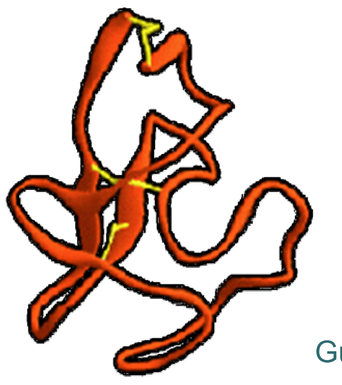


Projet GENOTO3D

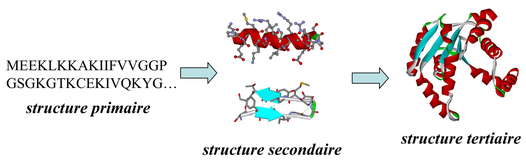
Apprentissage automatique appliqué à la prédiction de la structure tertiaire des protéines

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], Thomarat F[1]

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

Contexte biologique :

Les protéines, macromolécules indispensables à la vie, assurent des fonctions très diverses. Cette fonction biologique est étroitement liée à la structure 3D de la protéine (*cf.* figure ci-dessous). La complexité de la détermination expérimentale de la structure 3D et la croissance exponentielle des données de séquences disponibles nécessitent la mise au point de méthodes de prédiction de la structure 3D. Ce problème central en biologie permet d'aborder l'essentiel des grandes questions ouvertes en traitement de données séquentielles.



Les différents niveaux d'organisation structurale des protéines : des régions de la *structure primaire* (ou séquence) de la protéine s'organisent en un ensemble d'éléments structuraux périodiques constituant la *structure secondaire* qui s'agencent dans l'espace pour former la *structure 3D* (*tridimensionnelle* ou *tertiaire*) de la protéine.

L'**apprentissage automatique** doit permettre d'exploiter les **données de structures** déjà existantes pour développer des **méthodes de prédiction** efficaces.

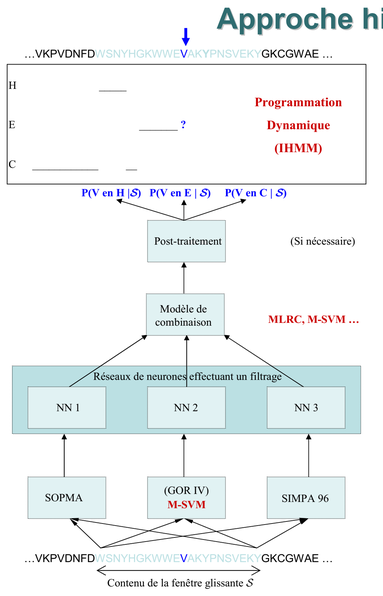
Le projet :

• **objectif** : dans le contexte de la prédiction de la structure tertiaire des protéines, mettre en évidence des problèmes de prédiction sur des séquences génériques et difficiles, et proposer des méthodes susceptibles de faire progresser l'état de l'art dans le domaine.

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

Approche hiérarchique de la prédiction de la structure secondaire

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



Machines à noyau

Noyau défini positif

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

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

Espace de Hilbert à noyau reproduisant

Soit $\mathcal{X}$ un espace quelconque et $(\mathrm{H}, \langle .,. \rangle_{\mathrm{H}})$ un espace de Hilbert de fonctions sur $\mathcal{X}(\mathrm{H} \subset \mathbb{R}^{\mathcal{X}})$. Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est appelée un *noyau reproduisant* (Aronszajn 1950) pour H si et seulement si :
- H contient toutes les fonctions de la forme : $\forall x \in \mathcal{X}, \kappa_x : t \mapsto \kappa(x,t)$
- $\forall x \in \mathcal{X}, \forall h \in \mathrm{H}, h(x) = \langle h, \kappa_x \rangle_{\mathrm{H}}$ (propriété de reproduisance)

Si un noyau reproduisant existe, H est appelé un *espace de Hilbert à noyau reproduisant* (RKHS).

