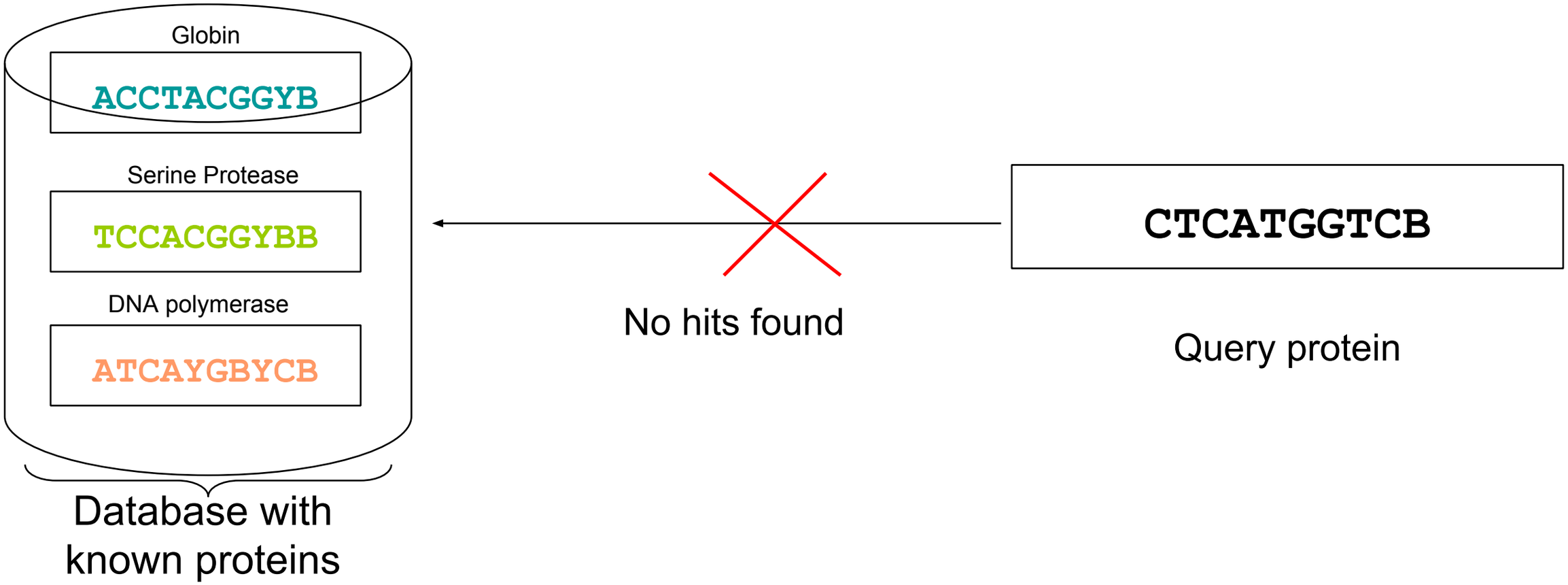
Homology Detection

Remote Homology Detection

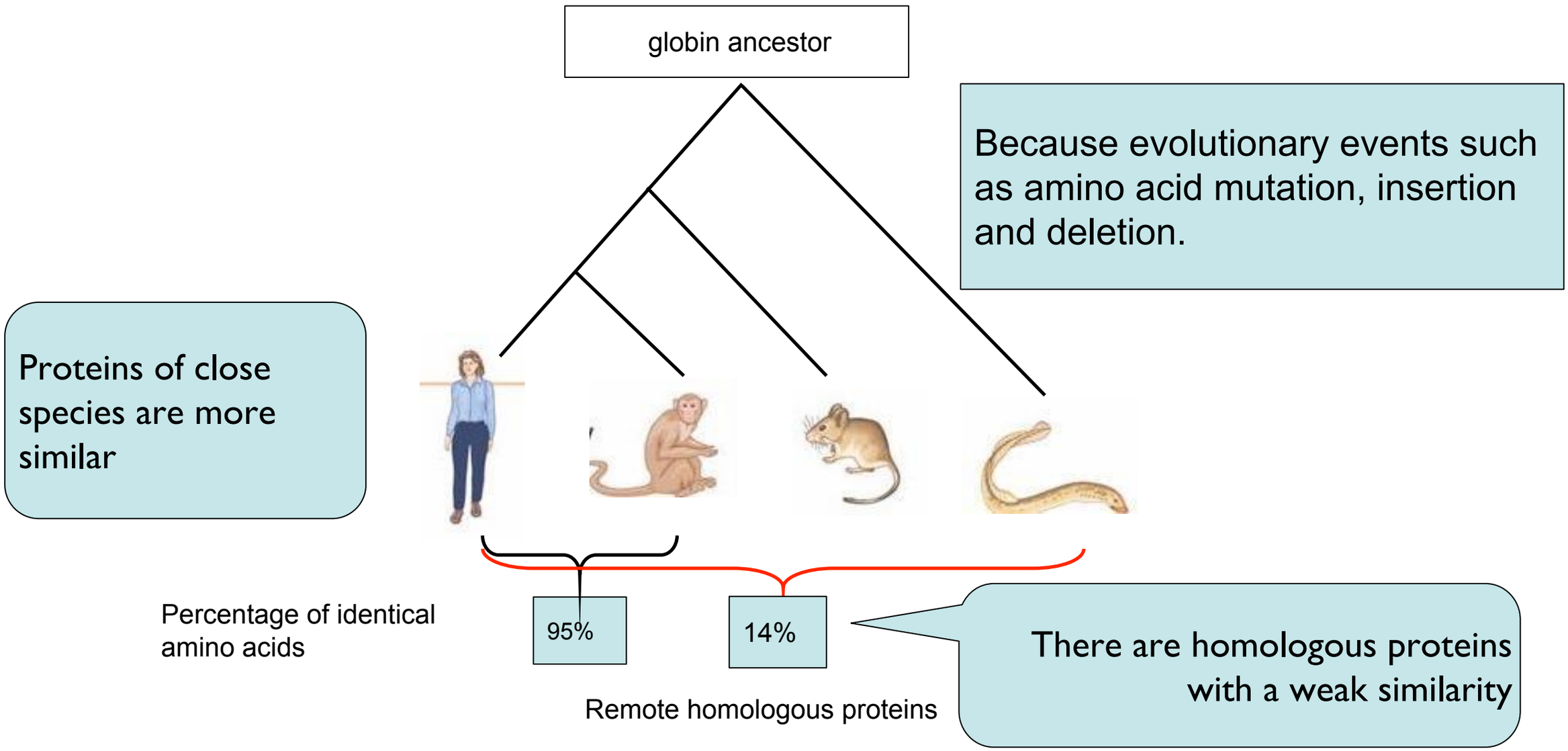


Remote Homology Detection

Why no hit is found?
Really, there is no hit in the database.
Remote homology detect methods are not efficient

60% of *P. faciparum* genes have not known function

Remote Homology Proteins



Remote Homology Detection

Which pairs of proteins are homologs?

```
1fleI   AQEPVKGPVSTKPGSCPILIRCAMINPPN----RCLKDTDCPGIKKCCEGSCG-MACFVP-Q
1udkA   ----------NEKSGSCPDMSM---PIPPLGICKTLCNSDSGCPNVQKCCKNGCGSMTCTTPVP
                  ****          *       *  *  **   ***   ** * *  *
```

To detect homology in remote homologous proteins is hard when we use only sequence properties.

```
1fleI -AQEPVKGPVSTKPGSCPILIRCAMINPPNRCLKD---TDCPG-IKKCCEGSCGMACFVPQ----
1a0aA MKRESHKHAEQARRNRAVALHEMASMPAEWKQQNNVSSAAPSKAATTVEAACRMRHLQQNGST
          *  *              *   * * *              *        *  *      *
```

Hydrophobic aa are known to play an important role in protein structural stability

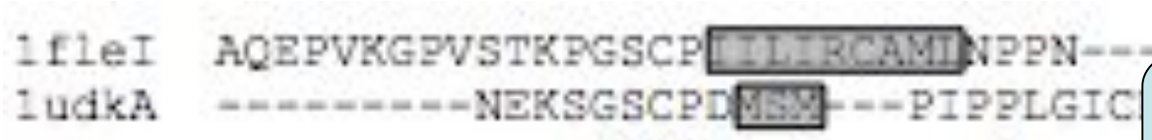
physico-chemical properties are conserved

