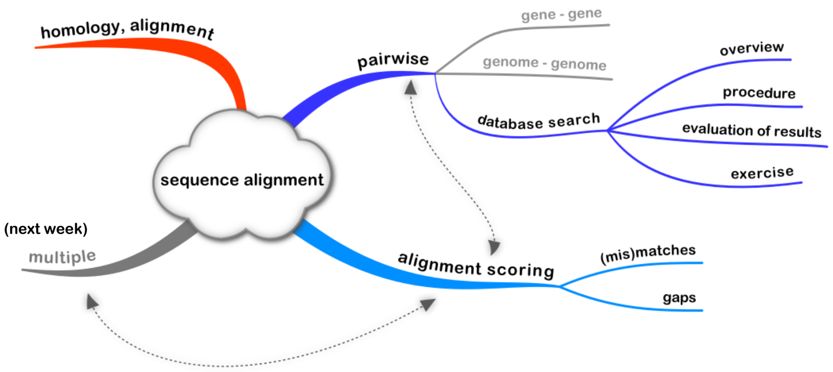
Database searches
Nov 15, 2019

heute!

10h: 3 Computer Pools (1a, 1b, 2a)

11: 2 Computer Pools (1a, 2a)

Today's topics



Analysis & comparison of sequences

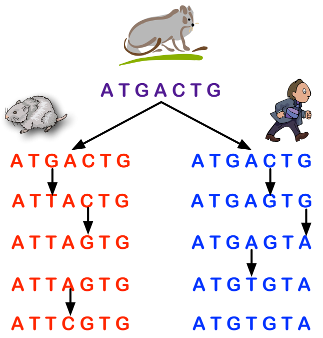
sequence data: abundant & relevant for many disciplines!

- database searching (pairwise alignments)
  - identify / collect homologous sequences or domains
- multiple sequence alignment
  - study conservation & variability of homologous sequences
- phylogenetic analysis
  - infer evolutionary history of homologous sequences
- genome alignment and assembly
- structure analysis & prediction
- …

Homology

homologs

- morphological structures, sequences, domains, ...
- derived from a common ancestor (and evolving!)
- all-or-nothing condition



similarity / identity

- a quantitative measure

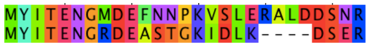
```
ATTCGTG
ATGTGTA
```

Sequence alignment

the comparison & relative arrangement of sequences by

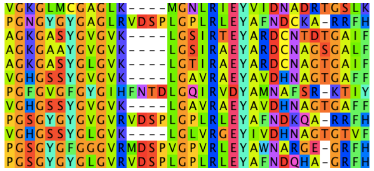
- searching for similarities between their characters (amino acids or nucleotides) and
- possibly inserting gaps in each sequence

pairwise sequence alignment



```
MYITENGMDEFNNPKVSLERALDDSNR
MYITENGRDEASTGKIDLK----DSER
```

multiple sequence alignment

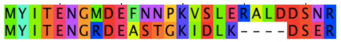


```
VGKGLMCGAGLK----MGNLRIEYVIDNADRTGSLK
PGNGYGYGAGLRVDSPLGPLRLEYAFNDCKA-RRFH
AGKGASYGVGVK----LGSIRTEYARDCNTDTGAIF
AGKGAAYGVGVK----LGSIRAEYARDCNAGSGALF
AGKGASYGLGVK----LGTIRAEYARDCNAGTGAIF
VGHGSSYGVGVK----LGAVRAEYAVDHNAGTGAFF
PGFGVGFGYGIHFNTDLGQIRVDYAMNAFSR-KTIY
VGHGSSYGVGVK----LGAVRAEYAVDHNAGTGAFF
PGSGYGYGVGVRVDSPLGPLRLEYAFNDKQA-RRFH
VGHGSSYGLGVK----LGLVRGEYIVDHNAGTGTVF
PGSGYGFGGGVRMDSPVGPVRLEYAWNARGE-GRFH
PGSGYGYGLGVRVDSPLGPLRLEYAFNDQHA-GRFH
```

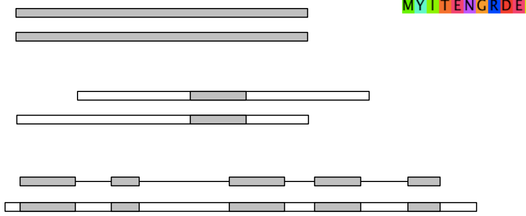
Pairwise alignment

two gene sequences

- global or local
- homologous sequences from different species
- homologous sequences from the same species
- genomic vs. cDNA sequence
- ...



```
MYITENGMDEFNNPKVSLERALDDSNR
MYITENGRDEASTGKIDLK----DSER
```