Une machine à noyau est un modèle d'apprentissage utilisant un n.d.p. κ comme *mesure de similarité* entre exemples. La classe de fonctions $\mathcal{H}$ qu'elle réalise est construite autour du RKHS induit par κ.

Architecture de la SVM multi-classe (M-SVM)

$\mathcal{H}$: famille des fonctions $h = (h_k)_{1 \le k \le Q}$ dont les fonctions composantes sont de la forme :

$$h_k(x) = \sum_{i=1}^{l_k} \beta_{ik} \kappa(x_{ik}, x) + b_k$$

où $s^{(k)} = \{x_{ik} : 1 \le i \le l_k\} \subset \mathcal{X}$ et les limites de ces fonctions lorsque $s^{(k)}$ devient dense dans $\mathcal{X}$. On note : $\mathcal{H} = \left(\left(\mathrm{H}, \langle .,. \rangle_{\mathrm{H}_k} \right) + \{1\} \right)^Q$ et $\bar{\mathcal{H}} = \left(\mathrm{H}, \langle .,. \rangle_{\mathrm{H}_k} \right)^Q$

La M-SVM est solution d'un problème de programmation convexe (quadratique) :

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\bar{h}\|_{\mathrm{H}}^2 + C \sum_{i=1}^{m} l_{M-SVM}(y_i, h(x_i)) \right\}$$

où $s_m = \{(x_i, y_i) : 1 \le i \le m\}$ est l'ensemble d'apprentissage et l_{M-SVM} la fonction de perte.

Dimension de Natarajan à marge γ

Soit $\mathcal{G}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\mathbb{R}^Q$. Pour $\gamma > 0$, un sous-ensemble $s_{\mathcal{X}^m} = \{x_i : 1 \le i \le m\}$ de $\mathcal{X}$ est dit être γ-*N-pulvérisé* (N-pulvérisé avec une marge γ) par $\Delta\mathcal{G}$ s'il existe un ensemble

$$I(s_{\mathcal{X}^m}) = \{(i_1(x_i), i_2(x_i)) : 1 \le i \le m\}$$

de m couples d'indices distincts dans $\{1,...,Q\}$ et un vecteur $v_b = (b_i)$ dans $\mathbb{R}^m$ tels que, pour tout vecteur binaire $v_y = (y_i) \in \{-1,1\}^m$, il existe une fonction g_y dans $\mathcal{G}$ satisfaisant

$$\forall i \in \{1,...,m\}, \begin{cases} si\ y_i = 1, & \Delta g_{y, i_1(x_i)}(x_i) - b_i \ge \gamma \\ si\ y_i = -1, & \Delta g_{y, i_2(x_i)}(x_i) + b_i \ge \gamma \end{cases}$$

La *dimension de Natarajan à marge* γ de la famille $\Delta\mathcal{G}$, $\mathrm{N} - \dim(\Delta\mathcal{G}, \gamma)$, est le cardinal maximal d'un sous-ensemble de $\mathcal{X}$ γ-N-pulvérisé par $\Delta\mathcal{G}$, si ce cardinal est fini, ou l'infini dans le cas contraire.

Lemme de Sauer-Shelah généralisé

Soit $\mathcal{G}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}^Q$. Pour toute valeur de γ dans $]0,1]$ et toute valeur entière de m satisfaisant $2m \ge \mathrm{N} - \dim(\Delta_\gamma \mathcal{G}, \gamma/24)$, on dispose de la borne suivante :

