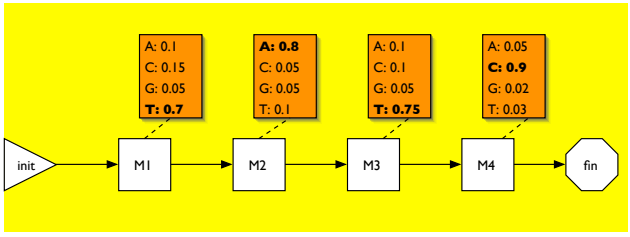
5 Applications de HMMs

5.1 Motif de séquence conservé

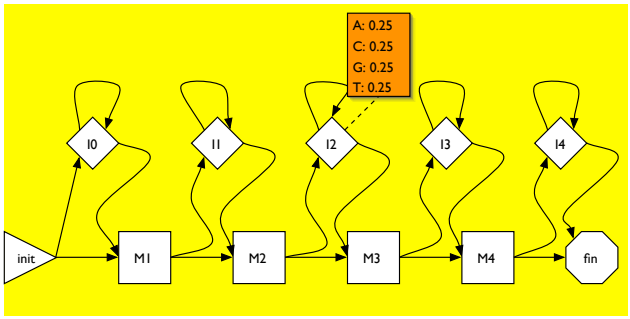
Un alignment multiple de protéines peut révéler des motifs conservés :

```
Q9DAK4/4-132      QLQG.TWKSVSCDNF.ENYMKELGVGRASRK.LGCLAK.......PTVT
Q8QHA8/4-132      HFVG.TWKLLSSENF.EDYMKELGVGFATRK.MAGVAK.......PNLT
FABA_RAT/3-131    AFVG.TWKLVSSENF.DDYMKEVGVGFATRK.VAGMAK.......PNLI
MYP2_HUMAN/3-131  KFLG.TWKLVSSENF.DDYMKALGVGLATRK.LGNLAK.......PTVI
TLBP_MOUSE/4-132  PFLG.TWKLVSSENF.ENYVRELGVECEPRK.VACLIK.......PSVS
O57663/4-133      KFVG.TWKMISSDNF.DDYMKAIGVGFATRQ.VGNRTK.......PNLV
FABE_RAT/6-134    DLEG.KWRLVESHGF.EDYMKELGVGLALRK.MGAMAK.......PDCI
FABE_BOVIN/6-134  QLVG.RWRLVESKGF.DEYMKEVGVGMALRK.VGAMAK.......PDCI
```

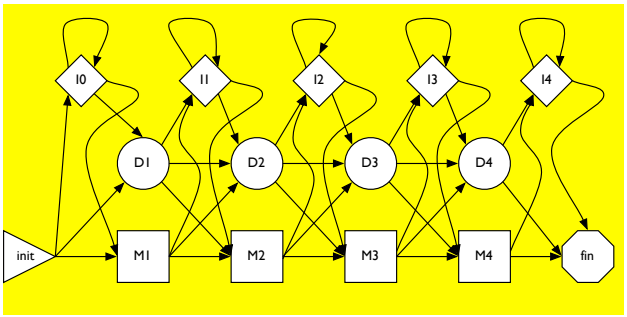
HMM de profile. On construit un modèle HMM dont les états correspondent aux colonnes de l'alignement :

1. 

substitutions : colonnes de résidus alignés

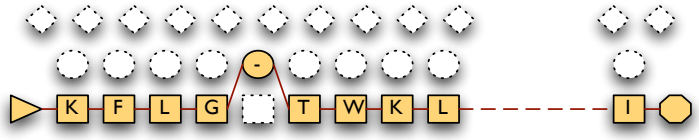
2. 

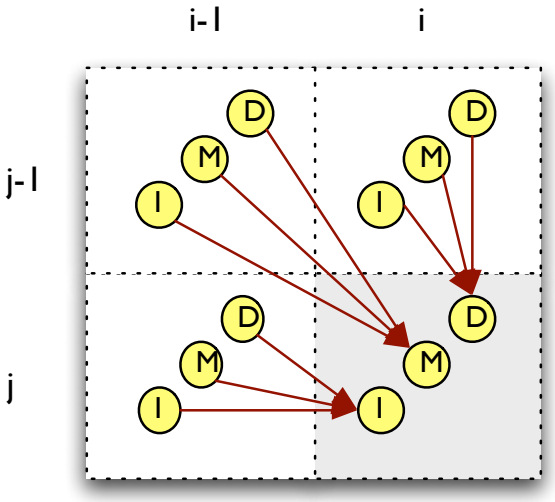
insertions : résidus entre colonnes conservées

3. 

suppressions : états sans émission !

