Prédiction de novo de la structure 3D des protéines avec les HMM

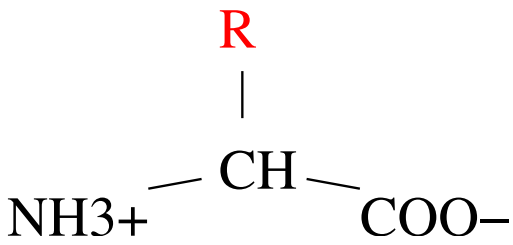
Thèse co-encadrée par

Jean-François Gibrat et François Rodolphe.

Juliette Martin

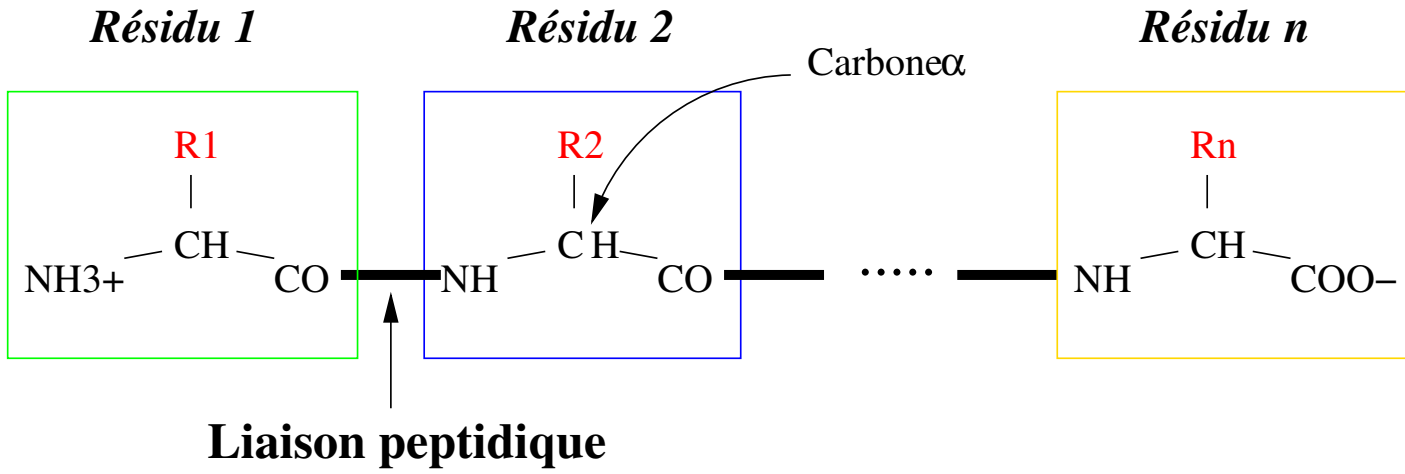
Unité MIG - INRA Jouy-en-Josas

Acides Aminés



20 acides aminés naturels principaux

Séquence

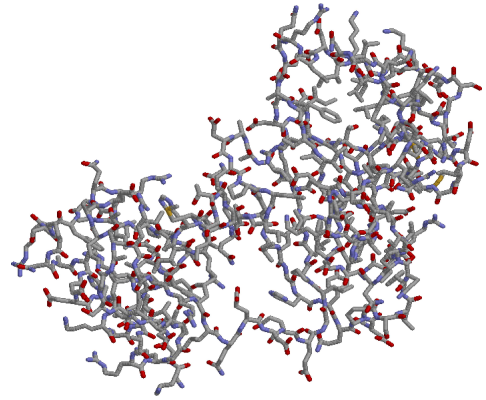


Séquence

MKALTARQQEVFDLIRDHISQTGMPPTRAEIAQRLGFRSPNAA
EEHLKALARKGVIEIVSGASRGIRLLQEEEEGLPLVGRVAADE

⇓

Structure 3D = fonction



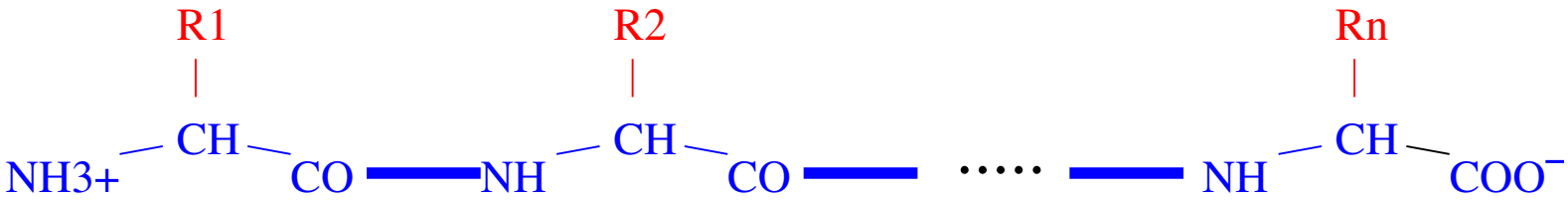
1,3 million de séquences ...mais 23 000 structures dans la Protein Data Bank (PDB) regroupés en environ 760 classes

Description de la structure 3D

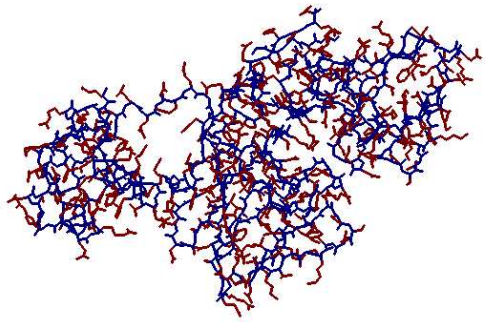
- Fichier PDB: coordonnées (x,y,z)des atomes lourds
- Séquence de 200 acides aminés $\rightarrow$ 1500 atomes lourds = 1500 coordonnées (x,y,z)

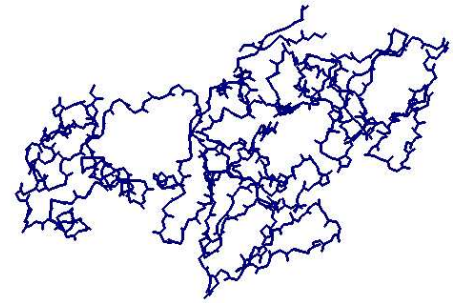
$\Rightarrow$ Simplifications pour décrire/prédire les structures 3D

Simplification(1)



