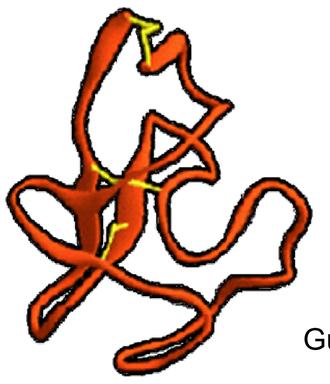
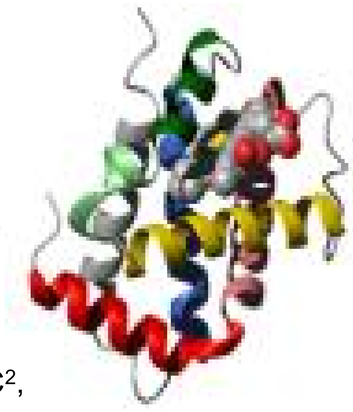
Projet GenoTo3D

Apprentissage automatique appliqué à la prédiction de la structure tertiaire des protéines





Guermeur Y[1], Benabdelsem K[2], Bréhélin L[3], Capponi C[4], Coste F[6], Darcy Y[1], Deléage G[2], Denis F[4], Gascuel O[3], Geourjon C[2], Gibrat JF[5], Jacquemin I[6], Magnan C[4], Marin A[5], Martin J[5], Monfrini E[1], Nicolas J[6], Ralaivola L[4], Taly JF[5]

1 : LORIA-Nancy, 2 : IBCP-Lyon, 3 : LIRMM-Montpellier, 4 : LIF-Marseille, 5 : MIG-Jouy en Josas, 6 : IRISA-Rennes

Prédiction de la structure et apprentissage automatique
Problème d'apprentissage sur de grandes quantités de données

Contexte biologique Exploitation fonctionnelle des informations provenant des grands programmes de séquençage des génomes : passe par la connaissance de la structure 3D des protéines. C'est cette structure 3D qui est responsable de la fonction biologique.

1. Arrivée massive de séquences protéiques (croissance exponentielle des bases)

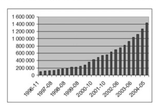


FIG. 1 – Croissance de la banque internationale TREMBL de 1996 à 2004

2. Détermination expérimentale de la structure 3D : tâche très lourde... lorsqu'elle est réalisable
⟹ Nécessité de passer d'une approche biochimique à une approche prédictive

Problème central en biologie permettant d'aborder
l'essentiel des grandes questions ouvertes en traitement de données séquentielles

Différents niveaux d'organisation structurale des protéines

- **Séquence ou structure primaire** **(1 536 117 séquences connues)**

`MEEKLKKAKIIFVVGGPGSGKGTQCEKIVQKYGYTHLSTC...`

- **Structure secondaire**

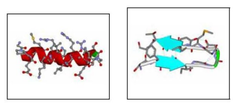


FIG. 2 – Elements structuraux périodiques : hélice α (à gauche) et brins β (à droite)

- **Structure tertiaire** **(27 112 structures 3D connues)**

Projet GENOTO3D de l'ACI "Masses de Données"

- **Apprentissage automatique appliqué à la prédiction de la structure tertiaire des protéines**

"L'objet de ce projet est, dans le contexte de la prédiction de la structure tertiaire des protéines, de mettre en évidence des problèmes de prédiction sur des séquences génériques et difficiles, et de proposer des méthodes susceptibles de faire progresser l'état de l'art dans le domaine."

- **Participants**

1. Projet MODBIO du LORIA (Nancy)
2. Laboratoire de Bioinformatique et RMN Structurales (LBRS) de l'IBCP (Lyon)
3. Equipe "Bases de Données et Apprentissage Automatique" du LIF (Marseille)
4. Projet Symbiose de l'IRISA (Rennes)
5. Equipe "Méthodes et algorithmes pour la bioinformatique" (MAB) du LIRMM (Montpellier)
6. Unité Mathématique, Informatique et Génome (MIG) de l'INRA, centre de Jouy-en-Josas

Approche modulaire et hiérarchique

- **Un ensemble de sous-problèmes et de reformulations du problème**
  - Prédiction des ponts disulfures et des ponts salins : IBCP, IRISA, LIF, LORIA
  - Prédiction de la structure secondaire (feuillets β...) : LIRMM, LORIA, MIG
  - Prédiction par homologie ou analogie et reconnaissance des cœurs structuraux : IBCP
  - Prédiction par *threading* : IRISA, MIG
  - Prédiction *ab initio* (*de novo*) : MIG

