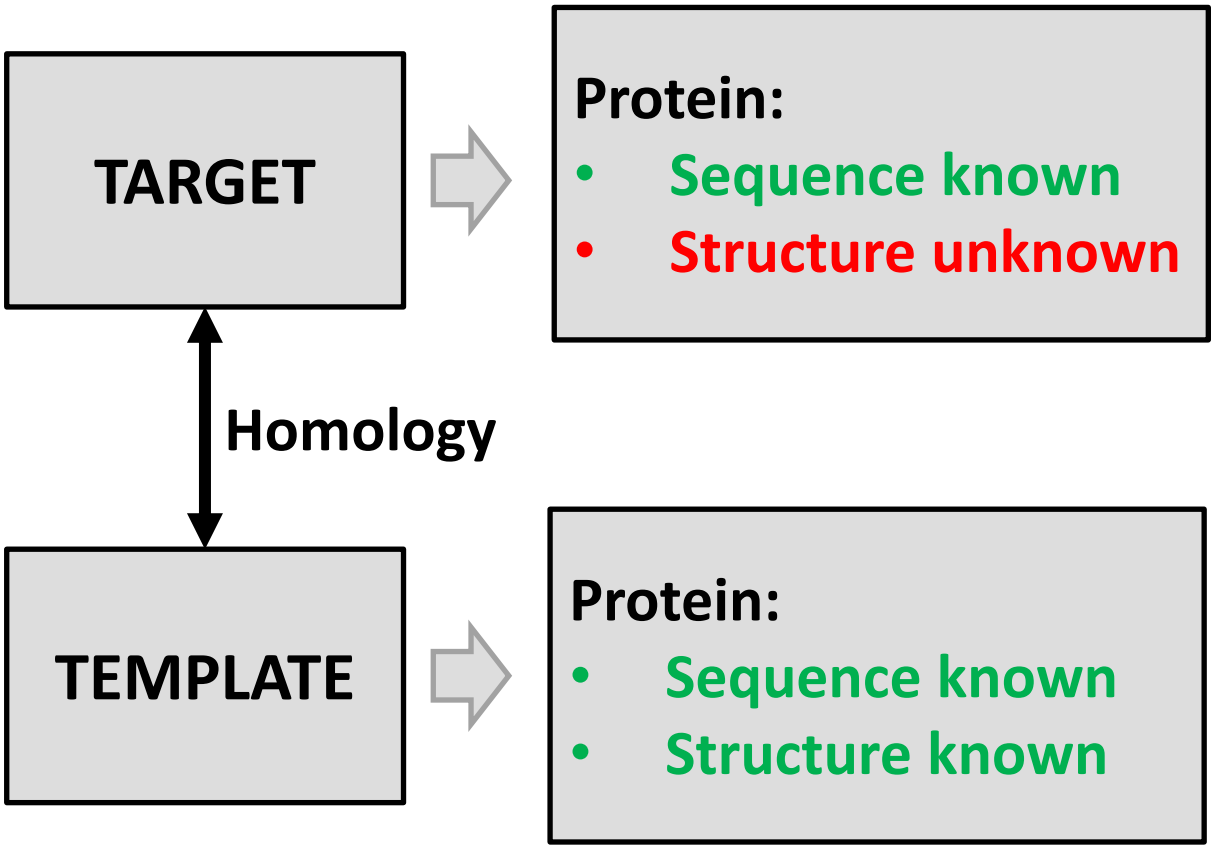
Structural biology

Practice 1: BLAST

Course 2023-2024

Target and template



Target and template

Using homology modeling we can use the structure of the template to build structural models of the target

Target and template

If two proteins are homologs they will have similar sequences

How can we know if two protein sequences are similar?

