

# Modélisation de motifs biologiques

Hélène Touzet

`touzet@lifl.fr`

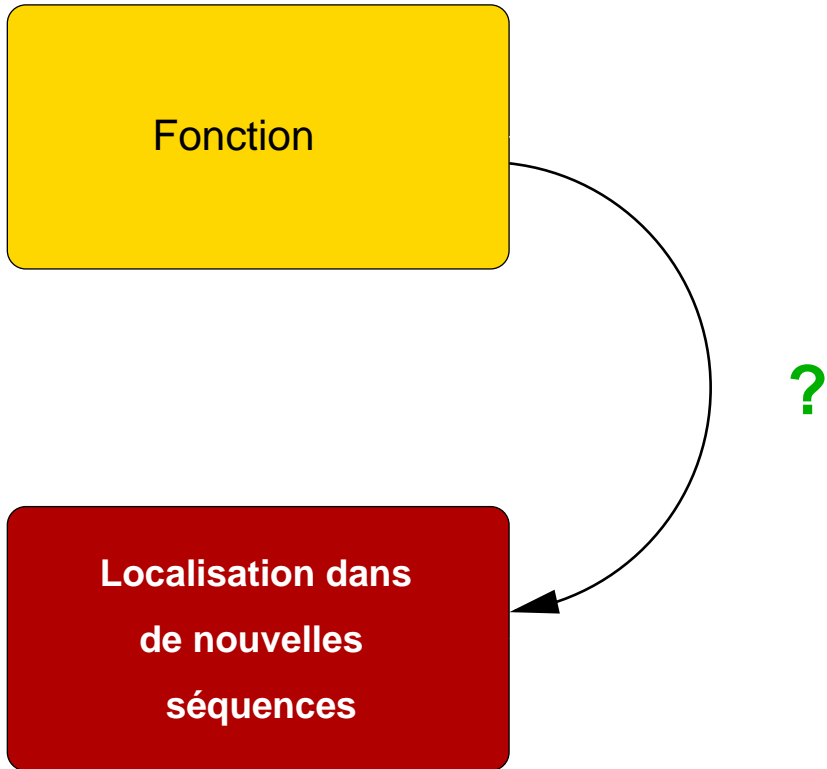
## ▷ Motifs ADN

- séquences consensus d'éléments répétés chevauchants (SINEs et LINEs)
- signaux d'expression (promoteurs, sites d'initiation de la traduction, ...)
- domaines compositionnels (îlots CpG)