Pairwise alignment

two entire genomes

- global or local
- computation, display

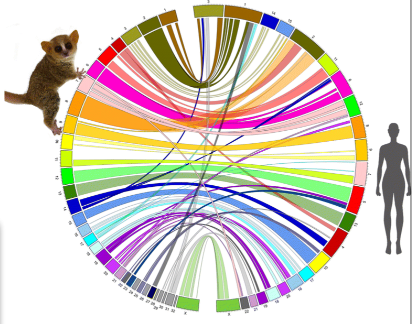
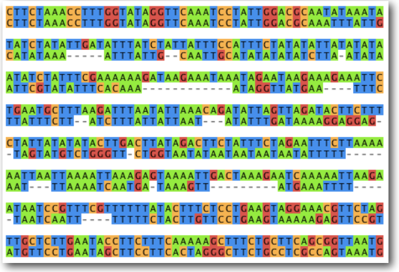


image taken from http://tinyurl.com/y9e6su4x



```
CTTCTAAACCTTTGGTATAGGTTCAAATCCTATTGGACGCAATATAAATA
CTTCTAAACCTTTGGTATAGGTTCAAATCCTATTGGACGCAAATTTATTG

TATCTATATTGATATTTATCTATTATTTCCATTTCTATATATTATATATA
CATATAAA------ATTTATTG--CAATTGCATATATATATCTTA-ATATA

ATATCTATTTCGAAAAAAGATAAGAAATAAATAGAATAAGAAAGAAATTC
ATTCGTATATTTCACAAA------------ATAGGTTATGAA----TTTC

TGAATGCTTTAAGATTTAATATTAAACAGATATTAGTTAGATACTTCTTT
TTATTTCTT--ATCTTTATTATTAAT---ATATTTGATAAAAGGAGGAG-

CTATTATATATACTTGACTTATAGACTTCTATTTCTAGAATTTCTTAAAA
-TAGTATGTCTGGGTT-CTGGTAATATAATAATAATAATATTTTT-----

AATTAATTAAAATTAAAGAGTAAAATTGACTAAAGAATCAAAAATTAAGA
AAT---TTAAAATCAATGA-TAAAGTT---------ATGAAATTTT----

ATAATCCGTTTCGTTTTTATACTTTCTCCTGAAGTAGGAAACGTTCTAG
-TAATCAATT----TTTTTCTACTTGTTCCTGAAGTAAAAAGAGTTCCGT

TTGCTCTTGAATACCTTCTTTCAAAAAGCTTTCTGCTTCAGCGGTTAATG
ATGTTCCTGAATAGCTTCCTTCACTAGGGCTTCTGCCTCGCCAGTAAATG
```

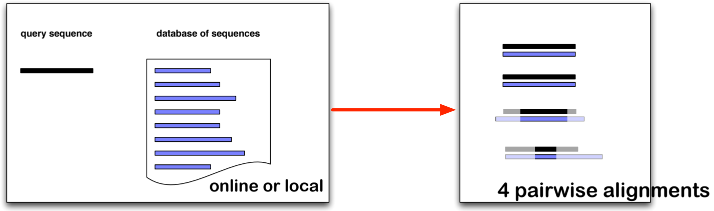
Database search

very frequent problem

- I have a starting (query) sequence
- I want to compare it to homologous sequences
- ...but which other sequences are homologous to it?

solution: database search!

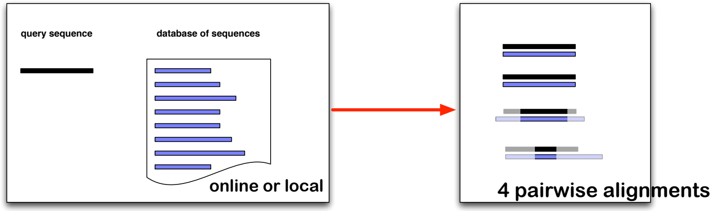
- compare query to a collection of sequences
- identify / collect homologs for further analysis



Database search using BLAST

BLAST {
- compute & score pairwise alignments between a starting (query) sequence and sequences in a database
- statistically evaluate the alignments, return the best alignments

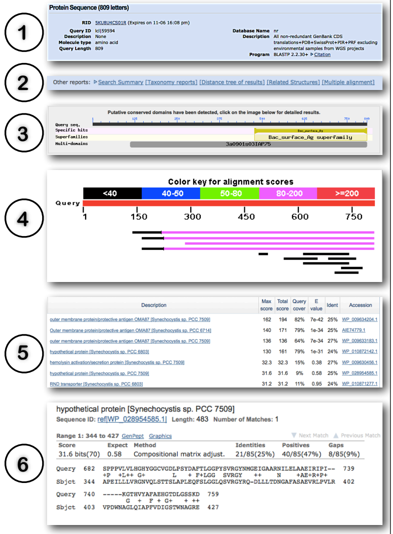
YOU →
- decide if the similarity is the result of homology (shared ancestry) or of chance



BLAST results

1. Header
2. Other reports:
   Search Summary
   Taxonomy reports
   Distance tree of results
   Multiple alignment
3. Conserved Domains
4. Graphic Summary
5. Summary table of alignments
6. Alignments

online only



Protein Sequence (809 letters)

RID 5KU8U4CS01R (Expires on 11-06 16:08 pm)
Query ID lcl|59594
Description None
Molecule type amino acid
Query Length 809
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.30+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]

Putative conserved domains have been detected, click on the image below for detailed results.

