Démo 4 :

Les Alignements Multiples

Retrouver le maximum de similarité entre plus de deux séquences. En essayant de placer le maximum de résidus dans une même colonne, en introduisant les trous (correspondant aux insertions et suppressions).
Donc à partir d'un alignement multiple, il sera possible de voir les mutations qui se sont déroulées durant l`évolution de la séquence et donc dériver l'ancestrale commune.

Utilité des alignements multiples

- ? Trouver l'homologie entre les séquences (des résidus conservés qui ont un sens biologique) peuvent impliqués une homologie.
- ? Repérer des résidus fortement conservés qui ont de fortes chances d'appartenir à des sites important pour la structure ou la fonction.
- ? Prédire la fonction
- ? Prédire la structure (souvent utilisés pour la comparaison des structures prédites de séquences avec celles qui sont homologues dont la structure est connue).
- ? Etc..

Exemple de la méthode progressive avec CLUSTALW

Etapes :

1. Calcul de scores deux à deux entre les séquences.
2. Construire un arbre guide reflétant la similarité entre les séquences à partir de la matrice des scores en utilisant une méthode de regroupement.
3. Aligner les séquences en suivant l'arbre.

Alignement progressif avec CLUSTALW des 5 séquences suivantes :

>Seq1
ATCTCGAGA
>Seq2
ATCCGAGA
>Seq3
ATGTCGACGA
>Seq4
ATGTCGACAGA
>Seq5
ATTCAACGA

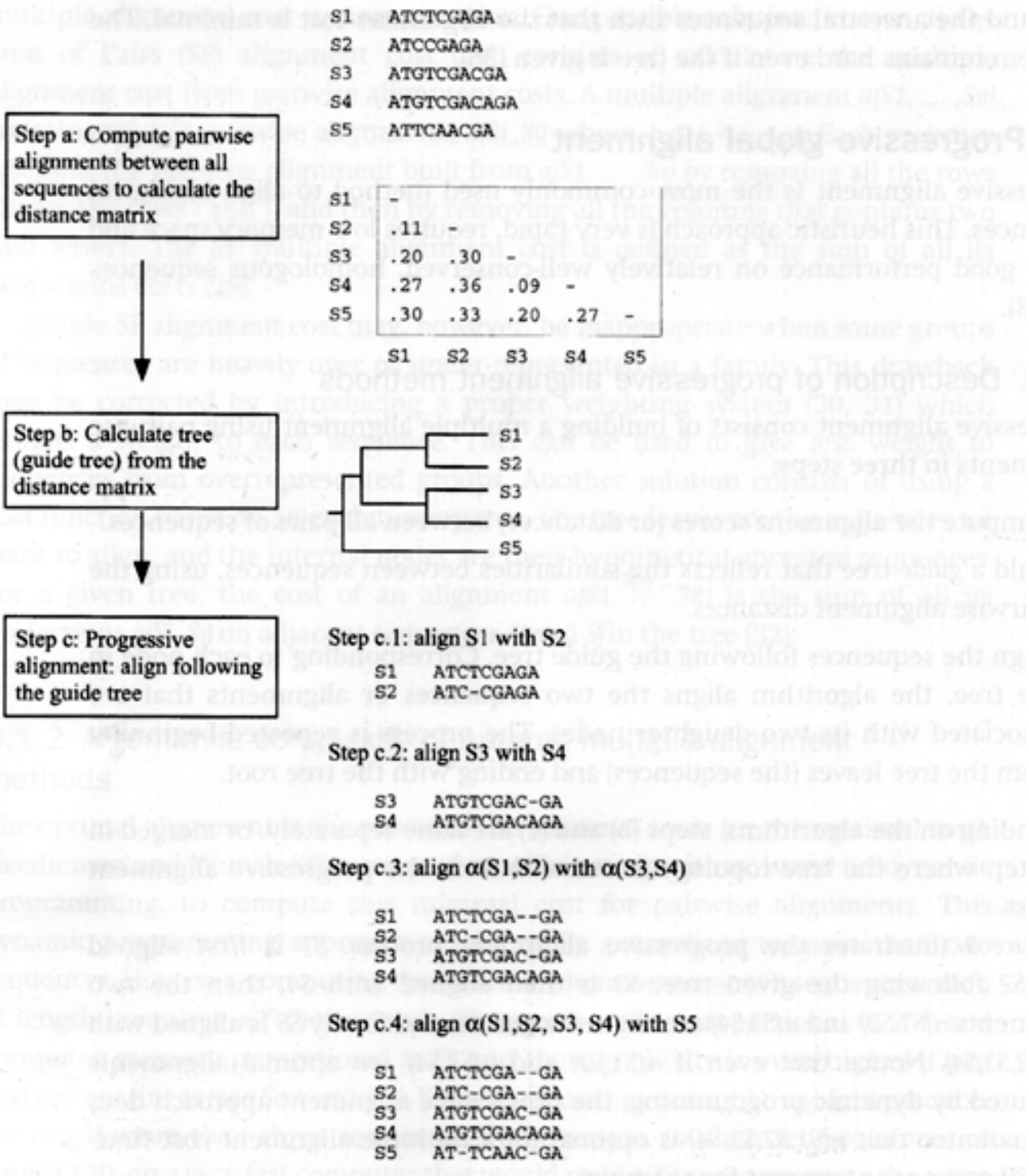
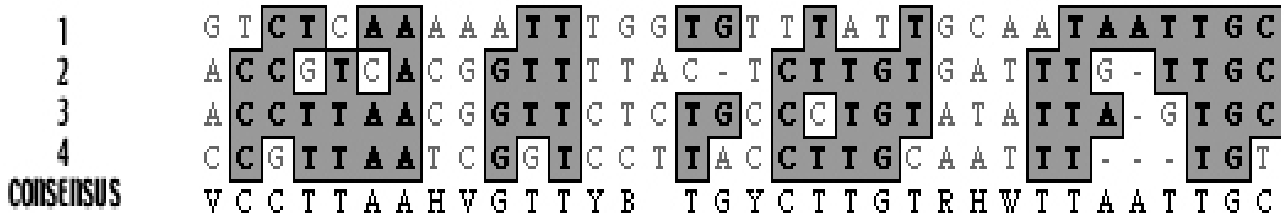


