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)

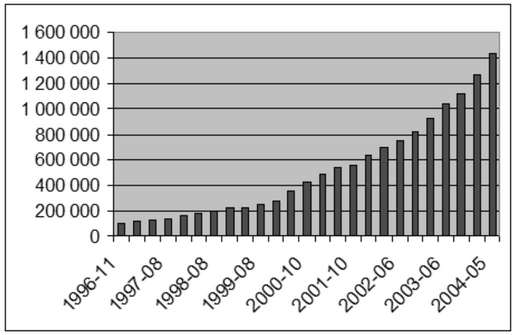


FIG. 1 – Croissance de la banque internationale TREMBL de 1996 à 2004

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**

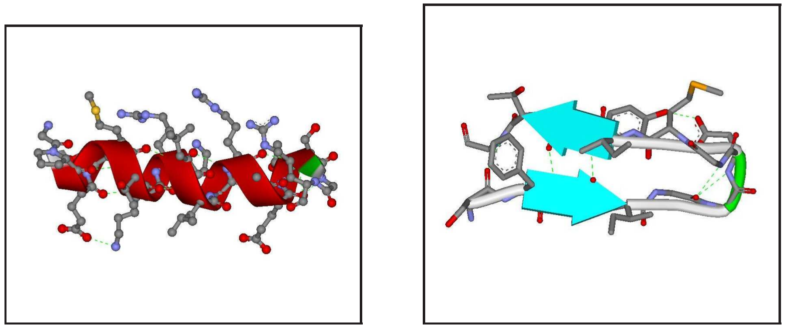
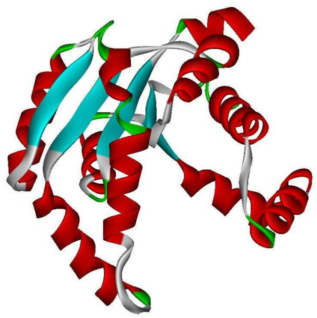
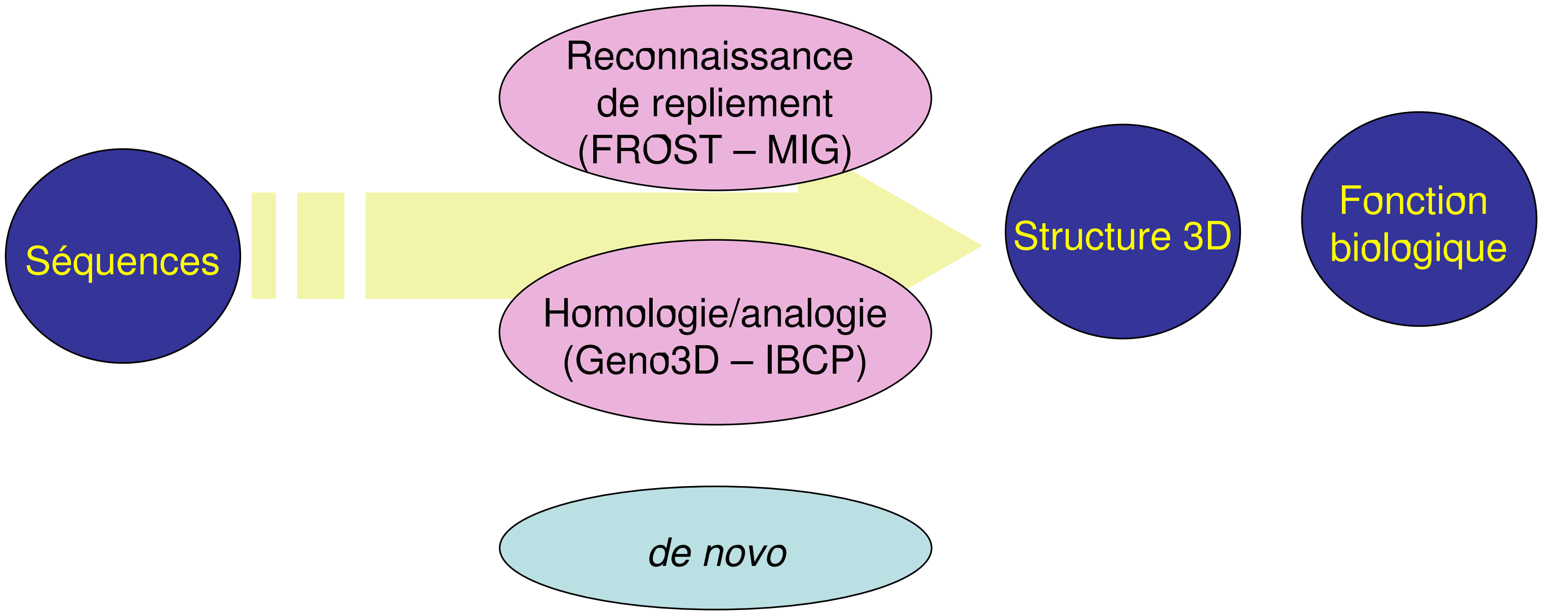
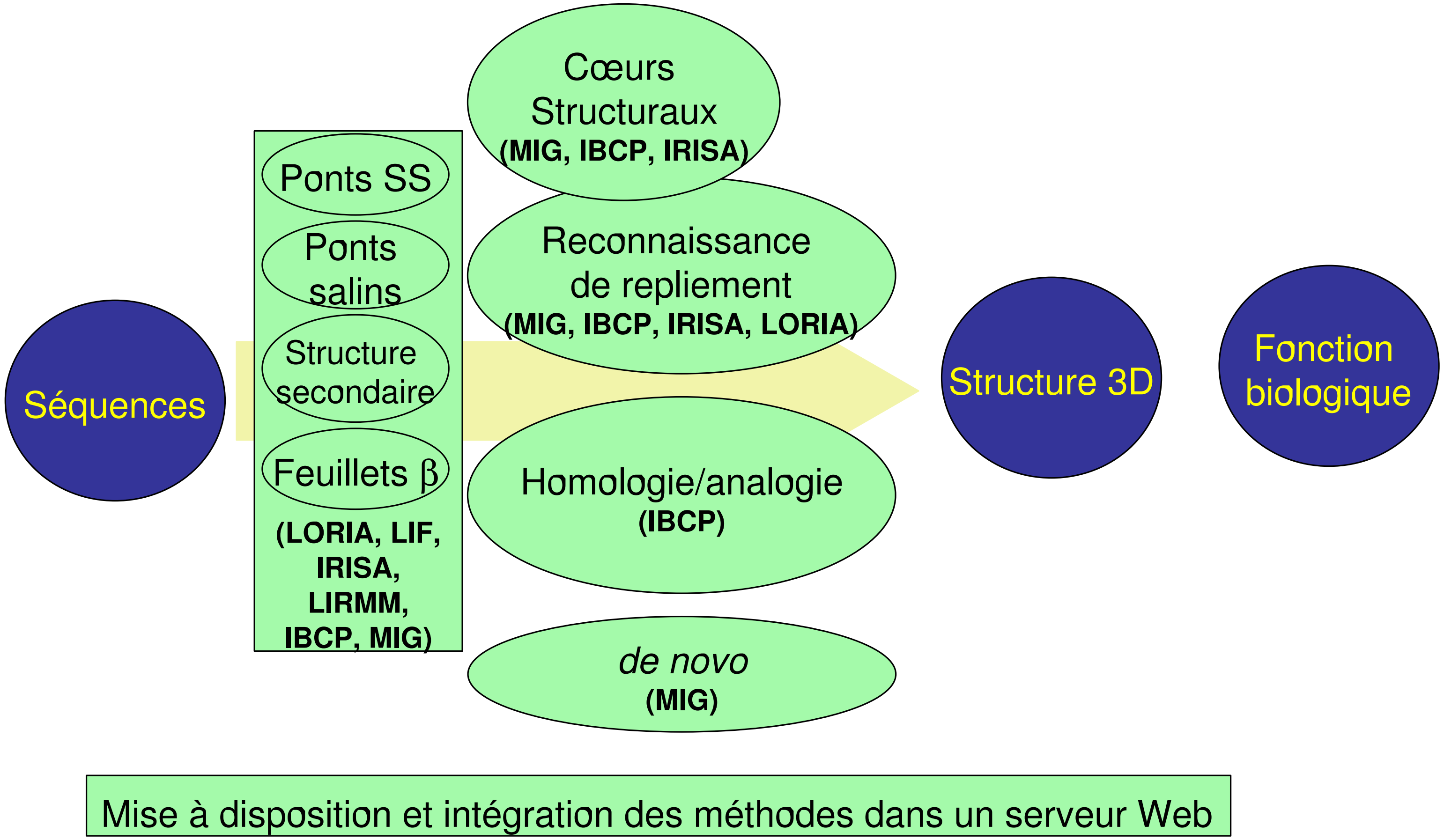