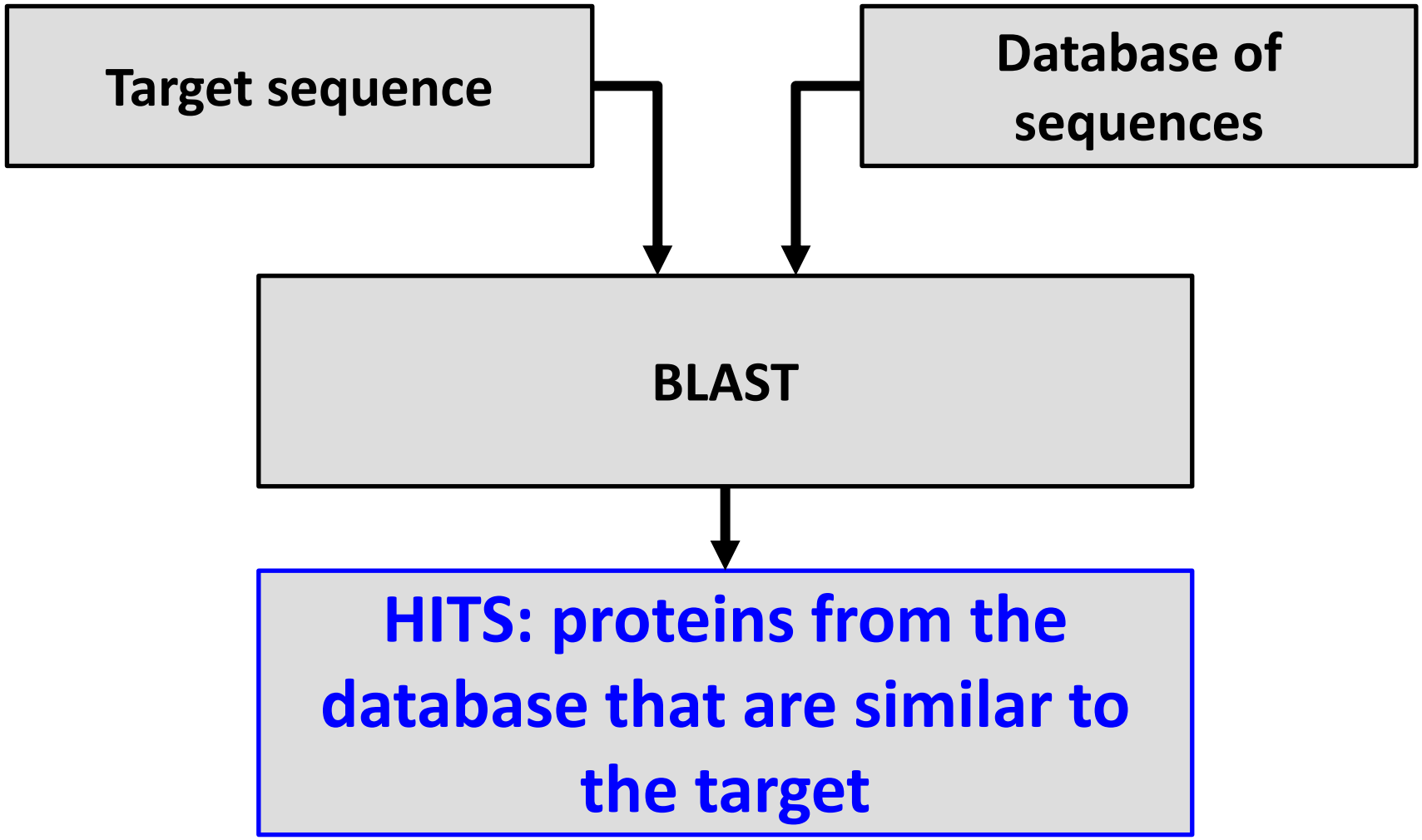
Target and template

If two proteins are homologs they will have similar sequences

How can we know if two protein sequences are similar?

Using sequence alignments

How BLAST works?



How BLAST works?

For each protein in the input database BLAST makes a local alignment

TG: AGVHK

Pn: AAVHR

How does BLAST to know what alignments are good or bad?

It uses <u>substitution matrices</u>

Substitution matrices

Substitution matrices contain scores associated to the frequency of substitution between different amino acids

