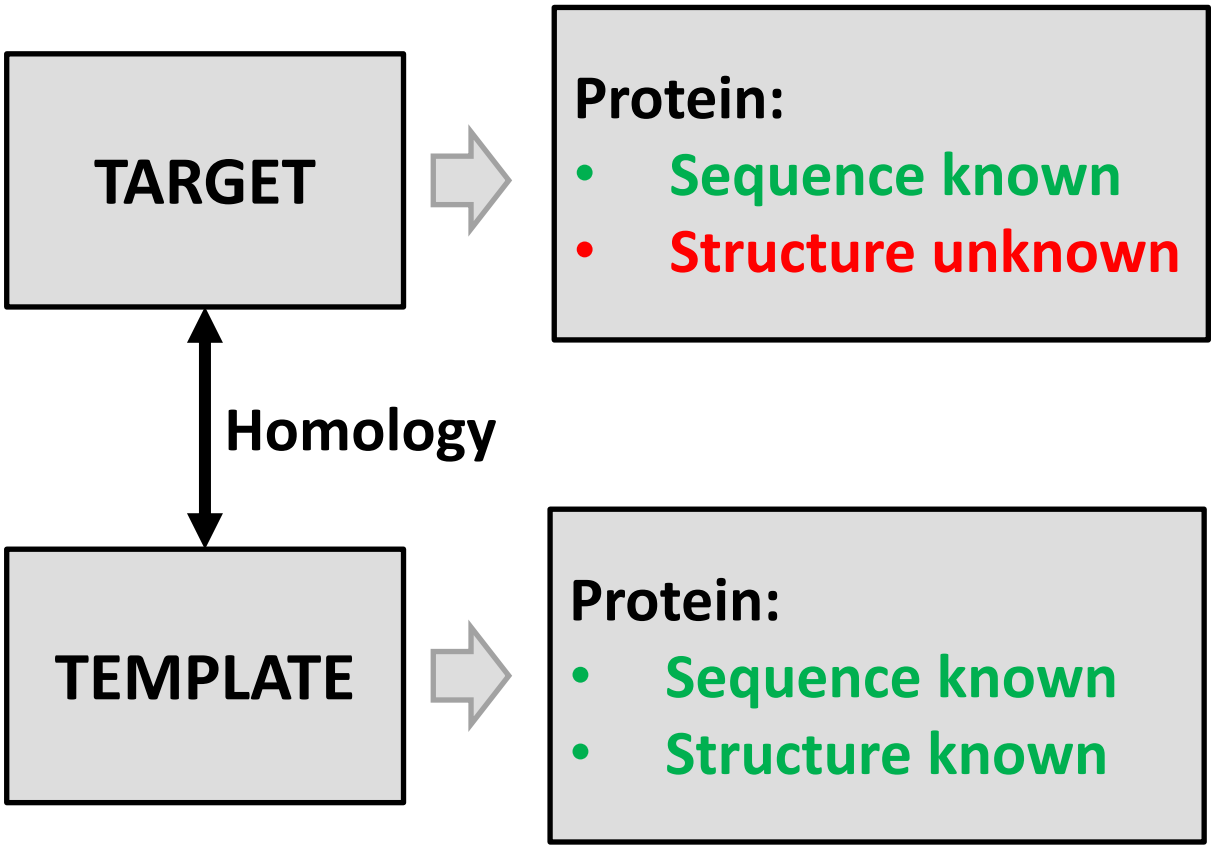
Structural biology

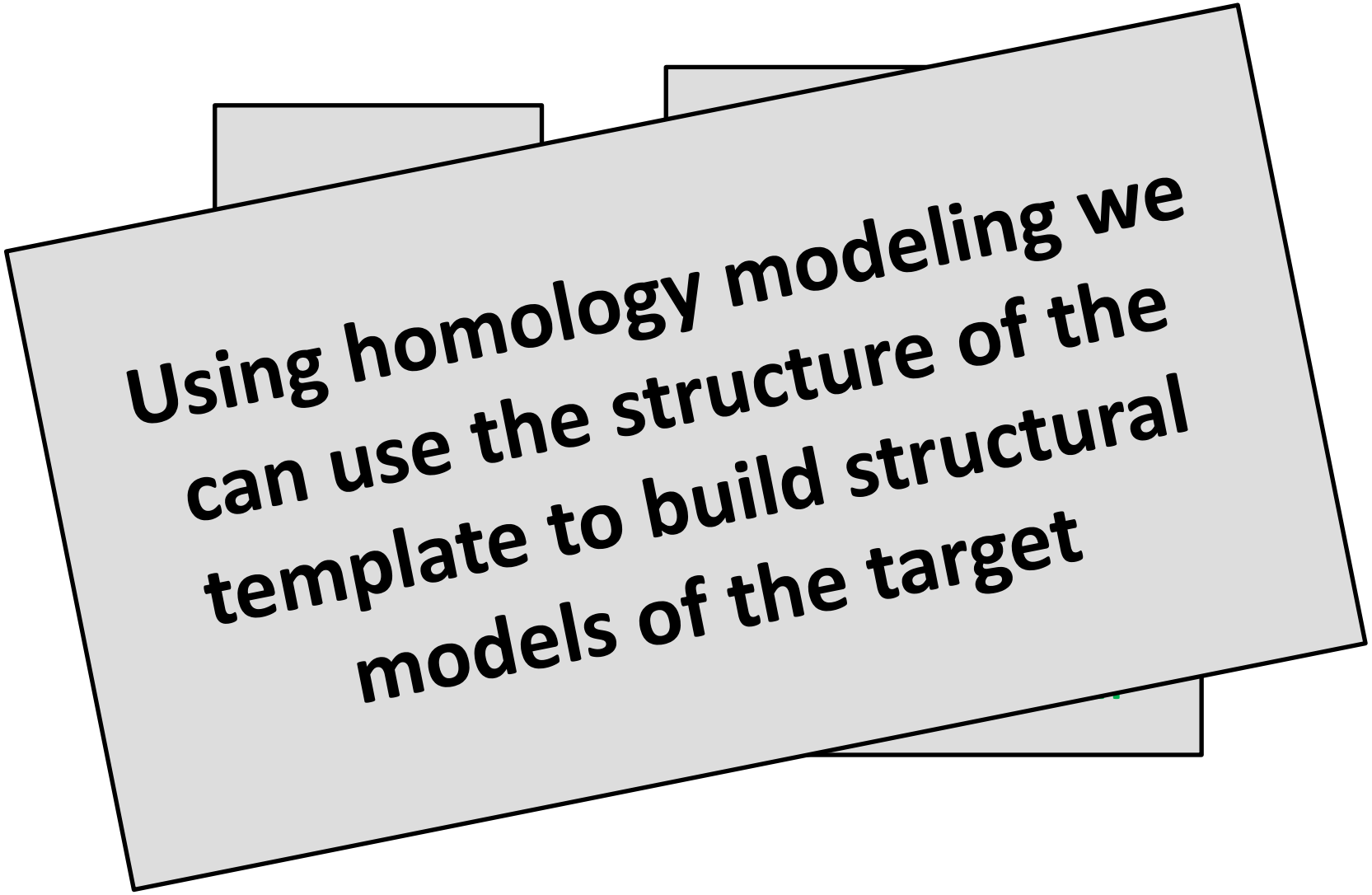
Practice 1: BLAST

Course 2023-2024

Target and template



Target and template



Target and template

If two proteins are homologs they will have similar sequences

How can we know if two protein sequences are similar?

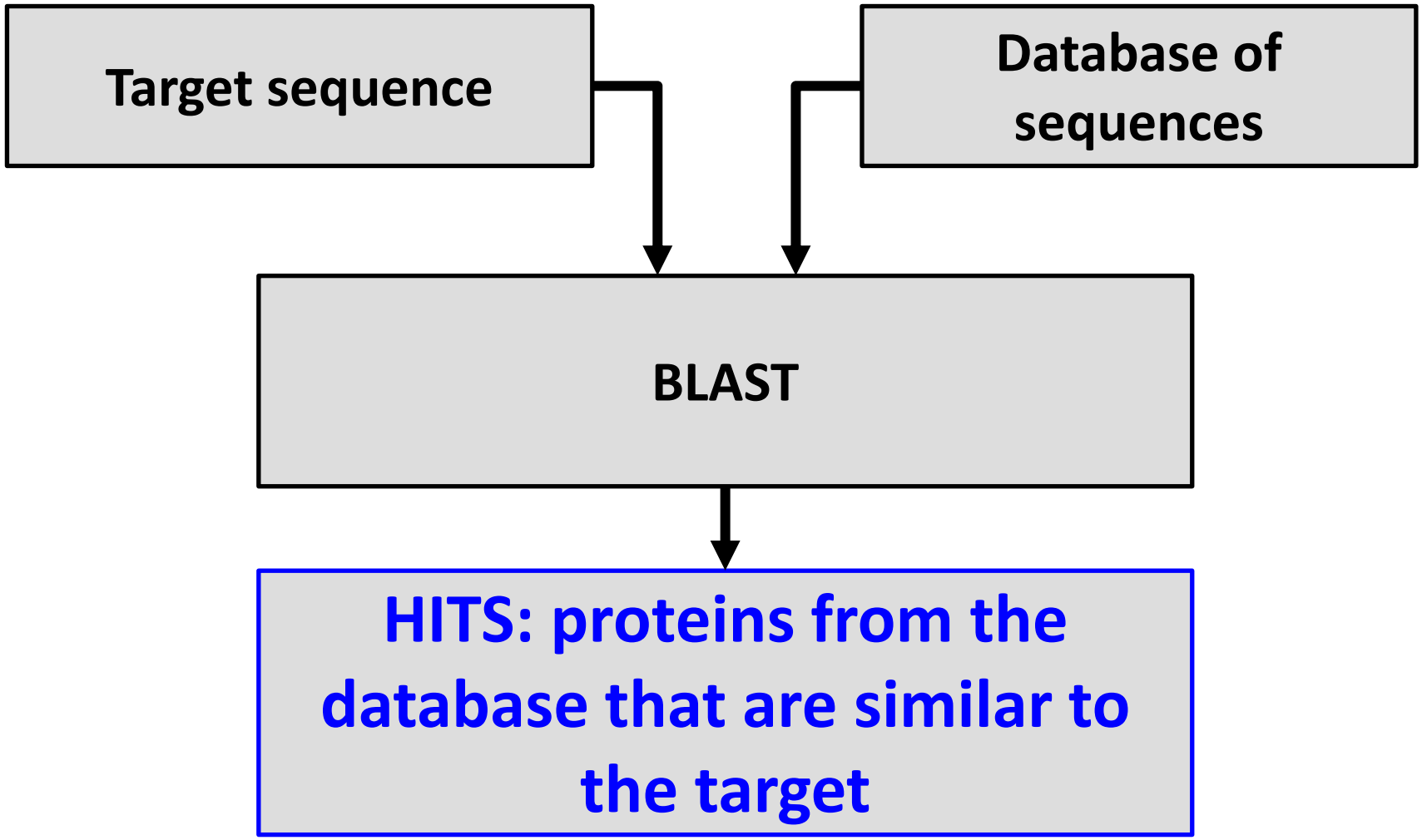
Target and template

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

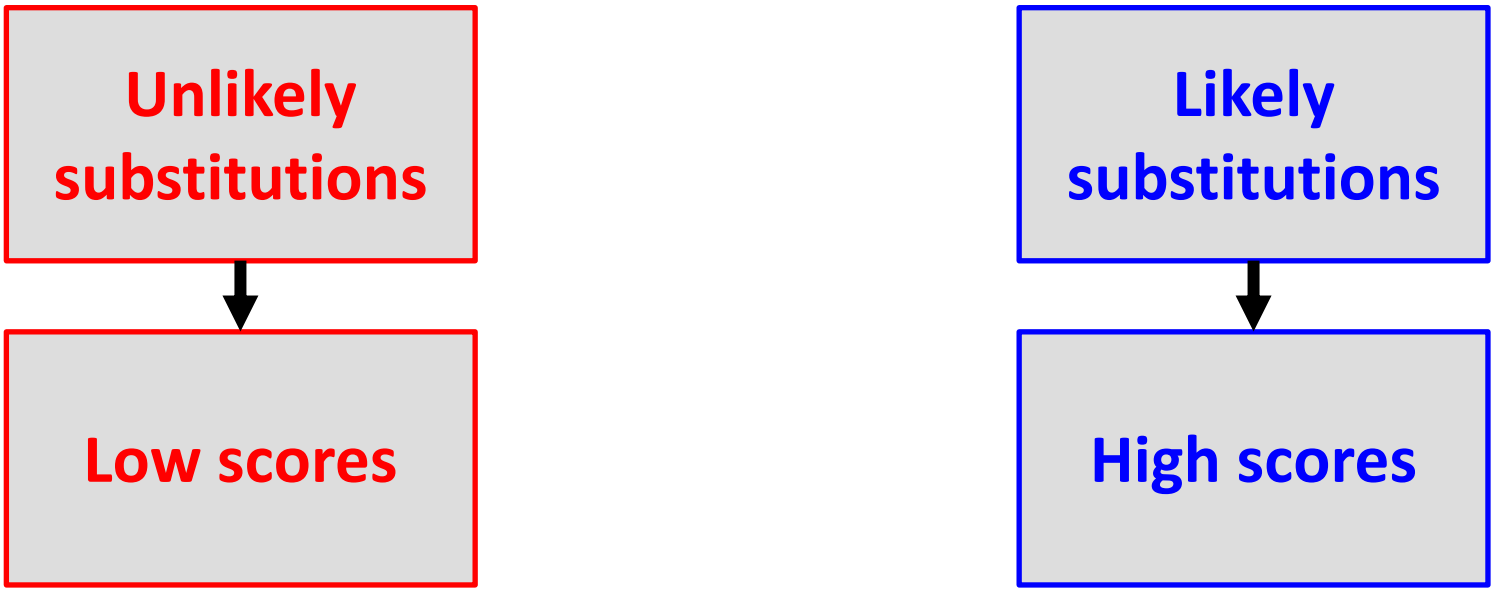
Pn: AAVHR

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

It uses substitution matrices

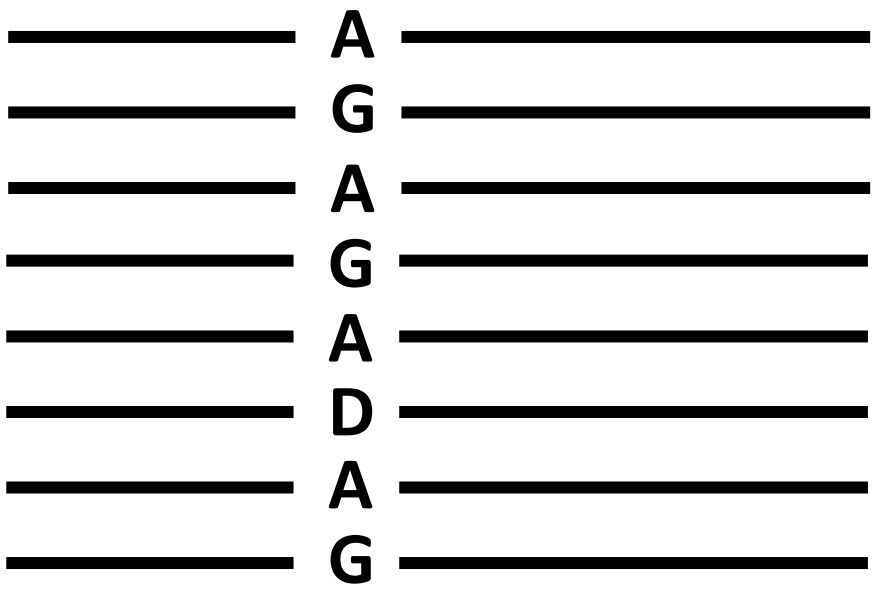
Substitution matrices

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



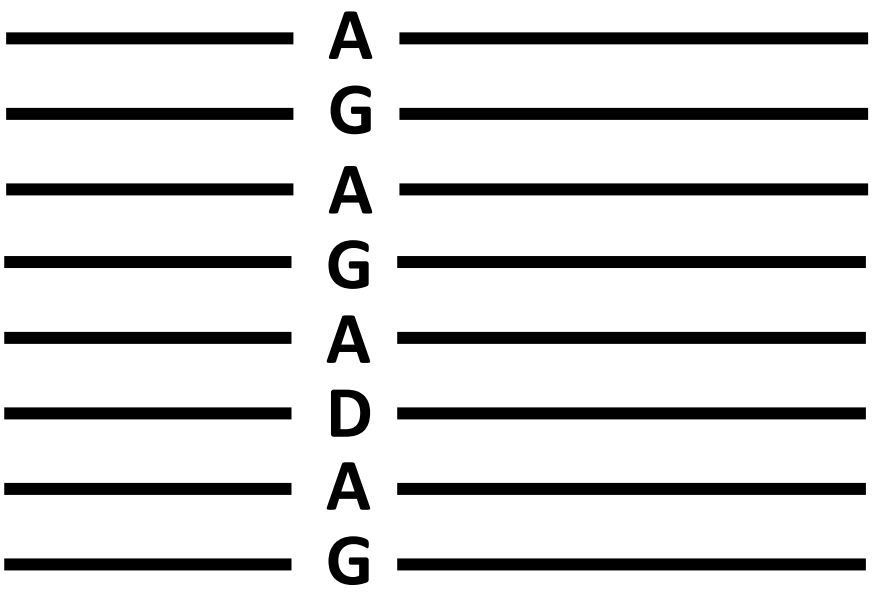
Substitution matrices

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



Substitution matrices

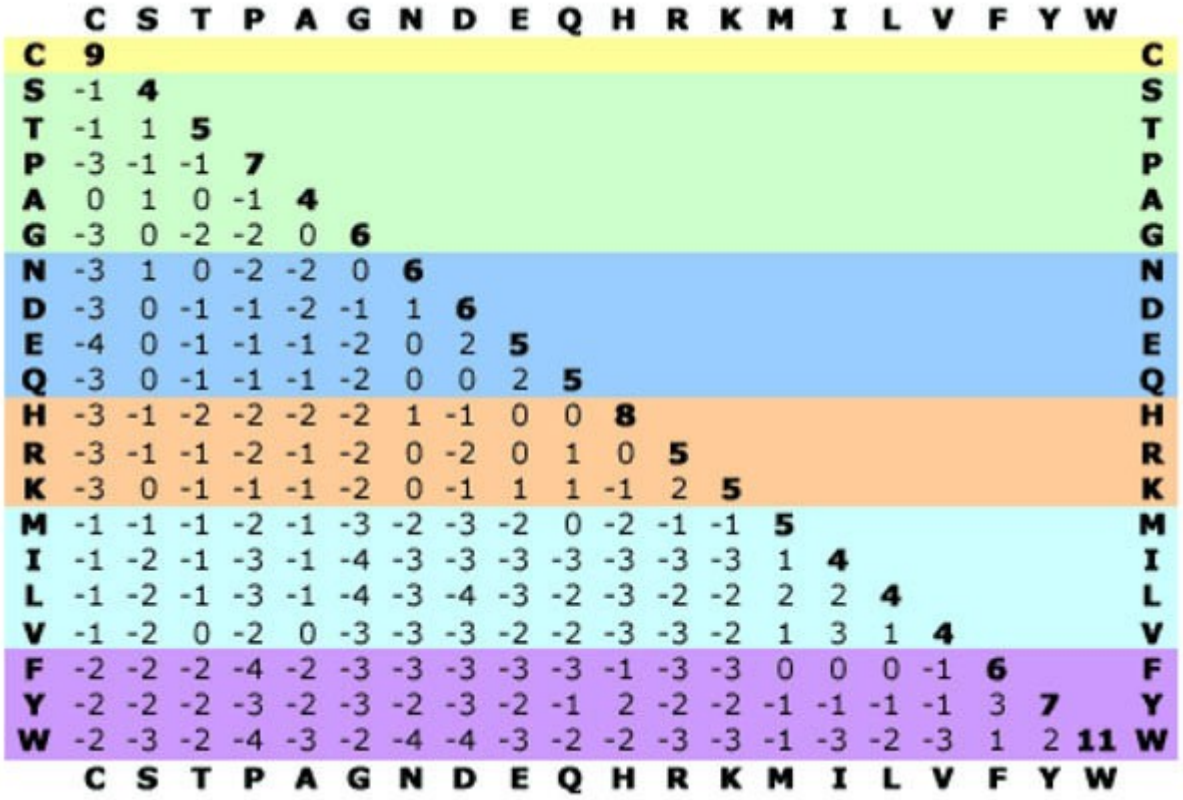
Substitution matrices contain scores associated to the frequency of substitution between different amino acids



A-G substitution is likely → High scores

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

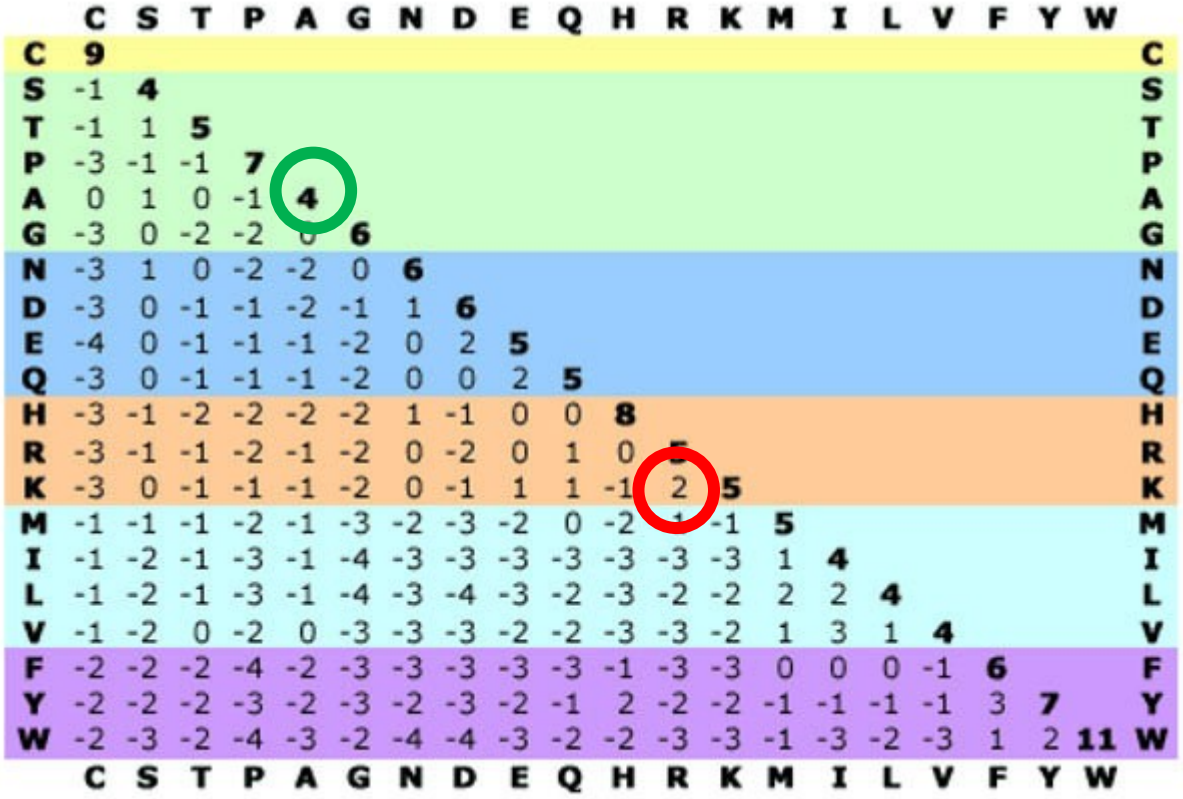
Substitution matrices



	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



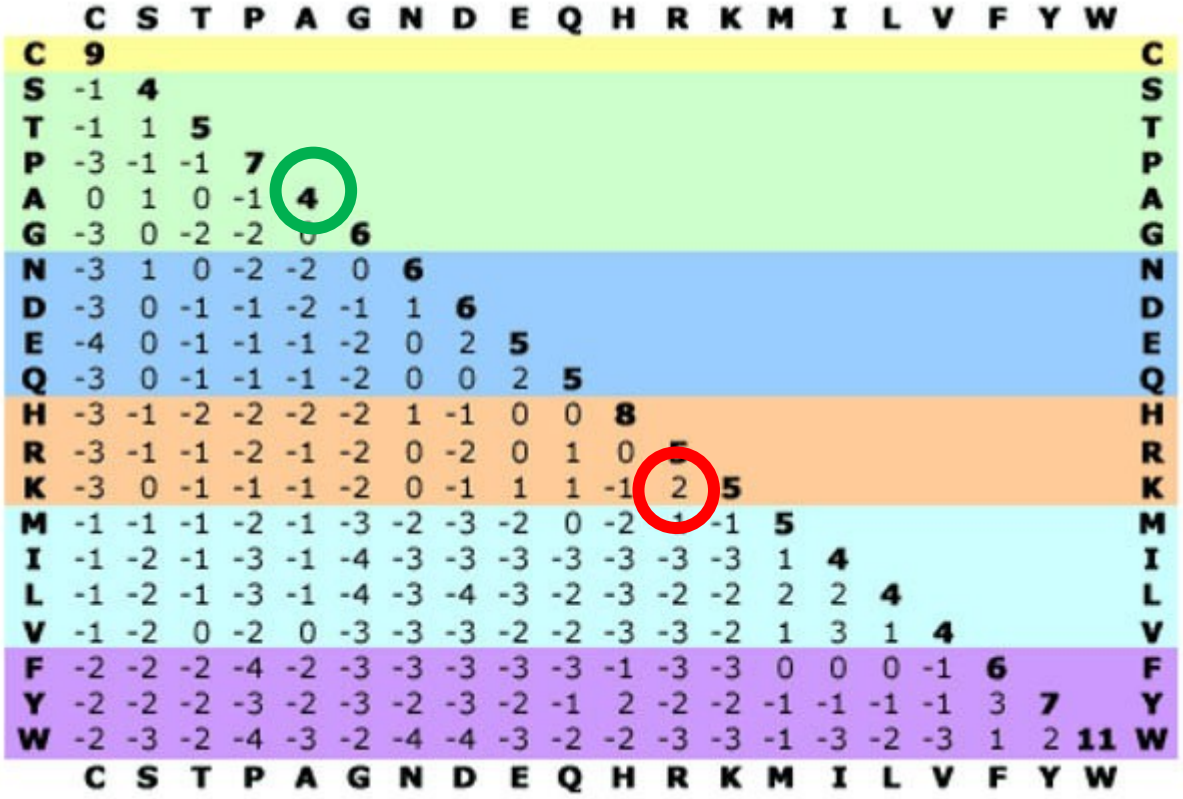
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	**9**																				C
S	-1	**4**																			S
T	-1	1	**5**																		T
P	-3	-1	-1	**7**																	P
A	0	1	0	-1	**4**																A
G	-3	0	-2	-2	0	**6**															G
N	-3	1	0	-2	-2	0	**6**														N
D	-3	0	-1	-1	-2	-1	1	**6**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