FIG. 2 – Elements structuraux périodiques : hélice α (à gauche) et brins β (à droite)

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




Reconnaissance de repliement (FROST – MIG)
Séquences
Structure 3D
Fonction biologique
Homologie/analogie (Geno3D – IBCP)
de novo


Séquences
Ponts SS
Ponts salins
Structure secondaire
Feuillets β
(LORIA, LIF, IRISA, LIRMM, IBCP, MIG)
Cœurs Structuraux
(MIG, IBCP, IRISA)
Reconnaissance de repliement
(MIG, IBCP, IRISA, LORIA)
Homologie/analogie
(IBCP)
de novo
(MIG)
Structure 3D
Fonction biologique
Mise à disposition et intégration des méthodes dans un serveur Web

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 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

Apprentissage de modèles syntaxiques : Caractérisation de protéines par automates

- **Problème**

A partir d'exemples de séquences protéiques E_+ (et éventuellement de contre exemples E_-) partageant une même structure, caractériser et modéliser par un automate ce repliement.

But :
– Classification (prédiction de la classe de repliement d'une nouvelle séquence)
– Modélisation (expliciter les positions importantes, structure d'automates probabilistes, nouveaux cœurs . . .)

- **Une nouvelle approche**

Développement d'une approche originale par fusion de fragments significativement similaires

Approche par fusion de fragments significativement similaires

- **Principe**

1. Sélection des paires de fragments significativement similaires
2. Ordonnancement de ces paires de fragments suivant leur représentativité dans la famille de séquences.
   Support (E_+), indice d'implication (E_+, E_-) ou pouvoir de discrimination (Multiclasse).
3. Fusion des paires de fragments en préservant les fragments précédents
4. Post-généralisation
   – Suppression des états non caractéristiques
   – Identification des propriétés physico-chimiques

- **Une thèse débute...**

Goulven Kerbellec (Bourse Région Bretagne)
– Développement de l'approche
– Test sur des familles fonctionnelles et structurelles (+ stage Master2) de protéines.
– . . .

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

Prédiction des ponts disulfures

Les ponts disulfures sont des liaisons covalentes entre deux cystéines oxydées qui contraignent le repliement de la protéine.

- **Problème :** Etant donnée une protéine (séquence d'acides aminés) de $2N$ cystéines, prédire les paires de cystéines formant un pont.
- **Deux types d'information possibles pour la prédiction :**

1. Informations locales : les "fenêtres" d'acide aminés qui entourent les cystéines
2. Informations globales : les informations de type topologique (configuration des ponts, rang des cystéines, espacement entre les cystéines, etc.)

- **On s'intéresse à deux problématiques :**

– Caractérisation de l'information de type 1 : recherche d'un modèle probabiliste de l'appariement.