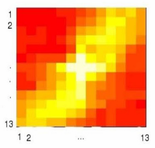
Prédiction des ponts disulfures

Modélisation probabiliste des appariements d'acides aminés autours des cystéines

- Extraction de la PDB des cartes de contact de chaque pont disulfure

L'extraction est paramétrée par la taille K des fenêtres et la distance minale d entre AA sous laquelle on considère qu'il y a un contact

- Modèle probabiliste des contacts :



$\theta(k,l)$ est la probabilité d'un contact à la position (k,l)

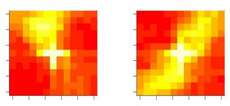
- blanc: la probabilité du contact à cette position est élevée ;
- rouge: la probabilité est faible.

On peut calculer la probabilité d'une carte sous ce modèle :

$$P(F_i,F_j|\Theta)=\Pi_{k=1}^{K}\Pi_{l=1}^{K}\theta(k,l)P(F_{ik},F_{jl}|contact)+(1-\theta(k,l))P(F_{ik},F_{jl}|non\,contact)$$

- Les cartes sont utilisées pour apprendre :
  - une distribution de paires d'AA pour les contacts et les non contacts : $P(a,b \mid contact)$ et $P(a,b \mid non\ contact)$;
  - un modèle de mélange de C classes (grâce à un algorithme de type EM).

Exemple de modèle de mélange à 2 classes :

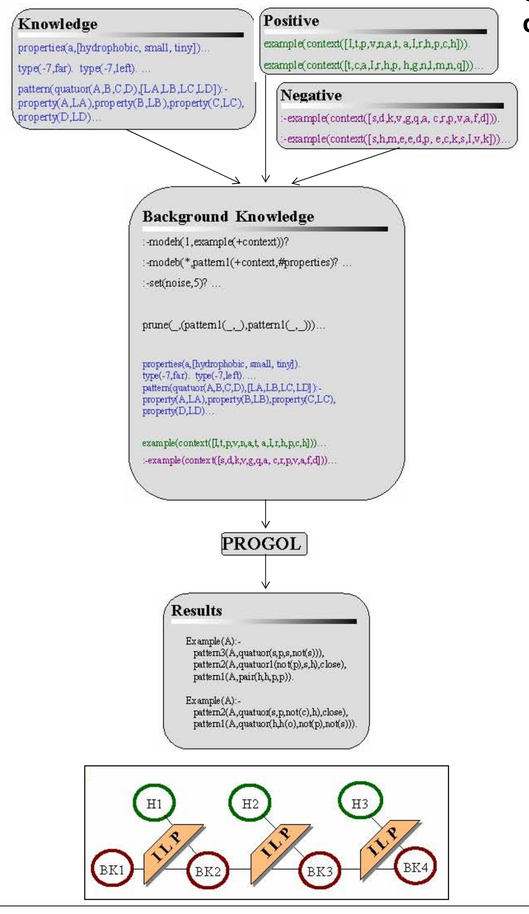


- La probabilité d'observer deux fenêtres d'AA en contact est alors :

$$P(F_i,F_j)=\Sigma_{c=1}^{C}P(c)P(F_i,F_j|\theta_c)$$

- Etant donnée une protéine à 2N cystéines, on peut utiliser cette modélisation pour prédire les N ponts disulfures les plus probables.

Prédiction des ponts disulfures par PLI



Protocole pour détecter la présence d'information locale pour la prédiction de contacts entre acides aminés

Motivations

Les ponts disulfures sont des liaisons covalentes entre cystéines oxydées qui forment un élément du repliement d'une protéine. Mais il n'est pas clair si les ponts participent au repliement ou en sont la cause. En particulier :

les voisinages des cystéines oxydées sur les séquences primaires contiennent-ils une information sur les ponts qu'elles forment ?

Problématique : poser un cadre formel pour détecter la présence d'une information au voisinage de deux acides aminés permettant de (contribuer à) prédire s'ils sont en contact ou non.

Modélisation
(déclinée dans le cadre de la prédiction des ponts disulfures.)

Protéines : (Σ^*, P) où $\Sigma = \{\text{acides aminés}\}$ et P dist. de probabilité.
$\mathcal{P} \subset \Sigma^*$: protéines contenant un nombre pair de cystéines oxydées.

