Vers la prédiction séquence - structure 3D: Extraction des cœurs structuraux

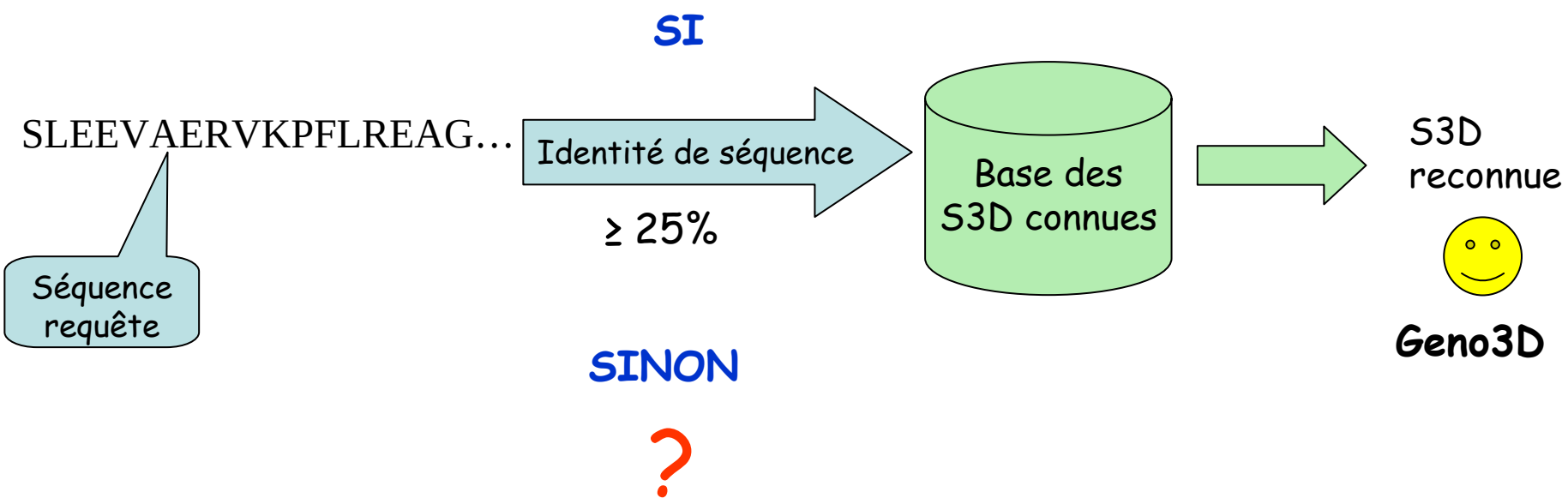
Présenté par
Khalid BENABDESLEM
kbenabde@ibcp.fr

En collaboration avec
Christophe GEOURJON
c.geourjon@ibcp.fr

Projet
GENOTO3D – ACI: Masse de données

Problématique

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

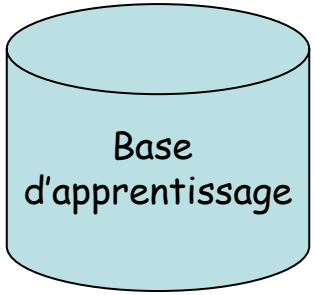


Solution envisagée

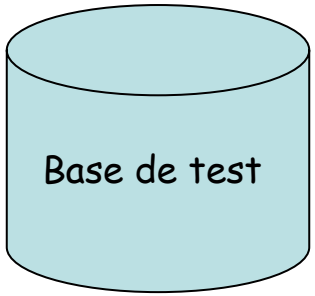
Système prédictif modulaire à base d'apprentissage

Système prédictif modulaire à base d'apprentissage

Principe général



Construction de F



Qualité du modèle

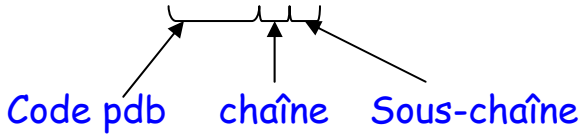
F est construite à partir d'une base d'apprentissage contenant des séquences tels que:

