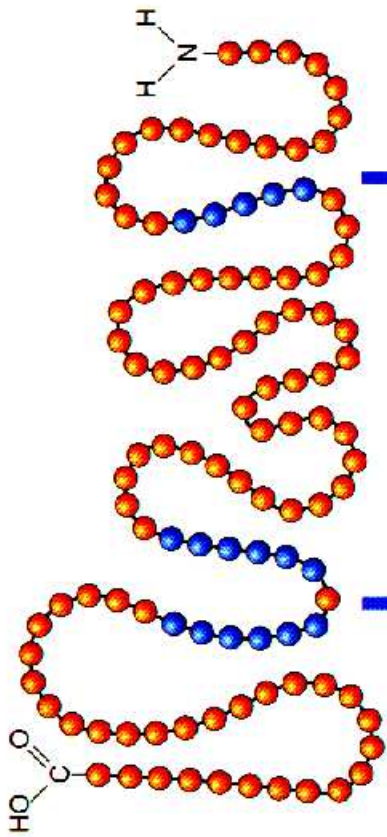


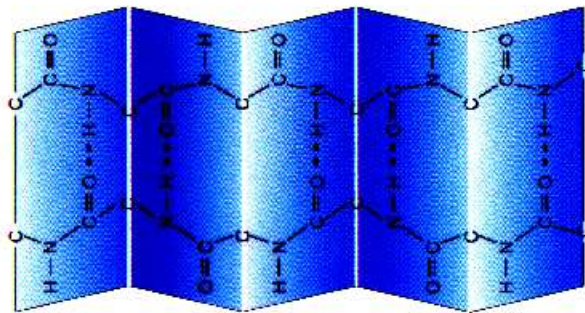
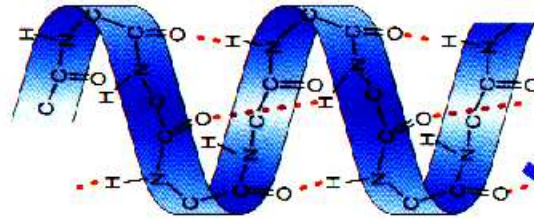
Prédiction de la Structure des Protéines

Hélène TOUZET

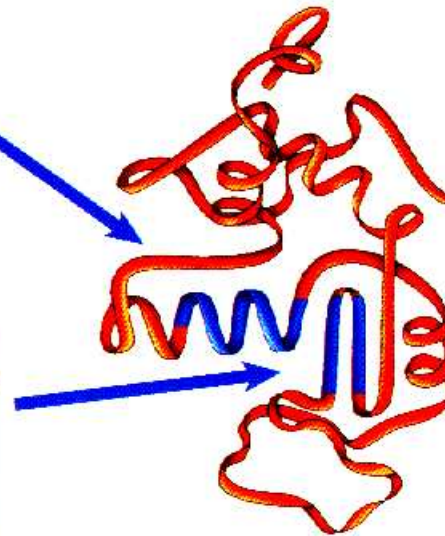
`touzet@lifl.fr`



Structure primaire



Structure secondaire



Structure tertiaire



Structure quaternaire

Exemple : la structure secondaire de la transthyretine (homme)

GPTGTGESKCPMLVKVLDVAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTTEEQFVEGIYKVEI

EEEEEEE EE EEEEEEE EEE EEEE EE EEEEEEE

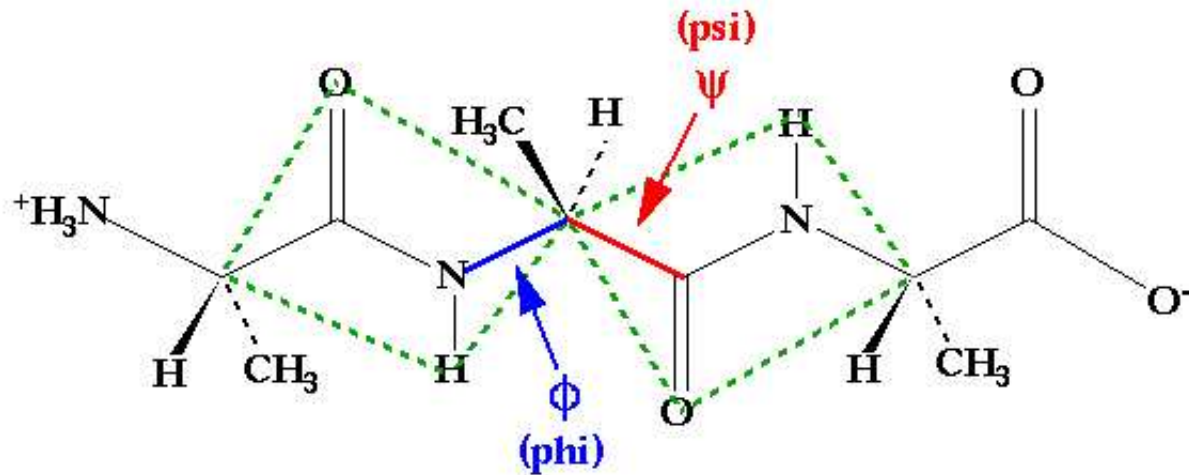
DTKSYWKALGISPFHEHAEEVFTANDSGPRRYTIAALLSPYSYSTTAVVTN

HHHHHH EEEEEEE EEEEEEEEE EEEEEEEEE

E : feuillet β , H : hélice α

Diagramme de Ramachandran: de la 2D à la 3D

▷ graphe des angles de torsion ψ (liaison $C\alpha - C$) et ϕ (liaison $N-C\alpha$).



