Apprentissage Automatique Appliqué à la Prédiction de la Structure Tertiaire des Protéines

Projet 2003-96 : GENOTO3D

LORIA - IBCP - LIF - IRISA - LIRMM - MIG

http://www.loria.fr/~guermeur/ACIMD

De la biologie structurale à l'apprentissage automatique

- **Un problème de biologie**
- **Un problème d'apprentissage automatique**
- **Travaux effectués**
- **Perspectives et objectifs**

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)

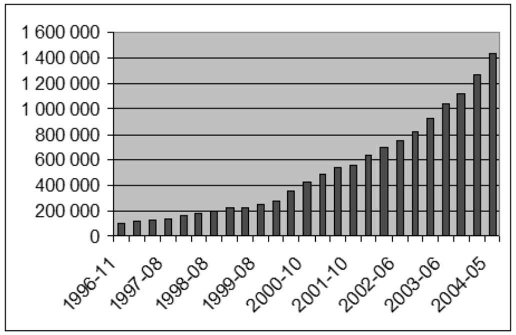


FIG. 1 – Croissance de la banque internationale TREMBL de 1996 à 2004

2. Détermination expérimentale de la structure 3D : tâche très lourde... lorsqu'elle est réalisable

$\Longrightarrow$ 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**

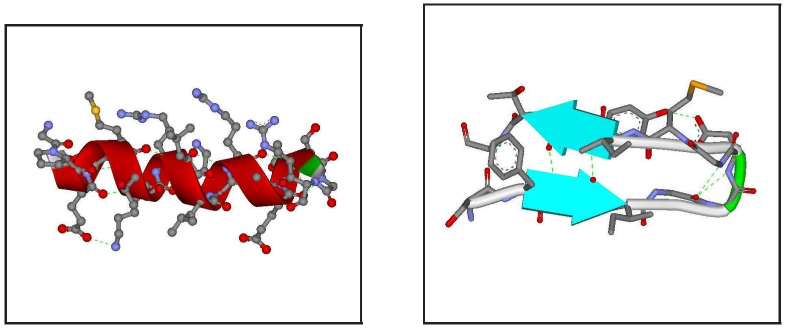
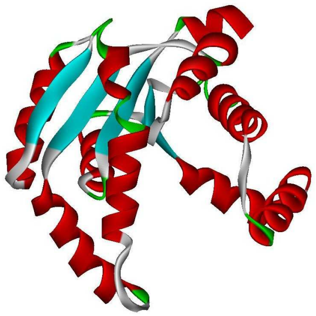


FIG. 2 – Elements structuraux périodiques : hélice α (à gauche) et brins β (à droite)

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



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

- **Intégration des modules - serveur de prédictions en ligne**

– Développement de méthodes de combinaison de modèles inhomogènes
– Réalisation d'un serveur disponible sur le site web du PBIL-IBCP

Un ensemble de domaines de l'apprentissage

- **Apprentissage de modèles syntaxiques**

  – Inférence de structures d'automates (probabilistes) : IRISA, LIF, LIRMM
  – Modèles probabilistes génératifs : IBCP, LIRMM, LORIA, MIG

- **Apprentissage numérique**

  – Théorie statistique de l'apprentissage - Méthodes à noyau - Apprentissage semi-supervisé : IRISA, LIF, LORIA
  – Systèmes connexionnistes supervisés (*feed-forward* et récurrents) et non supervisés : IBCP, LORIA

Premiers résultats obtenus - Prédiction des ponts disulfures

- **Différentes tâches**
  - Prédire l'état de chaque cystéine (oxydée - non oxydée)
  - Prédire les ponts (couples de cystéines impliquées dans un même pont)
- **Méthodes utilisées (caractérisation de l'information pertinente)**
  - Statistiques sur des alignements multiples (voisinages favorisés) : IBCP
  - SVM et perceptrons multi-couches (sélection / pondération de prédicteurs) : LORIA
  - Inférence grammaticale de langages de contrôle (interactions à longue portée) : IRISA
  - Bayes naïf semi-supervisé (utilisation de paires de cystéines ne formant pas un pont) : LIF
  - Programmation logique inductive (extraction de connaissance) : IRISA