Figure 2 Progressive alignment process. (a) All sequences are compared to each other S_2. (b) A guide tree is calculated from the pairwise distance matrix. (c) Sequences are progressively aligned following the guide tree.

```
CLUSTAL W (1.82) Multiple Sequence Alignments



Sequence format is Pearson
Sequence 1: Seq1              9 bp
Sequence 2: Seq2              8 bp
Sequence 3: Seq3             10 bp
Sequence 4: Seq4             11 bp
Sequence 5: Seq5              9 bp
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  62
Sequences (1:3) Aligned. Score:  66
Sequences (1:4) Aligned. Score:  77
Sequences (1:5) Aligned. Score:  33
Sequences (2:3) Aligned. Score:  37
Sequences (2:4) Aligned. Score:  50
Sequences (2:5) Aligned. Score:  37
Sequences (3:4) Aligned. Score:  80
Sequences (3:5) Aligned. Score:  66
Sequences (4:5) Aligned. Score:  44
Guide tree        file created:   [/ebi/extserv/old-
work/116342.26829.dnd]
Start of Multiple Alignment
There are 4 groups
Aligning...
Group 1: Sequences:   2      Score:115
Group 2: Sequences:   3      Score:116
Group 3: Sequences:   2      Score:124
Group 4: Sequences:   5      Score:106
Alignment Score 191
CLUSTAL-Alignment file created  [/ebi/extserv/old-
work/116342.26829.aln]



CLUSTAL W (1.82) multiple sequence alignment


Seq1            ATCTCGAGA-- 9
Seq2            ATC-CGAGA-- 8
Seq4            ATGTCGACAGA 11
Seq3            ATGTCGACGA- 10
Seq5            AT-TCAACGA- 9
                **  * *
```