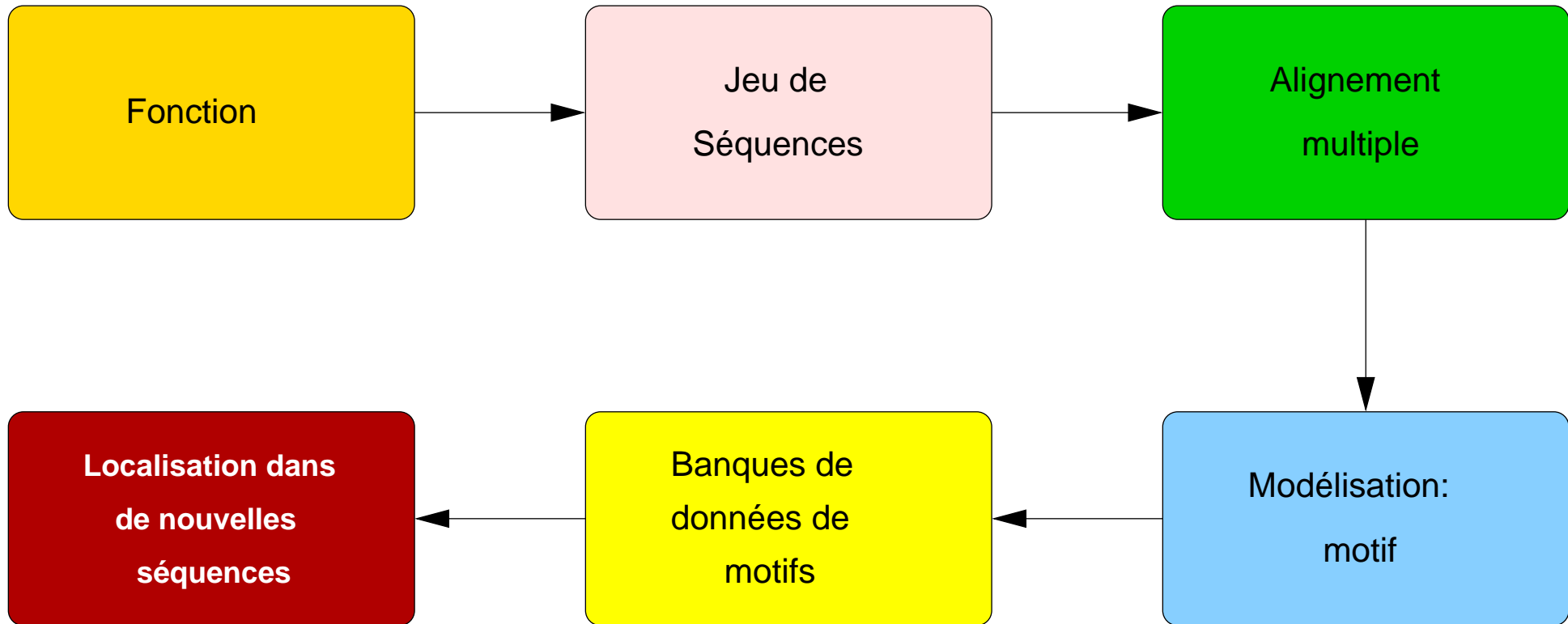
## ▷ Motifs protéiques

- signature de familles de protéines
- domaines structurels ou fonctionnels dégénérés ( (domaines immunoglobine, SH2, SH3, ... )
- 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.)
- domaine compositionnels (sites d'activation riches en glutamines, répétitions riches en leucines, ...)

# Découverte de motifs



# Découverte de motifs



- **Problème 1:** trouver une représentation des motifs à partir des alignements multiples
- **Problème 2:** concevoir des algorithmes pour localiser les occurrences des motifs dans une nouvelle séquence

# Les séquences consensus

**Code IUPAC** (International Union of Pure and Applied Chemistry)

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

**Séquence consensus** : à chaque position de l'alignement multiple, on retient la lettre majoritaire

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

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 → **N C M G G A W G Y N**

# Représentation d'un motif par une matrice

- ▷ ligne → position de l'alignement
- ▷ colonne → acide nucléique (4 colonnes)

**Exemple :** *c-Ets-1*

GCCGGAAGTG  
 ACCGGAAGCA  
 GCCGGATGTA  
 ACCGGAAGCT  
 ACCGGATATA  
 CCCGGAAGTG  
 ACAGGAAGTC  
 GCCGGATGCA  
 TCCGGAAGTA  
 ACAGGAAGCG  
 ACAGGATATG  
 TCCGGAACC  
 ACAGGATATC  
 CAAGGACGAC  
 TCTGGACCCT

	A	C	G	T	
	7	2	3	3	<b>N</b>
	1	14	0	0	<b>C</b>
	5	9	0	1	<b>M</b>
	0	0	15	0	<b>G</b>
	0	0	15	0	<b>G</b>
	15	0	0	0	<b>A</b>
	8	2	0	5	<b>W</b>
	4	1	10	0	<b>G</b>
	1	6	0	8	<b>Y</b>
	5	4	4	2	<b>N</b>



**Position Frequency Matrix**

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

## Des PFM aux PWM

- ▷ **PWM** : Position Weight Matrix
- ▷ Poids positif : les bases qui apparaissent plus que la moyenne
- ▷ Poids négatif : les bases qui apparaissent moins que la moyenne

Cela doit vous rappeler les matrices de score PAM, BLOSUM,...

- ▷ Poids de la base  $x$  dans une colonne de l'alignement :

$$\log_2 \left( \frac{f(x)}{0.25} \right)$$

$f(x)$  est la fréquence de  $x$  dans la colonne considérée

0.25 suppose que les quatre bases sont la même probabilité d'apparition

- ▷ Le problème des 0 : ajout d'un pseudo-compte pour éviter qu'il y ait sur-adaptation

$$\log_2 \left( \frac{f(x) + 0.05}{0.25} \right)$$

## Suite de l'exemple :

Position Frequency Matrix

0.91	-0.94	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



Position Weight Matrix

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

## Suite de l'exemple :

Position Frequency Matrix				Position Weight Matrix				
0.91	-0.94	0.2	0.2	0.91	-0.94	-0.32	-0.32	T
0.07	0.93	0	0	-1.8	1.9	-2.3	-2.3	A
0.33	0.6	0	0.07	0.4	1.26	-2.3	-1.8	C
0	0	1	0	-2.3	-2.3	2	-2.3	G
0	0	1	0	-2.3	-2.3	2	-2.3	G
1	0	0	0	2	-2.3	-2.3	-2.3	A
0.53	0.13	0	0.33	1.1	-0.94	-2.3	0.4	T
0.27	0.07	0.67	0	0.11	0.07	1.42	-2.3	A
0.07	0.4	0	0.53	-1.8	0.4	0	1.1	C
0.33	0.27	0.27	0.13	0.4	0.11	0.11	-0.94	G

