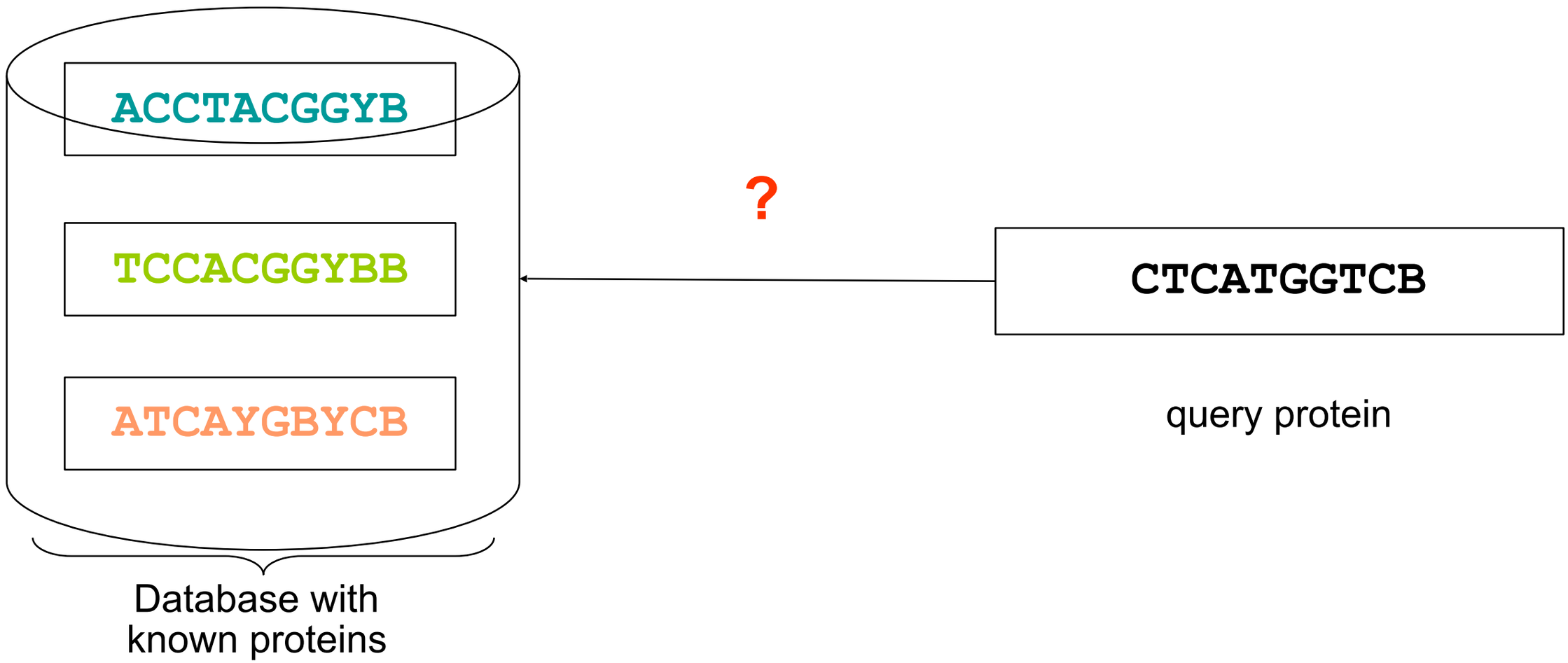
Protein domain Annotation

Juliana Bernardes
Riccardo Vicedomini

Protein Annotation

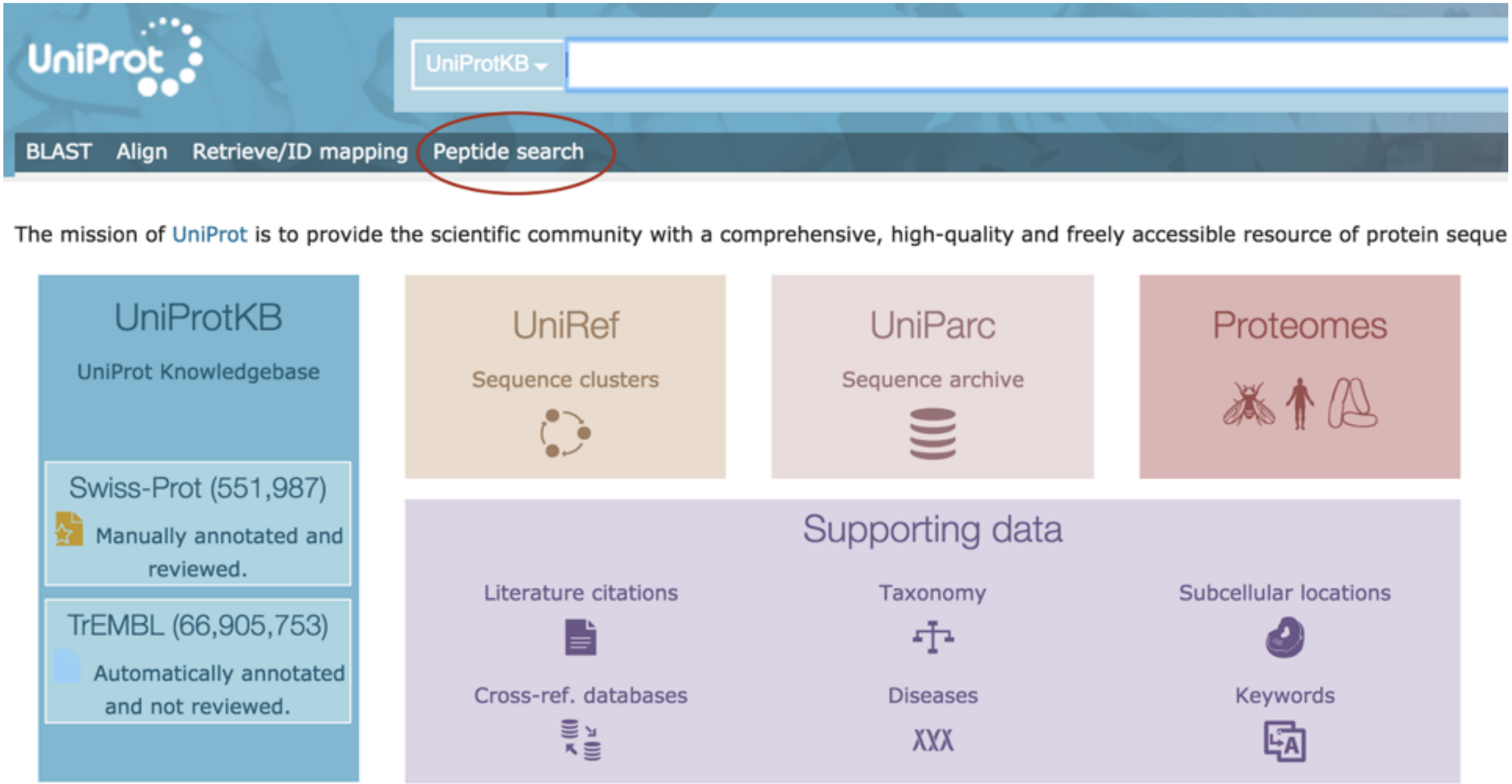
- Exponential **increase of the number of proteins** being identified by sequence genomic projects
- Impossible to perform **functional assay** for every new protein
- Need of **computational methods** for annotating the huge volume of sequences being produced

Protein Annotation



Why compare a sequence against a database of known proteins?

Database with annotated proteins



How to compare protein sequences?

Pairwise sequence alignment

How to aligner two sequences?

- Three elementary events:
  - match
  - mismatch
  - Indel (Insertion/Délétion) (Gaps)

Match.: 2, MisMatch.: -1, Gap: -2

```
ACGGCTAT      ACGGCTAT      ACGGCTAT
              ||  |         ||  |  |||
ACTGTAT       ACTGTAT-      ACTG-TAT
```

How to evaluate an alignment?

- **ACGGCTAT** et **ACTGTAT**

Match.: 2, MisMatch.: -1, Gap: -2

A1 =>

```
ACGGCTAT
||  |
ACTGTAT-
```

Score A1 = 2+2-1+2-1-1-1-2 = 0

A2 =>

```
ACGGCTAT
||  |  |||
ACTG-TAT
```

Score A2= 2+2-1+2-2+2+2+2 = 9

Two ways to align sequences

- global alignment - end-to-end alignment
- local alignment - local similarity

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

★divide and conquer

The dynamic programming solves the original problem by dividing the problem into smaller independent sub problems

An alignment of size L is the best one

if the alignment of size L-1 is the best sub-alignment

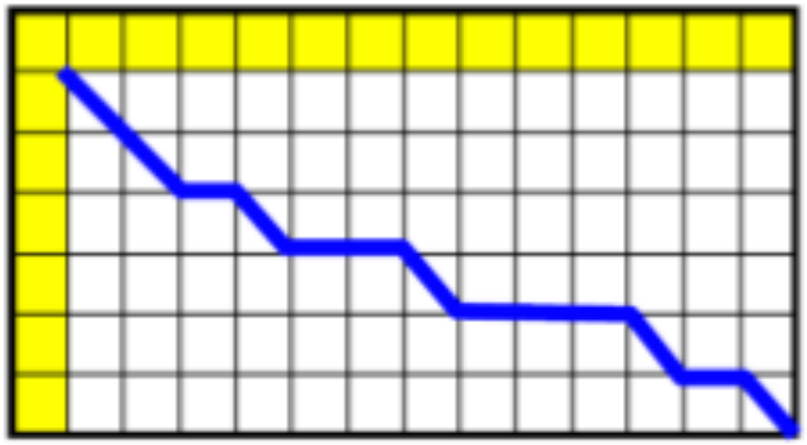
Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

- Represents the two sequences by using a matrix
- An alignment is a unique path in the matrix
- A score is associated to each path (or alignment)
- We are looking for the alignment with best score

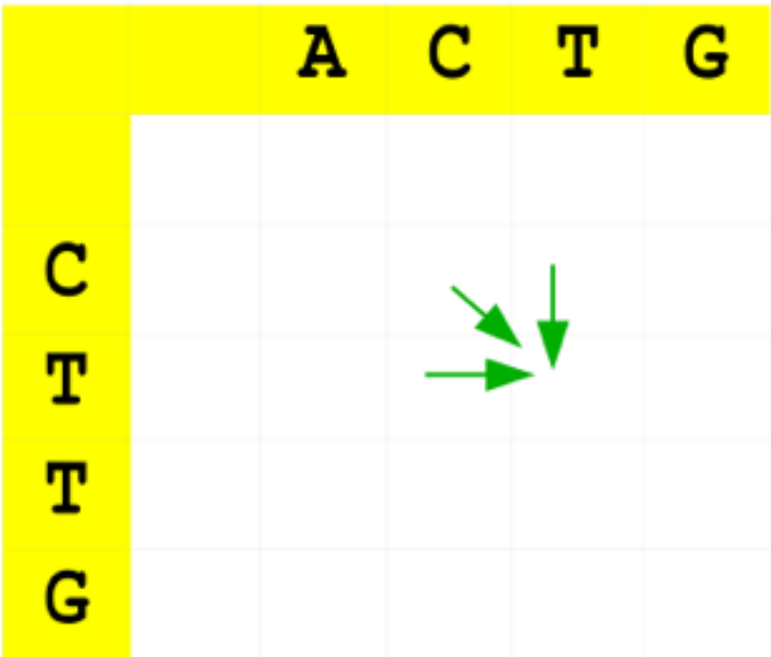
```
AGTCAGTGCGTGC
AG_C__T___T_C
```



Global alignment : Dynamic programming

Needleman & Wunsch - 1970

- How to fill the matrix?



Rule 1: We compute a score for each cell, and the final score is given by last cell

Rule 2: the score of a given cell is computed from scores on top, diagonal and left of the previous cells.

Rule 3: top and left insert a gap, diagonal align two lettres.

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

		A	C	T	G
	0	- 4	- 8	- 12	- 16
C	- 4				
T	- 8				
T	- 12				
G	- 16				

Match = +4
Mismatch=-4
Gap = -4

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

		A	C	T	G
	0	-4	-8	-12	-16
C	-4				
T	-8				
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

alignement AC → score = 0-4 = -4

insertion de gap → score = -4-4 = -8

insertion de gap → score = -4-4 = -8

Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

We fill all cells and we keep a pointer for the previous cell with the best score

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4			
T	-8				
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

alignement AC → score = 0-4 = -4

insertion de gap → score = -4-4 = -8

insertion de gap → score = -4-4 = -8

Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

We fill all cells and we keep a pointer for the previous cell with the best score

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0		
T	-8				
T	-12				
G	-16				

Score:

gap: -4

mismatch: -4

match: +4

alignement CC → score = -4+4 = 0

insertion de gap → score = -8-4 = -12

insertion de gap → score = -4-4 = -8

Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0		
T	-8	-8			
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

match: +4

alignement AT → score = -4-4 = -8

insertion de gap → score = -4-4 = -8

insertion de gap → score = -8-4 = -12

Global alignment : Dynamic programming

Needleman & Wunsch - 1970

		A	C	T	G
	0	- 4	- 8	- 12	- 16
C	- 4	- 4	0	- 4	- 8
T	- 8	- 8	- 4	4	0
T	- 12	- 12	- 8	0	0
G	- 16	- 16	- 12	- 4	4

Best score

Score:

gap: -4

mismatch: -4

match: +4

How many optimal alignment we have?

Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming

Follow the arrows back to the original cell to obtain the path for the best alignment.

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0	-4	-8
T	-8	-8	-4	4	0
T	-12	-12	-8	0	0
G	-16	-16	-12	-4	4

2 paths =
2 optimal alignments

```
AC-TG      ACT-G
-CTTG      -CTTG
```

score: +4

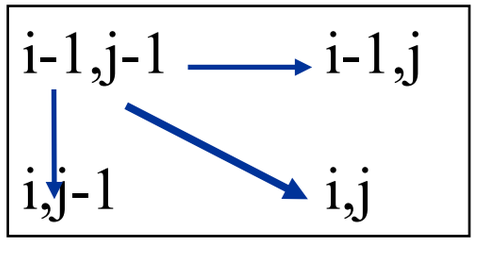
24 scores computes
3^{4+4} = 6561 different paths

Global alignment sequences were aligned end-to-end

Source : Aix-Marseille Université - LUMINV

Global alignment : Dynamic programming: formalisation

- 2 seq A = $(a_1,\ldots,a_n)$ et $B(b_1,\ldots,b_m)$
- $S_{i,j}$ = maximum score between 2 aligned sub-sequences a_i et b_j tel que :
- $$S_{i,j} = \max \left| \begin{array}{l} S_{i-1,j-1} + w(a_i,b_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{array} \right.$$



Local Alignement

- What are the most conserved regions of two sequences ?