Avec Consensus



Un exemple avec l'alignement de la globine de différents organismes ou on n'observe pas une forte conservation des résidus, par contre les protéines ont 7 hélices de la structure tertiaires se conservent. (La présences des 2 histidines 2 H) qui relient un groupe

```
HBB_HUMAN       -------MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQR
HBB_HORSE       --------VQLSGEEKAAVLALWDKV--NEEEVGGEALGRLLVVYPWTQR
HBA_HUMAN       --------MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT
HBA_HORSE       --------MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKT
MYG_PHYCA       ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
GLB5_PETMA      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQE
LGB2_LUPLU      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKD
                         *:  :   :   *  .           :  .:   * :

HBB_HUMAN       FFESFGDLSTPDAVMGNPKVKAHGKKVL-----GAFSDGLAHLDNLKGTF
HBB_HORSE       FFDSFGDLSNPGAVMGNPKVKAHGKKVL-----HSFGEGVHHLDNLKGTF
HBA_HUMAN       YFPHF-DLS-----HGSAQVKGHGKKVA-----DALTNAVAHVDDMPNAL
HBA_HORSE       YFPHF-DLS-----HGSAQVKAHGKKVG-----DALTLAVGHLDDLPGAL
MYG_PHYCA       KFDRFKHLKTEAEMKASEDLKKHGVTVL-----TALGAILKKKGHHEAEL
GLB5_PETMA      FFPKFKGLTTADQLKKSADVRWHAERII-----NAVNDAVASMDDTEKMS
LGB2_LUPLU      LFSFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATL
                 *  :   .       . .:: *.  :       :.   :

HBB_HUMAN       ATLSEL---HCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQK
HBB_HORSE       AALSEL---HCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQK
HBA_HUMAN       SALSDL---HAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDK
HBA_HORSE       SNLSDL---HAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDK
MYG_PHYCA       KPLAQS---HATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNK
GLB5_PETMA      MKLRDLSGKHAKSFQVDPQYFKVLAAVI----A-----DTVAAGDAGFEK
LGB2_LUPLU      KNLGSV---HVSKGVAD-AHFPVVKEAILKTIKEVVGAKWSEELNSAWTI
                  * .    *  .       : .:   :          .     ...
```

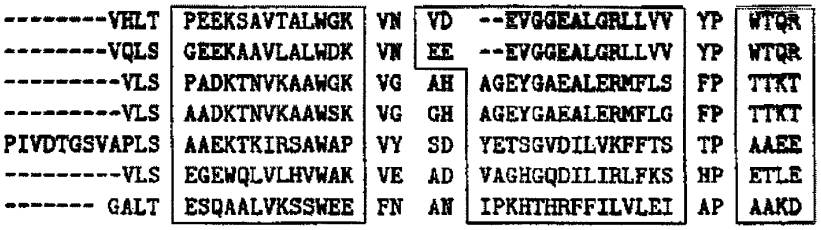
```
HBB_HUMAN       VVAGVANALAHKYH------
HBB_HORSE       VVAGVANALAHKYH------
HBA_HUMAN       FLASVSTVLTSKYR------
HBA_HORSE       FLSSVSTVLTSKYR------
HYG_PHYCA       ALELFRKDIAAKYKELGYQG
GLB5_PETMA      LMSMICILLRSAY-------
LGB2_LUPLU      AYDELAIVIKKEMNDAA---
                    .    :
```