– Utilisation de l'information potentiellement contenue dans les paires de cystéines **non appariées** : algorithme d'apprentissage semi-supervisé.

Recherche d'un modèle probabiliste de l'appariement

- Les règles d'appariement sont fortement corrélées avec des informations de type structurel : longueur des protéines, espacement entre les cystéines, ordre des cystéines, etc.
- Certains travaux (*e.g.* **(Vullo et Frasconi, 99), (Fariselli et Casadio, 01)**) semblent montrer que des informations locales (relatives aux acides aminés qui entourent les cystéines) peuvent aider à la prédiction des paires appariées.
- Différents facteurs peuvent biaiser ces résultats (difficile de distinguer ce qui relève d'une information réellement locale). Pas de modèle d'appariement entre les paires de fenêtres d'acides aminés qui soit satisfaisant.

- **Approche adoptée pour mettre à jour l'information locale :**

Recherche d'une modélisation probabiliste cohérente des paires de fenêtres intervenant dans les ponts.

Techniques utilisées : Théorie de l'information, Tests statistiques, Classification supervisée et non-supervisée, Algorithmes EM.

Prédiction de l'appariement par classification semi-supervisée

Schéma de connexion : graphe non orienté pondéré dont les sommets sont les cystéines oxydées et les arcs représentent une interaction potentielle.

Pour calculer les poids : inférence d'un modèle probabiliste décrivant la probabilité qu'une paire de cystéines forme un pont.

Contrainte forte sur le nombre de ponts : $nbPonts \leq nbCys/2$.

Hypothèse : les paires de cystéines qui ne forment pas de ponts, pourraient peut-être en former un dans d'autres conditions, et recèlent une information qui doit pouvoir être exploitée par les modèles prédictifs.

Développement

- $x_i = \phi(c_i, P)$: une information locale à la cystéine c_i de la protéine P.

Exemple : une fenêtre de diamètre $2k+1$ centrée sur c_i, et modélisée par une chaîne d'AA aléatoire X ;

- x_1 et x_2 sont *compatibles* ($C(x_1, x_2)$) s'il (peut) existe(r) une protéine P contenant 2 cystéines c_1 et c_2 formant un pont et tq $x_1 = \phi(c_1, P)$ et $x_2 = \phi(c_2, P)$;

- $Z = (X_1, X_2)$, 2 copies indépendantes de X

- $P(B|\mathbf{z})$: probabilité que la paire $\mathbf{z}$ forme un pont.

Supposition : $P(\cdot|B) = P(\cdot|C)$.

Propriété : $P(B) < P(C)$ et $P(B|\mathbf{z}) \neq 0 \Rightarrow C(\mathbf{z})$.

$$P(\mathbf{x}|\overline{B}) = \frac{P(\mathbf{x}|B)(P(C) - P(B)) + P(\mathbf{x}|\overline{C})P(\overline{C})}{1 - P(B)}.$$

Sous ces hypothèses

- les ponts observés sont un échantillon de la distribution $P(\cdot|C)$

- les "paires non appariées" sont un échantillon d'un mélange des distributions $P(\cdot|C)$ et $P(\cdot|\overline{C})$.

- Un *modèle probabiliste* est un triplet (P^+, P^-, λ) où P^+ et P^- modélisent $P(\cdot|C)$ et $P(\cdot|\overline{C})$ et où λ représente un coef. de mélange.

- θ_{ML} : le modèle qui maximise la vraisemblance

$$L(D) = \prod_{\mathbf{z} \in D_B} P^+(\mathbf{z}) \prod_{\mathbf{z} \in D_{\overline{B}}} [\lambda P^+(\mathbf{z}) + (1-\lambda) P^-(\mathbf{z})].$$

- Remarque : en général, ce problème d'optimisation est mal-posé.

- Mais, le principe est consistant pour certaines familles de modèles : θ_{ML} converge vers l'unique solution avec le nombre de données disponibles.

Un modèle étant inféré, utiliser $P(C|\mathbf{z})$ pour prédire si $\mathbf{z}$ peut ou pourrait former un pont.

Expériences et résultats

Données **(Errami, Geourjon et Deléage, 03)** : 227 protéines extraites de la PDB, dont les cystéines sont oxydées et toutes impliquées dans des ponts de la même chaîne ; degré d'homologie entre séquences inférieur à 25%.

Σ : alphabet composé de paires d'acides aminés. Une paire de fenêtre $(\mathbf{a}, \mathbf{b})$ est représentée par le multi-ensemble $\{\{a_i, b_j\} | -k \leq i, j \leq k\}$.

Méthode de classification : naive Bayes + EM.

#Pts	#Prot	$Q_p(R)$	$Q_p(NB_{sup})$	$\mathbf{Q_p(NB_{ssup})}$	Q_{Freq}	Q_{dist}
2	51	33.3%	39.4%	61.0%	49.0%	56.3%
3	50	20.0%	20.4%	34.9%	27.6%	31.7%

Conclusion : l'hypothèse semble validée ; à développer avec des classifieurs moins naïfs et une information plus riche.

Apprentissage de règles avec la Programmation Logique Inductive (ILP)

- **Logiciel utilisé -> PROGOL : Muggleton (∼ 98)**
- **Background knowledge**
- **Echantillon -> fenêtres d'acides aminés : 722 positifs, 50 négatifs**
- **Format des règles**

Les exemples contiennent un contexte gauche appelé *left()* et un contexte droit appelé *right()*. Ces 2 contextes représentent les séquences d'acides aminés de chaque côté de la cystéine dans la protéine Figure 3.