GGCTGACCACCTT et GATCACTTCCATG

Local Alignment

Match.: 2, MisMatch.: -1, Gap: -2

- two sequences :
  - GGCTGACCACCTT et GATCACTTCCATG

- global alignement :

```
1 GGCTGACCACC-TT 13
  |   | ||   || |
1 GA-TCACTTCCATG 13
```

Score = 5

- local Alignement :

```
5 GACCACCTT 13
  || ||| ||
1 GATCAC-TT 8
```

Score = 11

Local Alignement

Smith & Waterman (1981)

- Local alignment algorithm of Smith & Waterman is based on the global alignment proposed by Needleman & Wunsch
- Score of the first row and column are set to zero;
- **Traceback** from the cell with the best score.

Local Alignment : Score of the first row and column are set to zero;

		A	C	G	G	C	T	A	T
	0	0	0	0	0	0	0	0	0
A	0								
G	0								
C	0								
T	0								
T	0								
T	0								
C	0								

Local Alignement : Compute cell scores

Match.: 2, MisMatch.: -1, Gap: -2

max [↓ 0-2= -2 → 0-2= -2 ↘ 0+2 = **2**]

		A	C	G	G	C	T	A	T
	0	0	0	0	0	0	0	0	0
A	0	2							
G	0								
C	0								
T	0								
T	0								
T	0								
C	0								

Local Alignment : compute all scores

Match.: 2, MisMatch.: -1, Gap: -2

		A	C	G	G	C	T	A	T
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	0	0	2	1
G	0	0	1	2	2	0	0	0	1
C	0	0	2	0	1	4	2	0	0
T	0	0	0	1	0	2	6	4	2
T	0	0	0	0	0	0	4	5	6
T	0	0	0	0	0	0	2	3	7
C	0	0	2	0	0	2	0	1	5

Local Alignment : Traceback from the best score

Match.: 2, MisMatch.: -1, Gap: -2

		A	C	G	G	C	T	A	T
	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	0	0	2	1
G	0	0	1	2	2	0	0	0	1
C	0	0	2	0	1	4	2	0	0
T	0	0	0	1	0	2	6	4	2
T							4	5	6
T							2	3	7
C							0	1	5

```
GCTAT
||| |
GCTTT
```

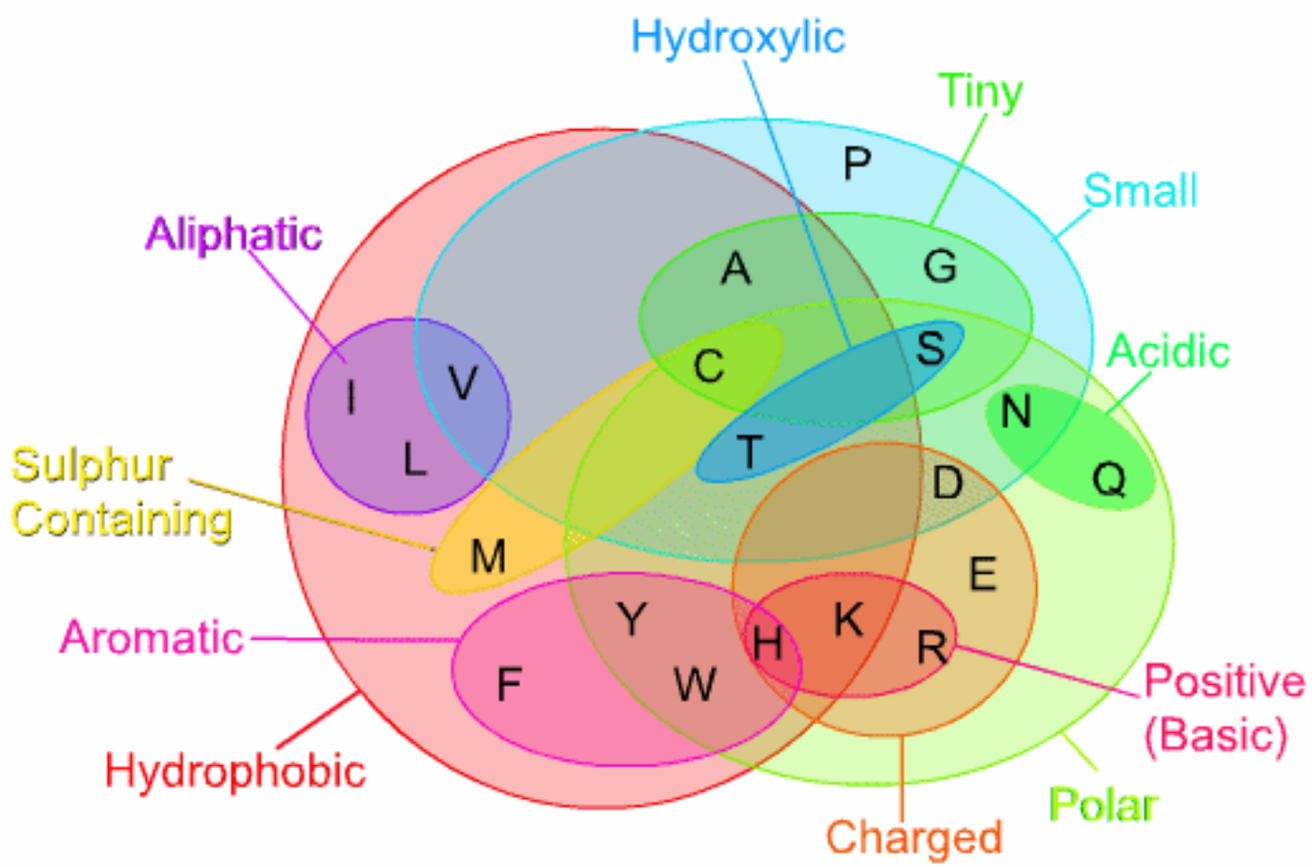
Score = 2+2+2-1+2=7

Alignment score

- alignement score = sum of scores (Match, Mismatch, Gaps) of each position.
- Improvements:
  - Substitution matrix (different Match and Mismatch) => Give different scores according to residues mutability.

Substitution matrices of amino acids

Over a longer period of evolutionary time. Each amino acid is more or less likely to mutate into various other amino acids.



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

Substitution matrices of amino acids

- Scoring matrix 20x20
- $S_{i,j}$ represents the gain/penalty due to substituting AA_j by AA_i (i–line,j–colomn)

BLOSUM 62 scoring matrix

(positive values are shaded)

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

When we search a query sequence against a large database which type of alignment we should use?

Global or local alignment

BLAST: Heuristic algorithm

Smith & Waterman (1981), exact algorithm producing optimal alignments.

Problem :

if 1 alignment with SW takes 15 ms
On SwissProt (> 500 000 sequences) will take 2h
=> Very slow ! We need some **heuristic** to run much faster than optimal alignment approaches.

Blast: Basic Local Alignement Search Tool

Altschul & al., 1990

- BLAST is one of the most widely used bioinformatics programs for sequence searching
- BLAST compares a **query** sequence with a database of sequences (Subject), and select those above a certain threshold.
- It emphasis on speed is vital to making the algorithm practical on the huge genome databases

BLAST cannot "guarantee the optimal alignments of the query and database sequences" as Smith-Waterman algorithm does.

Blast: Three main steps

Step 0 : Indexing the database

Step 1 : matching exact words

Step 2 : Extend the alignment

Step 3 : Compute alignment score

Blast: Step 0 - Indexing the database

Prepare the database for posterior searches (Performed just once)

database :

```
>PrSub1
EKFKAAMLLKSDTRCLGYRNVCKEG
>PrSub2
YYDDVGLLCEKADTRALMAQFVPPL
>PrSub3
SACILSTVNHSILKKSVHCLGYRSV
```

Blast: Step 0 - Indexing the database

k is the word size to index the database

database : k=5

```
>PrSub1
EKFKAAMLLKSDTRCLGYRNVCKEG
>PrSub2
YYDDVGLLCEKADTRALMAQFVPPL
>PrSub3
SACILSTVNHSILKKSVHCLGYRSV
```

Index :

```
EKFKA PrSub1   1
```

Blast: Step 0 - Indexing the database

database : **k=5**

```
>PrSub1
EKFKAAMLLKSDTRCLGYRNVCKEG
>PrSub2
YYDDVGLLCEKADTRALMAQFVPPL
>PrSub3
SACILSTVNHSILKKSVHCLGYRSV
```

Index :

```
EKFKA PrSub1   1
KFKAA PrSub1   2
```

Blast: Step 0 - Indexing the database

database : k=5

```
>PrSub1
EKFKAAMLLKSDTRCLGYRNVCKEG
>PrSub2
YYDDVGLLCEKADTRALMAQFVPPL
>PrSub3
SACILSTVNHSILKKSVHCLGYRSV
```

Index :

EKFKA	PrSub1	1
KFKAA	PrSub1	2
FKAAM	PrSub1	3

Blast: Step 0 - Indexing the database

database : k=5

```
>PrSub1
EKFKAAMLLKSDTRCLGYRNVCKEG
>PrSub2
YYDDVGLLCEKADTRALMAQFVPPL
>PrSub3
SACILSTVNHSILKKSVHCLGYRSV
```

```
Index :

EKFKA PrSub1   1
KFKAA PrSub1   2
FKAAM PrSub1   3

   …
   …
   ...
CLGYR PrSub1 15 PrSub3 19
   ...
```

Blast: Step 0 - Indexing the database

Index :

```
EKFKA PrSub11
KFKAA PrSub12
FKAAM PrSub13
   …
CLGYR PrSub115
   …
   ...
CLGYR PrSub115  PrSub3  19
   ...
```

Sort

Index sorted :

```
AAMLL PrSub1 5
ACILS PrSub3 2
...
CILST PrSub3 3
CLGYR PrSub1 15 PrSub3 19
DDVGL PrSub2 3
DTRAL PrSub2 13
DTRCL PrSub1 12
 . . .
KSDTR PrSub1 10
KSVHC PrSub3 15
etc.
```

Blast: Three main steps

Step 0 : Indexing the database

Step 1 : matching exact words

Step 2 : Extend the alignment

Step 3 : Compute alignment score

Blast: Step 1 - Matching exact words

Match exact words in the query and indexed database

Query k=5

```
>ProtQ
SKCDKSDTRALLAQYIPSTVNHPIL
```

Index sorted :

```
AAMLL PrSub1 5
ACILS PrSub3 2
...
CILST PrSub3 3
CLGYR PrSub1 15 PrSub3 19
DDVGL PrSub2 3
DTRAL PrSub2 13
DTRCL PrSub1 12
 . . .
KSDTR PrSub1 10
KSVHC PrSub3 15
etc.
```

Blast: Step 1 - Matching exact words

Match exact words in the query and indexed database

Query k=5

```
>ProtQ
SKCDKSDTRALLAQYIPSTVNHPIL

SKCDK => 0
```

Index sorted :

```
AAMLL PrSub1 5
ACILS PrSub3 2
...
CILST PrSub3 3
CLGYR PrSub1 15 PrSub3 19
DDVGL PrSub2 3
DTRAL PrSub2 13
DTRCL PrSub1 12
 . . .
KSDTR PrSub1 10
KSVHC PrSub3 15
etc.
```

Blast: Step 1 - Matching exact words

Match exact words in the query and indexed database

Query

```
>ProtQ
SKCDKSDTRALLAQYIPSTVNHPIL

SKCDK => 0
KCDKS => 0
CDKSD => 0
DKSDT => 0
```

Index sorted :

```
AAMLL PrSub1 5
ACILS PrSub3 2
...
CILST PrSub3 3
CLGYR PrSub1 15 PrSub3 19
DDVGL PrSub2 3
DTRAL PrSub2 13
DTRCL PrSub1 12
 . . .
KSDTR PrSub1 10
KSVHC PrSub3 15
etc.
```

Blast: Step 1 - Matching exact words

Match exact words in the query and indexed database

Query

```
>ProtQ
SKCDKSDTRALLAQYIPSTVNHPIL

SKCDK => 0
KCDKS => 0
CDKSD => 0
DKSDT => 0
KSDTR => PrSub1 10
```

```
    SKCDKSDTRALLAQYIPSTVNHPIL
EKFKAAMLLKSDTRCLGYRNVCKEG
```

Index sorted :

```
AAMLL PrSub1 5
ACILS PrSub3 2
...
CILST PrSub3 3
CLGYR PrSub1 15 PrSub3 19
DDVGL PrSub2 3
DTRAL PrSub2 13
DTRCL PrSub1 12
 . . .
KSDTR PrSub1 10
KSVHC PrSub3 15
etc.
```

Blast: Three main steps

Step 0 : Indexing the database

Step 1 : matching exact words

Step 2 : Extend the alignment

Step 3 : Compute alignment score

Blast: Step 2 - Extending the alignment

a) Extending the alignment (wo indel) => high-scoring segment pair

k=5

```
ProtQ        SKCDKSDTRALLAQYIPSTVNHPIL
PrSub1   EKFKAAMLLKSDTRCLGYRNVCKEG
```

Extend the segment while Score ≥ M (Threshold)

```
ProtQ        SKCDKSDTRALLAQYIPSTVNHPIL
PrSub1   EKFKAAMLLKSDTRCLGYRNVCKEG
```

```
ProtQ        SKCDKSDTRALLAQYIPSTV
PrSub1       AMLLKSDTRCLGYRNVCKEG
```