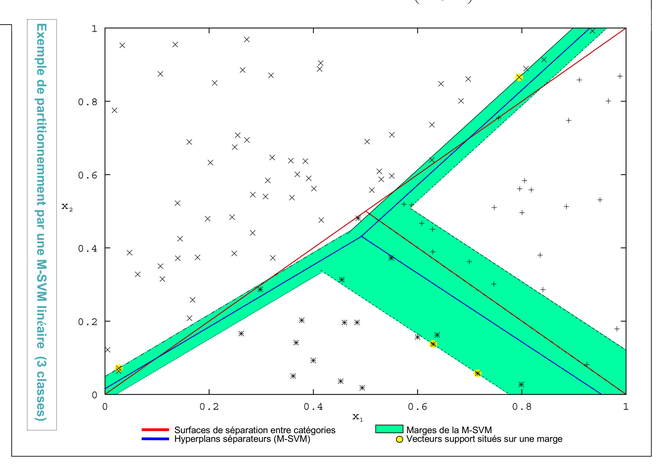
$$N_{m,\infty}^{(\gamma)}(\gamma/4, \Delta_\gamma^* \mathcal{G}, 2m) < 2\left(288\, m\, Q^2 (Q-1)\right)^{\left\lfloor d \log_2 (23 e m Q(Q-1)/d) \right\rfloor}$$

où $d = \mathrm{N} - \dim(\Delta_\gamma \mathcal{G}, \gamma/24)$.

Théorème

Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une SVM à Q-catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ de $E_{\Phi(\mathcal{X})}$, que les vecteurs w_k définissant les hyperplans séparateurs vérifient $1/2 \max_{1 \le k < l \le Q} \|w_k - w_l\|_{H_k} \le \Lambda_w$ et que $b = 0_Q$. Alors, pour toute valeur strictement positive d'ε, on dispose de la borne suivante :

$$\mathrm{N} - \dim(\Delta\bar{\mathcal{H}}, \varepsilon) \le C_Q^2 \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\varepsilon} \right)^2$$



Prédiction de la structure locale des protéines

Application des modèles de chaînes de Markov cachées au problème de la prédiction de structure secondaire

Les modèles de chaînes de Markov cachées ont été utilisés pour construire des modèles génératifs et prédictifs de la structure secondaire des protéines.
Le problème de la sélection de modèle a été abordé en utilisant des critères statistiques (Bayesian Information Criterion) et de performance (taux de bonne prédiction), aboutissant à un modèle à 36 états cachés (Martin et al, IEEE Intelligent System, 2005).

Un nouvel outil d'analyse

Le modèle optimal, en plus de fournir des résultats satisfaisants en prédiction, est aussi un nouvel outil d'analyse de l'architecture interne des protéines. A la différence des réseaux de neurones, il reflète l'information utile pour la prédiction.

(a): Représentation graphique de modèle optimal à 36 état cachés.

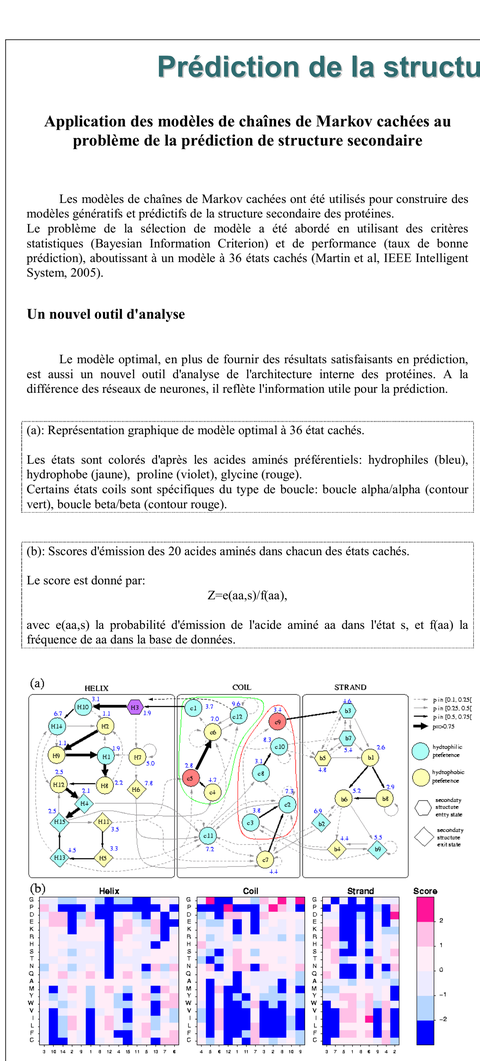
Les états sont colorés d'après les acides aminés préférentiels: hydrophiles (bleu), hydrophobe (jaune), proline (violet), glycine (rouge).
Certains états coils sont spécifiques du type de boucle: boucle alpha/alpha (contour vert), boucle beta/beta (contour rouge).