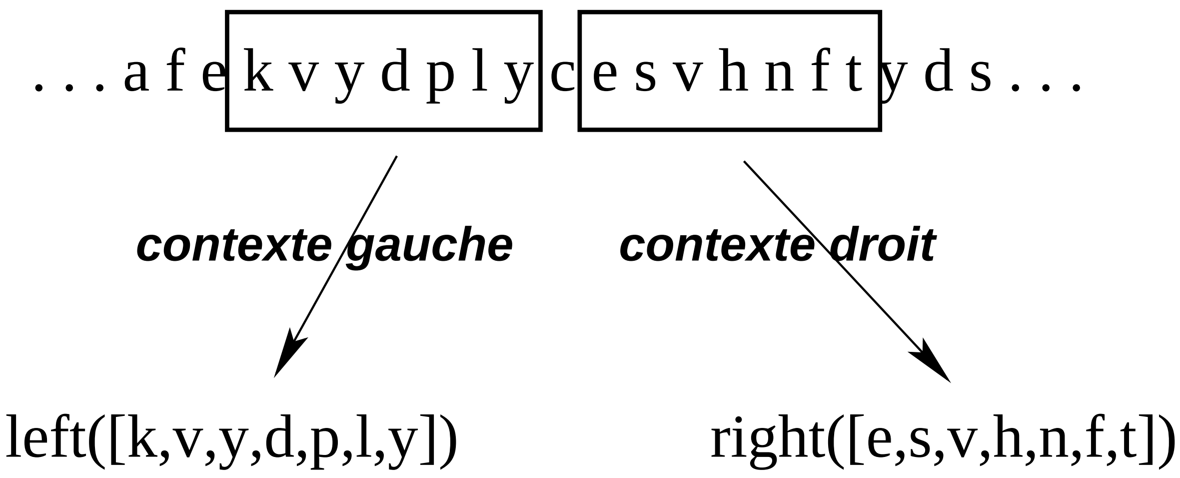


FIG. 3 – Sélection des fenêtres autour d'une cystéine

Déclaration de l'en-tête de la clause :

:– modeh(1,example(+lcontext,+rcontext)) ?

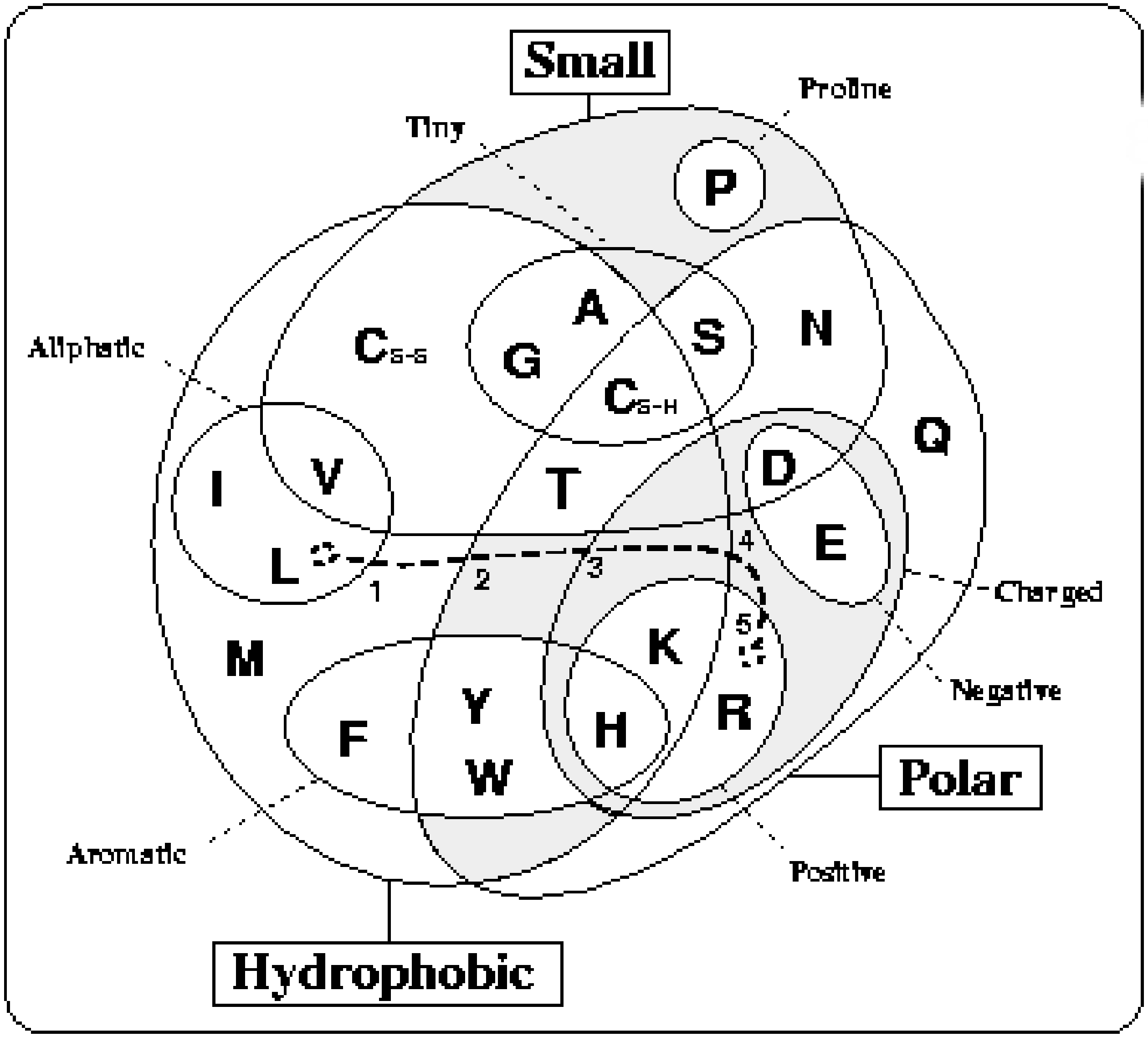


Fig. 4 – Diagramme de Taylor

- **Prédicats autorisés dans le corps des hypothèses :**

:−modeb(∗,lwindow(+lcontext,−pos,#llist_property)) ?
:−modeb(∗,rwindow(+rcontext,#pos,#rlist_property)) ?

