Modélisation de motifs

Hélène Touzet

helene.touzet@univ-lille.fr

CNRS, Bonsai, CRIStAL

Définition

- suite de nucléotides ou d'acides aminés
- résultat d'une pression sélective : le motif est mieux conservé que le reste de la séquence
  - site actif
  - contrainte spatiale (structure 3D)
  - site de liaison
- caractériser ces motifs participent à l'analyse fonctionnelle des séquences qui les portent

Exemples de motifs nucléiques

- signaux liés à la structure d'un gène
  - codon d'initiation
  - codon de terminaison
  - signaux de transcription (TATA box, etc.)
  - signaux d'épissage
- sites de fixation de facteurs de transcription

Exemples de motifs protéiques

- signature de familles de protéines
- sites enzymatiques
- cystéines impliquées dans des ponts di-sulfures
- régions impliquées dans la liaison à un autre molécule ou une autre protéine

  (ADP/ATP, GDP/GTP, calcium, ADN, etc.)

Comment modéliser un motif biologique?

Quatre types de modèle

- séquence consensus
- expression Prosite
- matrices positionelles
- profils HMM

Code IUPAC pour l'ADN

- International Union of Pure and Applied Chemistry
- alphabet à 15 lettres qui décrit toutes les combinaisons de nucléotides possibles

A	adenine
C	cytosine
G	guanine
T	thymine
U	uracile
R	G A (purine)
Y	T C (pyrimidine)
K	G T (groupe keto)

M	A C (groupe amino)
S	G C (strong)
W	A T (weak)
B	G T C (pas A)
D	G A T (pas C)
H	A C T (pas G)
V	G C A (pas T)
N	A G C T

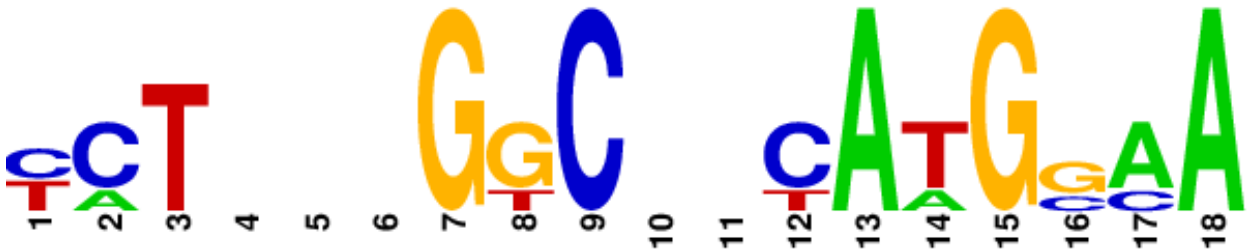
```
sequence 1      C C T A T G G G C T A C A A G C C A
sequence 2      C A T C C T G T C C C T A T G G A A
sequence 3      T C - - A A G G C C G C A T G - A A
sequence 4      T C - - A A G G C A G C A T G G A A

IUPAC           Y M - - H D G K C H V Y A W G - M A
```

Sequence logo

```
sequence 1      C C T A T G G G C T A C A A G C C A
sequence 2      C A T C C T G T C C C T A T G G A A
sequence 3      T C - - A A G G C C G C A T G - A A
sequence 4      T C - - A A G G C A G C A T G G A A

IUPAC           Y M - - H D G K C H V Y A W G - M A
```



https://weblogo.berkeley.edu

Exemple: site de fixation du facteur de transcription *c-Ets-1* chez les murins (15 séquences, TRANSFAC M00032)

```
G C C G G A A G T G
A C C G G A A G C A
G C C G G A T G T A
A C C G G A A G C T
A C C G G A T A T A
C C C G G A A G T G
A C A G G A A G T C
G C C G G A T G C A
T C C G G A A G T A
A C A G G A A G C G
A C A G G A T A T G
T C C G G A A A C C
A C A G G A T A T C
C A A G G A C G A C
T C T G G A C C C T
```

Séquence consensus $\rightarrow$ N C M G G A W G Y N

Expression Prosite

`C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H`

`C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C`

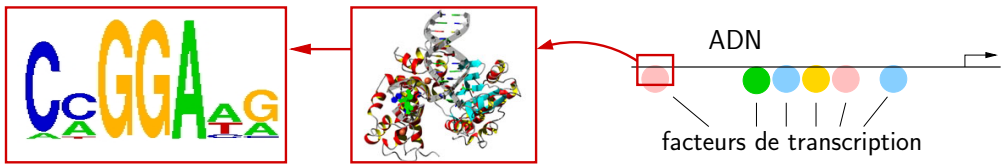
- motifs protéiques
- syntaxe

`-`	:	séparation des éléments
`x`	:	n'importe quel acide aminé
`(3,5)`	:	nombre d'occurrences (entre 3 et 5)
`[NHG]`	:	alternative (N, H ou G)