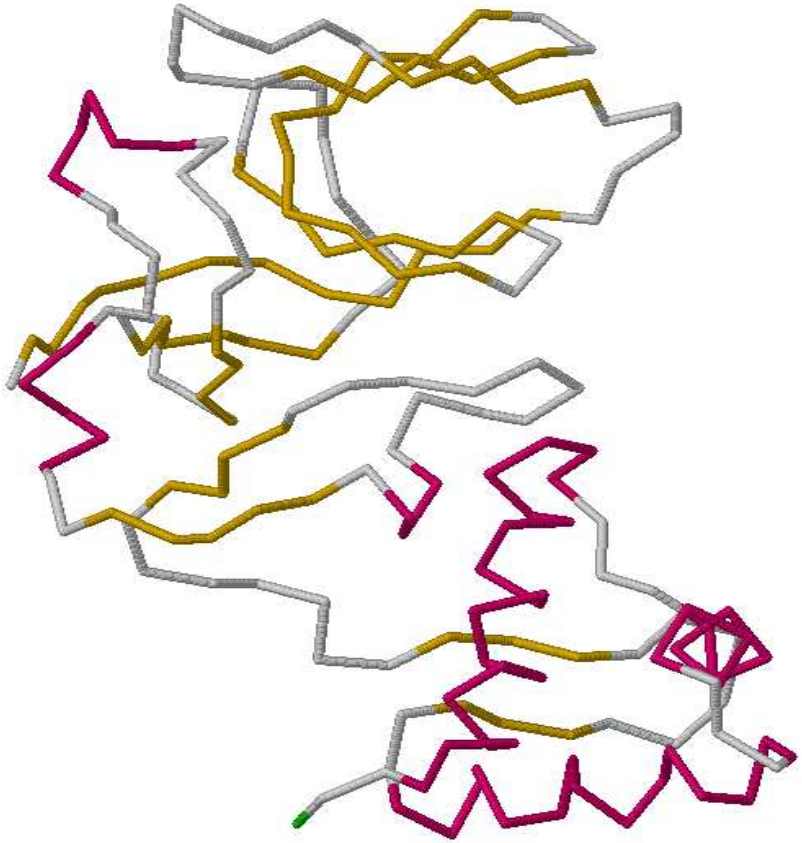
Structure complète - chaînes latérales =
chaîne principale ou squelette protéique