**Score d'un mot dans le modèle:** somme des poids

T A C G G A T A C G → 6.16

## Contenu informationnel d'une PFM

- ▷ Permet de visualiser les colonnes qui sont conservées
- ▷ Mesure d'**incertitude** d'une colonne de l'alignement :

$$H = - \sum_{x \in \{A,C,G,T\}} f(x) \log_2 f(x)$$

**Incertitude maximale** : les quatre bases ont chacune une fréquence de 25%

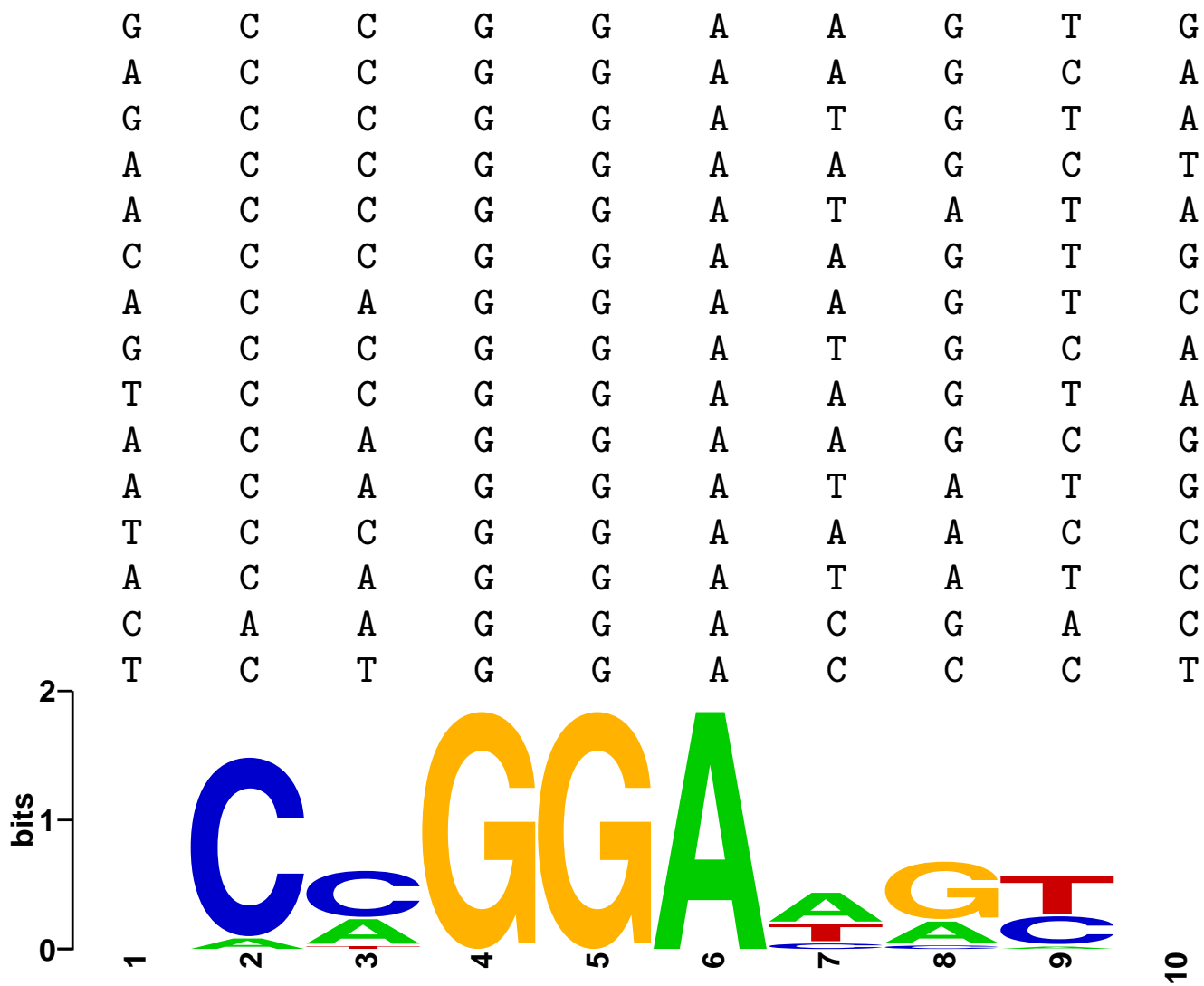
$$H = - \sum_1^4 \frac{1}{4} \log_2\left(\frac{1}{4}\right) = \sum_1^4 \frac{1}{4} \log_2(4) = \log_2(4) = 2$$

**Incertitude minimale** : la conservation est parfaite, il y n'y a qu'une seule base.

$$H = - \sum_1^3 0 \log_2(0) + 1 \log_2(1) = 0$$

- ▷ Quantité d'**information** d'une colonne :  $2 - H$
- ▷ **Contribution du nucléotide**  $x$  dans cette colonne:  $f(x) \times H$

# Exemple pour le site de C-ets-1:



webllogo.berkeley.edu

<http://webllogo.berkeley.edu/logo.cgi>

# Modélisation par pattern

Exemple : hormone pancréatique (PP)