Description	Max score	Total score	Query cover	E value	Ident	Accession
outer membrane protein/protective antigen OMA87 [Synechocystis sp. PCC 7509]	162	194	82%	7e-42	25%	WP_009634204.1
Outer membrane protein/protective antigen OMA87 [Synechocystis sp. PCC 6714]	140	171	79%	1e-34	25%	AIE74779.1
outer membrane protein/protective antigen OMA87 [Synechocystis sp. PCC 7509]	136	136	64%	7e-34	27%	WP_009633183.1
hypothetical protein [Synechocystis sp. PCC 6803]	130	161	79%	1e-31	24%	WP_010872142.1
hemolysin activation/secretion protein [Synechocystis sp. PCC 7509]	32.3	32.3	15%	0.38	27%	WP_009630456.1
hypothetical protein [Synechocystis sp. PCC 7509]	31.6	31.6	9%	0.58	25%	WP_028954585.1
RND transporter [Synechocystis sp. PCC 6803]	31.2	31.2	11%	0.95	24%	WP_010871277.1

hypothetical protein [Synechocystis sp. PCC 7509]
Sequence ID: ref|WP_028954585.1| Length: 483 Number of Matches: 1

Range 1: 344 to 427 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
31.6 bits(70)	0.58	Compositional matrix adjust.	21/85(25%)	40/85(47%)	8/85(9%)

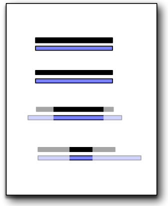
```
Query  682   SPPPVLVLHGHYGGCVGDLPSYDAFTLGGPYSVRGYNMGEIGAARNILELAAEIRIPI--   739
             +P  +L++ G+        L   + F+LGG  SVRGY    ++     N    +AE+R+P+
Sbjct  344   APEILLLVRGNVQLSTTSLAPLEQFSLGGLQSVRGYRQ-DLLLTDNGAFASAEVRLPVLR   402

Query  740   -----KGTHVYAFAEHGTDLGSSKD   759
                  G  +  F + G+   + ++
Sbjct  403   VPDWNAGLQIAPFVDIGSTWNAGRE   427
```

5

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

TOC75_ORYSJ RecName: Full=Protein TOC75, chloroplastic; AltName:...  1048   0.0
TC754_ARATH RecName: Full=Protein TOC75-4, chloroplastic; AltNam...   493   e-139
MATK_LEPPR RecName: Full=Maturase K; AltName: Full=Intron maturase;    35   0.14
MATK_MACCO RecName: Full=Maturase K; AltName: Full=Intron maturase;    33   0.46
IF3C_EUGGR RecName: Full=Translation initiation factor IF-3, chl...    33   0.54
PDI_MEDSA RecName: Full=Protein disulfide-isomerase; Short=PDI; ...    30   2.8
```

6

```
>TOC75_ORYSJ RecName: Full=Protein TOC75, chloroplastic; AltName: Full=75 kDa
          translocon at the outer-envelope-membrane of
          chloroplasts; Flags: Precursor;
         Length = 817

 Score = 1048 bits (2709), Expect = 0.0,   Method: Compositional matrix adjust.
 Identities = 487/695 (70%), Positives = 579/695 (83%), Gaps = 4/695 (0%)

Query: 119 FWSRILSPARAIADEPKSEDWDSHELPADITVLLGRLSGFKKYKISDILFFDRNKKSKVE 178
           FWSRI S   A ADE   S DWD H LPA+I V + +LSG K+YKIS++ FFDR
Sbjct: 123 FWSRIFSGGAAHADEKSSGDWDPHGLPANINVPMTKLSGLKRYKISELKFFDRAAGGGGA 182

Query: 179 ---TQDSFLDMVSLKPGGVYTKAQLQKELESLATCGMFEKVDMEGKTNADGSLGLTISFA 235
              +DSF +MV+L+PGGVYTK+QL KELE+L +CGMFE+VD+EGK   DG+LGLT+SF
Sbjct: 183 FTGPEDSFFEMVTLQPGGVYTKSQLLKELETLVSCGMFERVDLEGKAKPDGTLGLTVSFV 242

Query: 236 ESMWERADRFRCINVGLMGQSKPVEMDPDMSEKEKIEFFRRQEREYKRRISSARPCLLPT 295
           ES+W  A +F+CINVGLM QS  V+ D DM+E+EK+++ R+QER+Y++R+  A+PC+LP
Sbjct: 243 ESVWSAAKQFKCINVGLMSQSGQVDFDQDMTEREKMDYLRKQERDYQQRVRGAKPCILPD 302
Sbjct: 103 -HFVVVKEALLKTIKEVSEDKWSEELNTAWEIAYDGLASAI 142


>PDI_MEDSA RecName: Full=Protein disulfide-isomerase; Short=PDI; EC=5.3.4.1;
          Flags: Precursor;
         Length = 512

 Score = 30.4 bits (67), Expect = 2.8,   Method: Compositional matrix adjust.
 Identities = 16/37 (43%), Positives = 21/37 (56%), Gaps = 1/37 (2%)

Query: 738 PIKGTHVYAFAEHGTDLGSSKDVKGNPTV-VYRRMGQ 773
           P+    V A  EH  DL S  DVKG PT+ ++R  G+
Sbjct: 85  PVVLAKVDANEEHNKDLASENDVKGFPTIKIFRNGGK 121
```

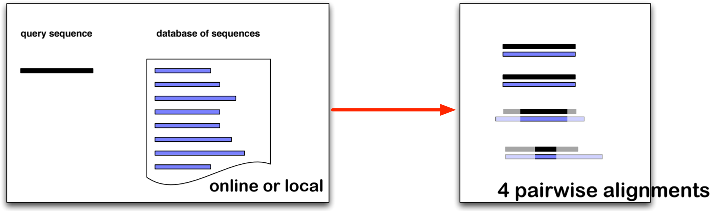


Database search using BLAST

considerations & challenges:

- score similarities?
- efficiently search the database?
  - query: 1 sequence, 500-1000 bp or aa
  - GenBank: 209,656,636 sequences, 279,668,290,132 bases
- statistically evaluate if the similarity could be due to chance?