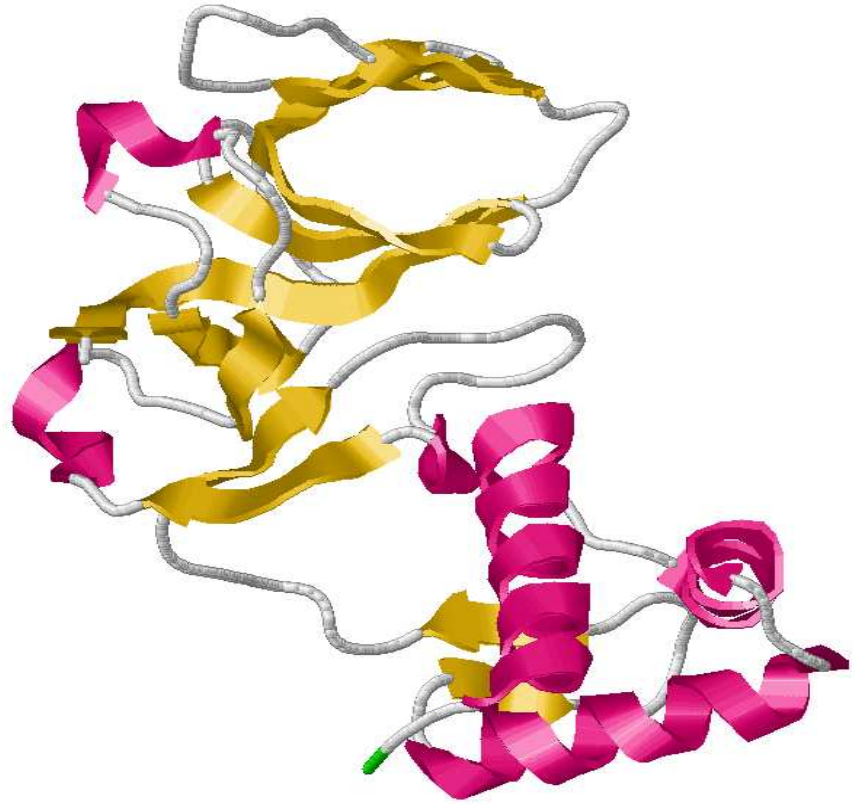




Simplification(2) : Structures secondaires

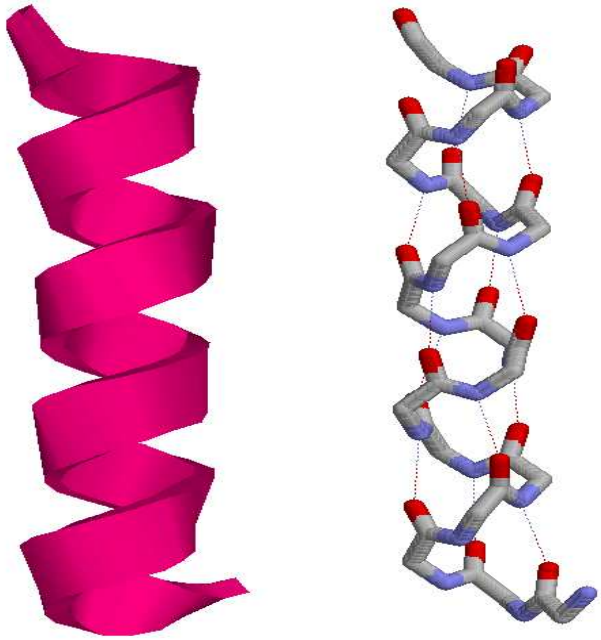
Sous-structures régulières



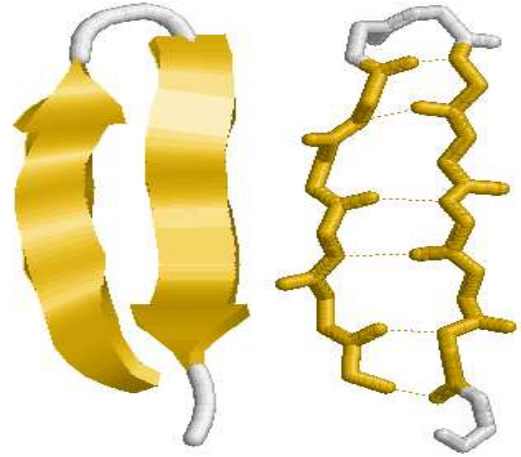


Structures secondaires

Hélices α



Feuillets β



Importance des structures secondaires

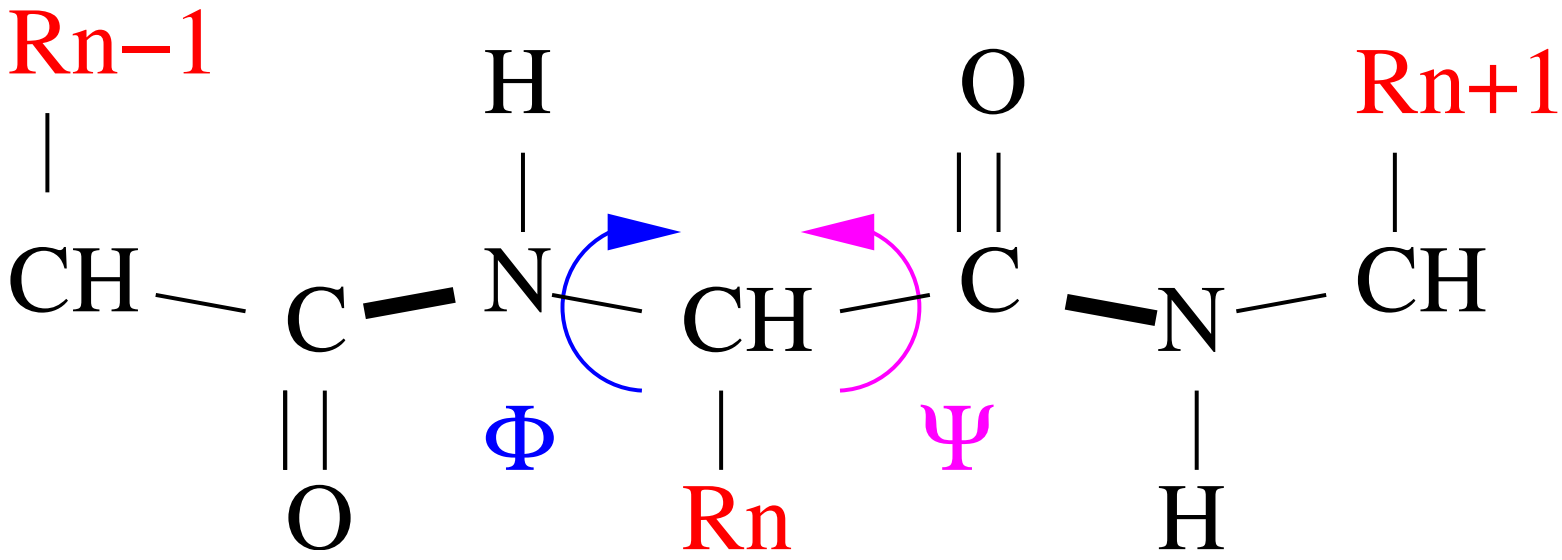
- Description simplifiée 1D de la structure 3D :

```
EEHLKALARKGVIEIVSGASRGIRLLQEEEEGLPLVGRVAADE

                        ↓

HHHHHHHHH..bbb.......bbb.......bbbb........
```

- Recouvrent en moyenne 50% des résidus des protéines
- Premier pas vers structure 3D
- Nombreuses méthodes de prédiction de structures secondaires

Simplification(3) : Angles dièdres

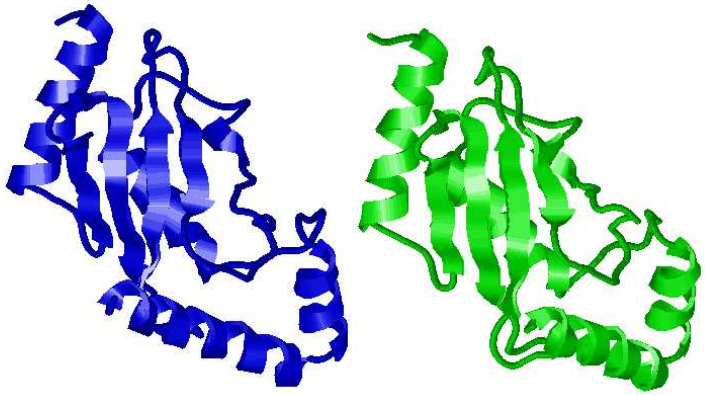
Φ et Ψ décrivent la chaîne principale



Protéine de n résidus $\rightarrow$ n-1 couples (Φ,Ψ)

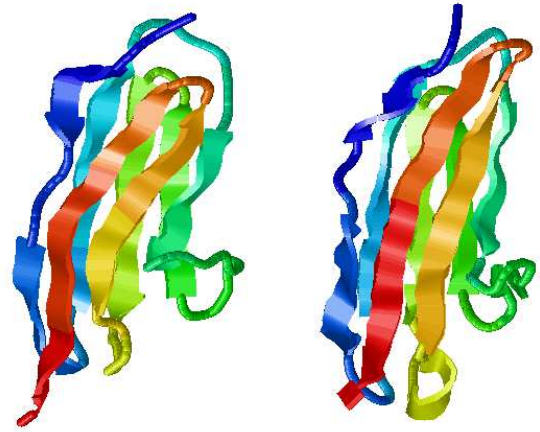
Relation séquence/structure

- La séquence détermine la structure 3D
- 2 protéines homologues (=ancêtre commun) ont des structures similaires, même à très bas taux d'identité



homologie détectable : 63% d'identité de séquence

Enzyme de conjugaison à l'ubiquitine chez la S. cerevisiae et A. thaliana



homologie lointaine : 9% d'identité de séquence

Twitchin (protéine musculaire) chez C elegans et chez l'homme

- Mais aussi : structures très proches mais homologie incertaine

Méthodes de prédiction de structure

1. Modélisation par homologie
2. Reconnaissance de repliements (threading)
   utilisent les structures disponibles
3. Modélisation de novo :
   - ne nécessite pas de structure homologue
   - détermination d'un nouveau repliement (environ 30% des protéines d'un génome)
   - méthodes récentes (Baker 2001)