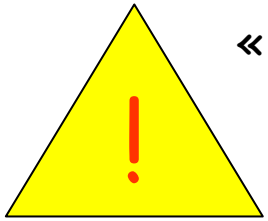
Pour tout $1 \leq i,j \leq N$, $i \neq j$, $\forall Si, \nexists Sj$ / $Si \cong Sj$ (N=nombre total de séquence)

$\geq 25\%$

Données
Bases de données CATH

- Utilisation des bases de données fournies par CATH (Class – Architecture -Topology – Homology)
- Particularité
  - Code à 6 lettres: 1fnnA1





« 1 » peut être plusieurs segments non consécutifs de « A »

Exemple: 1fnnA1

AIVVDDSVFSPSYVPKRYTKDQIFDILLDRAKAGLAEGSYSEDILQMIADITGAQTPLDTNRGDARLAID DSEQS
ILYRSAYAAQQNGRKHIAPEDVRKSSKEVLF

A : Pdb Start = 1 , Pdb End = 17

A : Pdb Start = 192, Pdb End = 275

Construction de F
Démarche

- Extraction des cœurs structuraux
  - Classification CATH
  - 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 la taille de la fenêtre de prédiction
  - -Sélection de variable pour l'optimisation

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
- Enrichissement des cœurs : Alignement de séquences , BLAST (Basic local alignement search tool)

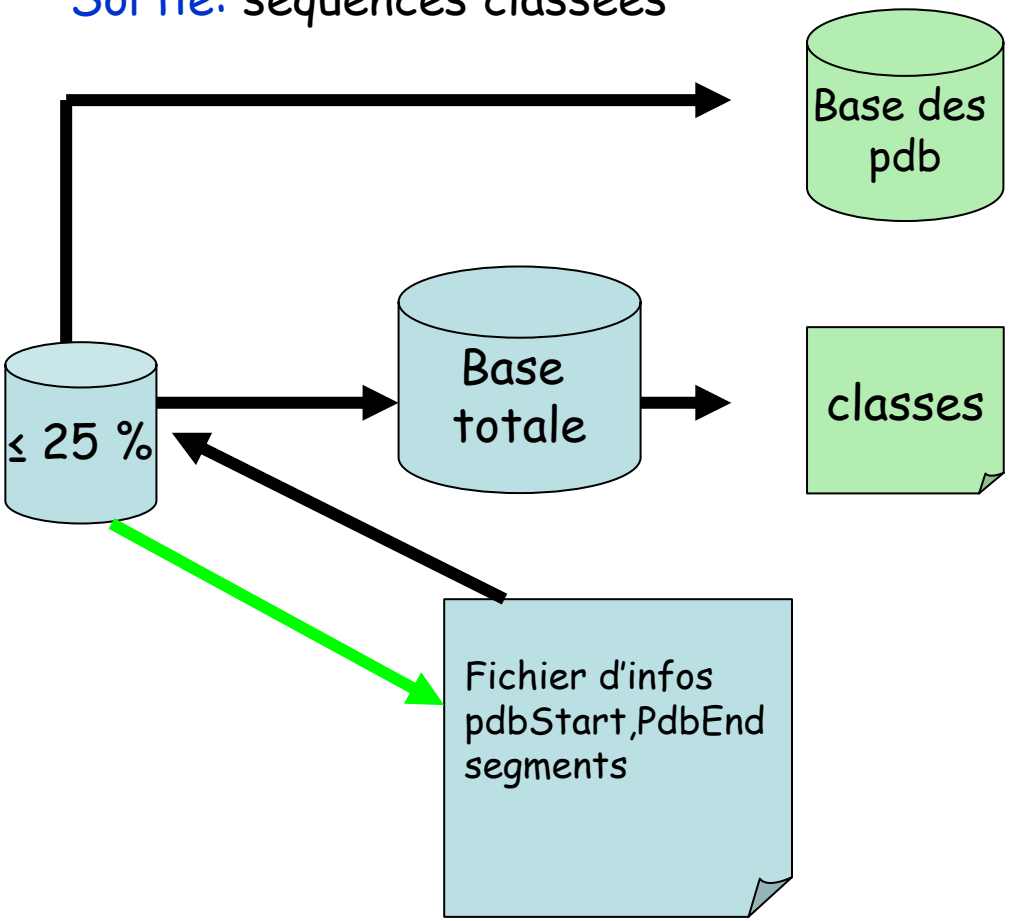
Extraction des cœurs structuraux
Classification CATH (1)

