# Extraction de cœurs structuraux et reconnaissance de repliements de protéines

**Khalid BENABDESLEM**

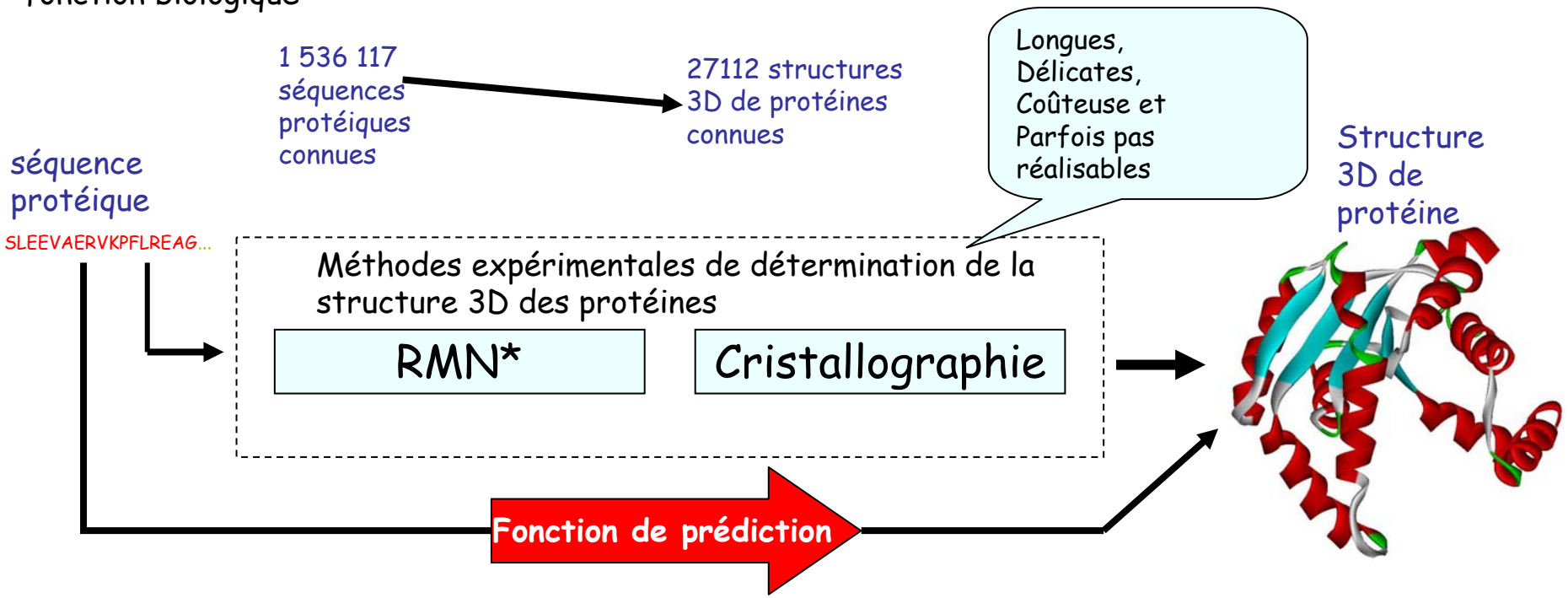
kbenabde@bat710.univ-lyon1.fr

IRISA , Le 12/10/2006

# Problématique Générale

## Contexte biologique

Exploitation fonctionnelle des informations provenant des grands programmes de séquençage des génomes: passe par la connaissance de la structure 3D des protéines. Cette structure 3D conditionne la fonction biologique



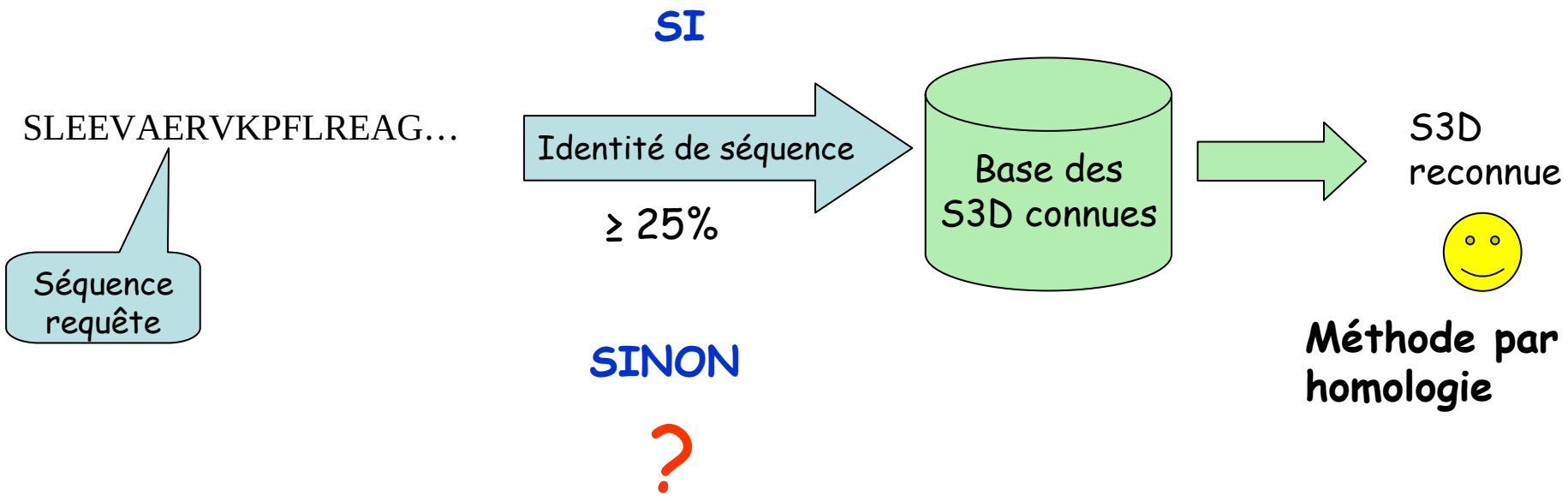

Problème centrale en biologie permettant d'aborder des grandes questions ouvertes en traitement de données séquentielles

RMN*: Résonance magnétique nucléaire

# Prédiction de la structure 3D

## Problématique informatique

- Volume des données : Important
- Nature des données : Séquentielles
- Fonction de prédiction (F): Complexe et non linéaire




Solution envisagée

Système prédictif modulaire à base d'apprentissage

# Système de prédiction

## Démarche

1) Extraction hiérarchique des cœurs structuraux à partir de familles de protéines (**ASCE**: Automatic structural cores extraction)

2) Alignement sur les cœurs structuraux

3) Reconnaissance de repliements: Modélisation et apprentissage

# Extraction des cœurs structuraux
## Méthodologie

### Objectif

-Construction d'un noyau pour chaque famille de structures à partir de la base des ≤ 25%

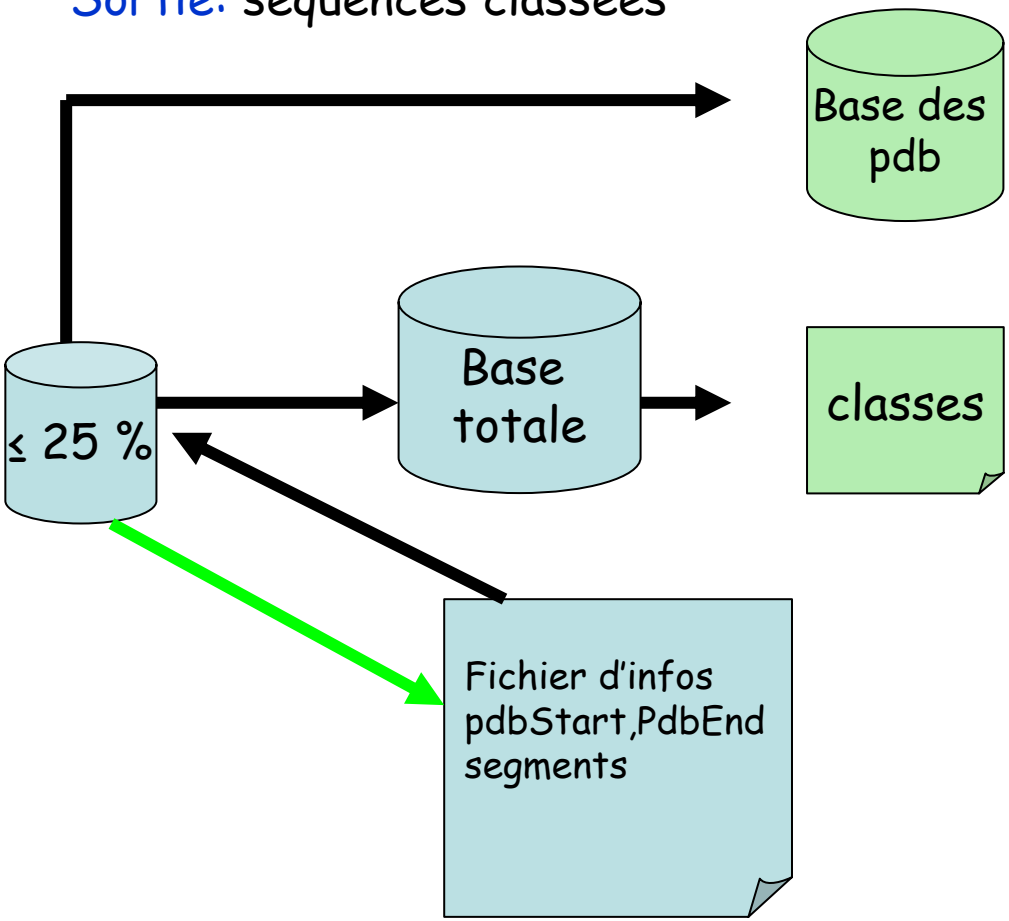
### Démarche

- Classification CATH : Correspondance entre différentes tables de différents fichiers
- Alignement structural : CE (Combinatorial Extension ), matrice de dissimilarités
- CAH : dendogramme à partir de la matrice
- Sélection des cœurs : Calcul de RMSDs locaux