Matrices positionelles

Point de départ: alignement multiple

```
G C C G G A A G T G
A C C G G A A G C A
G C C G G A T G T A
A C C G G A A G C T
A C C G G A T A T A
C C C G G A A G T G
A C A G G A A G T C
G C C G G A T G C A
T C C G G A A G T A
A C A G G A A G C G
A C A G G A T A T G
T C C G G A A A C C
A C A G G A T A T C
C A A G G A C G A C
```



sites de fixation du facteur de transcription *c-Ets-1*

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

▷

A	C	G	T
7	2	3	3
1	14	0	0
5	9	0	1
0	0	15	0
0	0	15	0
15	0	0	0
8	2	0	5
4	1	10	0
1	6	0	8
5	4	4	2

matrice de comptage
ligne : position de l'alignement
colonne : acide nucléique

matrice de comptage

A	C	G	T
7	2	3	3
1	14	0	0
5	9	0	1
0	0	15	0
0	0	15	0
15	0	0	0
8	2	0	5
4	1	10	0
1	6	0	8
5	4	4	2

▷

$$F_{ij} = \frac{C_{ij} + f_i * pw}{\sum_i C_{ij} + pw}$$

- i acide nucléique
- j position de l'alignement
- f fréquence génomique
- pw pseudo-poids

matrice de fréquences corrigées

A	C	G	T
0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

matrice de fréquences corrigées

A	C	G	T
0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

▷

$$P_{ij} = \log(\frac{F_{ij}}{f_i})$$

i acide nucléique
j position de l'alignement
f fréquence génomique

matrice de poids

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

- poids positif: les bases plus fréquentes que la moyenne
- poids négatif: les bases moins fréquentes que la moyenne

Recherche de motifs

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

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

Recherche de motifs

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

Score de TACGGATACG

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

T A C G G A T A C G

Recherche de motifs

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

Score de TACGGATACG

1. on repère le poids de chaque position dans la matrice

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

T A C G G A T A C G

Recherche de motifs

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

score de TACGGATACG

1. on repère le poids de chaque position dans la matrice
2. le score est la somme des poids

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

T A C G G A T A C G score : 6.16

Recherche de motifs

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

on recommence à la position suivante

score de ACGGATACGT

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

T A C G G A T A C G score : 6.16

A C G G A T A C G T score : -1.86

- matrices nucléiques: sites de fixation de facteurs de transcription

  JASPAR: `http://jaspar.genereg.net`

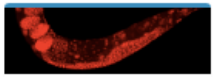
  TRANSFAC

- matrices protéiques: signatures protéiques

  PROSITE : `https://prosite.expasy.org` $\rightarrow$ InterPro

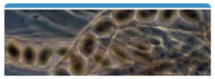
Browse JASPAR CORE for six different taxonomic groups


Vertebrata


Nematoda


Insecta


Plantae

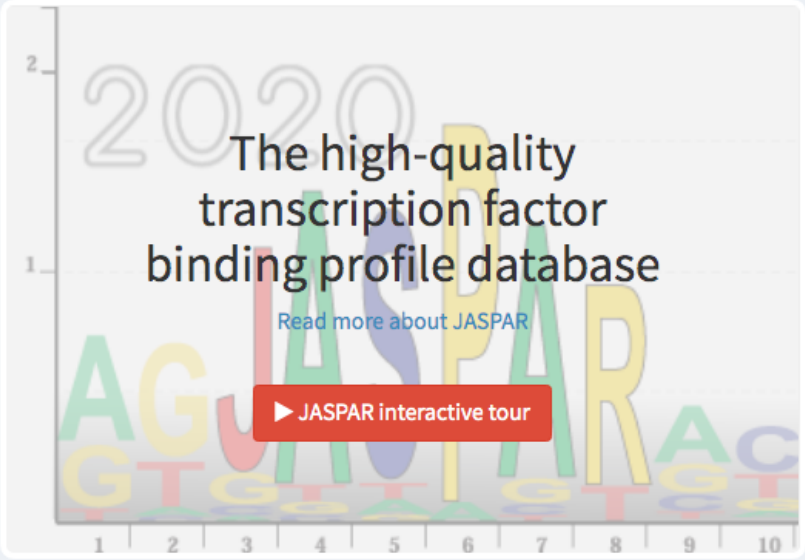

Fungi


Urochordata

JASPAR CORE & when should it be used?

Info about other collections

The JASPAR CORE contains a curated, non-redundant set of profiles, derived from published and experimentally defined transcription factor binding sites for eukaryotes. It should be used, when seeking models for specific factors or structural classes, or if experimental evidence is paramount.