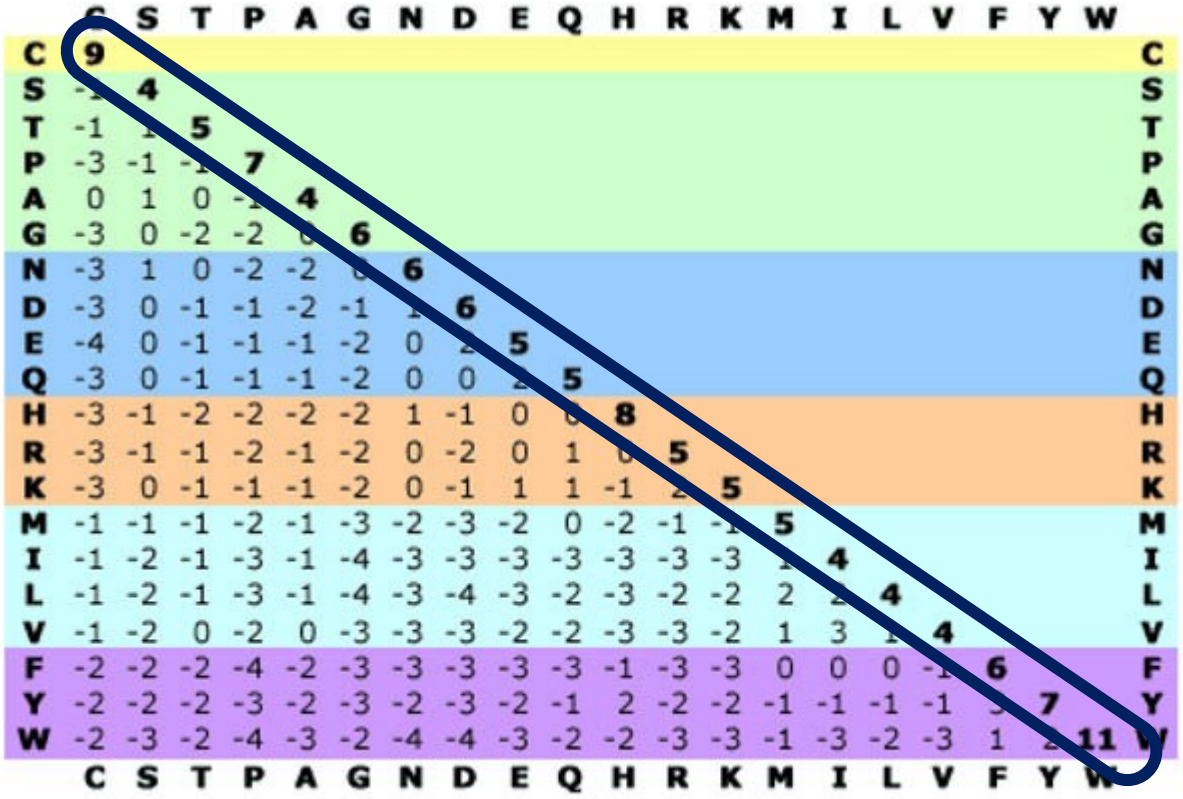
K and R are both basic amino acids, they have similar chemical properties

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

Sequences producing significant alignments:		Score (Bits)	E 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?

Sequences producing significant alignments:		Score (Bits)	E 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

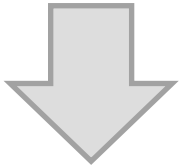
$$E(S) \approx \exp\left(-NmnKe^{-(S-\mu)}\right)$$

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

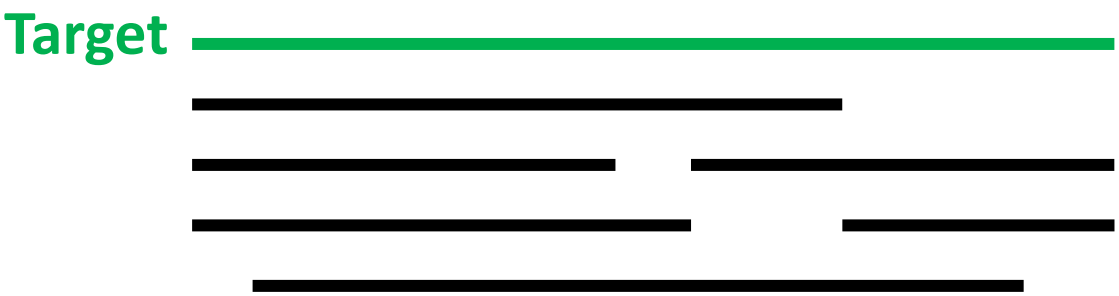
What is the problem pf using a BLO… …tion

- …
- D… …erent su…

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

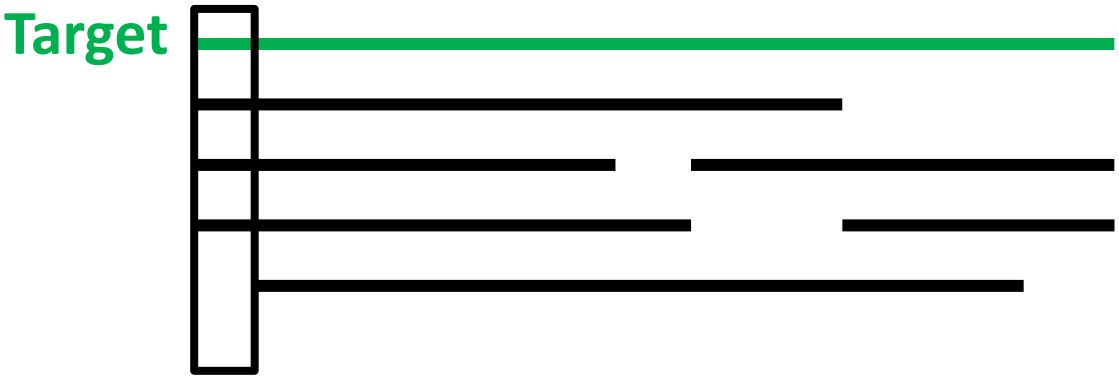
Position specific substitution matrices (PSSM)

PSSM are obtained from a MSA



Position specific substitution matrices (PSSM)

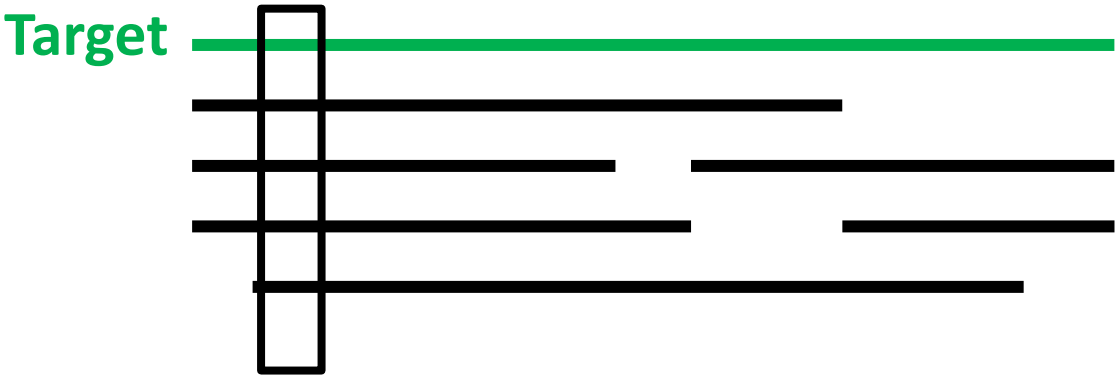
PSSM are obtained from a MSA



Position 1
Substitution
scores

Position specific substitution matrices (PSSM)

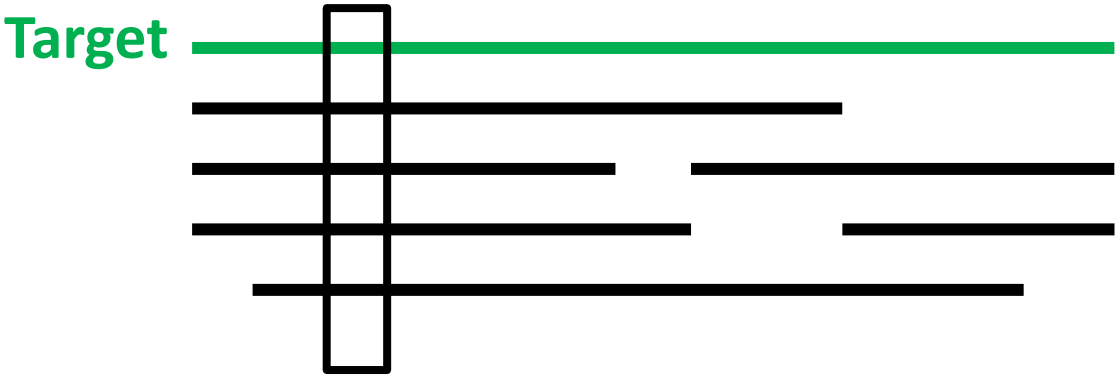
PSSM are obtained from a MSA



Position 2
Substitution
scores

Position specific substitution matrices (PSSM)

PSSM are obtained from a MSA



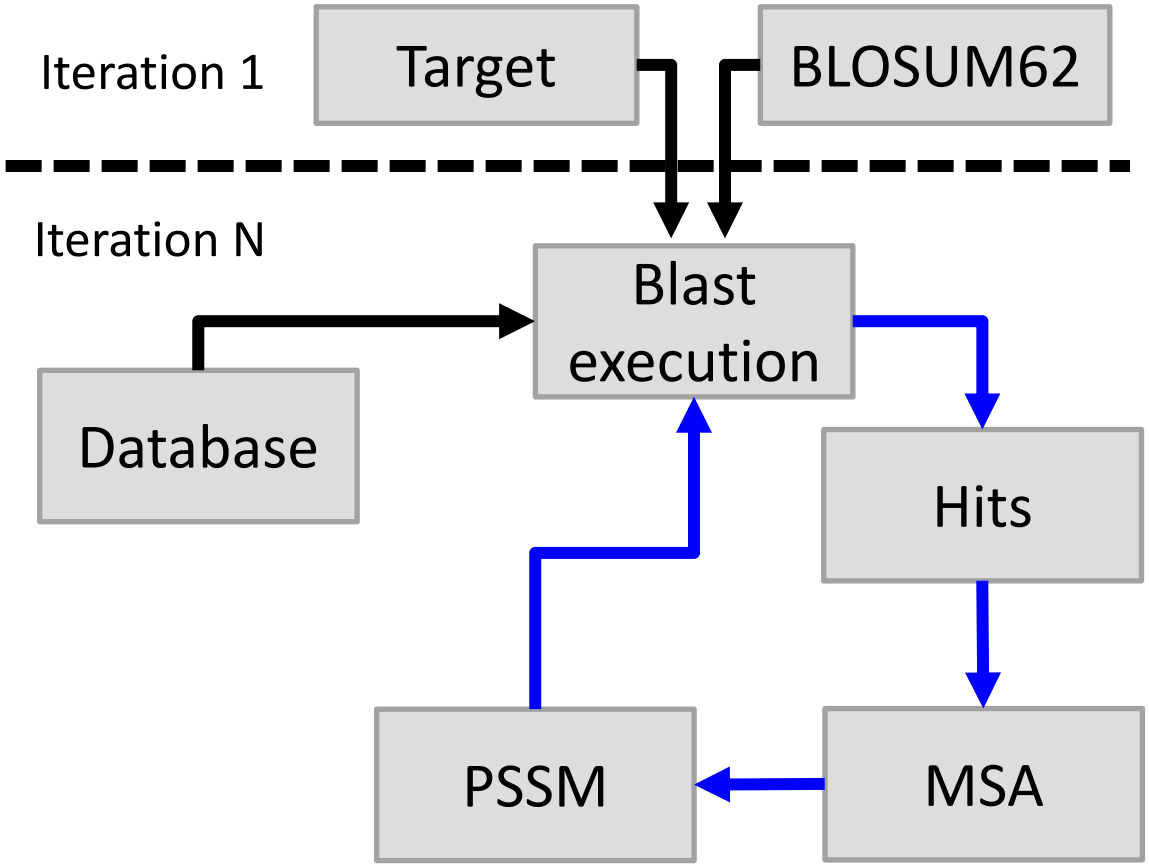
PSI-BLAST

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

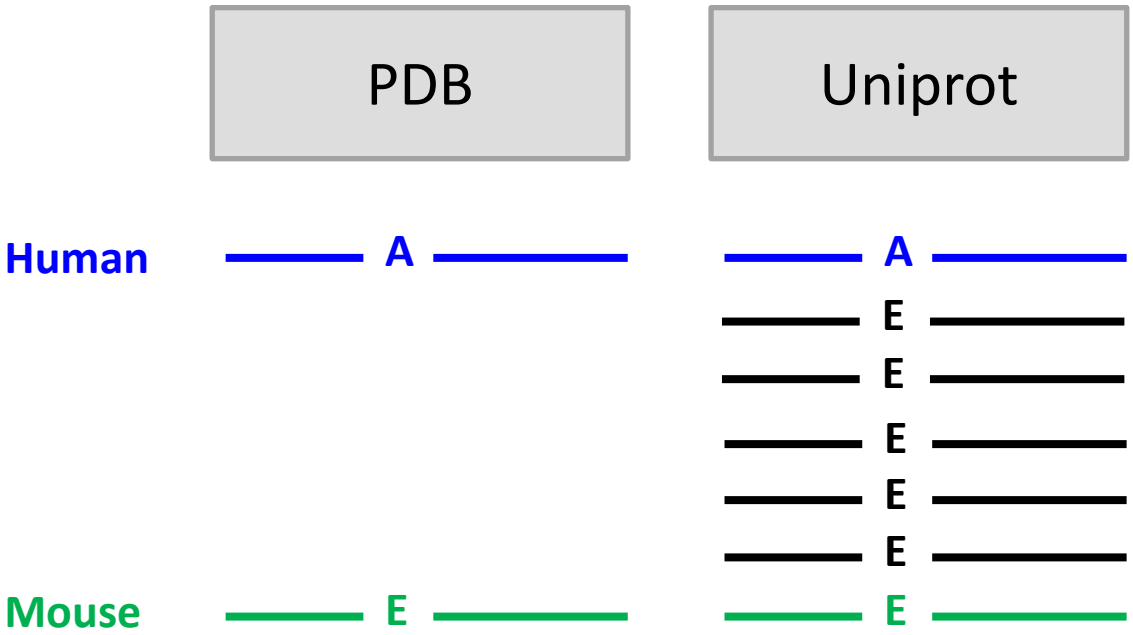
The same proteins are repeated several times

The PDB is very biased

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

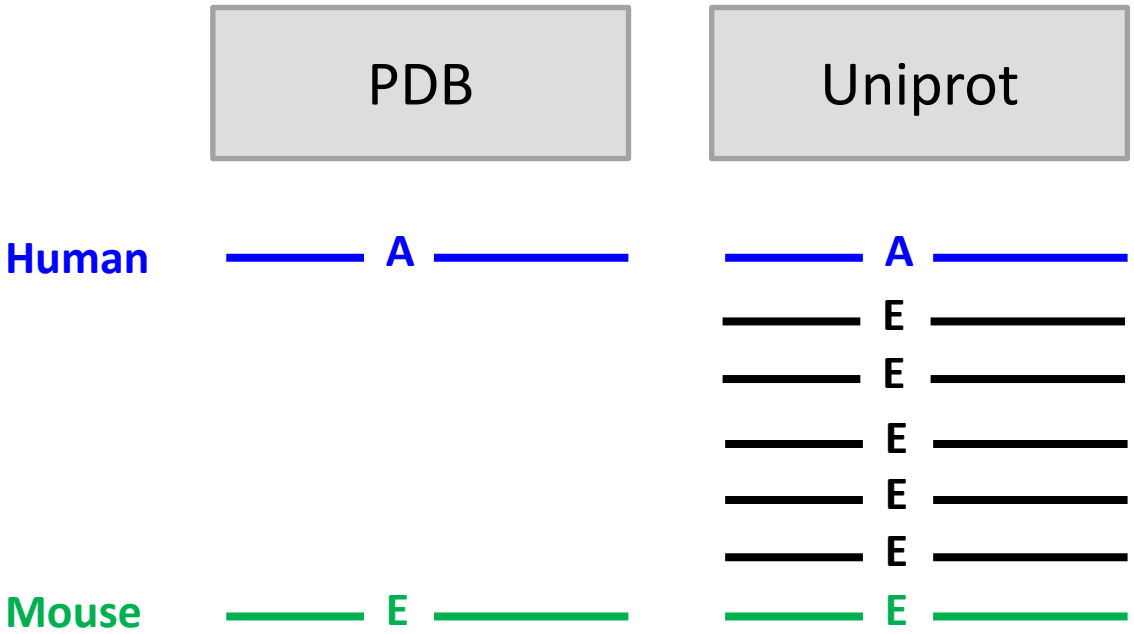
Databases of sequences

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

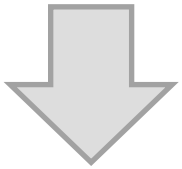
E-values and scores always depend on the substitution matrix you use

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?

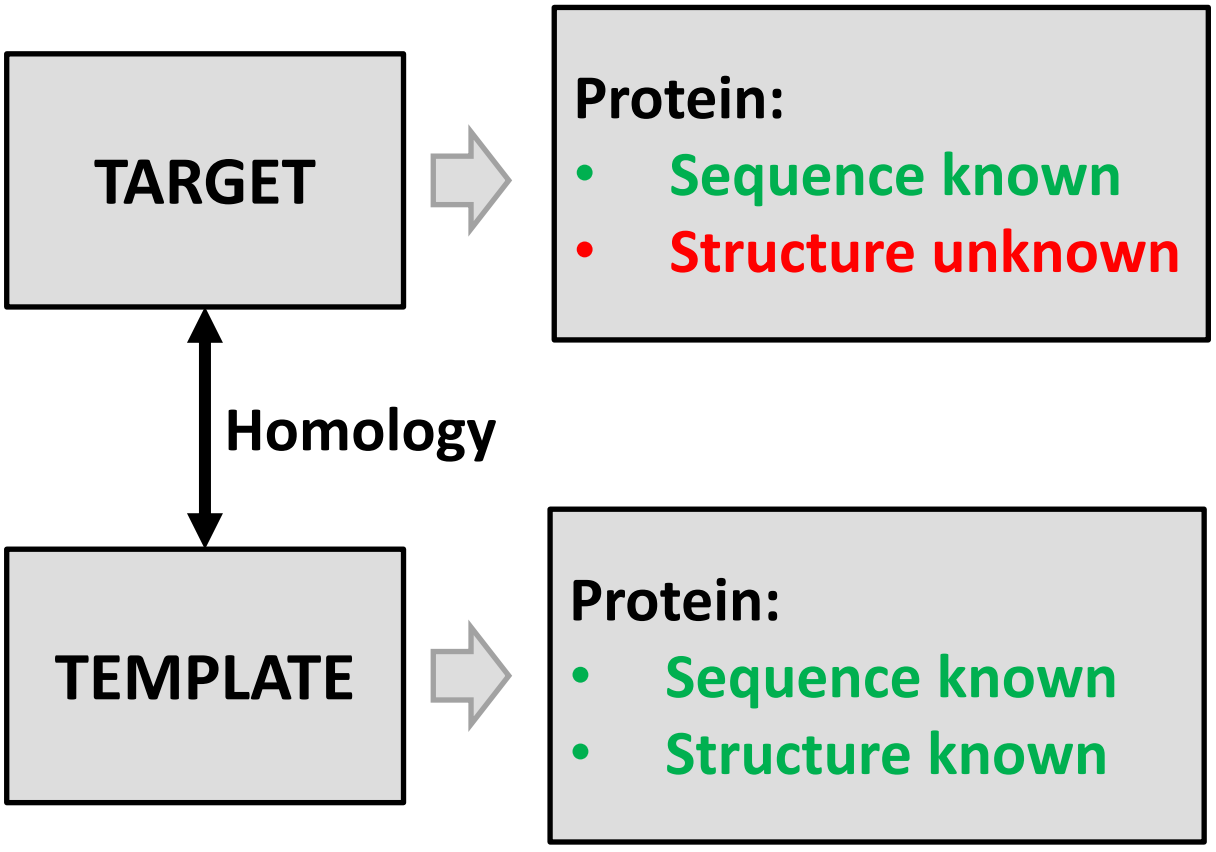
https://scop.mrc-lmb.cam.ac.uk/

Structural biology

Practice 2: Hidden Markov Models and HMMer

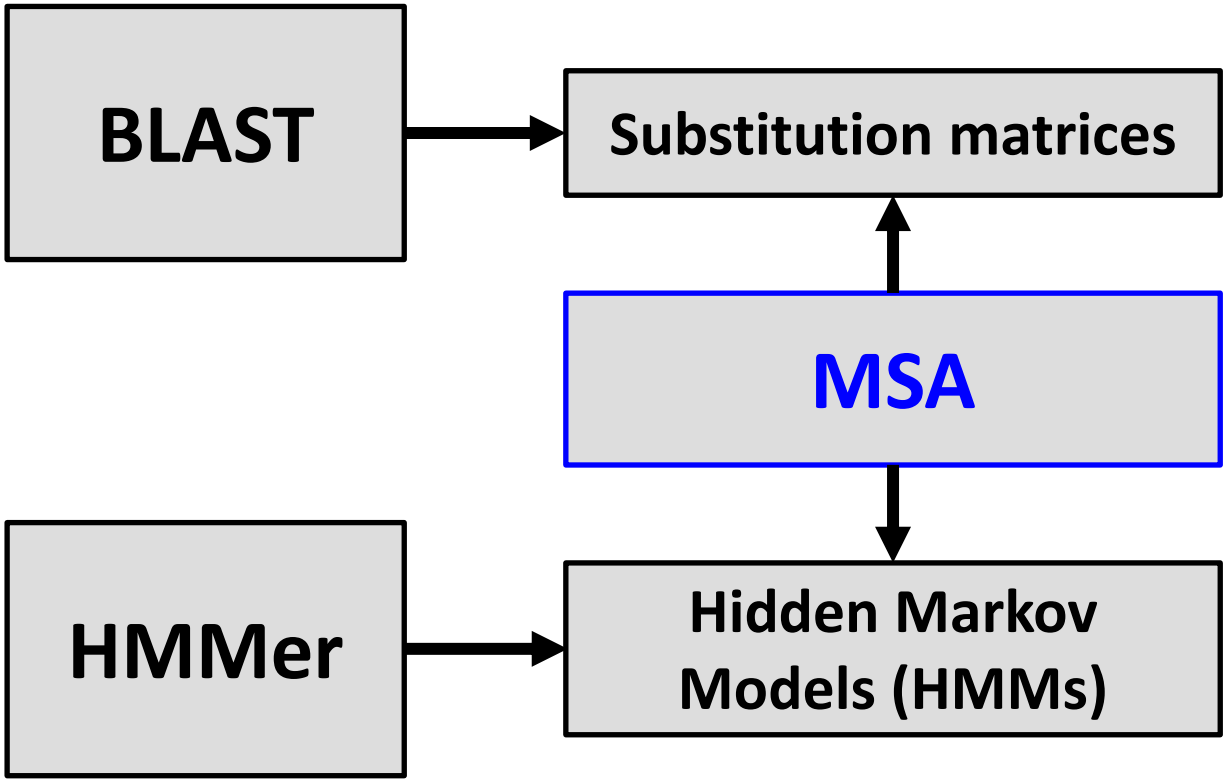
Course 2022-2023

Target and template



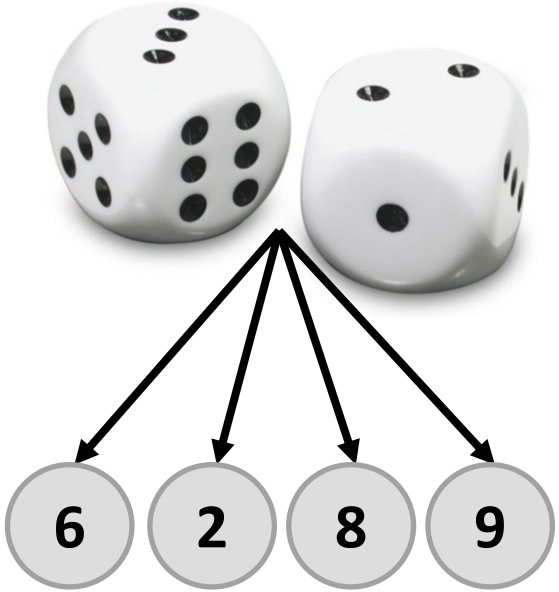
Hidden Markov Models and substitution matrices

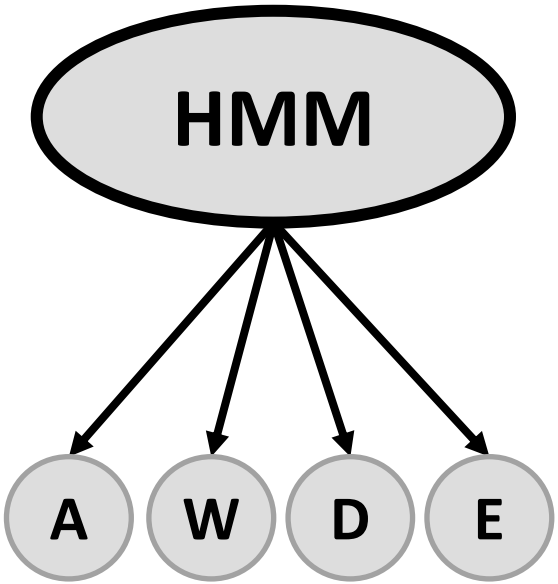
Hidden Markov Models (HMMs) are equivalent to substitution matrices



What is a HMM?