# Extraction des cœurs structuraux
## Classification CATH (1)

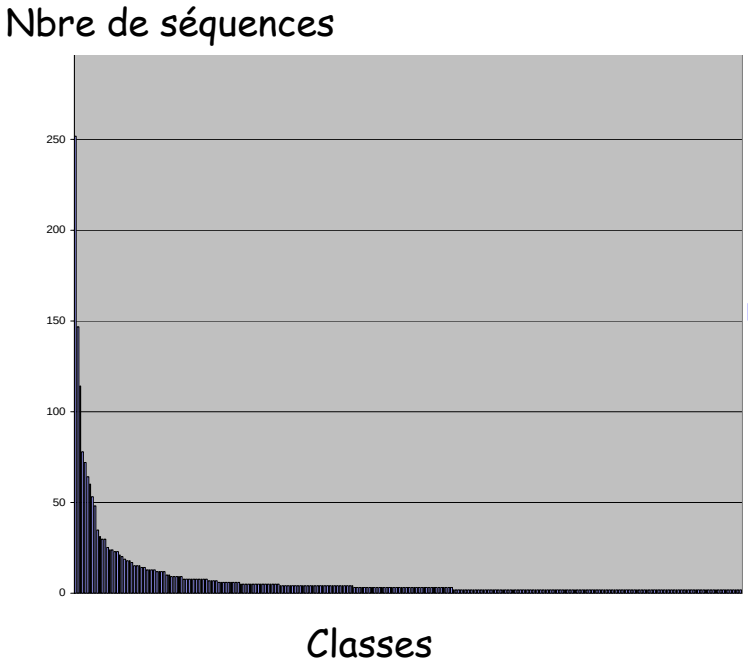
Entrée: Base de séquences de ≤ 25 %

Sortie: séquences classées




~ 3000 séquences

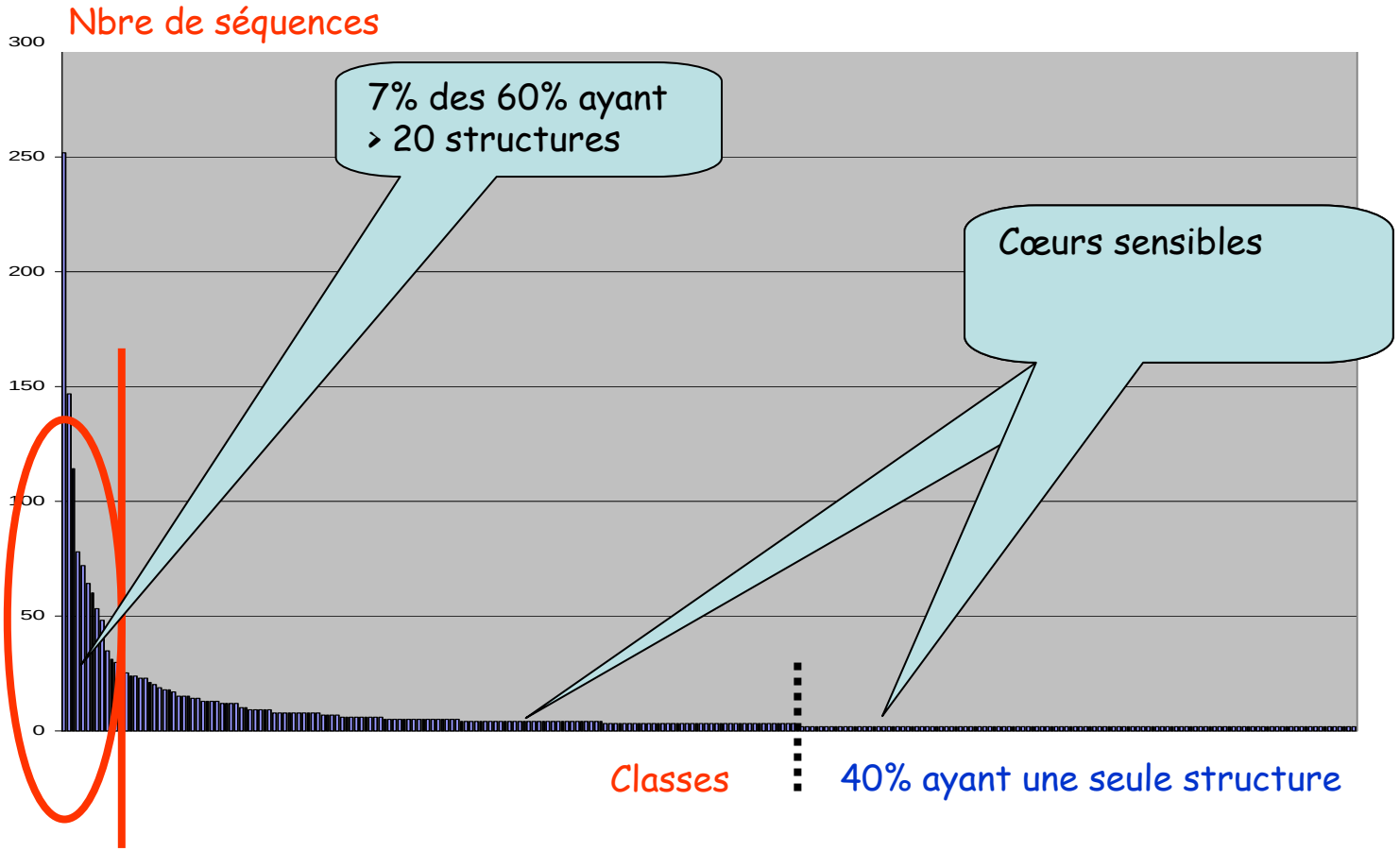
~ 600 classes



# Extraction des cœurs structuraux
## Classification CATH (2)

~ 3000 structures

~ 600 classes

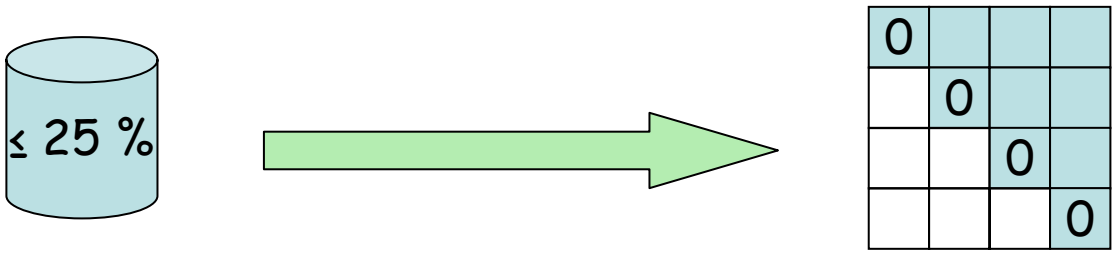


# Extraction des cœurs structuraux
## Alignement structural

Entrée: Famille de structures

Sortie: Matrice de dissimilarités structurales

Outil: CE (Z-Score, Rmsd, % de gaps, % identité de séquences, Alignement de séquences issu de l'alignement structural, matrice de Rotation - Translation




Règles: Si et Sj appartiennent à la même classes ssi:

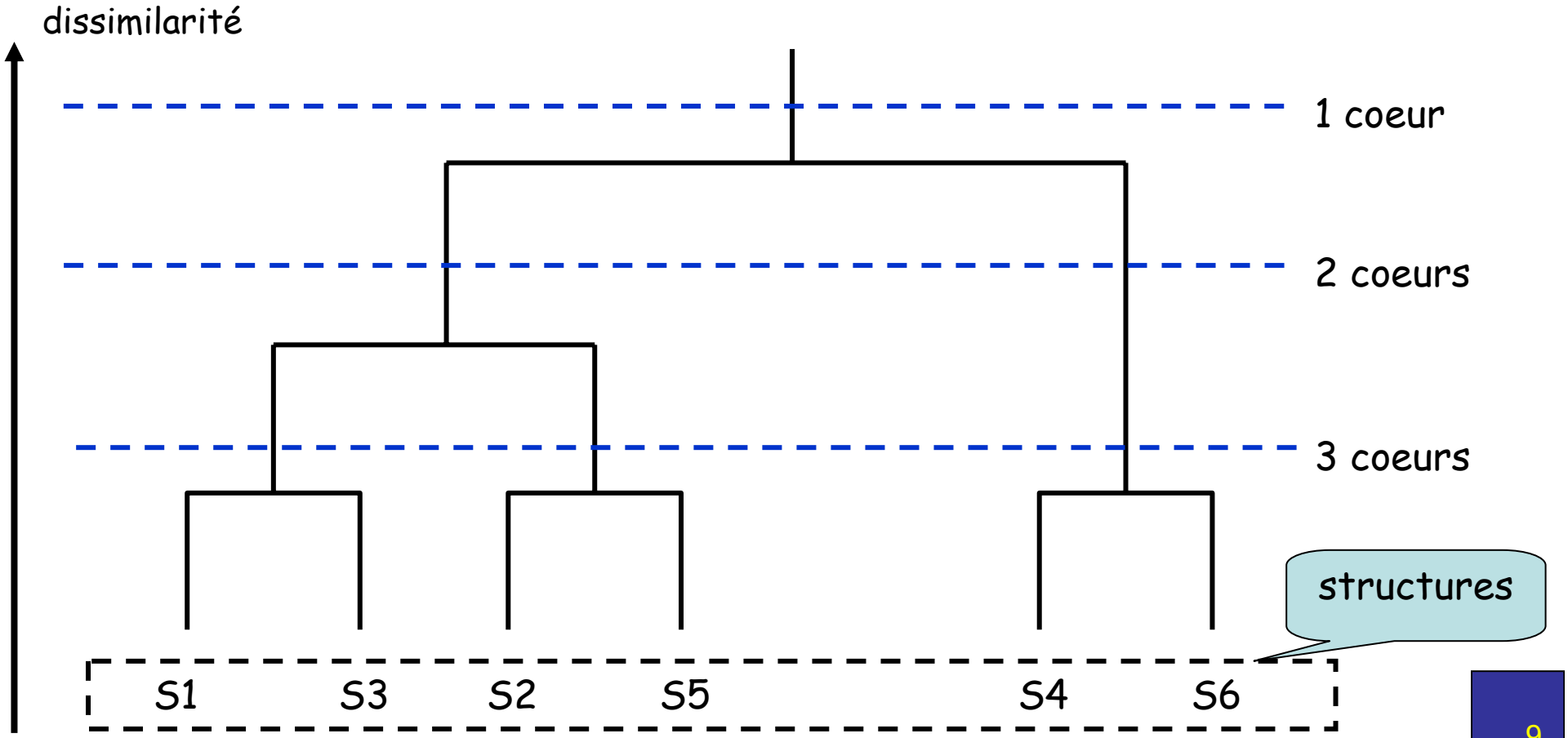
Z-Score ≥ 4.6 Ou

Z-Score < 4.6 Et Rmsd(Si,Sj) ≤ 2 Å

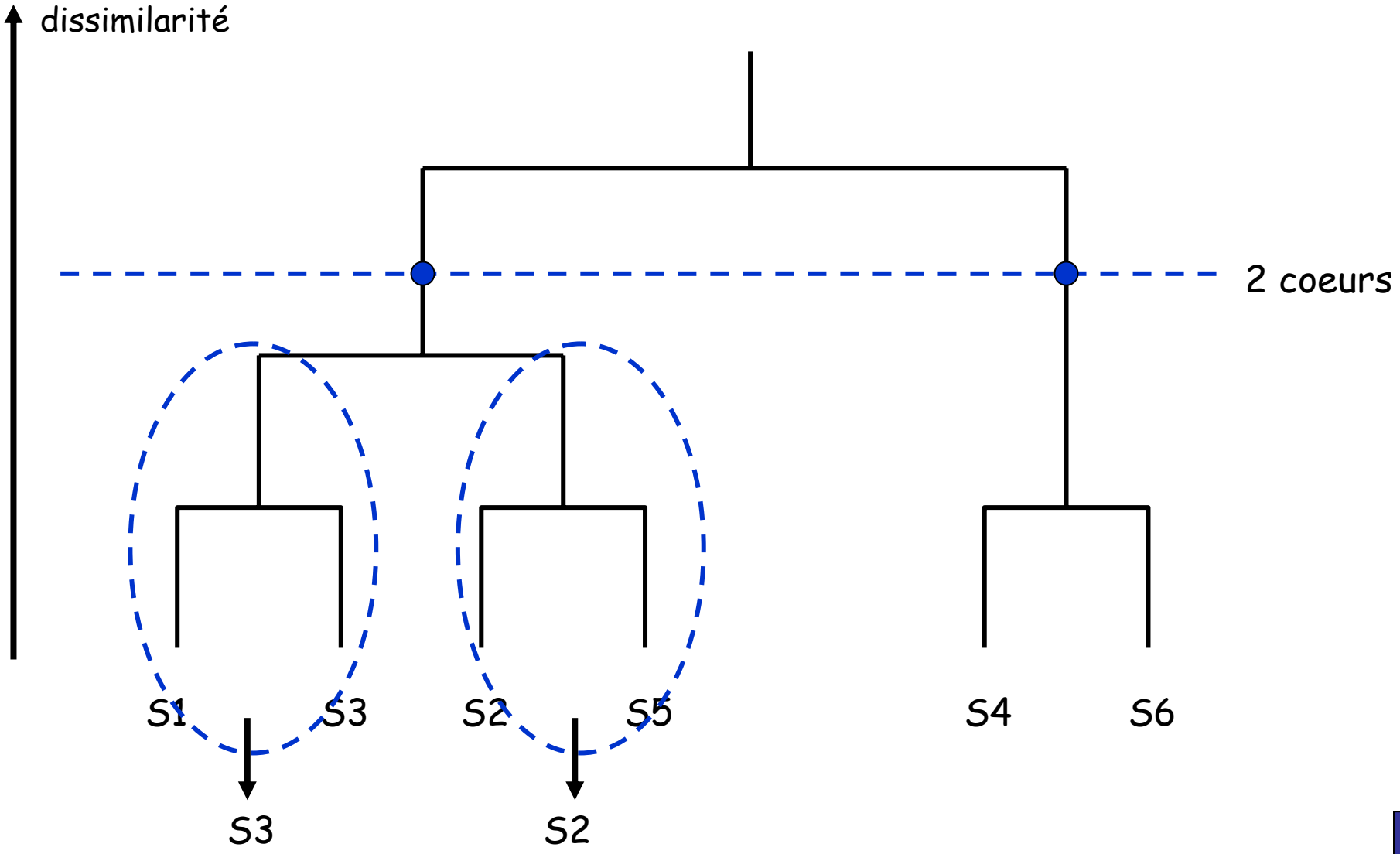
# Extraction des cœurs structuraux
## CAH - Principe