Unlikely substitutions

↓

Low scores

Likely substitutions

↓

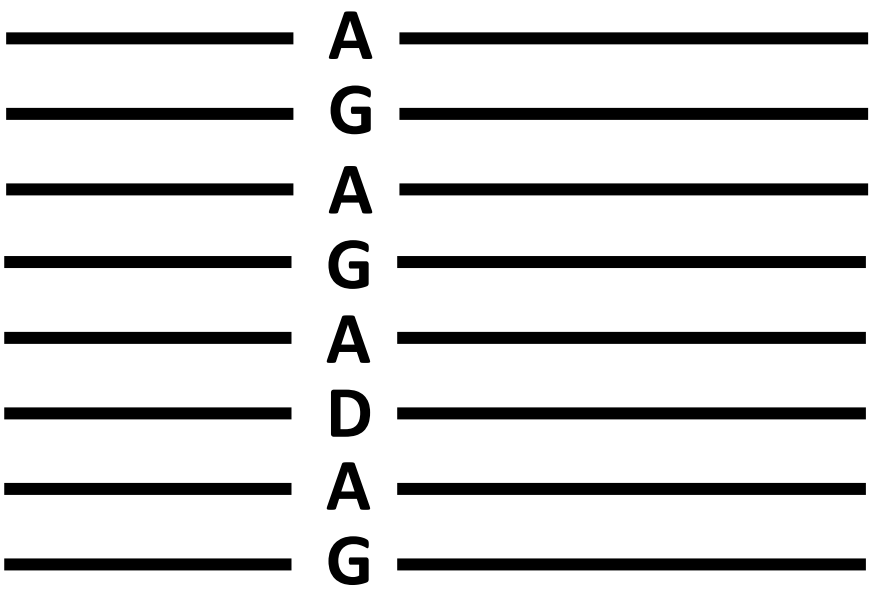
High scores

Substitution matrices

<u>Substitution matrices</u> contain scores associated to the frequency of substitution between different amino acids

Substitution matrices

Substitution matrices contain scores associated to the frequency of substitution between different amino acids



A-G substitution is likely → High scores

A-D and G-D substitutions are unlikely → Low scores

Substitution matrices

	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	

Substitution matrices are obtained from Multiple Sequence Alignments (MSAs)

Substitution matrices

	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	

TG: AGVHK

4 2

Pn: AAVHR

Substitution matrices

	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	

TG: AGVHK

4 2

Pn: AAVHR

K and R are both basic amino acids, they have similar chemical properties

Substitution matrices

	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	

This is the BLOSUM62 substitution matrix (used by BLAST by default)

Substitution matrices

	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	

Values on the diagonal correspond with amino acid conservation

BLAST outputs

After executing BLAST we obtain a list of hits

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

  1b0b_A mol:protein length:142 HEMOGLOBIN                          45.4    2e-06
  1ebt_A mol:protein length:142 HEMOGLOBIN                          45.1    3e-06
  1moh_A mol:protein length:142 MONOMERIC HEMOGLOBIN I              43.9    7e-06
  1flp_A mol:protein length:142 HEMOGLOBIN I (AQUO MET)             43.9    7e-06
  2olp_B mol:protein length:152 Hemoglobin II                       41.2    8e-05
  2olp_A mol:protein length:152 Hemoglobin II                       41.2    8e-05
  1eco_A mol:protein length:136 ERYTHROCRUORIN (CARBONMONOXY)       30.0    0.71
  1ecn_A mol:protein length:136 ERYTHROCRUORIN (CYANO MET)          30.0    0.71
  1ecd_A mol:protein length:136 ERYTHROCRUORIN (AQUO MET)           30.0    0.71
  1eca_A mol:protein length:136 ERYTHROCRUORIN (AQUO MET)           30.0    0.71
  2iyo_A mol:protein length:472 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_C mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_B mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_A mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
```

BLAST outputs

How do we know how good these hits are?

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

  1b0b_A mol:protein length:142 HEMOGLOBIN                          45.4    2e-06
  1ebt_A mol:protein length:142 HEMOGLOBIN                          45.1    3e-06
  1moh_A mol:protein length:142 MONOMERIC HEMOGLOBIN I              43.9    7e-06
  1flp_A mol:protein length:142 HEMOGLOBIN I (AQUO MET)             43.9    7e-06
  2olp_B mol:protein length:152 Hemoglobin II                       41.2    8e-05
  2olp_A mol:protein length:152 Hemoglobin II                       41.2    8e-05
  1eco_A mol:protein length:136 ERYTHROCRUORIN (CARBONMONOXY)       30.0    0.71
  1ecn_A mol:protein length:136 ERYTHROCRUORIN (CYANO MET)          30.0    0.71
  1ecd_A mol:protein length:136 ERYTHROCRUORIN (AQUO MET)           30.0    0.71
  1eca_A mol:protein length:136 ERYTHROCRUORIN (AQUO MET)           30.0    0.71
  2iyo_A mol:protein length:472 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_C mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_B mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
  2iyp_A mol:protein length:473 6-PHOSPHOGLUCONATE DEHYDROGENASE,...  27.7    7.6
```

BLAST outputs

How do we know how good these hits are?