▷ amplitudes observées

hélice α droite	$-60^\circ < \psi < +30^\circ$ $-120^\circ < \phi < -30^\circ$
feuillet β	$+90^\circ < \psi < +180^\circ$ $-180^\circ < \phi < -60^\circ$
hélice α gauche	$0^\circ < \psi < +60^\circ$ $+45^\circ < \phi < +90^\circ$

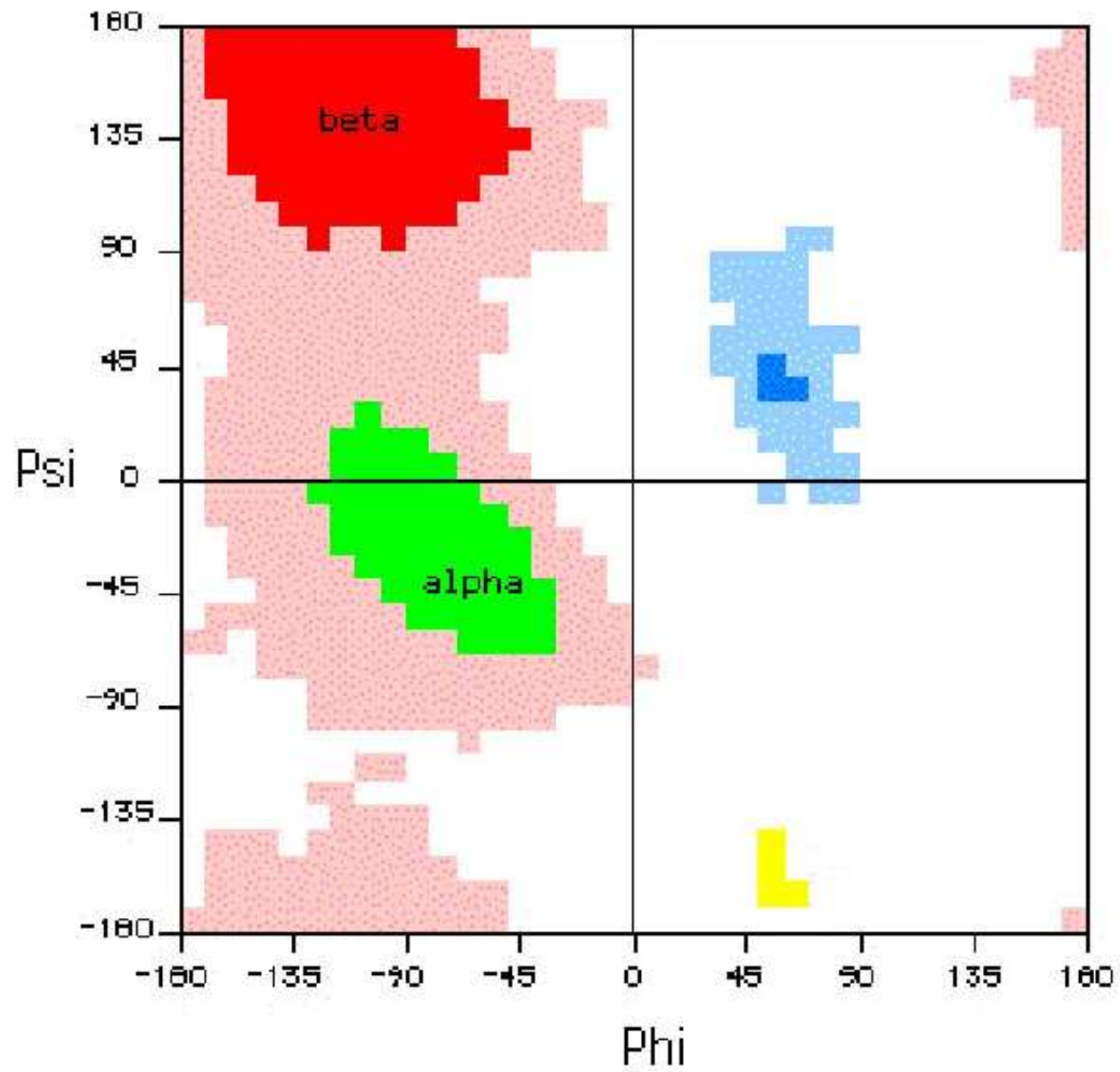


diagramme de Ramachandran

Algorithmes de prédiction de structures secondaires

- ▷ Approche statistique : *Méthode de Chou Fasman*
- ▷ Par apprentissage : *Réseaux de neurones*
- ▷ Par homologie : *Repliement inverse*
- ▷ Par optimisation : *Algorithmes génétiques*

NB : l'approche thermodynamique n'est pas fructueuse

Méthode de Chou-Fasman (1974)

Exploiter la répartition non uniforme des acides aminés dans les structures d'hélice α ou de feuillet β .