Premiers résultats obtenus - Prédiction des ponts disulfures

- **Différentes tâches**
  - Prédire l'état de chaque cystéine (oxydée - non oxydée)
  - Prédire les ponts (couples de cystéines impliquées dans un même pont)
- **Méthodes utilisées (caractérisation de l'information pertinente)**
  - Statistiques sur des alignements multiples (voisinages favorisés) : IBCP
  - SVM et perceptrons multi-couches (sélection / pondération de prédicteurs) : LORIA
  - Inférence grammaticale de langages de contrôle (interactions à longue portée) : IRISA
  - Bayes naïf semi-supervisé (utilisation de paires de cystéines ne formant pas un pont) : LIF
  - Programmation logique inductive (extraction de connaissance) : IRISA

Premiers résultats obtenus - Structure secondaire

- **Prédiction de l'appariement des brins β : MIG**

Recherche d'une information dans la séquence

- **Méthode de prédiction hybride : LORIA + IBCP**

– Ensemble d'experts effectuant une première prédiction à partir d'un alignement multiple
– SVM multi-classe incorporant un noyau dédié au traitement de séquences protéiques
– Post-traitement des prédictions par un HMM inhomogène

Premiers résultats obtenus - Structure secondaire

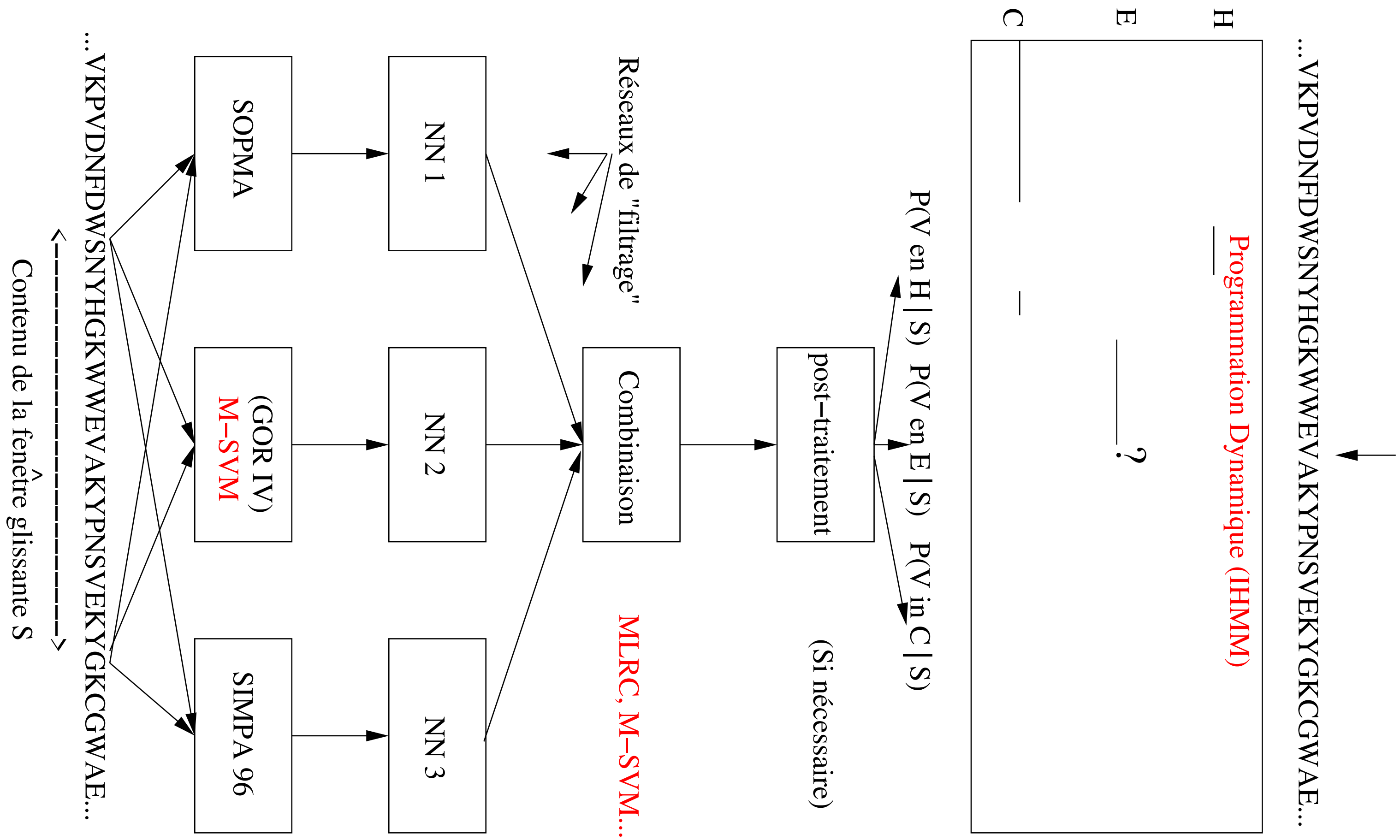


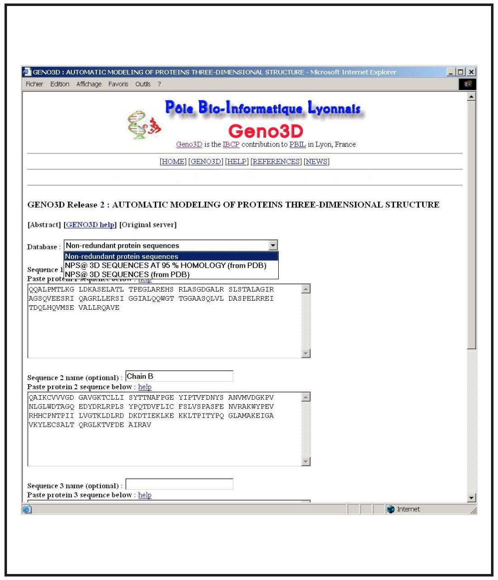
