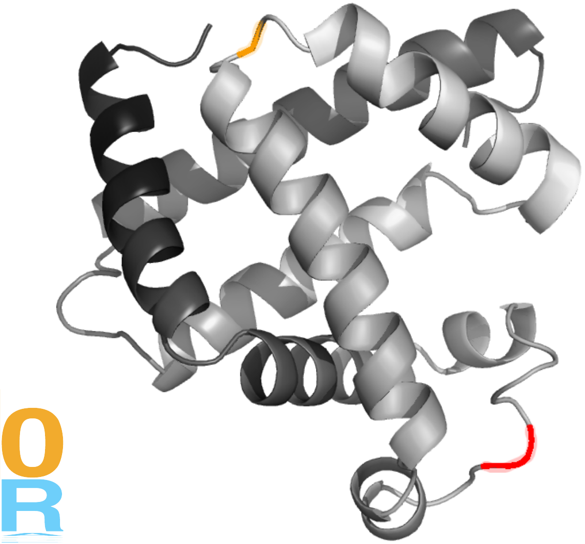
Prediction of patterns from protein primary sequence through structural alphabet

Christelle REYNES,
Leslie REGAD,
Robert SABATIER,
Anne-Claude CAMPROUX



JOBIM MONTPELLIER 2010

7-9 september 2010

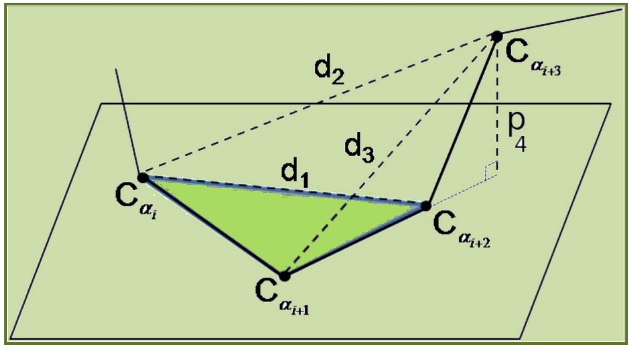
Pattern prediction: context

Objective : locate in an amino-acid sequence a pattern of interest (fonctional, turn,…) previously identified through HMM-SA (HMM-Structural Alphabet, Camproux *et al.*, 2004).

HMM-SA : alphabet made of 27 structural letters

➔ allows simplification of the space of all 4-amino-acid fragment possible conformations

Ex.:



argg

encoding → S

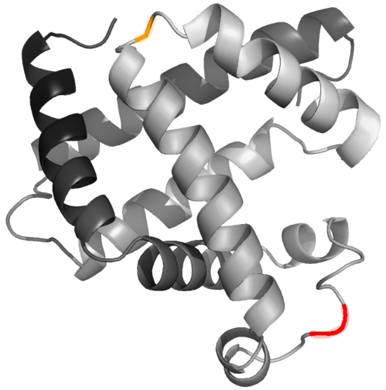
Pattern prediction: context

Objective : locate in an amino-acid sequence a pattern of interest (fonctionnal, turn,…) previously identified through HMM-SA (HMM-Structural Alphabet).

HMM-SA : alphabet made of 27 structural letters

➔ allows simplification of the space of all 4-amino-acid fragment possible conformations

Ex.:



SRTBGGDCAUHSTYUODBBBA
AAAAAAaaAAAaaHHFUSTOBB
BBBDUDO….

➔ simply describes in 1D the 3D conformation of a protein

Pattern prediction: context

Objective : locate in an amino-acid sequence a pattern of interest (fonctionnal, turn,…) previously identified through HMM-SA (HMM-Structural Alphabet).