Entrée: Matrice de dissimilarités

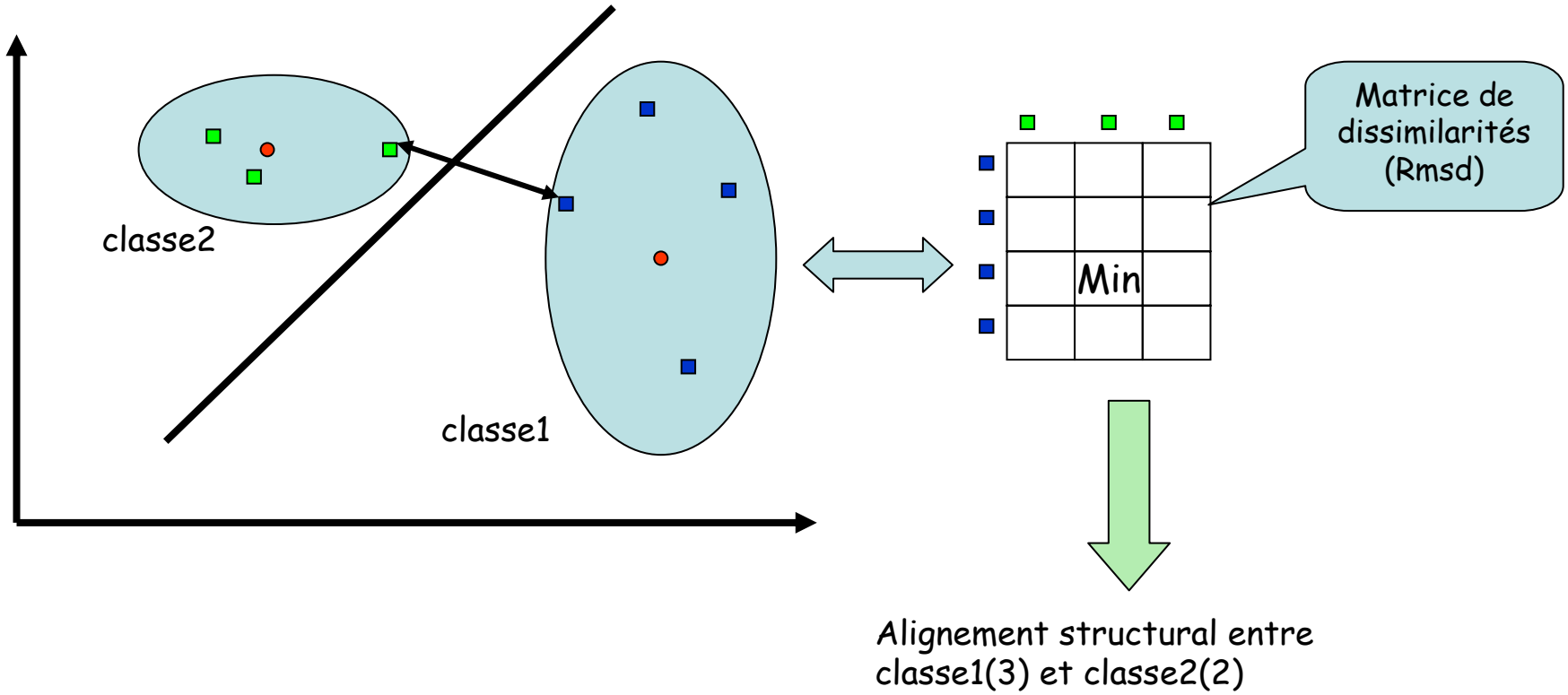
Sortie: Dendogramme



# Extraction des cœurs structuraux
## CAH - Élection



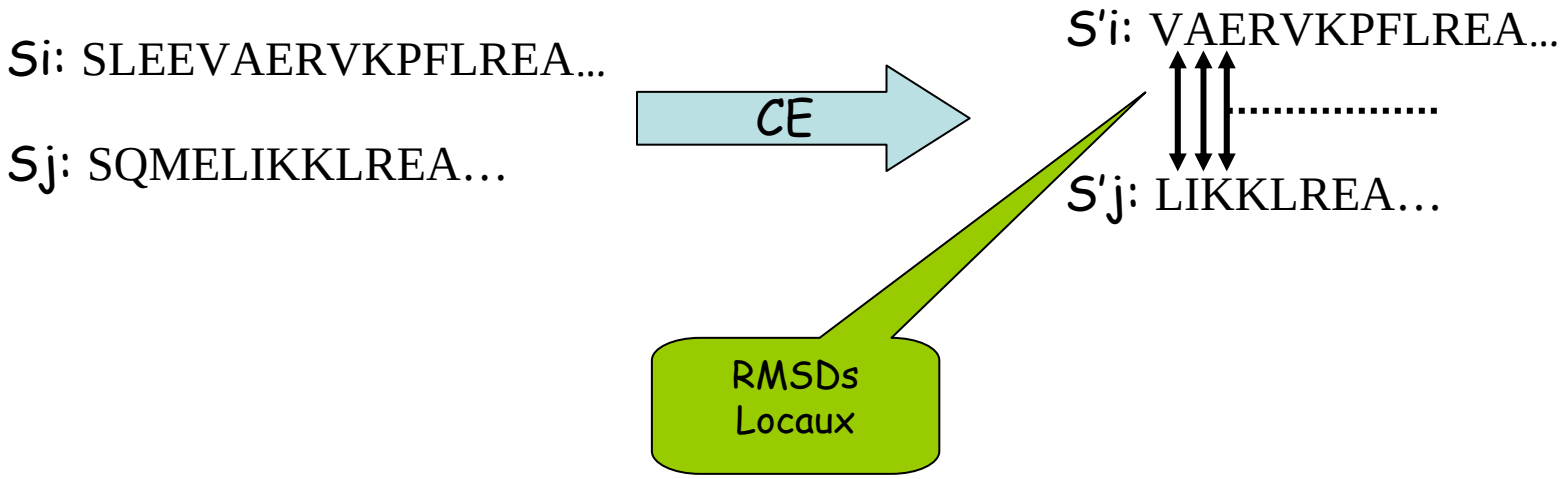
# Extraction des cœurs structuraux
## CAH – Élection - Stratégie