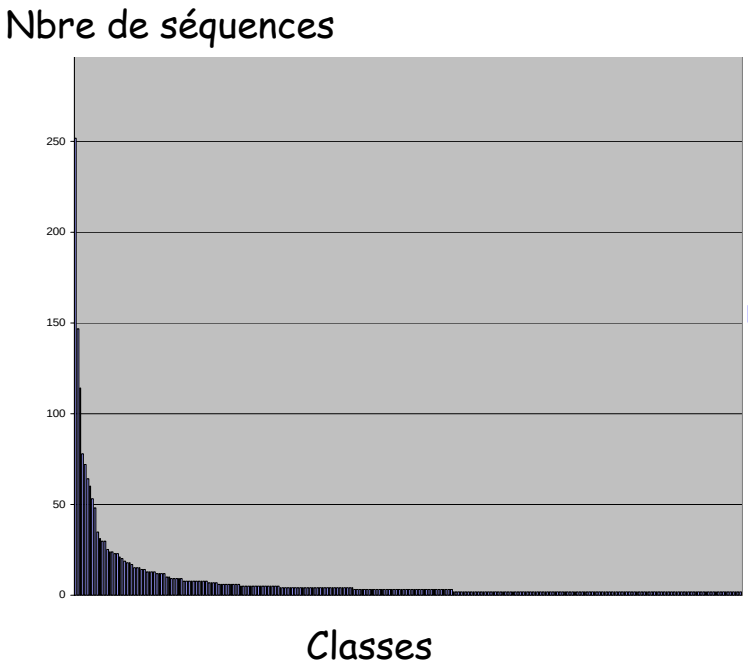
Entrée: Base de séquences de ≤ 25 %

Sortie: séquences classées

~ 3000 séquences

~ 600 classes

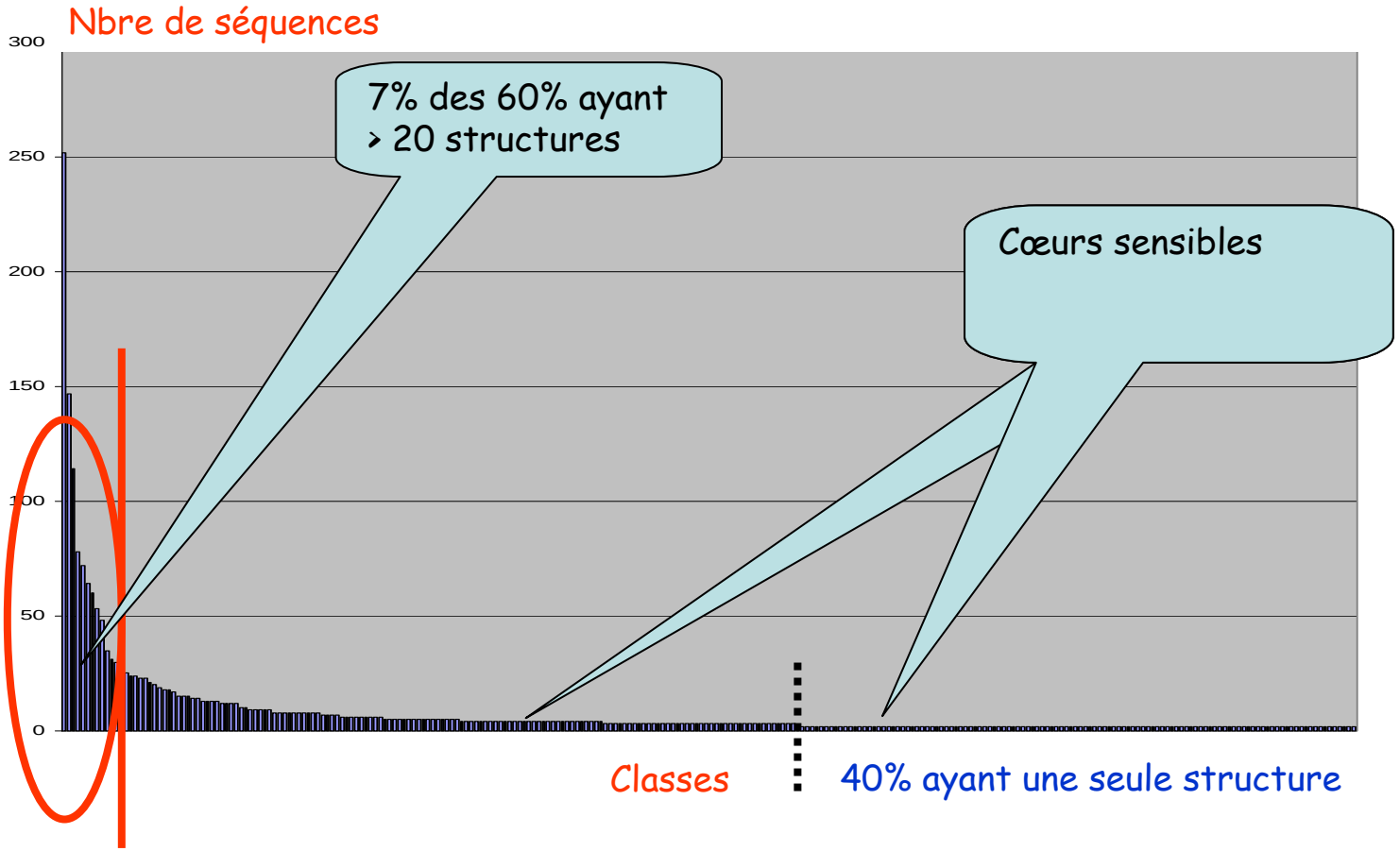




Extraction des cœurs structuraux
Classification CATH (2)