```
MYITENGMDEFNNPKVSLERALDDSNR
||||||| ||    |  |     || |
MYITENGRDEASTGKIDLK----DSER
```



Scoring an alignment

- matches
- mismatches
- gaps

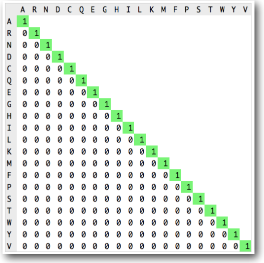
```
MYITENGMDEFNNPKVSLERALDDSNR
||||||| ||    |  |     || |
MYITENGRDEASTGKIDLK----DSER
```

- a scoring matrix is used for matches, mismatches,
- gap penalties are used for opening / extending gaps
- alignment score: sum of scores at each alignment position, $S = \sum s_{ij}$
- goal: identify the optimal alignment

Scoring an alignment

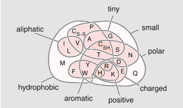
```
seq 1  MYITENGMDEYNNPKVSLERALDDSNR
       ||||||| ||    |  |     || |
seq 2  MYITENGRDEASTGKIDLK----DSER
```

score: 1+1+1+1+1+1+1+0+1+1+0+0+0+0+1+0+0+1+0+1+1+0+1 = 14

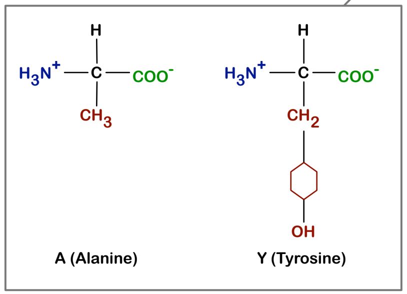


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1																			
R	0	1																		
N	0	0	1																	
D	0	0	0	1																
C	0	0	0	0	1															
Q	0	0	0	0	0	1														
E	0	0	0	0	0	0	1													
G	0	0	0	0	0	0	0	1												
H	0	0	0	0	0	0	0	0	1											
I	0	0	0	0	0	0	0	0	0	1										
L	0	0	0	0	0	0	0	0	0	0	1									
K	0	0	0	0	0	0	0	0	0	0	0	1								
M	0	0	0	0	0	0	0	0	0	0	0	0	1							
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

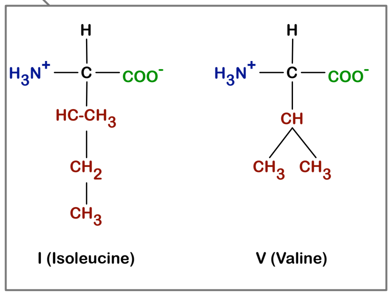
Scoring an alignment



```
seq 1  MYITENGMDEYNNPKVSLERALDDSNR
       ||||||| ||    |  |     || |
seq 2  MYITENGRDEASTGKIDLK----DSER
```



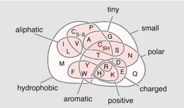
A (Alanine) Y (Tyrosine)

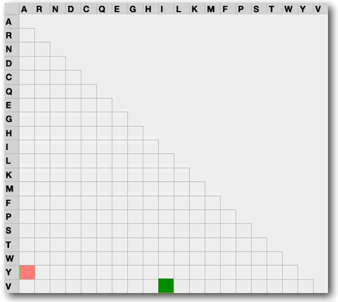


I (Isoleucine) V (Valine)

Scoring an alignment

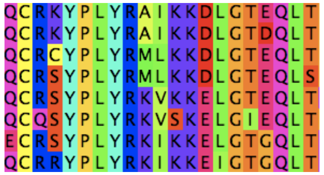
```
seq 1  MYITENGMDEYNNPKVSLERALDDSNR
       ||||||| ||    |  |     || |
seq 2  MYITENGRDEASTGKIDLK----DSER
```



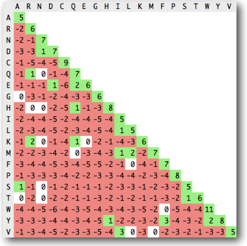


Scoring an alignment

very similar
homologous sequences

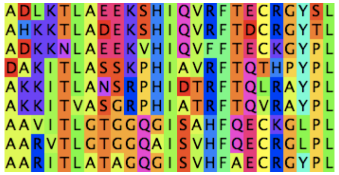


```
QCRKYPLYRAIKKDLGTEQLT
QCRKYPLYRAIKKDLGTDQLT
QCRCYPLYRMLKKDLGTEQLT
QCRSYPLYRMLKKDLGTEQLS
QCRSYPLYRKVKKELGTEQLT
QCQSYPLYRKVSKELGIEQLT
ECRSYPLYRKIKKELGTGQLT
QCRRYPLYRKIKKEIGTGQLT
```

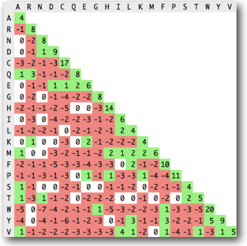


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-2	6																		
N	-2	-1	7																	
D	-3	-3	1	7																
C	-1	-5	-4	-5	9															
Q	-1	1	0	-1	-4	7														
E	-1	-1	-1	1	-6	2	6													
G	0	-3	-1	-2	-4	-3	-3	6												
H	-2	0	0	-2	-5	1	-1	-3	8		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

more divergent
homologous sequences