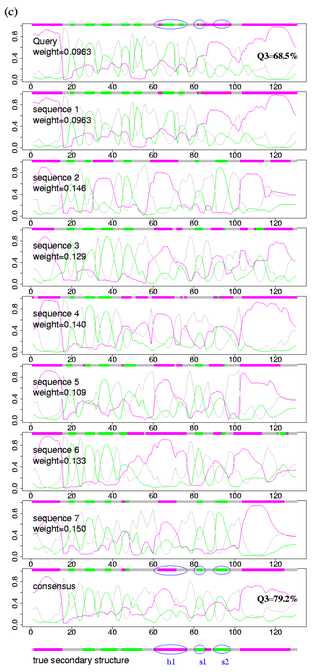
(b): Sscores d'émission des 20 acides aminés dans chacun des états cachés.

Le score est donné par:

Z=e(aa,s)/f(aa),

avec e(aa,s) la probabilité d'émission de l'acide aminé aa dans l'état s, et f(aa) la fréquence de aa dans la base de données.





Une intégration simple des séquences homologues

Notre modèle, lorsqu'il est utilisé avec un alignement multiple, fournit un score Q3 de 76%. Pour cela, les prédictions des séquences d'une famille donnée sont réunies en une prédiction consensus. La prédiction repose sur les probabilités postérieures du HMM et la pondération des séquences.

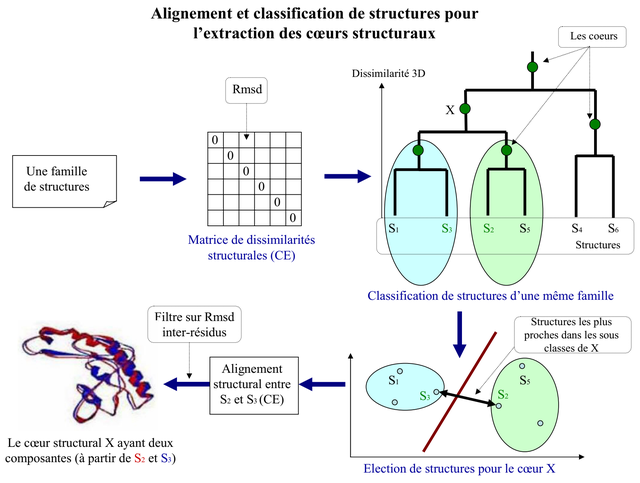
(c) Exemple de prédiction de la protéine 1jyoa.

Chaque graphique représente les probabilités postérieures de l'état hélice (magenta), brin (vert) et coil (gris). La structure secondaire prédite est indiquée sous forme de segment en haut de chaque graphique, la vraie structure secondaire figure dans la partie inférieure de la figure.

La prédiction sur la séquence seule (QUERY) est correcte à 68.5%. La prise en compte des séquences homologues (sequence 1 à séquence 7), munies de pondérations (weight) résulte en une prédiction consensus correcte à 79.2%.

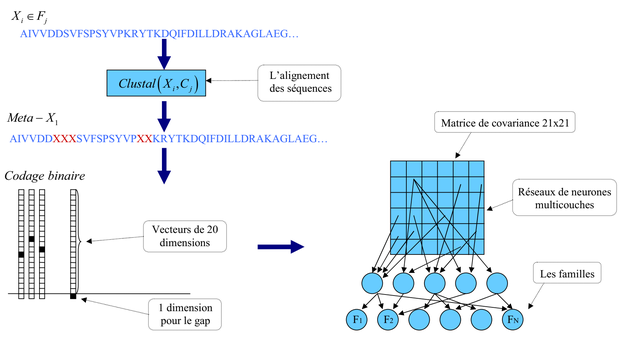
Apprentissage à partir des cœurs structuraux