* Identical amino acid

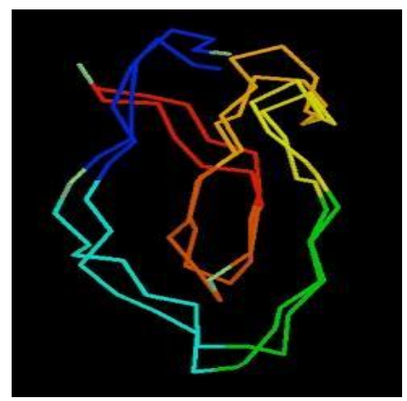
Hydrophobic amino acid

Remote Homology Detection

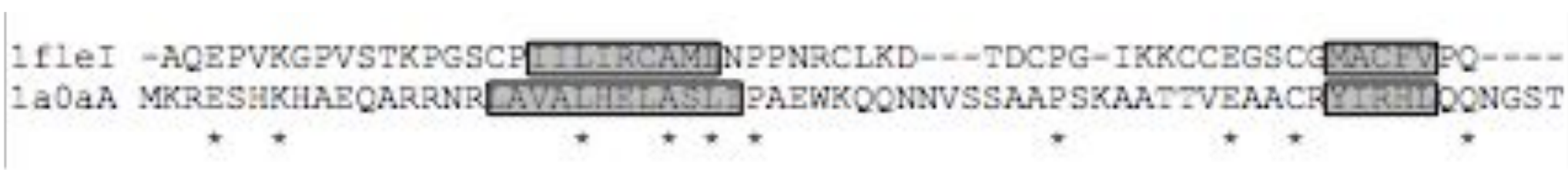
- An analysis of their structural alignment can detect homology