Blast: Step 2 - Extending the alignment

b) Grouping together HSPs

```
ProtQ      SKCDKSDTRALLAQYIPSTV
PrSub1     AMLLKSDTRCLGYRNVCKEG
```

```
Query: 27   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
            V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K         ++
Sbjct: 33   VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87   ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
             D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83   PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
                L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

Blast: Three main steps

Step 0 : Indexing the database

Step 1 : matching exact words

Step 2 : Extend the alignment

Step 3 : Compute alignment score

Blast: step 3: Compute alignment score

Compute the score S for each HSP by using a substitution matrix and gap penalty.

Compute the statistical significance of each HSP score by exploiting the Gumbel extreme value distribution

$$\text{E-value} = K_B * l_Q * 2^{-S}$$

where K_B depend on the database size

l_Q is the length of the Query

Blast: step 3: Compute alignment score

Interpretability of HSP scores

- The Expect value (E) is describes the number of HSP one can "expect" to find by chance when searching a database of a particular size.

- E-value = 3 means, If we compare a sequence to a random database of the same size and same composition as the original, we would expect to find 3 sequences with score S

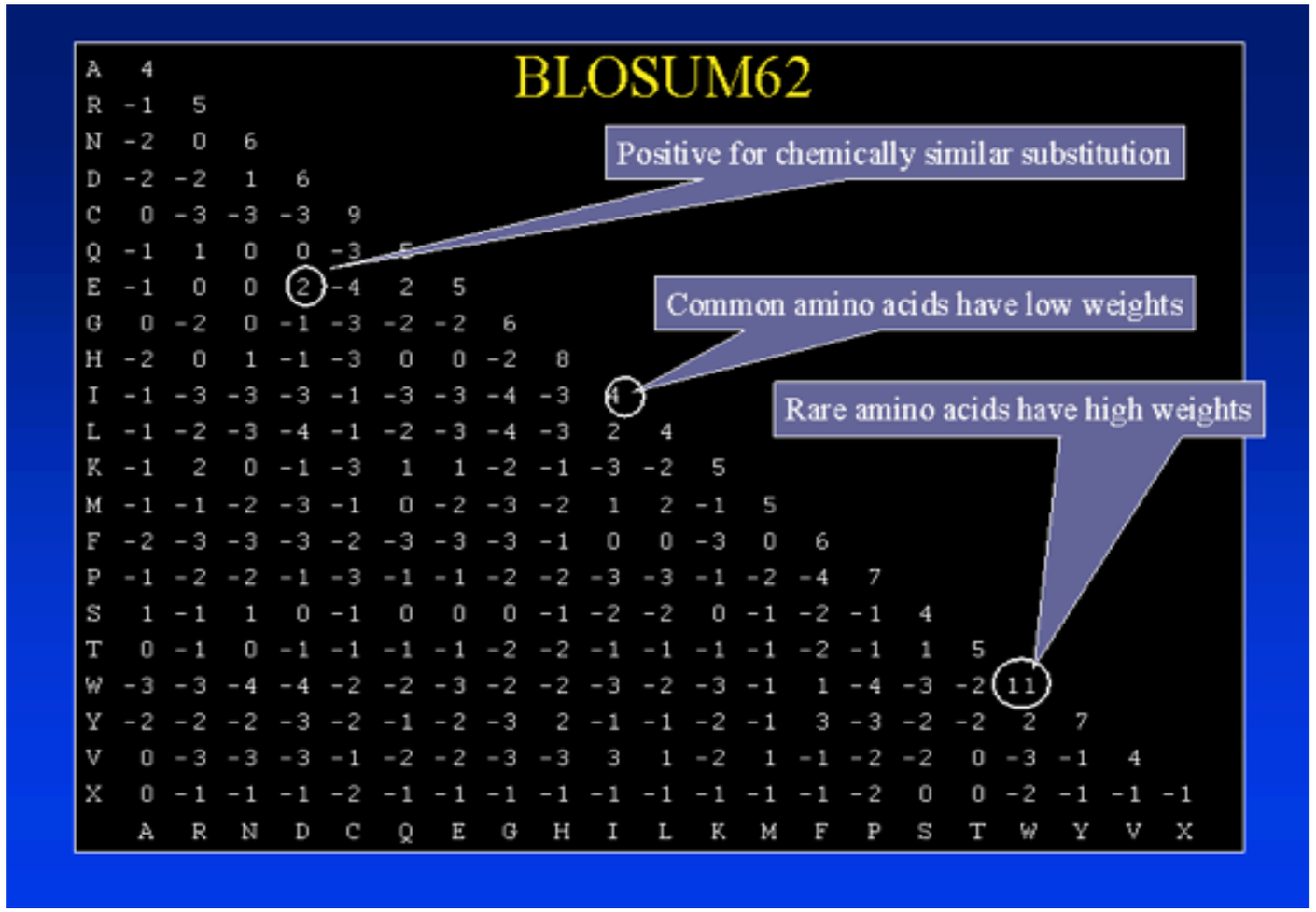
- The lower the E-value, or the closer it is to zero, the more "significant" the match is.

- The E-value depend on :
  - Substitution matrix;
  - The size and database

PSI-Blast

- Blast efficiency depend on the substitution matrix used to score HSP.

- It always uses the same matrix.



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

PSI-Blast

Sometimes a specific substitution matrix can improve the results

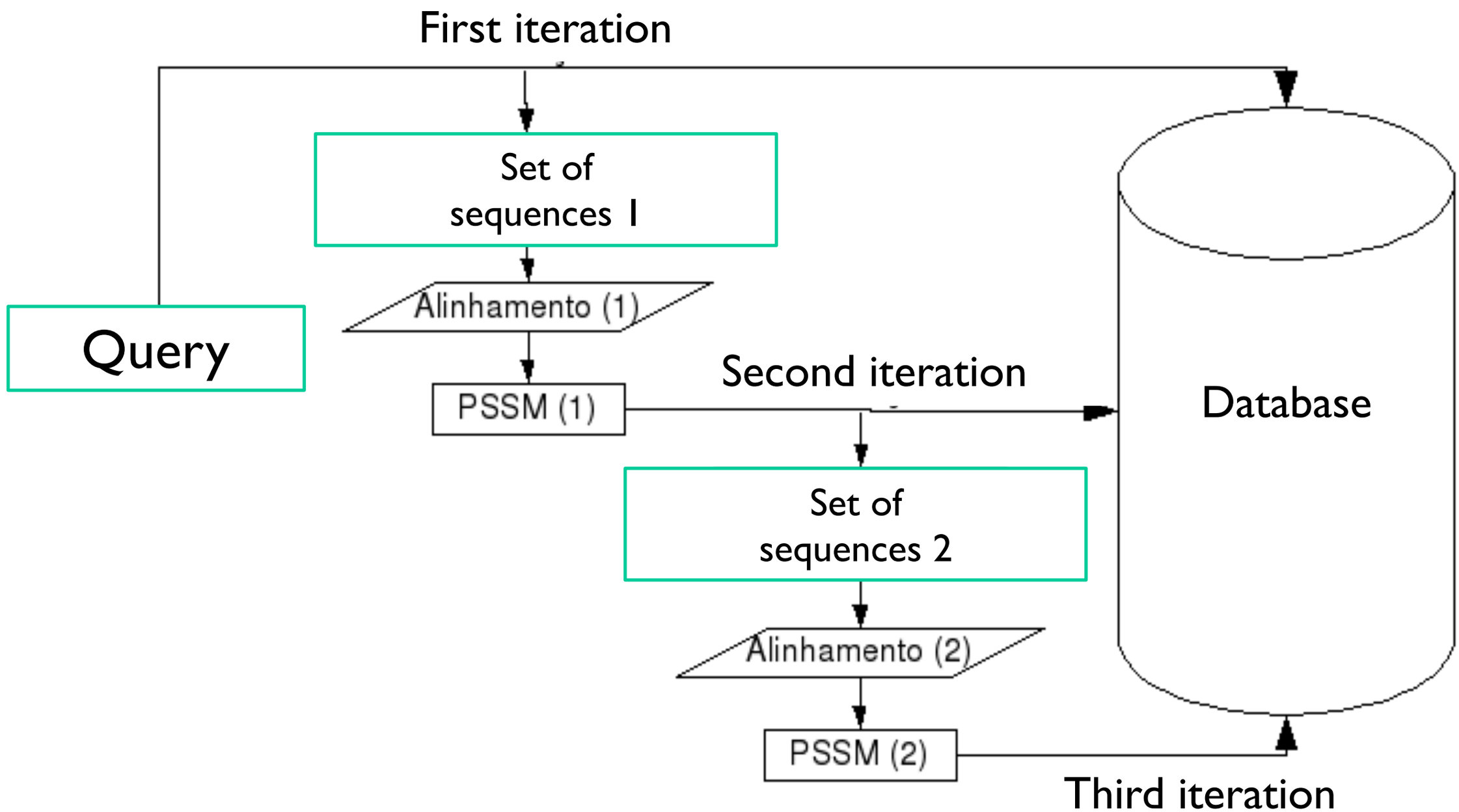
Position-Specific Scoring Matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	-7	-4	-4	-5	0	-4
213	N	-2	0	2	-1	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	-2	-5	-1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-4	-4	-6	-3	-5	-6	-6	-6
215	D	-5	-5	-2	9	-7	-4	-1	-5	-5	-7	-7	-4	-7	-7	-5	-4	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	[illegible]	6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218	G	-3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8	-6	-7	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219	P	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-6	-6	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4

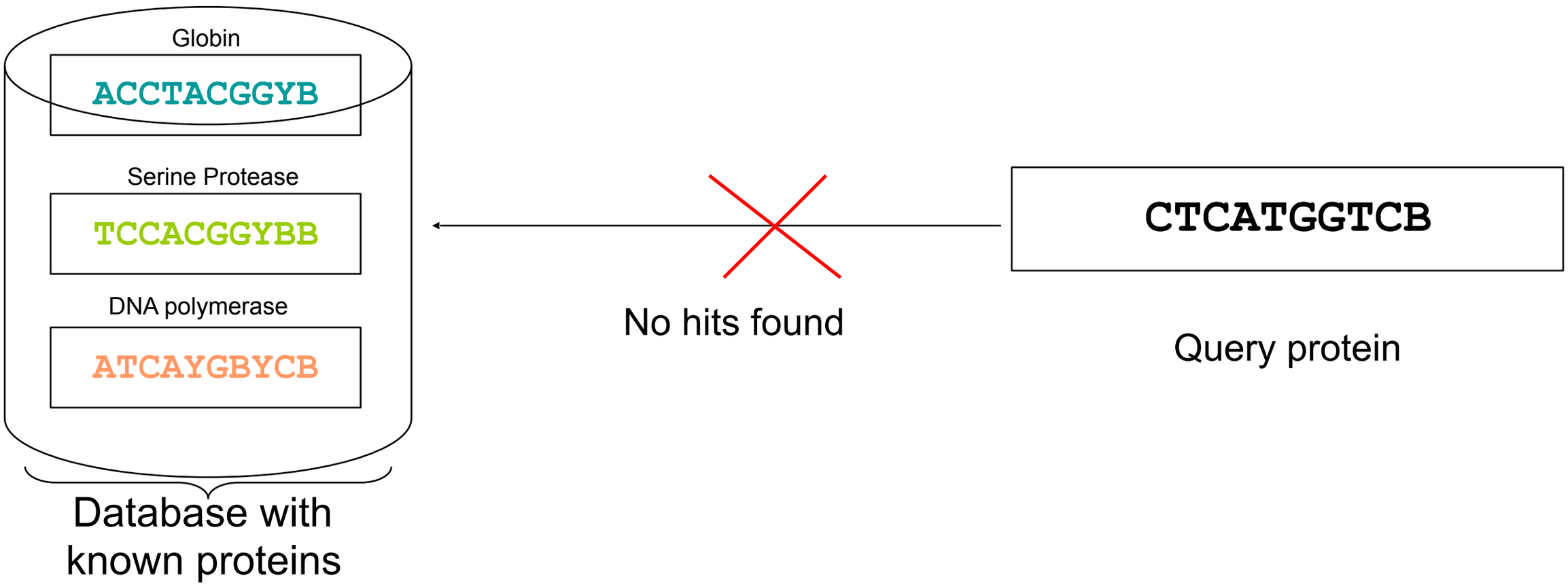
Serine scored differently in these two positions