Pour $w, w' \in \Omega_r = \{w \in \Sigma^{2r+1} | w[r+1] = \text{cystéine}\}$ (contextes locaux centrés sur une cystéine), on définit

$P(B(w,w')|CL(w,w',l))$: probabilité que w et w' forment un pont sachant que ce sont des contextes locaux distincts de cystéines oxydées d'une protéine en contenant $2l$.

Absence d'information locale
⇔
$\forall w, w' \in \Omega_r, \forall l, P(B(w,w')|CL(w,w',l)) = 1/(2l-1).$

Inenvisageable d'estimer directement $P(B(w,w')|CL(w,w',l))$. **Idée** : supposer l'existence d'une fonction d'affinité $g : \Omega_r \to Y$ ($|Y|$ petit) tq

$$g(w_1,w_2) = g(w_1',w_2') \Rightarrow$$
$$\forall l, P(B(w_1,w_2)|CL(w_1,w_2,l)) = P(B(w_1',w_2')|CL(w_1',w_2',l)).$$

Cas le plus simple : $Y = \{0,1\}$. Les paires de fenêtres se répartissent en deux classes, correspondant à deux niveaux d'affinités : *faible* (0) et *fort* (1). On devra alors pouvoir observer

$$P(B(w,w')|g(w,w')=1,l) > P(B(w,w')|g(w,w')=0,l).$$

Affinité, ponts et bruit de classification

Sous l'hypothèse de l'existence d'une telle fonction d'affinité g,

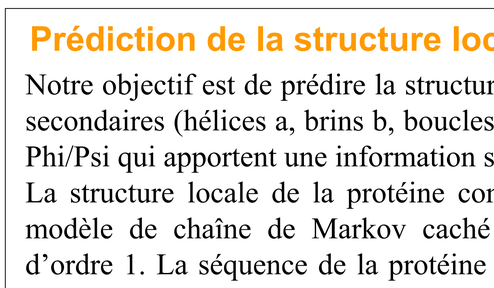
$$P(B(w,w')|CL(w,w',l)) = P(B(w,w')|g(w,w'),l) = \begin{cases} \alpha_1^l \text{ si } g(w,w')=1 \\ \alpha_0^l \text{ si } g(w,w')=0. \end{cases}$$

L'observation d'un pont (B) correspond à $g=1$ *avec un bruit de classification différencié* de paramètres : $\eta^+ = 1-\alpha_1^l$ et $\eta^- = \alpha_0^l$.

Prédiction de la structure locale des protéines

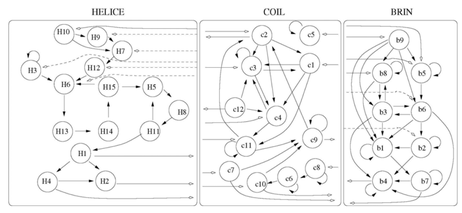
Notre objectif est de prédire la structure locale en terme de structure secondaires (hélices a, brins b, boucles) et de zones d'angles dièdres Phi/Psi qui apportent une information sur la structure des boucles.
La structure locale de la protéine constitue le processus caché du modèle de chaîne de Markov caché (HMM) avec une mémoire d'ordre 1. La séquence de la protéine est le processus observé. Les acides aminés sont émis indépendamment conditionnellement à la structure locale.
Chaque classe structurale est modélisée par un ensemble d'états cachés. Deux stratégies ont été mises en place pour construire ces modèles :

➢Modélisation des structures secondaires à partir d'a priori biologiques (modélisation des hélices amphiphiles) et d'une étude des mots exceptionnels dans les brins.



HMM à 21 états cachés. Les états carrés privilégient les résidus polaires, les états héxagonaux les résidus hydrophobes. Le taux de bonne prédiction obtenu est de **65%.**

➢-Choix du nombre d'états cachés selon des critères de performance et le critère BIC