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



Discrimination multi-classes : Modèle proposé (CAH+SVM= DSVM)

Décomposition Hiérarchique (*DSVM*)

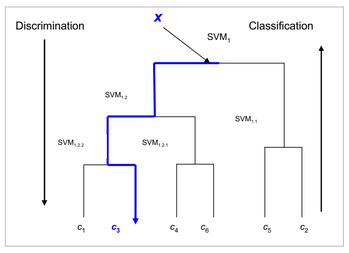
Soit une base d'apprentissage $x_1, x_2, \ldots, x_n$ classifiée en k classes , $c_1, c_2, \ldots, c_k$

→ **Principe**

1) Calcul de k centres de gravités des k classes
2) CAH sur les k classes => **Taxonomie**
3) Distribution de (k-1) *SVM* sur les nœuds de la taxonomie

Avantages

- *Un nombre optimal de SVM*
- *Une trace de reconnaissance*
- *Complexité réduite de reconnaissance*



Perspective:

Développement d'un noyau dynamique de la racine vers les feuilles

Prédiction des ponts disulfures

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. On ne sait pas 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 (*AA*) permettant de (contribuer à) prédire s'ils sont en contact ou non.

Modélisation

Protéines : (Σ^*, P) où $\Sigma^* = \{AA\}$ et P est une distribution de probabilités.
$\mathcal{P} \subset \Sigma^*$:protéines contenant un nombre pair de cystéines oxydées.

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

$$P\left(B(w,w') \,|\, CL(w,w',l)\right),$$

la probabilité que w et w' forment un pont sachant que ce sont des contextes locaux distincts de cystéines oxydées d'un protéine en contenant *2l*.

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

Mais il est inenvisageable d'estimer directement $P\left(B(w,w') \,|\, CL(w,w',l)\right)$.

Idée : supposer l'existence d'une fonction d'affinité $g : \Omega_r \to Y(|Y| \, petit)$ tq

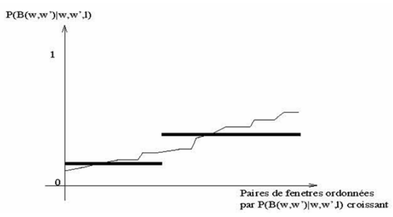
$$g(w_1, w_2) = g(w'_1, w'_2) \Rightarrow$$
$$\forall l, P\left(B(w_1, w_2) \,|\, CL(w_1, w_2, l)\right) = P\left(B(w'_1, w'_2) \,|\, CL(w'_1, w'_2, l)\right)$$

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 pouvoir observer alors :

$$P\left(B(w,w') \,|\, g(w,w') = 1, l\right) > P\left(B(w,w') \,|\, g(w,w') = 0, l\right)$$



Affinité, ponts et bruit de classification

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

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

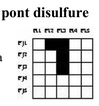
L'observation d'un pont *(B)* correspond à *g=1 avec un bruit de classification différencié* de paramètres :

$$\eta^+ = 1 - \alpha_1^l \text{ et } \eta^- = \alpha_0^l.$$

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 *d* minimale entre AA sous laquelle on considère qu'il y a 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 ;
- Vert : la probabilité est faible.

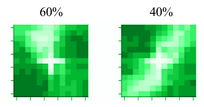
On peut calculer la probabilité d'une carte sous ce modèle :

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

Les cartes sont utilisées pour apprendre :

- Une distribution de paires d'AA pour les contacts et les non contacts : $P(a,b|contact)$ et $P(a,b|non\ contact)$;
- Un modèle de mélange de *C* classes (grâce à un algorithme de type *EM*).