BLAST provides two measurements:

- **Score (substitution matrix)**
- **E-value (how likely is to find such hit by chance)**

BLAST outputs

To calculate the E-value BLAST uses the next formula

$$E(S) \approx \exp\left(-NmnKe^{-(S-\mu)}\right)$$

Using BLAST

In what database can I search for proteins with available structure?

Using BLAST

In what database can I search for proteins with available structure?

In the PDB
(Protein Data Bank)

Using BLAST

Executing BLAST with our target sequence in the PDB

Step 1: Using BLAST

Within "exercise_2" you will find the subdirectory BLAST. Within this there is the sequence problem named "target.fa".

To look for proteins of known structure similar to the target protein, try:

```
blastp  -query  target.fa  -db  /mnt/NFS_UPF/soft/databases/blastdat/pdb_seq
-out  target_pdb.out
```

Example of BLAST usage:
➢ **blastp -query [target_fasta_format]-db [database]-out [output]**

You can see the result of the search in the output file target_pdb.out.

Substitution matrices

What is the problem pf using a BLOSUM62 substitution matrix?

- Is not especific for the substitutions that happen in the protein family of our target
- Different regions of the protein may have different substitution frequencies

Substitution matrices

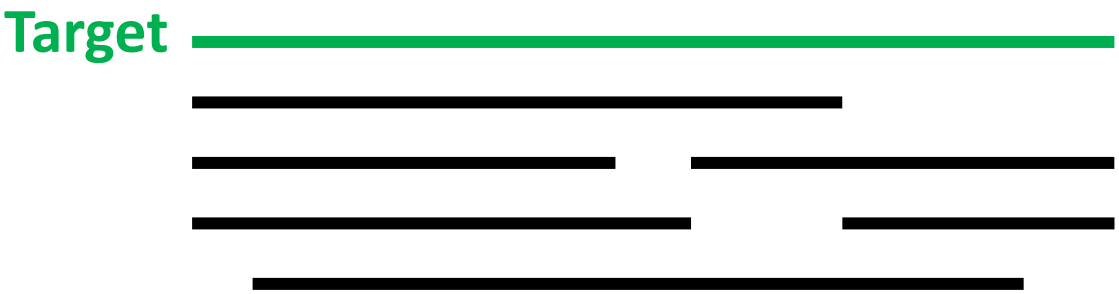
What is the problem pf using a BLOS tion

You can solve these problems using a position specific substitution matrix (PSSM)

- P
- D erent
su

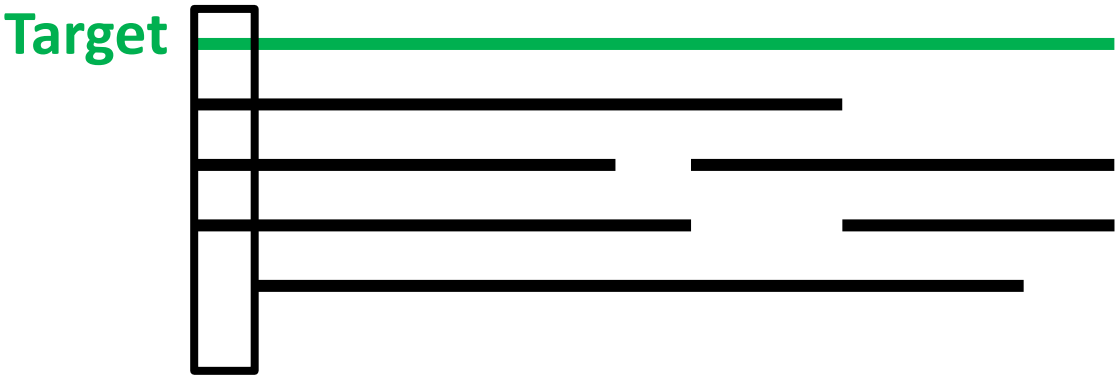
Position specific substitution matrices (PSSM)