HMM-SA : alphabet made of 27 structural letters

➔ allows simplification of the space of all 4-amino-acid fragment possible conformations

➔ simply describes in 1D the 3D conformation of a protein

Patterns identification: study of pattern exceptionality and link with annotation databases (see Regad *et al.*)

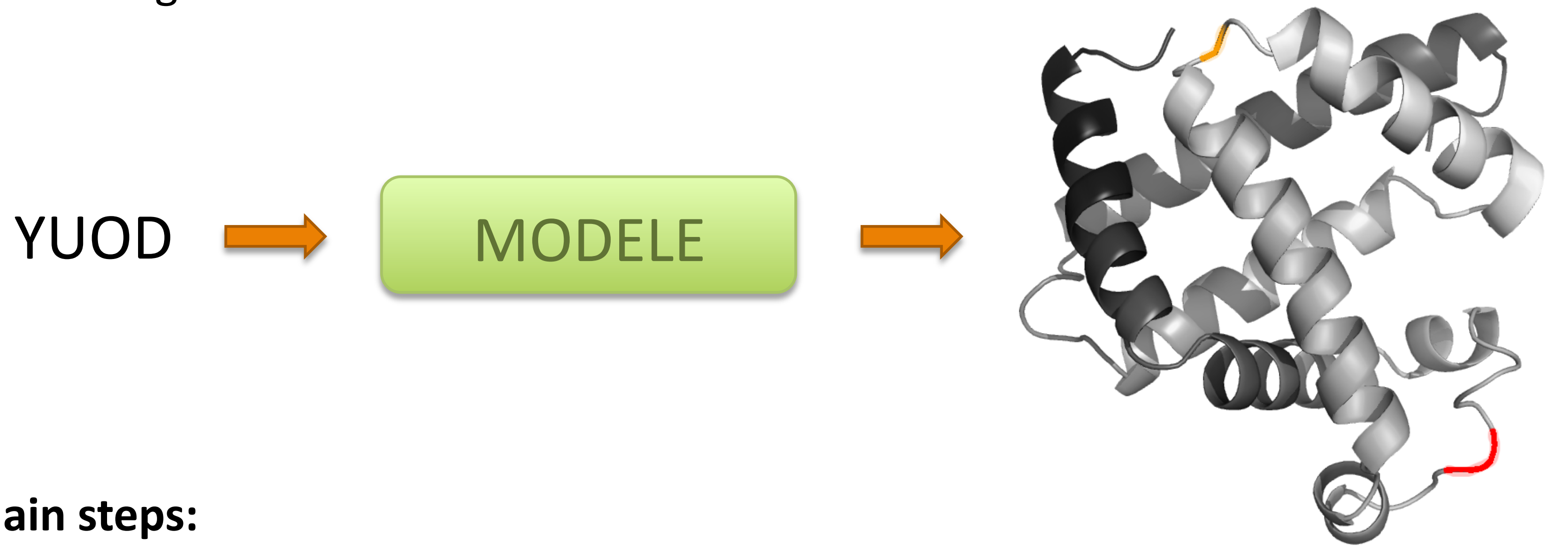
Selection of patterns likely to be predicted directly from amino-acid sequence: sequence specificity required

Design of a prediction method

Prediction method principle

Objective :

Locating from protein sequence zones likely to be encoded into a pattern identified as interesting :



YUOD → MODELE →

Two main steps:

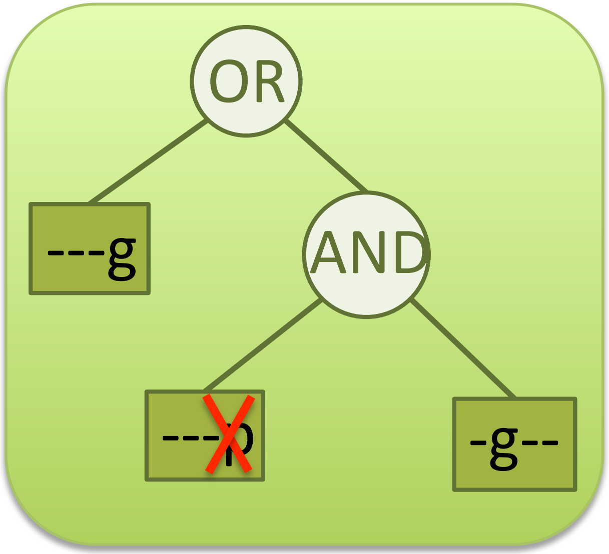
- finding correspondances between single structural letters (SL) and amino-acids (aa)
- construction of a pattern specific Hidden Markov Model