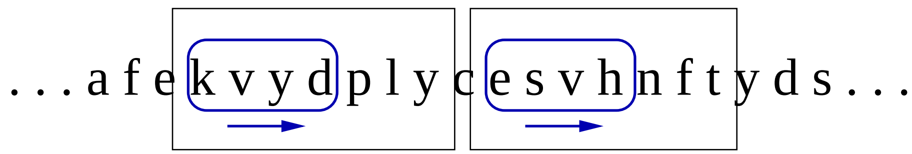


Fig. 5 – Fenêtre glissante pour la recherche de modèles de taille 4

- **But : obtenir des modèles physico-chimiques en relation dans chacun des deux contextes**
- **Résultats obtenus satisfaisants, règles du style :**

example(A,B) :− lwindow(A,C,D,lprop(quatuor(hydrophobic,polar,hydrophobic,small))),
rwindow(B,E,F,rprop(quatuor(hydrophobic,polar,small,hydrophobic))).

- **2e niveau d'apprentissage : Apprendre sur les règles obtenues afin de diminuer l'espace de Herbrand**
- **Gain de temps**
- **Autorisation de bruit**
- **5 règles couvrent 500 exemples positifs**

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

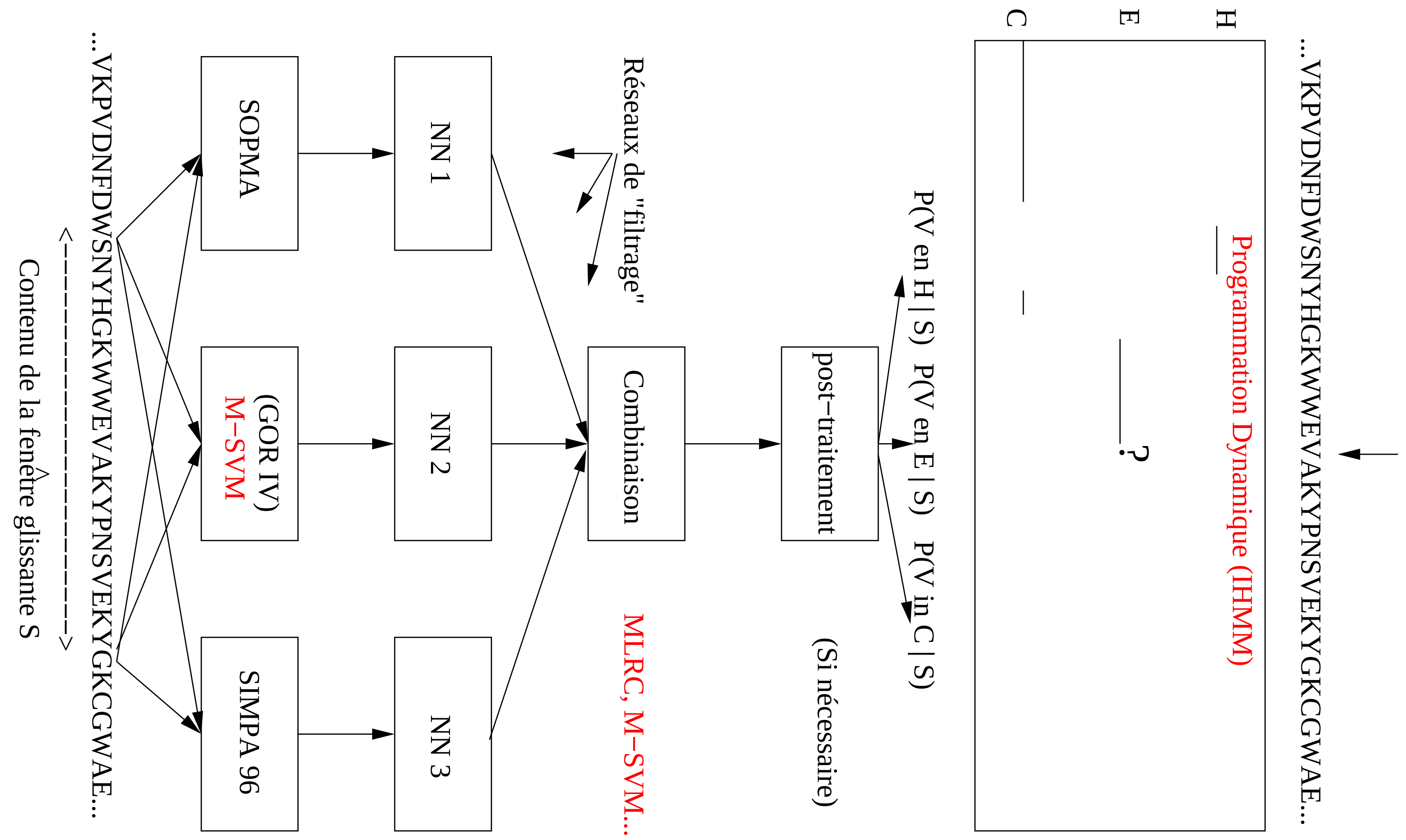


FIG. 6 – Architecture hiérarchique et modulaire pour la prédiction de la structure secondaire

Machines à vecteurs support multi-classes (M-SVMs) pour la prédiction de la structure secondaire

- **Architecture (commune à toutes les machines)**

L'ensemble $\mathcal{H}$ des fonctions $h = (h_k)$ calculées par l'architecture est défini par :

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \ldots, Q\},\ h_k(x) = \tilde{h}_k(x) + b_k = \langle w_k, \Phi(x) \rangle + b_k$$

où

– Φ est soit une transformation non linéaire dans l'*espace de représentation* soit l'identité