Stratégie de prédiction de novo

Stratégie du local vers le global

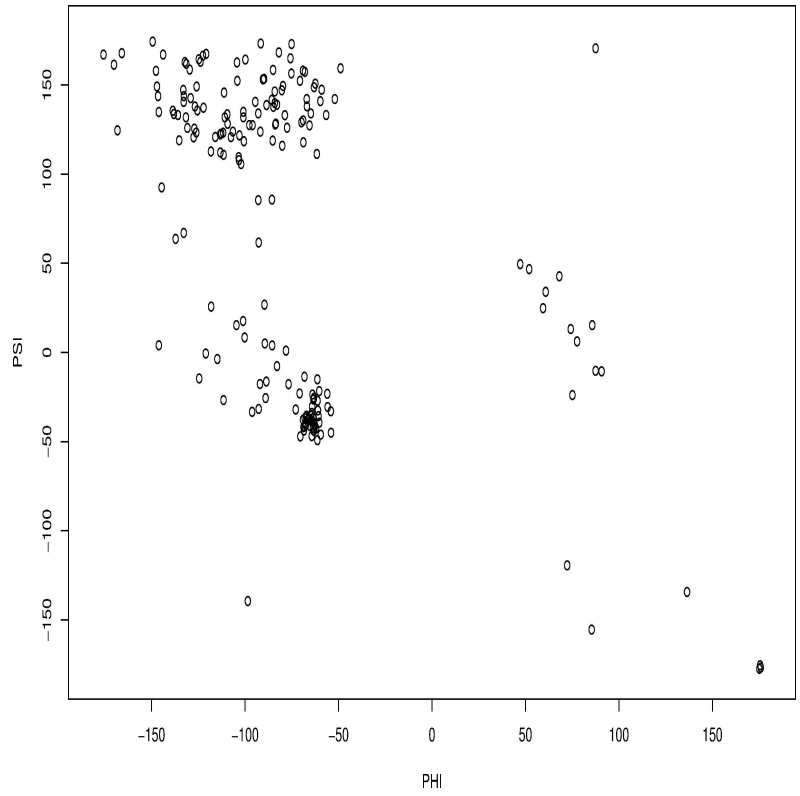
1. structures locales de fragments de la protéine
2. assemblage des fragments
3. sélection d'un modèle 3D

Étape 1 : utiliser la prédiction de structure locale pour choisir des fragments dans les structures connues
Prédiction de structure locale =

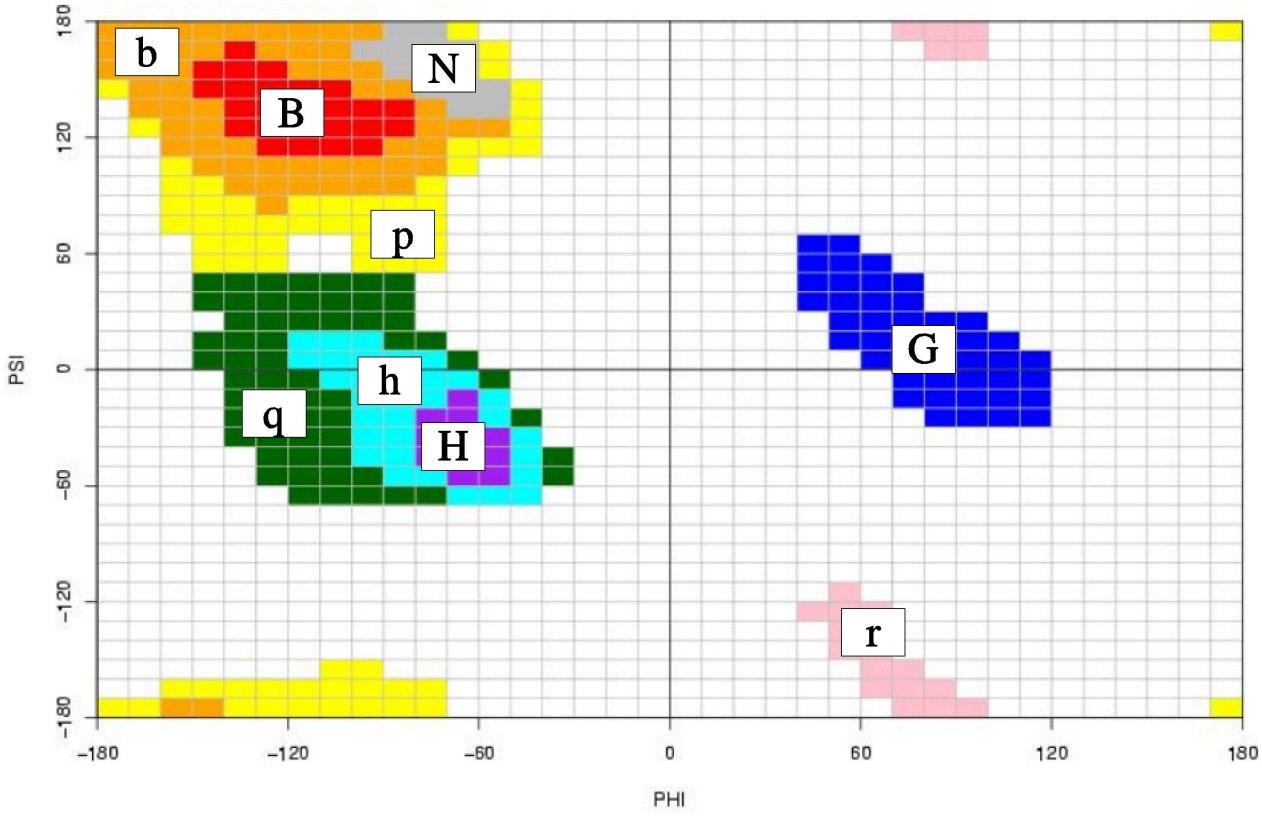
- Prédiction des angles Φ/Ψ
- Prédiction des structures secondaires

Prédiction des angles Φ/Ψ

Les angles Φ/Ψ ne prennent pas toutes les valeurs possibles:



Définition de zones d'angles



Puis recodage des structures en zones

Prédiction

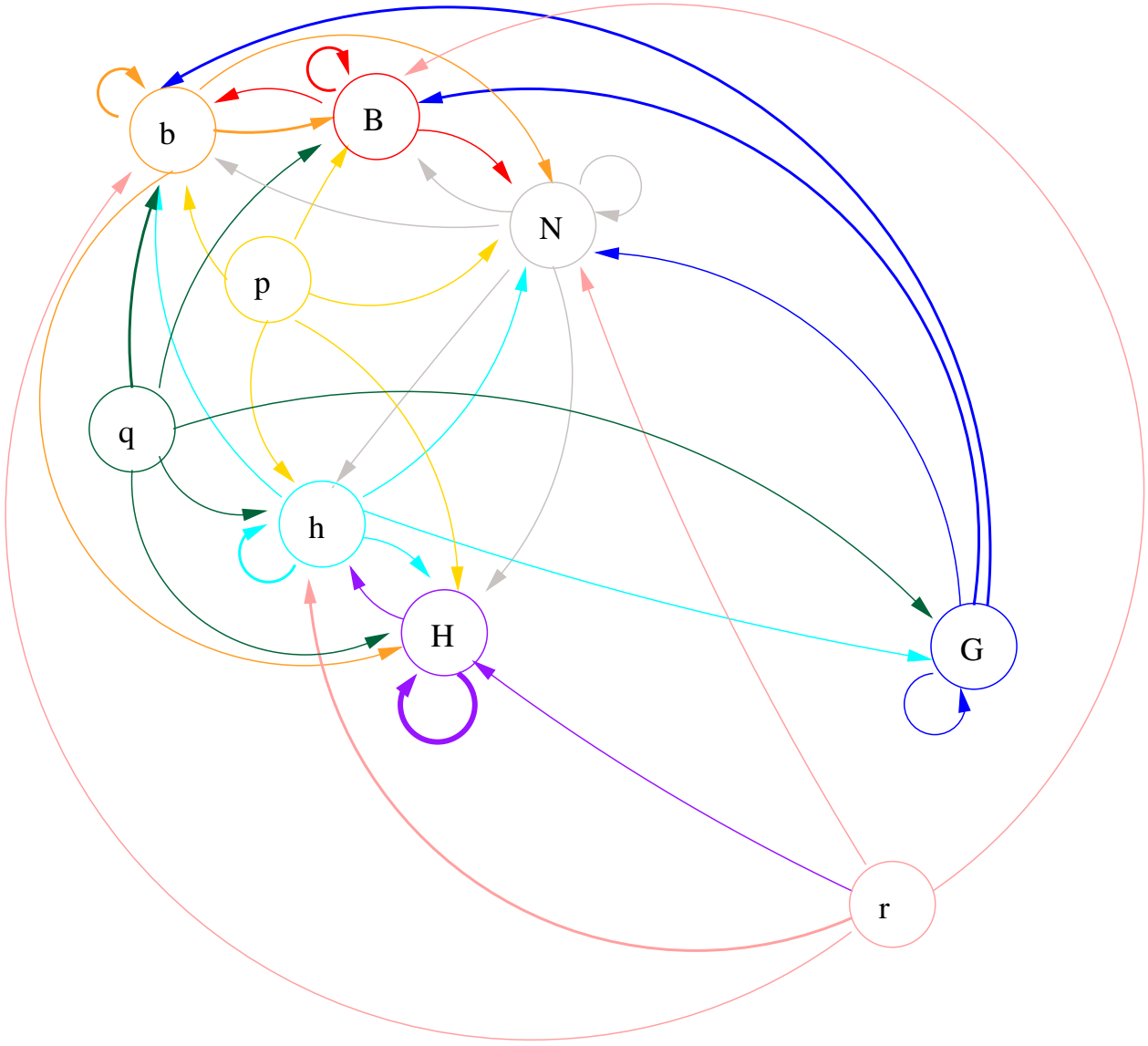
- But :

MKALTARQQEVFDLIRDHISQTGMPPTRAEIAQRLGFRSPNAA
EEHLKALARKGVIEIVSGASRGIRLLQEE

xxBBNHHHHHHHHHHHHHHHhqGBNNNHHHHHHHhGNHbHHHH
HHHHHHHHHhGhbBBBNrNNGpBBbbGHb

$\Rightarrow$ Utilisation des HMM

Modèle pour la prédiction des zones Φ/Ψ



- Modèle M1M0

$$\begin{array}{ccccccccc} S_1 & \rightarrow & S_2 & \rightarrow & S_3 & \rightarrow & S_4 & \rightarrow \ldots \rightarrow & S_n \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ X_1 & & X_2 & & X_3 & & X_4 & \ldots & X_n \end{array}$$

- Estimation des paramètres (9*9 transitions et 9*20 émissions) par comptage sur des structures connues
- Prédiction par forward/backward avec SHOW
- Performances : 40% des résidus correctement classés (20% aléatoire)
- Problème : état peu peuplés=sous-prédits

Étape intermédiaire: prédiction des structures secondaires

Prédiction de structures secondaires

De très nombreuses méthodes, divers modèles mathématiques.

1. Méthodes statistique prenant en compte les sites individuels (Chou et Fasman, 1974)
   $\rightarrow$ 55 % de bonne prédiction
2. Prise en compte de l'environnement local (GOR)
   $\rightarrow$ 65 % de bonne prédiction
3. Prise en compte des familles de protéines
   $\rightarrow$ 76 % de bonne prédiction

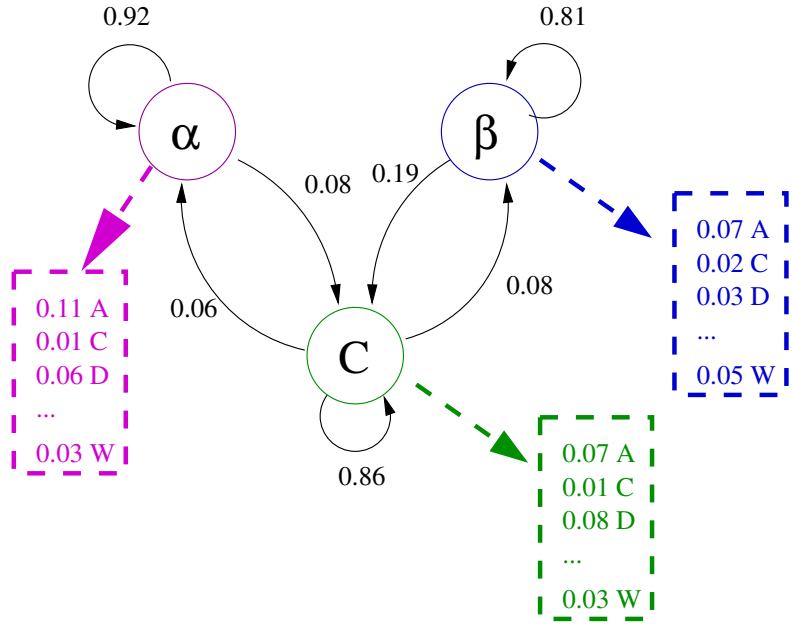
Prédiction avec un HMM

LAQQHIEGHYQVDPSLFKPN

↓

...HHH.bbbbb..HHHHH..