- ▷ Pour chaque acide aminé sont déterminés empiriquement des coefficients reflétant la propension de l'acide aminé pour une hélice α , un feuillet β ...
- ▷ $P(h)$ propension à figurer dans une hélice α , $P(f)$ propension à figurer dans un feuillet β

	P(h)	P(f)	
Alanine	142	83	
Arginine	98	93	
Aspartic Acid	101	54	...
Asparagine	67	89	
Cysteine	7	119	
Glutamic Acid	151	037	
Glutamine	111	11	
Glycine	57	75	
Histidine	10	87	
Isoleucine	108	16	
Leucine	121	13	
Lysine	114	74	
Methionine	145	105	
Phenylalanine	113	138	
Proline	57	55	
Serine	77	75	
Threonine	83	119	
Tryptophan	108	137	
Tyrosine	69	147	
Valine	106	17	

Règles pour une hélice α

- ▷ Identifier les régions de 6 résidus consécutifs pour lesquelles 2/3 des résidus satisfont $P(h) > 100$
- ▷ étendre ces régions dans les deux directions, jusqu'à atteindre 4 résidus consécutifs de $P(h)$ moyen < 100
- ▷ les régions pour lesquelles $P(h) > P(f)$ sont déclarées hélice α .

Règles pour un feuillet β

- ▷ Identifier les régions de 5 résidus consécutifs pour lesquelles trois résidus au moins satisfont $P(f) > 100$
- ▷ étendre ces régions
- ▷ les régions pour lesquelles $P(f) > 105$ et $P(f) > P(h)$ sont déclarées feuillets β .

Résultat : **60%** de correction

Réseaux neuronaux

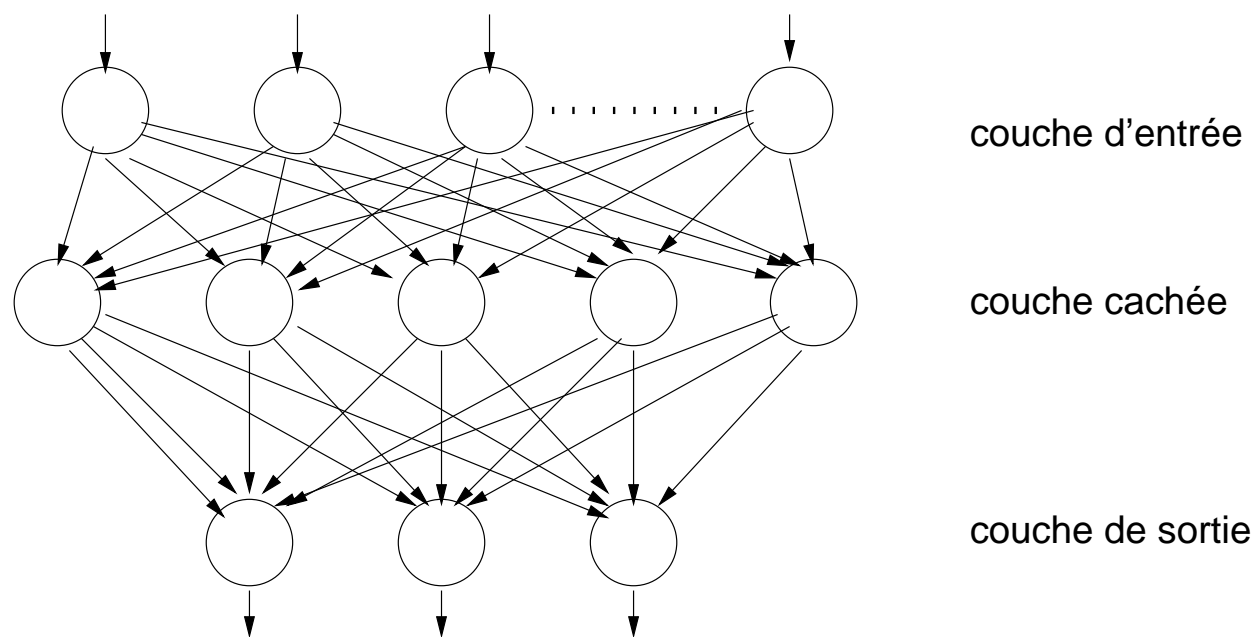
(Qian & Sejnowski - 1988)

I S L S A D Q I **S** T V Q A S F D K



? ? *hélice / feuillet / boucle* ? ?

Réseau **feed-forward**, totalement connecté



- ▷ Couche d'entrée : 17 nœuds (ou 21 nœuds)

Taille de la fenêtre centrée autour de l'acide aminé

- ▷ 1 Couche cachée : 5 nœuds
- ▷ Couche de sortie : 3 nœuds (*hélice / feuillet / boucle*)

Codage des informations

Format unaire

- ▷ Un acide aminé : vecteur de taille 21

20 acides aminé + un symbole de terminaison

Alanine 100000000000000000000
Cystéine 010000000000000000000

...

- ▷ Entrée : taille 17 * 21
- ▷ Sortie : *winner-takes-all*

Apprentissage

- ▷ Apprentissage supervisé (back propagation, scaled conjugate gradients)
- ▷ Cible : 100, 010, 001
- ▷ Une centaine de séquences protéiques non apparentées
- ▷ soit environ 15 000 paires d'apprentissage

Évolution du modèle - PHD

(Rost & Sander - 1994)

Utiliser des familles de protéines partageant la même structure

- ▷ Un alignement multiple par famille
- ▷ Paire d'apprentissage
 - motif : extrait d'un alignement multiple
 - cible : *hélice / feuillet / boucle*
- ▷ Codage d'une position de l'alignement :
ligne de la matrice des fréquences