Citing JASPAR 2020

PubMed | NAR | PDF

Fornes O, Castro-Mondragon JA, Khan A, et al. **JASPAR 2020: update of the open-access database of transcription factor binding profiles**. *Nucleic Acids Res*. 2019; doi: 10.1093/nar/gkz1001

http://jaspar.genereg.net

Profils HMM

- HMM= Hidden Markov model
- modèle probabiliste plus fin que les matrices
- prise en compte des insertions, des délétions
- mis en œuvre dans PFAM et dans Interpro (motifs protéiques)

V E D - - L I R Y

V E D - - L R R Y

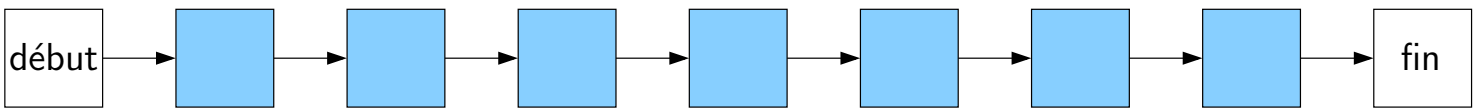
P N E - - L R R F

D N K A A L R R F

A E E - - L A - -

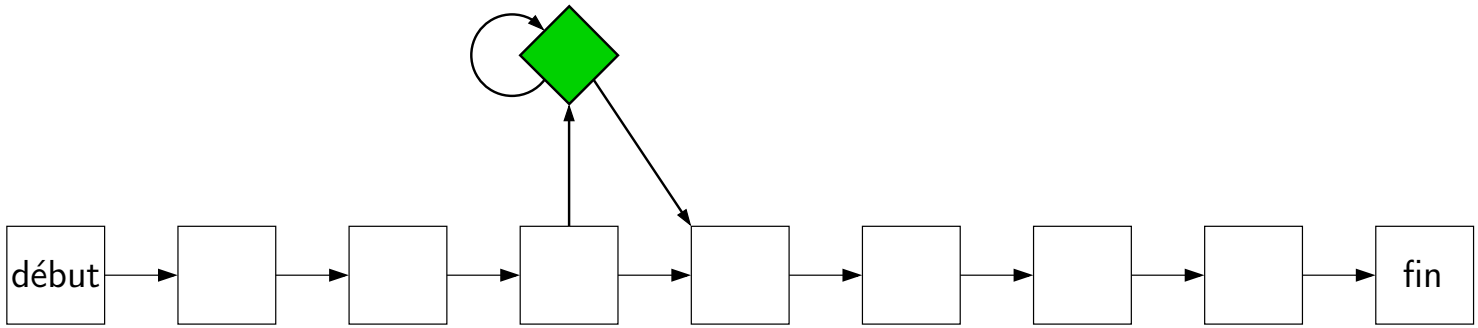
V	E	D	-	-	L	I	R	Y
V	E	D	-	-	L	R	R	Y
P	N	E	-	-	L	R	R	F
D	N	K	A	A	L	R	R	F
A	E	E	-	-	L	A	-	-

création d'un état par colonne



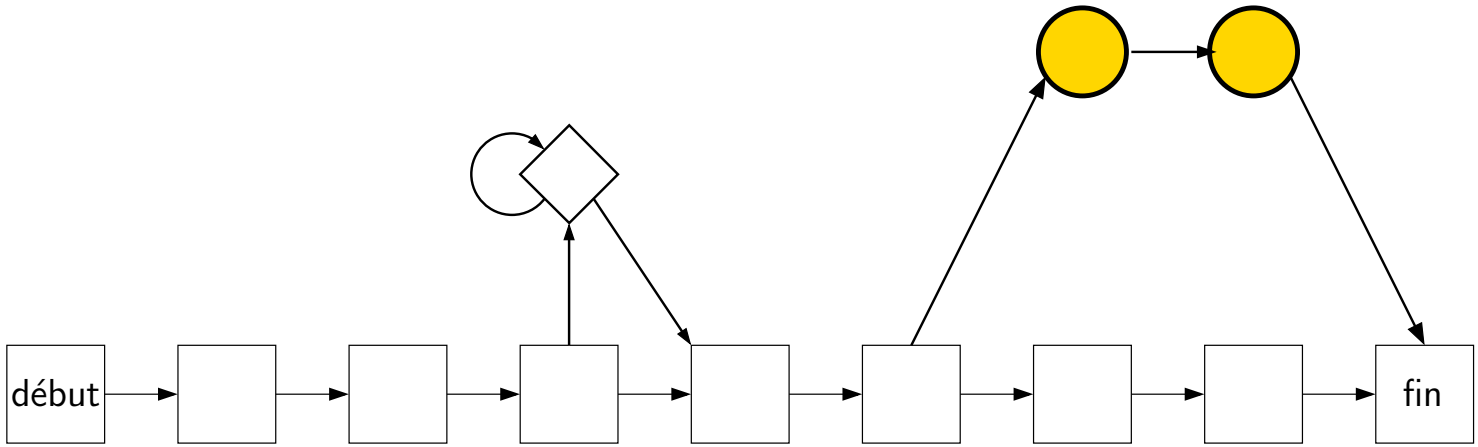
```
V E D - - L I R Y
V E D - - L R R Y
P N E - - L R R F
D N K A A L R R F
A E E - - L A - -
```