# Extraction des cœurs structuraux
## Sélection de composantes

Entrée: structures élues

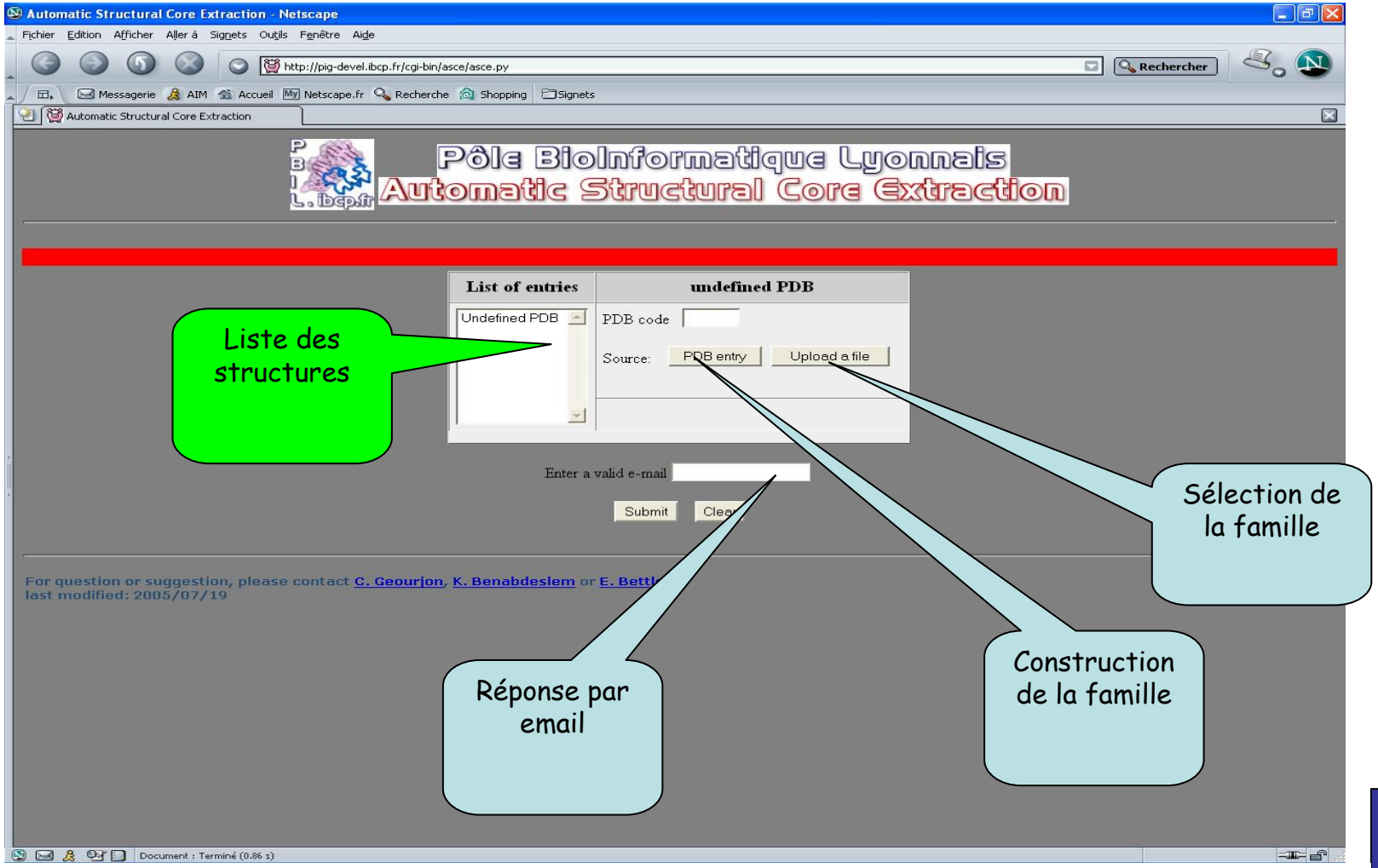
Sortie: Alignement structural optimal



# (Application Web) ASCE:
# Automatic structural cores extraction

**http://pig-pbil.ibcp.fr/cgi-bin/acse**

**C. Geourjon, K. Benabdeslem et E.bettler**



# Extraction des cœurs structuraux
# Exemple

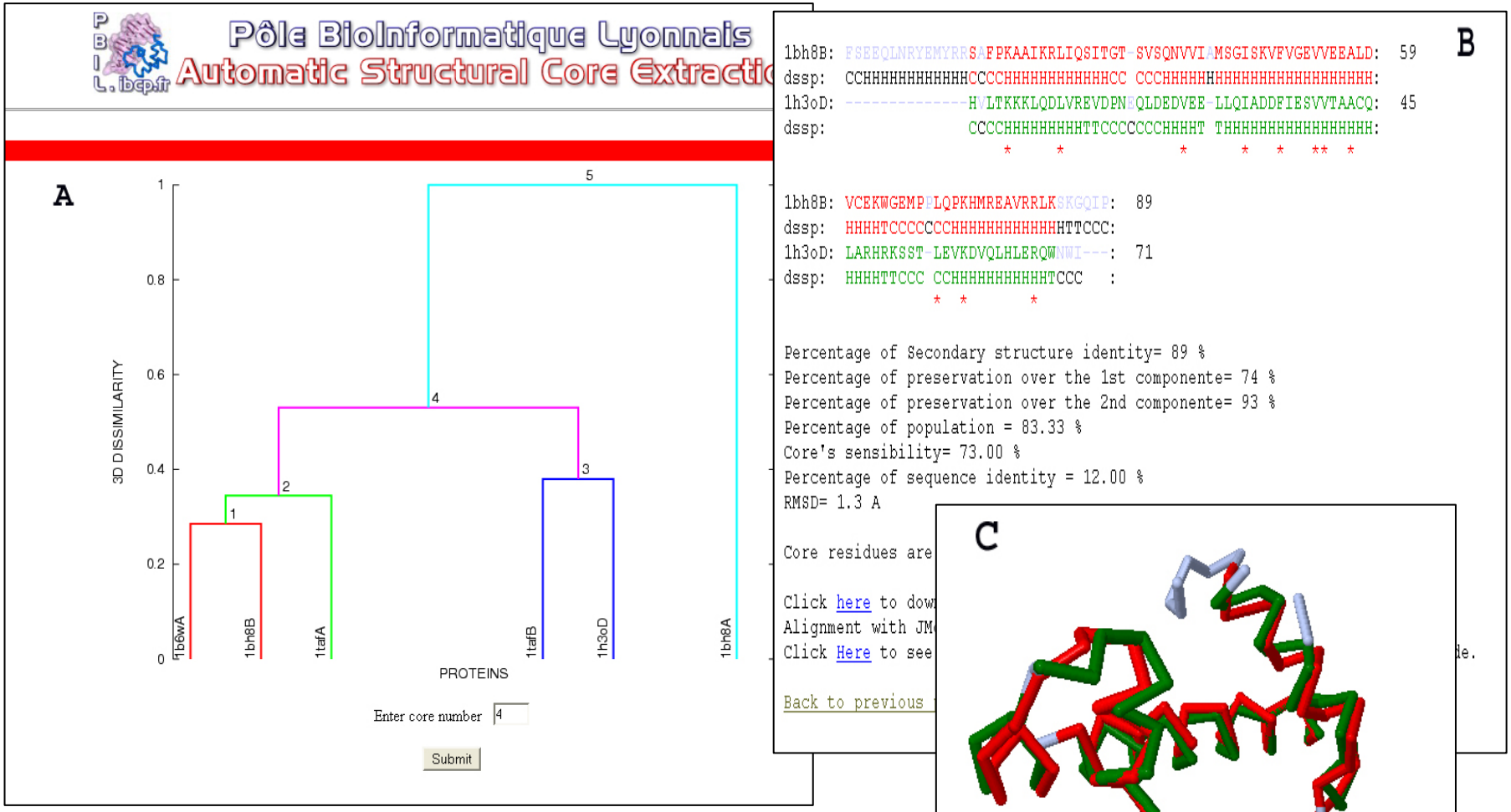



**Fig.1.** General interface of ASCE. (A) Clustering of structures. (B) Core's components with various statistics.
(C) 3D graphic of a structural core

# Système de prédiction

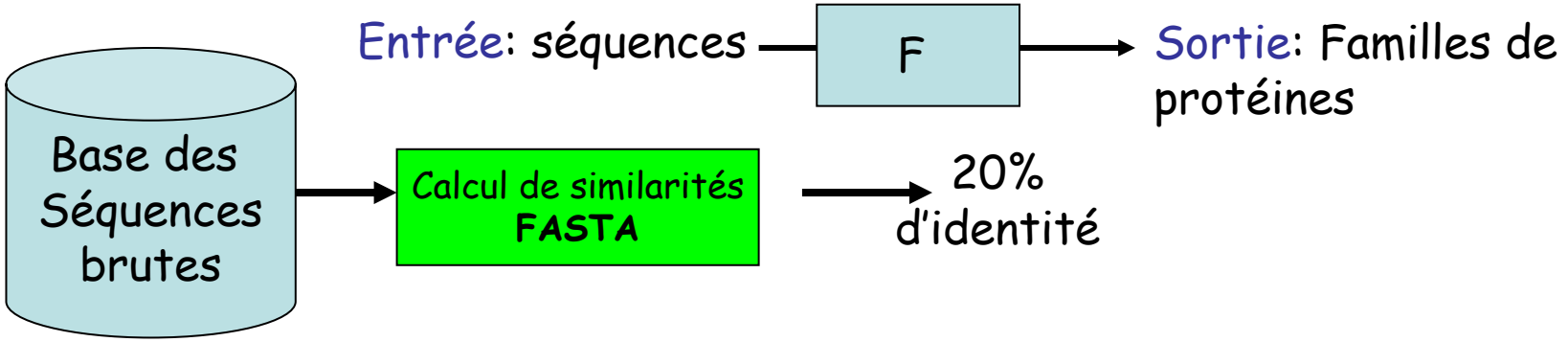
## Démarche

1) Extraction hiérarchique des cœurs structuraux à partir de familles de protéines (**ASCE**: Automatic structural cores extraction)

2) Alignement sur les cœurs structuraux

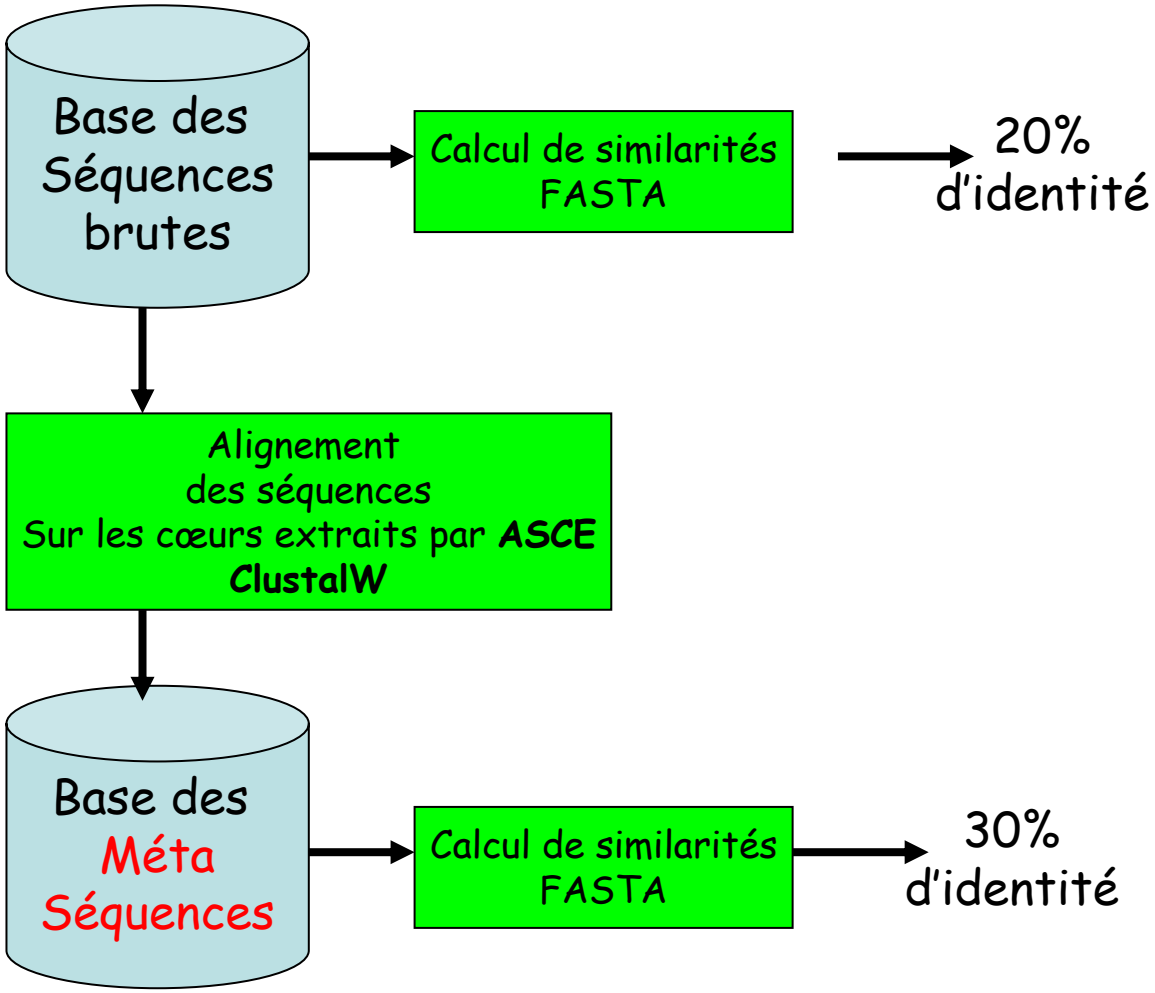
3) Reconnaissance de repliements: Modélisation et apprentissage