PSSM are obtained from a MSA



Position specific substitution matrices (PSSM)

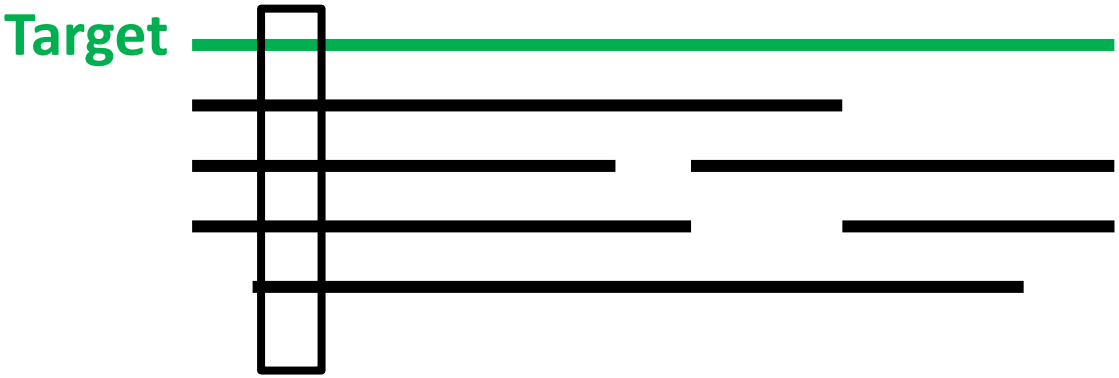
PSSM are obtained from a MSA



Position 1
Substitution
scores

Position specific substitution matrices (PSSM)

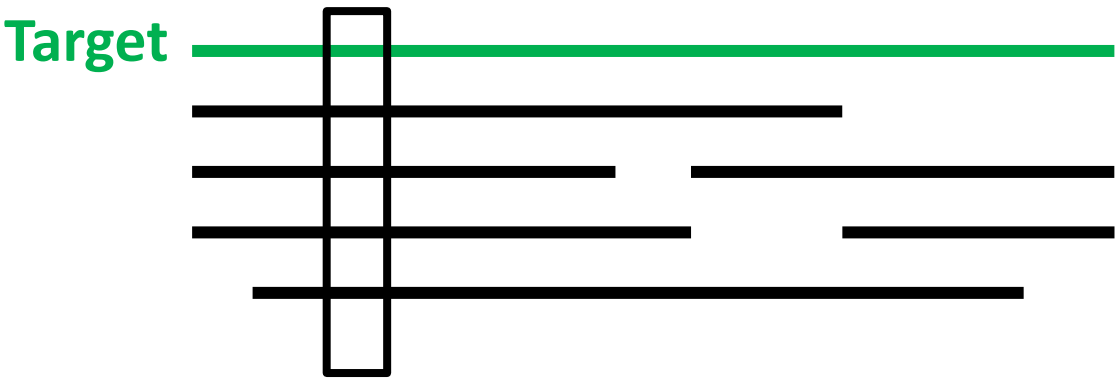
PSSM are obtained from a MSA



Position 2
Substitution
scores

Position specific substitution matrices (PSSM)