~ 3000 séquences

~ 600 classes

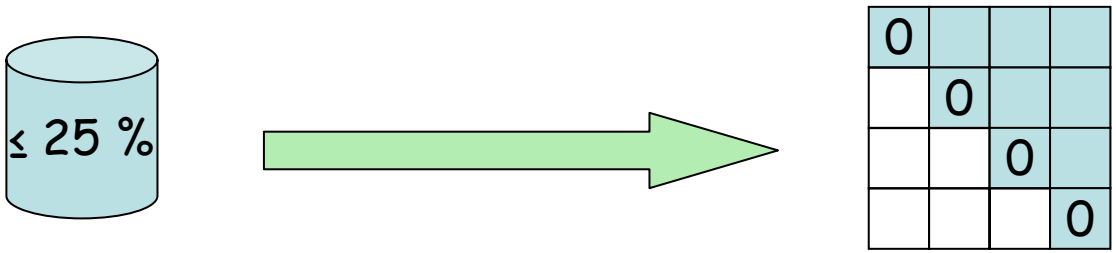


Extraction des cœurs structuraux
Alignement structural

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

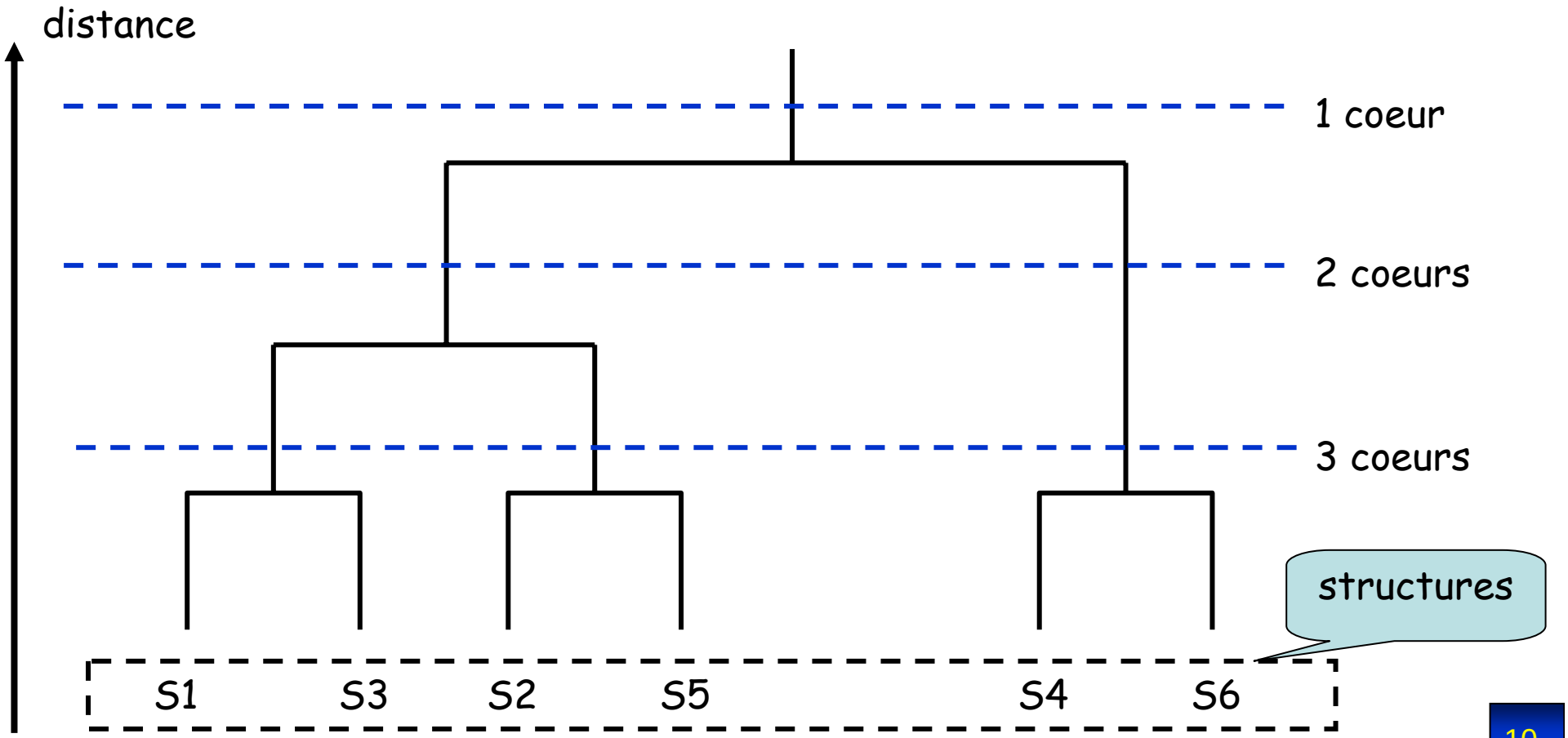
Z-Score < 4.6 Et Rmsd(Si,Sj) ≤ 2 Å

Extraction des cœurs structuraux