70 % de correction de la prédiction

Repliement inverse de protéine

threading, inverse folding

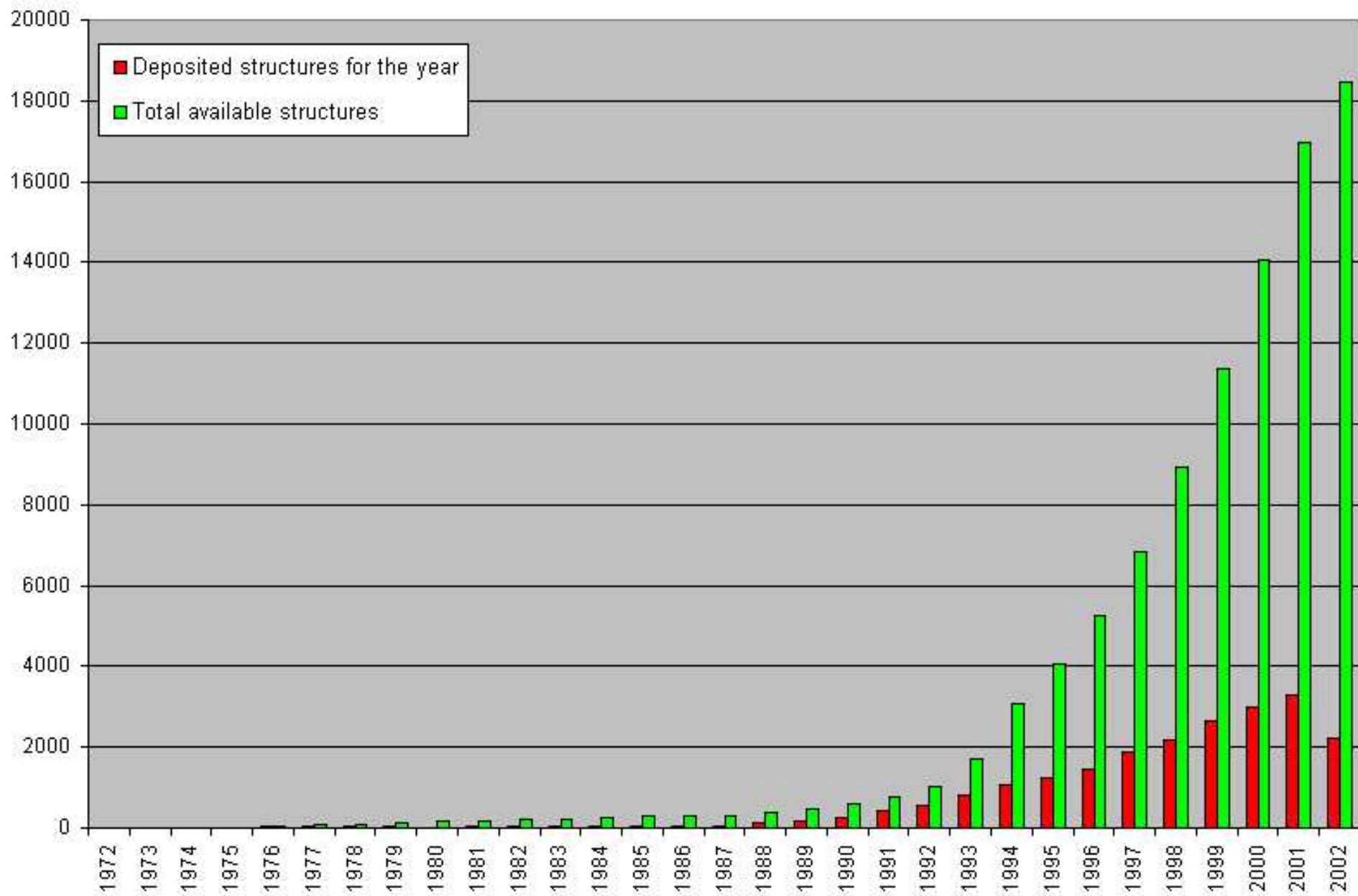
- ▷ Dans **PDB**, toutes les protéines avec plus de 25% d'identité partagent la même structure.
- ▷ Faire correspondre la séquence avec la structure de protéines homologues connues.

Données

- un modèle structural
- une fonction de score
- une séquence dont on veut déterminer la structure (2D ou 3D)

Résultat

Meilleur alignement entre le modèle et la séquence ("enfilage")



Évolution du nombre de structures répertoriées dans PDB

Modèle structural

- ▷ éléments de la structure secondaire
- ▷ coordonnées dans l'espace (PDB)