prise en compte des insertions

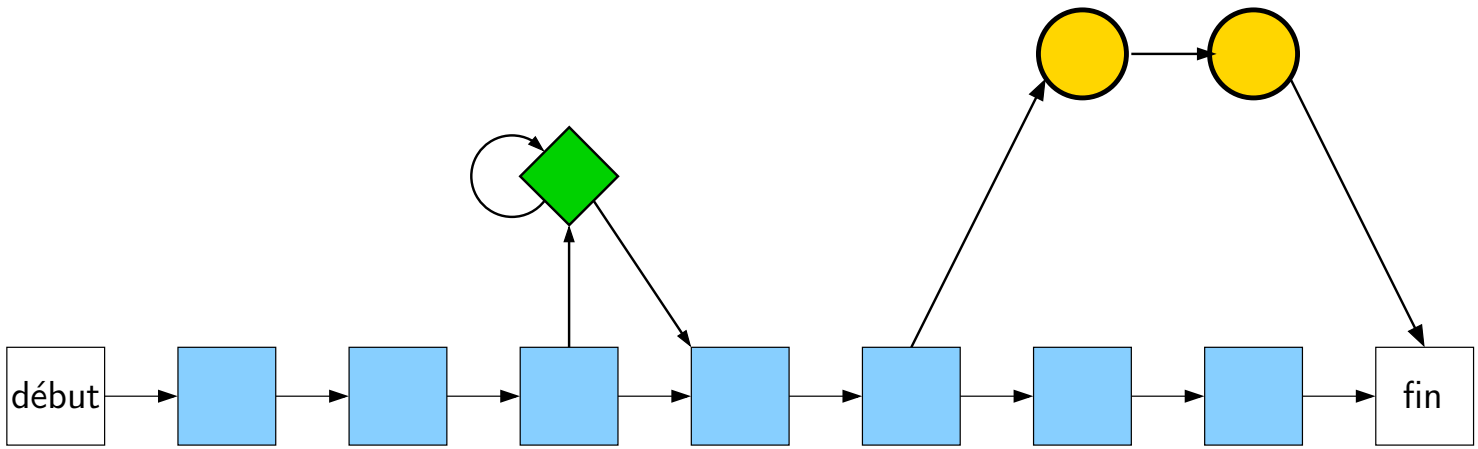


```
V E D - - L I R Y
V E D - - L R R Y
P N E - - L R R F
D N K A A L R R F
A E E - - L A - -
```