However, structural information is not available for most of existing proteins



Homologous



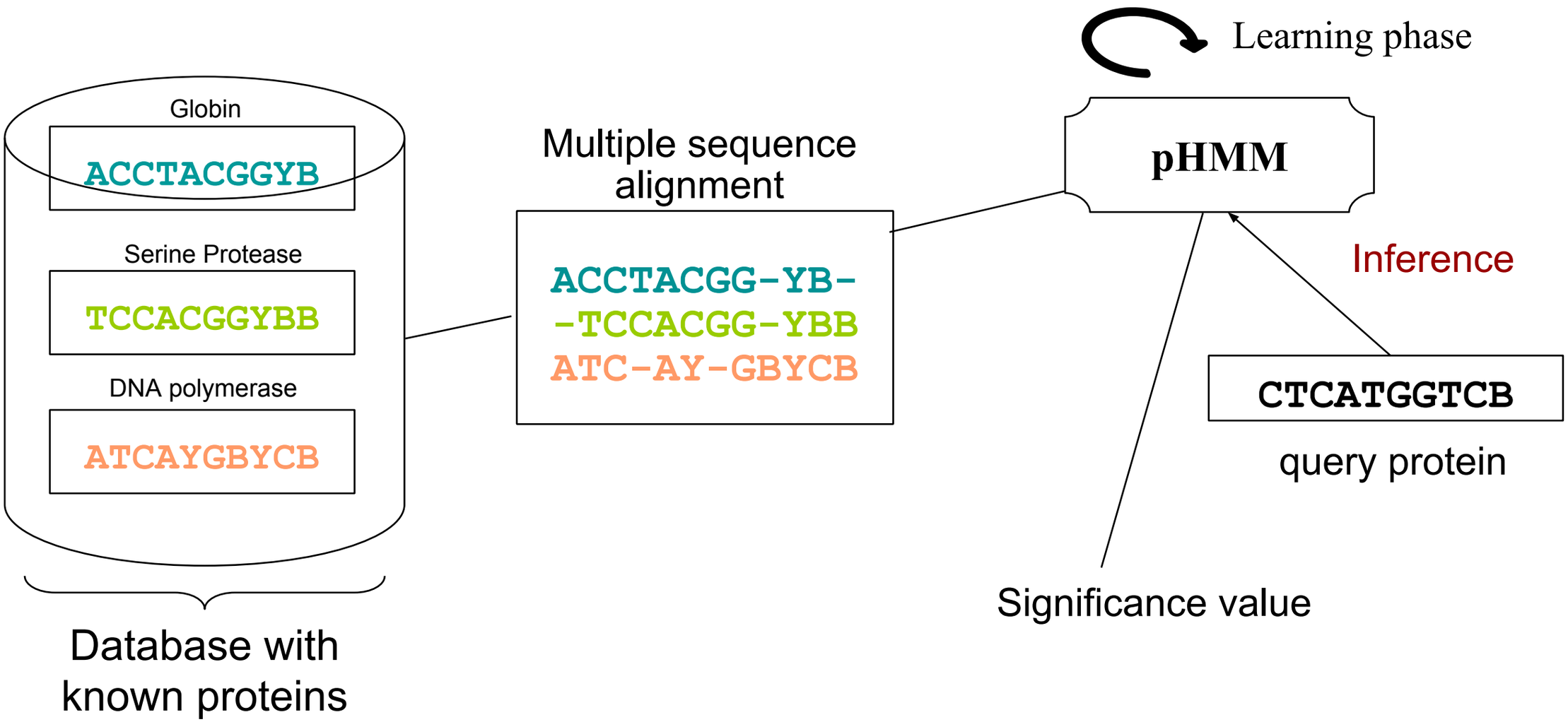
* Identical amino acid

Hydrophobic amino acid

Non-homologous

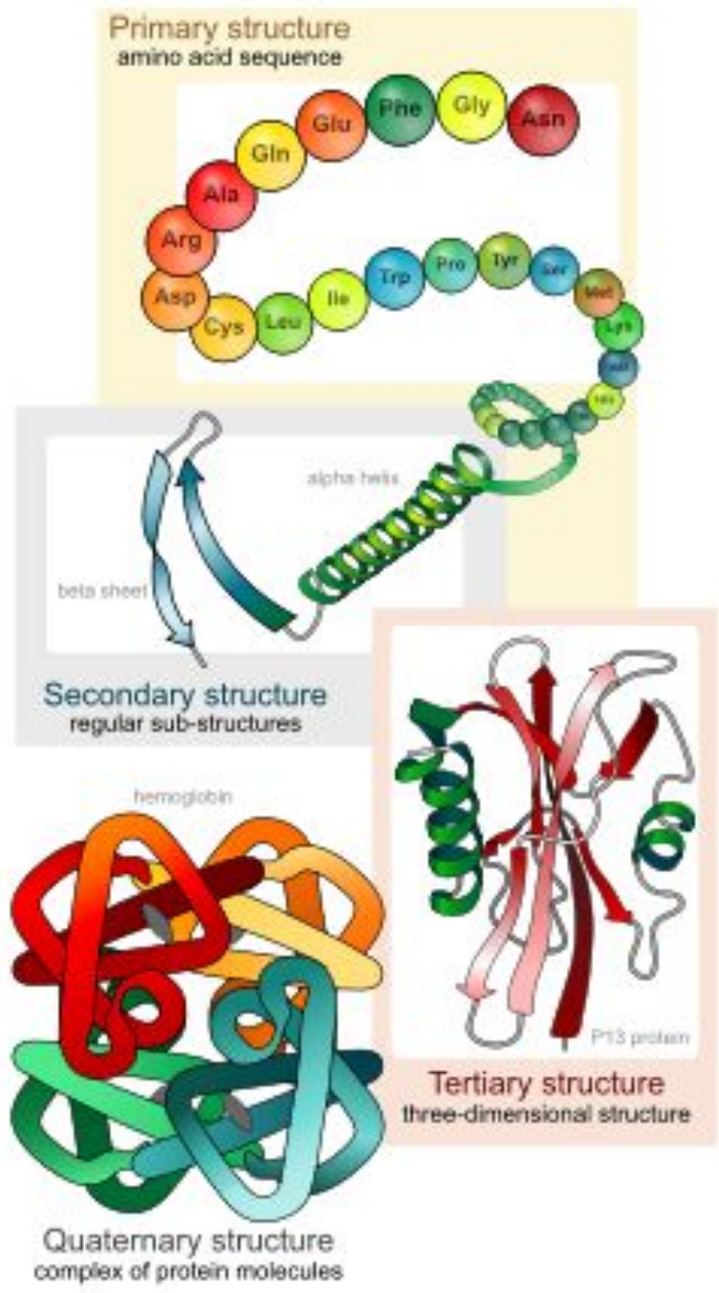
profile HMMs

Methods as profile HMMs outperform pairwise methods on remote homology



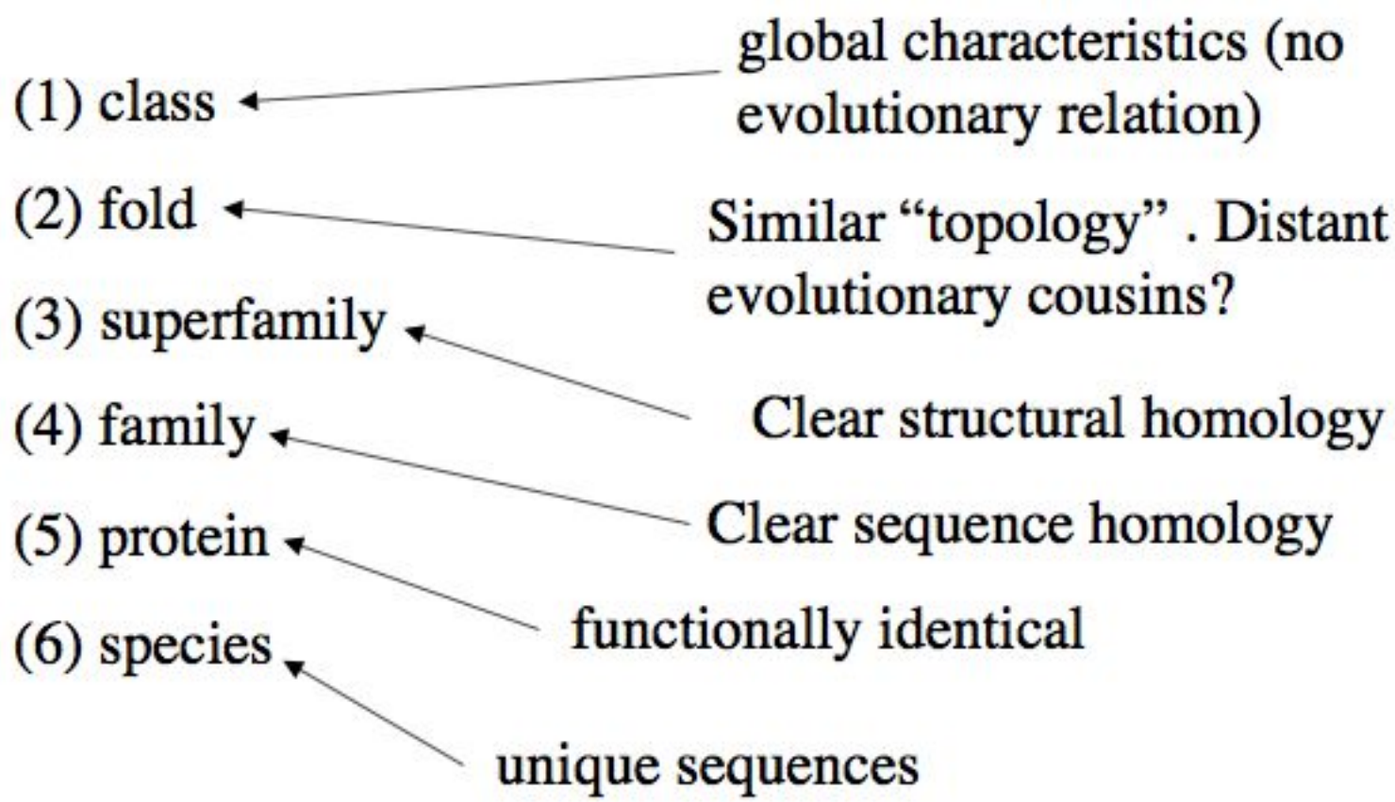
SCOP database

SCOP contains the domains of all PDB entries.



SCOP database

SCOP classifies all PDB protein structures according to their structural properties



SCOP database

Structural Classification of Proteins

Search the scop database [scop 1.75]

You can use this search engine to search the SCOP database using several access methods (
sophisticated options. Please read the release notes for a detailed explanation and example

By checking the PDB box, you can also search SCOP using the external MSDlite search eng
and MeSH terms from the primary citation). Please refer to MSDlite for more details.

d1hsja1

Search the **SCOP** database.

Search the **PDB** database using MSDlite.

Retrieve information

Clear the search form.

SCOP database

Protein: Staphylococcal accessory regulator A [1280]

Lineage:

1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: DNA/RNA-binding 3-helical bundle [46688]
   core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down
4. Superfamily: "Winged helix" DNA-binding domain [46785]
   contains a small beta-sheet (wing)
   Superfamily
5. Family: MarR-like transcriptional regulators [63379]
   The N- and C-terminal helical extensions to the common fold form the dimer interface
6. Protein: Staphylococcal accessory regulator A homolog, SarR [63472]
7. Species: Staphylococcus aureus [TaxId: 1280] [63473]

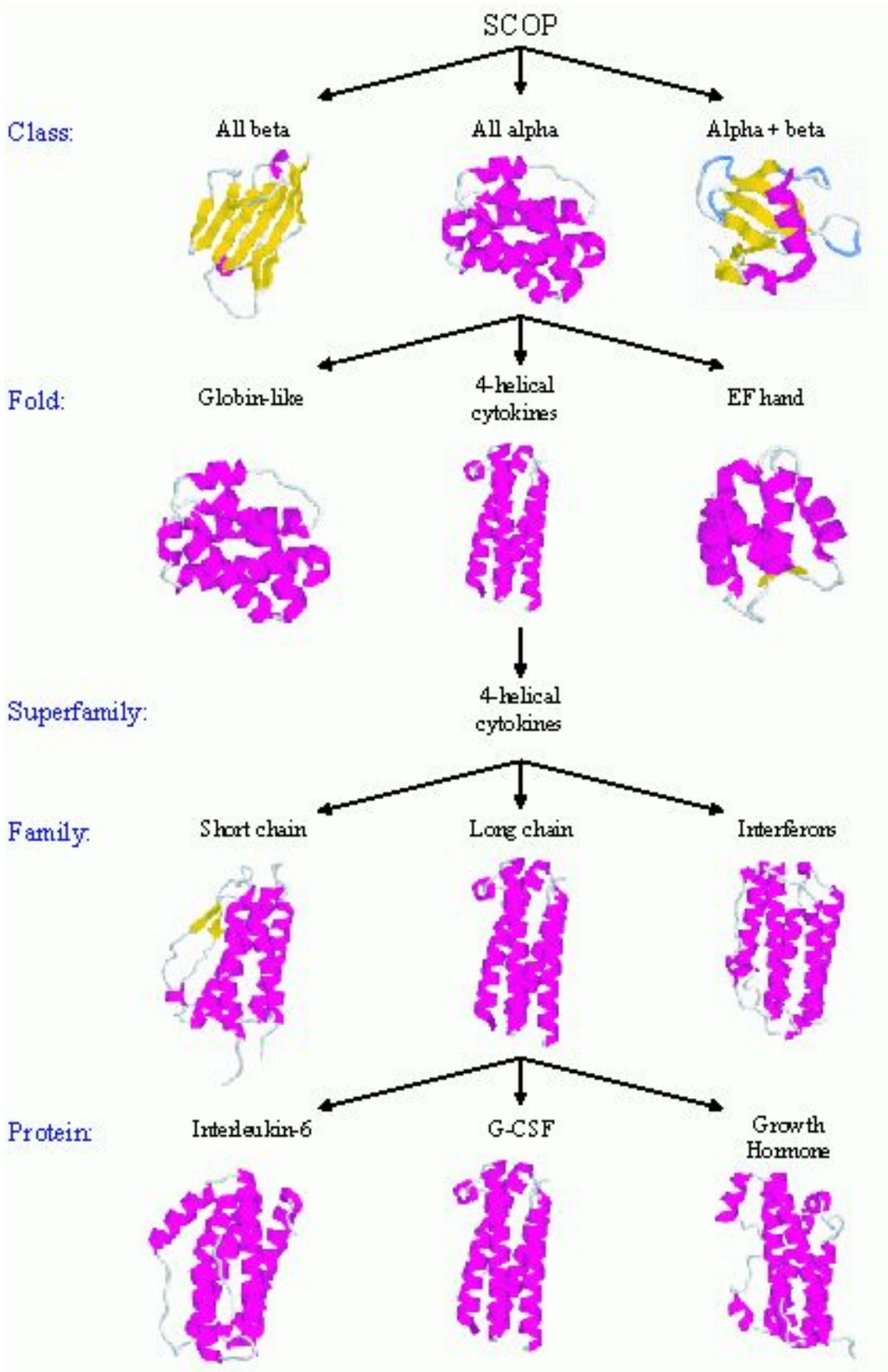
PDB Entry Domains:

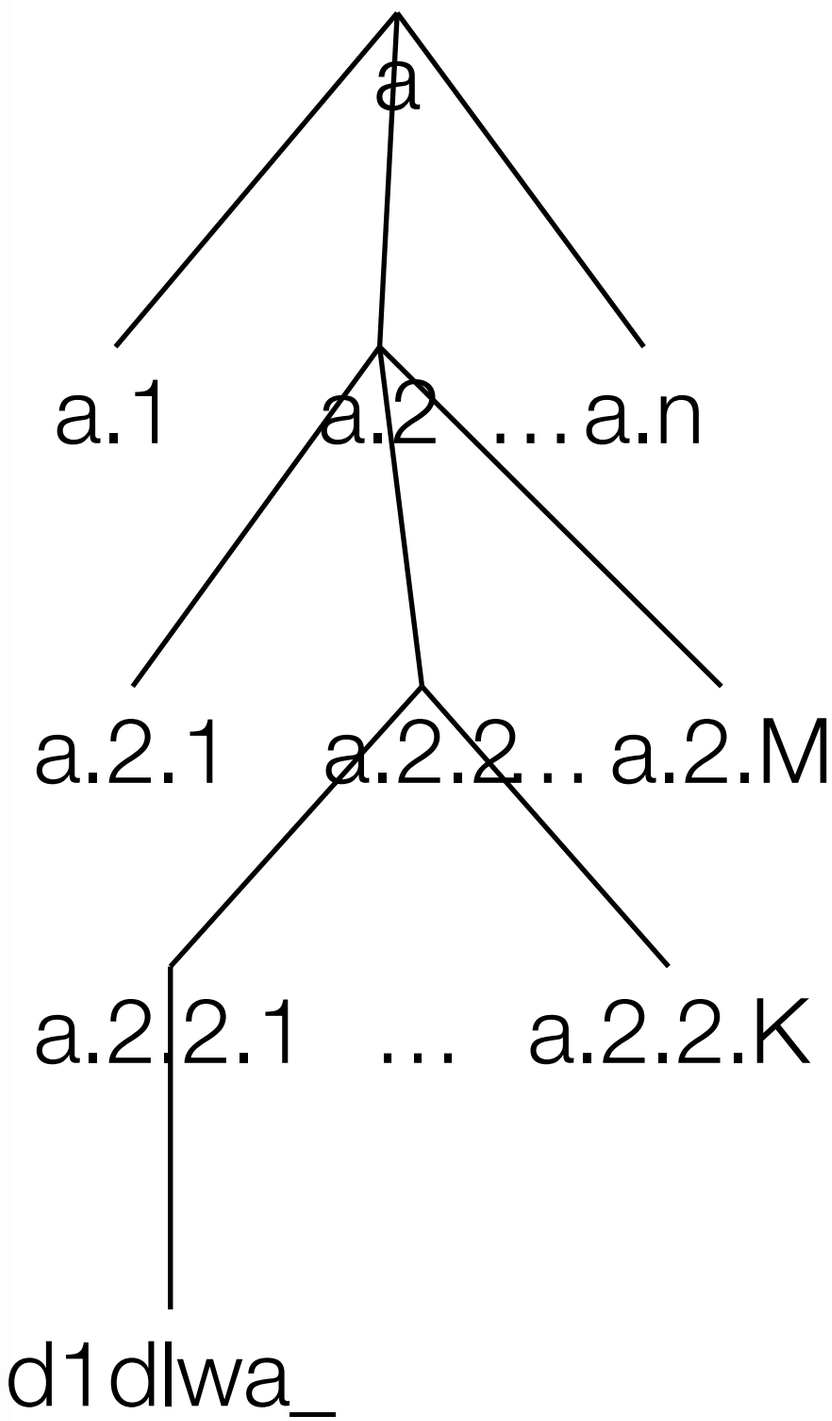
1. 1hsj
   Fusion protein with E. coli MBP
   complexed with glc
   1. region a:373-487 [61237]
   2. region b:373-487 [61239]

SCOP database

protein classes

1. all α (126) ← number of sub-categories
2. all β (81)
3. α/β (87)
4. α+β (151)
5. multidomain (21)
6. membrane (21)
7. small (10)
8. coiled coil (4)
9. low-resolution (4) ← possibly not complete, or erroneous
10. peptides (61)
11. designed proteins (17)


SCOP
Class:
All beta
All alpha
Alpha + beta
Fold:
Globin-like
4-helical cytokines
EF hand
Superfamily:
4-helical cytokines
Family:
Short chain
Long chain
Interferons
Protein:
Interleukin-6
G-CSF
Growth Hormone


a
a.1
a.2
…a.n
a.2.1
a.2.2…
a.2.M
a.2.2.1
…
a.2.2.K
d1dlwa_

How to use SCOP to evaluate homology detection tools?

Scop has subsets with different sequence identity

- ➡ Scop95 : at most 95% of sequence identity
- ➡ Scop90 : at most 90% of sequence identity

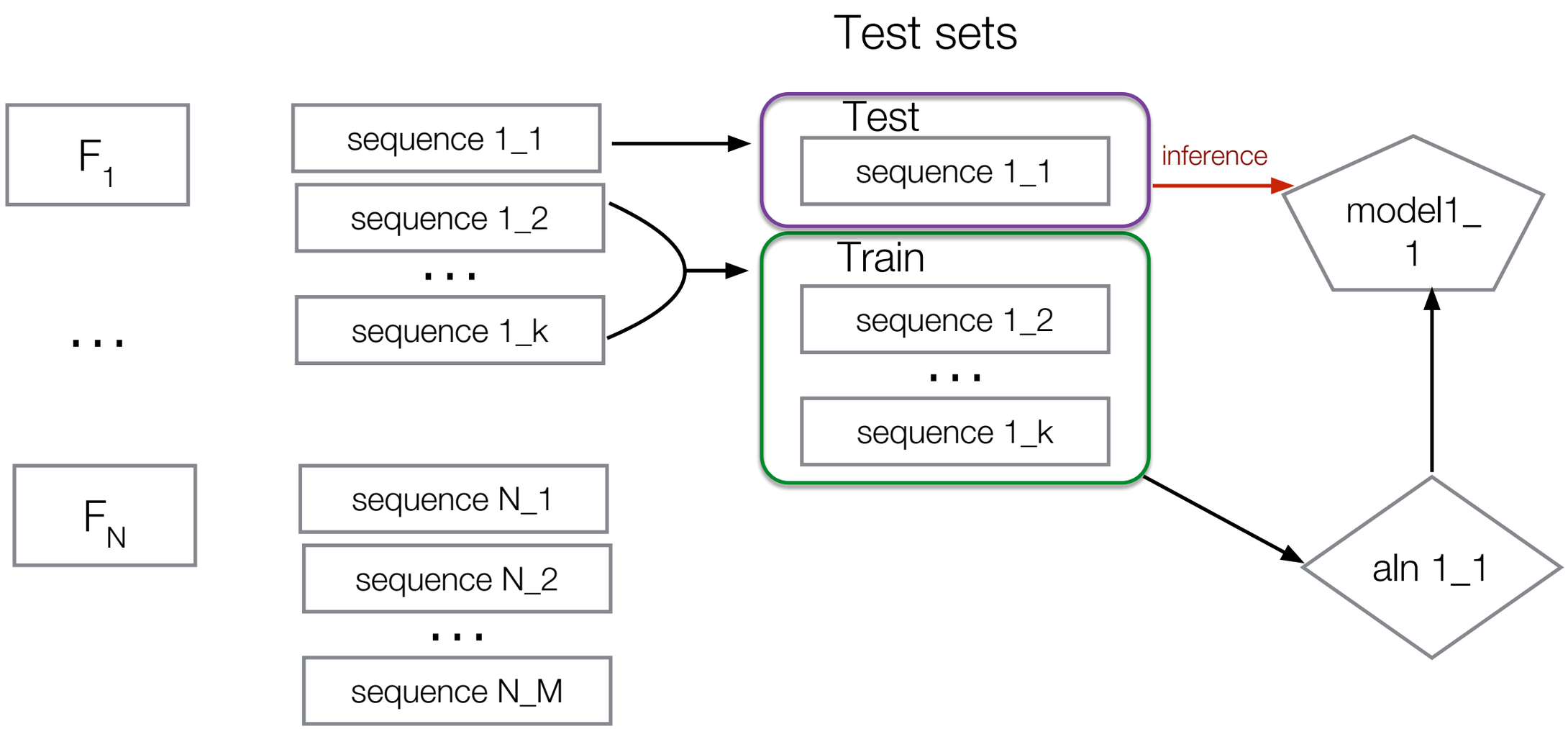
…

- ➡ Scop10 : at most 10% of sequence identity

Leave one-sequence-out experiment

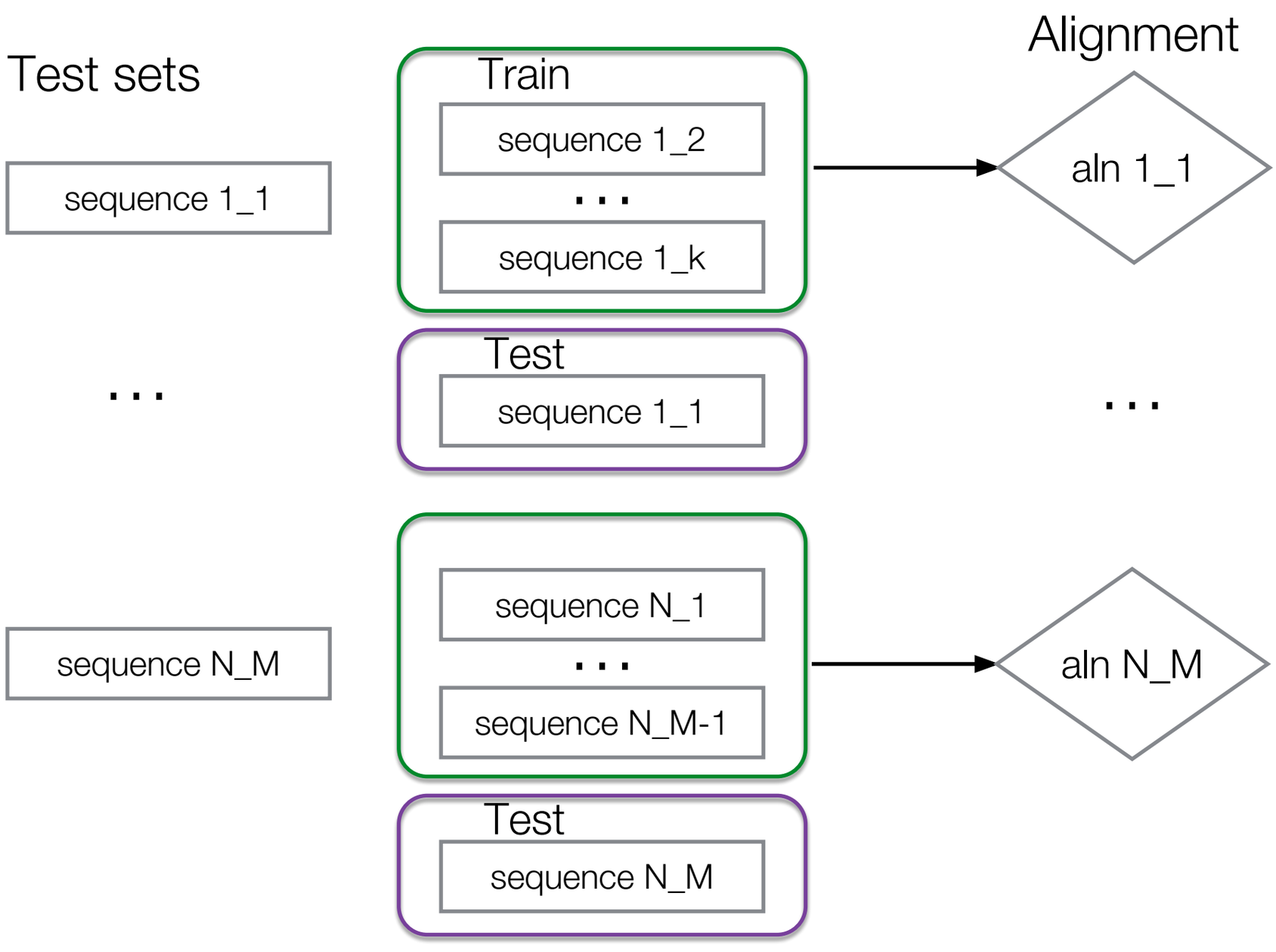
How to use SCOP to evaluate homology detection tools?