PSSM are obtained from a MSA



Position 3 Substitution scores

And so on and so forth…

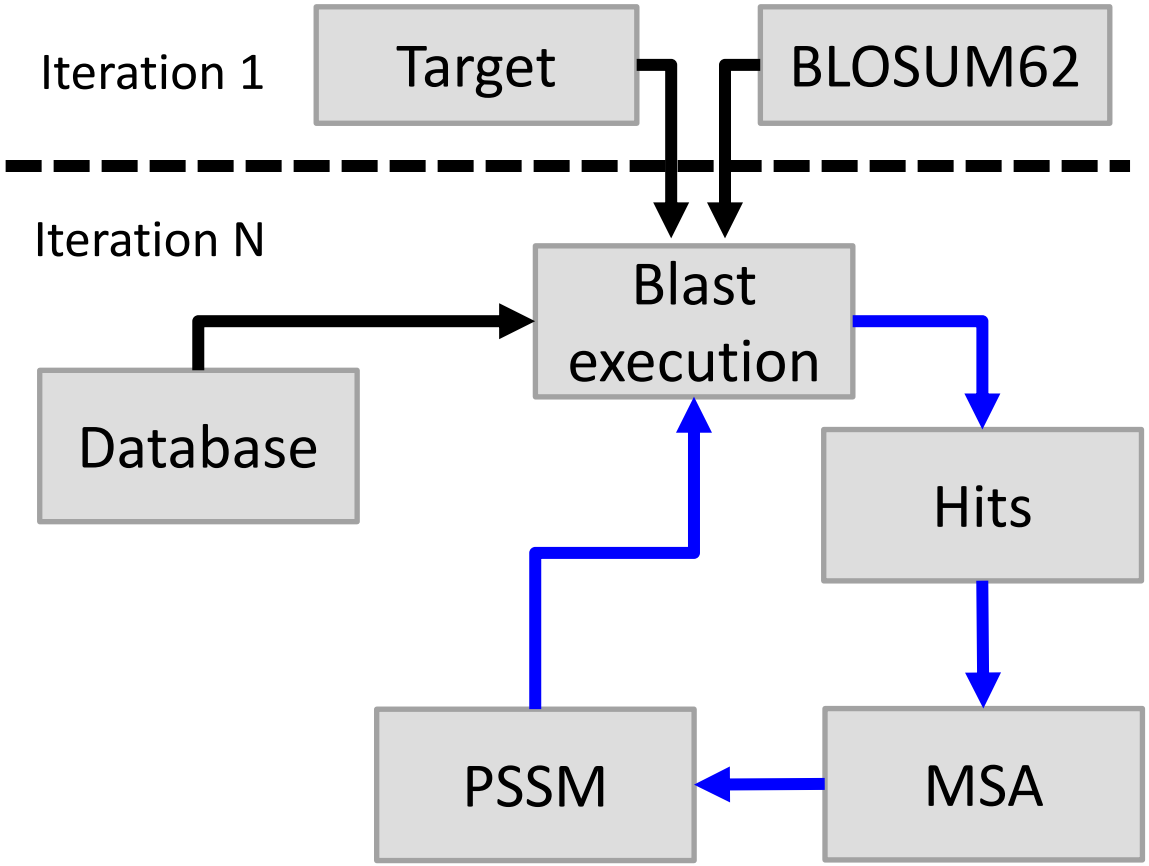
PSI-BLAST

How can we use PSSMs from the family of our target to search for templates?

Using PSI-BLAST

PSI-BLAST

PSI-BLAST creates a new PSSM at each iteration



PSI-BLAST

Executing PSI-BLAST with our target sequence in the PDB with 5 iterations

Example of PSI-BLAST usage:

➢ **psiblast -query [target_fasta_format] -db [database] -num_iterations [number of iterations] -out [output]**

Using the previous example, we can make an iterative search into the PDB database. We are going to make 5 iterations:

```
psiblast -query target.fa -db /mnt/NFS_UPF/soft/databases/blastdat/pdb_seq
-num_iterations 5 -out target_pdb_5.out
```

Databases of sequences

How can we improve our sequence search?