Active site nucleophile

PSI-Blast


First iteration
Set of sequences 1
Alinhamento (1)
Query
PSSM (1)
Second iteration
Database
Set of sequences 2
Alinhamento (2)
PSSM (2)
Third iteration

Protein annotation on highly divergent sequences

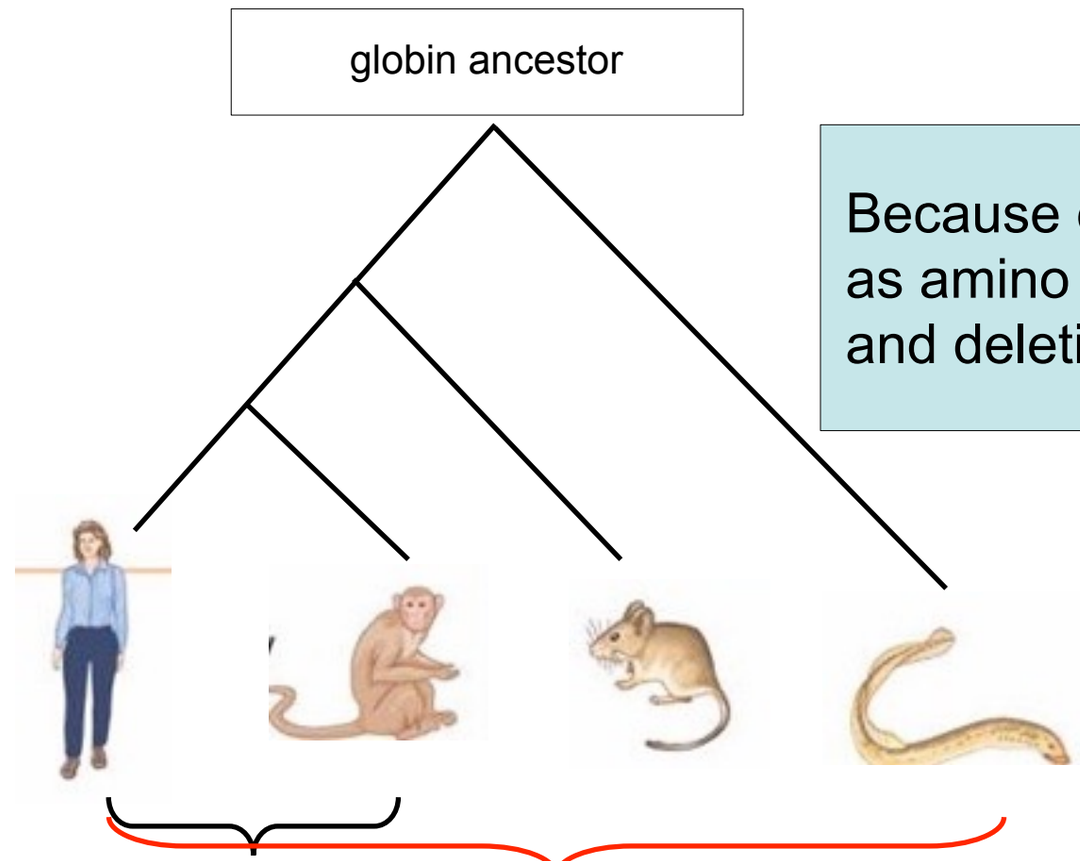


Protein annotation on highly divergent sequences

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

50% of *P. faciparum* genes have not known function

Protein annotation on highly divergent sequences



Because evolutionary events such as amino acid mutation, insertion and deletion.

Proteins of close species are more similar

Percentage of identical amino acids

95%

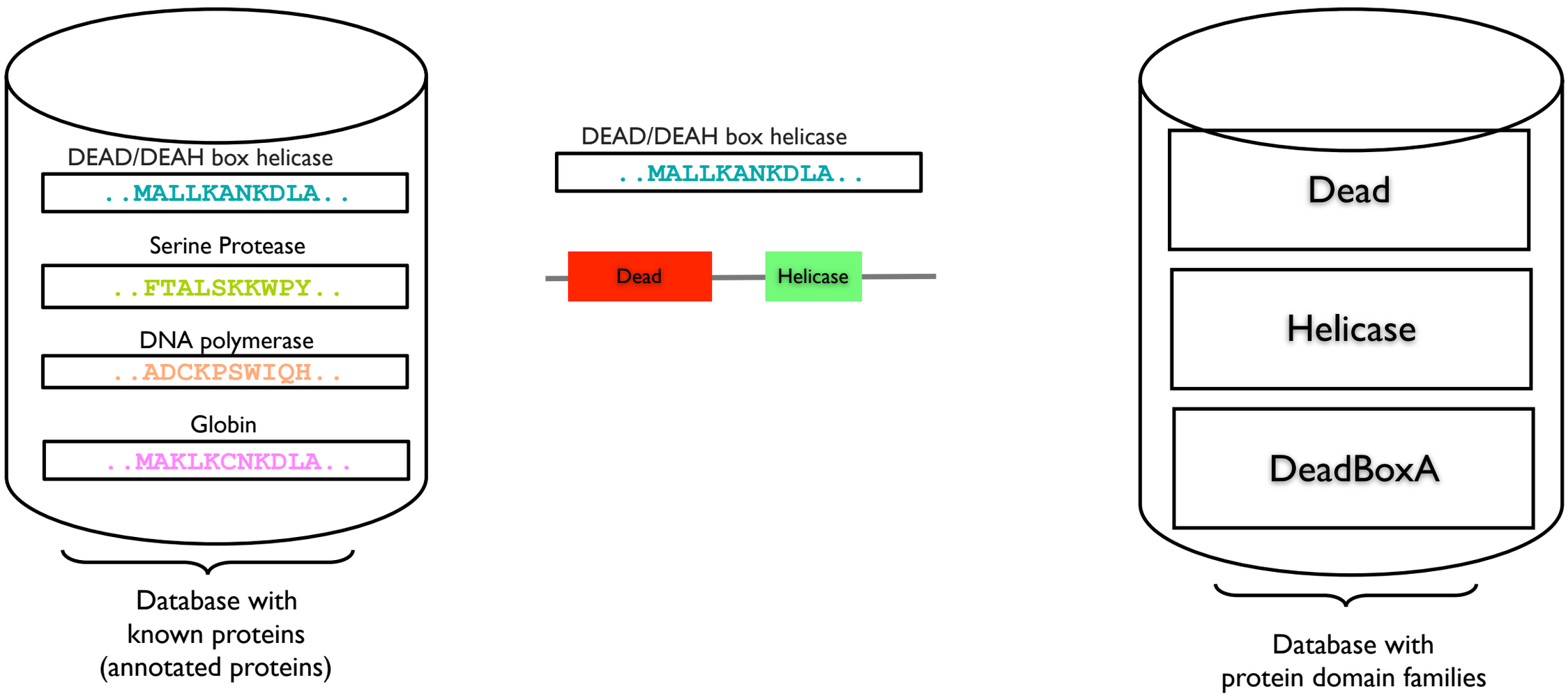
14%

Remote homologous proteins

There are homologous proteins with a weak similarity

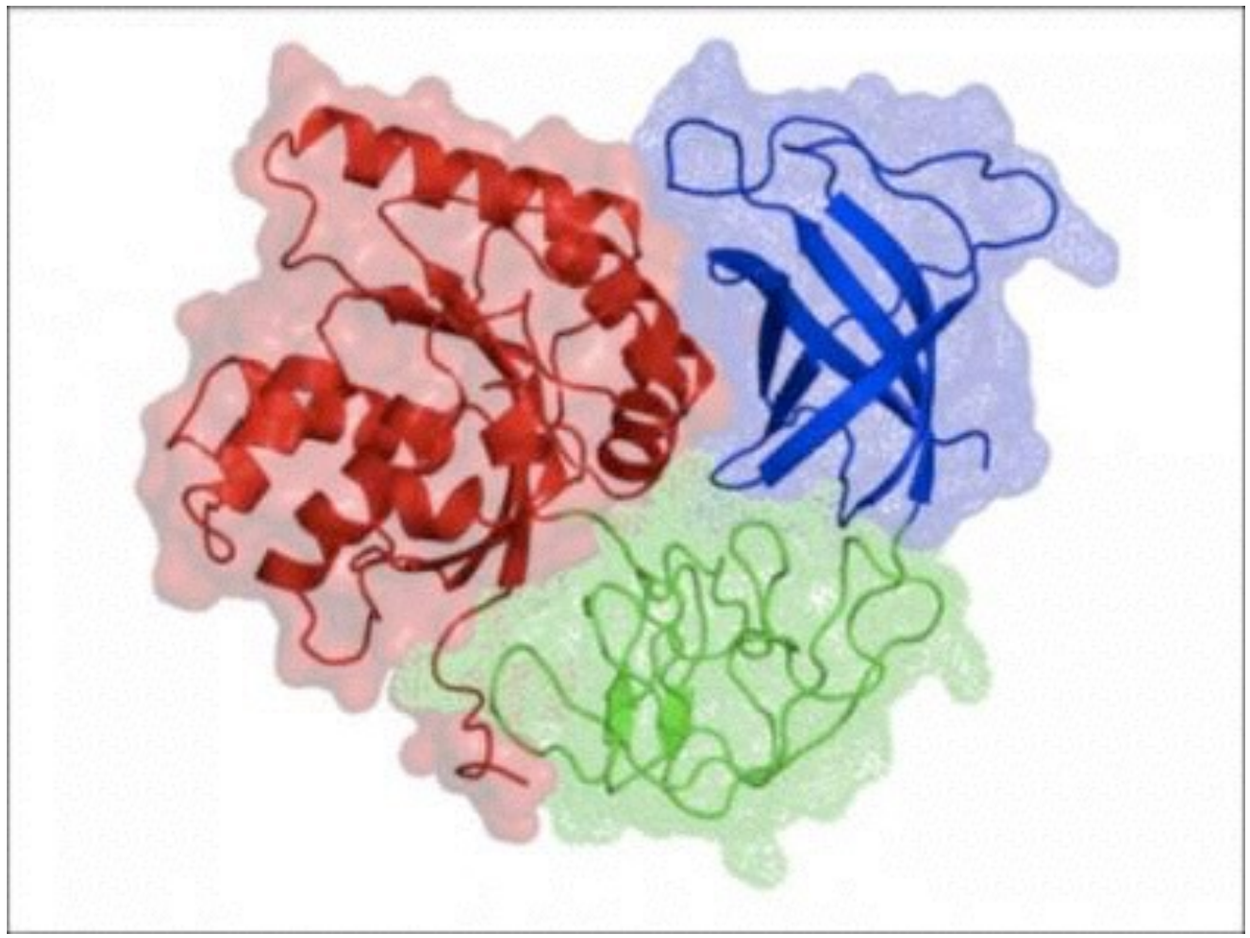
Protein Domais

- To improve protein annotation we can :
  - classify known protein sequences according to their functional regions (**domains**)
  - Search for conserved domains instead of conserved sequences



Protein Domains

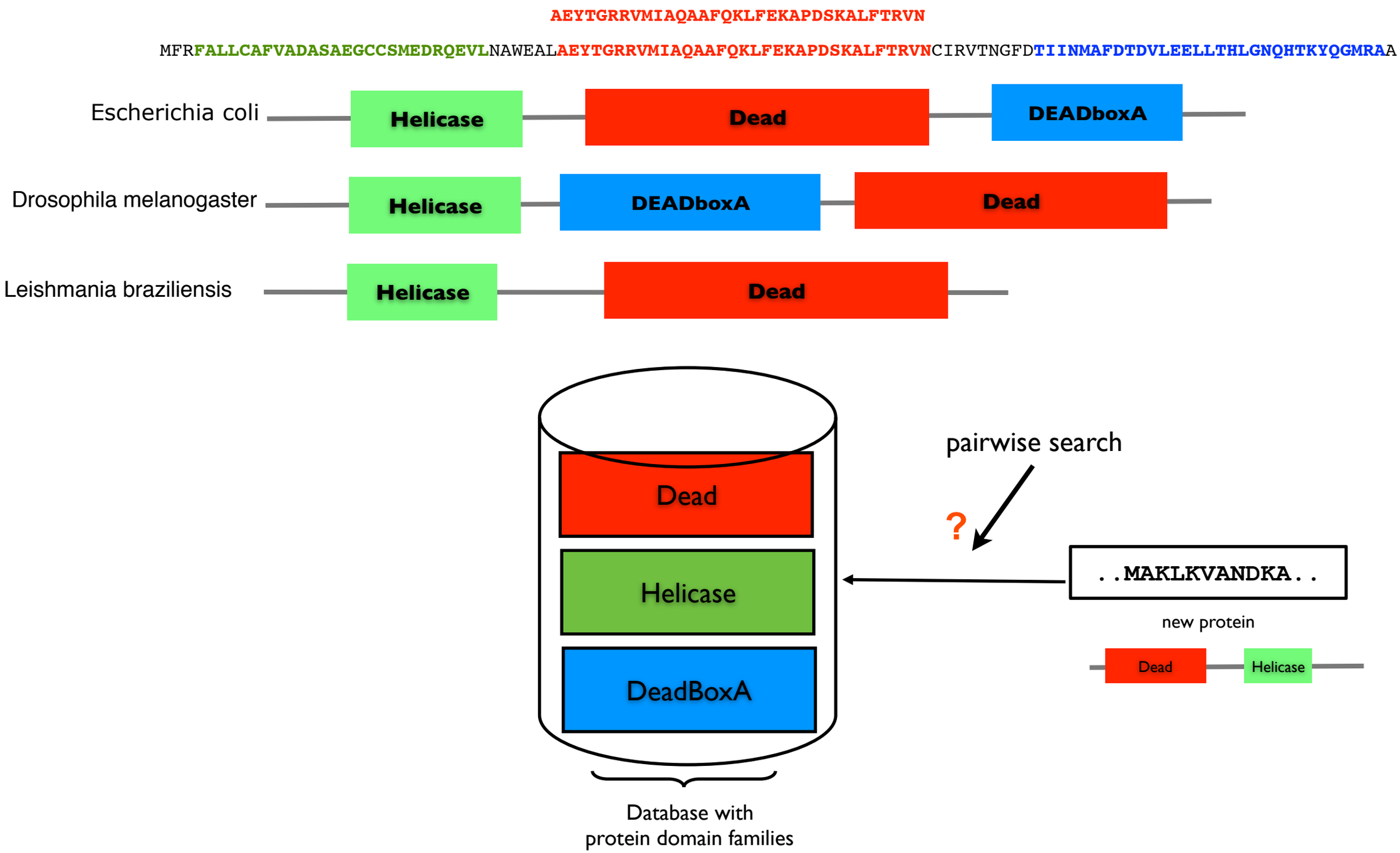
- Domains are the building blocks of proteins.



MFRFALLCAFVADASAEGCCSMEDRQEVLNAWEALWSAEYTGRRVMIAQAAFQKLFEKAPDSKALFTRVNVDNIGSPQFRAHCIRVTNGFDTIINMAFDTDVLEELLTHLGNQHTKYQGMRAA

Domain Recognition

- Identifying domains can help to determine protein function.



Domain Databases

EMBL-EBI

Pfam

keyword search Go

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS

SEQUENCE SEARCH

VIEW A PFAM ENTRY

VIEW A CLAN

VIEW A SEQUENCE

VIEW A STRUCTURE

KEYWORD SEARCH

JUMP TO

VIEW PFAM ANNOTATION AND ALIGNMENTS

Enter a entry identifier (e.g. *Piwi*) or accession (e.g. *PF02171*) to see all data for that entry.

PF00018 Go Example

You can also browse through the list of all Pfam families.

Family: *SH3_1* (PF00018)

695 architectures | 10749 sequences | 11 interactions | 444 species | 373 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Summary: SH3 domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: SH3 domain | Pfam | InterPro

This is the Wikipedia entry entitled "SH3 domain". **More...**

SH3 domain

Edit Wikipedia article

The **SRC Homology 3 Domain** (or **SH3 domain**) is a small protein domain of about 60 amino acids residues first identified as a conserved sequence in the viral adaptor protein v-Crk and the non-catalytic parts of enzymes such as phospholipase and several cytoplasmic tyrosine kinases such as Abl and Src.[1][2] It has also been identified in several other protein families such as: PI3 Kinase, Ras GTPase-activating protein, CDC24 and cdc25.[3][4][5] SH3 domains are found in proteins of signaling pathways regulating the cytoskeleton, the Ras protein, and the Src kinase and many others. They also regulate the activity state of adaptor proteins and other tyrosine kinases and are thought to increase the substrate specificity of some tyrosine kinases by binding far away from the active site of the kinase. Approximately 300 SH3 domains are found in proteins encoded in the human genome.