Modèle de mélange à 2 classes : 60% 40%

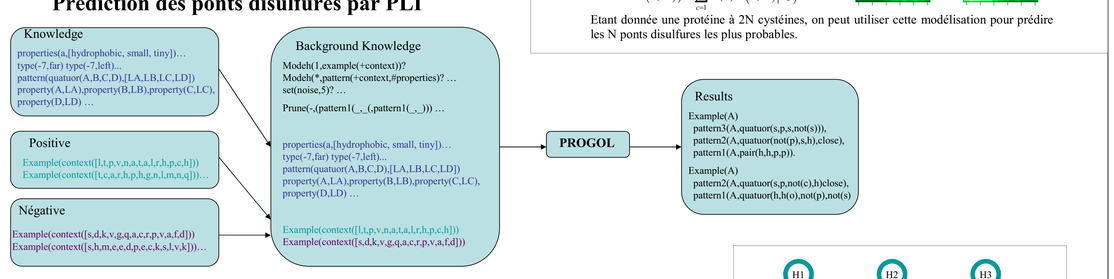


La probabilité d'observer deux fenêtres d'AA en contact est alors :

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

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



L'apprentissage effectué ici repose sur la programmation logique inductive (PLI) pour inférer des motifs caractéristiques autour des cystéines oxydées et expliquant l'appariement de celles-ci. Pour cela de nombreuses contraintes sont fixées dans notre connaissance a priori (BK) sur le format des motifs attendus.
L'idée est d'effectuer plusieurs phases d'apprentissage en tenant compte des résultats obtenus aux phases précédentes. Ceci permet de diminuer considérablement la combinatoire dans les espaces d'hypothèses logiques car à chaque étape notre algorithme construit des règles de plus en plus discriminantes.

Références

- 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.
- Khalid Benabdeslem & Younès Bennani. Dendogram based SVM for multi-class classification. **Journal of Computing and Information Technology** – **CIT 1r4**, 4, 2006, 291-297.
- Khalid Benabdeslem, Gilbert Deléage & Christophe Geourjon. Alignement structural et classification hiérarchique pour l'extraction des coeurs structuraux. Atelier : Extraction et gestion de connaissances appliquées aux données biologiques, **EGC**,Lille, Janvier 2006, 9-17.
- François Denis, Yann Esposito & Amaury Habrard. Learning Rational stochastic languages. Proc. of the 19th Annual Conference on Learning Theory, **LNAI**, 4005, 2006, 274–288.
- Elodie Duprat, Marie-Paule Lefranc & Olivier Gascuel. A simple method to predict protein-binding from aligned sequences – application to MHC superfamily and 2-microglobulin. **Bioinformatics, 2006, 22, 4, 453-459.**
- Yann Guermeur. Large Margin Multi-category Discriminant Models and Scale-sensitive Psi-dimensions. **Rapport de recherche INRIA**, RR-5314, 2004 (révisé en 2006).
- 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.
- Christophe Magnan. Asymmetrical Semi-Supervised Learning and Prediction of Disulfide Connectivity in Proteins. **RIA**, 2006, à paraître.
- Juliette Martin, Jean-François Gibrat & François Rodolphe. Analysis of an optimal hidden Markov model for secondary-structure prediction. Soumis.
- Juliette Martin, Jean-François Gibrat & François Rodolphe. Choosing the optimal hidden Markov model for secondary-structure prediction. **IEEE Intelligent Systems**, 20, 2005, 19-25.
- Liva Ralaivola, François Denis & Christophe Magnan. CN=CPCN. **Proc. of the 23rd Int. Conf. on Machine Learning**, 2006, 721–728.
- Nicolas Sapay, Yann Guermeur & Gilbert Deléage. Prediction of amphipathic in-place membrane anchors in monotopic proteins using a SVM classifier. **BMC Bioinformatics**, 2006, Vol.7, 255.
- Raluca Uricaru, Eric Rivals & Laurent Bréhélin, Hidden Markov models for the detection of motifs repeats in protein sequences, **IPG**, Lyon, novembre-décembre 2006.

INRIA

INRA
Institut National de la Recherche Agronomique

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ministère délégué à la recherche