NEUY_CARAU/29-64	AEE..LAKYYSALRH $\color{magenta}Y$ INLIT $\color{red}RQR$ $\color{red}Y$
PYY_HUMAN/29-64	PEE..LNRYYASLRH $\color{magenta}Y$ LNLVTR $\color{red}QR$ $\color{red}Y$
PMY_PETMA/1-36	PEE..LSKYMLAVRN $\color{magenta}Y$ INLIT $\color{red}RQR$ $\color{red}Y$
PPY_LOPAM/1-36	PED..WASYQAAVRH $\color{magenta}Y$ VNLIT $\color{red}RQR$ $\color{red}Y$
PAHO_BOVIN/30-65	PEQ..MAQYAAELRR $\color{magenta}Y$ INMLTR $\color{red}PR$ $\color{red}Y$
PAHO_CHICK/26-61	VED..LIRFYNDLQQ $\color{magenta}Y$ RLNVVTR $\color{red}HR$ $\color{red}Y$
PAHO_ANSAN/1-36	VED..LRFYYDNLQQ $\color{magenta}Y$ RLNVFTR $\color{red}HR$ $\color{red}Y$
NPF_HELAS/4-39	PNE..LRQYLKELNE $\color{magenta}Y$ YAIMGR $\color{red}TR$ $\color{red}F$
NPF_MONEX/1-39	DNKAALRDYLRQINE $\color{magenta}Y$ FAIIGR $\color{red}PR$ $\color{red}F$

\*\*\*\*\*

Source : Prosite, entrée PS00265

$\color{red}[FY] -x(3) - [LIVM] -x(2) - Y -x(3) - [LIVMFY] -x - R -x - R - [YF]$

## 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

### 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)

## Qu'est-ce un bon pattern ?

Suffisamment tolérant

pas de sur-adaptation  
limiter le nombre de *faux négatifs*

limiter le nombre de *faux positifs*

Suffisamment discriminant

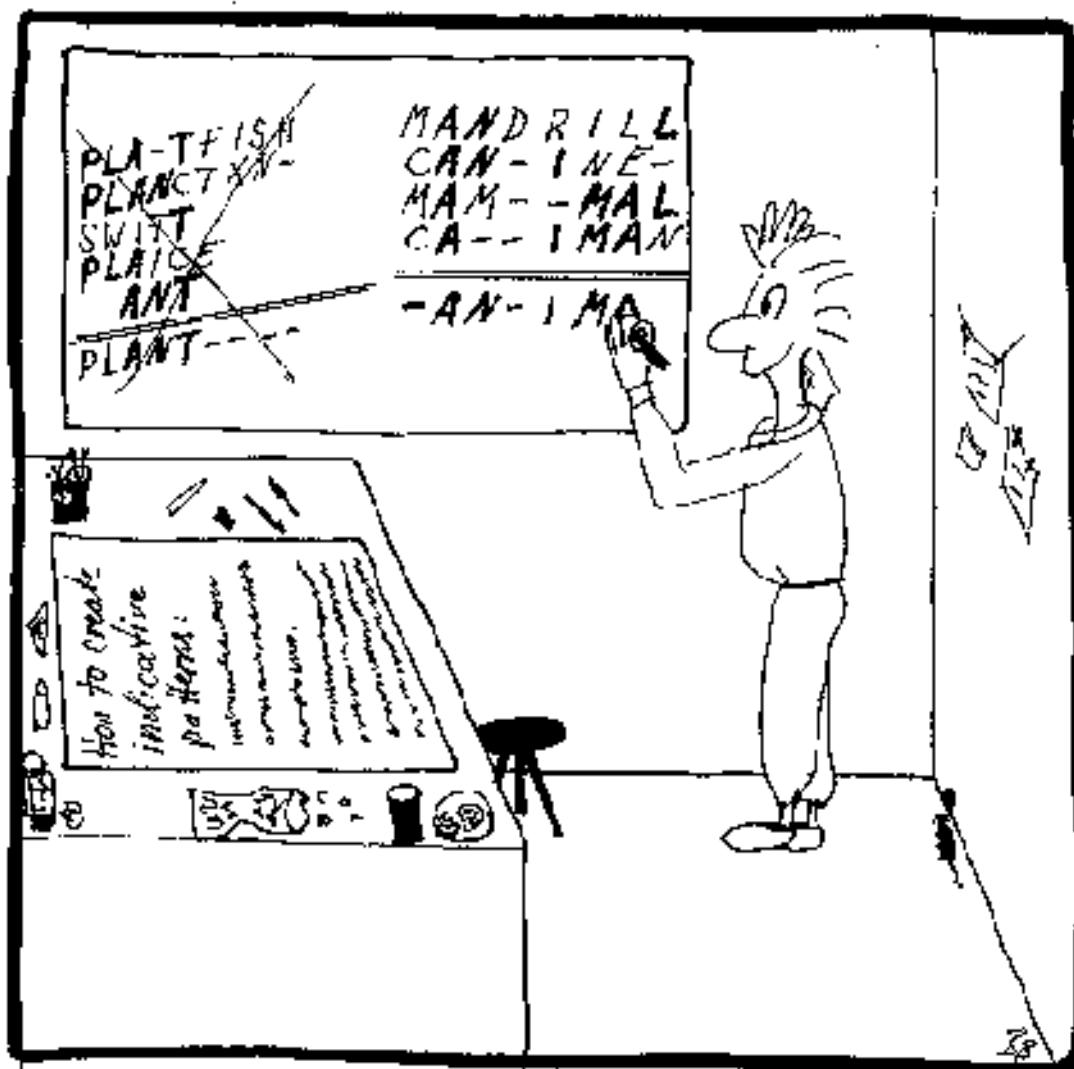
## Construction d'un pattern

- ▷ à la main
- ▷ PRATT (<http://www.ebi.ac.uk/pratt/>)