SH3 domain

Ribbon diagram of the SH3 domain, alpha spectrin, from chicken (PDB accession code 1SHG), colored from blue (N-terminus) to red (C-terminus).

Contents [hide]

Identifiers	
Symbol	SH3_1
Pfam	PF00018
Pfam clan	CL0010
InterPro	IPR001452
SMART	SM00326
PROSITE	PS50002
SCOP	1shf
SUPERFAMILY	1shf
CDD	cd00174

Available protein structures: [show]

Structure

The SH3 domain has a characteristic beta-barrel fold that consists of five or six β-strands arranged as two tightly packed anti-parallel β sheets. The linker regions may contain short helices. The SH3-type fold is an ancient fold found in eukaryotes as well as prokaryotes.[6]

Peptide binding

The classical SH3 domain is usually found in proteins that interact with other proteins and mediate assembly of specific protein complexes, typically via binding to proline-rich peptides in their respective binding partner. Classical SH3 domains are restricted in humans to intracellular proteins, although the small human MIA family of extracellular proteins also contain a domain with an SH3-like fold.

Many SH3-binding epitopes of proteins have a consensus sequence that can be represented as a regular expression or Short linear motif:

PFAM Database

Family: *SH3_1* (PF00018)

695 architectures | 10749 sequences | 11 interactions | 444 species | 373 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to...

enter ID/acc Go

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four representative proteomes (RP) sets, the NCBI sequence database, and our metagenomics sequence database. **More...**

View options

We make a range of alignments for each Pfam-A family. You can see a description of each **above**. You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (61)	Full (10749)	Representative proteomes				NCBI (20245)	Meta (89)
			RP15 (1639)	RP35 (2410)	RP55 (4041)	RP75 (5929)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	–	✓	✓	✓	–	✗	✗
PP/heatmap	✗[1]	–	✓	✓	✓	–	✗	✗
Pfam viewer	✓	✓	✗	✗	✗	✗	✗	✗

[1]Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, ✗ not generated, – not available.

Format an alignment

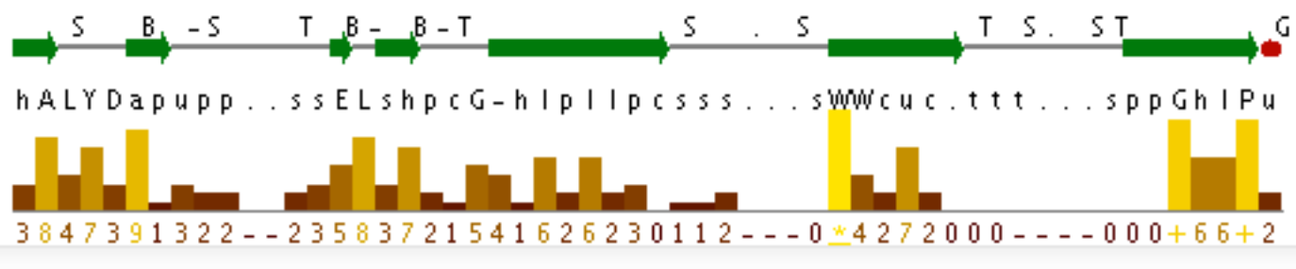
PFAM Database

EMBL-EBI

View seed alignment for *PF00018* using Jalview

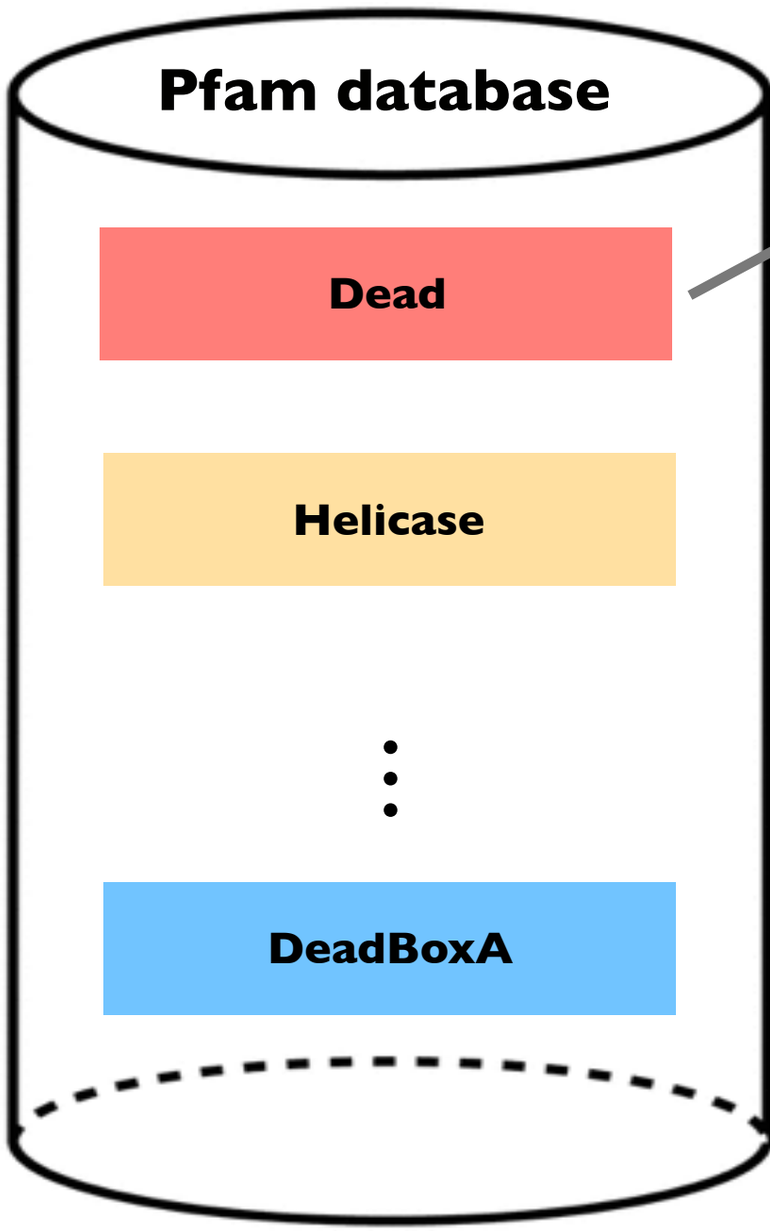
File Edit Select View Format Colour Calculate Help

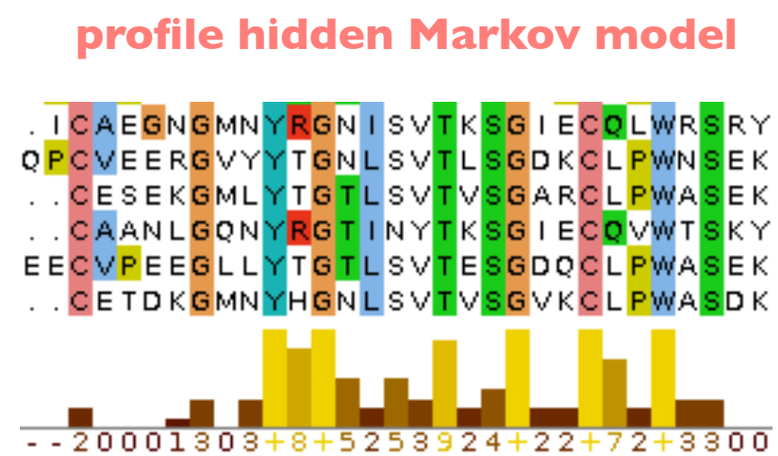
```
                                   10        20        30        40        50
SCD2_SCHPO/30-78          RALYDYTARK..ATEVSFAKGDFFHVIGREND..KAWYEVCNPAA...GTRGFVPV