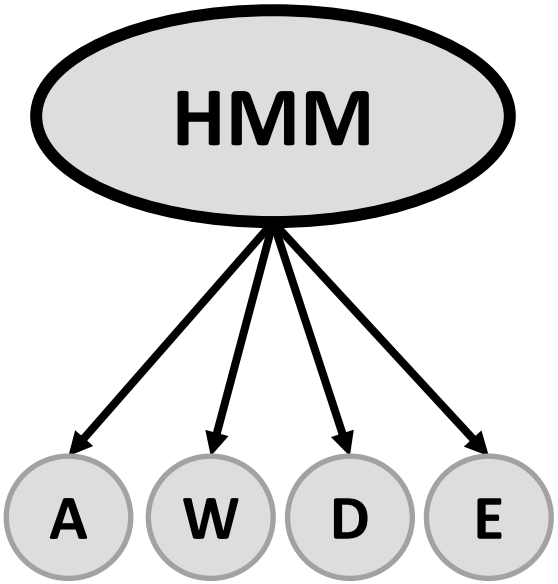
The same way that dice generate numbers, HMM generate amino acids





What is a HMM?

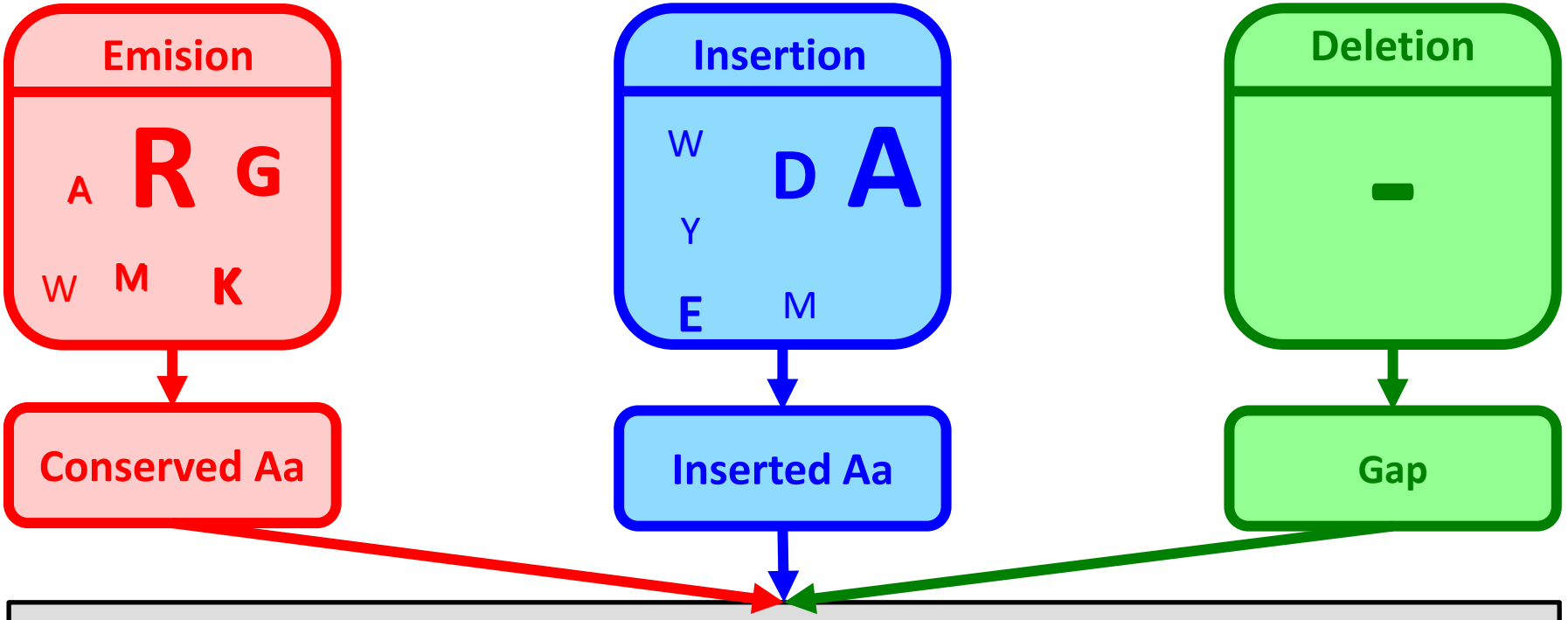
Each amino acid is produced with a specific probability contained inside the HMM



P(prot) = P(A) x P(W) x P(D) x P(E)

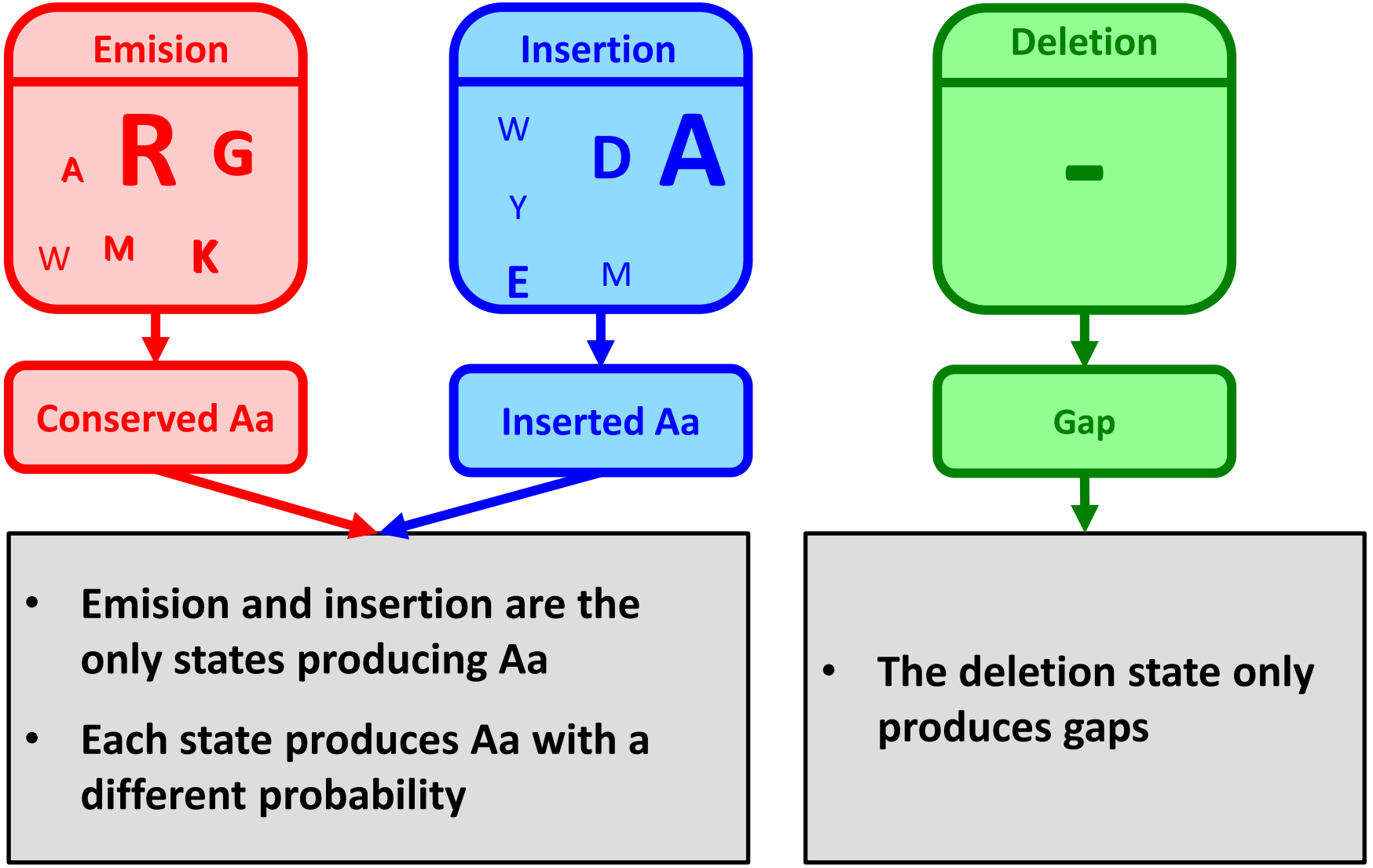
What is a HMM?

HMMs have states, each state has its own probabilities for producing amino acids

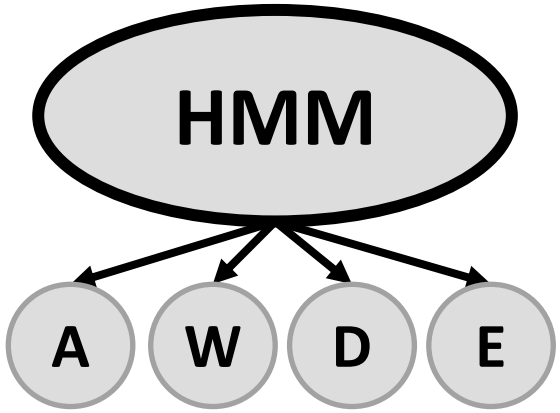


In reference with the sequences contained in the MSA used to create the HMM

What is a HMM?

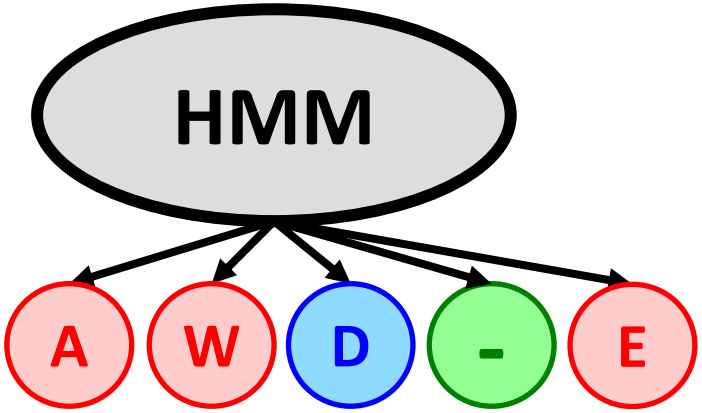

Emision
A R G
W M K
Insertion
W D A
Y
E M
Deletion
-
Conserved Aa
Inserted Aa
Gap
• Emision and insertion are the only states producing Aa
• Each state produces Aa with a different probability
• The deletion state only produces gaps

What is a HMM?



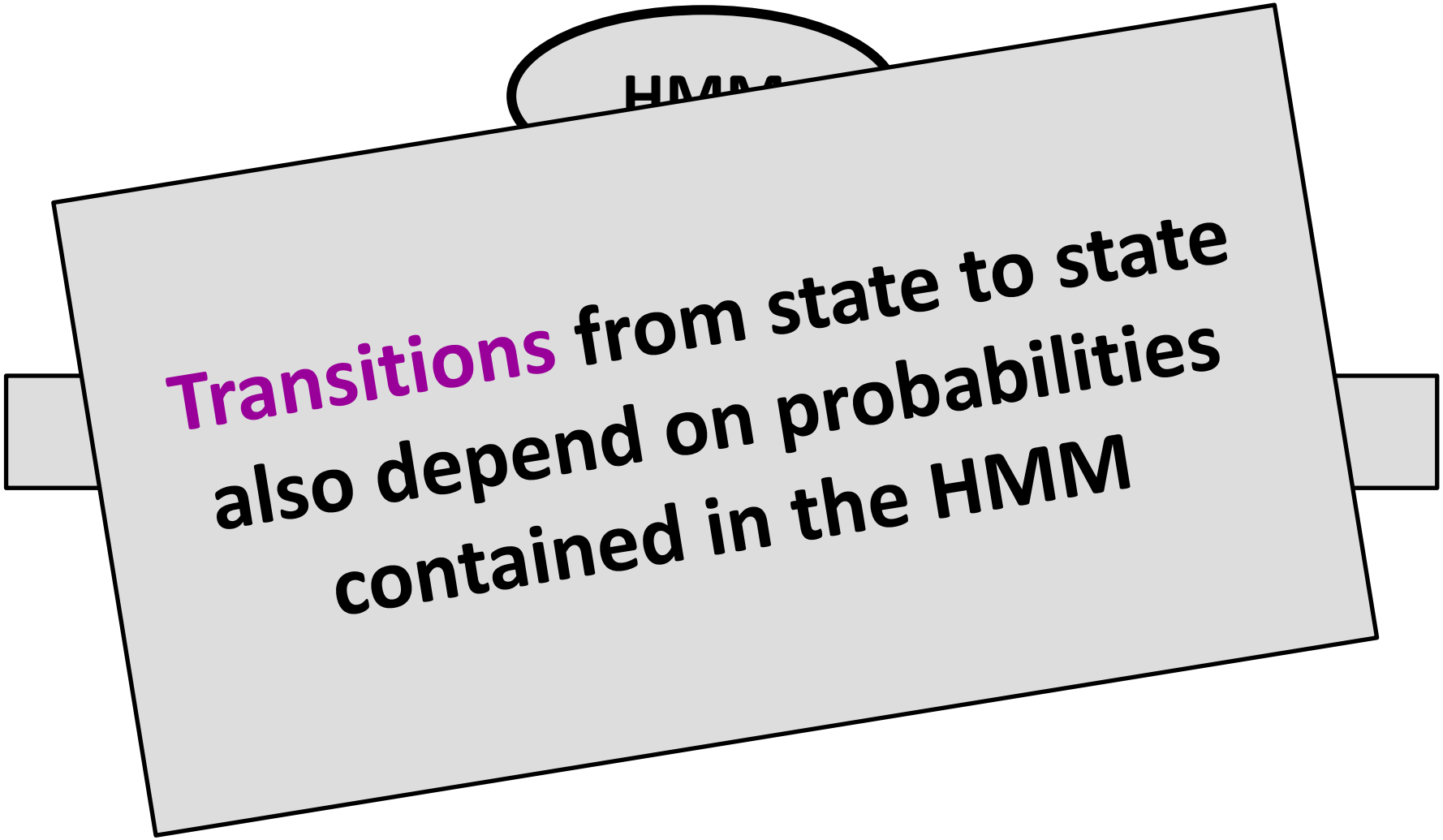
$$P(prot) = P(A) \times P(W) \times P(D) \times P(E)$$

What is a HMM?

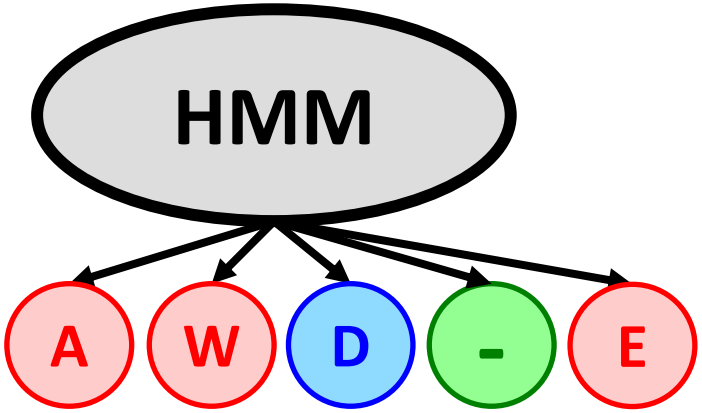


$$P(prot) = P_e(A) \times P_e(W) \times P_i(D) \times P_e(E)$$

What is a HMM?

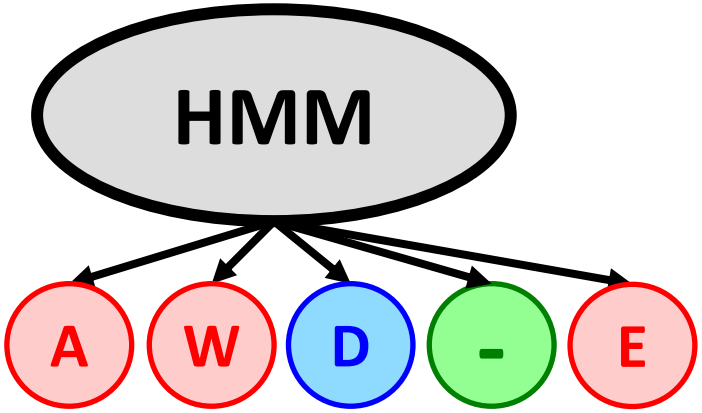

Transitions from state to state
also depend on probabilities
contained in the HMM

What is a HMM?



$$P(prot) = Pt(ee) \times P_e(A) \times Pt(ee) \times P_e(W) \times Pt(ei) \times P_i(D) \times Pt(id) \times Pt(de) \times P_e(E)$$

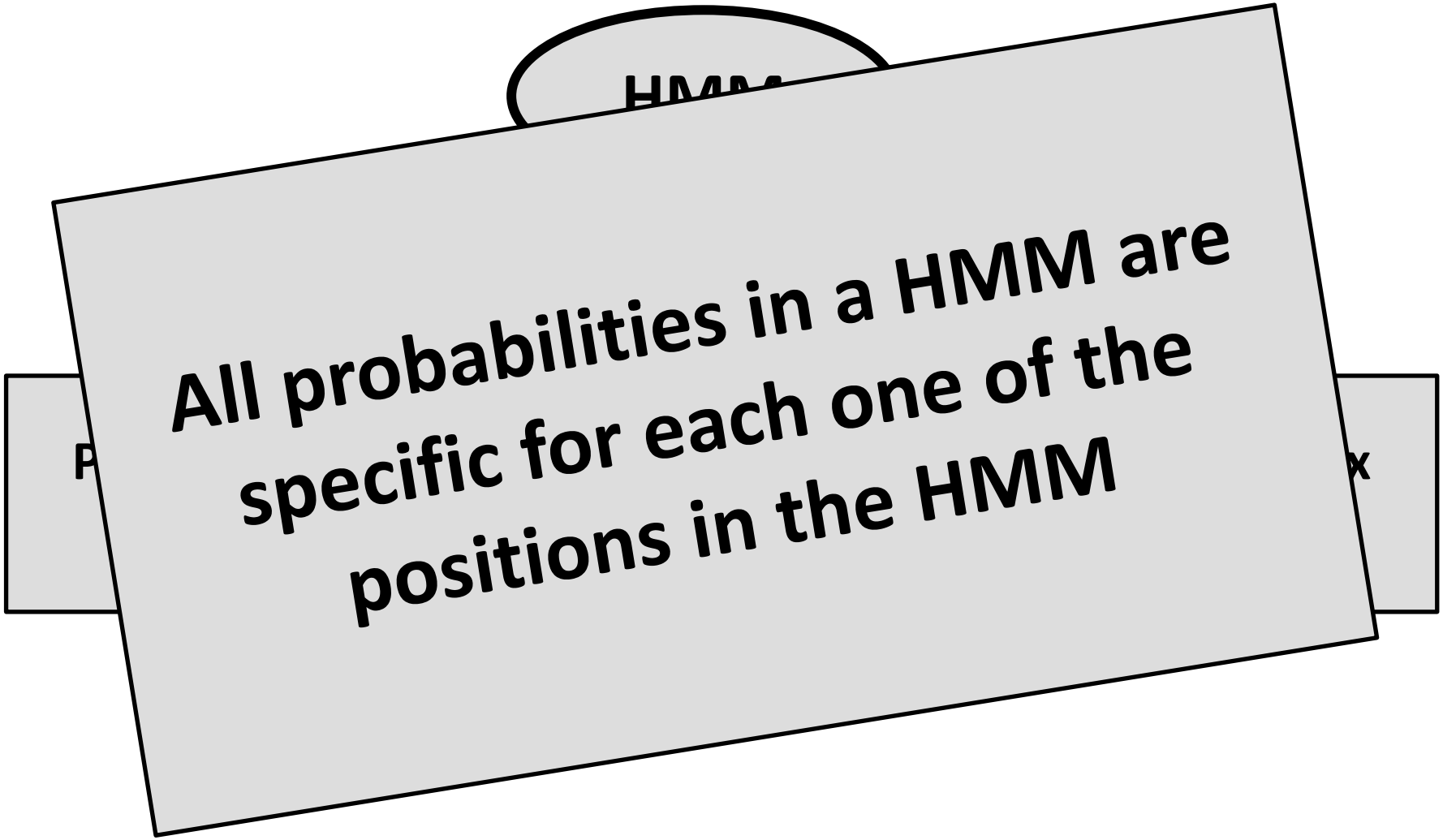
What is a HMM?



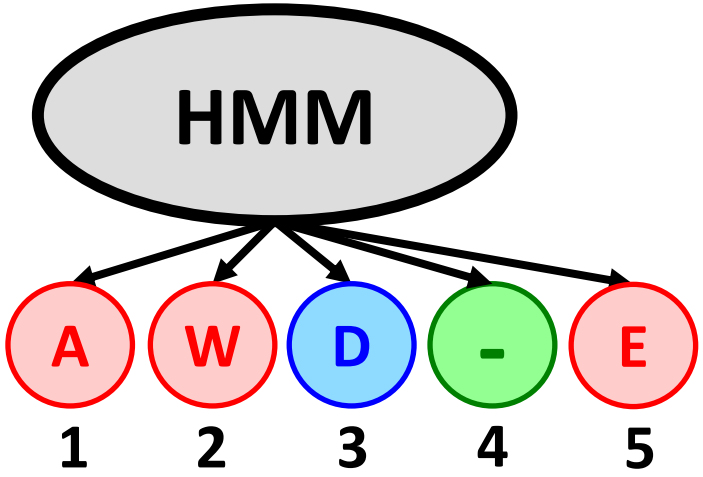
$$P(prot) = Pt(ee) \times P_e(A) \times Pt(ee) \times P_e(W) \times Pt(ei) \times P_i(D) \times Pt(id) \times Pt(de) \times P_e(E)$$

When the HMM introduces gaps in the sequence only considers the probability of moving inside the deletion state

What is a HMM?


All probabilities in a HMM are specific for each one of the positions in the HMM

What is a HMM?



$$P(prot) = P_{t1}(ee) \times P_{e1}(A) \times$$

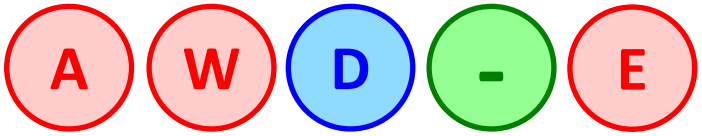
$$P_{t2}(ee) \times P_{e2}(W) \times P_{t3}(ei) \times P_{i3}(D) \times$$

$$P_{t4}(id) \times P_{t5}(de) \times P_{e5}(E)$$

What is a HMM?

A HMM can create the same sequence using different state paths

Path 1:



Path 2:

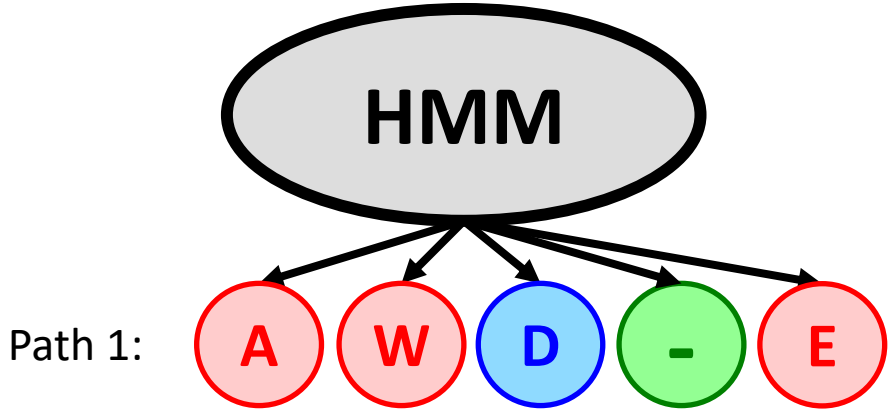
Path 3:

Path 4:

Etc…

What is a HMM?