```
ADLKTLAEEKSHIQVRFTECRGYSL
AHKKTLADEKSHIQVRFTDCRGYTL
ADKKNLAEEKVHIQVFFTECKGYPL
DAKITLASSKPHIAVRFTQTHPYPL
AKKITLANSRPHIDTRFTQLRAYPL
AKKITVASGRPHIATRFTQVRAYPL
AAVITLGTGGQGISAHFQECKGLPL
AARVTLGTGGQAISVHFQECRGLPL
AARITLATAGQGISVHFAECRGYPL
```



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	8																		
N	0	-2	8																	
D	0	-1	1	9																
C	-3	-2	-1	-3	17															
Q	1	3	-1	-1	-2	8														
E	0	-1	-1	1	1	2	6													
G	0	-2	0	-1	-4	-2	-2	8												
H	-2	-1	-1	-2	-5	0	0	-3	14											
I	0	-3	0	-4	-2	-2	-3	-1	-2	6										
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4									
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4								
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6							
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10						
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11					
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4				
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5			
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20		
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5

BLOSUM matrices

- BLOcks amino acid SUbstitution Matrices
- based on local alignments of divergent sequences
- different BLOSUM matrices are based on observed alignments with different degrees of similarity
  - BLOSUM60 matrix:

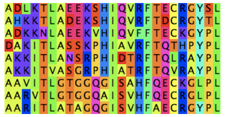
    derived from and best used for sequences that are 60% identical

	for proteins of
BLOSUM45	45% identity
BLOSUM60	60%
BLOSUM90	90%

Substitution matrices

➡ which amino acids occur together in the alignment columns more often than expected by chance?

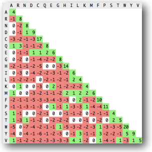
trusted alignment of
homologous sequences



```
ADLKTLAEEKSHIQVRFTECRGYSL
AHKKTLADEKSHIQVRFTDCRGYTL
ADKKNLAEEKVHIQVFFTECKGYPL
DAKITLASSKPHIAVRFTQTHPYPL
AKKITLANSRPHIDTRFTQLRAYPL
AKKITVASGRPHIATRFTQVRAYPL
AAVITLGTGGQGISAHFQECKGLPL
AARVTLGTGGQAISVHFQECRGLPL
AARITLATAGQGISVHFAECRGYPL
```

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

$$\longrightarrow$$



p_{ab} : observed frequency of residues a and b aligned

q_a, q_b : frequencies of residues a and b

M:	0.01
L:	0.1
ML:	0.002

$$s(M,L) = log(\frac{0.002}{0.01{*}0.1}) = +1$$

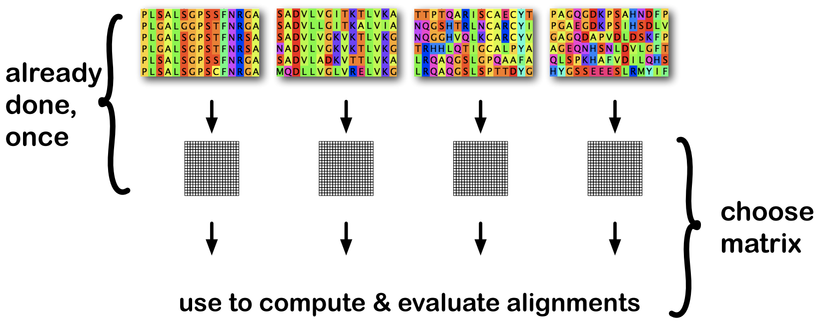
BLOSUM62

gaps: -2 each

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