## Détection des séquences correspondant au pattern

- ▷ Avec un automate particulièrement simple  
(Expressions régulières sans \* et avec + limité)

## How we develop Prosite patterns!



Brigitte Boeckmann / 1995

# Modélisation par profil

Point de départ : matrice des positions (20 colonnes non nulles)

```
1 0 0 1 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 2 0 0 0 0 0
0 0 0 0 7 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 3 4 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 7 1 0 0 0 0 0 0 0 0 1 0 0 0 0
3 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 3 1 0 0 0 0 0 0 0
0 0 0 1 0 1 0 0 0 0 2 0 0 0 0 0 2 2 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0
1 0 0 0 0 0 0 0 0 0 2 1 0 0 0 1 0 0 0 0 0 0 0 4 0 0
3 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0
3 0 0 1 2 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 6 0 0 0 0 0 0 0 0 2 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 2 5 0 0 0 0 0 0 0 0 0
0 0 0 0 2 0 0 3 0 0 0 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0
0 0 0 0 0 1 0 0 3 0 0 2 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0
2 0 0 0 0 0 0 0 0 0 0 1 0 6 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 2 0 0 4 1 1 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 4 0 0 1 1 0 0 0 0 0 0 0 0 3 0 0 0 0
0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 2 4 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0
0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0
```

## Apporter plus de souplesse . . .

- ▷ *autoriser des insertions, des délétions*

Ajouter des pénalités particulières pour l'insertion ou la délétion d'un acide aminé → [colonne supplémentaire](#)

- ▷ *autoriser des substitutions (entre acides aminés voisins)*

$M(j, k)$  : similarité entre les deux acides aminés  $j$  et  $k$   
(PAM, BLOSUM, . . .)

$f_{ik}$  : fréquence du  $k$ ième l'acide aminé à la position  $i$

$$\text{Profil}_{ij} = \sum_{k=1}^{20} f_{ik} M(j, k)$$

5.2	0	0.8	2.1	...	...	4.55
2.8	0	-5.3	9.1	...	...	4.55
2.6	0	-5.7	10	...	...	4.55
1.7	0	0.3	0.3	...	...	4.55
1.7	0	0.3	0.3	...	...	4.55
-1.7	0	-8.2	-5.6	...	...	4.55
4.7	0	0.6	1.7	...	...	4.55
-0.0	0	-3.3	2.8	...	...	4.55
-3.2	0	8.8	-5.6	...	...	4.55
0.3	0	1.7	-2.8	...	...	4.55
5.6	0	-1.1	3.3	...	...	4.55
6.9	0	-1.2	6.4	...	...	4.55
-0.2	0	-4.7	-4	...	...	4.55
-0.8	0	-3.7	2.7	...	...	4.55
0.7	0	-3.7	5.6	...	...	4.55
-3	0	10	-5	...	...	4.55
-1.2	0	-0.2	-3.7	...	...	4.55
4.6	0	-2.2	4.1	...	...	4.55
0	0	-3.9	-2.7	...	...	4.55
0.6	0	0	-2.6	...	...	4.55
3.4	0	1.7	1.6	...	...	4.55
-3.0	0	-3	0	...	...	4.55
2.2	0	-2.4	4	...	...	4.55
-3.0	0	-3	0	...	...	4.55
-3.4	0	7.6	-6.1	...	...	4.55

5.2

=

$$\frac{1}{9}M(A, A) + \frac{1}{9}M(A, D) + \frac{5}{9}M(A, P) + \frac{2}{9}M(A, V)$$

colonne des indels → 4.55

## Recherche avec un profil

	S	E	Q	U	E	N	C	E
p	.	.	.	.	.	.	.	.
r		.	.	.	.	.	.	.
o	.	\	.	.	.	.	.	.
f	.	.	-	-	.	.	.	.
i	.	.	.	.	\	.	.	.
l	.	.	.	.	.		.	.
e	.	.	.	.	.	.	\	.
.	.	.	.	.	.	.	.	.