Modèle M1M0 :



3*3 transitions + 3*20 émissions

Performances : 59.6% de bonne prédiction (40% aleatoire)

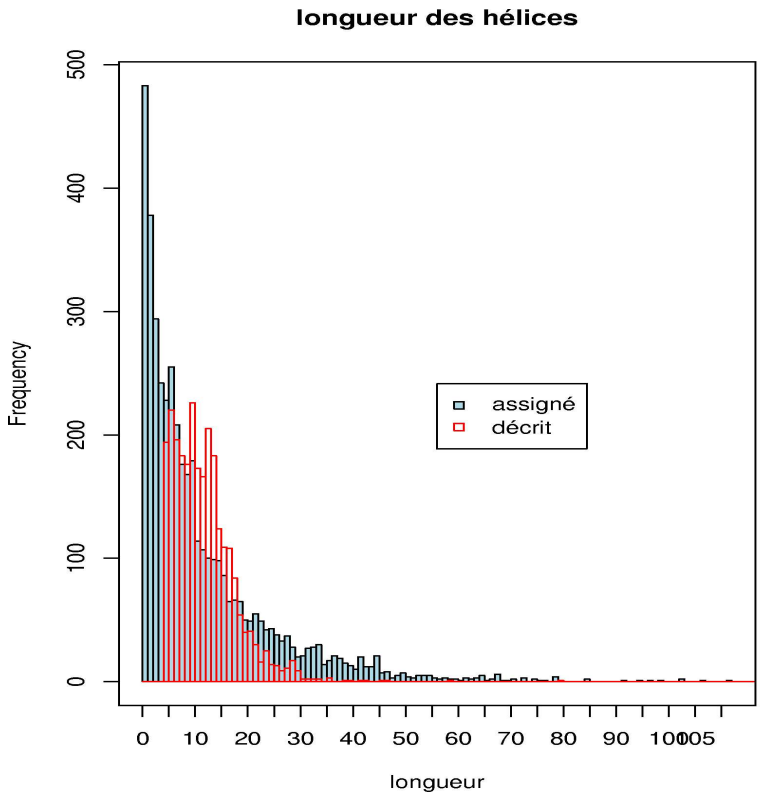
Augmentation de l'ordre du modèle

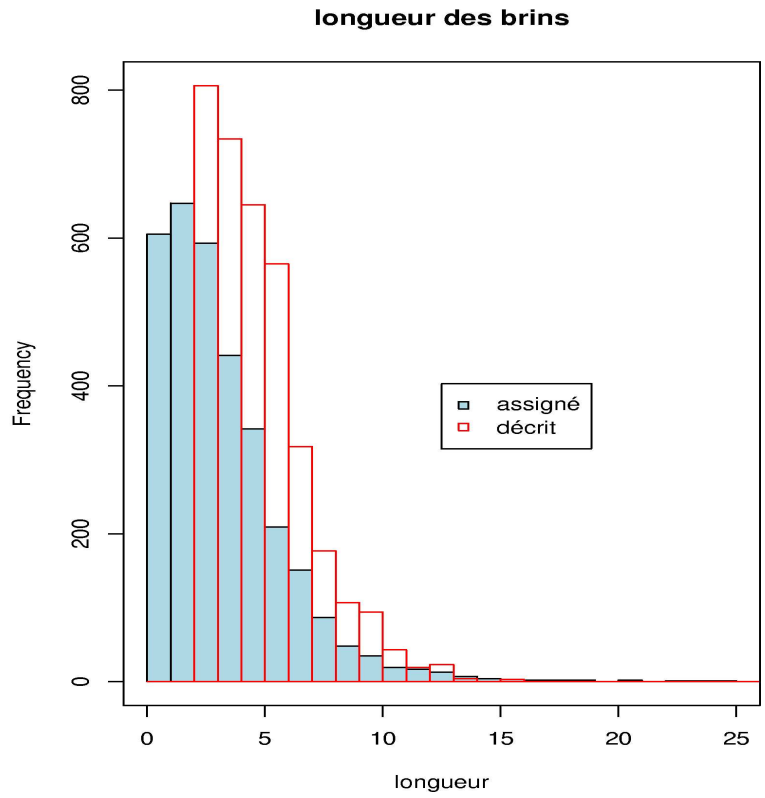
- Modèle M1M1
  - 3*3 transitions + 3*20*20 émissions

$$\begin{array}{ccccccccc} S_1 & \rightarrow & S_2 & \rightarrow & S_3 & \rightarrow & S_4 & \rightarrow \ldots \rightarrow & S_n \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ X_1 & \rightarrow & X_2 & \rightarrow & X_3 & \rightarrow & X_4 & \rightarrow \ldots \rightarrow & X_n \end{array}$$

  - $\rightarrow$ Pas d'améliorations
- Modèle M1M2
  - 3*20*20*20 émissions
  - $\rightarrow$ Sur-apprentissage : 59% sur un set indépendant / 68% sur le set d'apprentissage