Fonction de score

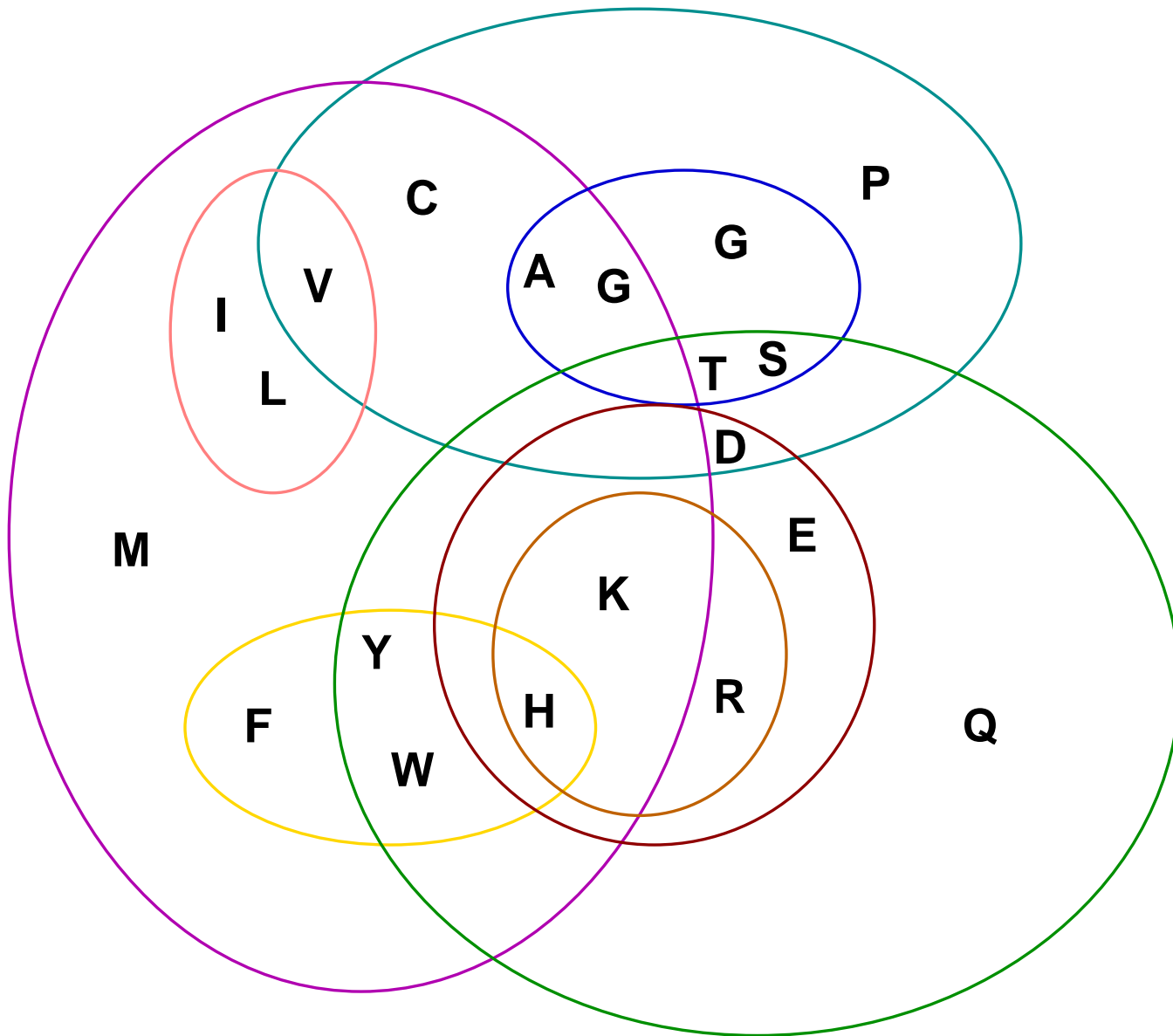
- ▷ Propriétés physico-chimiques

Bowie *et al.* - 1991

- ▷ Potentiel de force moyenne

Sippl *et al.* - 1995

Correction : proche de 100% quand les séquences sont identiques à plus de 25%, moins de 40% quand les séquences sont des "homologues distants"



aliphatiques

petits

très petits

positifs

non polarisés

aromatiques

chargés

polarisés

Principales classes d'acides aminés

Les algorithmes génétiques

Problème d'optimisation :

Trouver une solution qui minimise un coût, une distance, maximise un score, etc., sous contraintes

- ▷ Exploration probabiliste de l'espace des solutions
- ▷ Création d'un ensemble initial de solutions potentielles arbitraires
- ▷ Évolution de l'ensemble par itérations successives

solution potentielle = **individu**
ensemble d'individus = **population**
nouvelle population = **génération**