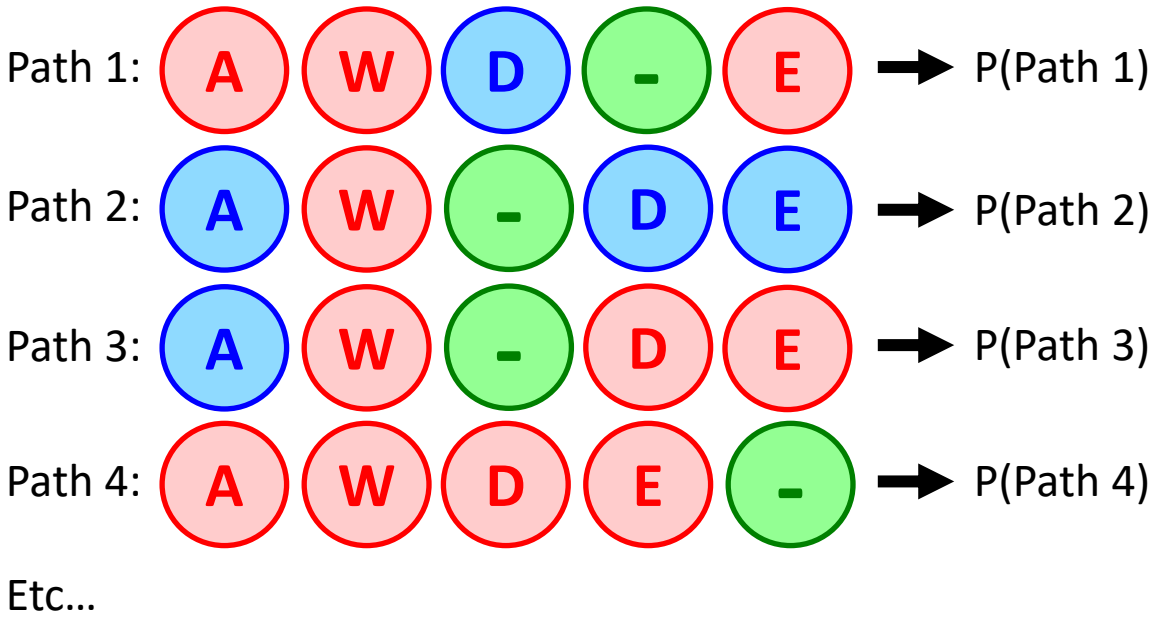
The probability of a HMM making one protein sequence through one specific path is the product of all the probabilities involved



P(path1) = P_{t1}(ee) x P_{e1}(A) x P_{t2}(ee) x P_{e2}(W) x P_{t3}(ei) x P_{i3}(D) x P_{t4}(id) x P_{t5}(de) x P_{e5}(E)

What is a HMM?

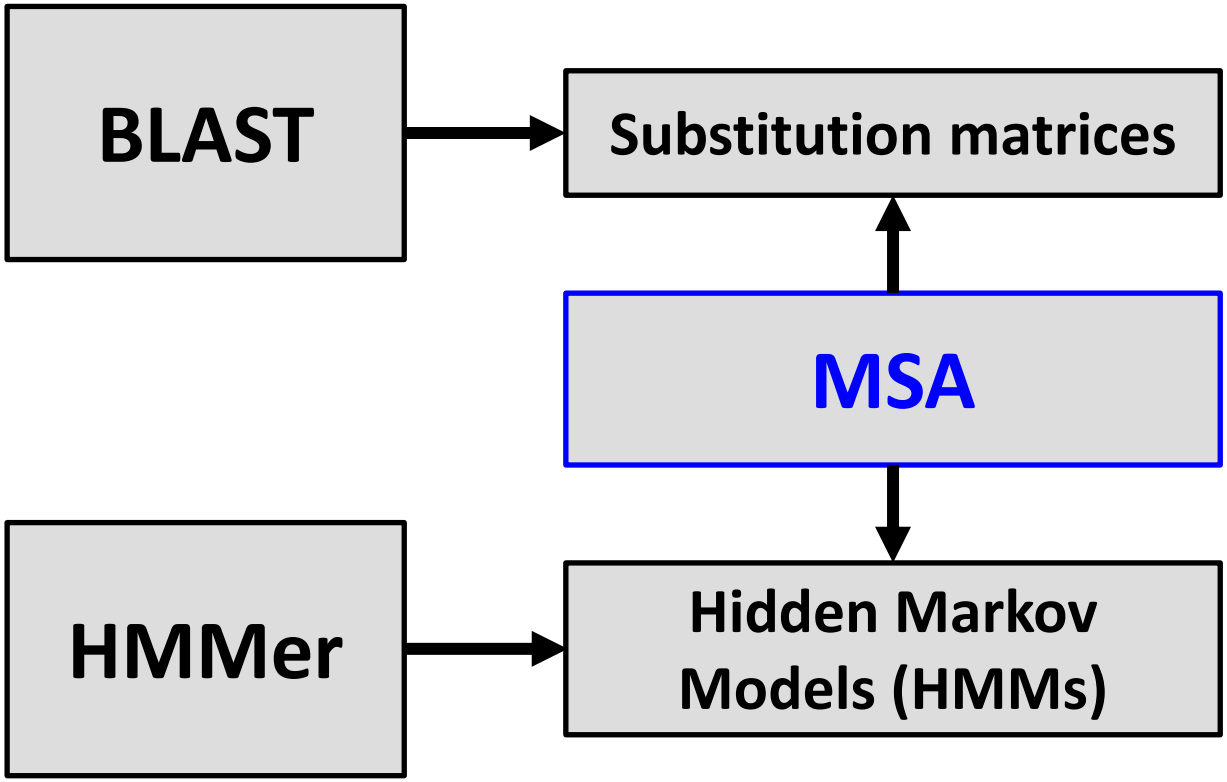
The probability of a HMM making one protein sequence through any path is the addition of the probabilities for each path



P(total) = P(path1) + P(path2) + P(path3) + P(path4) + Etc…

What is a HMM?

Hidden Markov Models (HMMs) are equivalent to substitution matrices



Create a HMM with hmmbuild

Create a HMM from a MSA using hmmbuild

Step 1: Creating a HMM using hmmbuild

To generate a HMM of a particular family of sequences we need a previous alignment of these sequences. This MSA, named seed, will be turn into a HMM by using the program hmmbuild. Here is an example of HMM usage:

- **hmmbuild [model_HMM] [alignment]**

The alignment has to be in STOCKHOLM format, like **globins4.sto**. You will find the required files in the folder HMMER within the directory of exercise_2. We run this as an example:

```
hmmbuild  globins4.hmm  globins4.sto
```

Create a HMM with hmmbuild

How does a HMM look from the inside?

```
HMMER3/f [3.1b2 | February 2015]
NAME  globins4
LENG  149
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Tue Jan  5 18:22:24 2021
NSEQ  4
EFFN  0.964844
CKSUM 2027839109
STATS LOCAL MSV       -9.9014  0.70957
STATS LOCAL VITERBI  -10.7224  0.70957
STATS LOCAL FORWARD   -4.1637  0.70957
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q
     R        S        T        V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.36553  4.52577  2.96709  2.70473  3.20818  3.02239  3.41069  2.90041  2.55332  2.35210  3.67329  3.19812  3.45595  3.16091
 3.07934  2.66722  2.85475  2.56965  4.55393  3.62921
         2.68640  4.42247  2.77497  2.73145  3.46376  2.40504  3.72516  3.29302  2.67763  2.69377  4.24712  2.90369  2.73719  3.18168
 2.89823  2.37879  2.77497  2.98431  4.58499  3.61525
         0.57544  1.78073  1.31293  1.75577  0.18968  0.00000        *
     1   1.70038  4.17733  3.76164  3.36686  3.72281  3.29583  4.27570  2.40482  3.29230  2.54324  3.63799  3.55099  3.93183  3.61602
 3.56580  2.71897  2.84104  1.67328  5.32720  4.10031      9 v - - -
         2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
 2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
         0.03156  3.86736  4.58970  0.61958  0.77255  0.34406  1.23405
     2   2.62748  4.47174  3.31917  2.82619  3.63815  3.49607  2.75382  3.03401  2.75280  2.74783  3.65114  3.24714  2.62341  3.12082
 3.11124  2.79244  2.89355  1.88003  5.06315  3.77128     10 v - - -
```

Create a HMM with hmmbuild

How does a HMM look from the inside?

```
HMMER3/f [3.1b2 | February 2015]
NAME  globins4
LENG  149
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Tue Jan  5 18:22:24 2021
NSEQ  4
EFFN  0.964844
CKSUM 2027839109
STATS LOCAL MSV       -9.9014  0.70957
STATS LOCAL VITERBI  -10.7224  0.70957
STATS LOCAL FORWARD   -4.1637  0.70957
```

General information

Probabilities

```
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q
     R       S        T        V        W        Y
                m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.36553  4.52577  2.96709  2.70473  3.20818  3.02239  3.41069  2.90041  2.55332  2.35210  3.67329  3.19812  3.45595  3.16091
  3.07934  2.66722  2.85475  2.56965  4.55393  3.62921
          2.68640  4.42247  2.77497  2.73145  3.46376  2.40504  3.72516  3.29302  2.67763  2.69377  4.24712  2.90369  2.73719  3.18168
  2.89823  2.37879  2.77497  2.98431  4.58499  3.61525
          0.57544  1.78073  1.31293  1.75577  0.18968  0.00000        *
      1   1.70038  4.17733  3.76164  3.36686  3.72281  3.29583  4.27570  2.40482  3.29230  2.54324  3.63799  3.55099  3.93183  3.61602
  3.56580  2.71897  2.84104  1.67328  5.32720  4.10031      9 v - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03156  3.86736  4.58970  0.61958  0.77255  0.34406  1.23405
      2   2.62748  4.47174  3.31917  2.82619  3.63815  3.49607  2.75382  3.03401  2.75280  2.74783  3.65114  3.24714  2.62341  3.12082
  3.11124  2.79244  2.89355  1.88003  5.06315  3.77128     10 v - - -
```

Create a HMM with hmmbuild

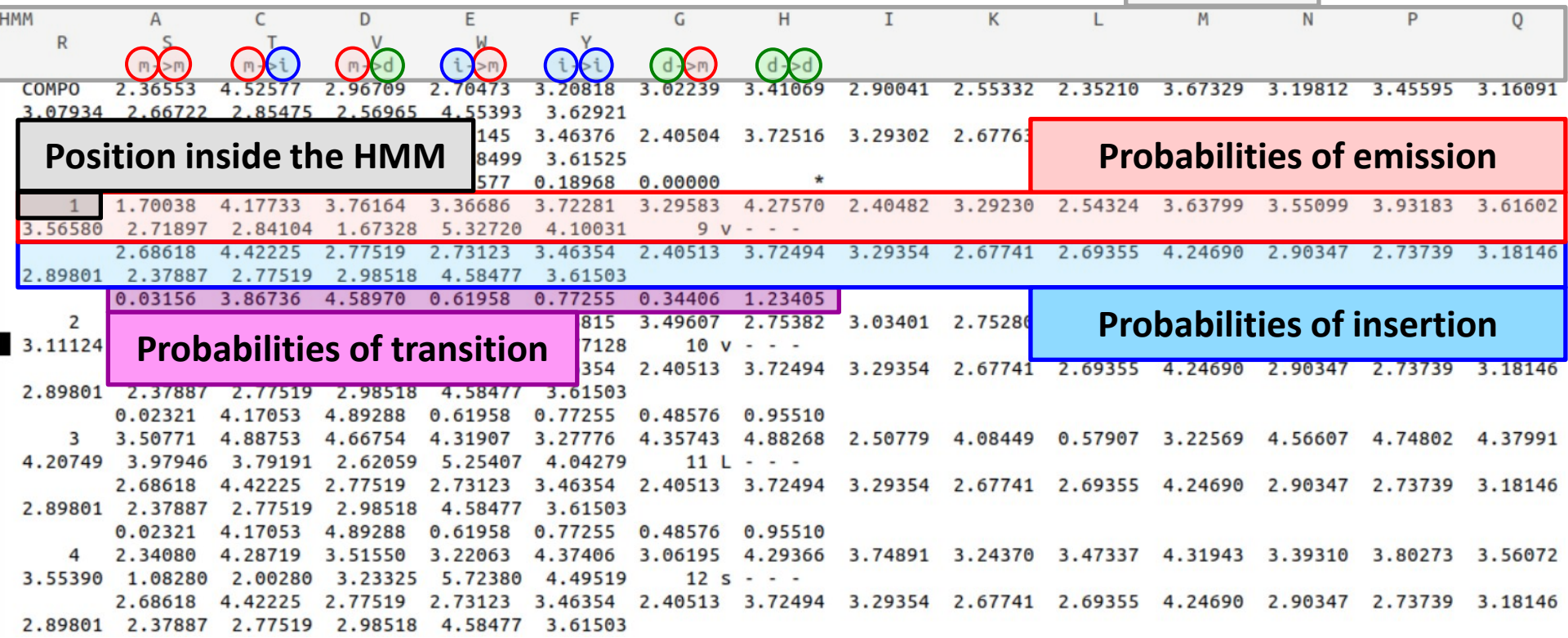
How does a HMM look from the inside?

```
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q
     R       S        T        V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.36553  4.52577  2.96709  2.70473  3.20818  3.02239  3.41069  2.90041  2.55332  2.35210  3.67329  3.19812  3.45595  3.16091
 3.07934  2.66722  2.85475  2.56965  4.55393  3.62921
          2.68640  4.42247  2.77497  2.73145  3.46376  2.40504  3.72516  3.29302  2.67763  2.69377  4.24712  2.90369  2.73719  3.18168
 2.89823  2.37879  2.77497  2.98431  4.58499  3.61525
          0.57544  1.78073  1.31293  1.75577  0.18968  0.00000        *
     1    1.70038  4.17733  3.76164  3.36686  3.72281  3.29583  4.27570  2.40482  3.29230  2.54324  3.63799  3.55099  3.93183  3.61602
 3.56580  2.71897  2.84104  1.67328  5.32720  4.10031      9 v - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
 2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03156  3.86736  4.58970  0.61958  0.77255  0.34406  1.23405
     2    2.62748  4.47174  3.31917  2.82619  3.63815  3.49607  2.75382  3.03401  2.75280  2.74783  3.65114  3.24714  2.62341  3.12082
 3.11124  2.79244  2.89355  1.88003  5.06315  3.77128     10 v - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
 2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02321  4.17053  4.89288  0.61958  0.77255  0.48576  0.95510
     3    3.50771  4.88753  4.66754  4.31907  3.27776  4.35743  4.88268  2.50779  4.08449  0.57907  3.22569  4.56607  4.74802  4.37991
 4.20749  3.97946  3.79191  2.62059  5.25407  4.04279     11 L - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
 2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02321  4.17053  4.89288  0.61958  0.77255  0.48576  0.95510
     4    2.34080  4.28719  3.51550  3.22063  4.37406  3.06195  4.29366  3.74891  3.24370  3.47337  4.31943  3.39310  3.80273  3.56072
 3.55390  1.08280  2.00280  3.23325  5.72380  4.49519     12 s - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146
 2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
```

Create a HMM with hmmbuild

How does a HMM look from the inside?

Legend



Position inside the HMM

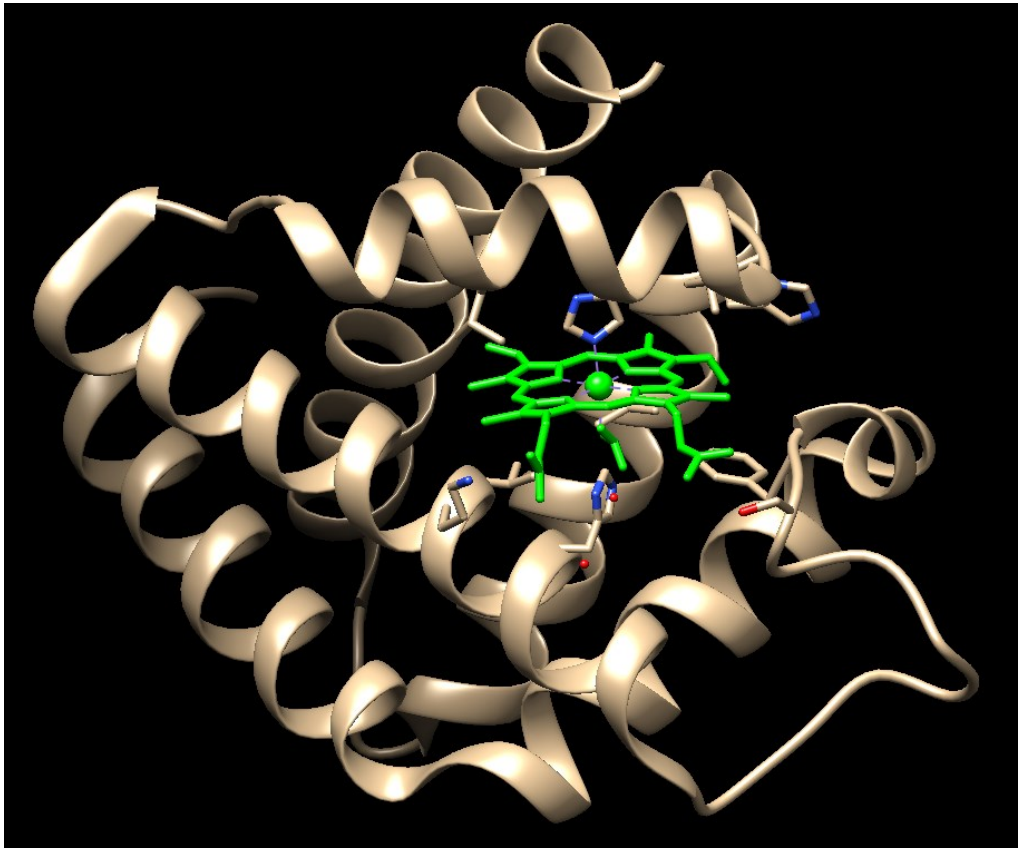
Probabilities of emission

Probabilities of transition

Probabilities of insertion

Create a HMM with hmmbuild

We created a HMM that is informative for the globin domain



Create a HMM with hmmbuild

It is common to use HMMs that are informative for specific protein domains

We can call them profiles

Find sequences using HMMs with hmmsearch

Search for templates using hmmsearch

```
hmmsearch  globins4.hmm  /mnt/NFS_UPF/soft/databases/blastdat/pdb_seq
>  globins_pdb.out
```

hmmsearch finds proteins in a database that match a HMM

Finds sequences that are likely to be produced by the input HMM

Find sequences using HMMs with hmmsearch

Take a look to the hmmsearch output

```
Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence Description
    ------- ------ -----    ------- ------ -----   ---- --  -------- -----------
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1abw_A    mol:protein length:283  HEMOGLOBIN-BASED BLOOD SUBS
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1aby_A    mol:protein length:283  HEMOGLOBIN
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1c7c_A    mol:protein length:283  PROTEIN (DEOXYHEMOGLOBIN (A
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1o1p_A    mol:protein length:283  Hemoglobin Alpha chain
     5e-119  396.4   8.1    8.1e-59  201.0   0.9    2.0  2  1c7d_A    mol:protein length:284  PROTEIN (DEOXYHEMOGLOBIN (A
   8.2e-117  389.2   8.0    1.1e-57  197.3   0.9    2.0  2  1o1n_A    mol:protein length:285  Hemoglobin Alpha chain
   1.6e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1j_A    mol:protein length:283  Hemoglobin Alpha chain
   1.7e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1m_A    mol:protein length:285  Hemoglobin Alpha chain
   5.7e-114  379.9   7.0    3.4e-56  192.4   0.7    2.0  2  1o1l_A    mol:protein length:283  Hemoglobin Alpha chain
    1.4e-65  222.9   3.3    1.6e-65  222.7   3.3    1.0  1  1cp5_A    mol:protein length:154  PROTEIN (MYOGLOBIN)
```

Find sequences using HMMs with hmmsearch

Take a look to the hmmsearch output

```
Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence Description
    -------  ------ -----   -------  ------ -----   ---- --  -------- -----------
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1abw_A    mol:protein length:283  HEMOGLOBIN-BASED BLOOD SUBS
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1aby_A    mol:protein length:283  HEMOGLOBIN
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1c7c_A    mol:protein length:283  PROTEIN (DEOXYHEMOGLOBIN (A
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1o1p_A    mol:protein length:283  Hemoglobin Alpha chain
     5e-119  396.4   8.1    8.1e-59  201.0   0.9    2.0  2  1c7d_A    mol:protein length:284  PROTEIN (DEOXYHEMOGLOBIN (A
   8.2e-117  389.2   8.0    1.1e-57  197.3   0.9    2.0  2  1o1n_A    mol:protein length:285  Hemoglobin Alpha chain
   1.6e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1j_A    mol:protein length:283  Hemoglobin Alpha chain
   1.7e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1m_A    mol:protein length:285  Hemoglobin Alpha chain
   5.7e-114  379.9   7.0    3.4e-56  192.4   0.7    2.0  2  1o1l_A    mol:protein length:283  Hemoglobin Alpha chain
    1.4e-65  222.9   3.3    1.6e-65  222.7   3.3    1.0  1  1cp5_A    mol:protein length:154  PROTEIN (MYOGLOBIN)
```

Why do we have different results for the full sequence and for the best domain?

Find sequences using HMMs with hmmsearch

Take a look to the hmmsearch output

```
Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence Description
    -------  ------ -----   -------  ------ -----   ---- --  -------- -----------
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1abw_A    mol:protein length:283  HEMOGLOBIN-BASED BLOOD SUBS
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1aby_A    mol:protein length:283  HEMOGLOBIN
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1c7c_A    mol:protein length:283  PROTEIN (DEOXYHEMOGLOBIN (A
   4.9e-119  396.4   8.1      8e-59  201.0   0.9    2.0  2  1o1p_A    mol:protein length:283  Hemoglobin Alpha chain
     5e-119  396.4   8.1    8.1e-59  201.0   0.9    2.0  2  1c7d_A    mol:protein length:284  PROTEIN (DEOXYHEMOGLOBIN (A
   8.2e-117  389.2   8.0    1.1e-57  197.3   0.9    2.0  2  1o1n_A    mol:protein length:285  Hemoglobin Alpha chain
   1.6e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1j_A    mol:protein length:283  Hemoglobin Alpha chain
   1.7e-114  381.7   7.3    1.7e-56  193.4   0.7    2.0  2  1o1m_A    mol:protein length:285  Hemoglobin Alpha chain
   5.7e-114  379.9   7.0    3.4e-56  192.4   0.7    2.0  2  1o1l_A    mol:protein length:283  Hemoglobin Alpha chain
    1.4e-65  222.9   3.3    1.6e-65  222.7   3.3    1.0  1  1cp5_A    mol:protein length:154  PROTEIN (MYOGLOBIN)
```

Why do we have different results for the full sequence and for the best domain?