Prediction method principle

First step: single structural letters prediction (pattern independent)

From a database, 1 vs 1 comparison of aa sequences associated to a given SL thanks to boolean functions (qualitative variables)

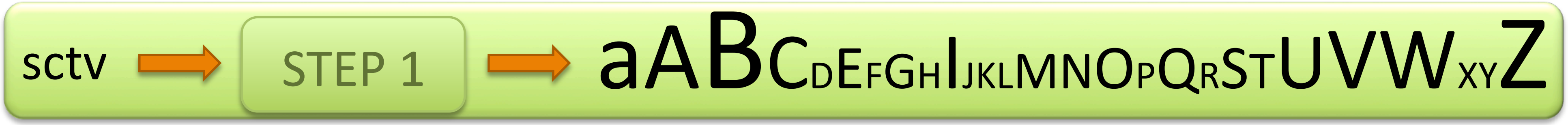
Link modeling seen as a tree:



One tree for each SL couple

➔ 351 (27 x 26 /2) trees to be optimized

Optimization by genetic programming (mutation, crossover, selection)

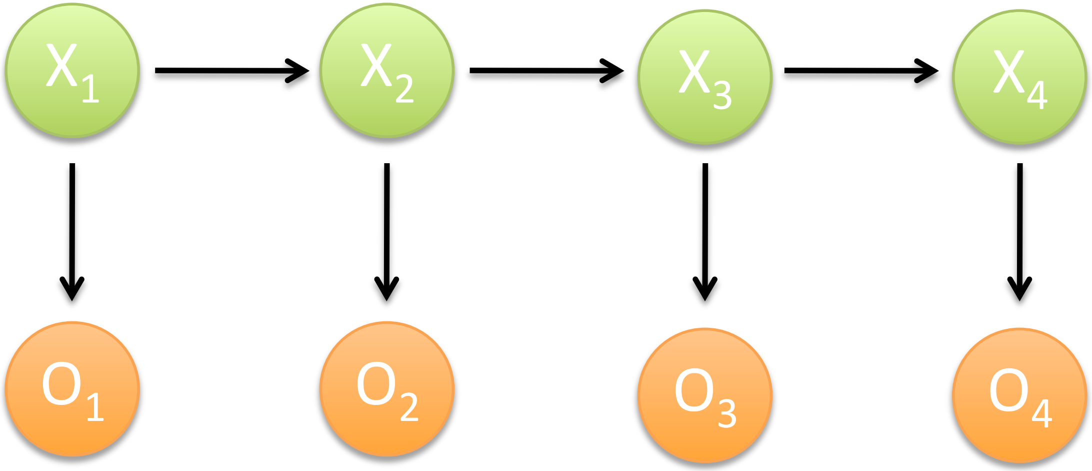


sctv → STEP 1 → aABCDEFGHIJKLMNOPQRSTUVWXYZ

Prediction method principle

Second step: construction of a pattern specific Hidden Markov Model

Problem: no simple link between aa an SL ➔ additional information required



hidden states:
« true » SL

observed states:
step 1 outputs

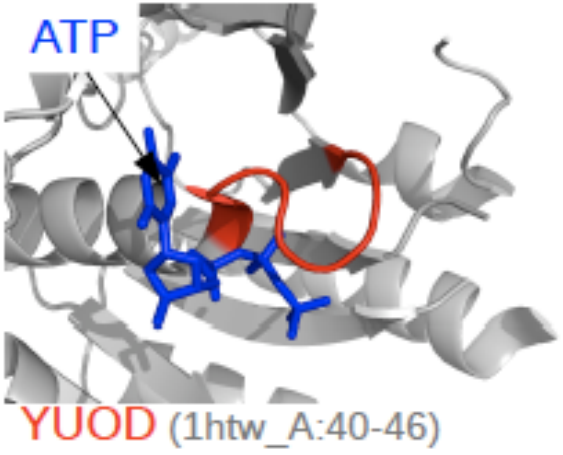
with $O_i = \left(o_i^1, o_i^2, \ldots, o_i^{351}\right)$ and $X_i \in \{a, A, B, \ldots, Z\}$