# Alignement sur les cœurs



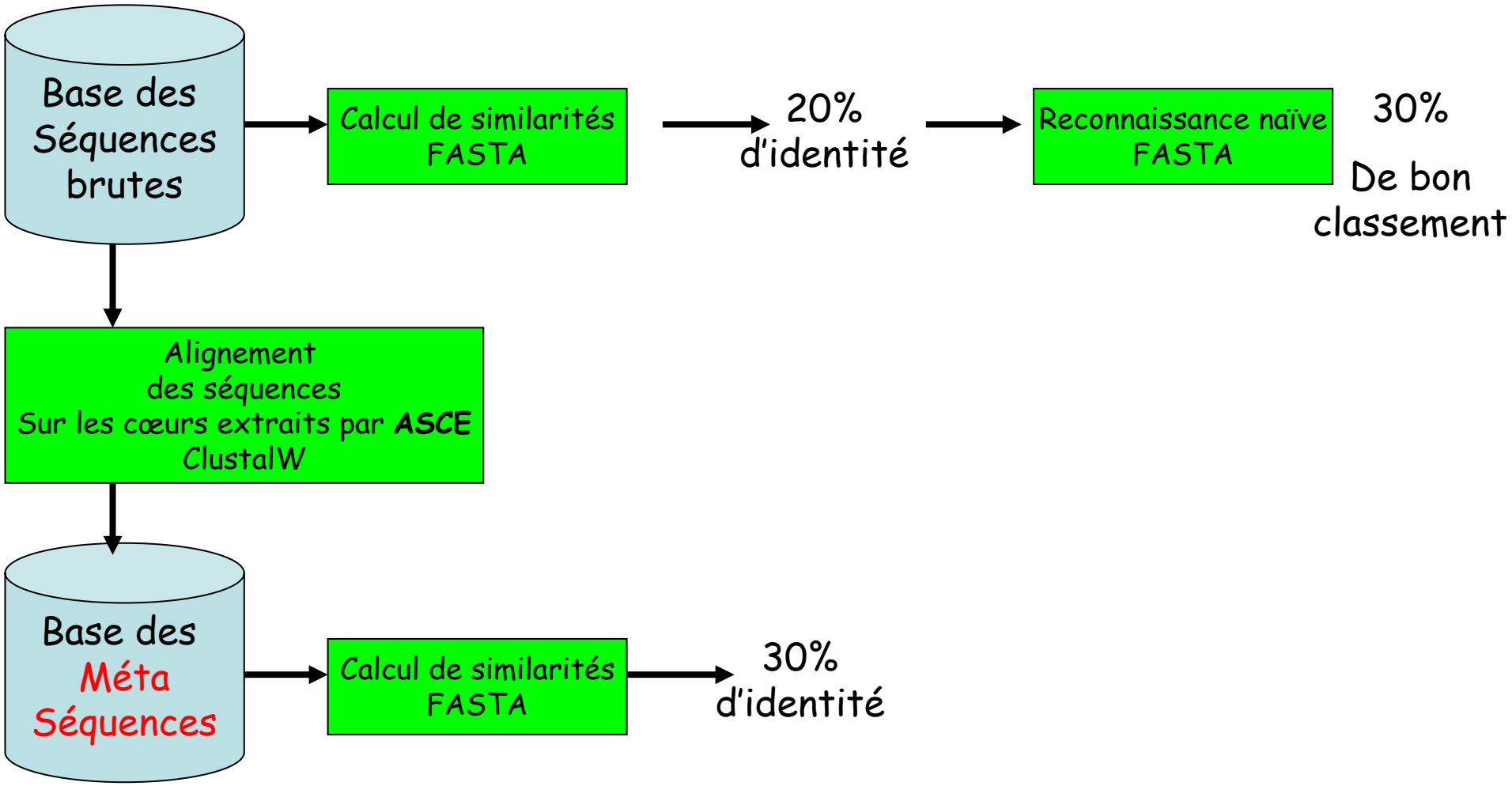

**Pearson W. R. (1990) Rapid and Sensitive Sequence Comparison with FASTP and FASTA, Methods in Enzymology 183, 63 - 98**

# Alignement sur les cœurs

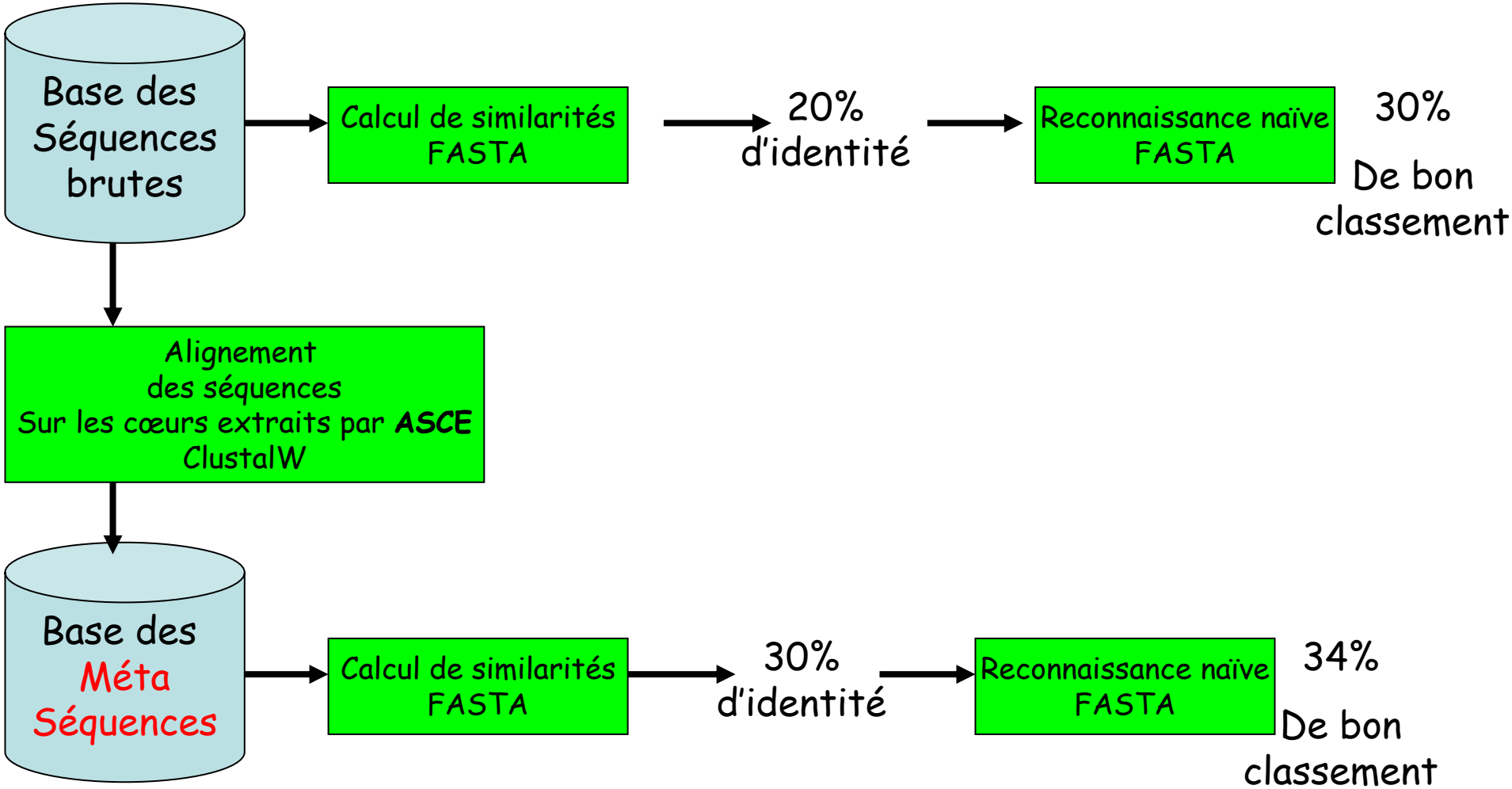



**Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22, 4673-4680**

# Alignement sur les cœurs



# Alignement sur les cœurs



# Système de prédiction

## Démarche

1) Extraction hiérarchique des cœurs structuraux à partir de familles de protéines (**ASCE**: Automatic structural cores extraction)

2) Alignement sur les cœurs structuraux

3) Reconnaissance de repliements: Modélisation et apprentissage

# Discrimination

## Codage matriciel & modélisation neuronale

***Sortie:*** ***Classes***: *Familles de proteines ($F_i$)*

***Entrée:*** X: $x_1, x_2, \ldots x_i, \ldots, x_N$ *(Meta séquence: avec des gaps)*

binaire

N composantes

.....