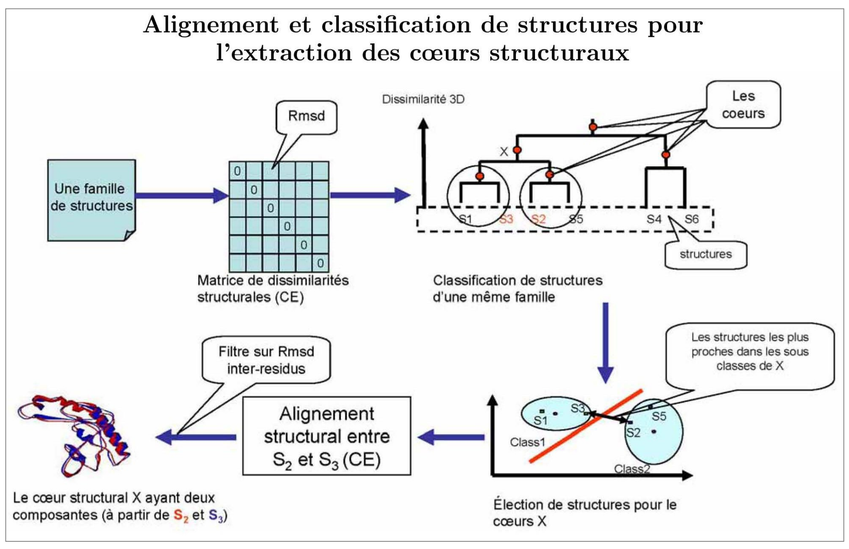


HMM à 36 états cachés. Seules les transitions les plus probables sont indiquées. Le taux de bonne prédiction obtenue est de **68%.**

La prédiction est améliorée grâce aux séquences homologues. Les contributions des séquences homologues sont combinées avec les pondération de Henikoff. Le taux de bonne prédiction atteint ainsi **76%**. La même méthodologie appliquées aux zones d'angles permet d'atteindre un taux de prédiction de **78%**.

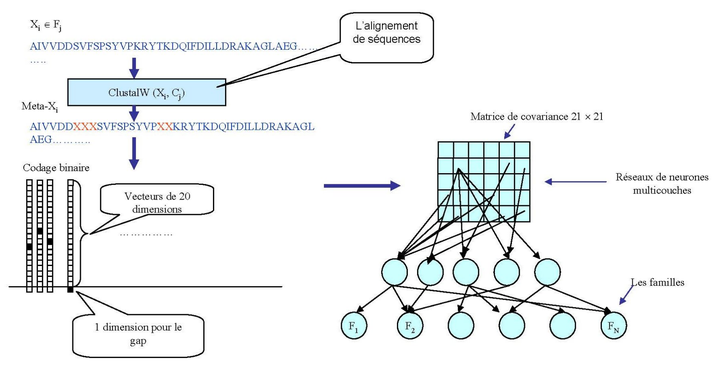
Apprentissage à partir des cœurs structuraux

Alignement et classification de structures pour l'extraction des cœurs structuraux



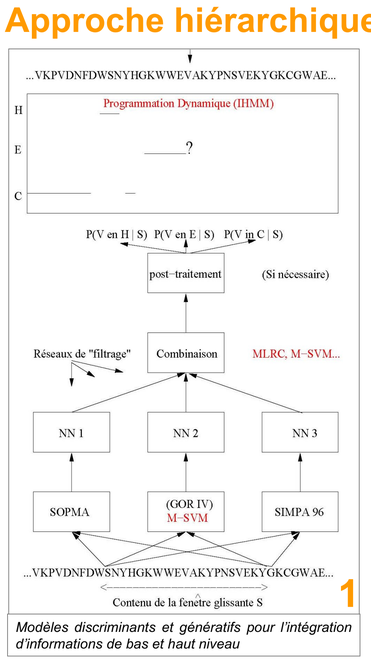
Codage matriciel et modélisation connexionniste

1. Apprentissage : alignement de chaque séquence avec le cœur structural extrait de *sa* famille
2. Test : alignement de chaque séquence avec *chacun* des cœurs structuraux déterminés
3. Modélisation de chaque séquence alignée par une matrice de covariance présentée à un PMC



GenoTo3D

Approche hiérarchique de la prédiction de la structure secondaire



Modèles discriminants et génératifs pour l'intégration d'informations de bas et haut niveau

Machines à noyau

Noyau défini positif

Un *noyau défini positif* (n.d.p.) sur l'ensemble $\mathcal{X}$ est une fonction $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ symétrique satisfaisant

$$\forall n \in \mathbb{N}^*,\ \forall (x_i)_{1\le i\le n} \in \mathcal{X}^n,\ \forall (a_i)_{1\le i\le n} \in \mathbb{R}^n,\ \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \kappa(x_i,x_j) \ge 0.$$

Machine à noyau

Une machine à noyau est une machine utilisant un n.d.p. comme *mesure de similarité* entre exemples. La classe de fonctions qu'elle calcule appartient à un espace de Hilbert à noyau reproduisant. **2**

Espace de Hilbert à noyau reproduisant

Soit $\mathcal{X}$ un espace quelconque et $(\mathcal{H}, \langle .,. \rangle_{\mathcal{H}})$ un espace de Hilbert de fonctions sur $\mathcal{X}$ ($\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$).