TEC_HUMAN/185-231         VAMYDFQAAE..GHDLRLERGQEYLILEKNDV...HWWRARDKYG...N.EGYIPS
RASA1_BOVIN/282-330       RAILPYTKVP.DTDEISFLKGDMFIVHNELEDG.WMWVTNLRTD.....EQGLIVE
NCF2_HUMAN/246-291        RVLFGFVPET..KEELQVMPGNIVFVLKKGND...NWATVMFNG.....QKGLVPC
PLCG1_BOVIN/797-843       KALFDYKAQR..EDELTFTKSAIIQNVEKQEG...GWWRGDYGG....KKQLWFPS
MYSB_ACACA/1096-1141      KALYDYDAQT..GDELTFKEGDTIIVHQKDPA...GWWEGELN.....GKRGWVPA
YKA7_CAEEL/277-322        TAIYDYNSNE..AGDLNFAVGSQIMVTARVNE...EWLEGECF.....GRSGIFPS
LASP1_CAEEL/272-319       KAIYDYAAAD..KDEISFLEGDIIVNCEKIDD...GWMTGTVQRT...LQWGMLPA
BEM1_YEAST/161-209        IVLYDFKAEK..ADELTTYVGENLFICAHHNC...EWFIAKPIGR..LGGPGLVPV
CSK_CHICK/15-62           IAKYNFHGTA..EQDLPFSKGDVLTIVAVTKD..PNWYKAKNKV....GREGIIPA
HCLS1_HUMAN/434-479       VAVYDYQGEG..SDELSFDPDDVITDIEMVDE...GWWRGRCH.....GHFGLFPA
MYSC_ACACA/982-1027       RALYDFAAEN..PDELTFNEGAVVTVINKSNP...DWWEGELN.....GQRGVFPA
SEM5_CAEEL/160-205        QALFDFNPQE..SGELAFKRGDVITLINKDDP...NWWEGQLN.....NRRGIFPS
LYN_HUMAN/69-115          VALYPYDGIH..PDDLSFKKGEKMKVLE.EHG...EWWKAKSLLT...KKEGFIPS
STK_HYDVU/65-112          VALYDYEARI..SEDLSFKKGERLQIINTADG...DWWYARSLIT...NSEGYIPS
BUD14_YEAST/265-312       YALYAFNGHD..SSHCQLGQDEPCILLNDQDA...YWWLVKRITD...GKIGFAPA
BLK_MOUSE/58-104          VALFDYAAVN..DRDLQVLKGEKLQVLR.STG...DWWLARSLVT...GREGYVPS
SRK1_SPOLA/60-108         VGKYDYDSRT..DDDLSFKKGDLMYIISTDEG...DWWFARSKDT..AGKEGYIPS
ABP1_SACEX/563-610        TAEYDYEAGE..DNELTFAENDKIINIEFVDD...DWWLGELETT...GQKGLFPS
                          KAKYSYQAQT..SKELSFMEGEFFYVSGDEKD....WYKASNPST...GKEGVVPK
secondary structure       VALYDFVASG..DNTLSITKGEKLRVLGYNHN..GEWCEAQTKNG.....QGWVPS
seq                       VALYPFKAIE..GGDLSLEKNAEYEVIDDSQE...HWWKVKDALG...N.VGYIPS
                          IALYEYDAMQ..PDELTFKENDVINLIKKVDA...DWWQGELVRT...KQIGMLPS
                          IAKHDFSATA..DDELSFRKTQILKILNMEDD..SNWYRAELDGK.....EGLIPS
Conservation              VALYDYMPMN..ANDLQLRKGDEYFILEESNL...PWWRARDKN....GQEGYIPS
                          IALYDYEART..EDDLTFTKGEKFHILNNTEG...DWWEARSLSS...GKTGCIPS
                          EALFSYEATQ..PEDLEFQEGDIILVLSKVNE...EWLEGECK.....GKVGIFPK
                          QALYPFSSSN..DEELNFEKGDVMDVIEKPEND.PEWWKCRKING....MVGLVPK
                          RAIANYEKTS..GSEMALSTGDVVEVVEKSES...GWWFCQMKA.....KRGWIPA
                          MALYDFQARS..PREVTMKKGDVLTLLSSINK...DWWKVEAAD.....HQGIVPA
```



Successfully loaded file http://pfam.xfam.org/family/PF00018/alignment/seed

PFAM Database





?

..EKTQVTNLWGKP..

new protein / query sequence

Profile HMMs

Multiple sequence alignment

ACCTACGG-YB-
-TCCACGG-YBB
ATC-AY-GBYC-

Learning phase

2 How to estimate the parameters?

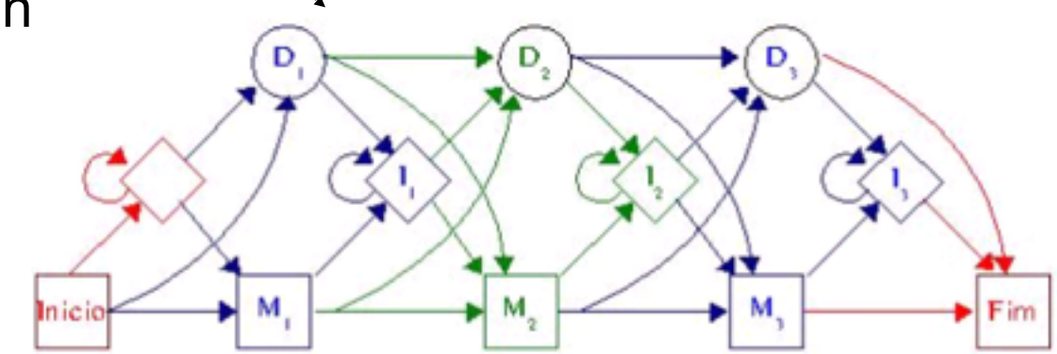
pHMM

Inference