Leave one-sequence-out experiment



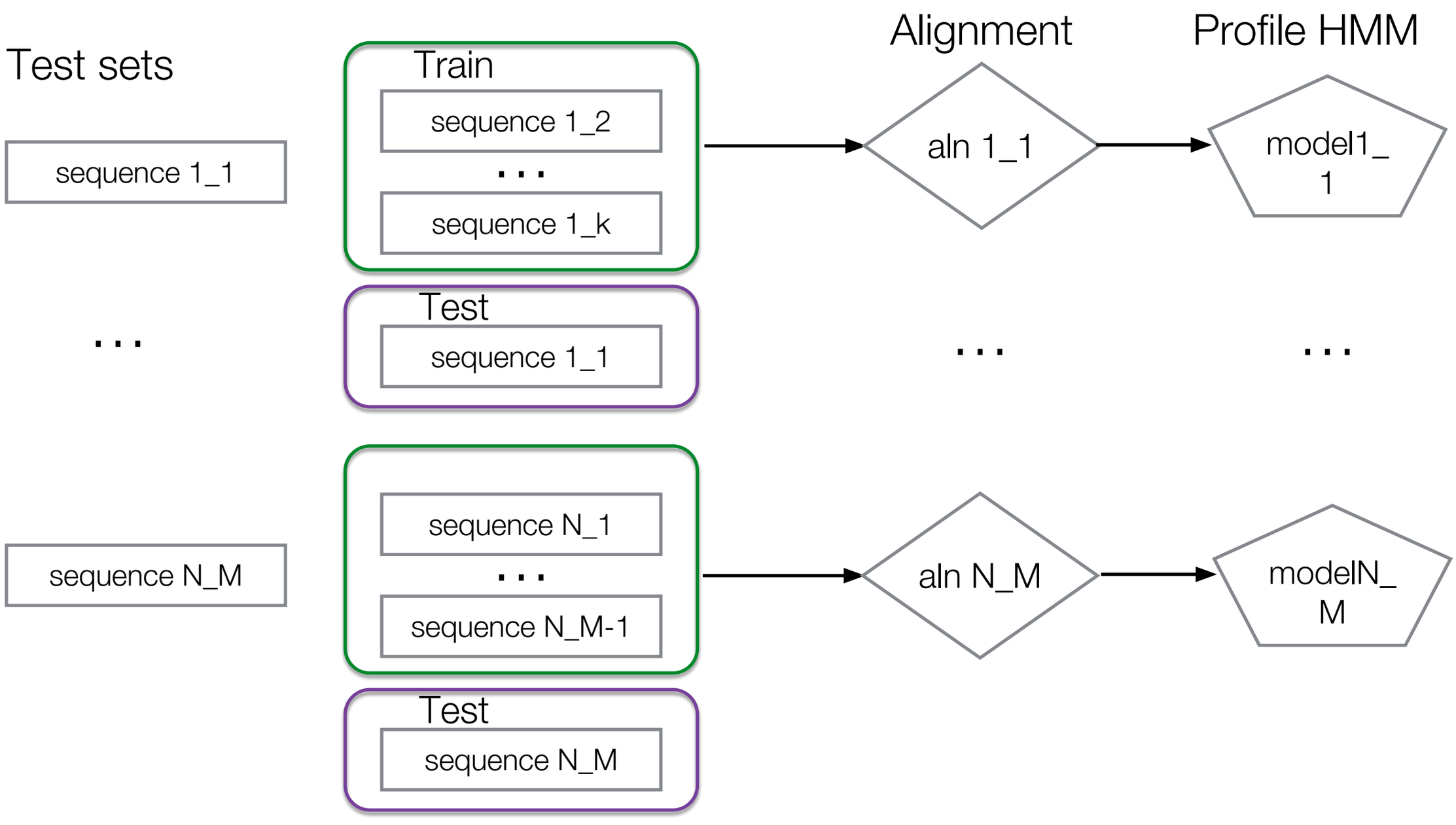
SCOP database

Leave one-sequence-out experiment



SCOP database

Leave one-sequence-out experiment

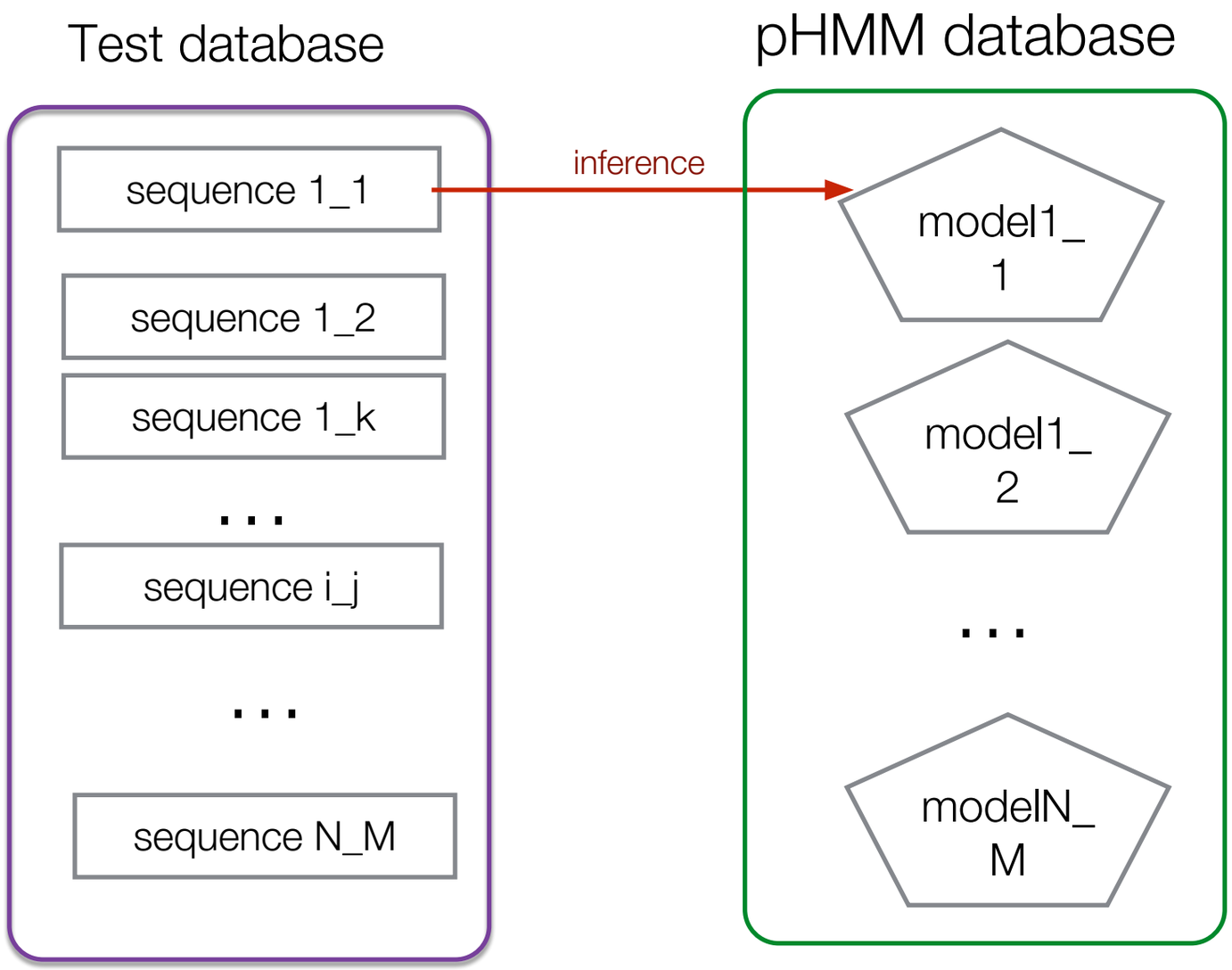


SCOP database

Constructing a unique test database

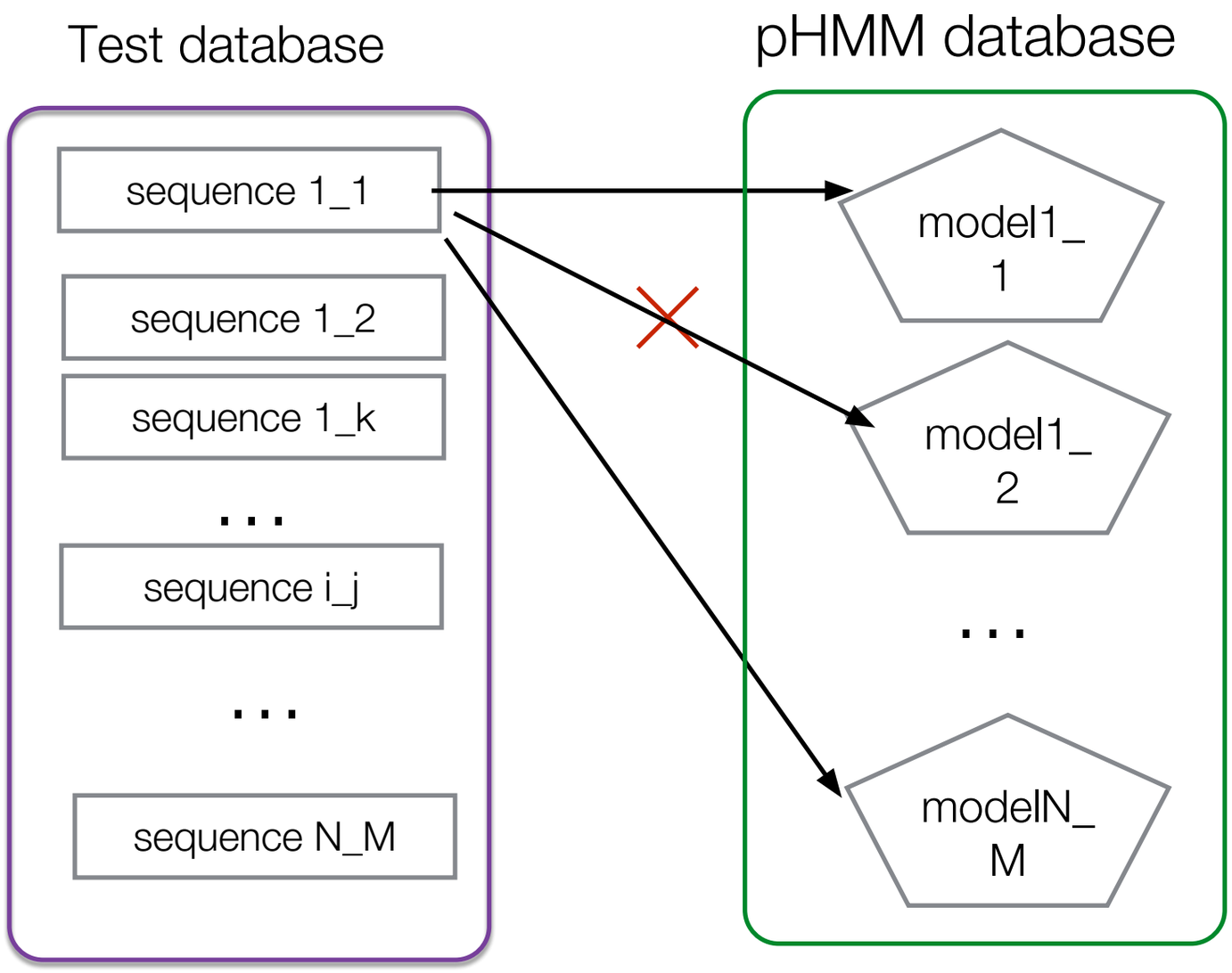
SCOP database

Leave one-sequence-out experiment



SCOP database

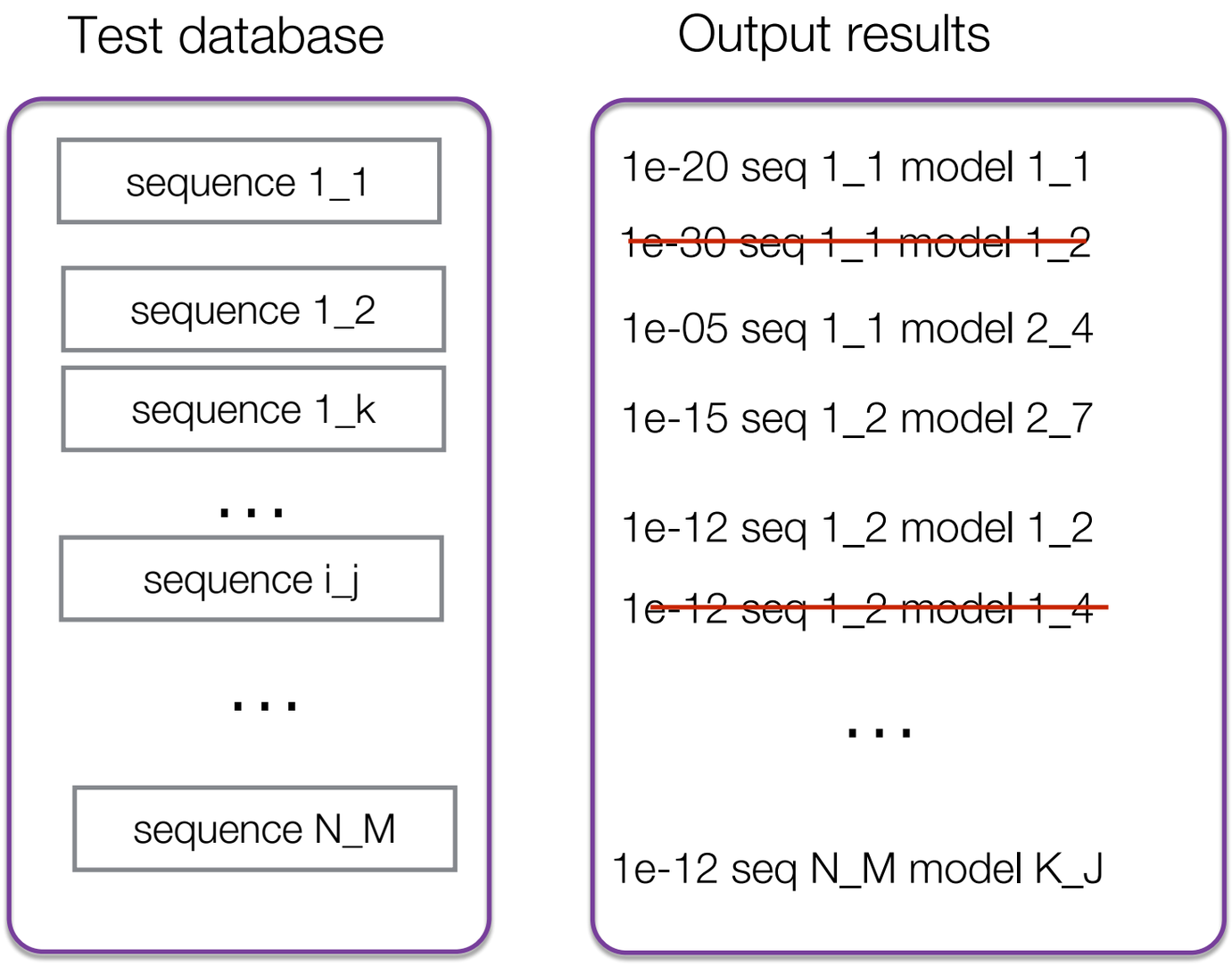
Leave one-sequence-out experiment



SCOP database

Leave one-sequence-out experiment

before evaluating the performance remove related hits



SCOP database

Leave one-sequence-out experiment

<u>Ranking and remove duplicated</u>

Output results

1e-20 seq 1_1 model 1_1

1e-05 seq 1_1 model 2_4

1e-15 seq 1_2 model 2_7

1e-12 seq 1_2 model 1_2

...

1e-12 seq N_M model N_M

1e-1 seq N_M model K_J

Output results

1e-20 seq 1_1 model 1_1

1e-15 seq 1_2 model 2_7

...

1e-12 seq N_M model N_M

SCOP database

Leave one-sequence-out experiment

<u>Ranking and remove duplicated</u>

Output results

1e-20 seq 1_1 model 1_1

1e-05 seq 1_1 model 2_4

1e-15 seq 1_2 model 2_7

1e-12 seq 1_2 model 1_2

...