→

Proteins can have more than one domain

Find domains using HMMs with hmmsearch

Search for fibronectin type-3 domains in a protein sequence using hmmsearch

```
hmmbuild  fn3.hmm  fn3.sto
```

```
hmmsearch  fn3.hmm  7LESS_DROME.fa  >  fn3.out
```

hmmsearch finds regions in protein sequences that match a HMM

Finds regions in the sequence that are likely to be produced by the input HMM

Find domains using HMMs with hmmsearch

Take a look to the hmmsearch output

```
Query:       fn3  [M=86]
Accession:   PF00041.13
Description: Fibronectin type III domain
Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence   Description
    ------- ------ -----    ------- ------ -----   ---- --  --------   -----------
    1.9e-57  178.0   0.4    1.2e-16   47.2   0.9    9.4  9  7LES_DROME  SEVENLESS PROTEIN (EC 2.7.1.112).

Domain annotation for each sequence (and alignments):
>> 7LES_DROME  SEVENLESS PROTEIN (EC 2.7.1.112).
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   -1.3   0.0      0.17      0.17      61      74 ..     396     409 ..     395     411 .. 0.85
   2 !   40.7   0.0   1.3e-14   1.3e-14       2      84 ..     439     520 ..     437     521 .. 0.95
   3 !   14.4   0.0     2e-06     2e-06      13      85 ..     836     913 ..     826     914 .. 0.73
   4 !    5.1   0.0    0.0016    0.0016      10      36 ..    1209    1235 ..    1203    1259 .. 0.82
   5 !   24.3   0.0   1.7e-09   1.7e-09      14      80 ..    1313    1380 ..    1304    1386 .. 0.82
   6 ?    0.0   0.0     0.063     0.063      58      72 ..    1754    1768 ..    1739    1769 .. 0.89
   7 !   47.2   0.9   1.2e-16   1.2e-16       1      85 [.    1799    1890 ..    1799    1891 .. 0.91
   8 !   17.8   0.0   1.8e-07   1.8e-07       6      74 ..    1904    1966 ..    1901    1976 .. 0.90
   9 !   12.8   0.0   6.6e-06   6.6e-06       1      86 []    1993    2107 ..    1993    2107 .. 0.89
```

Find domains using HMMs with hmmsearch

Take a look to the hmmsearch output

List of hits

```
Query:       fn3  [M=86]
Accession:   PF00041.13
Description: Fibronectin type III domain
Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence    Description
    ------- ------ -----    ------- ------ -----   ---- --  --------    -----------
    1.9e-57  178.0   0.4    1.2e-16   47.2   0.9    9.4  9  7LES_DROME  SEVENLESS PROTEIN (EC 2.7.1.112).
```

Results per domain

```
Domain annotation for each sequence (and alignments):
>> 7LES_DROME  SEVENLESS PROTEIN (EC 2.7.1.112).
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to     alifrom  ali to     envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   -1.3   0.0      0.17      0.17      61      74 ..     396     409 ..     395     411 .. 0.85
   2 !   40.7   0.0   1.3e-14   1.3e-14       2      84 ..     439     520 ..     437     521 .. 0.95
   3 !   14.4   0.0     2e-06     2e-06      13      85 ..     836     913 ..     826     914 .. 0.73
   4 !    5.1   0.0    0.0016    0.0016      10      36 ..    1209    1235 ..    1203    1259 .. 0.82
   5 !   24.3   0.0   1.7e-09   1.7e-09      14      80 ..    1313    1380 ..    1304    1386 .. 0.82
   6 ?    0.0   0.0     0.063     0.063      58      72 ..    1754    1768 ..    1739    1769 .. 0.89
   7 !   47.2   0.9   1.2e-16   1.2e-16       1      85 [.    1799    1890 ..    1799    1891 .. 0.91
   8 !   17.8   0.0   1.8e-07   1.8e-07       6      74 ..    1904    1966 ..    1901    1976 .. 0.90
   9 !   12.8   0.0   6.6e-06   6.6e-06       1      86 []    1993    2107 ..    1993    2107 .. 0.89
```

Find domains using HMMs with hmmsearch

Take a look to the hmmsearch output
(Results per domain section)

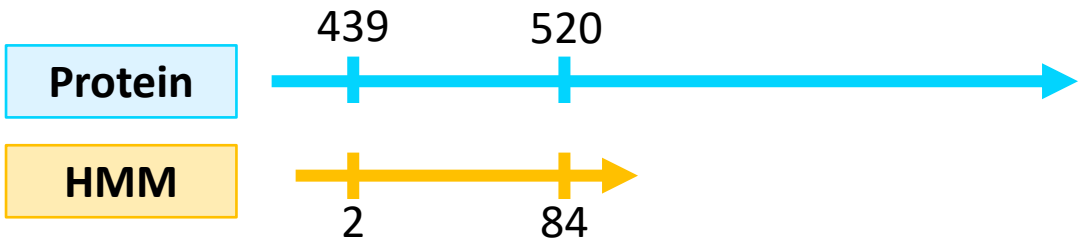
```
Domain annotation for each sequence (and alignments):
>> 7LES_DROME  SEVENLESS PROTEIN (EC 2.7.1.112).
```

#		score	bias	c-Evalue	i-Evalue	hmmfrom	hmm to		alifrom	ali to		envfrom	env to		acc
1	?	-1.3	0.0	0.17	0.17	61	74	..	396	409	..	395	411	..	0.85
2	!	40.7	0.0	1.3e-14	1.3e-14	2	84	..	439	520	..	437	521	..	0.95
3	!	14.4	0.0	2e-06	2e-06	13	85	..	836	913	..	826	914	..	0.73
4	!	5.1	0.0	0.0016	0.0016	10	36	..	1209	1235	..	1203	1259	..	0.82
5	!	24.3	0.0	1.7e-09	1.7e-09	14	80	..	1313	1380	..	1304	1386	..	0.82
6	?	0.0	0.0	0.063	0.063	58	72	..	1754	1768	..	1739	1769	..	0.89
7	!	47.2	0.9	1.2e-16	1.2e-16	1	85	[.	1799	1890	..	1799	1891	..	0.91
8	!	17.8	0.0	1.8e-07	1.8e-07	6	74	..	1904	1966	..	1901	1976	..	0.90
9	!	12.8	0.0	6.6e-06	6.6e-06	1	86	[]	1993	2107	..	1993	2107	..	0.89

Find domains using HMMs with hmmsearch

We can align HMMs with a protein sequence

hmmfrom	hmm to		alifrom	ali to
61	74	..	396	409
2	84	..	439	520
13	85	..	836	913
10	36	..	1209	1235
14	80	..	1313	1380
58	72	..	1754	1768
1	85	[.	1799	1890
6	74	..	1904	1966
1	86	[]	1993	2107



alifrom and ali to tell us where are the protein domains in our sequence

Find domains using HMMs with hmmsearch

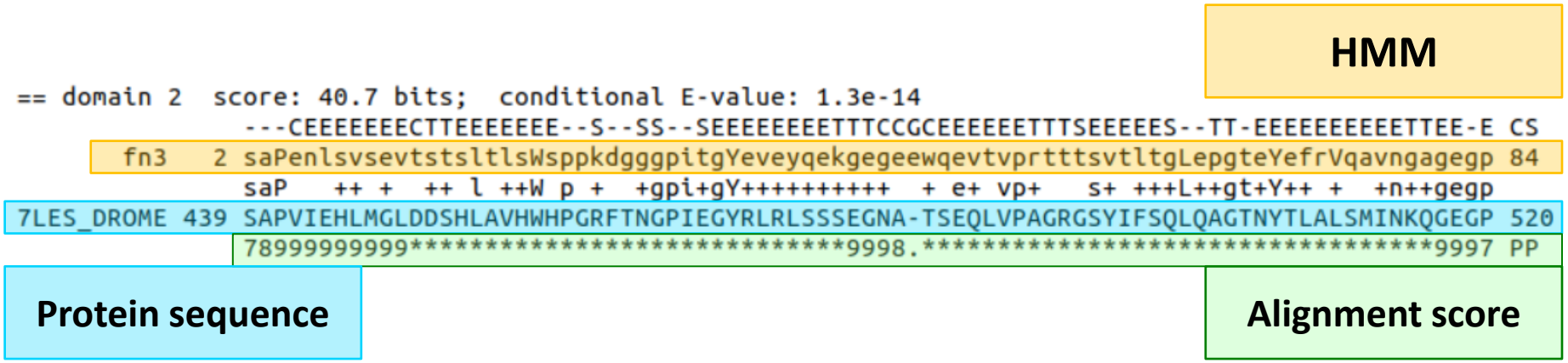
hmmsearch shows the alignment between the HMM and each of the domains

We can align HMMs with a protein sequence

```
== domain 2  score: 40.7 bits;  conditional E-value: 1.3e-14
               ---CEEEEEEECTTEEEEEEE--S--SS--SEEEEEEEETTTCCGCEEEEEETTTSEEEEES--TT-EEEEEEEEETTEE-E CS
       fn3   2 saPenlsvsevtstsltlsWsppkdgggpitgYeveyqekgegeewqevtvprtttsvtltgLepgteYefrVqavngagegp 84
               saP   ++ +  ++ l ++W p +  +gpi+gY++++++++++  + e+ vp+   s+ +++L++gt+Y++ +  +n++gegp
7LES_DROME 439 SAPVIEHLMGLDDSHLAVHWHPGRFTNGPIEGYRLRLSSSEGNA-TSEQLVPAGRGSYIFSQLQAGTNYTLALSMINKQGEGP 520
               78999999999**********************9998.**********************************9997 PP
```

Find domains using HMMs with hmmsearch

We can align HMMs with a protein sequence



HMMs contain probabilities of producing Aa on each position:

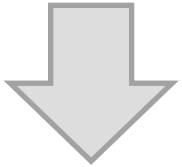
1. HMM position 1 aligns with Aa 1 in the protein sequence ➡ $score_1$
2. HMM position 2 aligns with Aa 2 in the protein sequence ➡ $score_2$

Etc…

Find HMMs that fit a sequence with hmmscan

There are several HMM databases on the internet

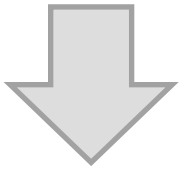
How can I know what HMM from a database fits my target sequence?



Using hmmscan

Find HMMs that fit a sequence with hmmscan

hmmscan finds what HMMs from a database match a protein sequence



- **What domains has my target sequence?**
- **Where are these domains in the sequence?**

Find HMMs that fit a sequence with hmmscan

Create a database of HMMs using hmmpress

```
hmmbuild  Pkinase.hmm  Pkinase.sto
```

Then, concatenate all the generated HMMs in one file:

```
cat  globins4.hmm  fn3.hmm  Pkinase.hmm  >  minifam
```

In order to check sequences and profiles very fast, we compress and index the database file using **hmmpress**. Here is a usage example:

- **hmmpress [database]**

We run then:

```
hmmpress  minifam
```

Find HMMs that fit a sequence with hmmscan

Execute hmmscan using this new database and the 7LES_DROME sequence

Now we can search what is the best profile for a given target sequence using the command hmmscan. Here is a usage example:

- **hmmscan (options) [Database_HMM] [sequence] > [output]**

For example we can use the sequence of 7LES_DROME to search for the best profile in the database previously generated. Run:

```
hmmscan minifam 7LESS_DROME.fa > 7LESS_DROME_minifam.out
```

Find HMMs that fit a sequence with hmmscan

Take a look to the hmmscan output

```
Query:       7LES_DROME  [L=2554]
Accession:   P13368
Description: SEVENLESS PROTEIN (EC 2.7.1.112).
Scores for complete sequence (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model    Description
    ------- ------ -----    ------- ------ -----   ---- --  -------- -----------
    5.6e-57  178.0   0.4    3.5e-16   47.2   0.9    9.4  9  fn3       Fibronectin type III domain
      3e-44  139.0   0.0    4.7e-44  138.3   0.0    1.3  1  Pkinase   Protein kinase domain
```

List of HMM matching our sequence

Find HMMs that fit a sequence with hmmscan

Take a look to the hmmscan output

Results per domain

```
Domain annotation for each model (and alignments):
>> fn3  Fibronectin type III domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   -1.3   0.0      0.33       0.5      61      74 ..     396     409 ..     395     411 .. 0.85
   2 !   40.7   0.0   2.6e-14   3.8e-14       2      84 ..     439     520 ..     437     521 .. 0.95
   3 !   14.4   0.0   4.1e-06   6.1e-06      13      85 ..     836     913 ..     826     914 .. 0.73
   4 !    5.1   0.0    0.0032    0.0048      10      36 ..    1209    1235 ..    1203    1259 .. 0.82
   5 !   24.3   0.0   3.4e-09     5e-09      14      80 ..    1313    1380 ..    1304    1386 .. 0.82
   6 ?    0.0   0.0      0.13      0.19      58      72 ..    1754    1768 ..    1739    1769 .. 0.89
   7 !   47.2   0.9   2.3e-16   3.5e-16       1      85 [.    1799    1890 ..    1799    1891 .. 0.91
   8 !   17.8   0.0   3.7e-07   5.5e-07       6      74 ..    1904    1966 ..    1901    1976 .. 0.90
   9 !   12.8   0.0   1.3e-05     2e-05       1      86 []    1993    2107 ..    1993    2107 .. 0.89
```

We already saw the results for the fibronectin type 3 domain

Find HMMs that fit a sequence with hmmscan

Take a look to the hmmscan output

Results per domain

```
>> Pkinase  Protein kinase domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  138.3   0.0   3.1e-44   4.7e-44       2     256 ..    2210    2479 ..    2209    2482 .. 0.85
```

Find HMMs that fit a sequence with hmmscan

Take a look to the hmmscan output

Alignment between HMM and domains

```
Alignments for each domain:
== domain 1  score: 138.3 bits;  conditional E-value: 3.1e-44
                 EEEEEEEEETTEEEEEEETTTTE....EEEEEEEEHHHCCCCCCHHHHHHHHHHHHHHSSSSB--EEEEEEETTEEEEEEE--TS-BHHHHHH.... CS
   Pkinase     2 elleklGsGsfGkVykakkktgk....kvAvKilkkeeekskkektavrElkilkklsHpnivkllevfetkdelylvleyveggdlfdllk.... 90
                 +ll+ lGsG+fG+Vy+++ k++ +    +vA+K l+k  ++  +     +E++++ +++H+niv+l++++ + +++ l++e++e+gdl ++l+
7LES_DROME 2210 KLLRFLGSGAFGEVYEGQLKTEDSeepqRVAIKSLRKGASEFAELL---QEAQLMSNFKHENIVRLVGICFDTESISLIMEHMEAGDLLSYLRaara 2303
                 67899*********887766554444448*******9998887764...4****************************************9988 PP

                 .......HHHST-HHHHHHHHHHHHHHHHHHHHTTEE-S--SGGGEEEETTTEE.......EE--GTT.E..EECSS-C-S--S..-GGGS-HHHHC CS
   Pkinase    91 .......kegklseeeikkialqilegleylHsngiiHrDLKpeNiLldkkgev.......kiaDFGlakklekssekllttlvg..treYmAPEvll 171
                        ls  e+  ++  +++g +yl +++++HrDL   N+L++++          ki DFGla+ ++ks+   ++  g    ++m+PE l
7LES_DROME 2304 tstqepqPTAGLSLSELLAMCIDVANGCSYLEDMHFVHRDLACRNCLVTESTGStdrrrtvKIGDFGLARDIYKSDYYRKEGEGllPVRWMSPESLV 2400
                 887766555666**********************************9554445999***********988887777666226 79******** PP

                 CS-CTHHHHHHHHHHHHHHHHH.SS-TTSSSHHCCTHHHHSSHHH......TTS.....HHHHHHHHHHT-SSGGGSTTHHHHHT CS
   Pkinase   172 kakeytkkvDvWslGvilyellt.gklpfsgeseedqleliekilkkkleedepkssskseelkdlikkllekdpakRltaeeilk 256
                  +   t+++DvW++Gv+++e+lt g+ p+  +   ++ e++++++++   ++ p  ++  e+l +l+  ++++dp +R++++++
7LES_DROME 2401 -DGLFTTQSDVWAFGVLCWEILTlGQQPYAAR---NNFEVLAHVKEGGRLQQ-PPMCT--EKLYSLLLLCWRTDPWERPSFRRCYN 2479
                 .9999***************999899999999...55666655555443333.33344..89*******************99887 PP
```

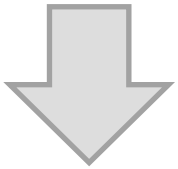
Comparing hmmsearch and hmmscan

The outputs of hmmsearch and hmmscan have the same organization

	hmmsearch	hmmscan
List of hits	Hits are protein sequences that fit the input HMM	Hits are HMMs that fit the input protein sequence
Results per domain	No difference	No difference
Alignments between HMMs and domains	No difference	No difference

Make MSAs with hmmalign

HMMs are more versatile tan substitution matrices



We can use HMMs to make MSAs

Make MSAs with hmmalign

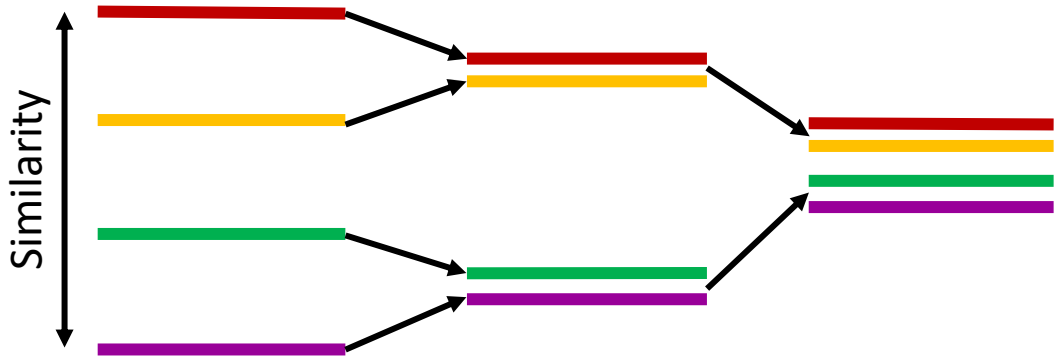
HMMs are better than agglomerative methods to make MSAs like clustalw

- **The alignment is made according with the specific information of the HMM**
- **The alignment is made faster**

Make MSAs with hmmalign

HMMs make alignments faster than clustalw

Clustalw

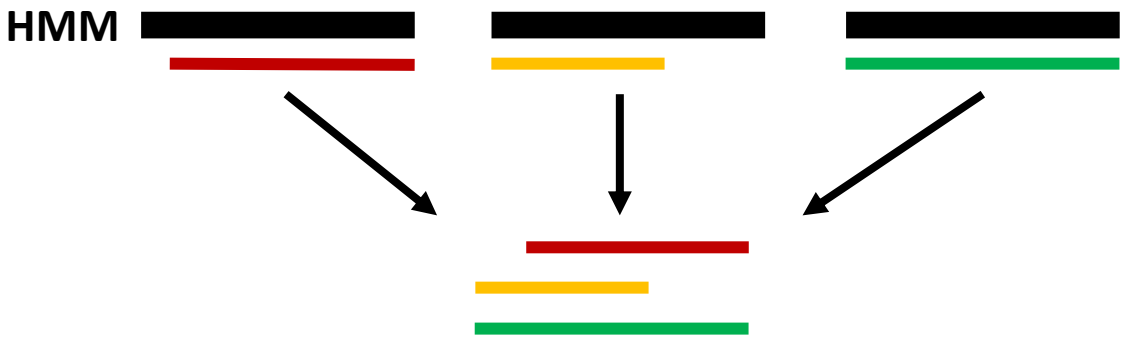


This takes a lot of time for a large number of sequences

Make MSAs with hmmalign

HMMs make alignments faster than clustalw

HMMs



Only one alignment per sequence

Make MSAs with hmmalign

Use hmmalign to make a MSA with globin sequences

- **hmmalign [model_HMM] [file_with_sequences] > [output]**

We can show this with the file globins45.fa. Run the following commands and test the speed of both approaches, hmmalign and clustalw:

```
hmmalign  globins4.hmm  globins45.fa  >  globins45_hmm.sto
```

```
clustalw  globins45.fa
```

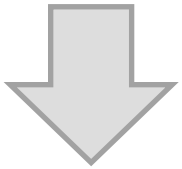
Make MSAs with hmmalign

Change the format of the output MSA

```
perl /mnt/NFS_UPF/soft/perl-lib/aconvertMod2.pl -in h -out c
<globins45_hmm.sto>globins45_hmm.clu
```

Find homologous proteins with phmmer and jackhmmer

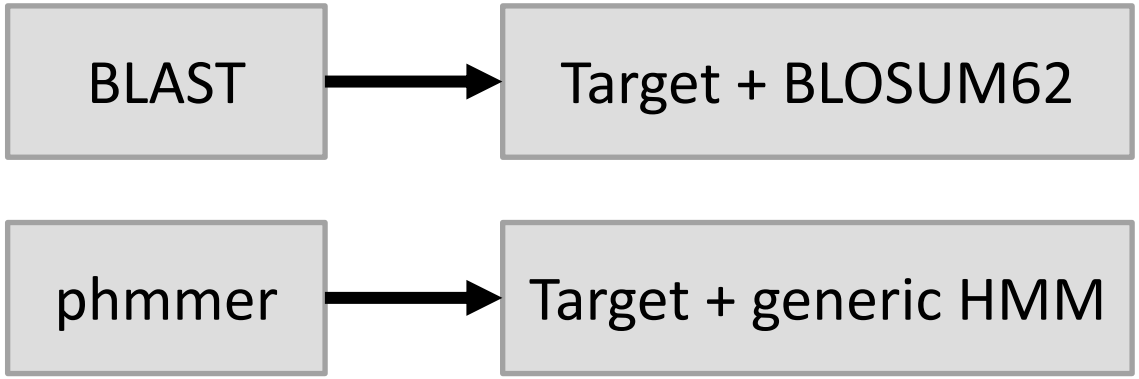
How can I find templates for my target using HMMs?



- **Using a HMM from the domain of my target**
- **Using phmmer or jackhmmer**

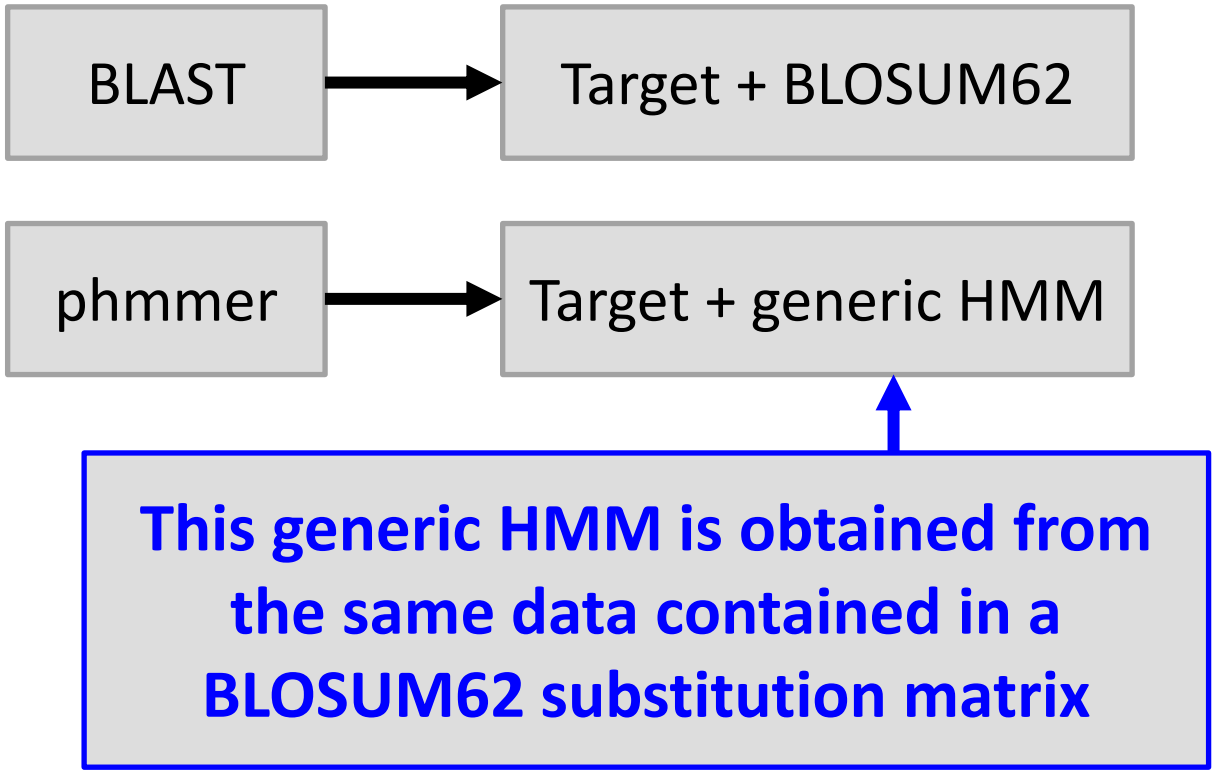
Find homologous proteins with phmmer and jackhmmer

phmmer is similar to BLAST



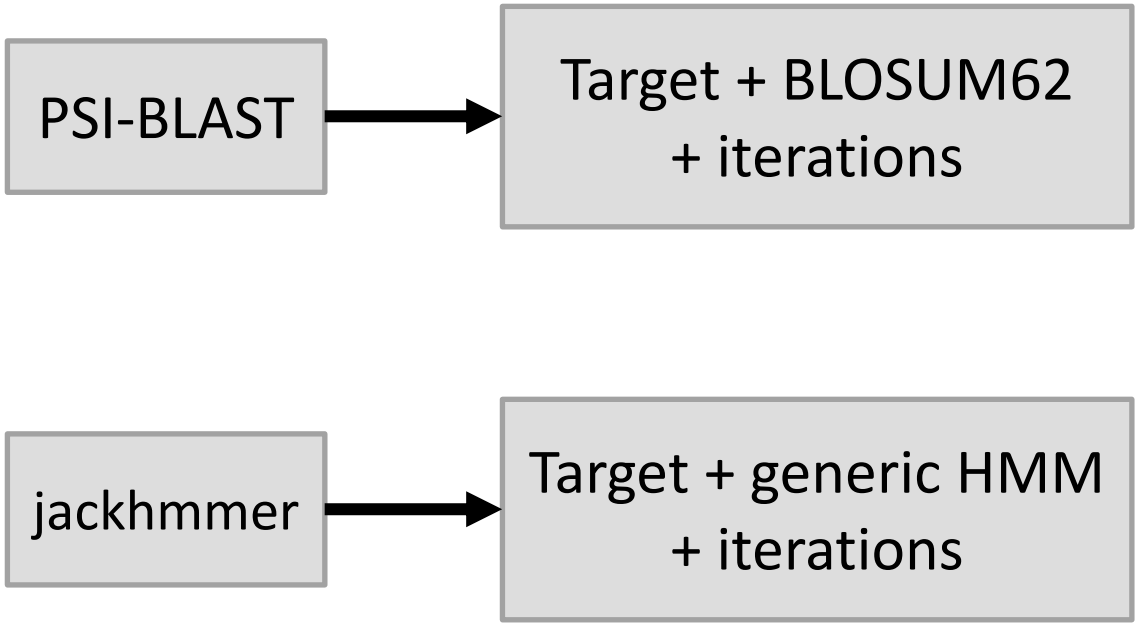
Find homologous proteins with phmmer and jackhmmer