- ▷ Score : alignement entre le profil et la séquence

Les pénalités de substitutions et de gaps sont données par le profil

- ▷ Seuil d'admission : **E-value**

Banque de référence : 59 021 séquences de longueur moyenne 359 - distribution  
34 de SwissProt

**Exemple** : la séquence NPF\_ARTTR contre le profil de l'hormone pancréatique

>NPF\_ARTTR

KVVHLRPRSSFSEDEYQIYLRNVSKYIQLYGRPRF

Local: Consensus vs NPF\_ARTTR

Score: 162.31

Consensus	1	PE.EAALAKYYAALRHYINLITRQRY	25
		:    ::  ::: : : : :	
NPF_ARTTR	13	SEDEY.YQIYLRNVSKYIQLYGRPRF	37

**Prosite** : 1226 entrées, avec 1665 patterns ou matrices (novembre 2003)

# Modélisation avec des HMM

HMM = Hidden Markov Model = Modèle de Markov caché

- ▷ un ensemble d'**états**
- ▷ des **probabilités de transitions** entre les états
- ▷ un ensemble d'**observations**
- ▷ une **probabilité d'émission** qui indique pour chaque état la probabilité d'y émettre telle information

# Profil HMM

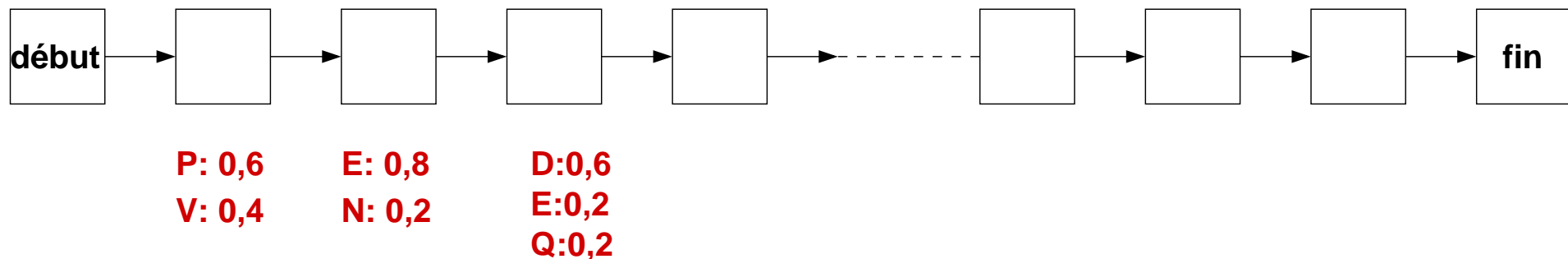
## 1. Si l'alignement n'a pas d'indels

PPY_LOPAM/1-36	PEDWASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65	PEQMAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61	VEDLIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36	VEDLRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39	PNELRQYLKELNEYAIMGRTRF

**1 observation** = 1 acide aminé

**1 état** = 1 colonne de l'alignement multiple

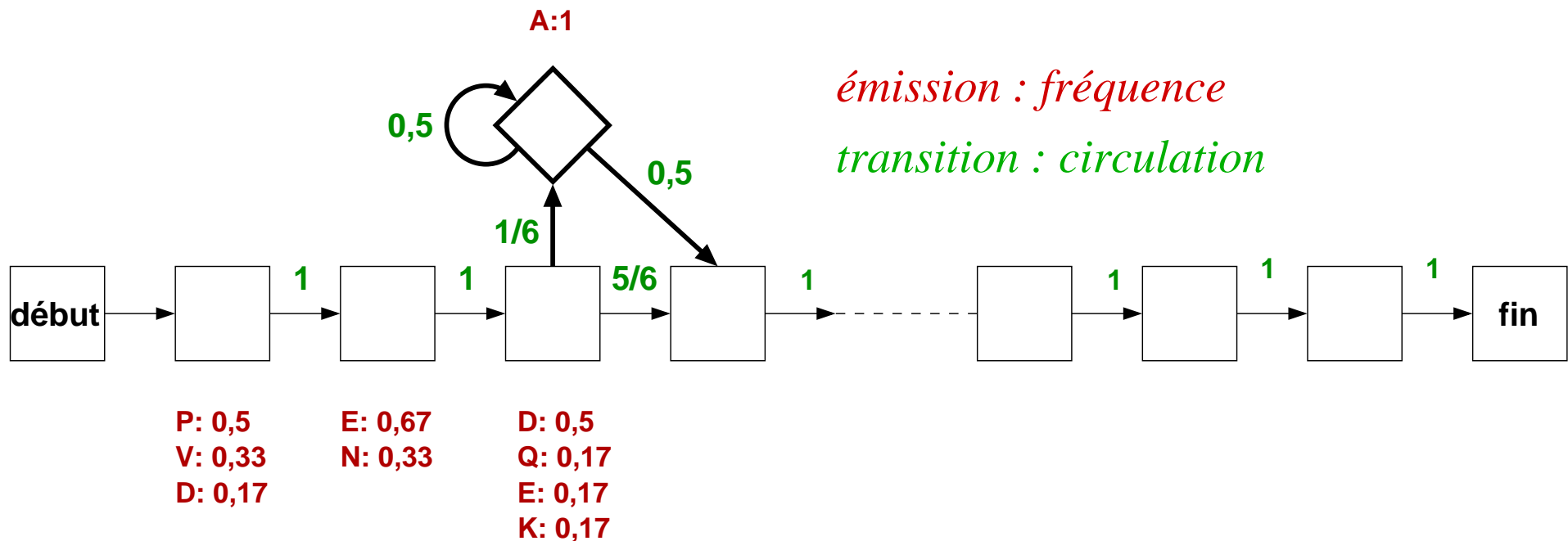
**émissions** = fréquences de chaque a.a.