```
MYITENGMDEFNNPKVSLERALDDSNR         MYITENGMDEFNNPKVSLERALDDSNR
                              →     ||||||| ||    |  |     || |    →   60
MYITENGRDEASTGKIDLKDSER             MYITENGRDEASTGKIDLK----DSER
```

(BLOSUM) matrices



already done, once

choose matrix

use to compute & evaluate alignments

Scoring matrices

identity matrices

- easy to understand and implement
- biologically not very realistic

matrices based on physico-chemical aa properties

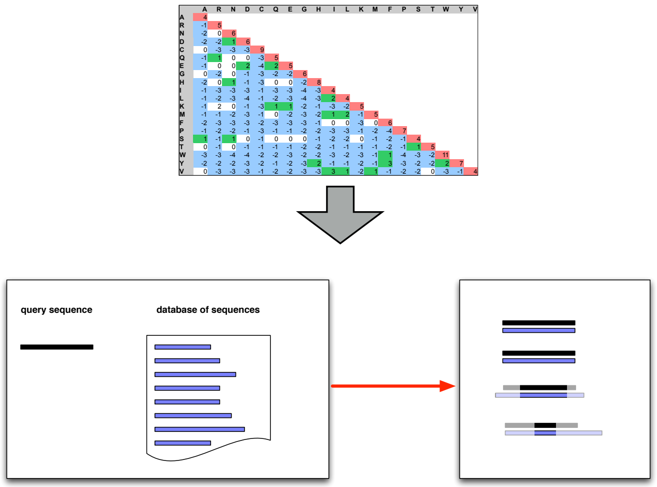
- better than identity scores
- different divergence times - same scores!

BLOSUM matrices

- evolutionary perspective:
  sequences accumulate changes over time

(many other matrices exist, some are very specific)

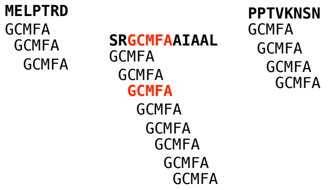
Finding & evaluating sequence matches



Pairwise alignment approaches

exhaustive search

- compare every position of the query with every position of every database sequence
- not practical



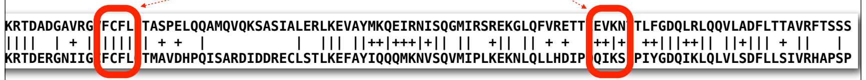
```
MELPTRD        SRGCMFAAIAAL        PPTVKNSN
GCMFA          GCMFA               GCMFA
 GCMFA          GCMFA               GCMFA
  GCMFA          GCMFA               GCMFA
                  GCMFA               GCMFA
                   GCMFA
                    GCMFA
                     GCMFA
                      GCMFA
```

heuristics

- most popular approach: BLAST
- exclude unpromising regions from the search

Main steps of BLAST

- pre-processing of the database to be searched
- seeding
  - homologous sequences contain (at least) short stretches of identical or high-scoring amino acids

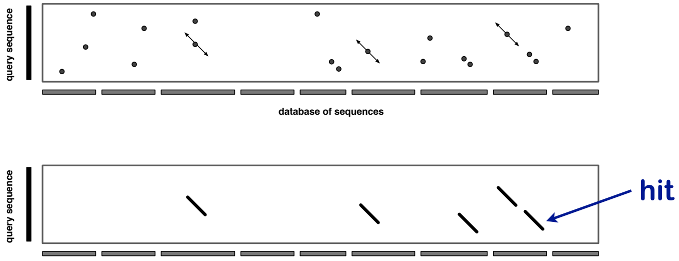


```
KRTDADGAVRG FCFL TASPELQQAMQVQKSASIALERLKEVAYMKQEIRNISQGMIRSREKGLQFVRETT EVKN TLFGDQLRLQQVLADFLTTAVRFTSSS
||||  | + | |||| | + +  |           |  ||| ||++|+++|+|| ||  +|| || + +   ++|+  ++|||++|| ||+||| + ||   |
KRTDERGNIIG FCFL TMAVDHPQISARDIDDRECLSTLKEFAYIQQQMKNVSQVMIPLKEKNLQLLHDIP QIKS PIYGDQIKLQLVLSDFLLSIVRHAPSP
```

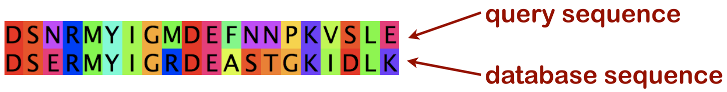
  - these will be identified and used as seeds
- extension to a good longer alignment
- evaluation of statistical significance
- ranking & presentation of alignments

Seeding & extension

- the locations of short, high scoring hits are identified
- these are used as alignment seeds
- seeds are extended into longer alignments (→“hits”)
- for each hit, an “E(xpect)-value” is computed
- hits are ranked by their E-values and reported



Evaluating alignments



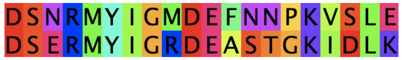
DSNRMYIGMDEFNNPKVSLE

DSERMYIGRDEASTGKIDLK

alignment raw score: $S = \sum s_{ij}$

- any two sequences (even if they are unrelated!) will have a “best” alignment score
- how high of a score can we expect from random (unrelated) sequences? is our current score better?
- statistics!
  - substitutions, aa distribution, seq lengths, etc
  - based on Gumbel extreme value distribution

Evaluating alignments



DSNRMYIGMDEFNNPKVSLE ← real query sequence

DSERMYIGRDEASTGKIDLK ← real database sequence

alignment raw score: $S = \sum s_{ij}$

$$\downarrow \quad E = Kmne^{-\lambda S}$$

$$\searrow \quad S' = \frac{(\lambda \times S) - ln(K)}{ln(2)}$$

$$\downarrow \quad E = mn(2^{-S'})$$

E: the number of alignments at least as good that are expected by chance

(under comparable conditions: aa compositon, db & query length, etc)

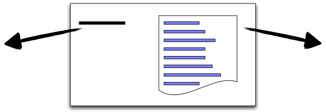
0 0.000000003 0.002 5 20

$3{\times}10^{-9}$

3e-09

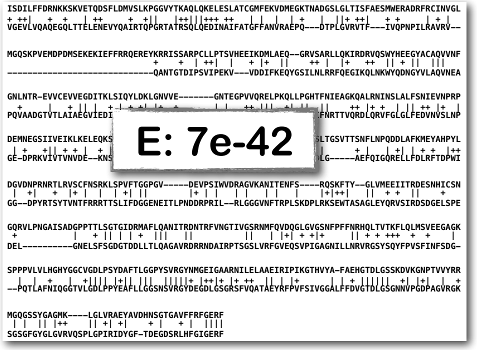
Evaluating alignments

query sequence:
one plant gene
sequence of 809 aa



database to be searched:
51,826,119 bacterial
sequences with a total of
18,610,688,330 aa

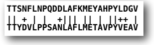
db seq: 763 aa; aln: 673 (287/42% | or +)