1e-12 seq N_M model N_M

1e-1 seq N_M model K_J

Output results

1e-20 seq 1_1 model 1_1 ← True positive

1e-15 seq 1_2 model 2_7 ← False positive

...

seq 7_5 is missing ← False negative

1e-12 seq N_M model N_M

Precision = #TP/(#TP + #FP)

Recall = #TP/(#TP + #FN)

F-score =2 * P * R/ (P + R)

SCOP database

Leave one-sequence-out experiment

TP= True positive

FP= False positive

FN= False negative

TN= True negative

Precision = #TP/(#TP + #FP)

Recall = #TP/(#TP + #FN)

True positive rate → TPR = #TP/(#TP + #FN)

False positive rate → FPR = #FP/(#FP + #TN)

F-score =2 * P * R/ (P + R)

SCOP database

Leave one-sequence-out experiment

ROC curve

Let's suppose we have N test-sequences

Sort e-values : 1e-20, 1e-18, 1e-16, …., 1

Output results

1e-20 seq 1_1 model 1_1

1e-18 seq 1_2 model 1_2

1e-16 seq 1_3 model 2_7

1e-15 seq 2_2 model 2_2

1e-14 seq 2_3 model 3_4

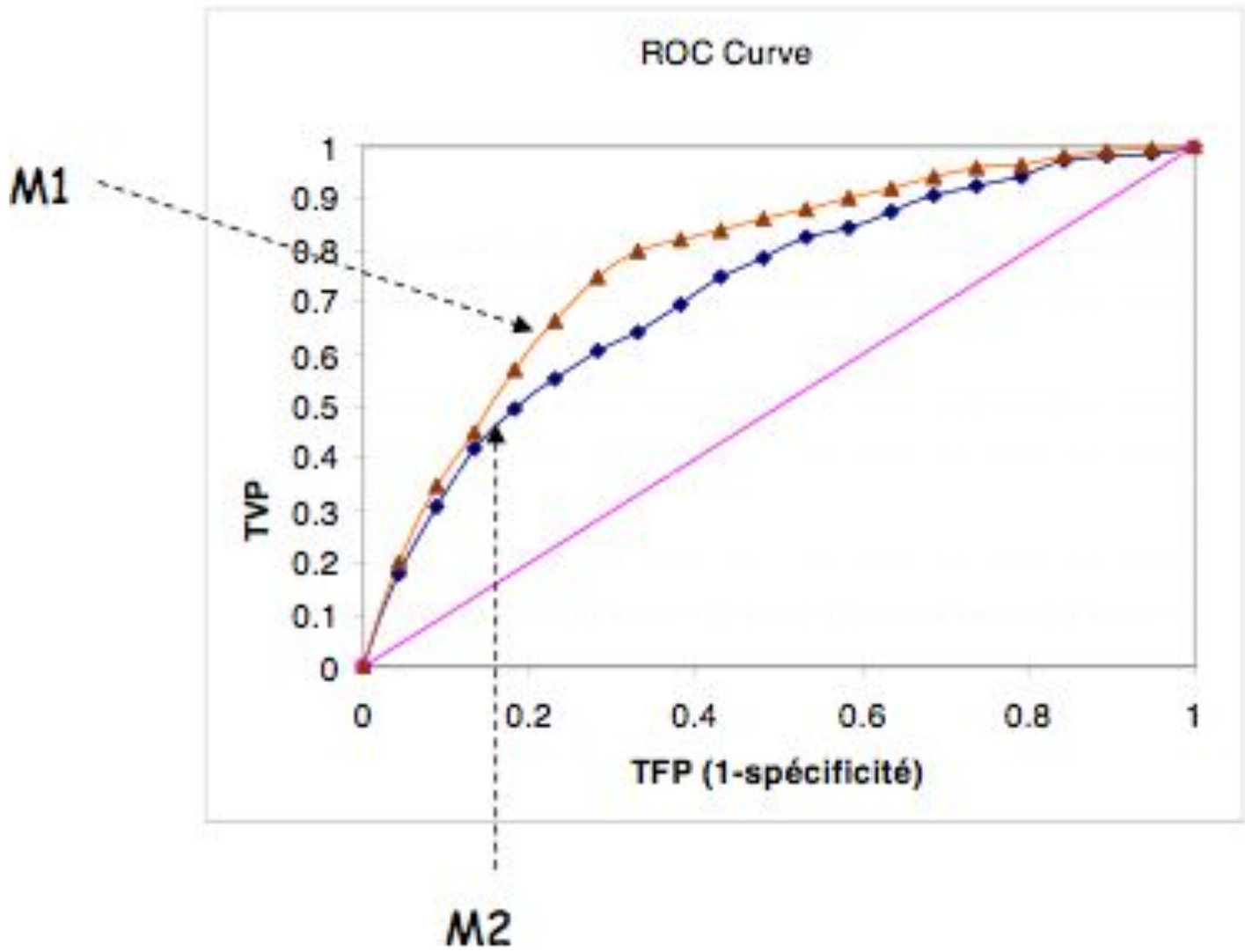
…

1e-12 seq N_M model K_J

	TP	FP
1e-20	1	0
1e-18	2	0
1e-16	2	1
1e-15	3	1
1e-14	3	2

…

Plot curves and compare tools



AUC = Area Under Curve

AUC_M1 > AUC_M2

SCOP database

Recall = #TP/(#TP + #FN) Precision = #TP/(#TP + #FP)

Precision Recall curves

Let's suppose we have N test-sequences

Output results

1e-20 seq 1_1 model 1_1

1e-18 seq 1_2 model 1_2

1e-16 seq 1_3 model 2_7

1e-15 seq 2_2 model 2_2

1e-14 seq 2_3 model 3_4

...

1e-12 seq N_M model K_J

	TP	FP	FN
1e-20	1	0	**N-1**
1e-18	2	0	**N-2**
1e-16	2	1	**N-2**
1e-15	3	1	**N-3**
1e-14	3	2	**N-3**

...

SCOP database

Recall = #TP/(#TP + #FN) Precision = #TP/(#TP + #FP)

Precision Recall curves

Let's consider N=100

Output results

1e-20 seq 1_1 model 1_1

1e-18 seq 1_2 model 1_2

1e-16 seq 1_3 model 2_7

1e-15 seq 2_2 model 2_2

1e-14 seq 2_3 model 3_4

...

1e-12 seq N_M model K_J

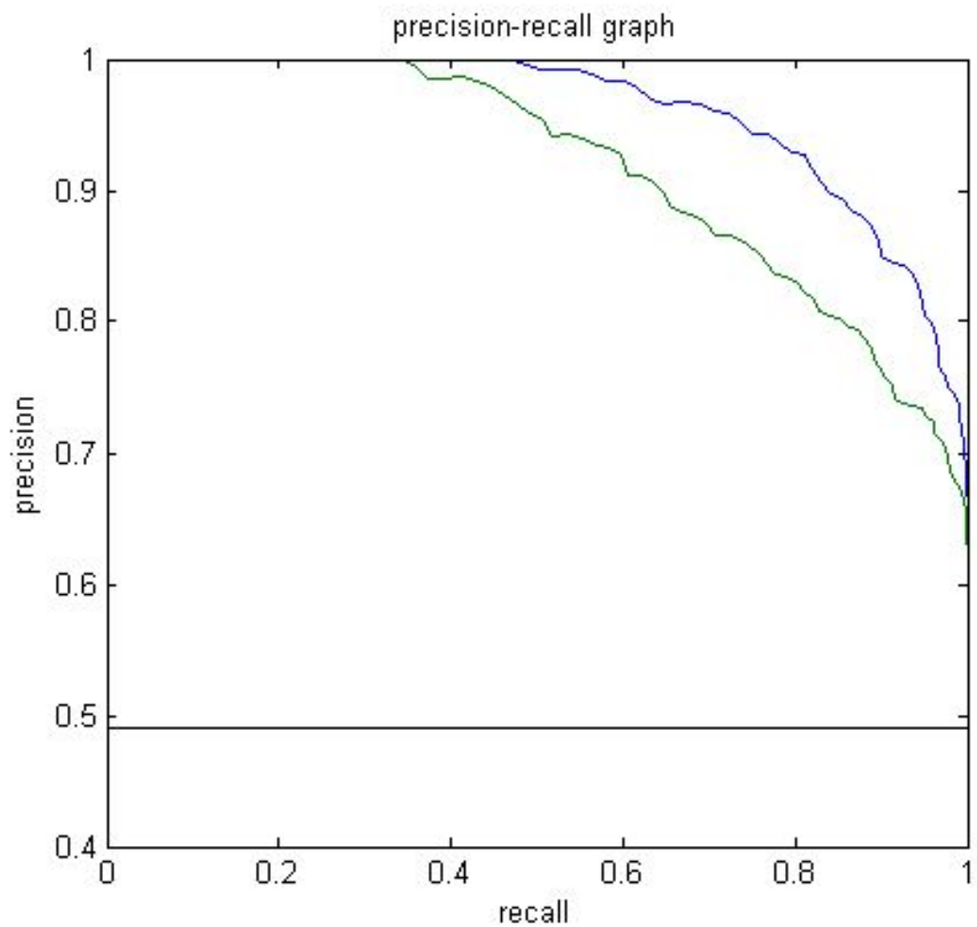
	TP	FP	FN
1e-20	1	0	**99**
1e-18	2	0	**98**
1e-16	2	1	**98**
1e-15	3	1	**97**
1e-14	3	2	**97**