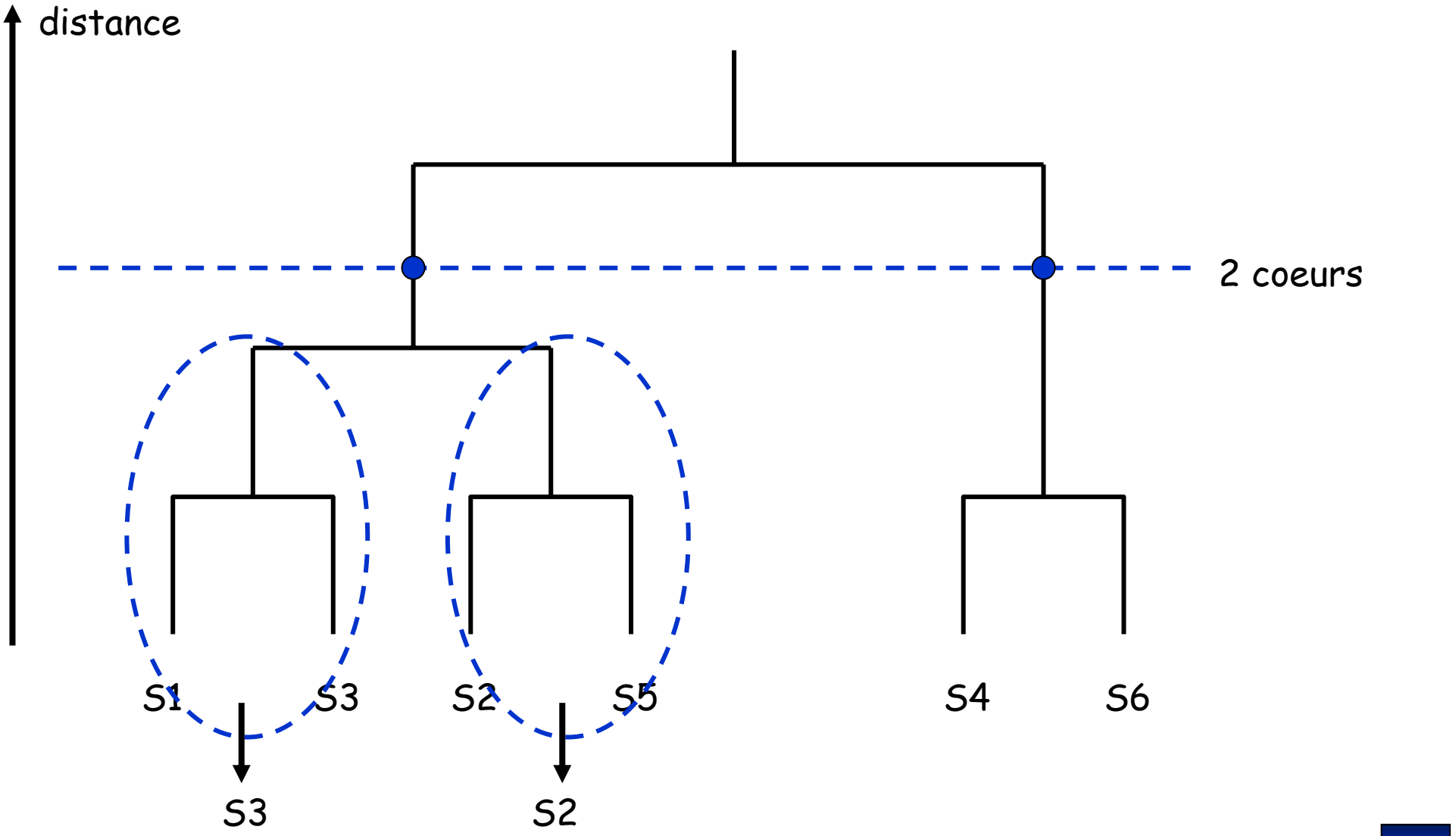
CAH - Principe

Entrée: Matrice de dissimilarités

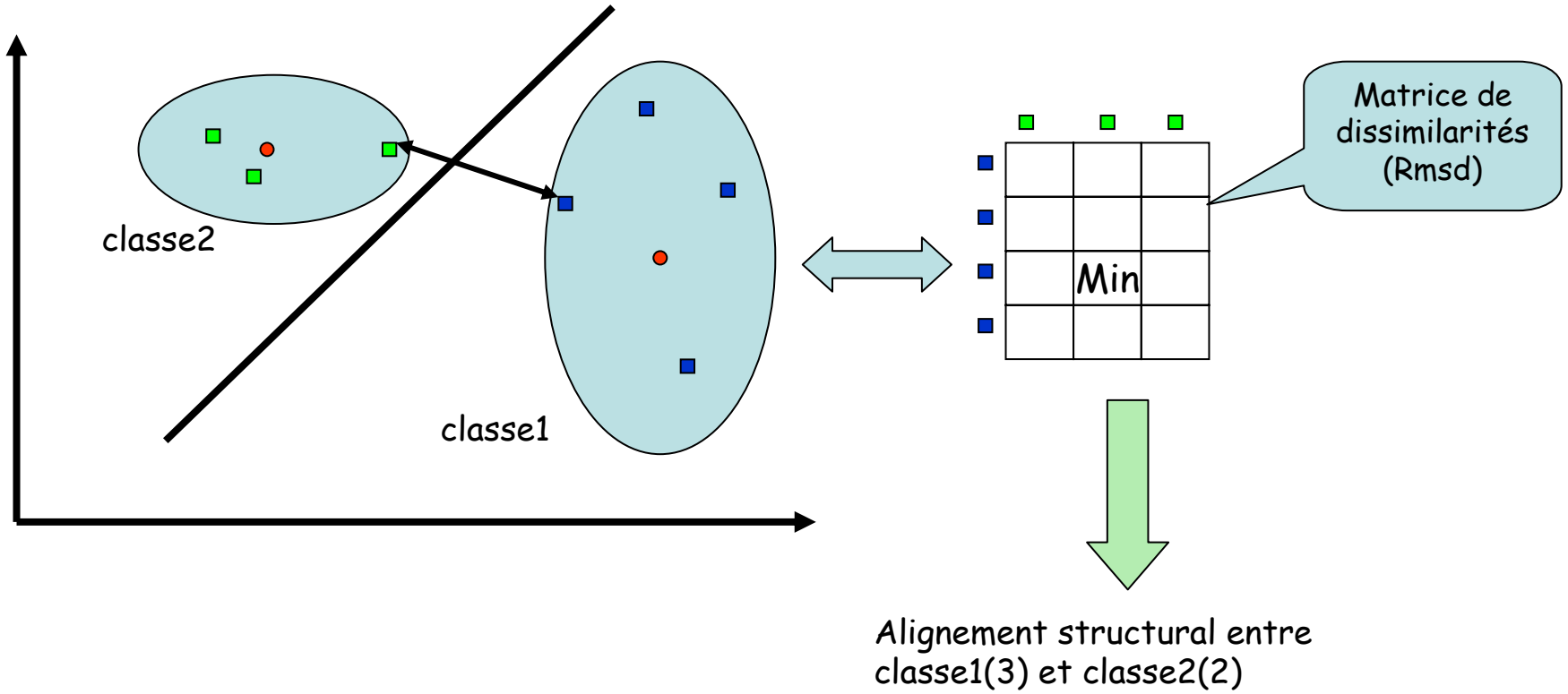
Sortie: Dendogramme