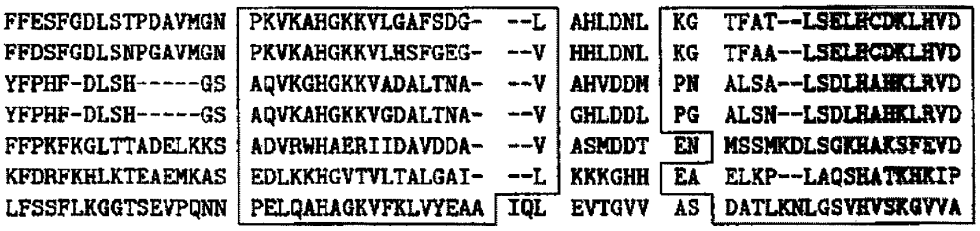
- *** = identity**
- **: = strongly conserveds**
- **. = weakly conserved**

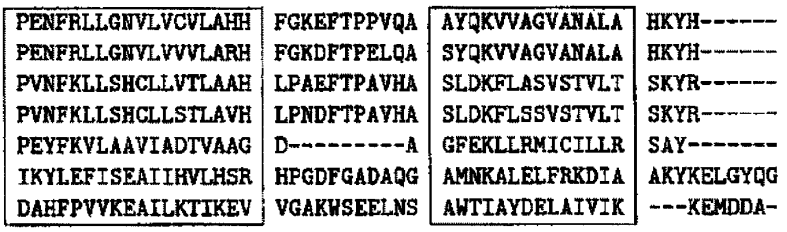
Exemple d'expression régulière à partir de l'alignement ci-dessus :