– $\tilde{\mathcal{H}} = \left\{ \tilde{h} = \left(\tilde{h}_k \right) \right\}$ est le produit de Q *espaces de Hilbert à noyau reproduisant* (RKHS)

– κ, le *noyau* associé à $\tilde{\mathcal{H}}$, est lié à Φ par la formule :

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

Apprentissage des M-SVM : principe général

Soit $s_m = ((x_1, C(x_1)), \ldots, (x_m, C(x_m)))$ l'échantillon d'apprentissage

- **Formulation primale (avec *marges molles*)**

Problème 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} + C \sum_{i=1}^{m} \ell\left(h(x_i), C(x_i)\right) \right\}$$

- **Formulation duale**

Un *théorème de représentation* établit que l'apprentissage (la résolution du Problème 1) revient à trouver les valeurs des coefficients β_{ik} dans :

$$\forall k \in \{1, \ldots, Q\}, \; h_k(x) = \sum_{i=1}^{m} \beta_{ik} \kappa(x_i, x) + b_k$$

Nouveau noyau pour le traitement des séquences protéiques

- **Expression analytique (structure primaire seule)**

(x, x') : vecteurs codant les contenus des deux fenêtres d'analyse à comparer

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

- **Extension au traitement des alignements multiples**

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

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

Estimation du vecteur θ

- **Alignement noyau-cible : cas multi-classe (Vert, 02)**

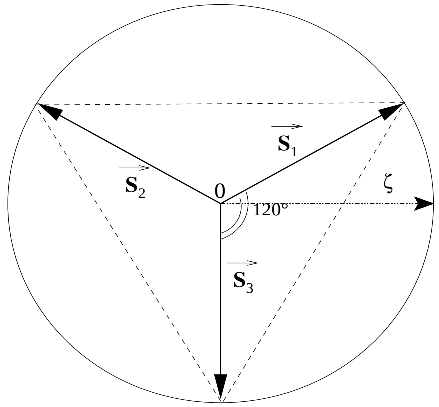


FIG. 7 – Choix optimal des représentants des catégories (cas de 3 classes)

$$\kappa_t(x, x') = \begin{cases} \zeta^2 & \text{si } C(x) = C(x') \\ -\zeta^2/(Q-1) & \text{sinon} \end{cases}$$