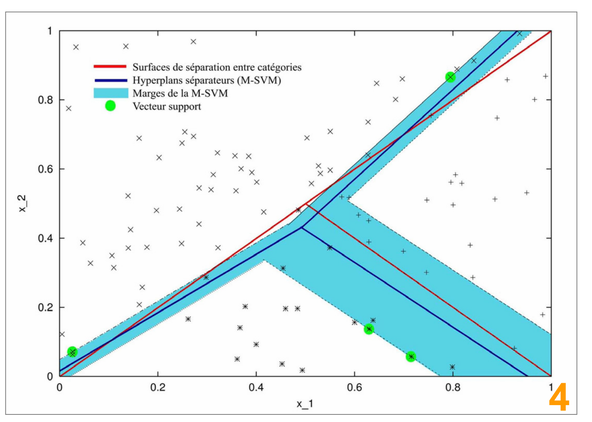
Noyau reproduisant (Aronszajn, 1950)

Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ est appelée un *noyau reproduisant* (noté n.r.) si et seulement si :

1. $\mathcal{H}$ contient toutes les fonctions de la forme
$$\forall x \in \mathcal{X},\quad \kappa_x : t \mapsto \kappa(x,t)$$
2. $\forall x \in \mathcal{X},\ \forall h \in \mathcal{H},\ h(x) = \langle h, \kappa_x \rangle_{\mathcal{H}}$

Espace de Hilbert à noyau reproduisant

Si un n.r. existe, $\mathcal{H}$ est appelé un *espace de Hilbert à noyau reproduisant* (e.h.n.r. ou RKHS) **3**



Nouveau noyau pour le traitement des séquences protéiques

Expression analytique (structure primaire seule)

(x,x') : vecteurs des descriptions de deux contenus de fenêtres (segments) à comparer

$$\kappa_{\mu,\theta,D}(x,x') = \exp\left(-\mu \sum_{i=-n}^{n} \theta_i^2 \|x_i - x_i'\|^2\right)$$

avec $D = (d_{jk})_{1\le j,k\le 22}$, $\|x_i - x_i'\|^2 = d_{jj} + d_{kk} - 2d_{jk}$ pour $x_i = a_j$ et $x_i' = a_k$

Extension pour le traitement des alignements multiples

Immédiate : x est remplacé par $\tilde{x}$ tel que $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$

$$\langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta'_{ik} a_k \rangle = \sum_{j=1}^{22}\sum_{k=1}^{22} \theta_{ij}\theta'_{ik} \langle a_j, a_k \rangle$$

5

Références

- Ingrid Jacquemin. Découverte de motifs relationnels en bioinformatique : application à la prédiction des ponts disulfures. Thèse de doctorat de l'Université Rennes 1, 2005.
- Ingrid Jacquemin & Jacques Nicolas. Modélisation de cystéines oxydées à l'aide de la programmation logique inductive. **JOBIM**, Lyon, juillet 2005, 331-340.
- Juliette Martin, Jean-François Gibrat & François Rodolphe. HMM for local protein structure. **ASMDA**, Brest, mai 2005, 180-187.
- Juliette Martin, Jean-François Gibrat & François Rodolphe. How to choose the optimal hidden Markov model for protein secondary structure prediction. **IEEE Intelligent Systems**, Special issue on Data Mining for Bioinformatics, accepté, à paraître en novembre/décembre 2005
- Khalid Benabdeslem, Christophe Geourjon, Yann Guermeur & Nicolas Sapay. Apprentissage automatique, application à la prédiction de la structure secondaire et tertiaire des protéines. Communication sur invitation présentée dans la session thématique : Bioinformatique II, **ASTI**, Clermont-Ferrand, octobre 2005.
- Khalid Benabdeslem, Gilbert Deléage & Christophe Geourjon. A neural network system based on structural alignment and clustering for proteins fold recognition. **ECCB**, Madrid, septembre 2005, 85-88.
- Khalid Benabdeslem, Gilbert Deléage & Christophe Geourjon. Cores extraction based neural network model for proteins fold recognition. **JOBIM**, Lyon, juillet 2005, 341-347.
- Yann Guermeur, A Lifchitz & Régis Vert. A kernel for protein secondary structure prediction. In **«Kernel Methods in Computational Biology»**, Editors : B. Schölkopf, K. Tsuda & Jean-Philippe Vert, The MIT Press, 2004, 193-206.