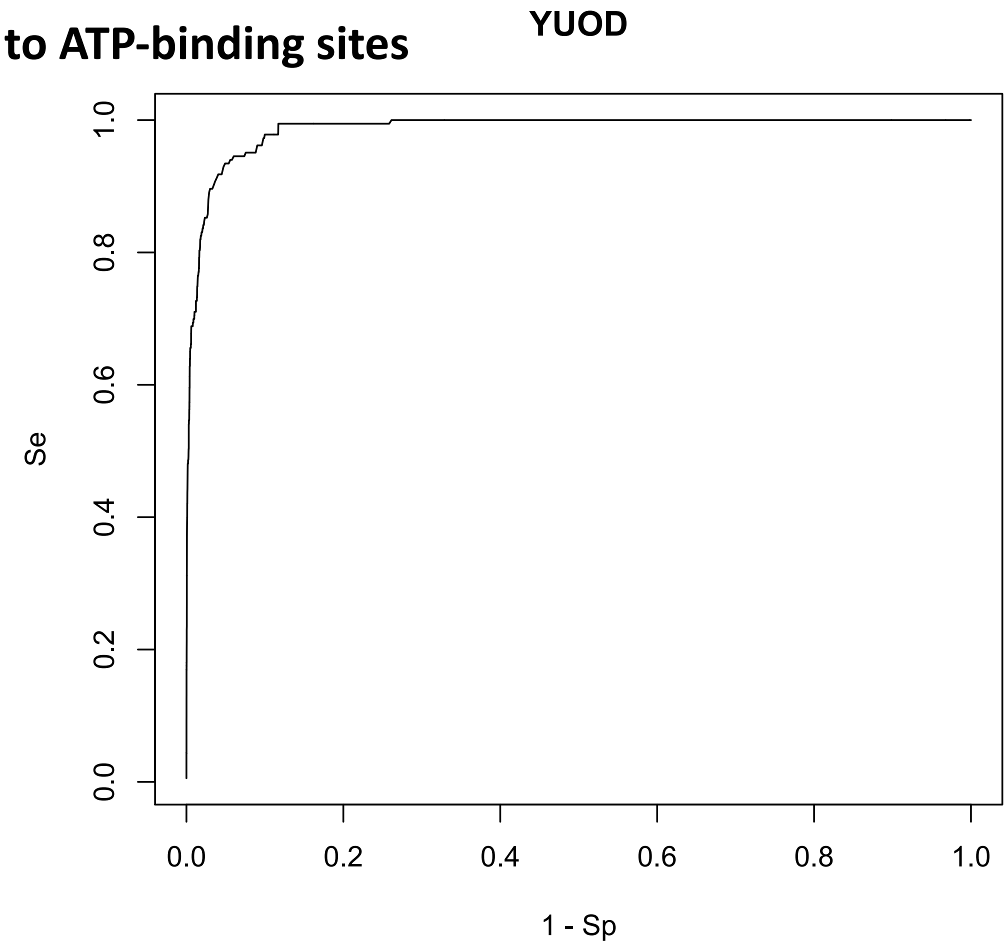
➔ computation of the probability to « **really** » find the pattern of interest given the SLs « **observed** » after step 1.