phmmer is similar to BLAST



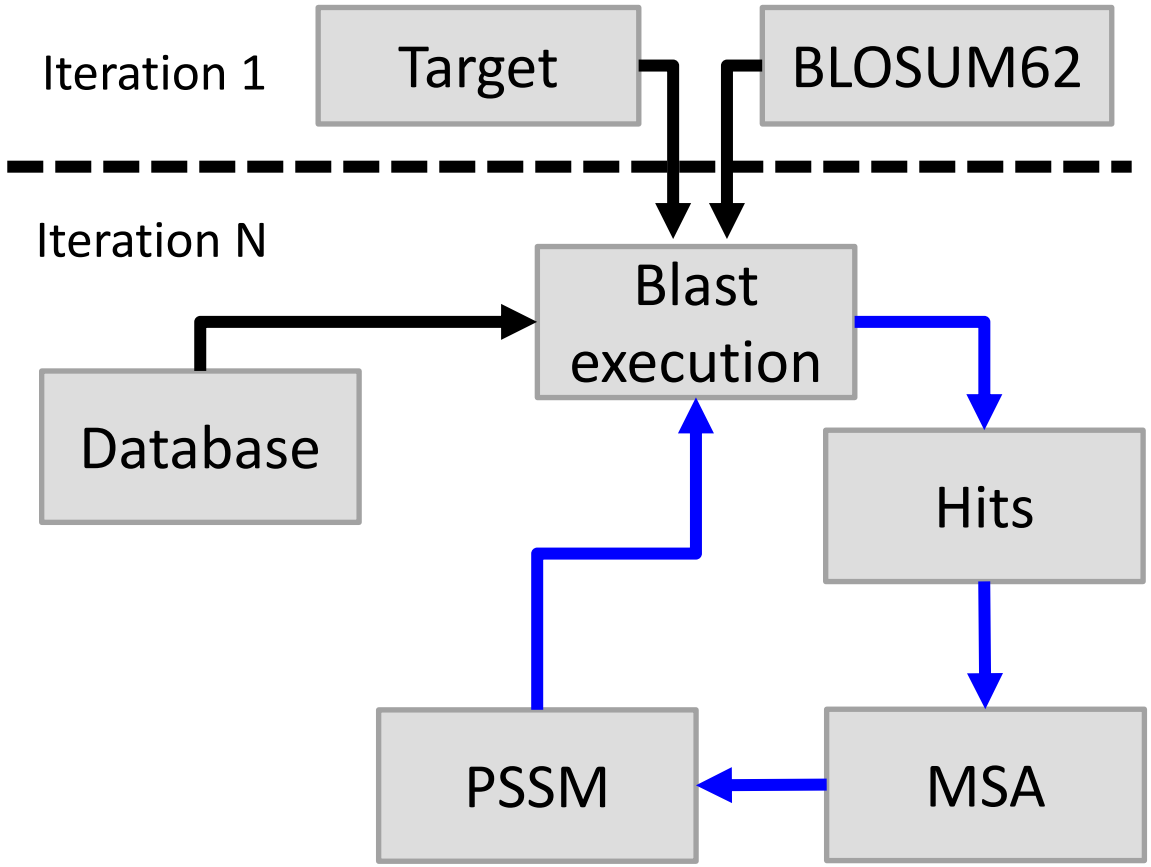
Find homologous proteins with phmmer and jackhmmer

jackhmmer is similar to PSI-BLAST



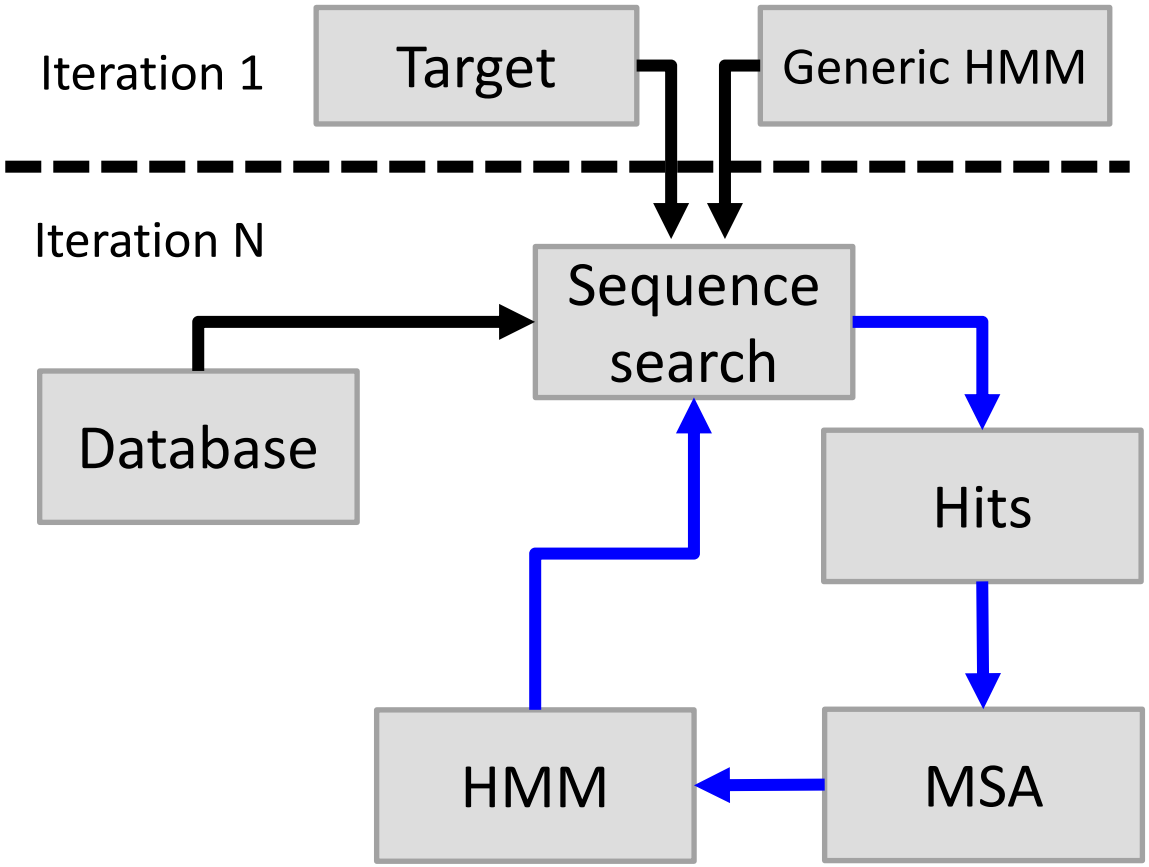
Find homologous proteins with phmmer and jackhmmer

PSI-BLAST creates a new PSSM at each iteration



Find homologous proteins with phmmer and jackhmmer

Jackhmmer creates a new HMM at each iteration



Find homologous proteins with phmmer and jackhmmer

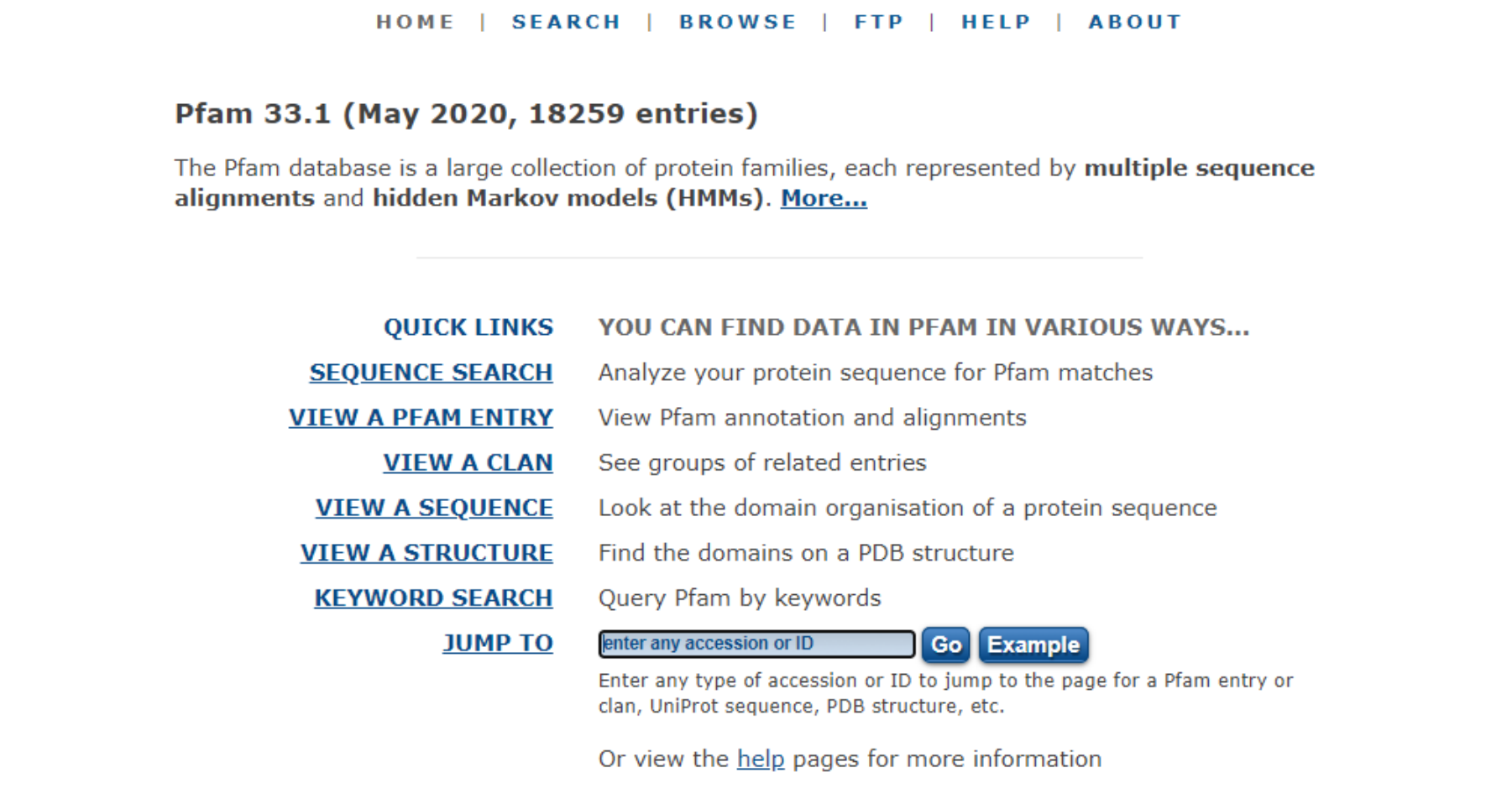
Execute phmmer and jackhmmer and compare the results

```
jackhmmer  hbb_human globins45.fa  >  globins_jackhmmer.out
```

```
phmmer  hbb_human  globins45.fa  >  globins_phmmer.out
```

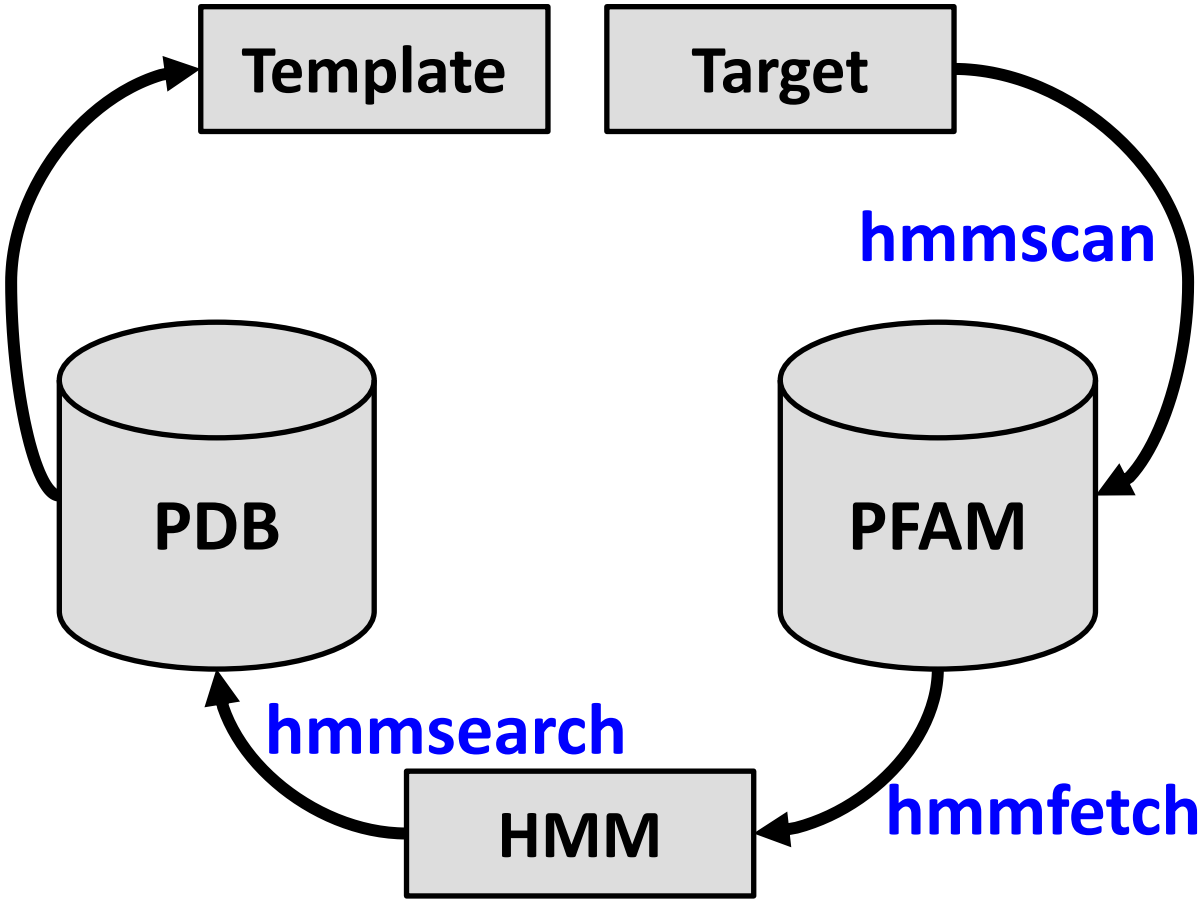
Using PFAM

PFAM is an extense and reliable database of HMMs



Using PFAM

How to use PFAM to find templates for our target?



Using PFAM

How to use PFAM to find templates for our target?

3 programs involved:

- **hmmscan:** finds what HMMs from a database match the input sequence
- **hmmfetch:** extracts a HMM from a database
- **hmmsearch:** finds what sequences from a database match the input HMM

Using PFAM

Execute hmmscan on the pfam database

```
hmmscan  /mnt/NFS_UPF/soft/databases/pfam-3/Pfam-A.hmm  hbb_human.fa
> hb_human_db.out
```

Using PFAM

Take a look to the hmmscan output

```
Query:       HBB_HUMAN  [L=146]
Description: Human beta hemoglobin.
Scores for complete sequence (score includes all domains):
   --- full sequence ---    --- best 1 domain ---     -#dom-
    E-value  score  bias     E-value  score  bias     exp  N  Model        Description
    ------- ------ -----     ------- ------ -----    ---- --  --------     -----------
   6.1e-30  103.6   0.0     4.5e-29  100.8   0.0     1.9  2  Globin        Globin
  ------ inclusion threshold ------
       0.12   11.9   0.7       0.35   10.4   0.0     2.1  2  BCA_ABC_TP_C  Branched-chain amino acid ATP-binding cassette


Domain annotation for each model (and alignments):
>> Globin  Globin
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  100.8   0.0   6.6e-33   4.5e-29       1     108 []       7     111 ..       7     111 .. 0.98
   2 ?    0.7   0.1     0.085   5.8e+02      52      72 ..     123     143 ..     116     145 .. 0.81
```

Using PFAM

Execute hmmfetch to extract the HMM we want from the PFAM database

- **hmmfetch [database_HMM] [name_HMM] > [file_HMM]**

Therefore, in our example, assuming we have found a domain_target:

Step 6.2) extract the profile(s) from PFAM that correspond to the domains of the target sequence which are found in the column indicated as "model" (see example in step 3 of this tutorial). Let's assume the name of the model we have found for hbb_human is "domain_hbb", then we execute the command:

```
hmmfetch  /mnt/NFS_UPF/soft/databases/pfam-3/Pfam-A.hmm  "domain_hbb"
                                    >  domain_hbb.hmm
```

Is this the name of the HMM that we want to get?

Using PFAM

Take a look to the hmmscan output

```
Query:       HBB_HUMAN  [L=146]
Description: Human beta hemoglobin.
Scores for complete sequence (score includes all domains):
   --- full sequence ---    --- best 1 domain ---     -#dom-
    E-value  score  bias     E-value  score  bias     exp  N  Model        Description
    ------- ------ -----     ------- ------ -----    ---- -- --------     -----------
   6.1e-30  103.6   0.0     4.5e-29  100.8   0.0     1.9  2  Globin        Globin
  ------ inclusion threshold ------
       0.12   11.9   0.7       0.35   10.4   0.0     2.1  2  BCA_ABC_TP_C  Branched-chain amino acid ATP-binding cassette


Domain annotation for each model (and alignments):
>> Globin  Globin
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  100.8   0.0   6.6e-33   4.5e-29       1     108 []       7     111 ..       7     111 .. 0.98
   2 ?    0.7   0.1     0.085   5.8e+02      52      72 ..     123     143 ..     116     145 .. 0.81
```

Using PFAM

Execute hmmfetch to extract the HMM we want from the PFAM database

- **hmmfetch [database_HMM] [name_HMM] > [file_HMM]**

Therefore, in our example, assuming we have found a domain_target:

Step 6.2) extract the profile(s) from PFAM that correspond to the domains of the target sequence which are found in the column indicated as “model” (see example in step 3 of this tutorial). Let’s assume the name of the model we have found for hbb_human is “domain_hbb”, then we execute the command:

```
hmmfetch  /mnt/NFS_UPF/soft/databases/pfam-3/Pfam-A.hmm
> domain_hbb.hmm
```

Globin

Using PFAM

Is this globins HMM different to the one we used at the beggining of the practice?

Using PFAM

Is this globins HMM different to the one we used at the beggining of the practice?

YES:

- **HMMs from the PFAM database are manually curated and very reliable**
- **The two HMMs are made with a different number of sequences**

Using PFAM

Execute hmmsearch to search for proteins containing the globin domain in the PDB

Step 6.3) Search for sequences with known structure that contain the same domain as our target using **hmmsearch**:

```
hmmsearch  domain_hbb.hmm  /mnt/NFS_UPF/soft/databases/blastdat/pdb_seq
                                > hbb_pdb_by_HMM.out
```

Programs from practical 1

BLAST	Find homologous sequences to a target
PSI-BLAST	Find homologous sequences to a target using iterations
clustalw	Make MSAs

Programs from practical 2

hmmbuild	Create HMMs from MSAs
hmmsearch	Find matches of a HMM in a database of sequences
hmmscan	Find matches of a sequence in a database of HMMs
hmmpress	Build a database of HMMs
hmmalign	Make MSAs using a HMM
phmmer	Find homologous sequences to a target
jackhmmer	Find homologous sequences to a target using iterations
hmmfetch	Extract a HMM from a database
aconvertMod2.pl	Change the format of MSAs

Databases from practicals 1 and 2

PDB

- Proteins with available structure
- Biased
- Redundant

SCOP

- Classification of protein structures (from the PDB) into domains

PFAM

- HMMs for protein domains

Uniprot (AKA SwissProt)

- Proteins with available sequence
- Non-biased
- Non-redundant

Exercises

You can try the exercises before the syncronic class

QUESTIONS FROM THE TUTORIAL

1) Compare the results of phmmer, jackhmmer with the results of hmmsearch using "domain_hbb.hmm" (see hbb_pdb_by_HMM.out) when searching homologs in pdb_seq for hbb_human.

2) If a protein sequence has more than one domain in PFAM, do you think the result of using hmmsearch and jackhmmer will be the same? Why? Test the example with 7LES_DROME in SwissProt.

3) In practice 2.1 we used PSI-BLAST to fish sequences in the database uniprot_sprot.fasta and generate a PSSM profile which was used for searching homologs in PDB. Check the manual of HMMER3.0 and create your own protocol in which you use the program jackhmmer in a similar approach: use SwissProt database to generate the HMM profile and perform the search in pdb_seq.

4) Use hmmscan to search the best model(s) for 7LES_DROME in PFAM and search the homologs in PDB with this/these model(s). Compare the results of this search with the results of your protocol search in question 3. What are the differences? Why?

5) Use your protocol described in question 3 to search homologs of 7LES_DROME in PDB and compare with the results of the protocol described in practice 2.1 when using PSI-BLAST.

6) Use the sequence target.fa from practice 2.1. Apply phmmer, jackhmmer and the protocols of questions 3 and 4 to find homologs in PDB. What's the fold of this sequence? Compare the result with the homologs found in practice 2.1

7) Use hmmalign and FetchFasta.pl to align the sequence of target.fa and its homologs of PDB

8) If you have to align the sequence 7LES_DROME and its homologs of PDB what's the best model to use? Produce the alignment with the models from question 4 and your protocol in question 3 to show your answer.

9) 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*

10) Find what are all the domains in the sequence 7LES_DROME. If you wanted to find templates for its Pkinase domain, what HMM would you choose and why?

Exercises

Exercise 10

```
Query:       7LES_DROME  [L=2554]
Accession:   P13368
Description: SEVENLESS PROTEIN (EC 2.7.1.112).
Scores for complete sequence (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model           Description
    ------- ------ -----    ------- ------ -----   ---- --  --------        -----------
    9.5e-92  306.6   0.0    1.6e-91  305.9   0.0    1.4  1  Pkinase_Tyr      Protein tyrosine kinase
    8.1e-52  173.1   0.6    3.9e-12   46.0   0.8    9.5  9  fn3              Fibronectin type III domain
    1.2e-40  139.2   0.0    1.8e-40  138.6   0.0    1.3  1  Pkinase          Protein kinase domain
     0.0047   16.8   0.0       0.17   11.7   0.0    3.5  4  Interfer-bind    Interferon-alpha/beta receptor, fibronectin
  ------ inclusion threshold ------
      0.054   13.4   0.1       0.24   11.3   0.1    2.1  2  CarboxypepD_reg  Carboxypeptidase regulatory-like domain
```

Exercises

Exercise 10

```
Query:       7LES_DROME  [L=2554]
Accession:   P13368
Description: SEVENLESS PROTEIN (EC 2.7.1.112).
Scores for complete sequence (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model          Description
    ------- ------ -----    ------- ------ -----   ---- --  --------       -----------
    9.5e-92  306.6   0.0    1.6e-91  305.9   0.0    1.4  1  Pkinase_Tyr     Protein tyrosine kinase
    8.1e-52  173.1   0.6    3.9e-12   46.0   0.8    9.5  9  fn3             Fibronectin type III domain
    1.2e-40  139.2   0.0    1.8e-40  138.6   0.0    1.3  1  Pkinase         Protein kinase domain
     0.0047   16.8   0.0       0.17   11.7   0.0    3.5  4  Interfer-bind   Interferon-alpha/beta receptor, fibronectin
  ------ inclusion threshold ------
      0.054   13.4   0.1       0.24   11.3   0.1    2.1  2  CarboxypepD_reg Carboxypeptidase regulatory-like domain
```

Exercises

Exercise 10

```
>> fn3  Fibronectin type III domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------     ----
   1 ?   -1.0   0.0      0.64   1.8e+03      60      73 ..     396     409 ..     395     411 .. 0.85
   2 !   40.9   0.0   5.5e-14   1.5e-10       1      83 [.     439     520 ..     439     521 .. 0.95
   3 !   14.4   0.0   9.8e-06     0.027      14      83 ..     838     912 ..     827     914 .. 0.72
   4 !    4.9   0.0    0.0094        26      10      35 ..    1210    1235 ..    1203    1259 .. 0.81
   5 !   23.4   0.0   1.5e-08   4.2e-05      13      79 ..    1313    1380 ..    1306    1385 .. 0.81
   6 ?    0.3   0.0      0.26   7.2e+02      57      72 ..    1754    1769 ..    1736    1769 .. 0.89
   7 !   46.0   0.8   1.4e-15   3.9e-12       1      84 [.    1800    1890 ..    1800    1891 .. 0.91
   8 !   18.0   0.0   7.4e-07     0.002       5      73 ..    1904    1966 ..    1901    1976 .. 0.90
   9 !    9.8   0.0   0.00027      0.73       1      85 []    1994    2107 ..    1994    2107 .. 0.87
```

Exercises

Exercise 10

```
>> Pkinase_Tyr  Protein tyrosine kinase
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  305.9   0.0   5.8e-95   1.6e-91       1     259 []    2209    2481 ..    2209    2481 .. 0.97

>> Pkinase  Protein kinase domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  138.6   0.0   6.7e-44   1.8e-40       2     256 ..    2210    2479 ..    2209    2482 .. 0.85
```

Why do we have two matches with HMMs that are informative for the Pkinase domain?