```
ISDILFFDRNKKSKVETQDSFLDMVSLKPGGVYTKAQLQKELESLATCGMFEKVDMEGKTNADGSLGLTISFAESMWERADRFRCINVGL
+ ++|    +   +    | ++       +   +||    |++|||+++ ++     |  |   |   | +    |   ||+ ++|     + +      | + |
VGEVLVQAQEGQLTTELENEVYQAIRTQPGRTATRSQLQEDINAIFATGFFANVRAEPQ---DTPLGVRVTF---IVQPNPILRAVRV--

MGQSKPVEMDPDMSEKEKIEFFRRQEREYKRRISSARPCLLPTSVHEEIKDMLAEQ--GRVSARLLQKIRDRVQSWYHEEGYACAQVVNF
                            +    +  | ++|  |    +  |+  ||     ++ |  |+   ++  || + ||  |||
----------------------------QANTGTDIPSVIPEKV---VDDIFKEQYGSILNLRRFQEGIKQLNKWYQDNGYVLAQVNEA

GNLNTR-EVVCEVVEGDITKLSIQYLDKLGNVVE-------GNTEGPVVQRELPKQLLPGHTFNIEAGKQALRNINSLALFSNIEVNPRP
  +    |   + || |   +                    | |   ++ |||   +|   ||     ++ |+ +  | || ++ |+  |
PQVAADGTVTLAIAEGVIEDI                                        FNRTTVQRDLQRVFGLGLFEDVNVSLNP

DEMNEGSIIVEIKLKELEQKS                                        SLTGSVTTSNFLNPQDDLAFKMEYAHPYL
 | +   +|| + + |  +                                               + |    | +| | + +  |++
GE-DPRKVIVTVNVDE--KNS                                        LG-----AEFQIGQRELLFDLRFTDPWI

DGVDNPRNRTLRVSCFNSRKLSPVFTGGPGV-----DEVPSIWVDRAGVKANITENFS----RQSKFTY--GLVMEEIITRDESNHICSN
 |  +|   +  |+ |  |  | +| ||        |+ | |   | |   | |   |    |+|++|    ||  + +  ||    +
GG--DPYRTSYTVNTFRRRTTSLIFDGGENEITLPNDDRPRIL--RLGGGVNFTRPLSKDPLRKSEWTASAGLEYQRVSIRDSDGELSPE

GQRVLPNGAISADGPPTTLSGTGIDRMAFLQANITRDNTRFVNGTIVGSRNMFQVDQGLGVGSNFPFFNRHQLTVTKFLQLMSVEEGAGK
 +           |   + || | | +  |||   ||           ||   | |+| + +|+     || + + +++  +  +    |
DEL----------GNELSFSGDGTDDLLTLQAGAVRDRRNDAIRPTSGSLVRFGVEQSVPIGAGNILLNRVRGSYSQYFPVSFINFSDG-

SPPPVLVLHGHYGGCVGDLPSYDAFTLGGPYSVRGYNMGEIGAARNILELAAEIRIPIKGTHVYA-FAEHGTDLGSSKDVKGNPTVVYRR
  |  |  +   |  +|||| |+|| |||  |||||+ |++|+ |+ ++  || | |+        | | + ||||||  +| |+|  |  +
--PQTLAFNIQGGTVLGDLPPYEAFLLGGSNSVRGYDEGDLGSGRSFVQATAEYRFPVFSIVGGALFFDVGTDLGSGNNVPGDPAGVRGK

MGQGSSYGAGMK----LGLVRAEYAVDHNSGTGAVFFRFGERF
 | |  || |++    || +| +|     + |   + |  ||||
SGSGFGYGLGVRVQSPLGPIRIDYGF-TDEGDSRLHFGIGERF
```

E: 7e-42

db seq: 409 aa;
aln: 26 (17/65% | or +)

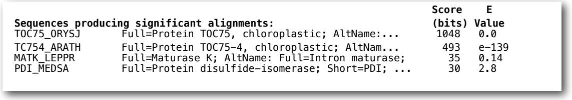


```
TTSNFLNPQDDLAFKMEYAHPYLDGV
|| +  | |  +|||| ||  |  ||++ |
TTYDVLPPSANLAFLMETAVPYVEAV
```

E: 9

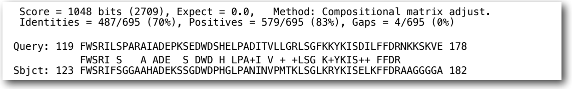
BLAST results

- pairwise alignments query–database sequence
  - overview table, ranked by E-value



```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value
TOC75_ORYSJ      Full=Protein TOC75, chloroplastic; AltName:...    1048   0.0
TC754_ARATH      Full=Protein TOC75-4, chloroplastic; AltNam...     493   e-139
MATK_LEPPR       Full=Maturase K; AltName: Full=Intron maturase;     35   0.14
PDI_MEDSA        Full=Protein disulfide-isomerase; Short=PDI; ...    30   2.8
```

  - alignments, ranked by E-value