Extraction des cœurs structuraux
CAH - Élection



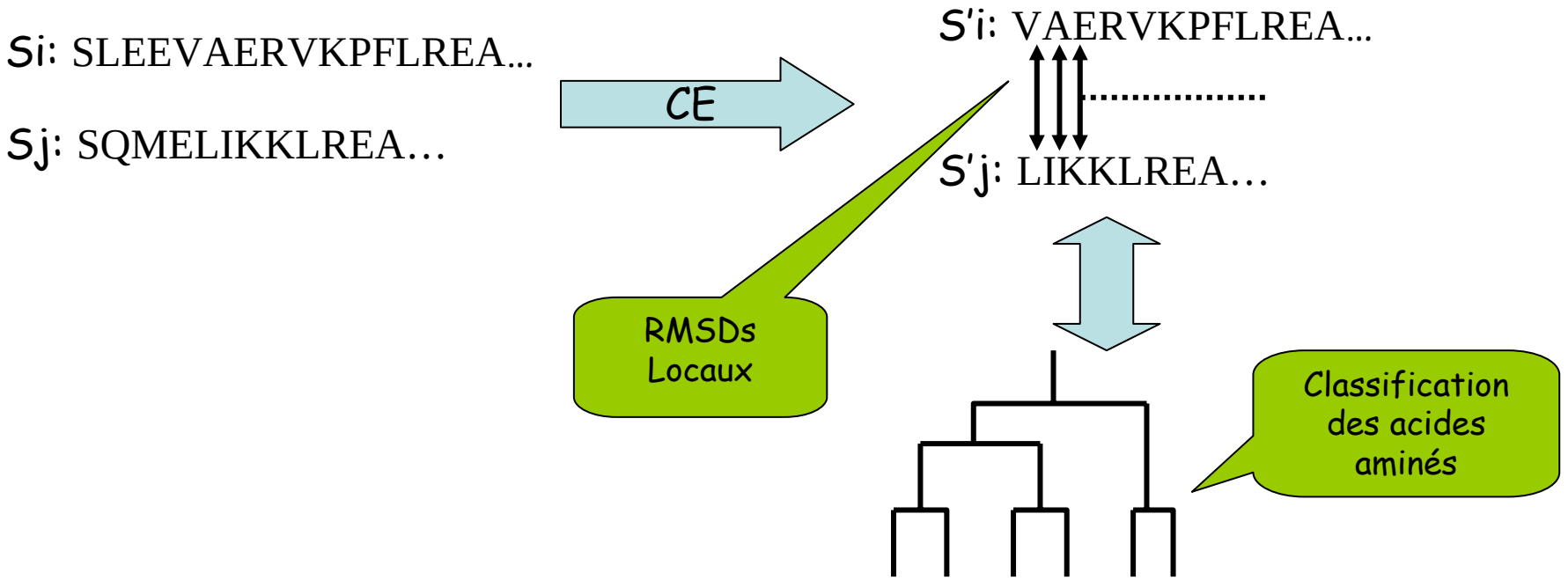
Extraction des cœurs structuraux
CAH – Élection - Stratégie