Exercises

Exercise 10

```
>> Pkinase_Tyr  Protein tyrosine kinase
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  305.9   0.0   5.8e-95   1.6e-91       1     259 []    2209    2481 ..    2209    2481 .. 0.97

>> Pkinase  Protein kinase domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  138.6   0.0   6.7e-44   1.8e-40       2     256 ..    2210    2479 ..    2209    2482 .. 0.85
```

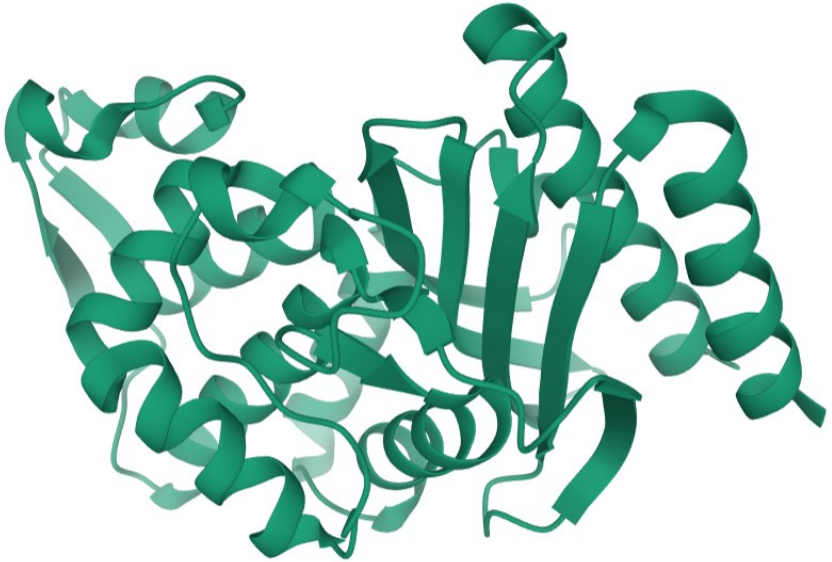
The two HMMs are recognizing the same domain: we select the HMM that recognizes this domain with best E-values

Structural biology

Practice 4: Statistical potentials and PROSA

Course 2023-2024

Practice 4: Statistical potentials and PROSA



Is this model correct?

Practice 4: Statistical potentials and PROSA

Is this model correct?

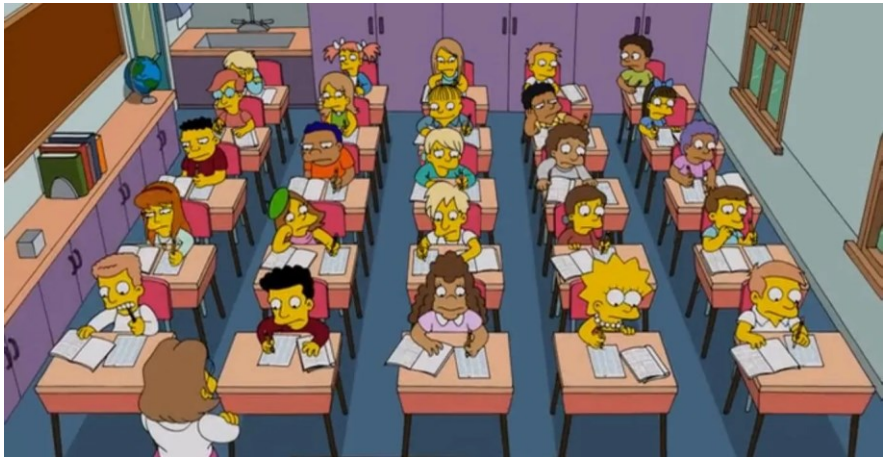
We can use statistical potentials to answer this question

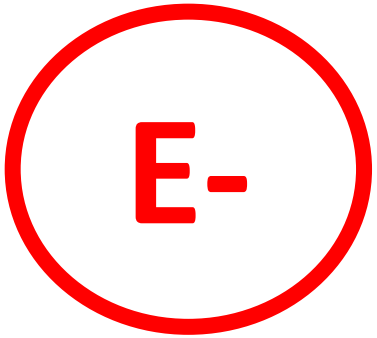
Practice 4: Statistical potentials and PROSA

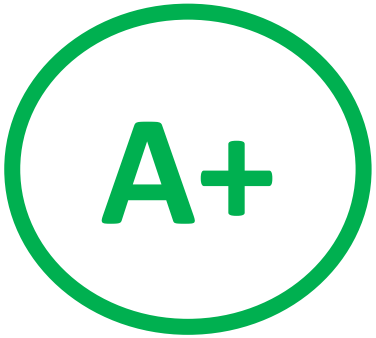
Statistical potentials are scoring functions that are derived from the analysis of experimental structures

Practice 4: Statistical potentials and PROSA

Statistical potentials are scoring functions that are derived from the analysis of experimental structures

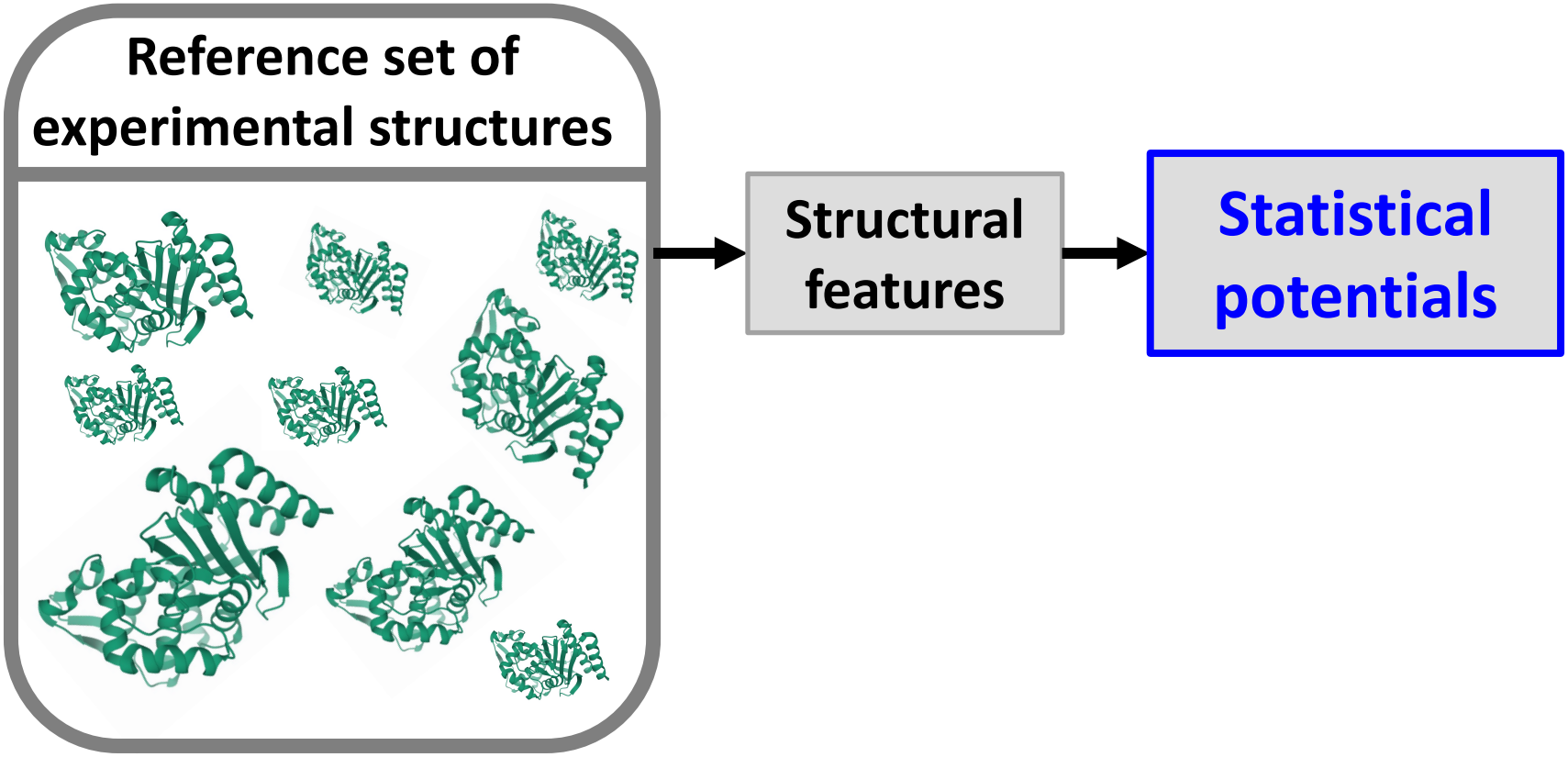




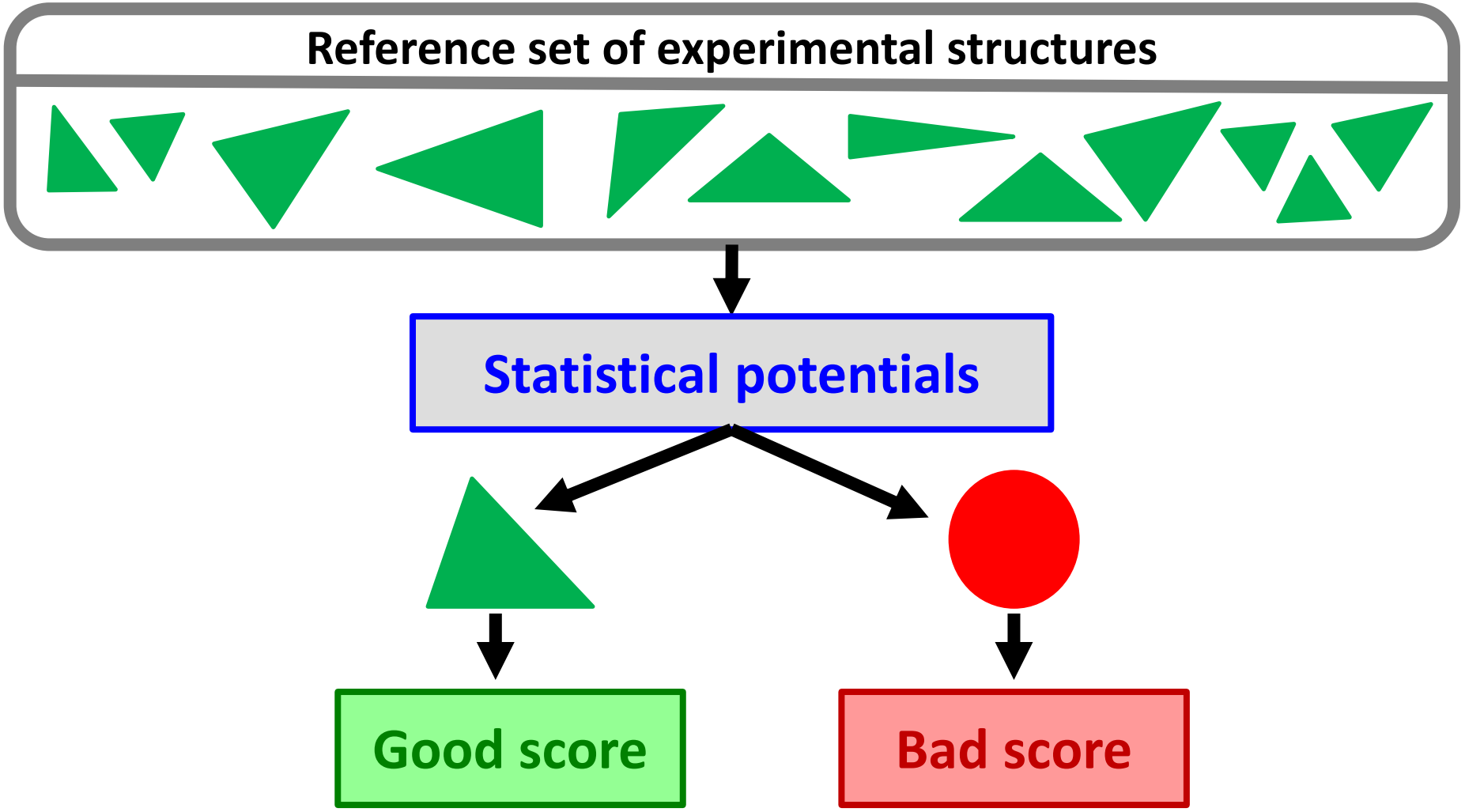


Practice 4: Statistical potentials and PROSA

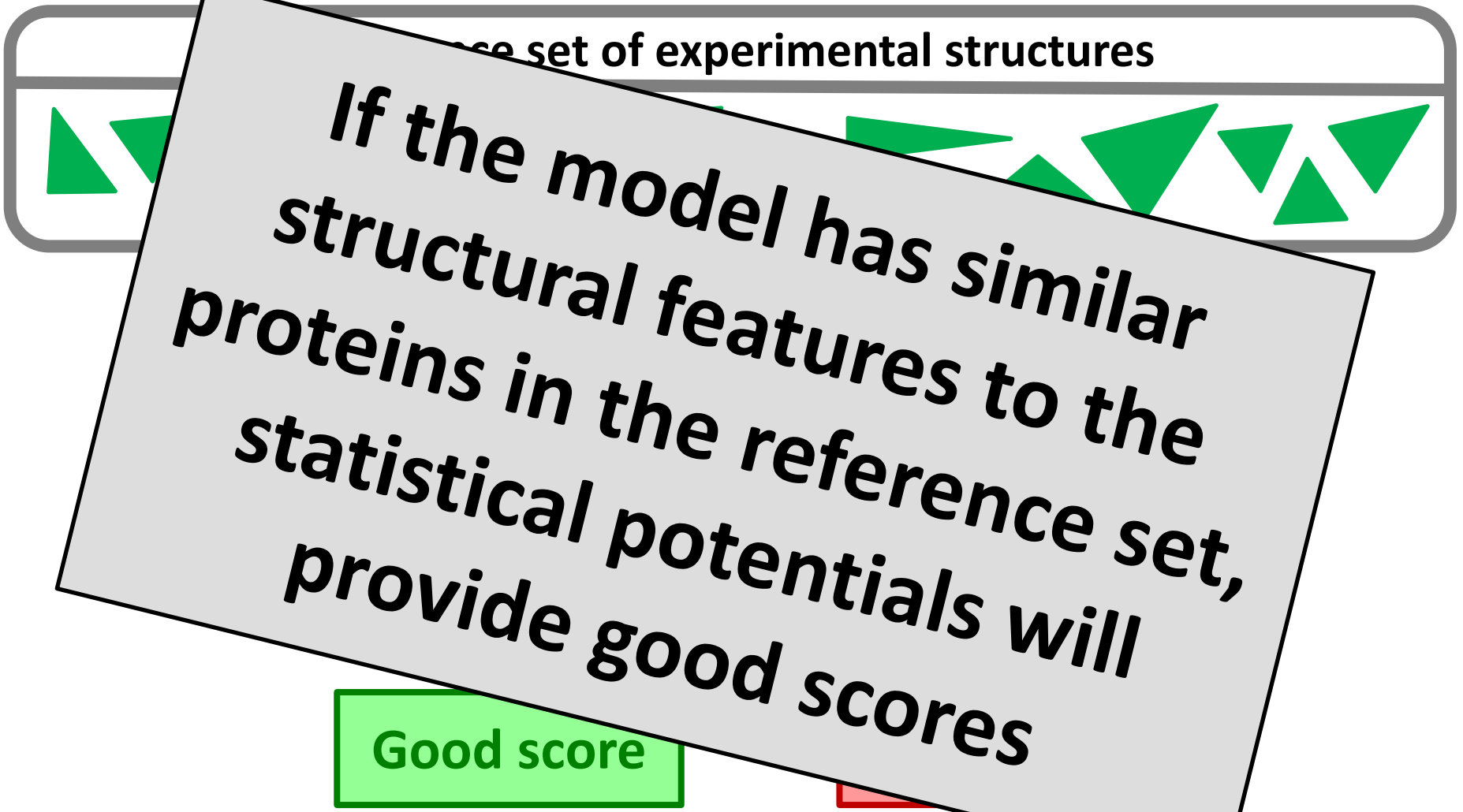
Statistical potentials are scoring functions that are derived from the analysis of experimental structures



Practice 4: Statistical potentials and PROSA



Practice 4: Statistical potentials and PROSA



Practice 4: Statistical potentials and PROSA

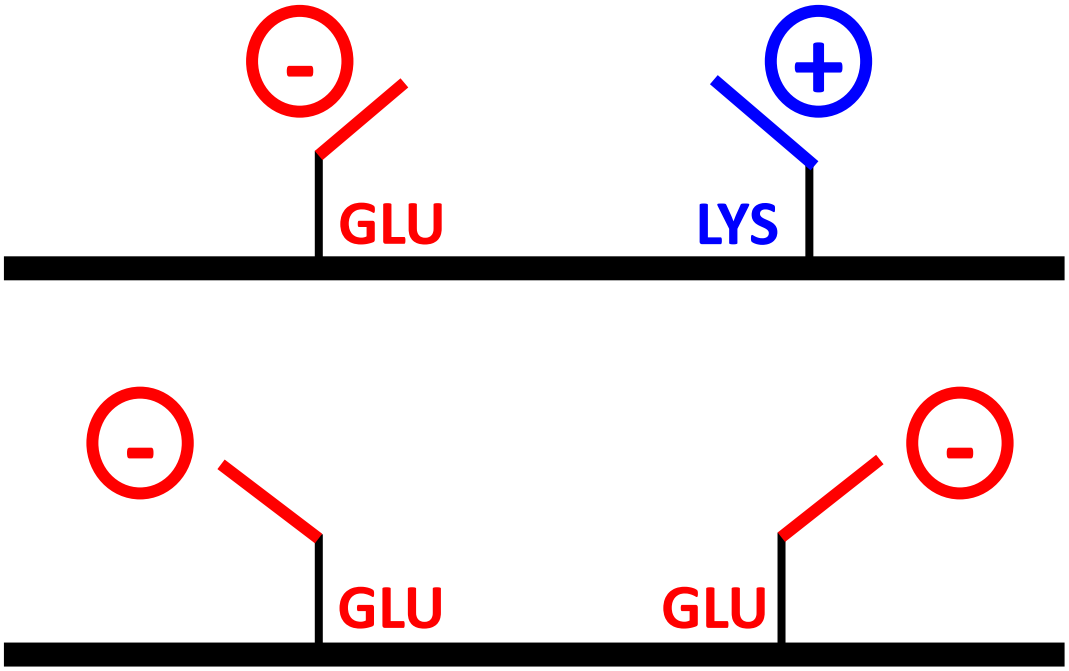
What are the structural features that statistical potentials use?

Amino acid contacts

Amino acid exposure

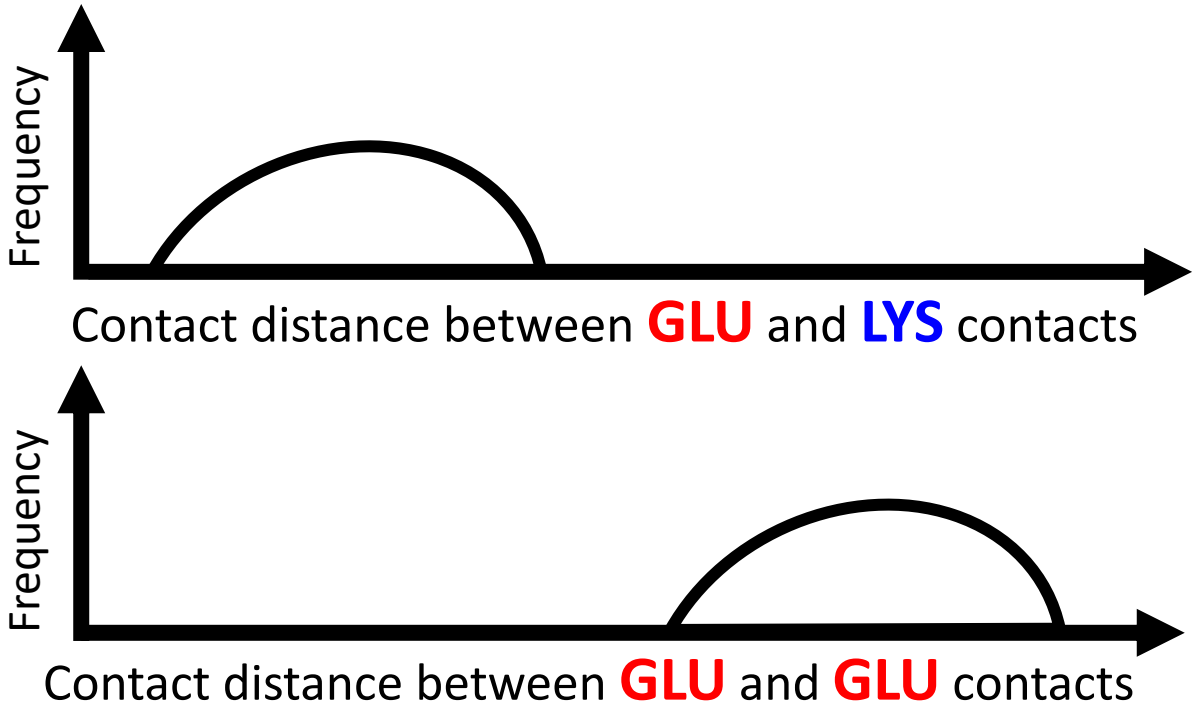
Practice 4: Statistical potentials and PROSA

Amino acid contacts



Practice 4: Statistical potentials and PROSA

Amino acid contacts



Practice 4: Statistical potentials and PROSA

Amino acid contacts

Amino acid charge is one of the factors that affect distances between pairs of amino acids

and **GLU** contacts

Practice 4: Statistical potentials and PROSA

What atom would you use to meaure distances between amino acids?

Practice 4: Statistical potentials and PROSA

What atom would you use to meaure distances between amino acids?

Beta carbon
(first carbon in the side chain)

This is a way of including information regarding side chain orientation into the potentials

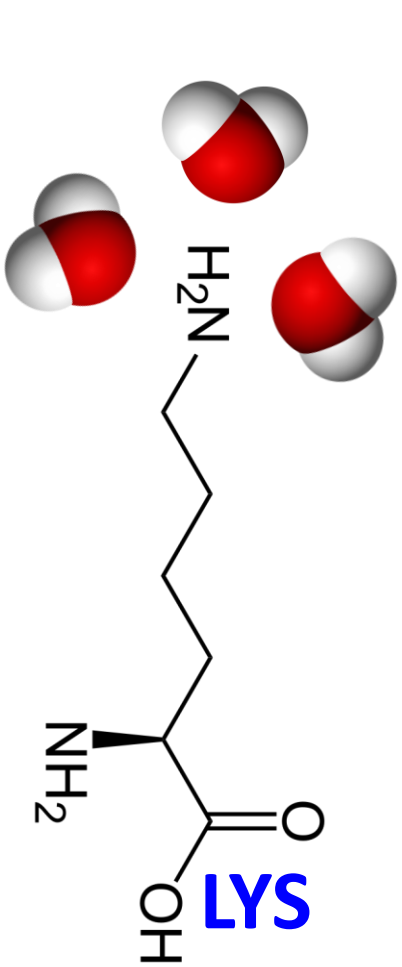
Practice 4: Statistical potentials and PROSA

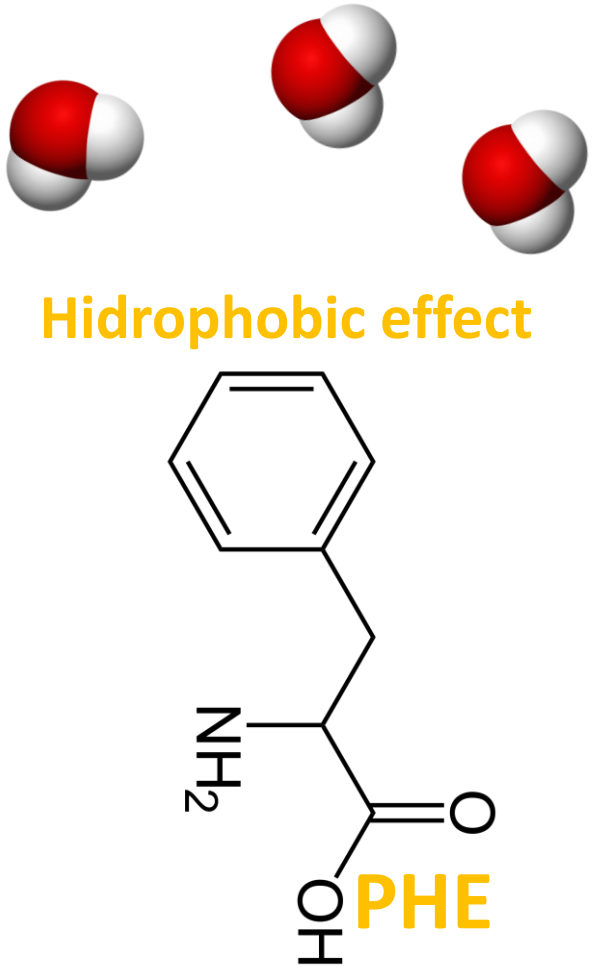
Amino acid exposure

Polar and charged are more likely to be exposed because of their tendency to interact with water molecules