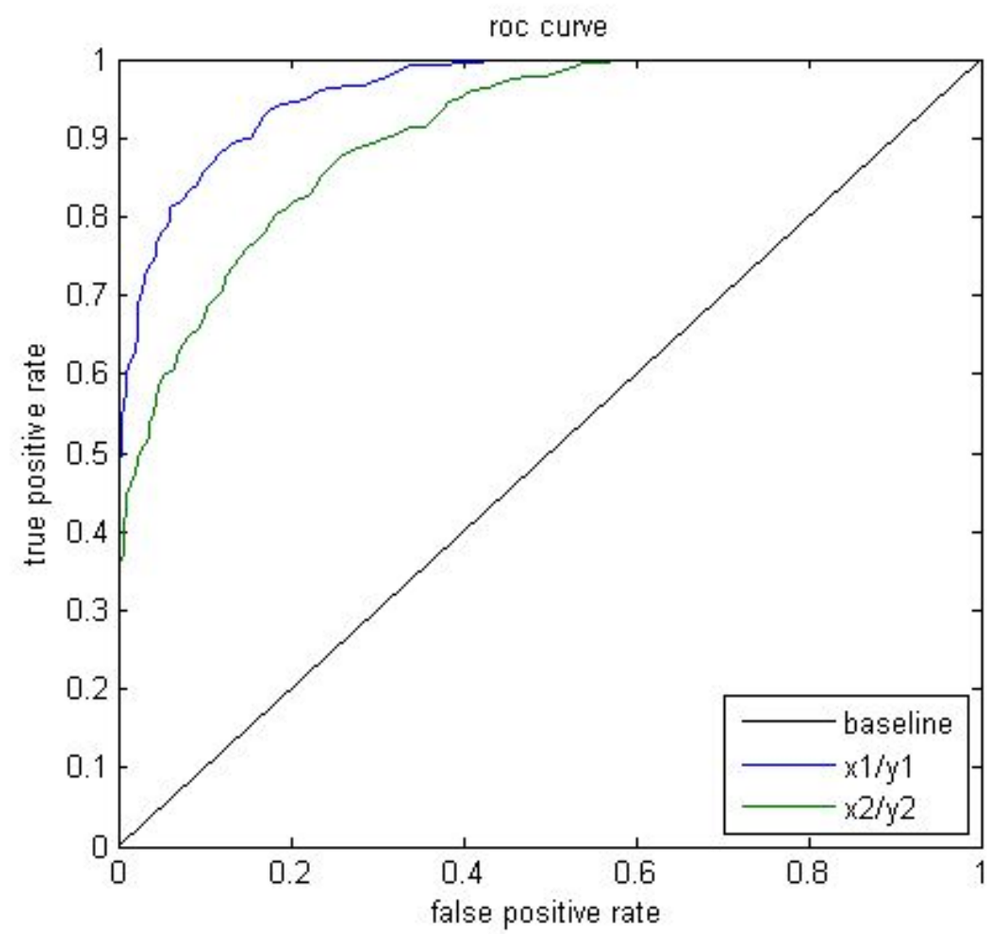
...

	Rec	Prec
1e-20	1/(99+1)	1/(1 +0)
1e-18	2/(98+2)	2/(2 +0)
1e-16	2/(98+2)	2/(2 +1)
1e-15	3/(97+3)	3/(3 +1)
1e-18	3/(97+3)	3/(3 +2)

...

Plot curves and compare tools





Some useful bash commands

- Concat your files results

```
cat *.results > all.results
```

- Remove lines with some pattern

```
#
# target name         accession   tlen query name            acce
#------------------- ---------- ----- -------------------- ----
d1atia2d.104.1.1     -            394 d12asa_d.104.1.1.aln -
d1b8aa2d.104.1.1     -            335 d12asa_d.104.1.1.aln -
d1wu7a2d.104.1.1     -            327 d12asa_d.104.1.1.aln -
d1h4vb2d.104.1.1     -            324 d12asa_d.104.1.1.aln -
d1z7ma1d.104.1.1     -            318 d12asa_d.104.1.1.aln -

sed '/^#/d' all.results > all.results.2
```

Some useful bash commands

- Taking some columns from a file

```
d1atia2d.104.1.1  -   394 d12asa_d.104.1.1.aln  -   312    2e-138  456.2   4.9   1   1  4.3e-140  2.4e-138  45
d1b8aa2d.104.1.1  -   335 d12asa_d.104.1.1.aln  -   312   6.6e-86  283.7   6.6   1   1   1.3e-87   7.4e-86  28
d1wu7a2d.104.1.1  -   327 d12asa_d.104.1.1.aln  -   312   6.3e-80  264.0   0.2   1   1   1.3e-81   7.1e-80  26
d1h4vb2d.104.1.1  -   324 d12asa_d.104.1.1.aln  -   312     1e-75  250.2   5.0   1   1     2e-77   1.1e-75  25
d1z7ma1d.104.1.1  -   318 d12asa_d.104.1.1.aln  -   312   1.9e-72  239.5   0.1   1   1   3.7e-74   2.1e-72  23
d1qf6a4d.104.1.1  -   291 d12asa_d.104.1.1.aln  -   312   3.6e-67  222.1   0.3   1   1     7e-69     4e-67  22
```

```
cut -d ' ' -f 1,4,12  all.results.2 > all.results.3
```

- It does not work with extra write space, remove them before cutting

```
cat all.results.2 | tr -s ' ' > all.results.3
cut -d ' ' -f 1,4,12  all.results.3 > all.results.4
```

```
d1atia2d.104.1.1  d12asa_d.104.1.1.aln  2e-138
d1b8aa2d.104.1.1  d12asa_d.104.1.1.aln  6.6e-86
d1wu7a2d.104.1.1  d12asa_d.104.1.1.aln  6.3e-80
d1h4vb2d.104.1.1  d12asa_d.104.1.1.aln  1e-75
d1z7ma1d.104.1.1  d12asa_d.104.1.1.aln  1.9e-72
```

Some useful bash commands

- Deleting some word from a file

d1atia2d.104.1.1 d12asa_d.104.1.1**.aln** 2e-138
d1b8aa2d.104.1.1 d12asa_d.104.1.1**.aln** 6.6e-86
d1wu7a2d.104.1.1 d12asa_d.104.1.1**.aln** 6.3e-80
d1h4vb2d.104.1.1 d12asa_d.104.1.1**.aln** 1e-75
d1z7ma1d.104.1.1 d12asa_d.104.1.1**.aln** 1.9e-72

```
sed 's/.aln//g'  all.results.4 > all.results.5
```

Bash Scripts

- Let's suppose we'd like to run hmmbuild for all .sto files in a directory

```
d12asa_d.104.1.1.aln.sto d1ixca2c.94.1.1.aln.sto  d1q0qa2c.2.1.3.aln.sto   d1vr5a1c.94.1.1.aln.sto
d1a3ca_c.61.1.1.aln.sto  d1ixha_c.94.1.1.aln.sto  d1q2ya_d.108.1.1.aln.sto d1vr9a3d.37.1.1.aln.sto
d2euia1d.108.1.1.aln.sto d2ozza1c.94.1.1.aln.sto  d1a5ta2c.37.1.20.aln.sto d1ixla_d.38.1.5.aln.sto
d1q33a_d.113.1.1.aln.sto d1vyra_c.1.4.1.aln.sto   d2f1ka2c.2.1.6.aln.sto   d2p0wa_d.108.1.1.aln.sto
d1a62a2b.40.4.5.aln.sto  d1j0ha1b.1.18.2.aln.sto  d1q35a_c.94.1.1.aln.sto  d1w0ha_c.55.3.5.aln.sto
d2f41a1d.38.1.5.aln.sto  d2p65a_c.37.1.20.aln.sto d1abaa_c.47.1.1.aln.sto  d1j0ha2b.71.1.1.aln.sto
d1q4ta_d.38.1.5.aln.sto  d1w23a_c.67.1.4.aln.sto  d2f5va1c.3.1.2.aln.sto   d2p74a_e.3.1.1.aln.sto
d1al3a_c.94.1.1.aln.sto  d1j5pa4c.2.1.3.aln.sto   d1q8ia1c.55.3.5.aln.sto  d1w4xa1c.3.1.5.aln.sto
d2f96a1c.55.3.5.aln.sto  d2plna_c.23.1.0.aln.sto
```

```
for FILE in *.sto; do
        NAME_HMM=${FILE/aln.sto/hmm}
        echo "Building $NAME_HMM"
        hmmbuild $NAME_HMM $FILE
done
```

Save it in a file: runHmmb.sh

To execute it, use the command:

```
bash runHmmb.sh
```

- After we can move all *.hmm files to a new directory

```
mkdir ../models
mv *.hmm ../models
```