2. Avec des insertions : Une insertion est un fragment de la séquence qui n'apparaît pas dans le modèle.

PPY_LOPAM/1-36	PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65	PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61	VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36	VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39	PNE..LRQYLKELNEYAIMGRTFR
NPF_MONEX/1-39	DNKAALRDYLRQINEYFAIIGRPRF

un nouvel état = un bloc d'insertion



### 3. Et finalement, avec les délétions

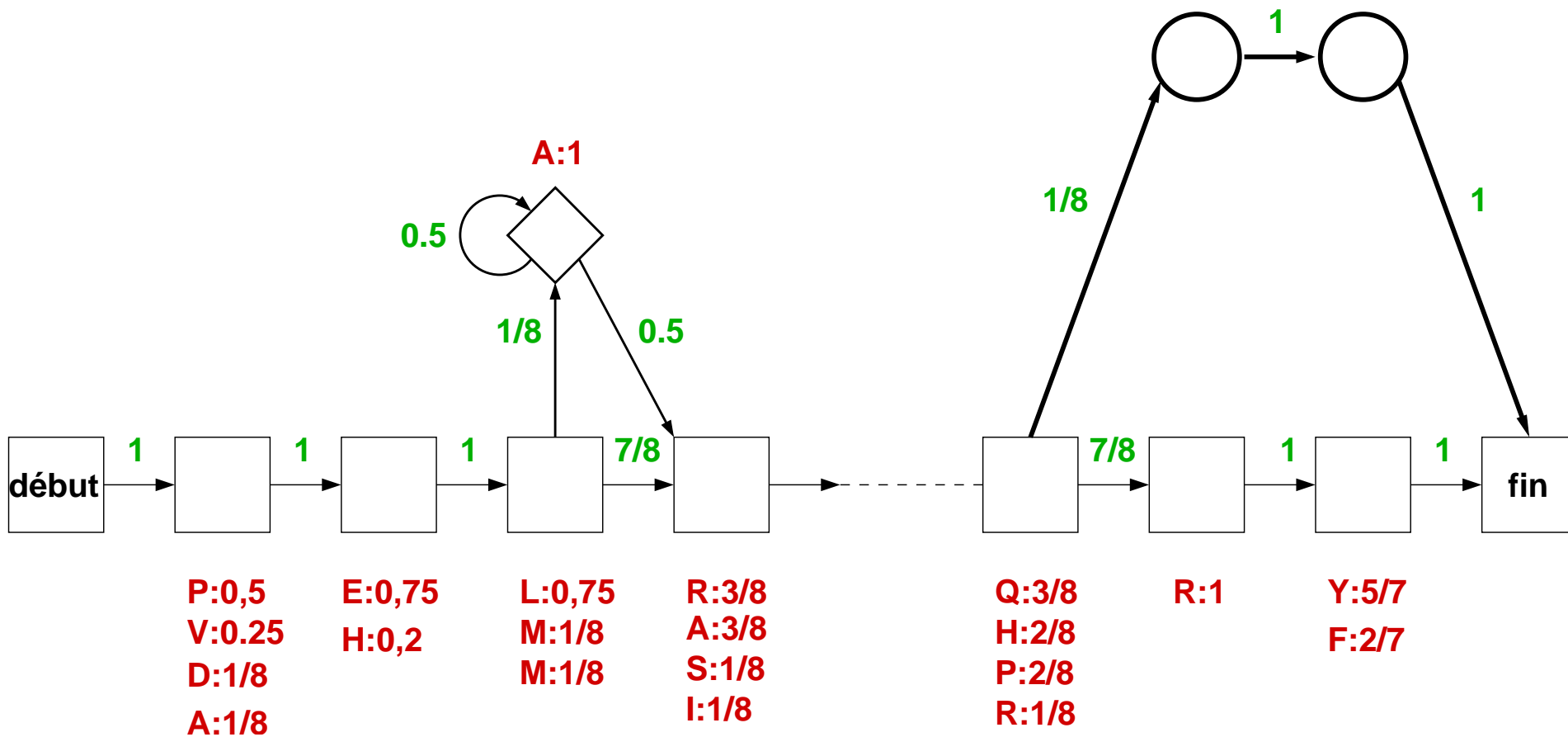
Une délétion est un fragment du modèle qui ne correspond à aucun acide aminé.