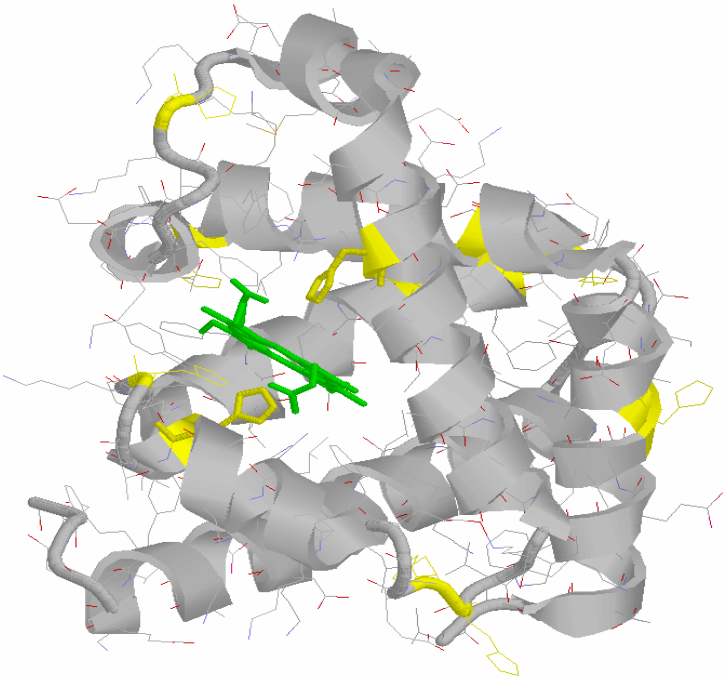
L-[TS]-x(2)-[EDQ]-x(3)-[VI]-x(3)-[W]…

Alors qu'elles ont les mêmes 7 hélices structures









Recherche d’un profile dans PROSITE

[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]

```
ID   MYRISTYL; PATTERN.
AC   PS00008;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-1990 (INFO
UPDATE).
DE   N-myristoylation site.
PA   G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}.
CC   /TAXO-RANGE=??E?V;
CC   /SITE=1,myristyl;
CC   /SKIP-FLAG=TRUE;
DO   PDOC00008;
//


ID   PROKAR_LIPOPROTEIN; RULE.
AC   PS00013;
```

```
DT   APR-1990 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO
UPDATE).
DE   Prokaryotic membrane lipoprotein lipid attachment site.
PA   {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C.
RU   Additional rules:
RU   (1) The cysteine must be between positions 15 and 35 of the
sequence in
RU       consideration.
RU   (2) There must be at least one charged residue (Lys or Arg) in the
first
RU       seven residues of the sequence.
NR   /RELEASE=40.7,103373;
NR   /TOTAL=449(448); /POSITIVE=416(415); /UNKNOWN=30(30);
/FALSE_POS=3(3);
NR   /FALSE_NEG=17; /PARTIAL=5;
CC   /TAXO-RANGE=AB?P?; /MAX-REPEAT=1;
CC   /SITE=5,lipid;
DR   P50927, 17KD_RICAM, T; P50928, 17KD_RICAU, T; P05372, 17KD_RICCN, T;
DR   Q52764, 17KD_RICJA, T; P50929, 17KD_RICMO, T; P50930, 17KD_RICPA, T;
DR   P16624, 17KD_RICPR, T; P50931, 17KD_RICRH, T; P22882, 17KD_RICTY, T;
DR   P31502, 19KD_MYCIT, T; P11572, 19KD_MYCTU, T; O83142, 5NTD_TREPA,
T;
DR   Q9KQ30, 5NTD_VIBCH, T; P22848, 5NTD_VIBPA, T; P31223, ACRA_ECOLI,
T;
……..

DO   PDOC00013;
//
```

```
ID   HSP20; MATRIX.
   AC   PS01031;
   DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); NOV-1995
(INFO UPDATE).
   DE   Heat shock hsp20 proteins family profile.
   MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY';
LENGTH=97;
   MA   /DISJOINT: DEFINITION=PROTECT; N1=2; N2=96;
```

Format :
/DISJOINT : DEFINITION=name; parameters;
notion de non jointures entre deux alignements entre le même profile et la même séquence

PROTECT N1 (int)
N2 (int)

Deux alignements profile-séquence sont disjjoints s'il n'y a pas de chevauchemenet. La région protégée du profile s`étend du match à la position N1 au match à la position N2.

```
   MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE;
```

GLE_ZSCORE

Y = (X/[R1*(1.0-exp(R2*SeqLen-R3))] –R4) / R5

```
   MA    R1=239.0; R2=-0.0036; R3=0.8341; R4=1.016; R5=0.169;
   MA   /CUT_OFF: LEVEL=0; SCORE=400; N_SCORE=10.0; MODE=1;
   MA   /DEFAULT: MI=-210; MD=-210; IM=0; DM=0; I=-20; D=-20;
   MA   /M: SY='R'; M=-12,-44,-11,-13,-13,-22,-2,-7,18,-12,5,-3,-
11,0,21,-6,-5,-11,-16,-34;
   MA   /M: SY='D'; M=1,-41,17,16,-41,-3,3,-11,-1,-22,-12,8,-
7,12,-7,0,-2,-19,-53,-36;
 …….
   MA   /I: MI=-55; MD=-55; I=-5;
   MA   /M: SY='P'; D=-5; M=1,-2,-1,0,-3,0,0,-1,-1,-2,-
2,0,4,0,0,1,0,-1,-4,-4;
   MA   /I: MI=-55; MD=-55; I=-5;
```

Les chaînes de Markov Cachées

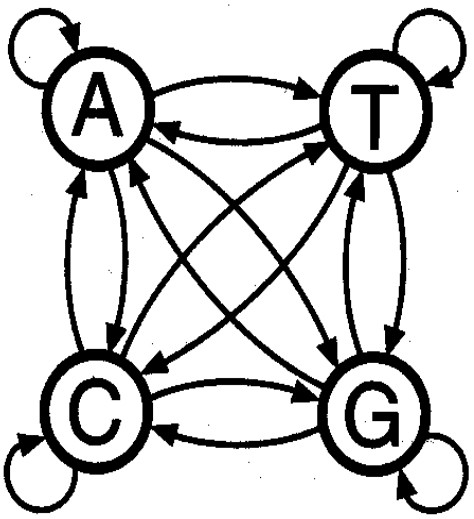
Le premier domaine dans lesquels ils ont été utilisés est celui de la reconnaissance de la parole.