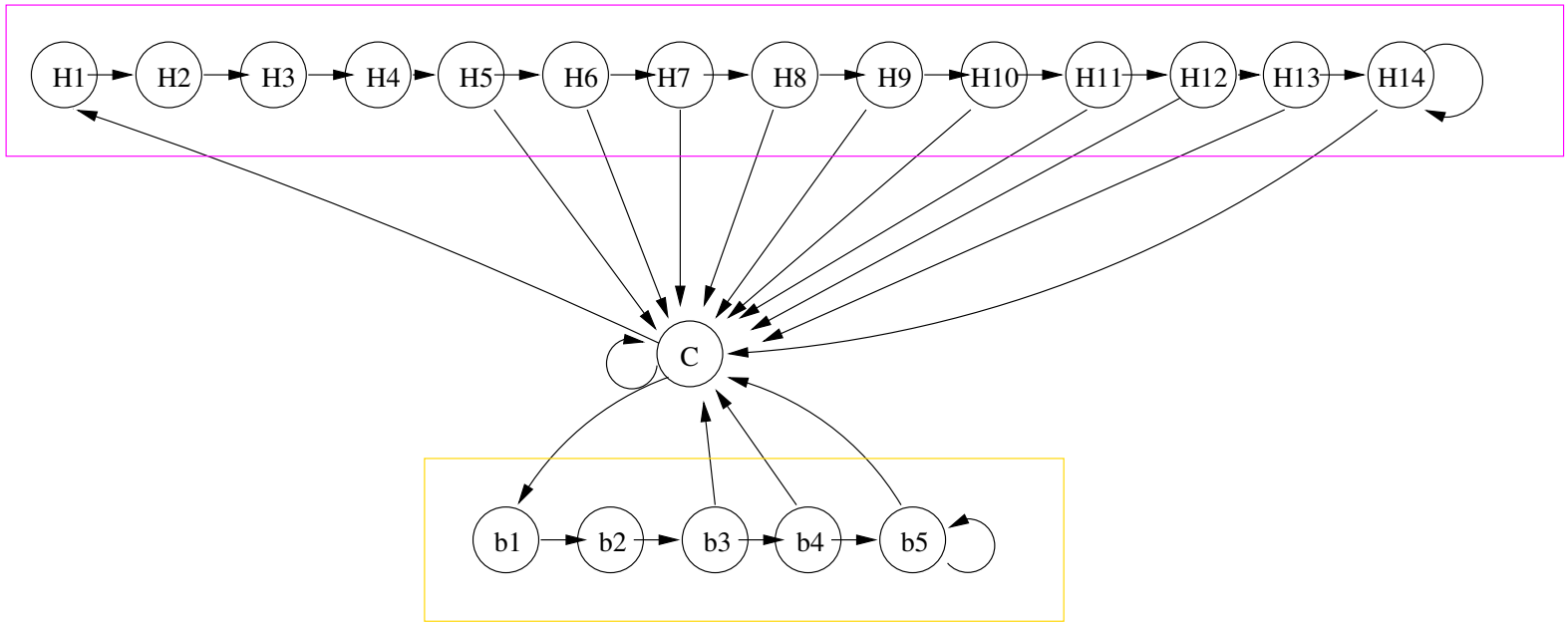
Longueur des hélices et des brins



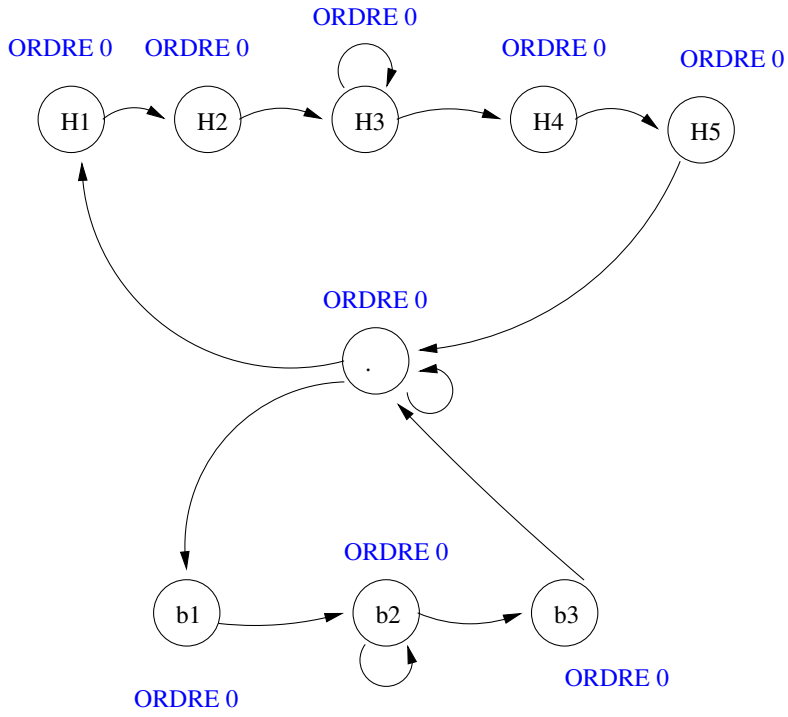


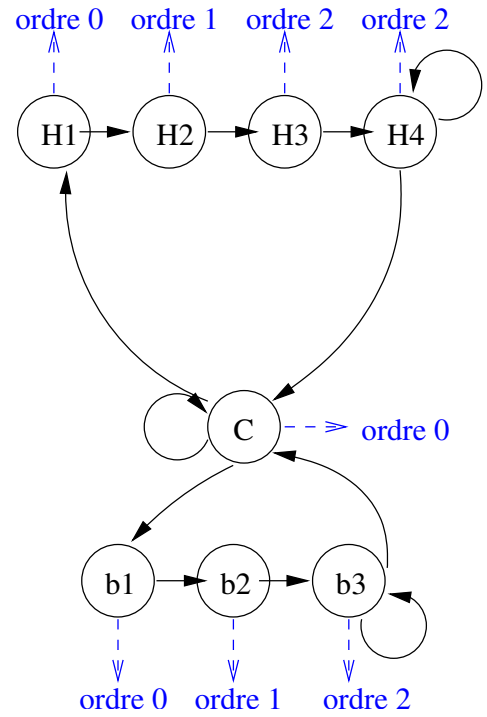
structures secondaires réelles

structures prédites par le modèle M1M0 a 3 états


H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
C
b1
b2
b3
b4
b5

Modèles avec contraintes de longueur

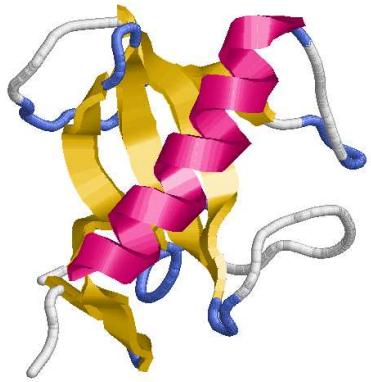


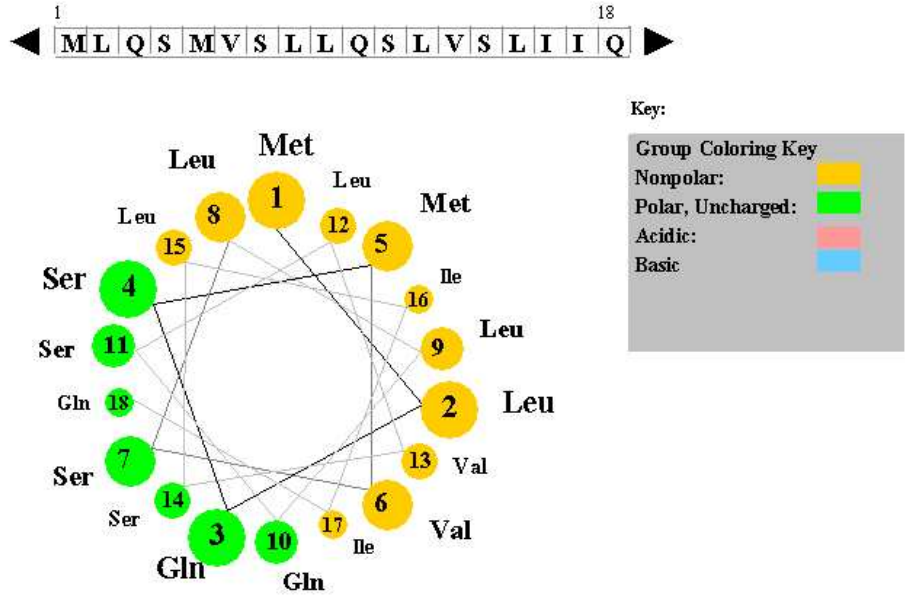


→ Pas d'amélioration. On prédit beaucoup moins souvent de l'hélice

→ il faut incorporer plus de connaissances sur la structure des protéines

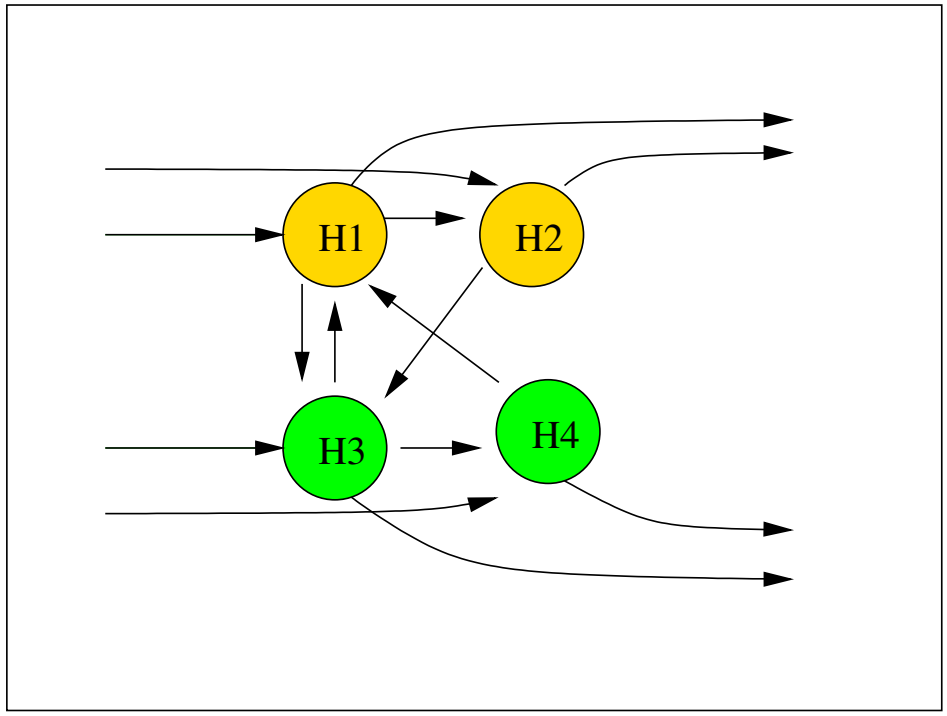
Hélices amphiphiles





Examen rapide des structures : 30 a 70% possedent une partie amphiphile

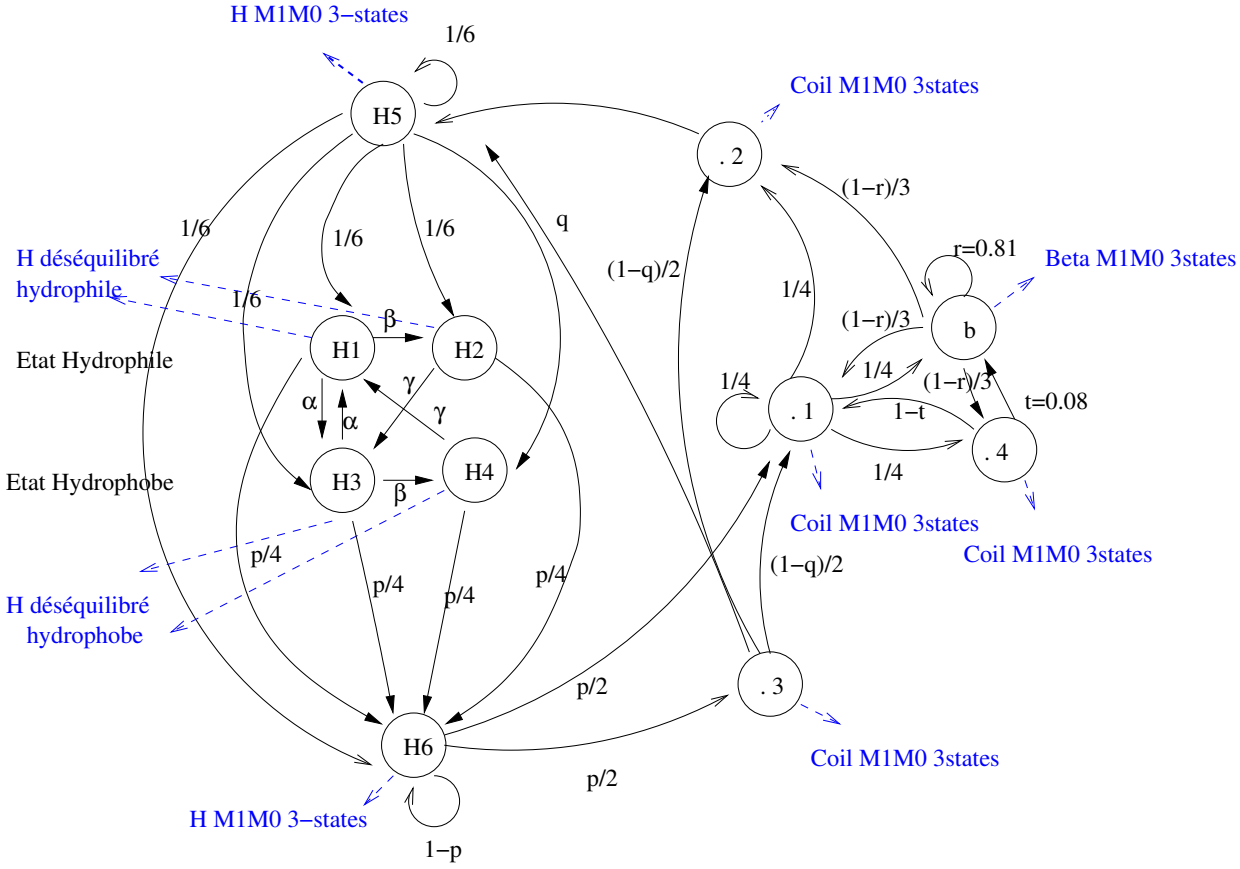
Hélices amphiphiles



- Un État Cache = un état structural
- Séquence observée = Séquence en acides aminés + structure secondaire

Pb : si on l'inclut tel quel on prédit moins d'hélice

Dernier modèle en date



q et p = d'apres le modele 3 etats

alpha=(1−P/4)/5
beta=(1−p/4)*4/5
gamma=1−p/4

q=0.0610914
p=0.0837484 p/4=0.0209371

alpha=0.1958126
beta= 0.7877817
gamma=0.9790629

Dernier Modèle en date

- Estimation des paramètres par EM avec SHOW avec la séquence d'acides aminés et la structure secondaire
- Prédiction F/B avec la séquence en acides aminés seule
- Performances : 64 % de bonne prediction

A suivre...

- Structure du modèle
- Mémoire d'ordre variable-MTD
- Conditionner le modèle par des informations externes (ex : scores d'appariement des brins)
- Utiliser l'information des protéines homologues