PMY_PETMA/1-36	PEE..LSKYMLAVRNYINLITRQRY
PPY_LOPAM/1-36	PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65	PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61	VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36	VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39	PNE..LRQYLKELNEYAIMGRTRF
NPF_MONEX/1-39	DNKAALRDYLRQINEYFAIIGRPRF
Q9PT97/29-62	AEE..LAKYYSALRHYINLITRQ..

*Option 1:* Ajouter des arcs entre les états matchants, mais nombre d'arcs quadratique

*Option 2:* Ajouter des états **silencieux**, qui n'émettent rien

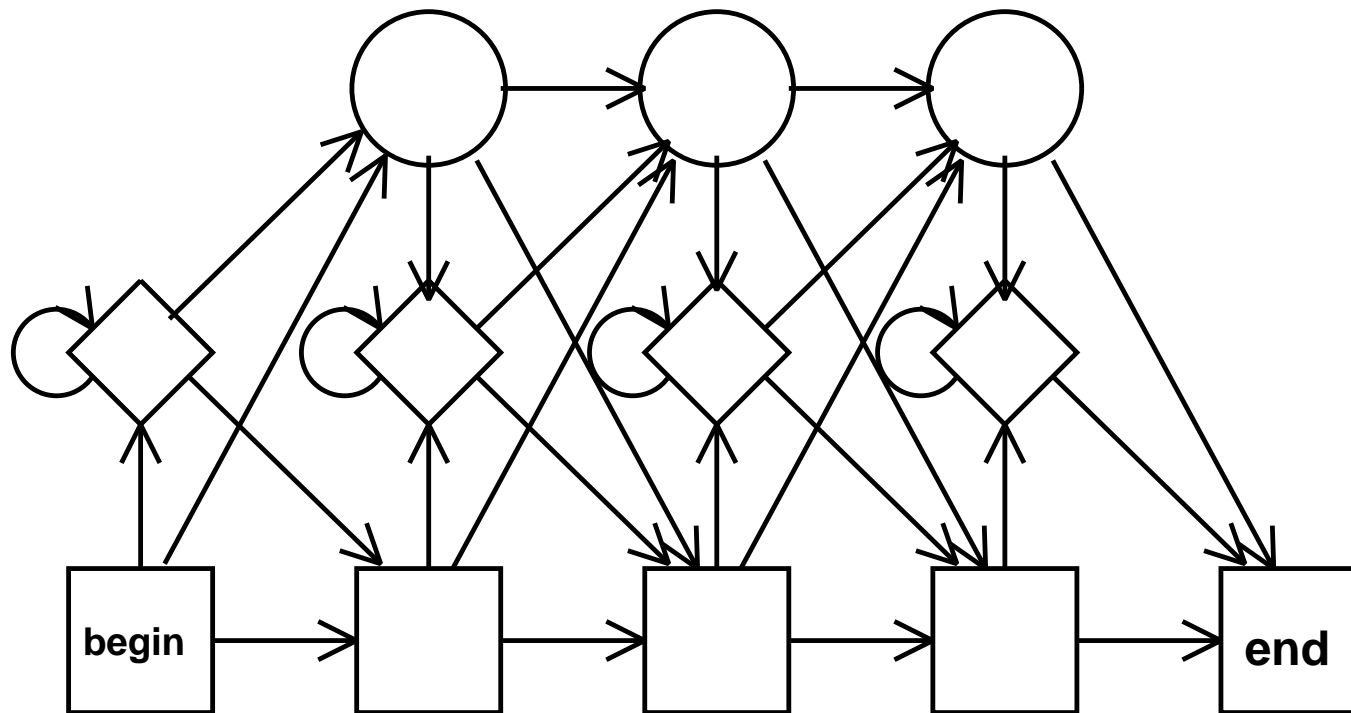
**un état = un a.a. délété**



## En résumé

- ▷ *États matchants* : colonnes avec moins de 50% de -
- ▷ *États d'insertion* : majorité de -
- ▷ *États 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-comptes* : +1 à chaque compte (loi de Laplace)

Modèle complet :



## Recherche avec un profil HMM

- ▷ Score : trouver la meilleure interprétation de la séquence dans le modèle

**Algorithme de Viterbi** (similaire à l'alignement deux à deux)

- ▷ Seuil d'admission : **E-value**

Banque de référence : 59 021 séquences de longueur moyenne 359 - (distribution 34 de SwissProt)

**Exemple :** la séquence NPF\_ARTTR contre le HMM de l'hormone pancréatique

Alignments of top-scoring domains:

hormone3: domain 1 of 1, from 3 to 36: score 48.4, E = 1.1e-13

```
*->yPskdfPenPGddaspEeelaqYlraLrqYinliTRpRY<-*  
      +++++      P++++s+E+e+++Ylr++++Yi+l++RpR+  
3    VHLR-----PRSSFSEDEYQIYLRNVSKYIQLYGRPRF      36
```

**Pfam:** alignements et HMM pour 7255 familles de protéines (novembre 2003)