Exemples

▷ une fonction $f : \mathbb{N} \rightarrow \mathbb{N}$ à maximiser

individu = entier

▷ voyageur de commerce

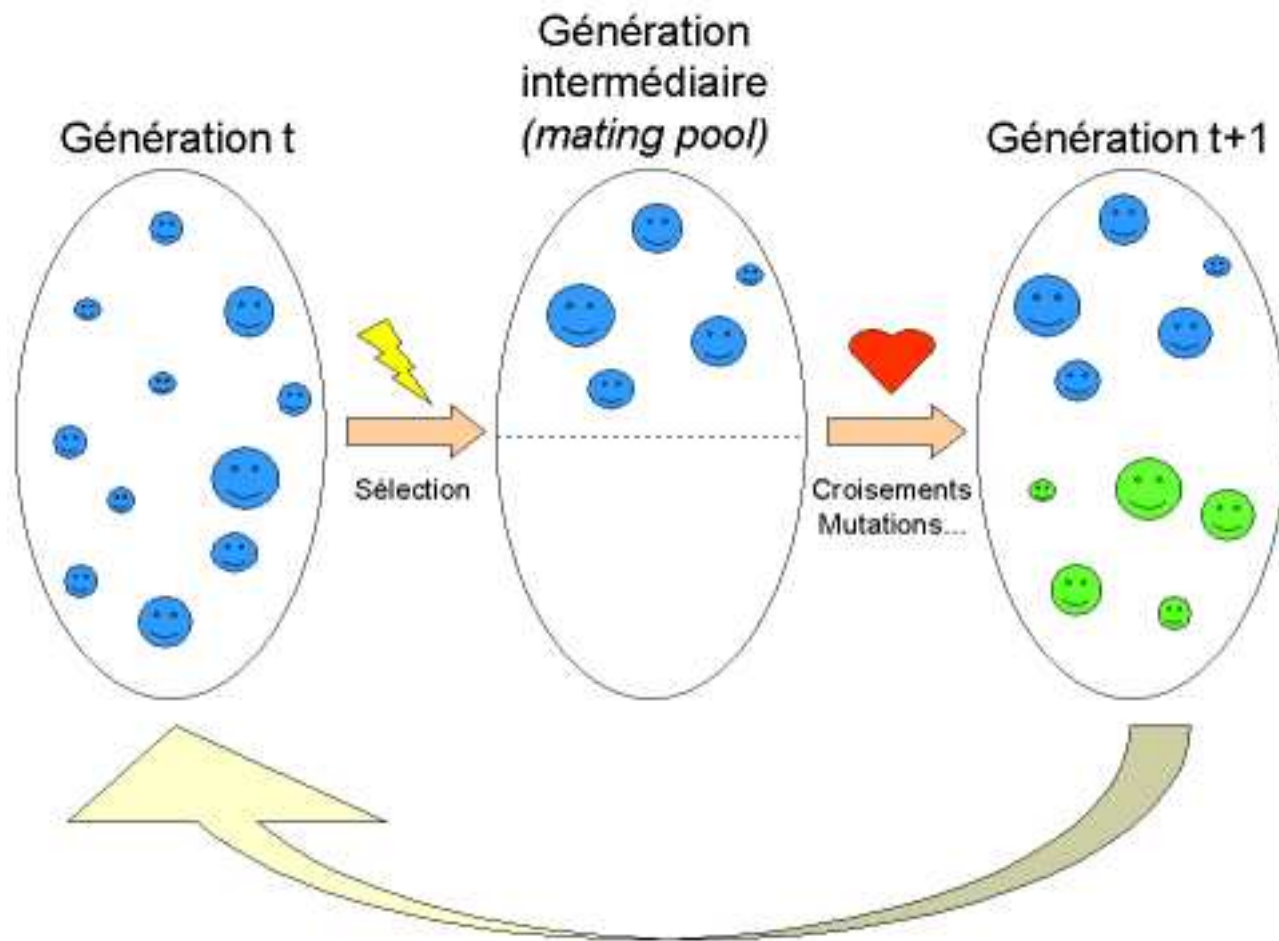
individu = tournée dans le graphe

▷ sac à dos

individu = sélection d'objets

▷ ordonnancement

individu = affectation



Le passage d'une génération à l'autre se fait par sélection des individus les plus adaptés, puis par reproduction.

Les algorithmes génétiques: l'orthodoxie

(Holland, De Jong - 1975)

- ▷ le problème à traiter est vu comme une boîte noire
- ▷ codage des données par des chaînes binaires
- ▷ transitions probabilistes

Construction de la génération intermédiaire

Étape 1 : *évaluation des individus*

▷ fonction d'objectif d'un individu i : f_i

Dépend du problème : valeur de la fonction à maximiser, coût d'une tournée, poids d'une affectation, temps global d'une exécution, etc.

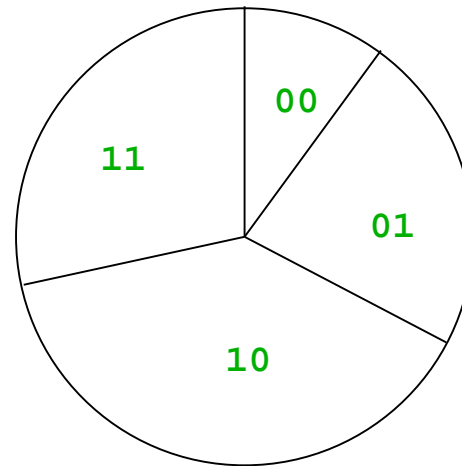
▷ fonction d'adéquation de l'individu (fitness)

$$\frac{f_i}{\text{moyenne des } f}$$

Étape 2: *sélection des individus de manière probabiliste en fonction de leur adéquation*

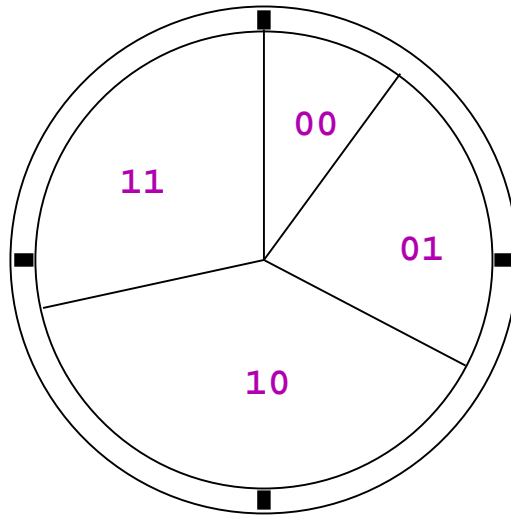
individus	0 0	0 1	1 0	1 1
adéquations	0.3	0.8	1.7	1.2

Sélection simple



- ▷ On tourne la roue autant de fois que l'on veut d'individus.
- ▷ À chaque tirage, un individu est sélectionné avec une probabilité proportionnelle à son adéquation.