- **Algorithme d'apprentissage**

Descente de gradient stochastique

Estimation de la “largeur de bande” μ

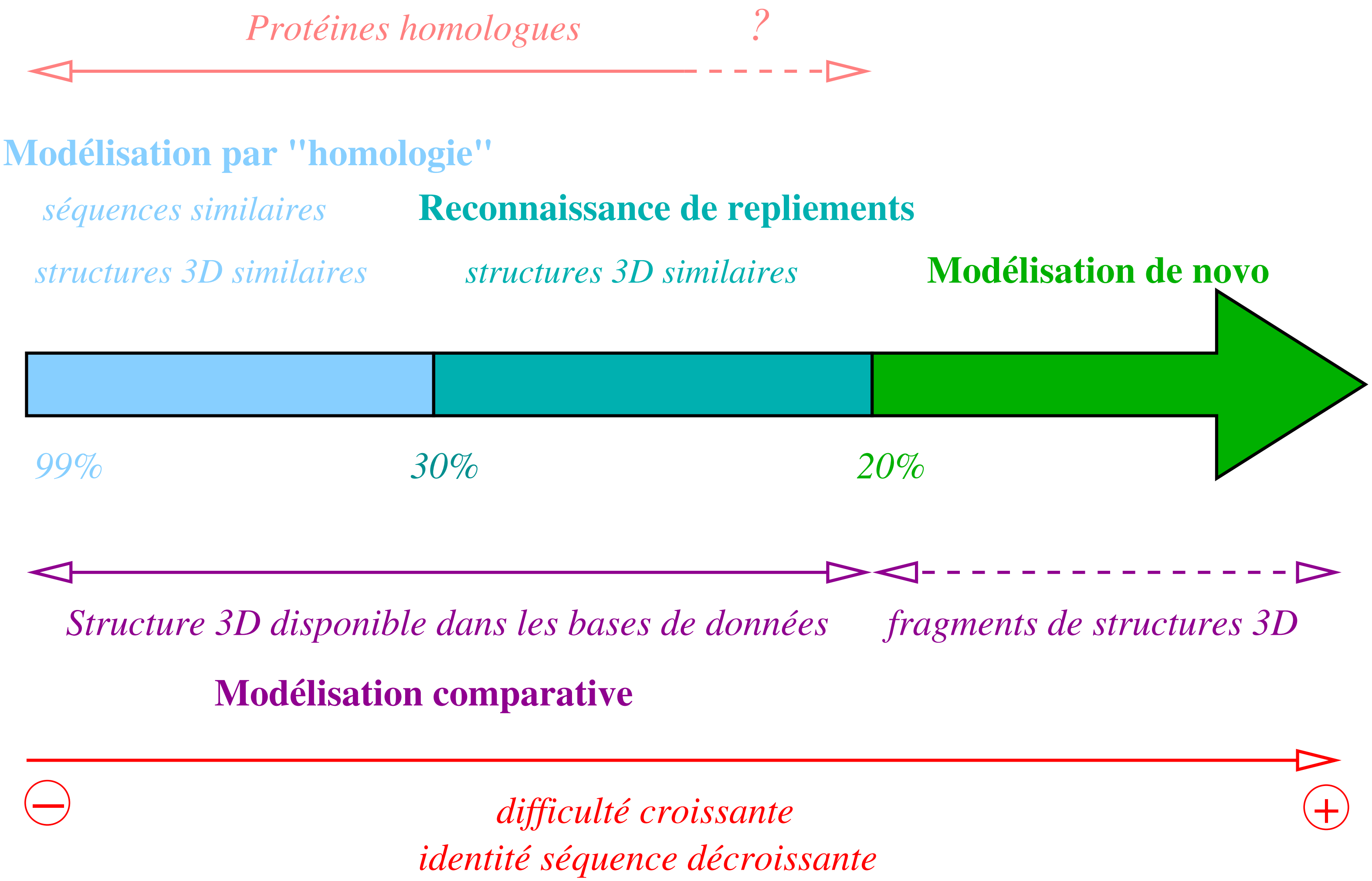
- **Utilité de procéder en deux temps**

1. ζ étant fixé (ici à 1), le vecteur θ est déterminé sous la contrainte $\theta_0 = 1$, en fonction de la valeur (arbitraire) de μ
2. $\Longrightarrow \mu$ est ensuite choisi suivant une procédure qui lui est propre (validation croisée ...)

- **Axe privilégié : mise en œuvre du principe inductif SRM**

– Utilisation de bornes *leave-one-out* dédiées aux M-SVM
– Utilisation de bornes *à la Vapnik* pour les systèmes discriminants multi-classes à grande marge
– Utilisation d’autres types de bornes (mesures de capacité empiriques ...)

Modélisation de ls structure 3D



Méthodes de reconnaissance de repliements

- **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

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 scores 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 de petites instances
2. Essayer de repérer des domaines structuraux dans de grandes séquences protéiques

- **Approche utilisée**

Reformulation du problème sous forme d'un modèle de programmation linéaire.

Méthode très performante : on peut maintenant traiter de très grandes instances (10^{40} alignements) en quelques secondes (Modèle plus performant que le modèle RAPTOR classé dans les premiers à CASP) **(Andonov *et al.*, 04)**.

FROST : nouveaux algorithmes d'alignement

- **Perspectives**
  - Développement d'un algorithme d'alignement semi-global
  - Amélioration de la bibliothèque de cœurs
  - Tests des modifications lors de CASP6
  - Intégration de Frost dans DEISA (Distributed European Infrastructure for supercomputing Applications)

Performances des techniques d'alignement

Template	**Nb Sol**	**Min**	**Mean**	**Max**
1BGLA0	$5.4\ 10^{27}$	0.95	0.97	1.02
	$1.2\ 10^{35}$	0.95	0.97	1.02
	$3.5\ 10^{58}$	35.6	45.2	73.2
	$1.3\ 10^{70}$	102.4	145.7	510.0
	$6.6\ 10^{77}$	203.8	291.7	797.4
1QBA_0	$1.6\ 10^{33}$	1.82	1.84	1.89
	$8.3\ 10^{37}$	1.82	1.84	1.89
	$5.2\ 10^{57}$	27.1	36.3	76.6
	$2.8\ 10^{68}$	68.4	101.4	354.8
	$7.2\ 10^{75}$	130.1	207.0	789.8
1ALO_0	$3.1\ 10^{33}$	0.85	0.87	0.89
	$6.0\ 10^{33}$	0.85	0.87	0.89
	$2.5\ 10^{57}$	25.8	40.8	135.2
	$1.6\ 10^{69}$	67.4	123.2	397.6
	$1.3\ 10^{77}$	139.9	262.2	735.0

Tests des modèles 3D

- **Problème à résoudre**

∗ Evaluer la qualité des modèles obtenus par modélisation comparative

– Mise en évidence des erreurs d'alignement (modélisation par homologie)
– Mise en évidence des erreurs de cible (modélisation par reconnaissance de repliements)

- **Approches utilisées**

– Développement d'une nouvelle fonction de score à la VERIFY3D
– Simulation de dynamique moléculaire des modèles pour en tester la stabilité

Méthodes *de novo*

- **Hypothèse**
  - On peut reconstituer une structure 3D à partir de fragments peptidiques
  - La PDB est un répertoire de tous les fragments possibles
- **Méthode**
  - Choix *judicieux* des fragments peptidiques
  - Assemblage des fragments peptidiques pour générer des structures 3D
  - Sélection de la “bonne” structure parmi celles générées

Développement de méthodes *de novo*

Choix des fragments peptidiques

- **Assignation des structures secondaires**

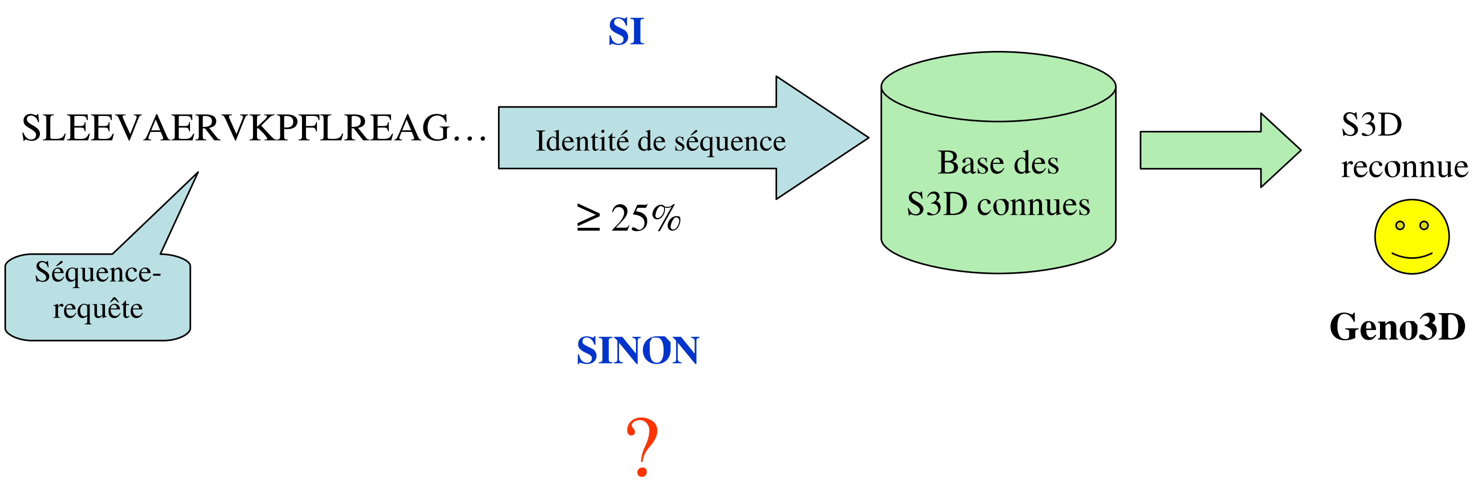
- **Prédiction des structures secondaires (zones apériodiques)**