Practice 4: Statistical potentials and PROSA

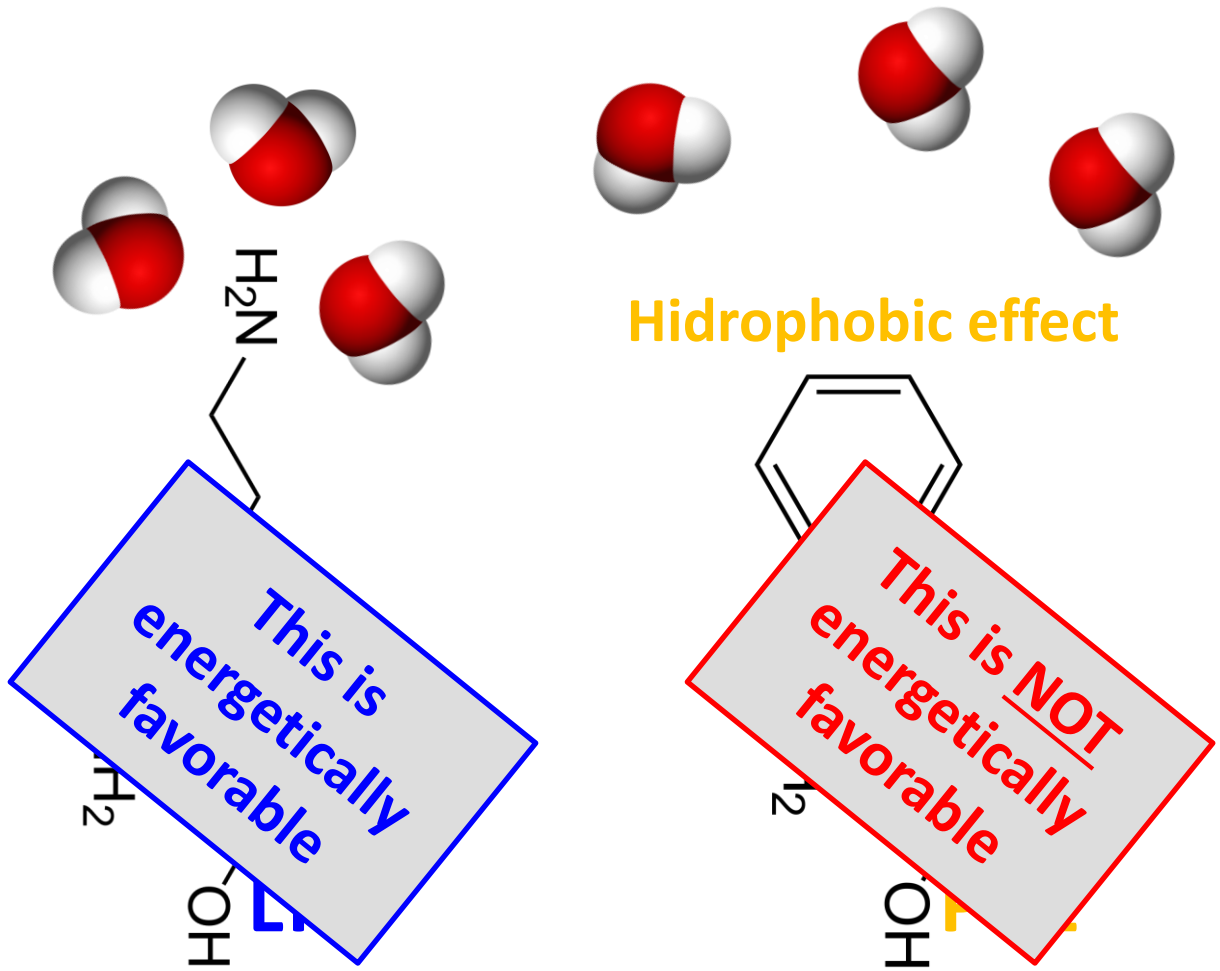
Amino acid exposure





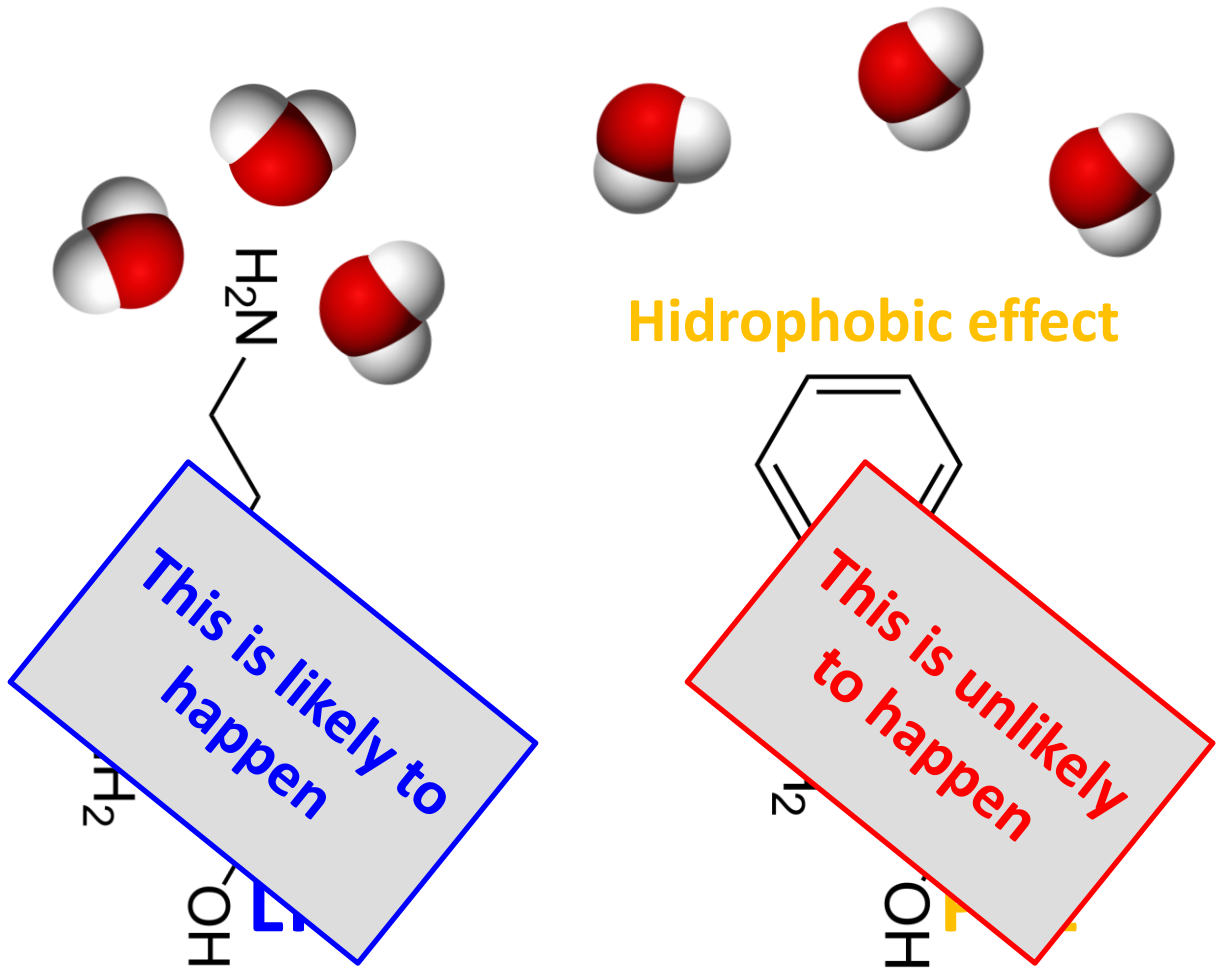
Practice 4: Statistical potentials and PROSA

Amino acid exposure



Practice 4: Statistical potentials and PROSA

Amino acid exposure



Practice 4: Statistical potentials and PROSA

Statistical potentials are computed using formulas coming from statistical thermodynamics

Boltzmann Law

$$P=(1/z)(e^{(-E/kT)})$$

Practice 4: Statistical potentials and PROSA

Statistical potentials are computed using formulas coming from statistical thermodynamics

Boltzmann Law

$$P = (1/z)(e^{(-E/kT)})$$

Probability

Energy

Practice 4: Statistical potentials and PROSA

Statistical potentials are computed using formulas coming from statistical thermodynamics

Boltzmann Law

$$P=(1/z)(e^{(-E/kT)})$$

High Probability	↔	Low Energy
Low Probability	↔	High Energy

Practice 4: Statistical potentials and PROSA

Statistical potentials are computed using formulas coming from statistical thermodynamics

Boltzmann Law

$$P = (1/z)(e^{(-E/kT)})$$

$$E = -KT\ln P + KT\ln Z$$

By operating with Boltzmann Law's equation we can isolate the energy

Practice 4: Statistical potentials and PROSA

St potentials are computed using formulas
statistical thermodynamics

Remember that probabilities come from our reference set of structures

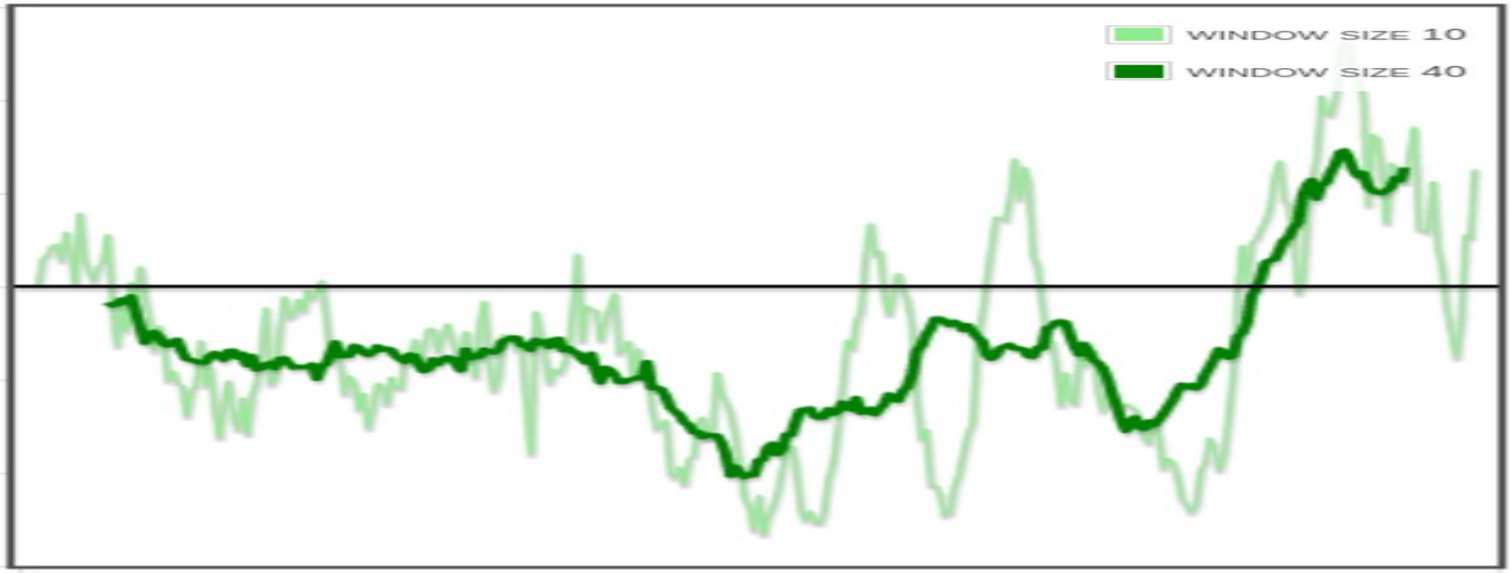
equation we

Practice 4: Statistical potentials and PROSA

We obtain one energy value for each amino acid and display it as a energy profile

$$P = (1/z)(e^{(-E/kT)})$$

$$E = -KT\ln P + KT\ln Z$$



Practice 4: Statistical potentials and PROSA

Statistical potentials are relative measurements

We use them to compare structures between them

Practice 4: Statistical potentials and PROSA

If you want to test the quality of your model, what structure would you compare with using statistical potentials?

The template you used in the modeling

Practice 4: Statistical potentials and PROSA

If you want to test the quality of your model, what structure would you compare with using statistical potentials?

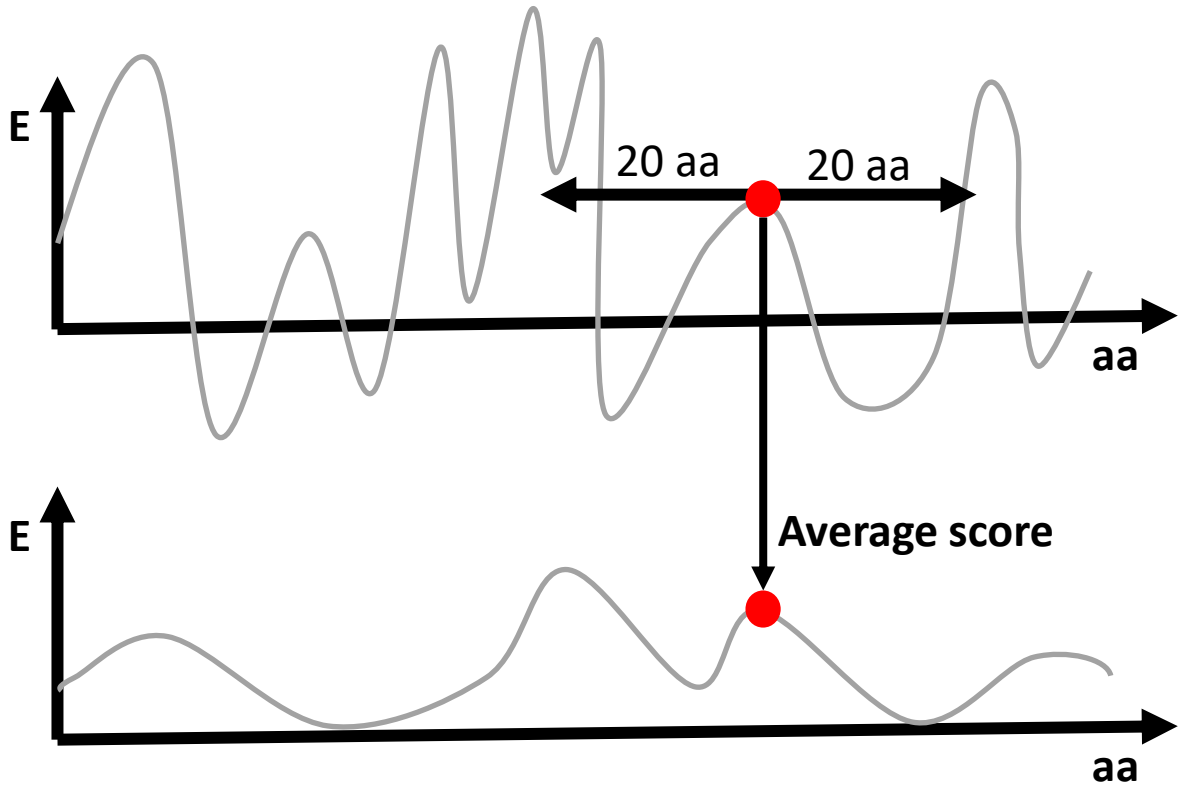
The template you used in the modeling:

- Is an experimental structure, therefore is a reference of what is right
- Is a similar protein to the one you modeled

Practice 4: Statistical potentials and PROSA

Using sliding windows

Sliding window = 40



Practice 4: Statistical potentials and PROSA

Working with Z-scores

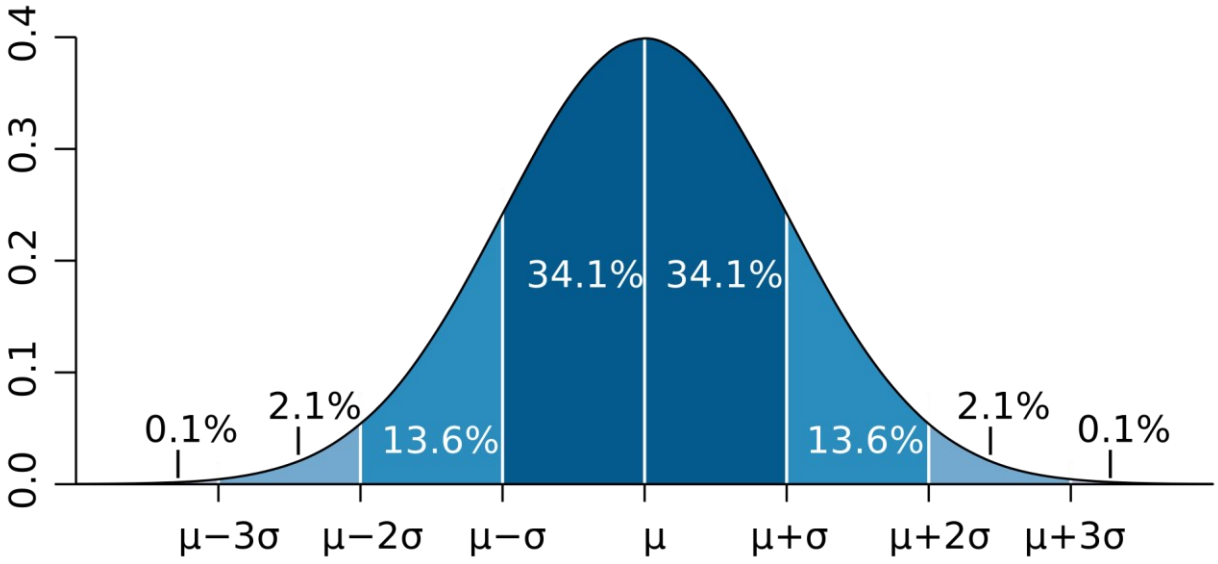
$$Z = \frac{(X - \ddot{X})}{SD}$$

How good am I in comparison with a reference distribution?

Practice 4: Statistical potentials and PROSA

Working with Z-scores

$$Z = \frac{(X - \ddot{X})}{SD}$$



Practice 4: Statistical potentials and PROSA

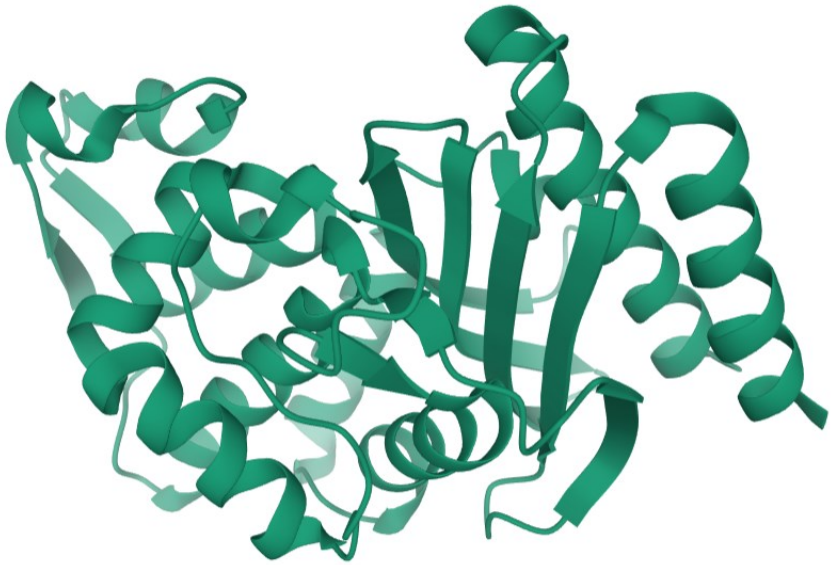
You created several models, how would you use PROSA to know what model is the best?

Practice 4: Statistical potentials and PROSA

You created several models, how would you use PROSA to know what model is the best?

You can compare the models by themselves and choose the one with best scores, you don’t need to compare to a template.

Practice 4: Statistical potentials and PROSA



Is this model correct?

Practice 4: Statistical potentials and PROSA

Is this model correct?

What does correct mean?

It depends on what question you want to answer with your model

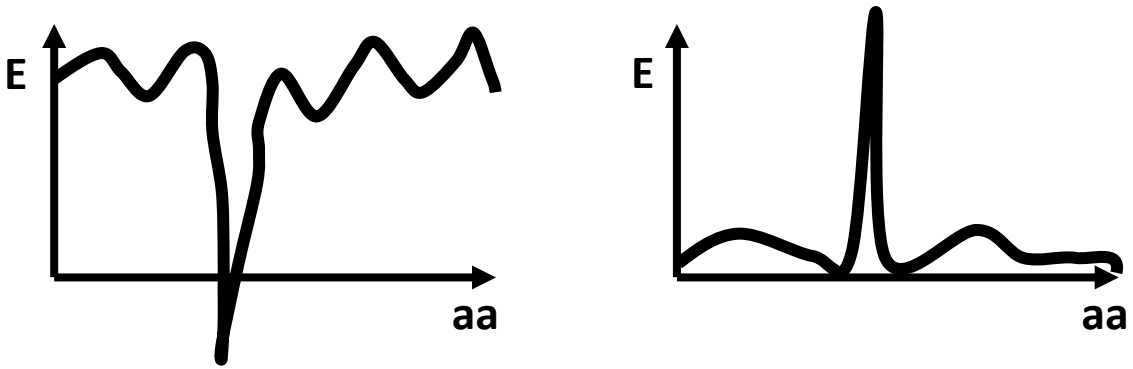
Practice 4: Statistical potentials and PROSA

It depends on what question you want to answer with your model

- **Am I interested in the model of the whole structure?**
- **Am I interested only in one region of the model?**
- **Am I interested only in one function of the model?**

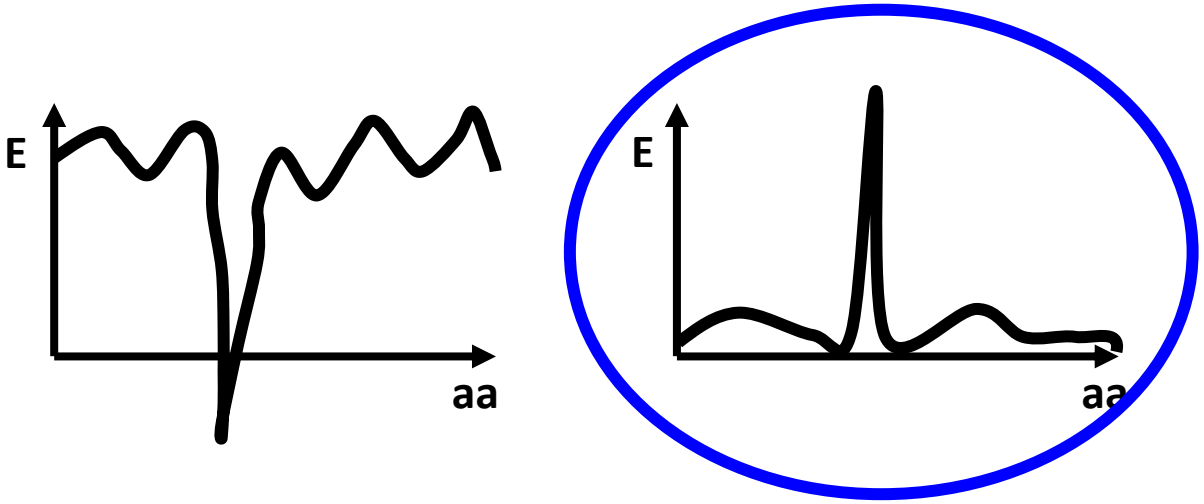
Practice 4: Statistical potentials and PROSA

What model is better?



Practice 4: Statistical potentials and PROSA

What model is better?

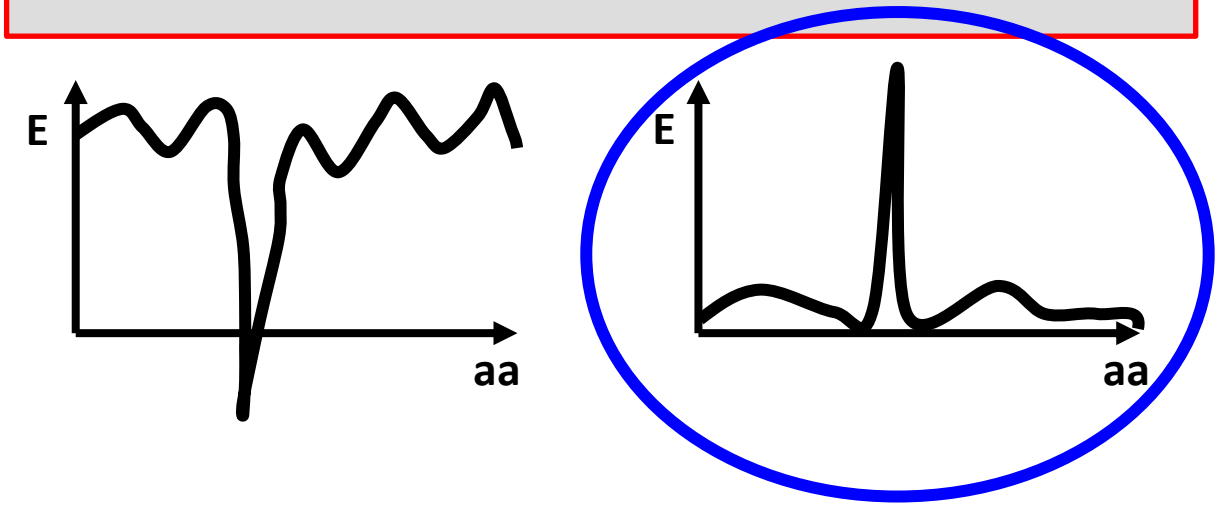


The second model only has one region with bad energies, while the rest of the model is alright. The wrong region could be corrected.

The first model only has one region right and the rest is wrong. Since most of the model is wrong it cannot be corrected. The best thing would be to make a new model.

Practice 4: Statistical potentials and PROSA

What model is better?



The second model only has one region with bad energies, while the rest of the model is alright. The wrong region could be corrected.

The first model only has one region right and the rest is wrong. Since most of the model is wrong it cannot be corrected. The best thing would be to make a new model.