Using a non-redundant and non-biased database to create the PSSMs

Databases of sequences

The PDB is very redundant

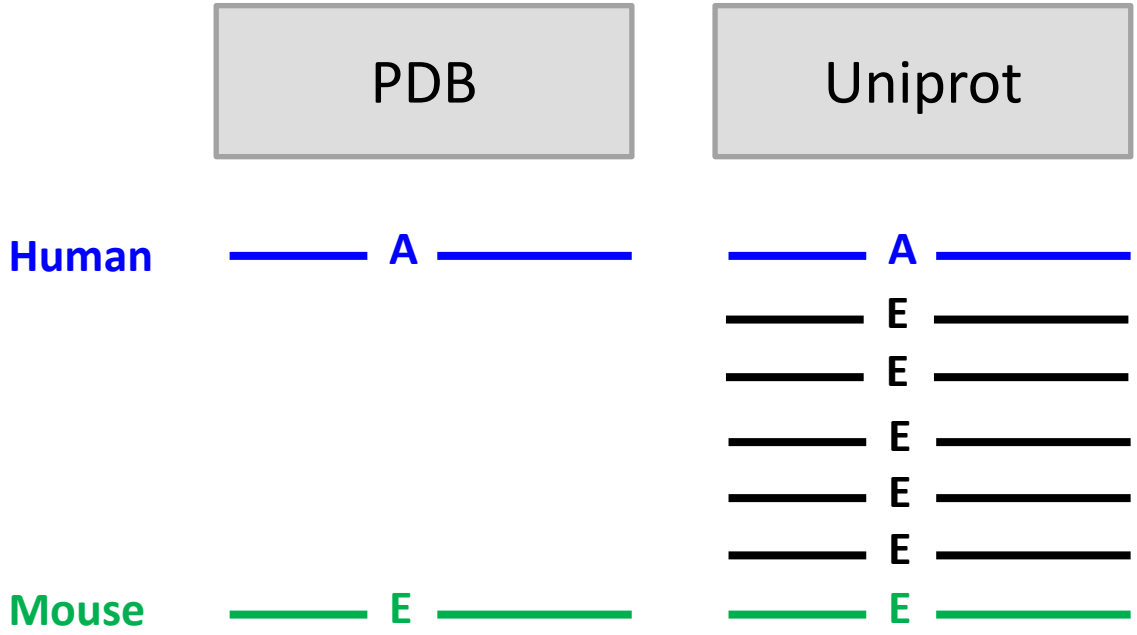
The same proteins are repeated several times

The PDB is very biased

Some protein families are overrepresented, others are not represented at all. Happens the same with species.

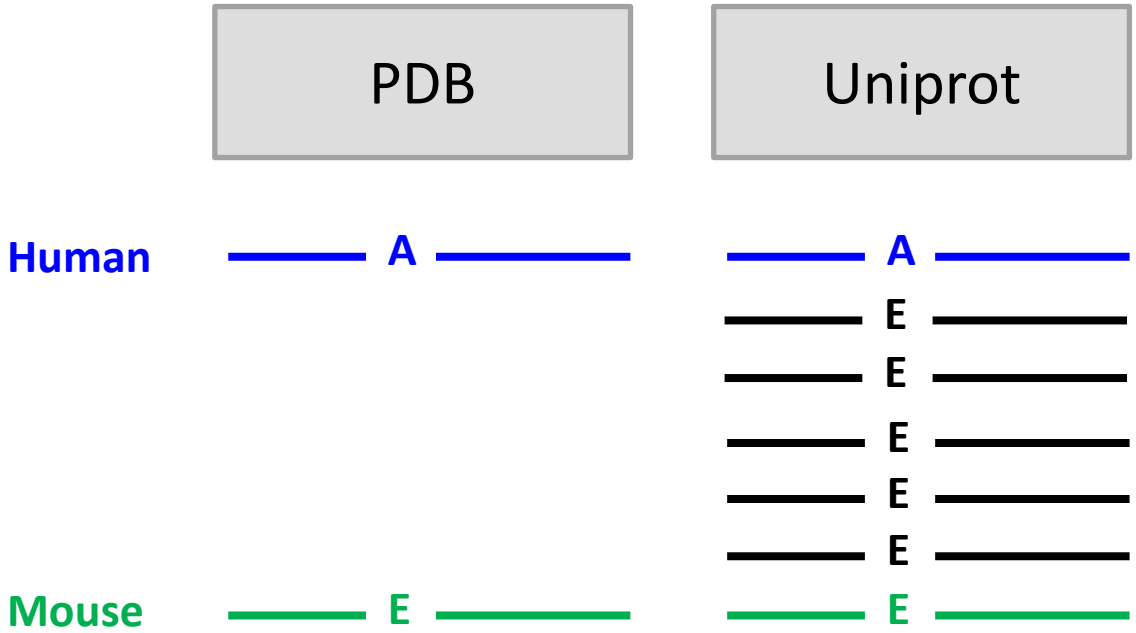
Databases of sequences

If we create our PSSM with a biased database our PSSM will be biased too



Databases of sequences

If we create our PSSM with a biased database our PSSM will be biased too



Using a non-biased and non-redundant database we can represent better the evolutionary relations between sequences in the same family

Databases of sequences and PSI-BLAST

Step1: Use Uniprot database to create an accurate PSSM

```
psiblast -query target.fa -num_iterations 5
-out_pssm target_sprot5.pssm -out target_sprot_5.out
-db /mnt/NFS_UPF/soft/databases/blastdat/uniprot_sprot.fasta
```