prise en compte des délétions

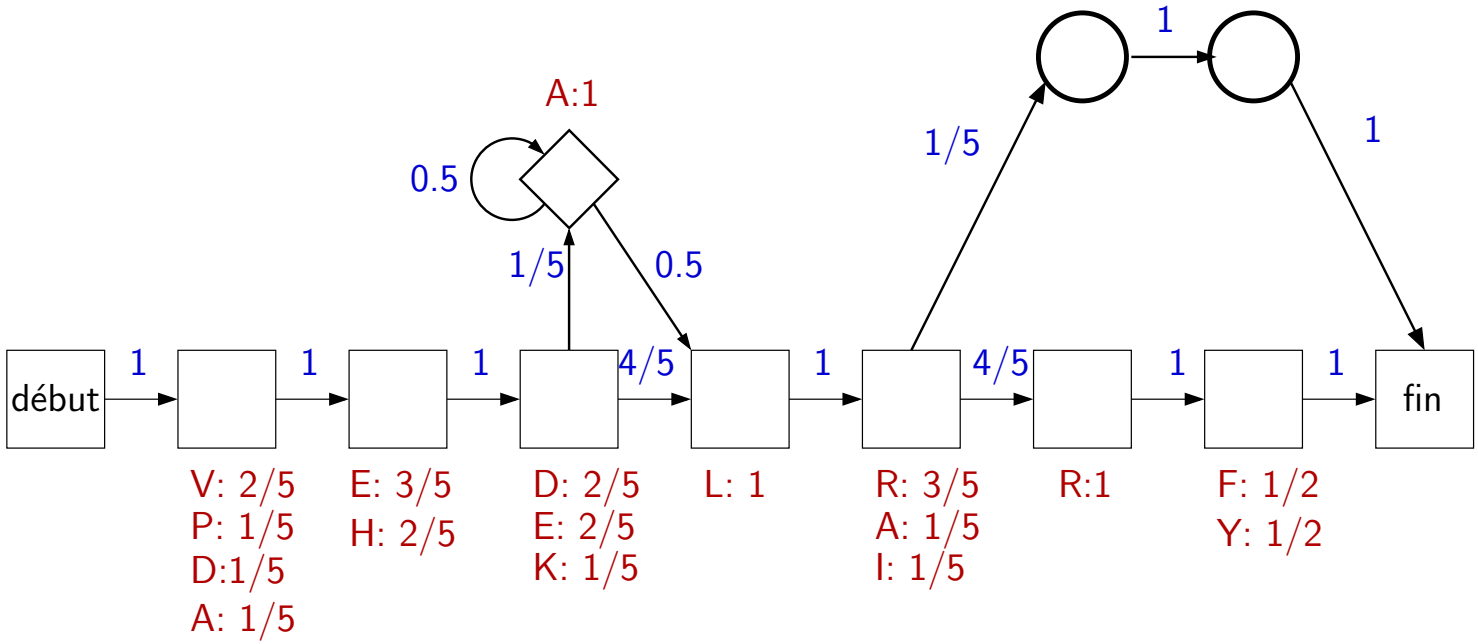


V	E	D	-	-	L	I	R	Y
V	E	D	-	-	L	R	R	Y
P	N	E	-	-	L	R	R	F
D	N	K	A	A	L	R	R	F
A	E	E	-	-	L	A	-	-



```
V E D - - L I R Y
V E D - - L R R Y
P N E - - L R R F
D N K A A L R R F
A E E - - L A - -
```

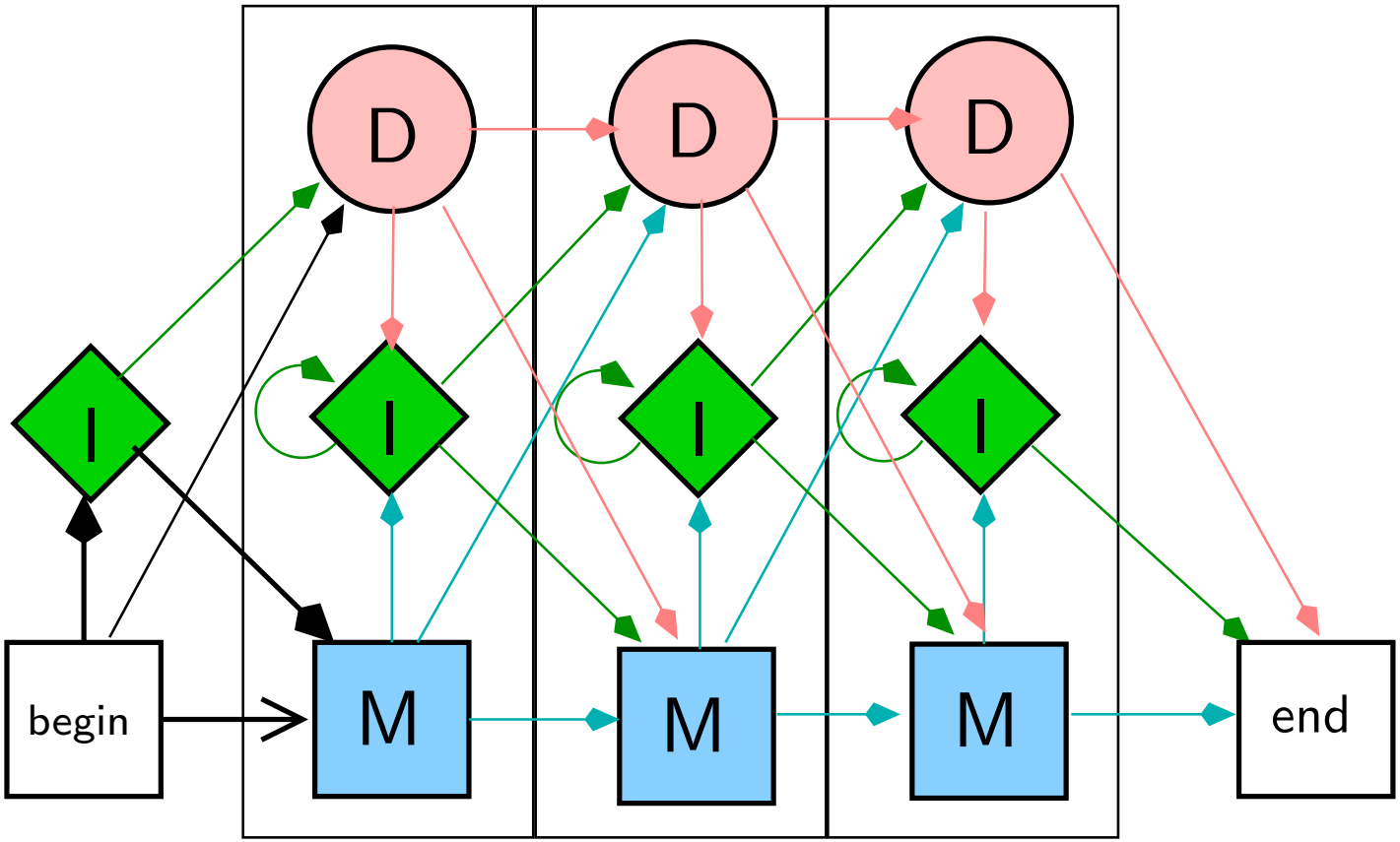
- émissions
  - fréquences des acides aminés
- transitions
  - circulation dans le modèle
  - indels