Sélection avec reste probabiliste



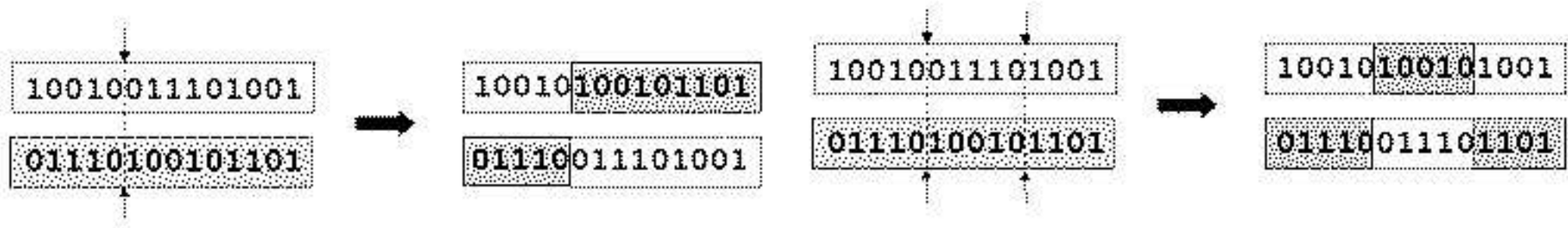
- ▷ Les bords de la roue sont marqués à intervalles réguliers d'autant de repères que l'on souhaite d'individus.
- ▷ Un unique tirage.

Création de la génération t+1

Application d'**opérateurs** probabilistes

▷ croisement: *2 parents* → *2 enfants*

Exemple: croisement en 1 point, en 2 points



▷ mutation: *1 individu* → *1 individu*

Exemple: modification aléatoire d'un bit

Conditions d'arrêt : convergence

Mise en pratique

- ▷ “kit heuristique” universel

Peut s’appliquer à tous les problèmes d’optimisation, y compris les problèmes NP-complets, sans analyse particulière du problème

- ▷ pas de garantie

Aucune prédiction sur le comportement de l’algorithme ou sur la qualité du résultat n’est possible.

- ▷ souvent très lent

Il ne faut pas avoir recours à un algorithme génétique si une méthode exacte efficace existe

- ▷ l’approche brute est souvent peu fructueuse

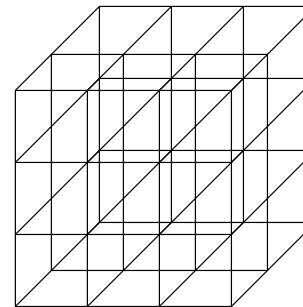
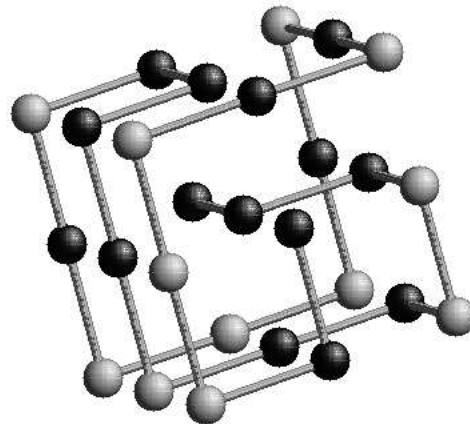
Ajouter des opérateurs spécifiques, changer la représentation des données

⇒ **Algorithmes évolutionnaires, hybrides**

Application à la prédiction de structures

Le Modèle HP

- ▷ classification des acides aminés: **H** *hydrophobe*, **P** *hydrophile (polarisé)*
- ▷ repliement: interactions hydrophobes
- ▷ discrétisation de l'espace