Longueur de la séquence: *N*

matriciel

**Matrice $COV_X$ modélisant la séquence**

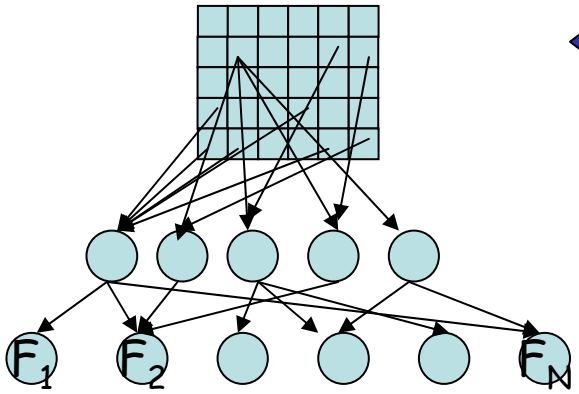
$N \times N$

$$Cov_X = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

dynamique

$$Cov^*_X = \frac{1}{N}\sum_{i=1}^{N}\lambda^{N-i}(x_i - \bar{x})(x_i - \bar{x})^T, 0 < \lambda < 1$$

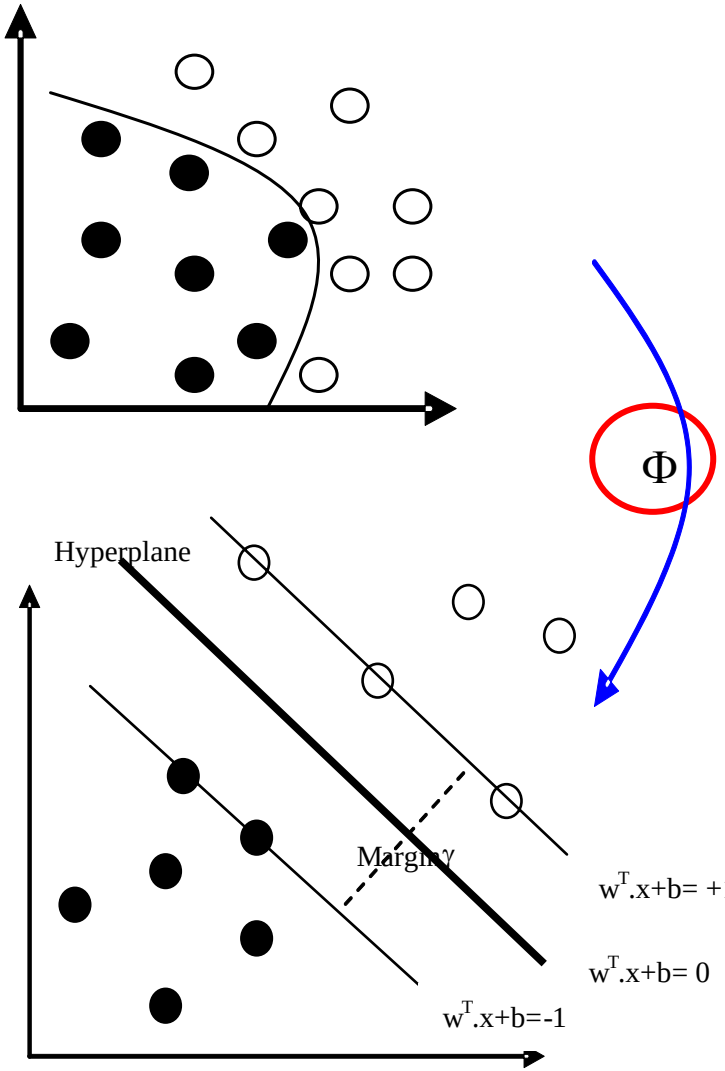
modélisation



# Discrimination

## Méthodes à noyaux (SVM): Principe général

SVM Bi-Classe




### *Données*

*x( un exemple d'apprentissage), X {l'ensemble d'apprentissage}*

*$y_i \in$ {-1, +1} (Les étiquettes des deux classes}*

### *Fonction de décision*

$$\sum_{i=1}^{n} w_i x_i + b = 0 \qquad \text{(HP: } W. \textbf{(1)})$$

$$y(x) = signe(\sum_{i=1}^{p} w_i x_i + b) = signe(\sum_{j=1}^{|X|} \alpha_j y_j k(x, x_j) + b) \qquad \textbf{(2)}$$

### *Apprentissage*

Pour chaque $x_j$ de $X$

$$M(x_j) = y_j(\frac{W}{\|W\|} x_j + \frac{b}{\|W\|}) \qquad \text{La marge de } x_j \text{ . } \textbf{(3)}$$

$$M(X) = \min_j M(x_j) \qquad \text{La marge de } X \text{ . } \textbf{(4)}$$

$$W = \arg MaxM(X)$$

$$Maximiser \sum_j \alpha_j - \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$s.t. \alpha_i \geq 0, \sum_j \alpha_j = 0$$