En résumé

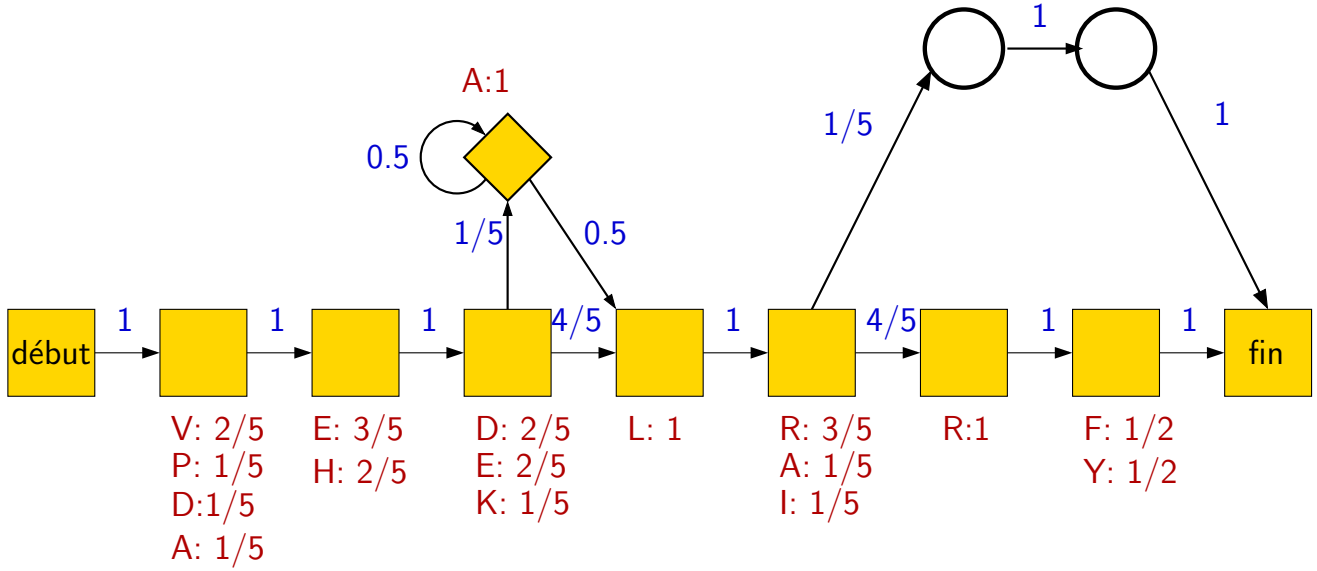
- Etats matchants : colonnes avec moins de 50% de -
- Etats d'insertion : majorité de -
- Etats de délétion : minorité de -
- Probabilités d'émission : on compte le nombre d'occurrences de chaque acide aminé
- Probabilités de transition : on compte le nombre de séquences empruntant la transition
- Correction avec les pseudo-poids : $+1$ à chaque compte

- Modèle complet



Recherche avec un profil HMM

- score : probabilité maximale d'un mot dans le modèle



Score de VHKALARY

$$1 \times \frac{2}{5} \times 1 \times \frac{2}{5} \times 1 \times \frac{1}{5} \times \frac{1}{5} \times 1 \times 0.5 \times 1 \times 1 \times \frac{1}{5} \times 1 \times 1 \times \frac{1}{2} \times 1$$