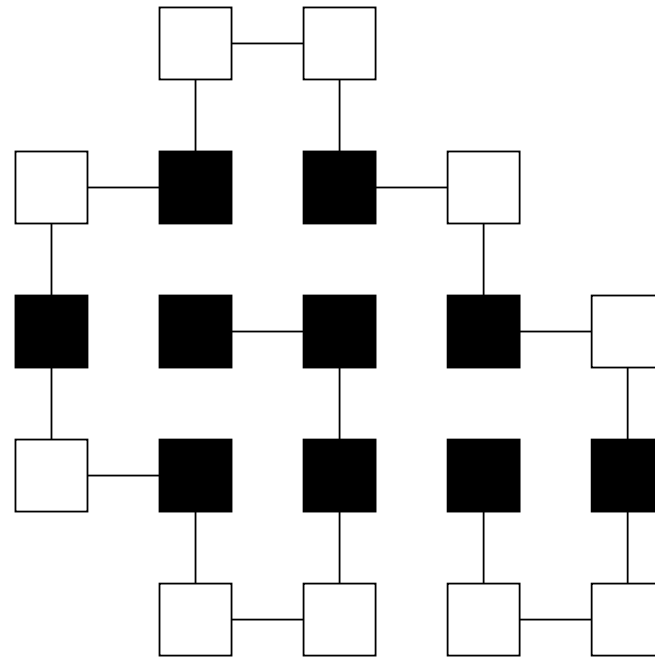


6 directions de déplacement

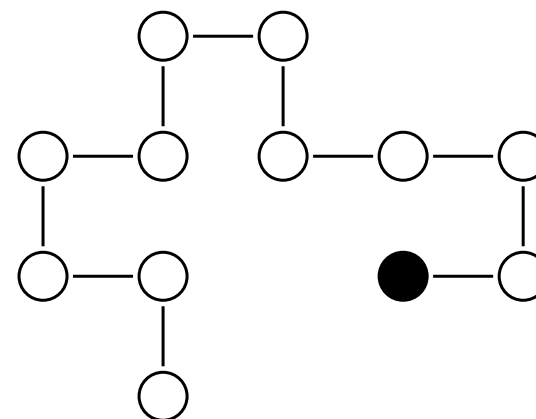
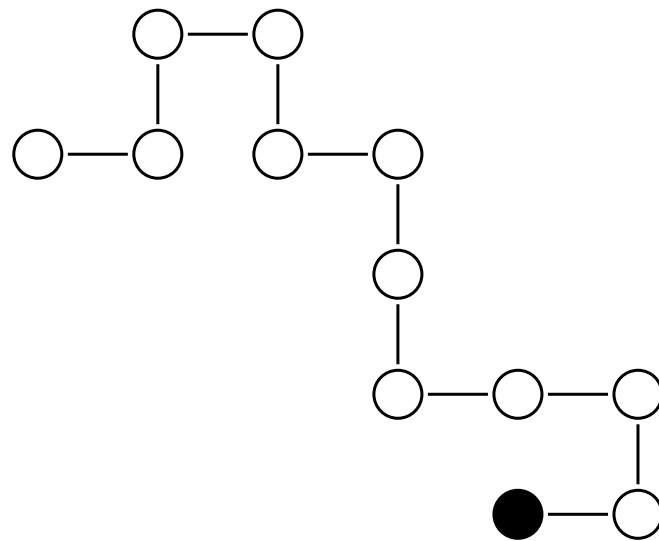
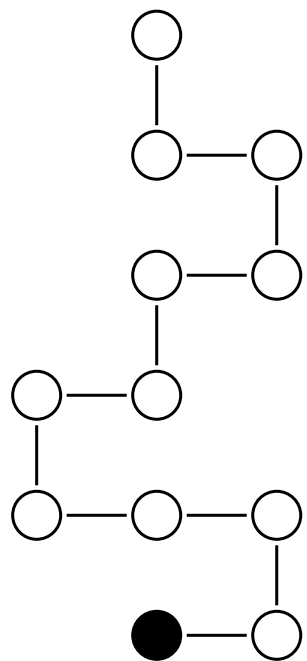
Énoncé du problème

- ▷ **Donnée:** une séquence sur $\{H, P\}^*$
- ▷ **Question:** quel est le chemin sans croisement qui maximise le nombre de contacts entre acides aminés hydrophobes ?

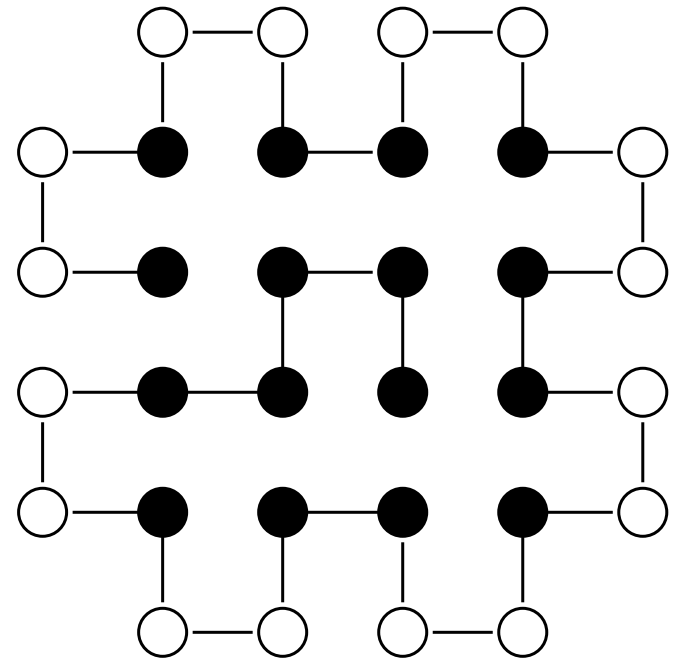
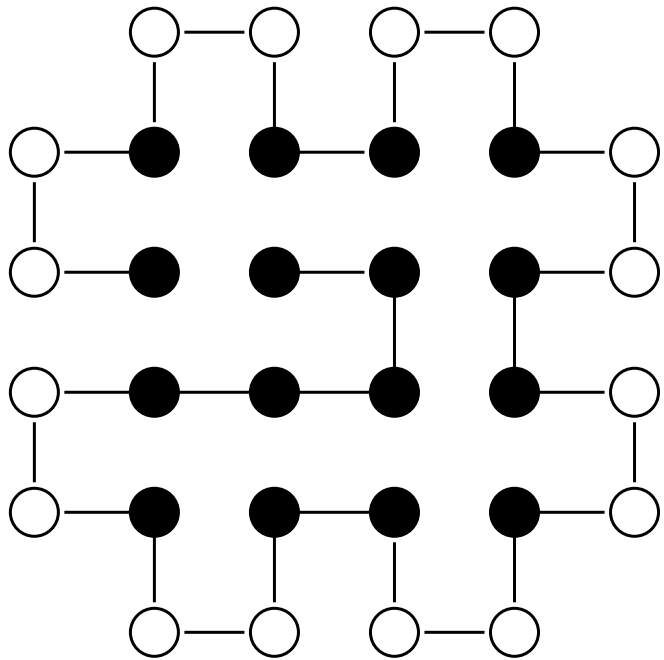
NP-complet



HHHPHPHPHPHPHPHPH



Notations absolues et relatives



Combien de mutations ?