FIG. 3 – Architecture hiérarchique et modulaire pour la prédiction de la structure secondaire

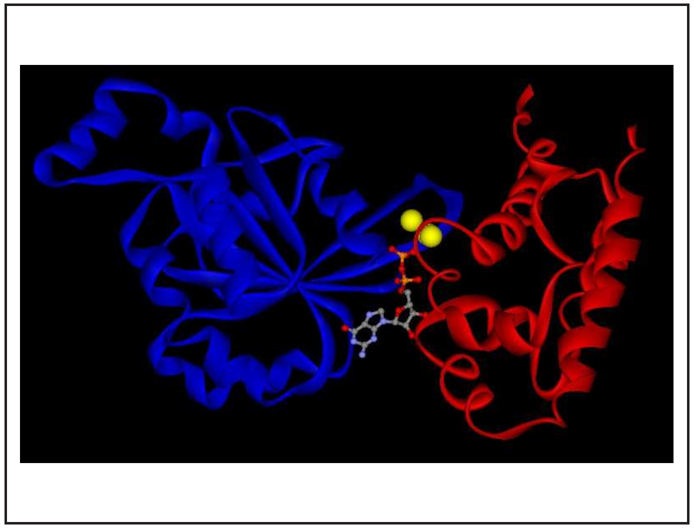
Premiers résultats obtenus - Geno3D - Version 2.0

- **Définition**

– Système automatique de modélisation de la structure 3D par homologie et analogie

– Sélection de l'empreinte validée suivant la compatibilité des structures secondaires prédites





- **Caractéristiques**

– Participe à la compétition CAFASP4

– Utilisé à grande échelle (modélisation de l'ensemble des protéines d'un protéome)

Méthodes de reconnaissance de repliements : FROST

- **But**

Détecter des relations entre homologues lointains

- **Principe**

Alignement séquence - structure 3D

- **Rationnel**

– Bonne conservation de la structure 3D
– Nombre limité de structures 3D possibles

- **Méthode**

1. Bibliothèque de structures 3D représentatives des structures connues
2. Fonctions de score mesurant l'adéquation de la séquence avec une structure 3D
3. Algorithme d'alignement d'une séquence sur une structure de *score optimal*
4. Evaluation de la significativité du score

- **Première version opérationnelle**

FROST : nouvelles fonctions de score

- **Problème à résoudre**

Difficile avec une seule fonction de score de capturer la relation séquence -structure 3D dans un modèle simple