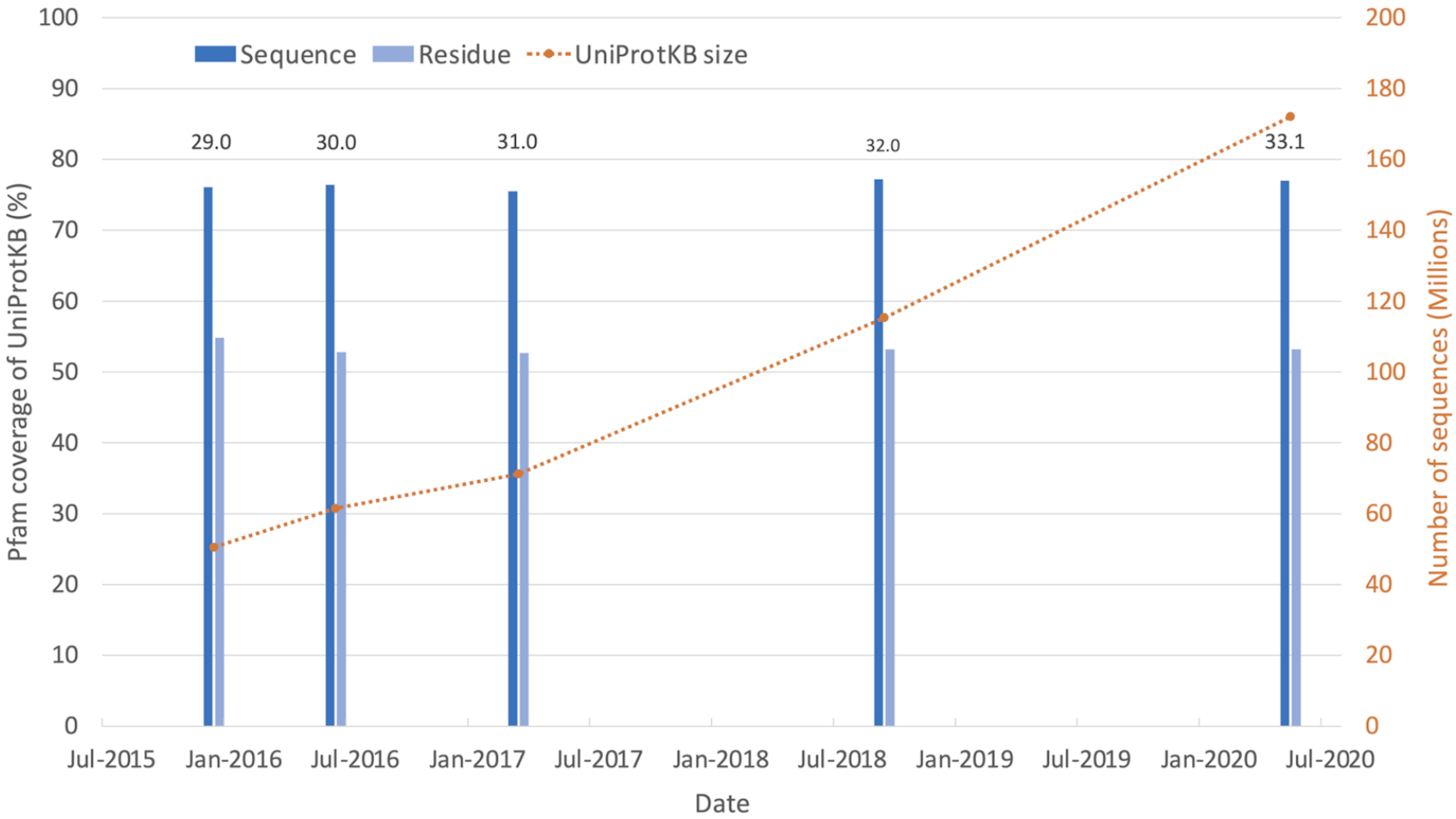
- localisation d'un motif dans une séquence: algorithme de Viterbi

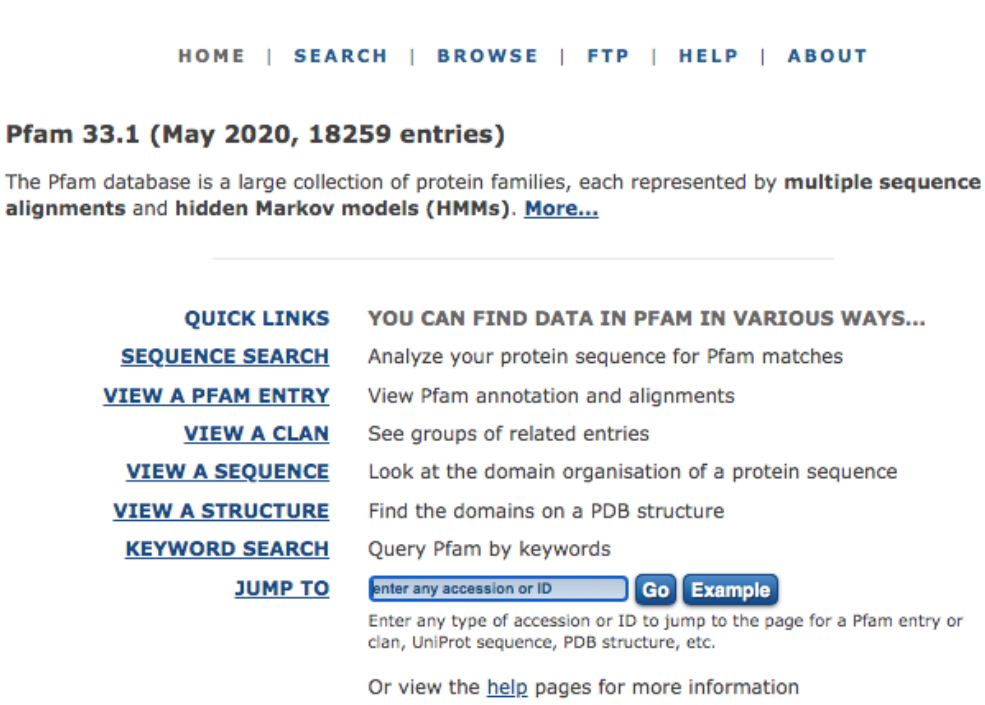
HMMER

- série d'utilitaires pour manipuler les profile HMM
  - recherche de motifs sur une séquence
  - construction d'un motif à partir d'un alignement multiple
  - etc.
- inteface web ou installation locale avec Conda
- `https://www.ebi.ac.uk/Tools/hmmer`

Pfam - protein families

- création en 1995
- version 33.1 (2020): 18 259 familles et 635 regroupements en clans
- pour chaque famille
  - ensemble de séquences
  - alignement multiple
  - profile HMM (contruits avec HMMER)
  - liens extérieurs: PDB, etc.