Problèmes de reconnaissance de textes manuscrits.
Analyses de séquences biologiques
Reconnaissance d'images.
Modélisation d'un signal musical. Etc…

Exemple des îles de CpG

Le problème de retrouver les paires de CG apparaissant de façon successive le long d'une séquence d'ADN.

Un modèle simple pour représenter les séquences est celui des chaînes de Markov



A chaque état correspond une des 4 lettres de l'alphabet, à chaque transition d'un résidu à un autre, correspond une probabilité de transition.

$$a_{st} = P(x_i = t \mid x_{i-1} = s)$$

La probabilité de chaque symbole x_i dépend uniquement de son prédécesseur et non toute la partie précédente de la séquence.

$$P(x) = P(x_1) \ ? \ a_{x_{i-1}x_i} \text{ de } i=2 \text{ à } L$$

En ajoutant deux états muets pour le début et la fin, on obtient le modèle suivant :

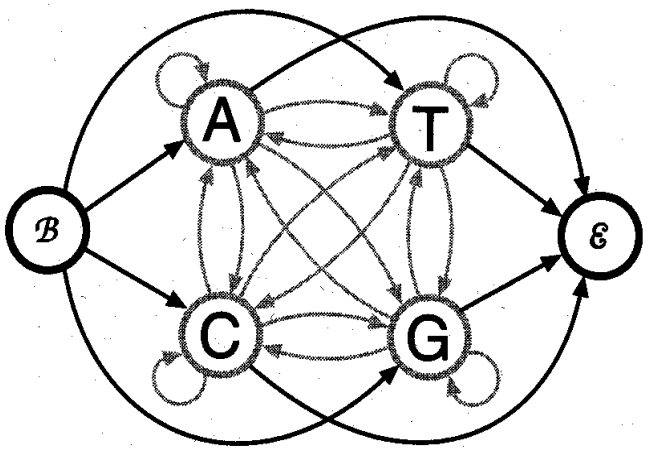


Figure 3.1 *Begin and end states can be added to a Markov chain (grey model) for modelling both ends of a sequence.*

Ceci aura pour effet de transformer la probabilité initiale en une probabilité de transition

Le problème des îles CpG peut être décomposé en deux modèles de markov Un pour les régions nommées les îles CpG (modèle +) et l'autre pour le reste de la séquence (modèle -).

+	A	C	G	T	-	A	C	G	T
A	0.180	0.274	0.426	0.120	A	0.300	0.205	0.285	0.210
C	0.171	0.368	0.274	0.188	C	0.322	0.298	0.078	0.302
G	0.161	0.339	0.375	0.125	G	0.248	0.246	0.298	0.208
T	0.079	0.355	0.384	0.182	T	0.177	0.239	0.292	0.292

Pour faire la discrimination entre ces deux modèles on dresse calcule le loge des probabilités de chacun des modèles pour la séquence x

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

$$= \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

β	A	C	G	T
A	−0.740	0.419	0.580	−0.803
C	−0.913	0.302	1.812	−0.685
G	−0.624	0.461	0.331	−0.730
T	−1.169	0.573	0.393	−0.679

Plus le score est élevé plus la probabilité que la séquence x est une île de CpG

Le deuxième problème sera la localisation de ces îles CpG :
Idée naïve : extraire une fenêtre d'une longueur très inférieure à celle de la séquence. Les séquences avec un score positif seront des îles CpG potentielles.

Désavantages : aucune information concernant le taille de ces îles.
La solution : combiner les deux modèles de Markov avec une petite probabilité de transition entre les deux chaînes.