Step 2: Use this accurate PSSM to search for templates in the PDB

```
psiblast -db /mnt/NFS_UPF/soft/databases/blastdat/pdb_seq -in_pssm
target_sprot5.pssm -out target_pdb_sprot5.out
```

Databases of sequences

You already know two databases. It is very important that you differentiate them:

PDB

- Contains proteins with available structure
- Biased
- Redundant
- Has few proteins

Uniprot (Swissprot)

- Contains proteins with available sequences
- Non-biased
- Non-redundant
- Has many proteins

E-values

E-values and scores always depend on the substitution matrix you use

If you use a non reliable or inappropriate substitution matrix your hits will be wrong altough you get good E-values

If you use a reliable substitution matrix you can trust your E-values

Create your own PSSM

Can we create a PSSM with the sequences that we want?

Yes, using programs to create MSA such as ClustalW or T-Coffee

Create your own PSSM

Get a set of sequences from uniprot

```
ORC1_DROME (O16810) ORIGIN RECOGNITION COMPLEX SUBUNIT 1 (DMORC1).    380   e-105
ORC1_SCHPO (P54789) ORIGIN RECOGNITION COMPLEX SUBUNIT 1.             295   2e-79
ORC1_CANAL (O74270) ORIGIN RECOGNITION COMPLEX SUBUNIT 1.             223   1e-57
ORC1_YEAST (P54784) ORIGIN RECOGNITION COMPLEX SUBUNIT 1 (ORIGIN...   189   1e-47
ORC1_KLULA (P54788) ORIGIN RECOGNITION COMPLEX SUBUNIT 1.             179   1e-44
CC18_SCHPO (P41411) CELL DIVISION CONTROL PROTEIN 18.                 115   2e-25
CC6_YEAST (P09119) CELL DIVISION CONTROL PROTEIN 6.                    87   8e-17
YPZ1_METTF (P29570) HYPOTHETICAL 40.6 KDA PROTEIN (ORF1').             60   1e-08
YPV1_METTF (P29569) HYPOTHETICAL 40.7 KDA PROTEIN (ORF1).              60   1e-08
SIR3_YEAST (P06701) REGULATORY PROTEIN SIR3 (SILENT INFORMATION ...    47   1e-04
G6PI_OENME (P54243) GLUCOSE-6-PHOSPHATE ISOMERASE, CYTOSOLIC (GP...    31   6.6
```

Then use the following command:

```
perl /mnt/NFS_UPF/soft/perl-lib/FetchFasta.pl  -i  file.list

-d  /mnt/NFS_UPF/soft/databases/blastdat/uniprot_sprot.fasta  -o  file.fasta
```

Put them in a MSA with the target using clustalw

```
cat  target.fa  >  pssm.fasta
```

```
cat  file.fasta  >>  pssm.fasta
```

Then run clustalw:

```
clustalw2  pssm.fasta
```

Create your own PSSM

Input the generated MSA into psiblast

```
psiblast -in_msa pssm.fa -out target_pdb_specific.out -db
/mnt/NFS_UPF/soft/databases/blastdat/pdb_seq
```

Some exercises

You can try the exercises to practice for the practical exams

QUESTIONS FROM THE TUTORIAL

Now we can compare all the results and answer the following questions:

1) Why are the e-values different in *target_pdb.out* than in the fifth iteration in *target_pdb_5.out*?

2) Why do we need to run psiblast with uniprot_sprot.fasta before searching in pdb_seq?

3) When obtaining the file *target_pdb_sprot5.out* why we didn't run 5 iterations as before?

4) Search in the SCOP database with the PDB code of the best match of the target sequence. Do all the files *target_pdb_specific.out*, *target_pdb_sprot5.out*, *target_pdb_5.out* and *target_pdb.out* produce the same result?

5) Can you use the file target_sprot5.out to obtain the name of the fold in SCOP? Why?

6) What are the folds of the following sequences?

   a. *problem1/serc_myctu.fa*

   b. *problem2/p72_mycmy.fa*

   c. *problem3/lip_staau.fa*

   d. *problem4/orc1_human.fa*

How to find the fold of one protein?

How can I know the fold of one protein?

Using the SCOP database

How to use the SCOP database?

https://scop.mrc-lmb.cam.ac.uk/