# Discrimination multi-classes

# Principe

Problèmes multi-classes

Méthode One.vs.All → K classes => K SVMs

Méthode One.vs.One →K classes => K(K-1)/2 SVMs

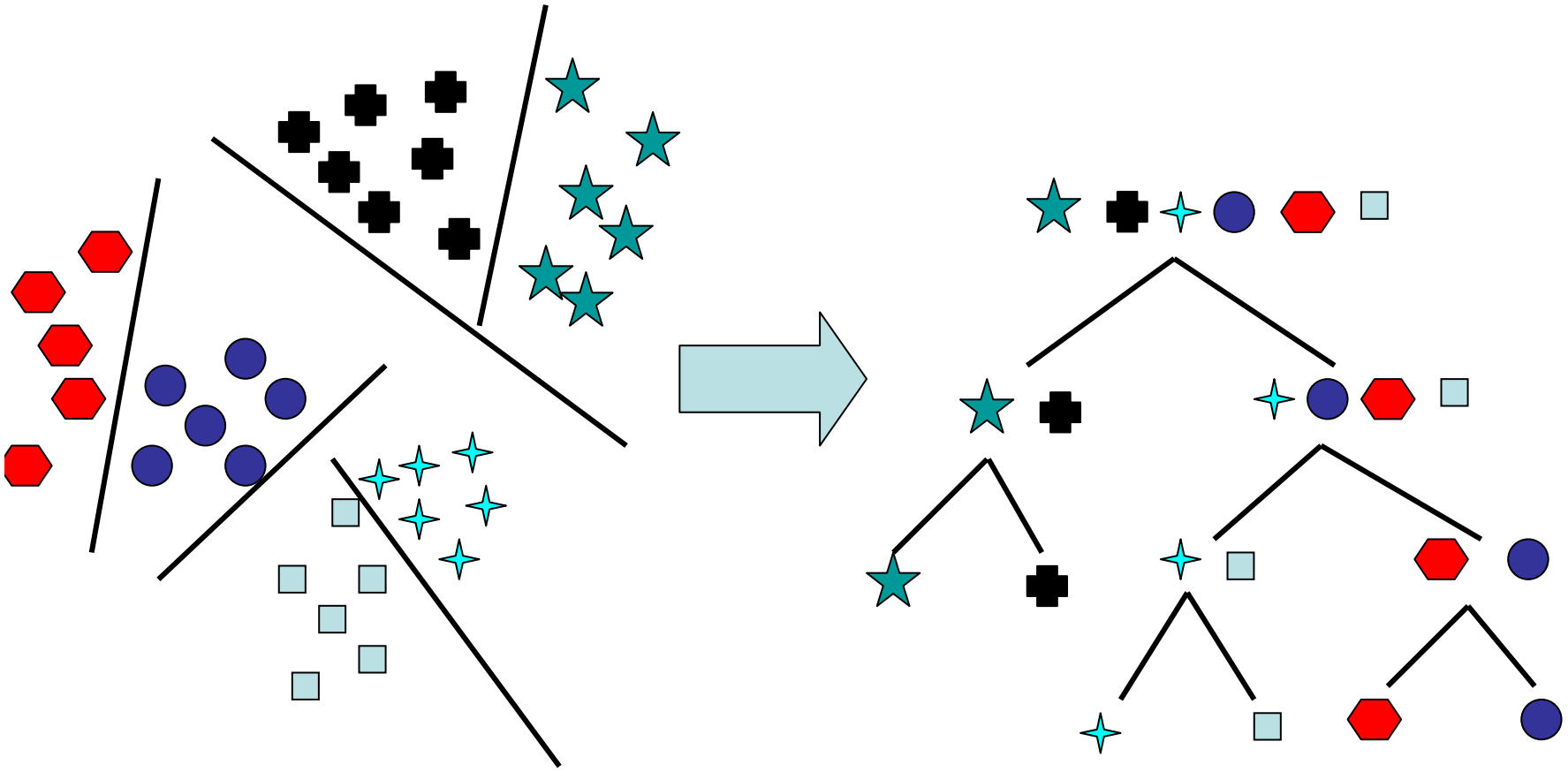
# Discrimination multi-classes

# Principe

Problèmes multi-classes

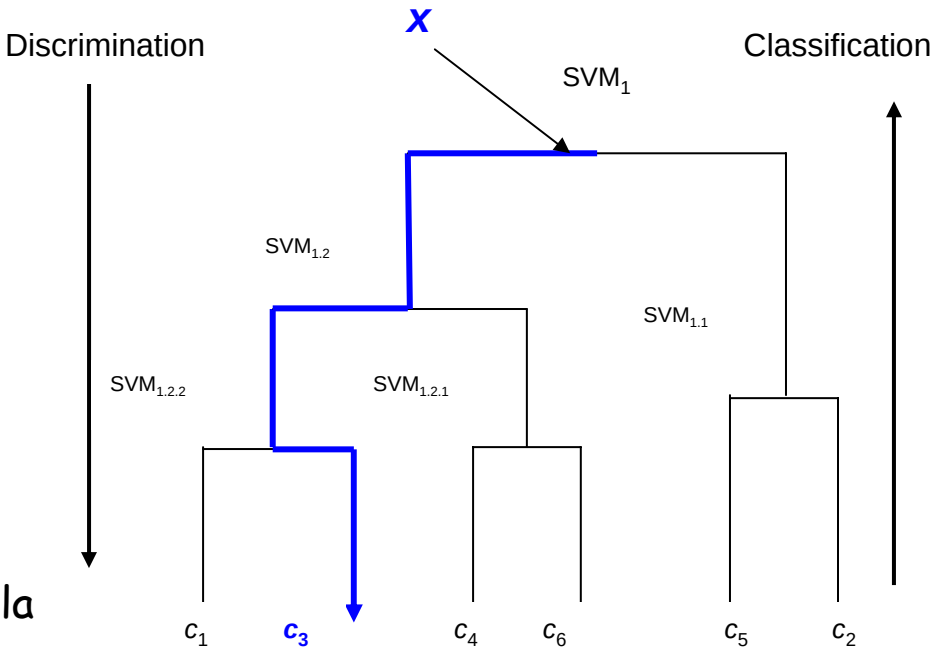
Méthode One.vs.All → K classes => K SVMs

Méthode One.vs.One →K classes => K(K-1)/2 SVMs



# Discrimination multi-classes

## Modèle proposé (CAH+SVM= DSVM)




Décomposition Hiérarchique (DSVM)

Soit une base d'apprentissage $x_1, x_2, \ldots, x_n$ classifiés 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) SVMs sur les nœuds de la taxonomie

*Avantages*

- *Un nombre optimal de SVMs*
- *Une Trace de reconnaissance*
- *Complexité réduite de reconnaissance*

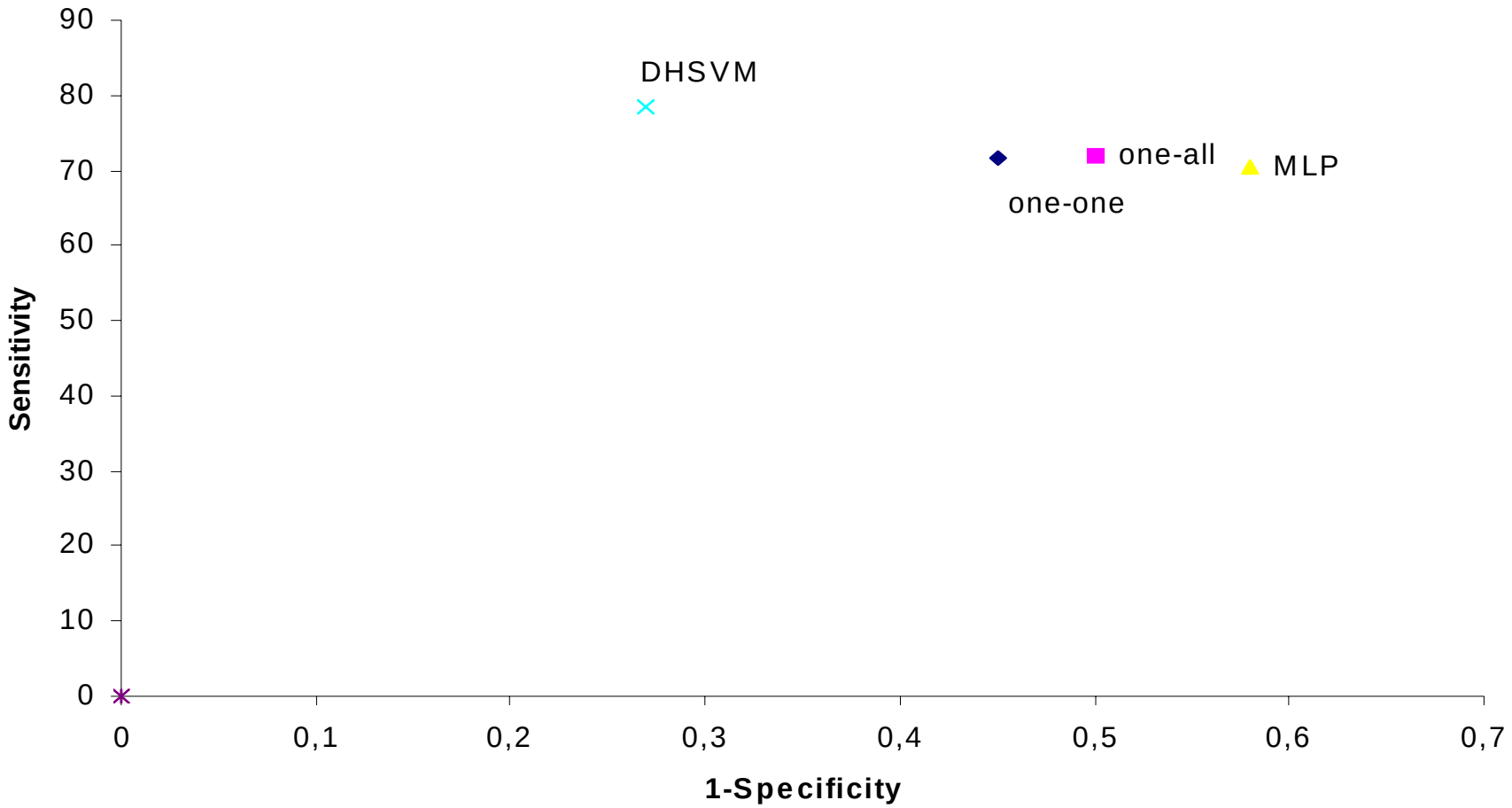
Perspective:

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

# Discrimination

## Résultats

>1000 séquences, 21 × 21 “variables” , 99 classes.



# Conclusion

# Publications et Logiciels

•**K.Benabdeslem**, G.Deléage and C.Geourjon. Bioinformatics."Structural cores extraction for fold recognition improvement". (Soumis)

•**K.Benabdeslem** and Y.Bennani. "Dendogram based SVM for multi-class classification". Journal of Computing and Information Technology (CIT). ( A paraitre)

•**K. Benabdeslem**, G. Deléage et C. Geourjon. BIO-EGC'06."Alignement structural et classification hiérarchique pour l'extraction des cœurs structuraux". Atelier: Extraction et gestion de connaissances appliquées aux données biologiques dans le cadre de la conférence EGC 2006, pp.09-17,Lille, Janvier 2006.

•**K. Benabdeslem**, G. Deléage and C. Geourjon. ECCB'05. "A Neural Network System based on Structural Alignment and Clustering for Proteins Fold Recognition". European conference on Computational biology, pp.85-88, Madrid, September 2005.

•**K. Benabdeslem**, C. Geourjon, Y. Guermeur et N. Sapay. ASTI'05. "Apprentissage automatique: Application à la prédiction de la structure secondaire et tertiaire des protéines". Communication sur invitation présentée dans la session thématique: Bioinformatique II, p.34, Clermont-Ferrand, Octobre 2005.

•**K. Benabdeslem**, G. Deléage and C. Geourjon. JOBIM05. "Cores extraction based Neural Network Model for Proteins fold recognition". Journées ouvertes en biologie, informatique et mathématiques, pp.341-347, Lyon, July 2005.

•Un serveur Web(**ASCE**) d'extraction des cœurs structuraux des protéines a été développé (http://pig-pbil.ibcp.fr/cgi-bin/asce/asce).

•Une première version de **DSVM** pour la discrimination multi-classes est disponible dans (http://www710.univ-lyon1.fr/~kbenabde/)