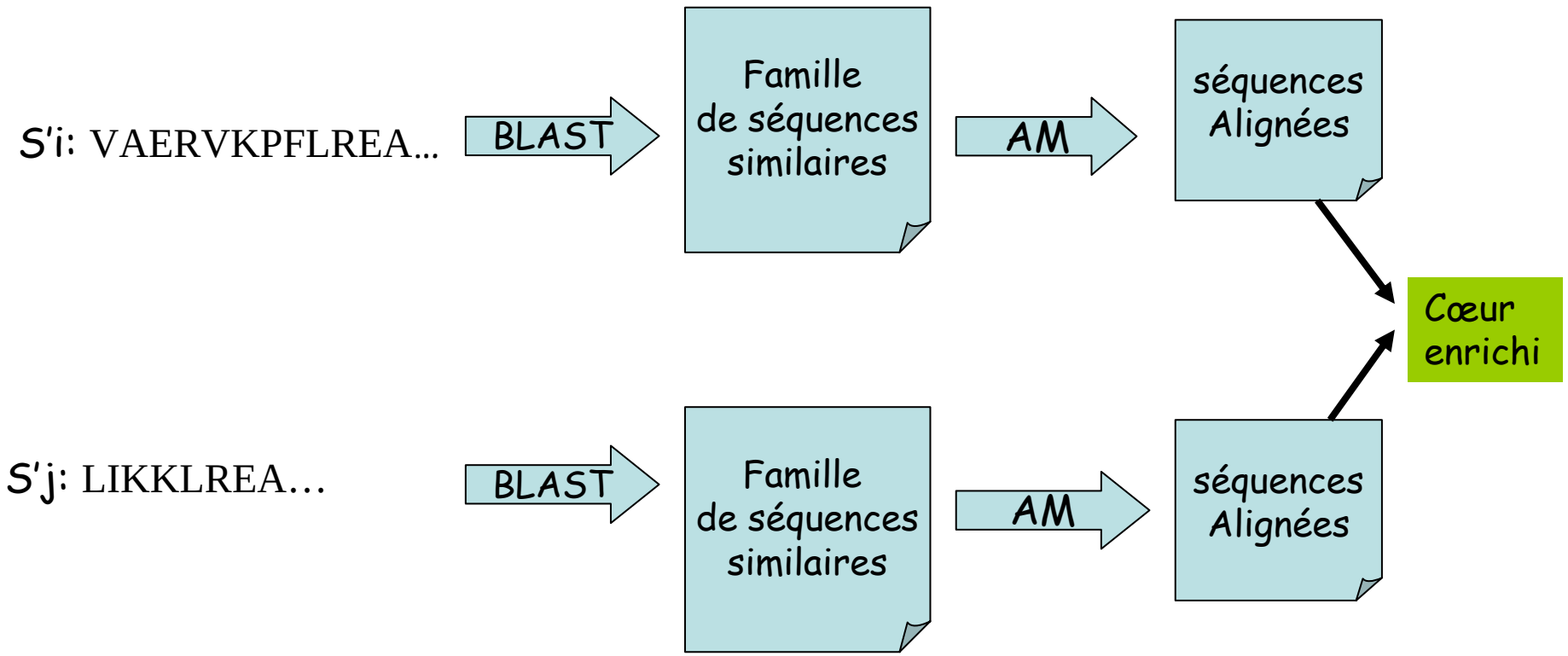
Extraction des cœurs structuraux
Sélection de composantes

Entrée: structures élues

Sortie: Alignement structural optimal



Extraction des cœurs structuraux
Enrichissement des coeurs



AM: Alignement multiple

BLAST: Basic local alignement search tool

Bilan & Perspectives

- Processus d'extraction automatique de cœurs structuraux à partir de familles de protéines.

- Modélisation à partir des cœurs structuraux
  - Modèles Markoviens
  - Modèles neuronaux
  - Machines à noyau

- Traitement de la dynamique
  - Heuristique de détermination de la taille de la fenêtre de prédiction
  - Sélection de variables pour l'optimisation