– Utilisation de modèles de Markov cachés (3 états)
– Essai de prédiction de l'appariement des brins bêta
– Utilisation d'alignements multiples dans les modèles de Markov cachés

De GENO3D à GENOTO3D
Problématique

- Volume des données : Important
- Nature des données : Séquentielles
- Identité de séquences: Non significative

Fonction de prédiction : Complexe et non linéaire



Solution envisagée

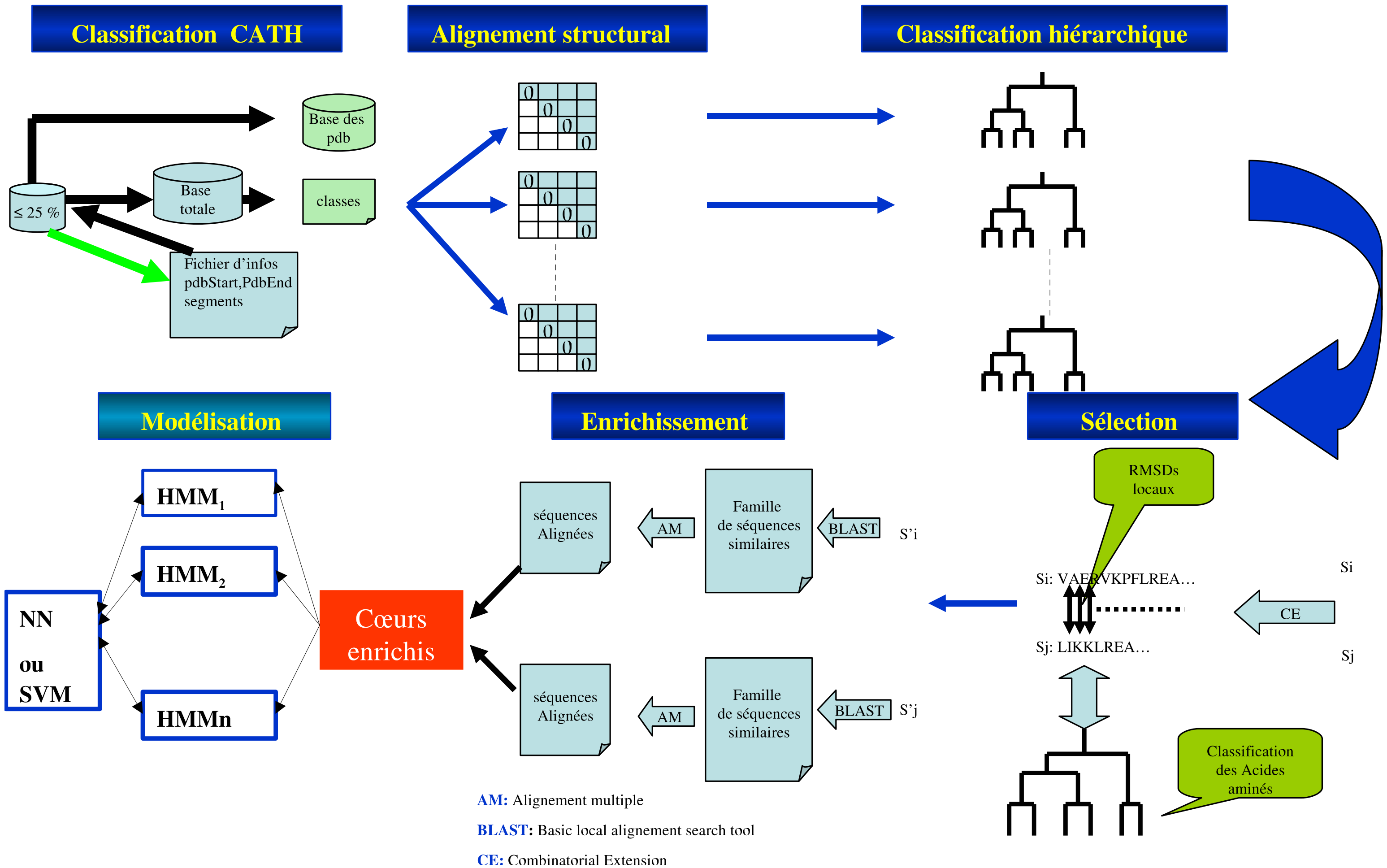
Système prédictif modulaire à base d'apprentissage

Composantes du système prédictif à base d'apprentissage

- Extraction des cœurs structuraux
  - Classification CATH (Class Architecture Topology Homology)
  - Alignement structural
  - Classification ascendante hiérarchique (CAH)
  - Sélection de composantes des cœurs
  - Enrichissement des cœurs
- Modélisation autour des cœurs structuraux
  - Modèle Hybride Markovien – Neuronal et/ou SVM (Y.Guermeur)
- Traitement de la dynamique
  - Heuristiques pour détermination de la taille de la fenêtre de prédiction
  - Sélection de variables pour l'optimisation

Extraction des cœurs structuraux pour la modélisation

Démarche



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.