Alignement à un profile. On calcule le chemin Viterbi pour aligner une nouvelle séquence au profile





Programmation dynamique :
- ⋆ états de match : M_j : $j = 1, \ldots, n$
- ⋆ états d'insertion : I_j : $j = 0, \ldots, n$
- ⋆ états de suppression : stD_j : $j = 1, \ldots, n$
- ⋆ sous-problèmes pour PD : $V^{(\mathsf{M})}(i,j), V^{(\mathsf{I})}(i,j), V^{(\mathsf{D})}(i,j)$ pour émission de $s_1 \cdots s_i$ en finssant dans un état M_j, D_j, I_j

Pondération par profile. Hypothèse null : $S[1..\ell]$ est une séquence de caractères iid $\mathbb{P}\{S[i] = a\} = r_a$ (souvent les mêmes probs d'émission aux états d'insertion). Hypothèse altérnatif :$S[1..\ell]$ est générée par notre HMM de profile $\mathcal{M}$. LODS (logarithme des chances) : comparer $p_0 = \mathbb{P}\Big\{ S[1..\ell] \;\Big|\; H_0 \Big\}$ et $p_{\mathcal{M}} = \mathbb{P}\Big\{ S[1..\ell] \;\Big|\; H_1 \Big\}$:

$$\mathsf{LODS} = \log \frac{p_{\mathcal{M}}}{p_0}$$

Calcul de LODS : par Viterbi, remplacer probabilité d'émission $p_q(a)$ par $p_q(a)/r_a$ + logarithme

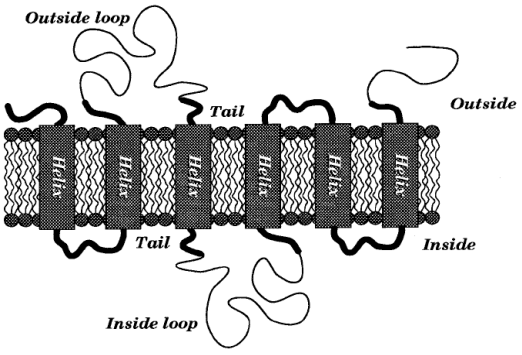
Bases de données de profiles. PFAM et PROSITE utilisent des HMMs de profile pour des familles de protéines. Exemple : `http://prosite.expasy.org/PS50071`. On trouve la pondération de substitutions par colonne, ainsi que la pondération de transitions entre états (en LODS).

```
ID   HOMEOBOX_2; MATRIX.
...
                A   B   C   D   E   ...
...
/M: SY='E'; M= -5,  2,-25,  3, 11, ...
/M: SY='Q'; M= -3, -4,-25, -4, 12, ...
...
/I:         I=-8; MI=-8; IM=-8; DM=-15; MD=-15;
```

5.2 Architectures spécialisées

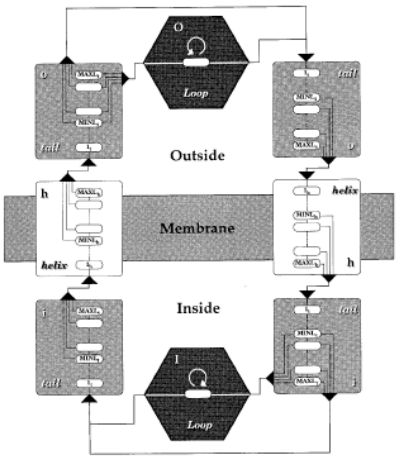
L'architecture de l'HMM (transitions permises entre états) peut capturer des contraintes spécifiques.

Protéines transmembranaires. On définit des états pour l'intérieur/extérieur de membrane.



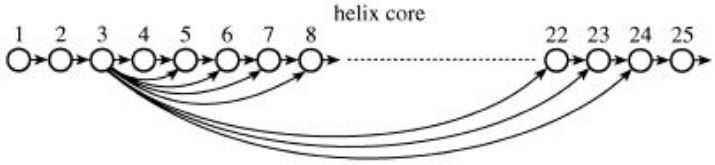
Amino acid seq: MGDVCDTEPGILVA...SVALRPRKHGRWIV...FWVDNGTEQ...PEHMTKLHMM...
State seq: oooooooohhhhh...hhhhiiiiiiihhh...hhhooooOO...OOOooooohhh...

(Tusnády & Simon, J.Mol.Biol. 283:489)



(Tusnády & Simon, J.Mol.Biol. 283:489)

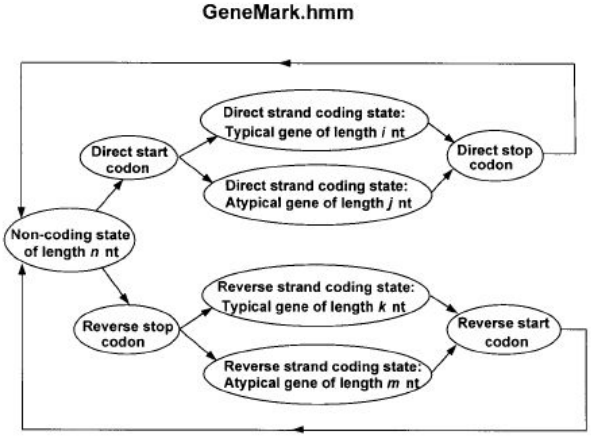
Modèles de duration. duration de séjour dans un état : p.e., longueur du segment restreinte pour hélices dans le membrane (min et max) :



5.3 Gènes

Les états de l'HMM modèlisent des composantes d'un gène.

GeneMark. Gènes compactes chez procaryotes. . .



GENSCAN. Eucaryotes : modèles plus compliqués.