- **Approche utilisée**

– Développement de plusieurs fonctions de score capturant chacune un aspect du problème
– Chaque fonction est utilisée comme un filtre

- **Combinaison des filtres**

Combinaison des résultats des différents filtres avec une SVM (DEA co-encadré IRISA-MIG)

FROST : nouveaux algorithmes d'alignement

- **Problèmes à résoudre**

  1. Alignement séquence-structure 3D : problème NP-complet très coûteux
     - La plupart des méthodes utilisent des heuristiques
     - Algorithme exact (*branch & bound*) : n'est utilisable en pratique que pour des petites instances
  2. Essayer de repérer des domaines structuraux dans de grandes séquences protéiques

- **Approche utilisée**

  1. Reformulation de l'algo. sous forme d'un problème de programmation linéaire en nombres entiers. Méthodes très performantes : on peut maintenant traiter de très grandes instances (10^{40} alignements) en quelques secondes
  2. Développement d'un algorithme d'alignement semi-global

- **Tests des modifications lors de CASP6**

Nos objectifs

- **Poursuite et synthèse des travaux dédiés aux différents aspects du problème**

Prédiction par homologie, analogie et *threading*

– Développement de Geno3D par incorporation d'un modèle d'apprentissage hybride pour la classification et l'extraction de la structure prototype

– Développement de la méthode de reconnaissance de repliements FROST

Vers la prédiction *de novo*

– Poursuite et intégration des travaux en prédiction des ponts disulfures et salins

– Poursuite et intégration des travaux en prédiction de la structure secondaire

- **Mise en place du serveur de prédictions**

– Mise à disposition des différentes réalisations logicielles des membres du projet

– Réalisation du serveur proprement dit autour de Geno3D et de FROST

$\Longrightarrow$ Participation au concours CAFASP5 (2006)

Bibliographie du projet

Revues internationales - Chapitres de livres

[1] R. Andonov, S. Balev, et N. Yanev, Protein Threading Problem : From Mathematical Models to Parallel Implementations. *INFORMS Journal on Computing, Special Issue on Computational Molecular Biology/Bioinformatics*, 6(4) 2004.

[2] P. Dupont, F. Denis et Y. Esposito. Links between Probabilistic Automata and Hidden Markov Models : probability distributions, learning models and induction algorithms *Pattern Recognition : Special Issue on Grammatical Inference Techniques & Applications* 2005. (à paraître).

[3] M. Errami, C. Geourjon et G. Deléage. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. *Bioinformatics*, 19 :506–512, 2003.

[4] M. Errami, C. Geourjon et G. Deléage. Conservation of amino acids into multiple alignments involved in pairwise interactions in three-dimensionnal protein structures. *Journal of Bioinformatics and Computational Biology*, 3 :505–520, 2003.

[5] Y. Guermeur, A. Lifchitz et R. Vert. A kernel for protein secondary structure prediction. Dans *Kernel Methods in Computational Biology*, édité par K. Tsuda, B. Schölkopf et J.-P. Vert, The MIT Press, 192–206, 2004.

[6] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy et P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56C :305–327, 2004.

[7] M. Jambon, A. Imberty, G. Deléage et C. Geourjon. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins, Structure function and genetics*, 52 :137-145, 2003.

Conférences internationales

[8] F. Denis et Y. Esposito. Learning classes of Probabilistic Automata. *17th Annual Conference on Learning Theory (COLT'04)*, LNAI 3120, 124–139, 2004.

Conférences nationales

[9] L. Bréhelin. Une approche bayésienne pour la classification de cinétiques d'expression de gènes. *Actes de JOBIM 2004*.

[10] F. Coste, G. Kerbellec, B. Idmont et D. Fredouille. Apprentissage d'automates par fusions de paires de fragments significativement similaires et premières expérimentations sur les protéines MIP. *Actes de JOBIM 2004*.

[11] F. Denis et Y. Esposito. Identification in the limit of Probabilistic Non Deterministic Automata and Undecidable problem for Multiplicity Automata. *Actes de CAP 2004*, Presses Universitaires de Grenoble, 81–96, 2004.