```
 Score = 1048 bits (2709), Expect = 0.0,   Method: Compositional matrix adjust.
 Identities = 487/695 (70%), Positives = 579/695 (83%), Gaps = 4/695 (0%)

Query: 119 FWSRILSPARAIADEPKSEDWDSHELPADITVLLGRLSGFKKYKISDILFFDRNKKSKVE 178
           FWSRI S   A ADE  S DWD H LPA+I V + +LSG K+YKIS++ FFDR
Sbjct: 123 FWSRIFSGGAAHADEKSSGDWDPHGLPANINVPMTKLSGLKRYKISELKFFDRAAGGGGA 182
```

- information about the query and the database
- information about statistical parameters
- different presentation online and in the terminal
- additional information & analysis options online

BLAST search: results

		online	terminal
Header		✗	✗
Other reports:	Search Summary	✗*	
	Taxonomy reports	✗	
	Distance tree of results	✗	
	Multiple alignment	✗	
Graphic Summary Conserved Domains Graphical Overview of Hits		✗	
Descriptions		✗	✗
Alignments		✗	✗
Footer			✗*

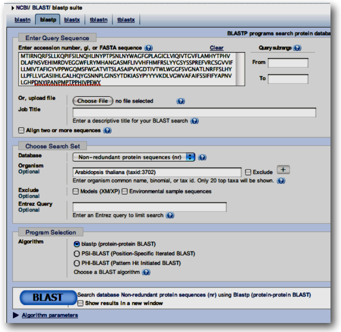
✗* equivalent information

BLAST programs

query	database	program
nucleotide	nucleotide	blastn
nucleotide (translated)	nucleotide (translated)	tblastx
nucleotide (translated)	peptide	blastx
peptide	peptide	blastp
peptide	nucleotide (translated)	tblastn

BLAST search in practice

- call the appropriate BLAST program
- select parameters for search
- select database to be searched
- select and submit query sequence



BLASTP programs search protein databa

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Clear

```
MTIRNQRFSLLKQPIFSILNQHLINYPTPSNLNYWAGFGPLAGICLVIQIVTGVFLAMHYTPHV
DLAFNSVEHIMRDVEGGWFLRYMHANGASMFLIVVHFHMFRSLYYGSYSSPREFVRCSGVVIF
LLMIVTAFIGYVPPWGQMSFWGATVITSLASAIPVVGDTIVTWLWGGFSVGNATLNRFFSLHY
LLPFLLVGASIIHLGALHQYGSNNPLGINSYTDKIASYPYYYVKDLVGWVAFAIFSSIFIFYAPNV
LGHPDNYIPANPMPTPPHIVPEWY
```

Query subrange: From ___ To ___

Or, upload file: Choose File no file selected

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Arabidopsis thaliana (taxid:3702) □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Environmental sample sequences

Entrez Query Optional: Enter an Entrez query to limit search

Program Selection

Algorithm:
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

□ Show results in a new window

▶ Algorithm parameters

```
blastp
   -db database
   -query querySeq
   -b xxx
   -e yyy
   ...
```

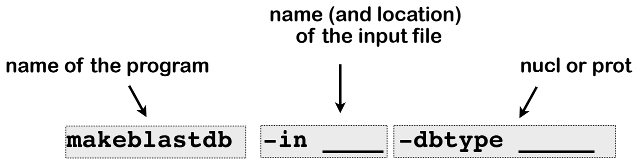
Today's exercise

query: human beta globin

database: *Medicago truncatula* genes

1. download the human beta globin in fasta format from NCBI
   - one file each: protein sequence, DNA sequence
2. download medicago DNA and protein sequences from Moodle
   - one file each: protein sequences, DNA sequences
3. convert the medicago sequences into a BLAST database
4. search if the human globin is similar to a medicago sequence
   - protein-protein (blastp)
   - DNA-DNA (blastn)

convert a fasta file into a database



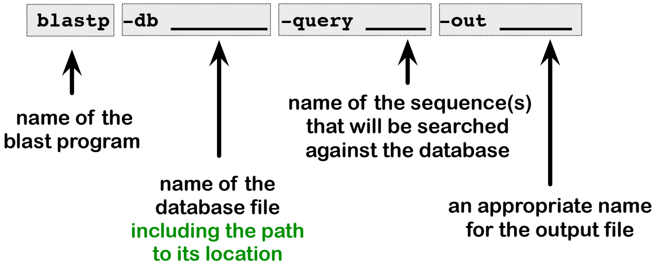
name of the program

name (and location) of the input file

nucl or prot

```
makeblastdb -in _____ -dbtype ______
```

command-line BLAST



```
blastp -db ________ -query _____ -out ______
```

name of the blast program

name of the database file including the path to its location

name of the sequence(s) that will be searched against the database

an appropriate name for the output file

heute!

10h: 3 Computer Pools (1a, 1b, 2a)

11: 2 Computer Pools (1a, 2a)

Key terms and concepts

- homology, similarity
- alignment
  - local, global
- BLOSUM substitution matrices:
  - what are they, how were they derived, how are they used in sequence alignment
- heuristics
- database searching / BLAST
  - principle, steps
  - evaluation of results
  - definition and interpretation of “E-value”