Sequence
Residue
UniProtKB size
29.0
30.0
31.0
32.0
33.1
Pfam coverage of UniProtKB (%)
Number of sequences (Millions)
Date
Jul-2015
Jan-2016
Jul-2016
Jan-2017
Jul-2017
Jan-2018
Jul-2018
Jan-2019
Jul-2019
Jan-2020
Jul-2020
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
120
140
160
180
200



https://pfam.xfam.org

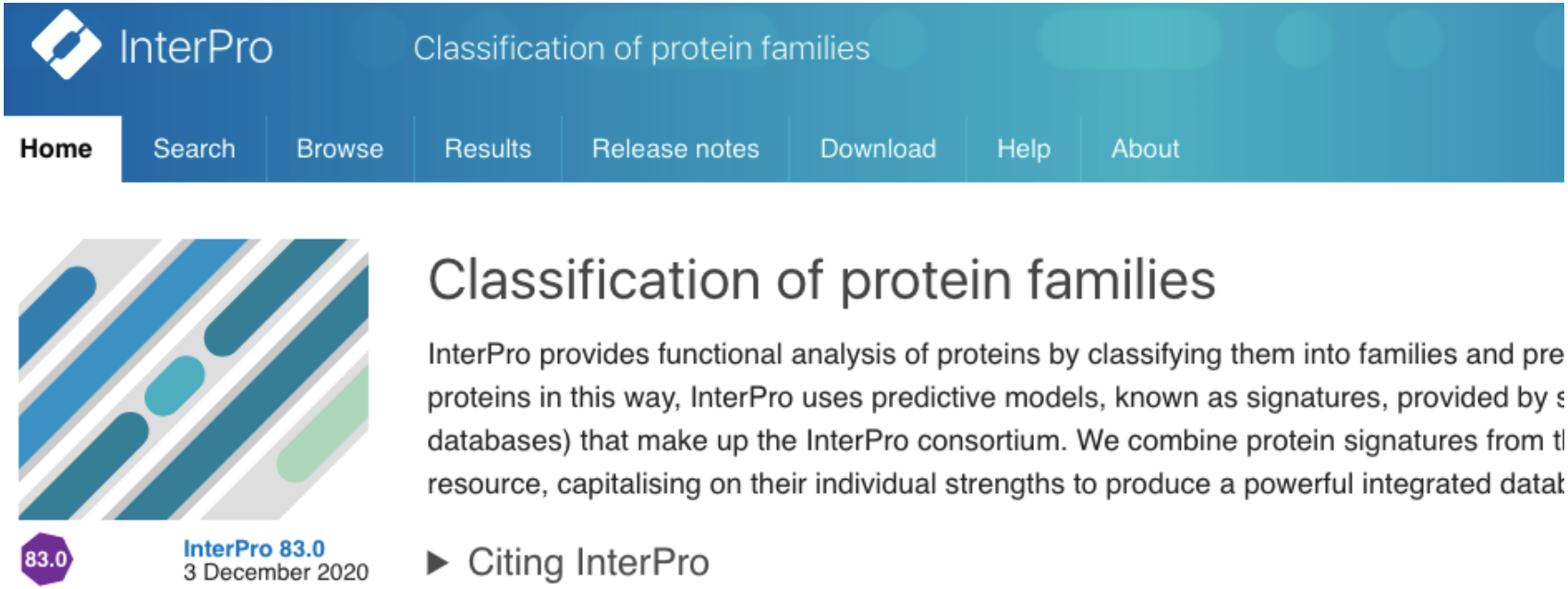
```
> protein
mvlspadktnvkaawgkvgahageygaealermflsfpttktyfphfdlshgsaqvkghg
kkvadaltnavahvddmpnalsalsdlhahklrvdpvnfkllshcllvtlaahlpaeftp
avhasldkflasvstvltskyr
```

InterPro

Protein sequence analysis & classification

- `http:// www.ebi.ac.uk/interpro`
- developed at EBI since 1999 (version 70)
- signatures for protein families, domains and functional sites collected from 14 databases (PFAM, Prosite and many more)

  35 020 entries based on 48 938 signatures
- mappings of InterPro entries to Gene Ontology (GO) terms (InterPro2GO)



https://www.ebi.ac.uk/interpro

```
> protein
mvlspadktnvkaawgkvgahageygaealermflsfpttktyfphfdlshgsaqvkghg
kkvadaltnavahvddmpnalsalsdlhahklrvdpvnfkllshcllvtlaahlpaeftp
avhasldkflasvstvltskyr
```