CTCATGGTCB

query protein

1 How to define the architecture?

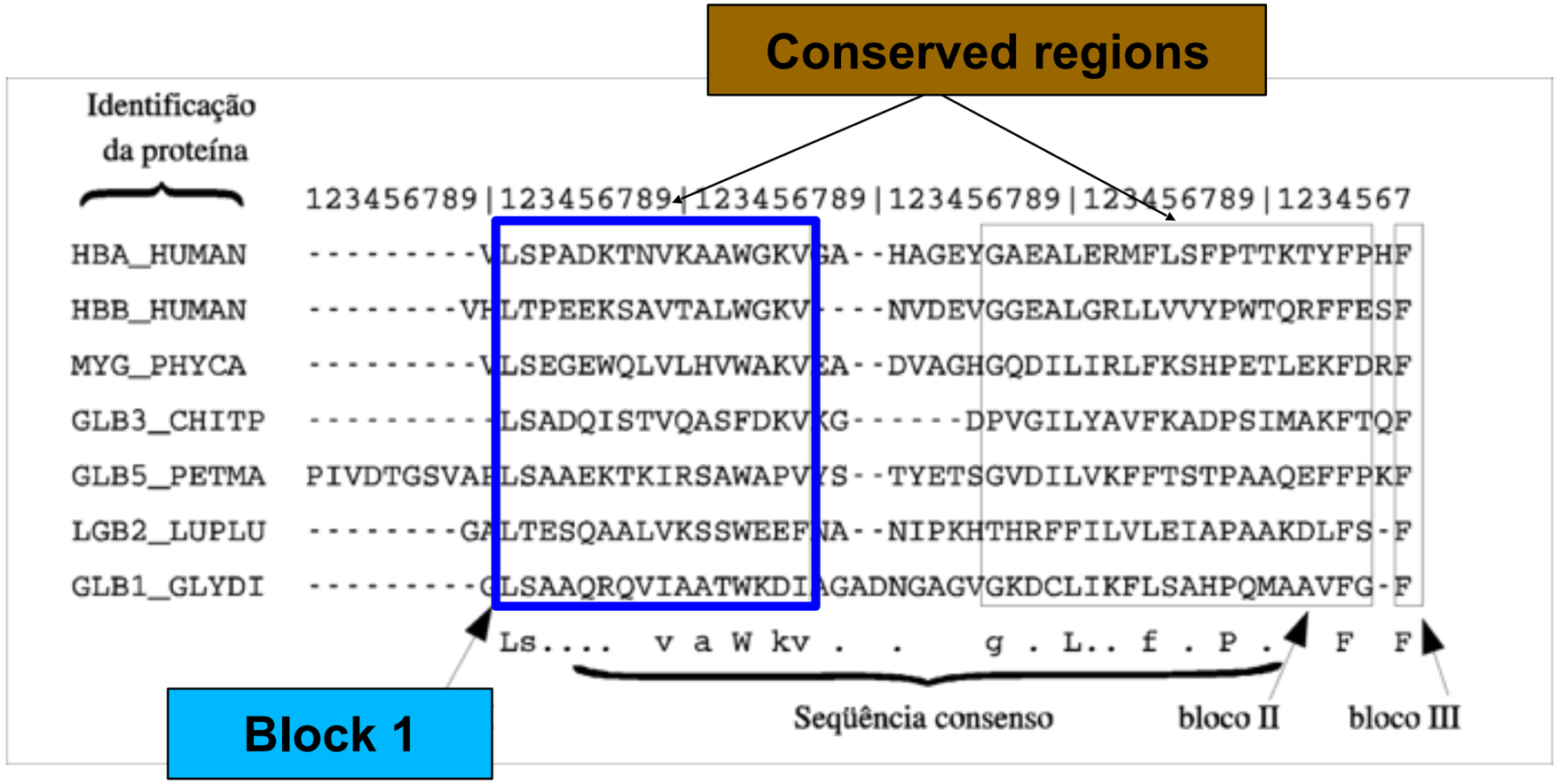


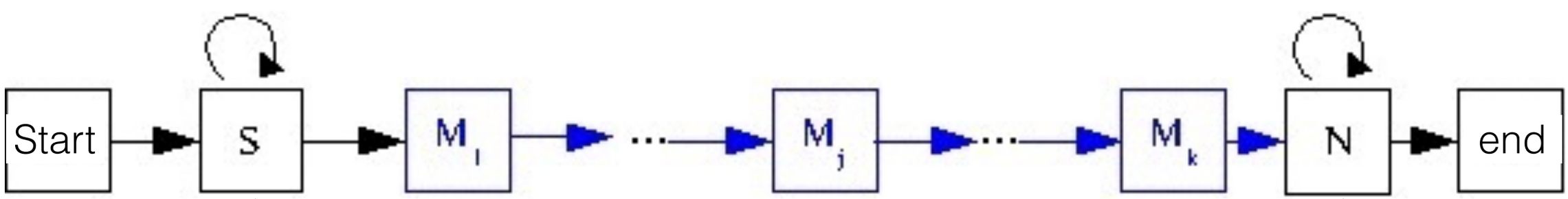
3 How to infer?

Viterbi

forward

1 How to define the HMM architecture to represent multiple sequence alignment conservation?



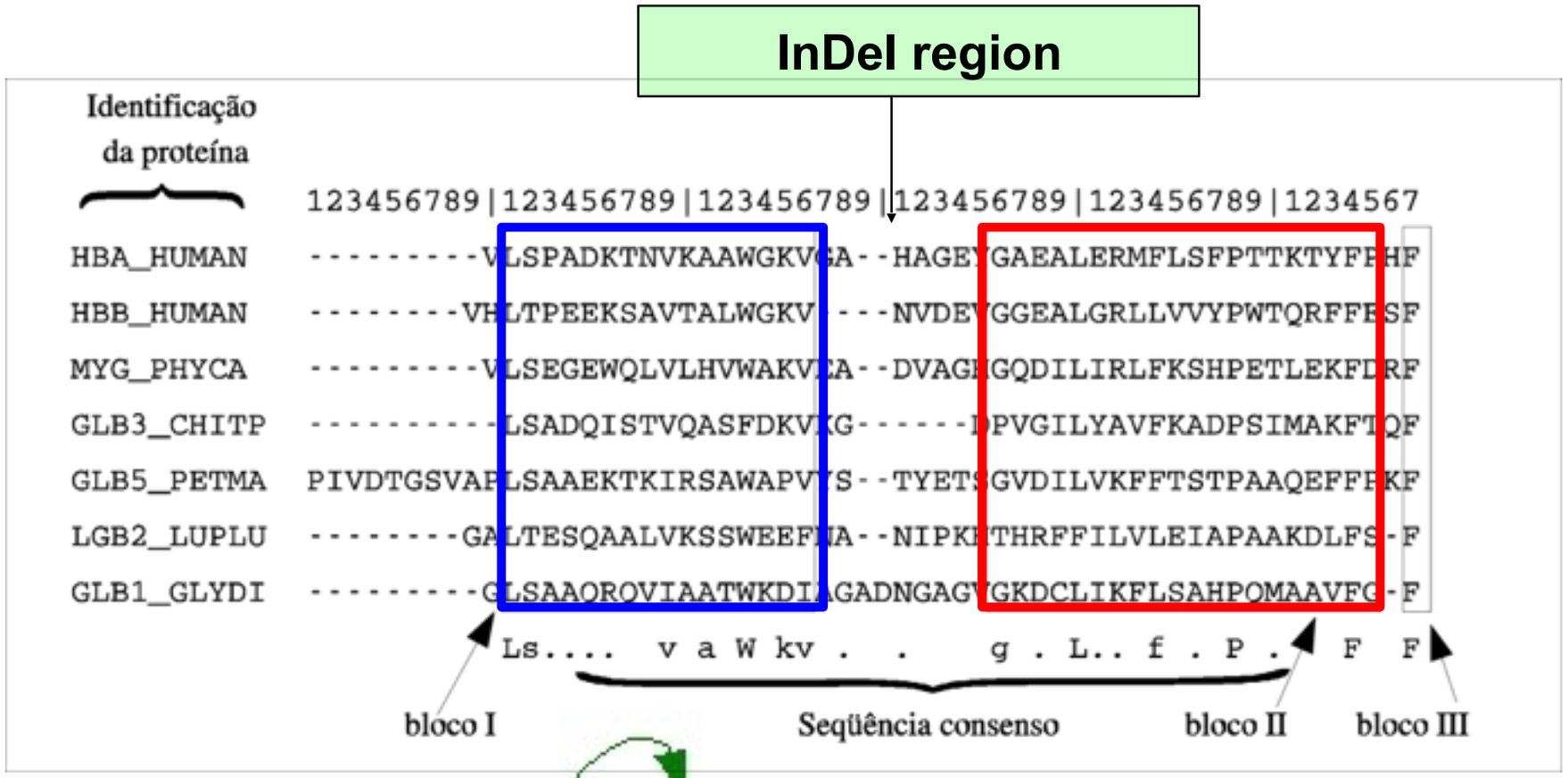


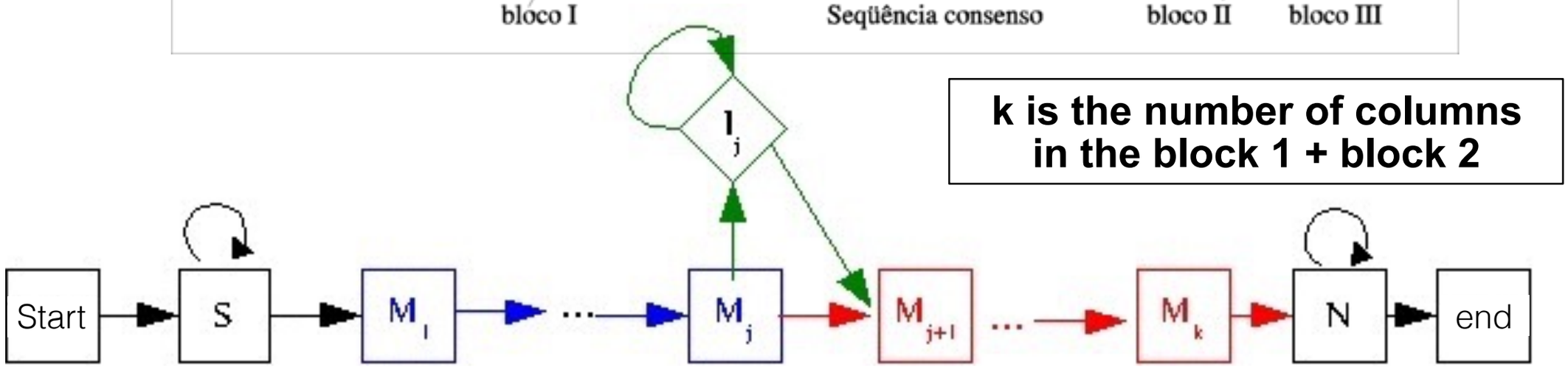
State S for amino acids before M_1

K is the number of columns in the block 1

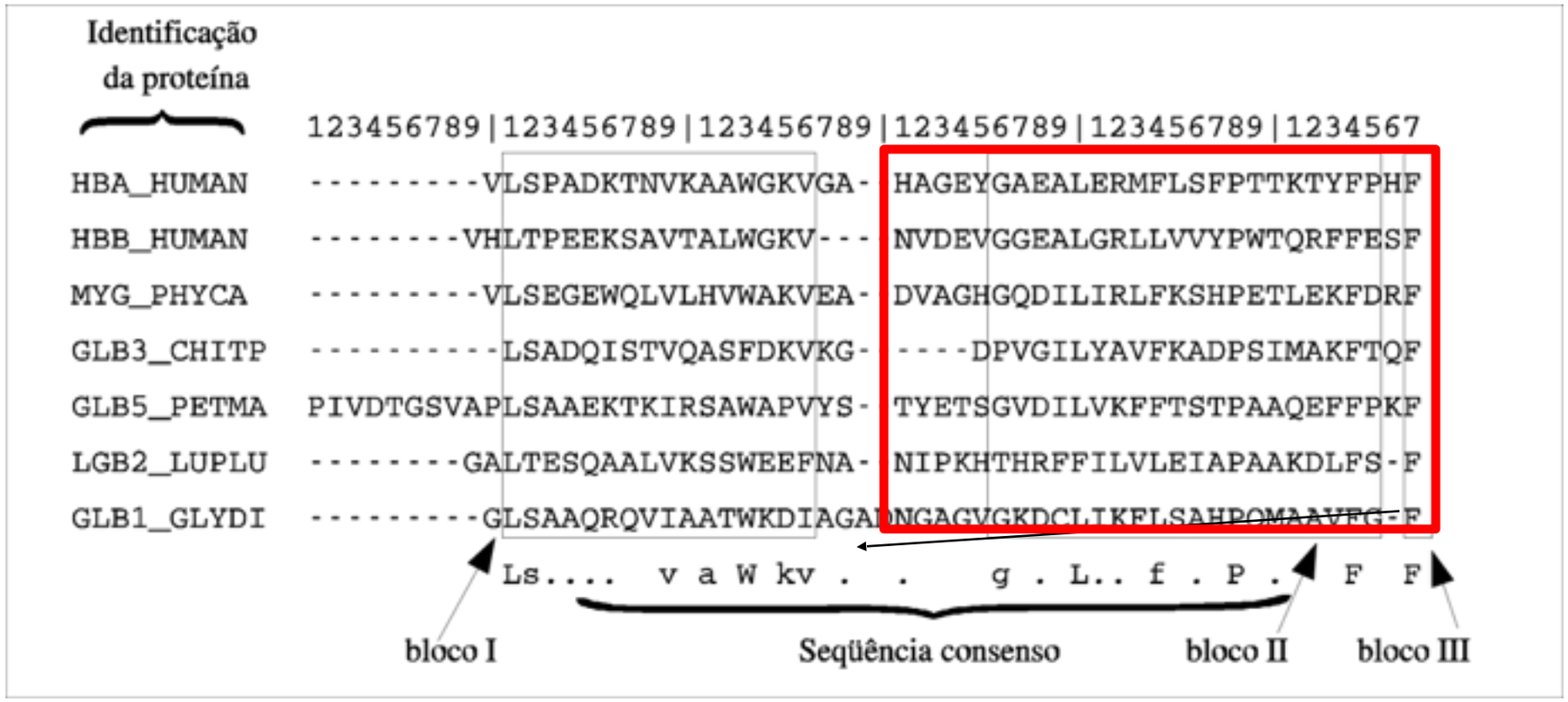
State N for amino acids after M_k

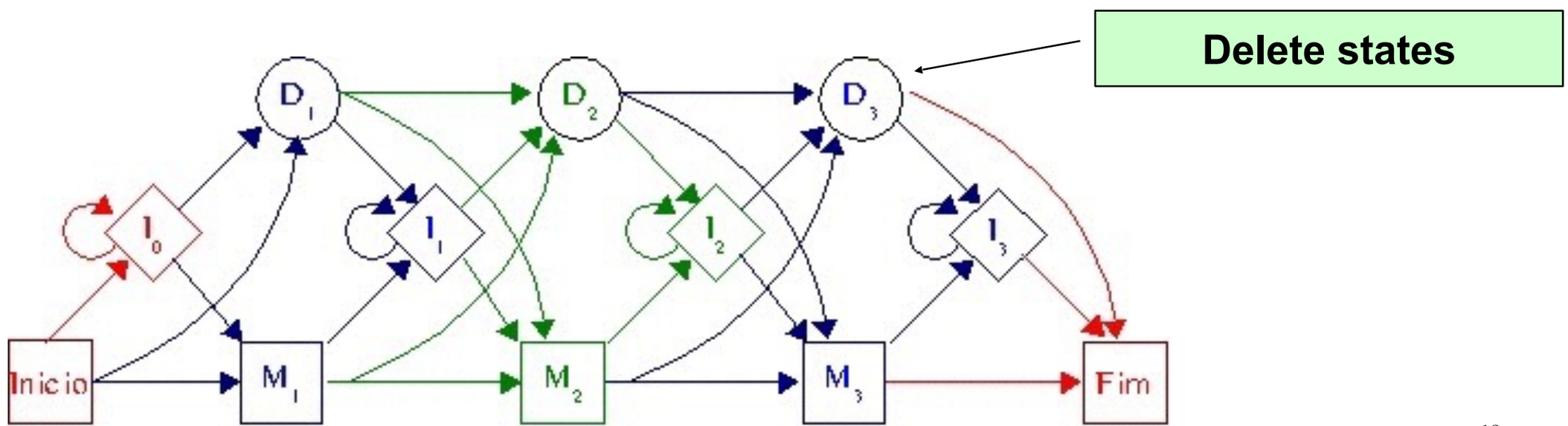
1 How to define the HMM architecture to represent multiple sequence alignment conservation?





1 How to define the HMM architecture to represent multiple sequence alignment conservation?





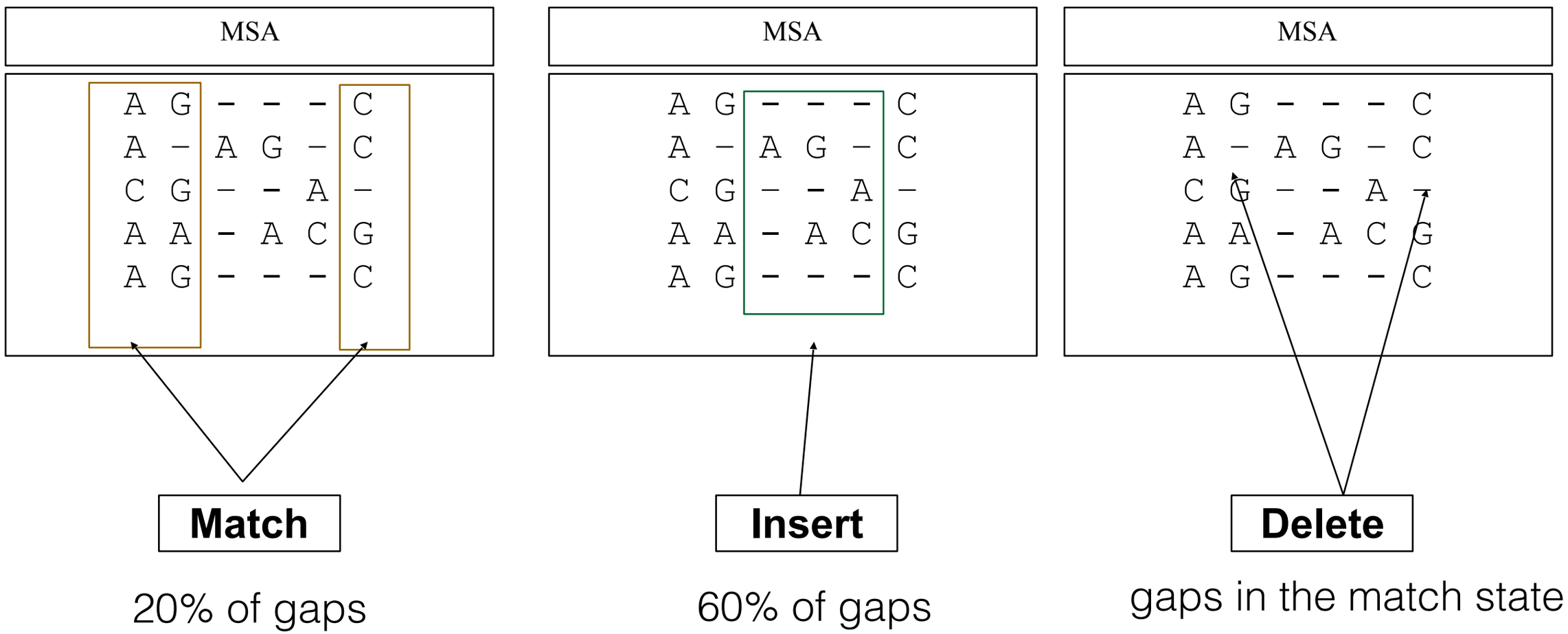
2

How to estimate the parameters?

- No aligned sequences =>Baum-Welch (BAUM, 1972)
- Aligned sequences:
  - Estimate *Match/Insert* states
  - Learn the probabilities by counting

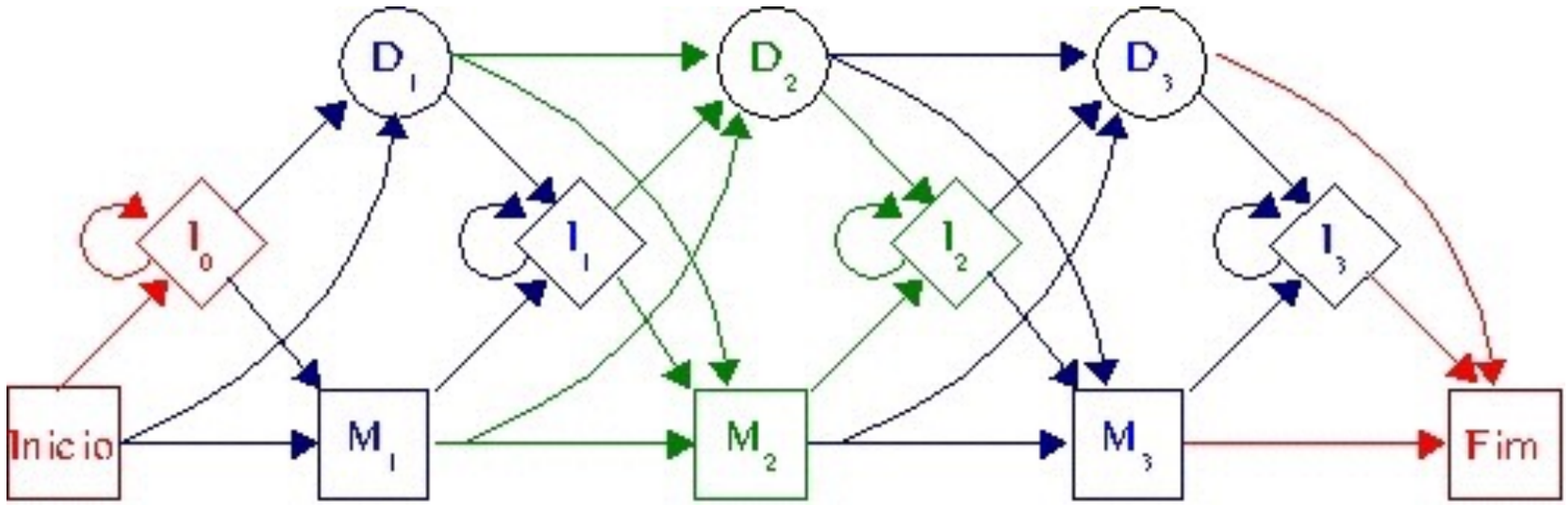
2 How to estimate the parameters?

- Assign cols >= w% of gaps as insert state, otherwise match
  - Example:
    - + 50% of gaps = insert and – 50% of gaps = match

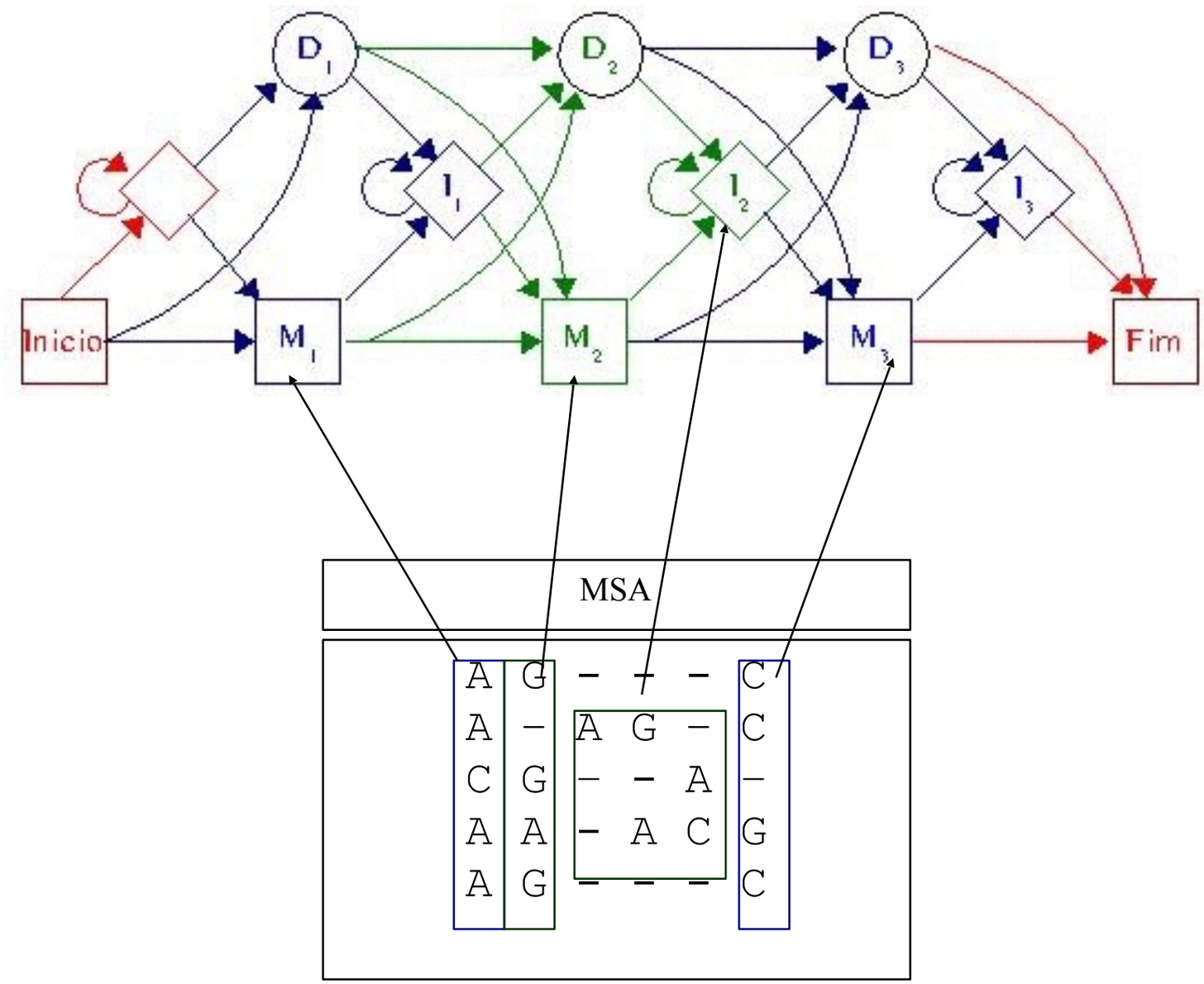


2 How to estimate the parameters?

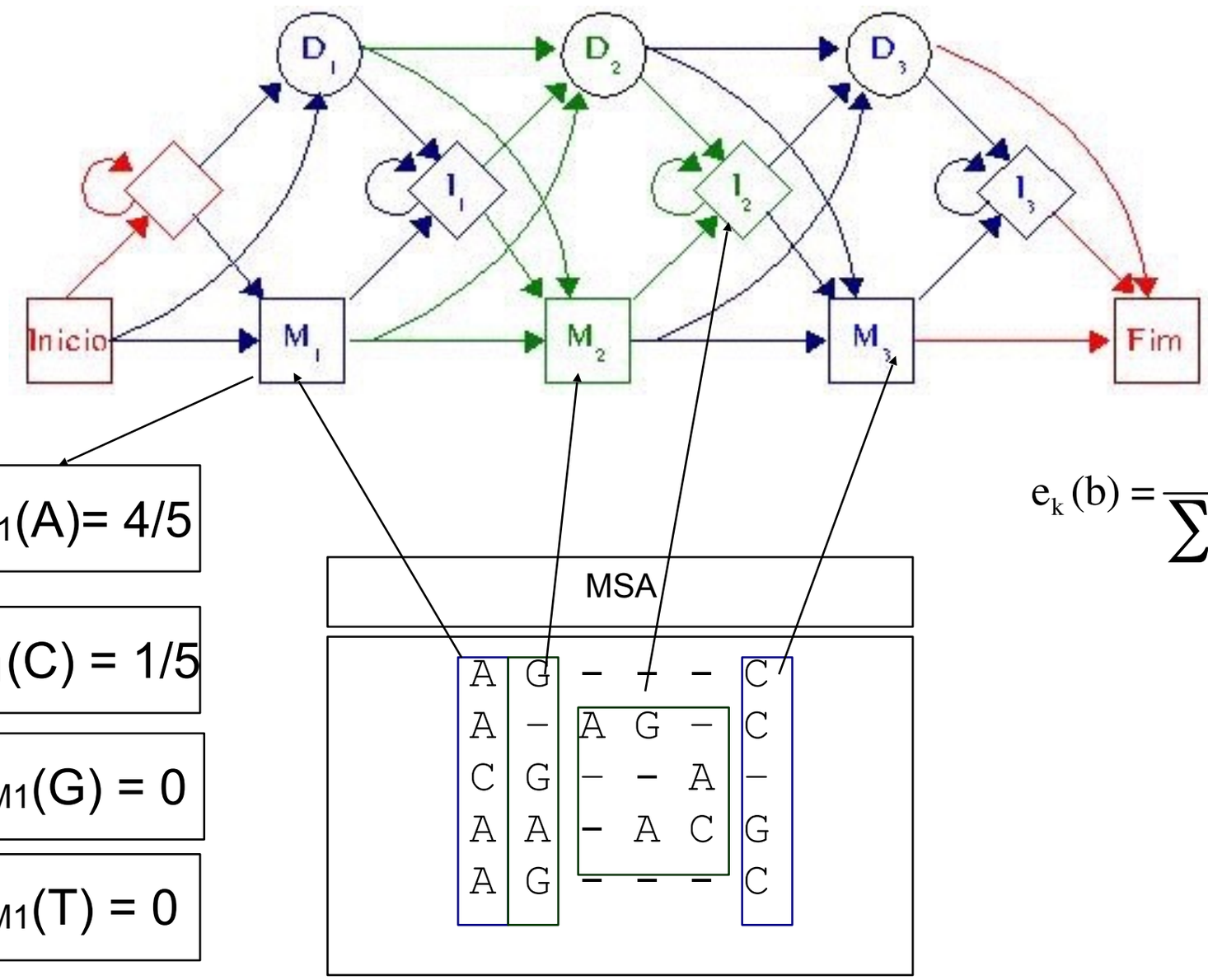
- Match and Insert states have **emission probabilities**.
- Delete states are silence.
- Arrows represent the **transition probabilities**.



- Match and Insert states



Emission probabilities



$e_{M1}(A) = 4/5$

$e_{M1}(C) = 1/5$

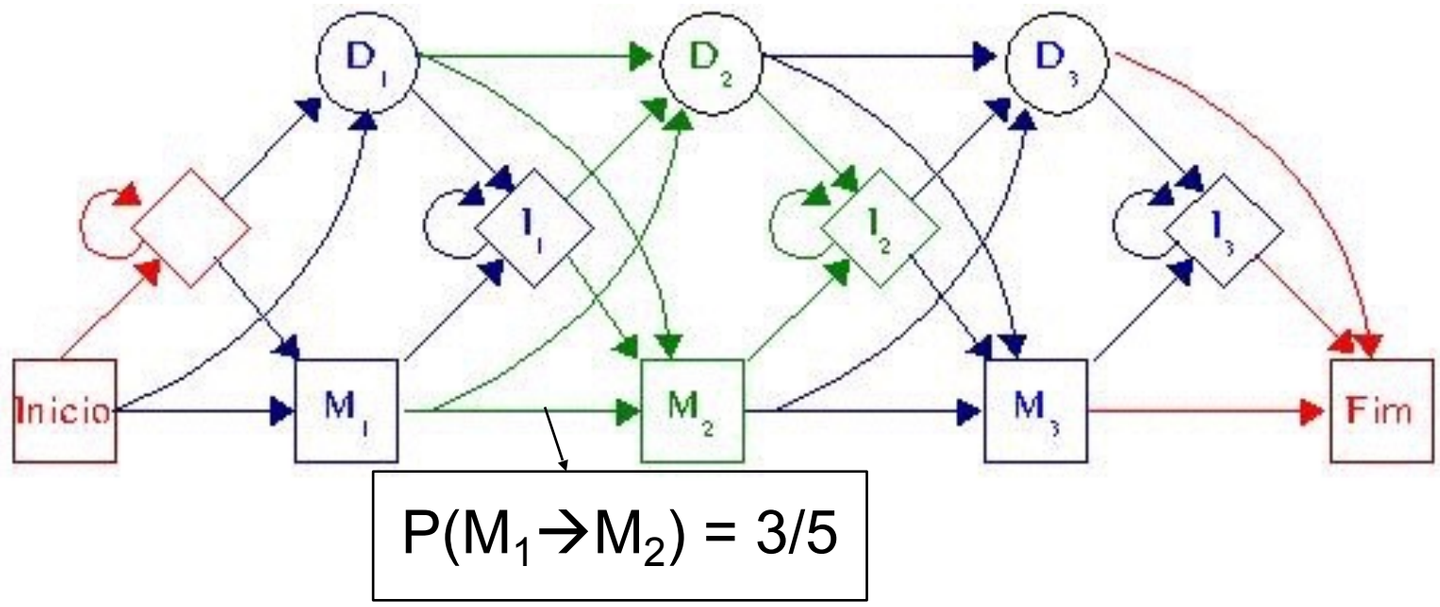
$e_{M1}(G) = 0$

$e_{M1}(T) = 0$

$$e_k(b) = \frac{E_k(b)}{\sum_{\sigma \in \Sigma} E_k(\sigma)}$$

We can use pseudo-count to avoid zero probabilities

Transition probabilities



$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}}$$

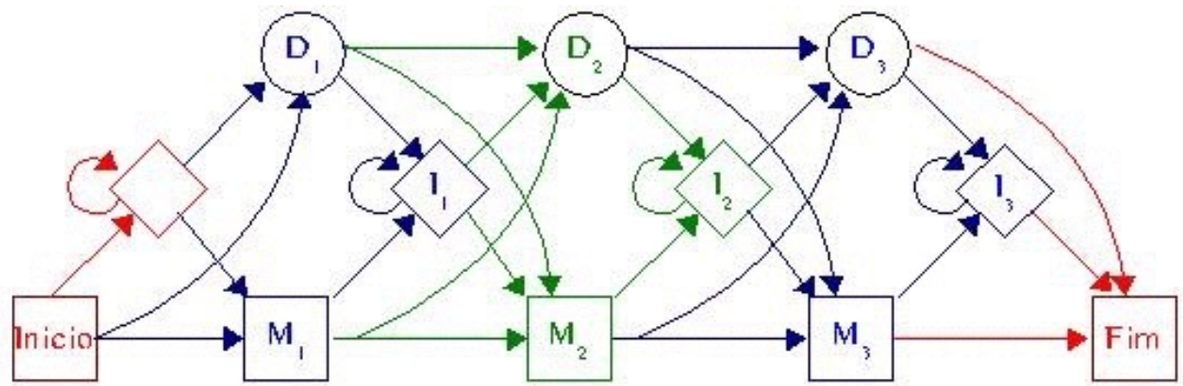
Alinhamento de entrada
A G - - - C
A - A G - C
A G - - A -
- A - A C G
A G - - - C

Alinhamento de entrada
M1 M2 - - - M3
M1 D2 I2 I2 - M3
M1 M2 - - I2 D3
D1 M2 - I2 I2 M3
M1 M2 - - - M3

A	G	-	-	-	C
A	-	A	G	-	C
A	G	-	-	A	-
-	A	-	A	C	G
A	G	-	-	-	C

M1	M2	-	-	-	M3
M1	D2	I2	I2	-	M3
M1	M2	-	-	I2	D3
D1	M2	-	I2	I2	M3
M1	M2	-	-	-	M3

$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}}$$



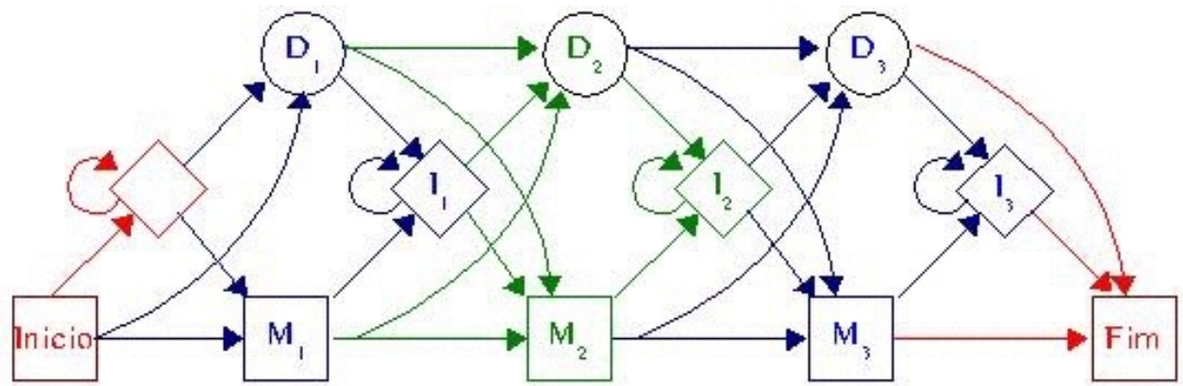
$$a_{M1M2} = \frac{A_{M1M2}}{A_{M1,M2} + A_{M1,I1} + A_{M1,D2}}$$

Match transitions **Insert transitions** **Delete transitions**

A	G	-	-	-	C
A	-	A	G	-	C
A	G	-	-	A	-
-	A	-	A	C	G
A	G	-	-	-	C

M1	M2	-	-	-	M3
M1	D2	I2	I2	-	M3
M1	M2	-	-	I2	D3
D1	M2	-	I2	I2	M3
M1	M2	-	-	-	M3

$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}}$$

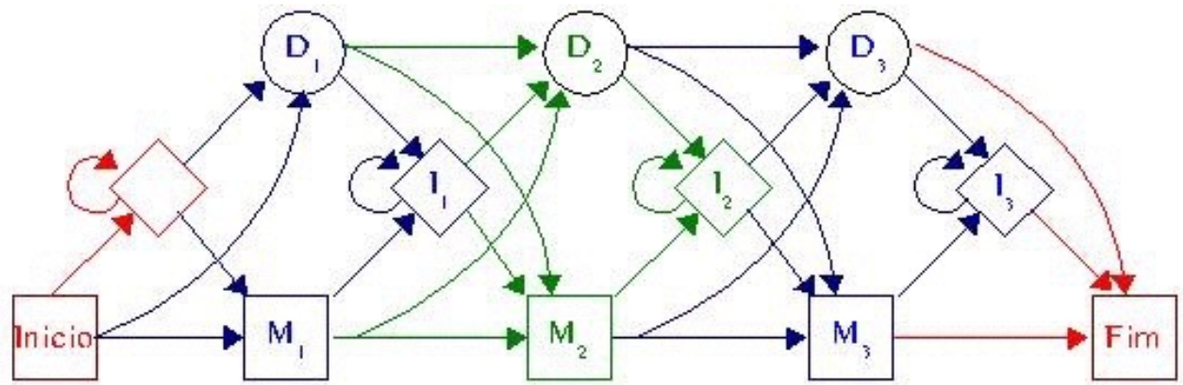


$$a_{M1M2} = \frac{A_{M1M2}}{A_{M1,M2} + A_{M1,I1} + A_{M1,D2}} = \frac{3}{3 + 0 + 1} = \frac{3}{4}$$