Exemple d'un HMM pour modéliser un casino malhonnête

Le croupier utilise un vrai dé la plupart du temps, mais des fois utilises un dé truqué. Le dé truqué a une probabilité 0.5 pour un 6 et de 0.5 pour chacun des nombres 1 à 5. Le croupier bascule entre un vrai dé et le truqué avec une probabilité de 0.05 et de faire le changement inverse avec une probabilité de 0.1.
Soit une série de lancé, on aimerait savoir à quel moment le vrai dé a été utilisé eu quand le faux a été utilisé.

Le HMM correspondant est :

Les états sont : Q = {F, L}
L'alphabet : S = {1, 2, 3, 4, 5, 6}

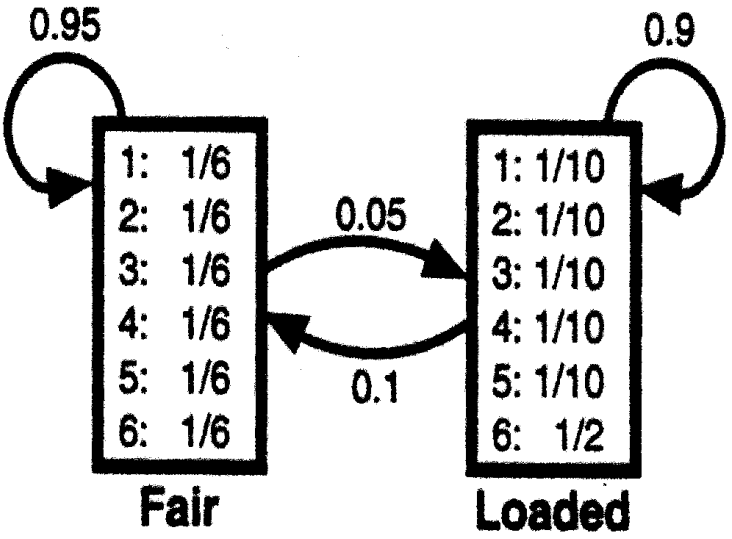


Figure 5.3: Source: [4]. HMM for the dishonest casino problem.

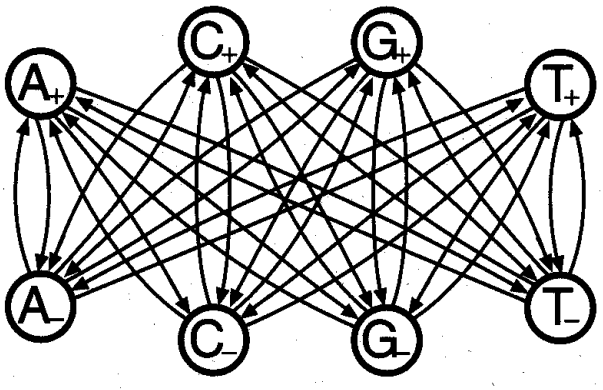


Figure 3.3 *An HMM for CpG islands. In addition to the transitions shown, there is also a complete set of transitions within each set, as in the earlier simple Markov chains.*

Remarque dans le cas d'un HMM : La probabilité de se déplacer à un état, ne dépend que de celui de l'état précédent. Par contre la correspondance état-symbole n'est plus valable. Donc il est définit la probabilités d'émission, qui est la probabilité que le symbole b soit vu quand on est dans un état k.

Le modèle contient 8 états

Etat : A+ C+ G+ T+ A- C- G- T-
Symbole émis : A C G T A C G T

```
Rolls   3151162464466442453113216311641521336251445436316566266566666
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL

Rolls   651166453132651245636664631636663162326455236266666625151631
Die     LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLPFFLLLLLLLLLLLLLLLFFFFFFFFF
Viterbi LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls   22255544166656656356432436413151346514635341112641462625335 6
Die     FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls   3661636664662325344136616611632525624622552652522664353533 6
Die     LLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi LLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
•

Rolls   233121625364414432335163243633665562466662632666612355245242
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
```