Application to a functional pattern

YUOD prediction: a pattern associated to ATP-binding sites

data: 181 proteins containing at least 1 YUOD

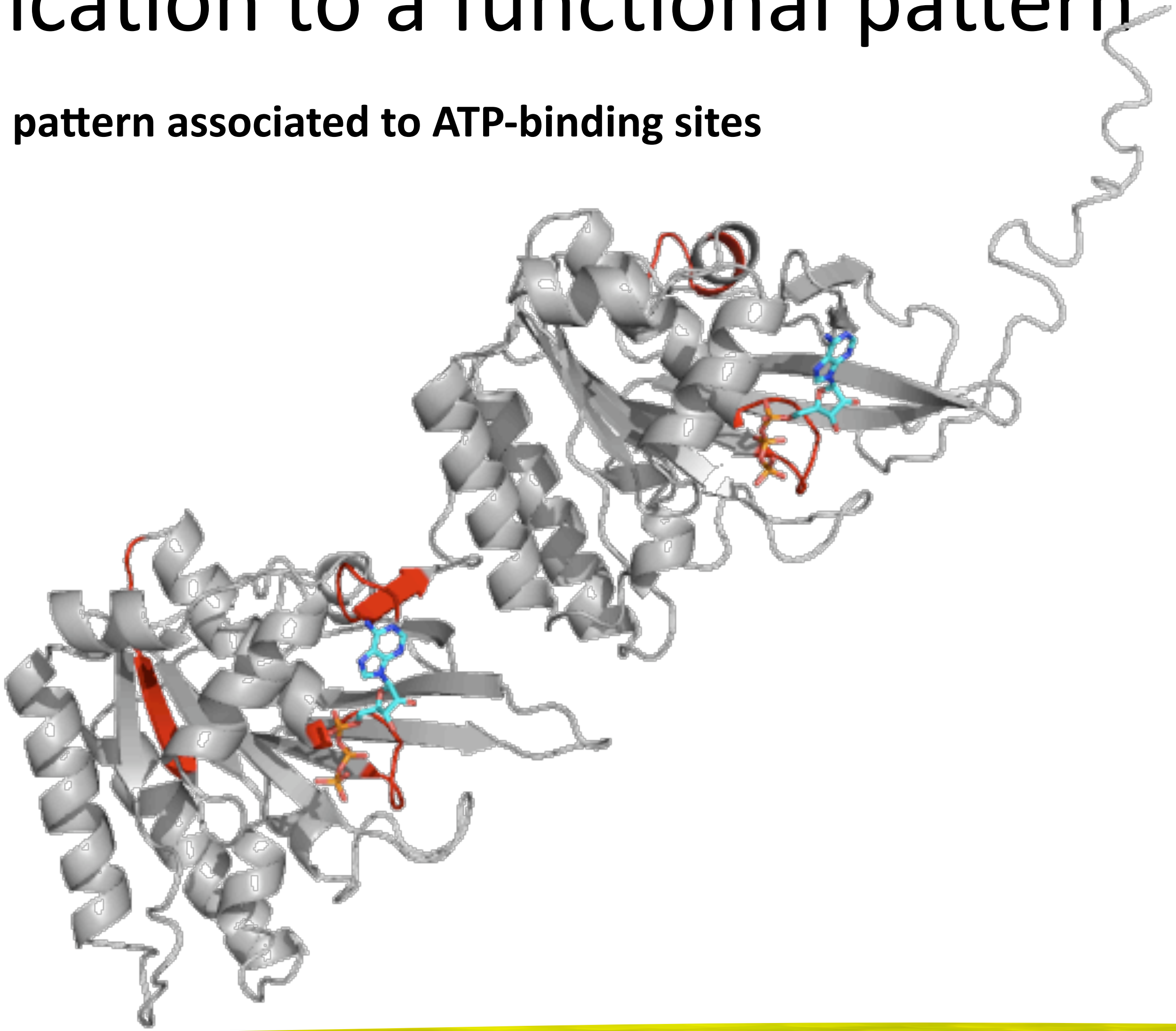




courbe ROC (AUC=0.9866)

Application to a functional pattern

YUOD prediction: a pattern associated to ATP-binding sites



Conclusion and perspectives

A powerful method to identify structural patterns directly from amino-acid sequences

Much information is extracted from data (dependencies between aa and SL, strength of this dependencies, structural dependencies between successive letters)

High adaptativity of the method to new patterns and alphabets

Limitation to patterns having sequence specificity

POSTER 20