```
A G - - - C
A - A G - C
A G - - A -
- A - A C G
A G - - - C
```

```
M1 M2 - - - M3
M1 D2 I2 I2 - M3
M1 M2 - - I2 D3
D1 M2 - I2 I2 M3
M1 M2 - - - M3
```

$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}}$$

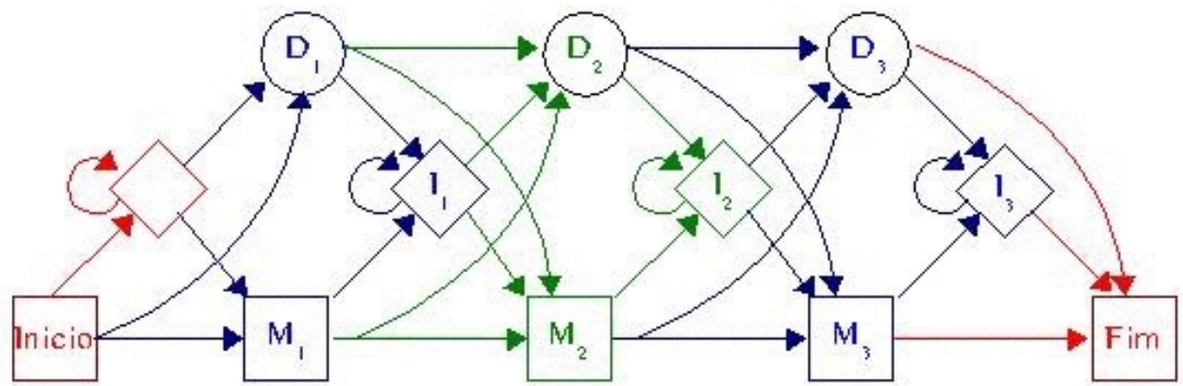


$a_{I2,M3}$?

A	G	-	-	-	C
A	-	A	G	-	C
A	G	-	-	A	-
-	A	-	A	C	G
A	G	-	-	-	C

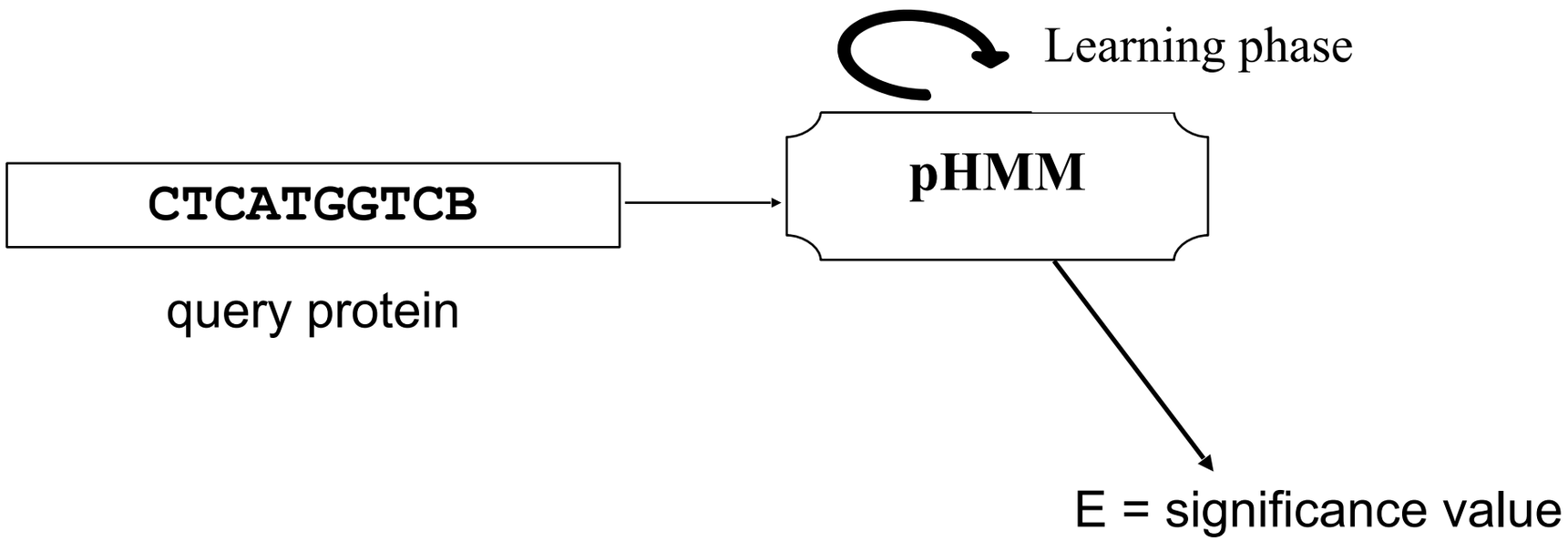
M1	M2	-	-	-	M3
M1	D2	I2	I2	-	M3
M1	M2	-	-	I2	D3
D1	M2	-	I2	I2	M3
M1	M2	-	-	-	M3

$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}}$$



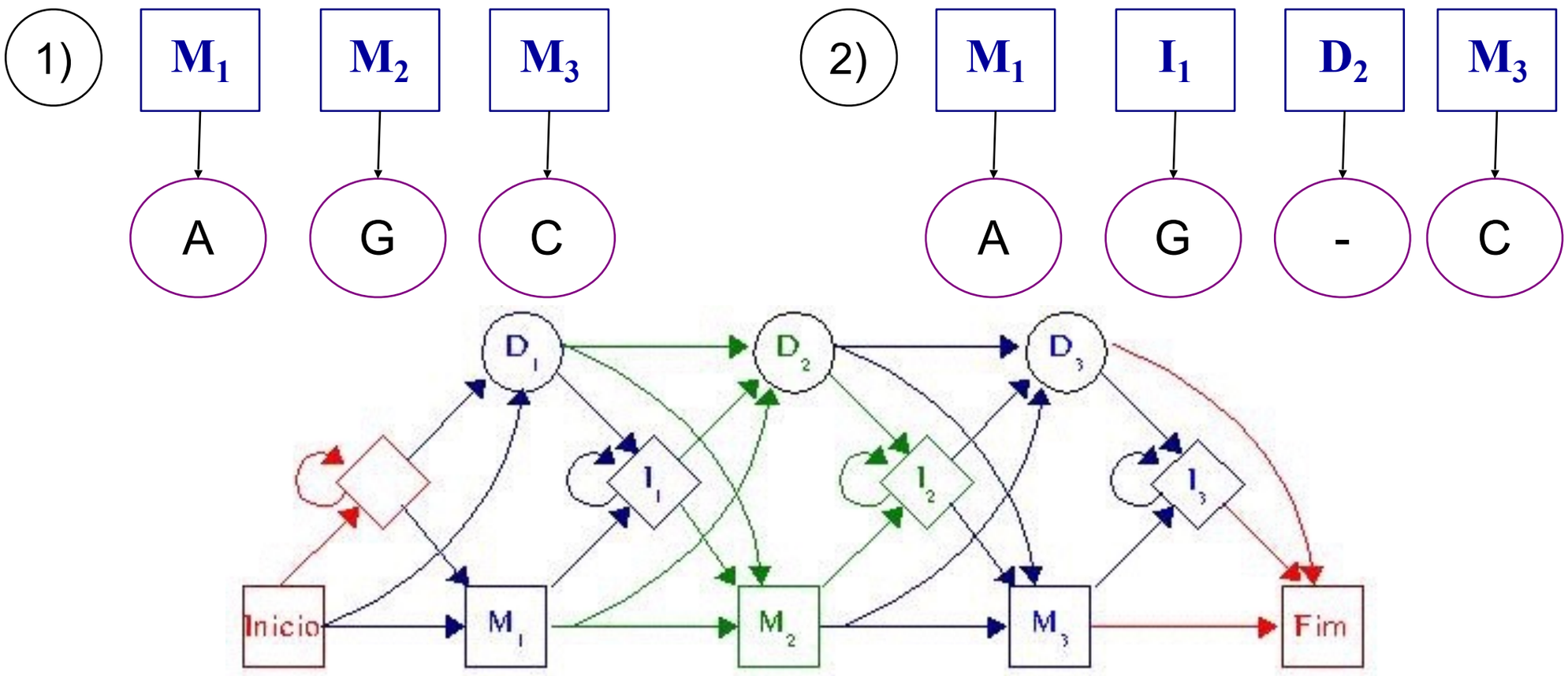
$$a_{I2,M3} = \frac{A_{I2M3}}{A_{I2,M3} + A_{I2,I2} + A_{I2,D3}} = \frac{2}{2 + 2 + 1} = \frac{2}{5}$$

3 How to infer?



3 How to infer?

- What is the path and probability of "agc" ?
- Several possibilities.



We apply Viterbi to obtain the most probable path and the probability associated

hmmer.janelia.org/software

Most Visited | Getting Started | Apple | iCloud | Facebook | Twitter | Wikipedia | Yahoo!

HMMER

biosequence analysis using profile hidden Markov models

Home | Search | Results | Software | Help | About

Current | Archive

The current version of HMMER

Download

The **current version** is HMMER **3.1b2** (05 March 2015).

Source:	[FTP]	[HTTP]	5.8 MB
with Linux/Intel ia32 binaries:	[FTP]	[HTTP]	18.1 MB
with Linux/Intel x86_64 binaries:	[FTP]	[HTTP]	20.2 MB
with MacOSX/Intel binaries:	[FTP]	[HTTP]	13.5 MB

If you are looking for **older versions** of the software, try the archive link at the top of the page.

Documentation

Release notes and User's Guide: [PDF, 116 pages].

Briefly, to compile from source:

```
% tar zxf hmmer-3.1b2.tar.gz
% cd hmmer-3.1b2
% ./configure
% make
% make check
```