Apprentissage d'automates sur les protéines

Approche par fusion de fragments significativement similaires (Jobim'04)

François Coste, Goulven Kerbellec, Boris Idmont, Daniel Fredouille
Christian Delamarche

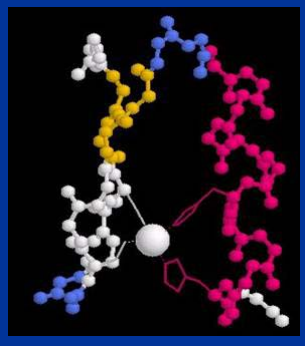
ACI Genoto3D, 22 juin 2004

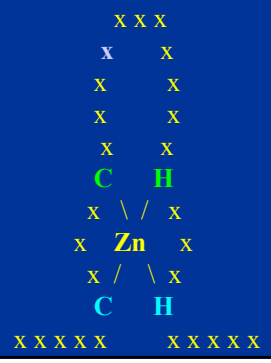
Caractérisation d'une famille de protéines

Exple : Protéines « en doigt de zinc »

Z	...YLGPLN**C**KS**C**WQK**F**DSFSKCHD**H**YLCR**H**CLNLLL...
ZFH2	...ILM**C**FI**C**KLS**F**GNVKSFSL**H**ANTE**H**RLNL...
ZNF236	...HK**C**EI**C**LLSFPKESQFQR**H**MRD**H**E...

C-x(2,4)-**C**-x(3)-[LIVMFYWC]-x(8)-**H**-x(3,5)-**H**





Motif C2H2 pour la famille Zinc finger

C-x(2,4)-**C**-x(3)-[LIVMFYWC]-x(8)-**H**-x(3,5)-**H**

- 416 séquences Zinc finger
- motif C2H2 contenu dans :
  - ✓ 372 Zinc finger
  - ✓ 34 protéines non Zinc finger
  - ✓ 6 protéines candidates Zinc finger

Caractérisation de séquences

Motifs (A. Brazma) :

Classe	Exemple	
A	t-c-t-t-g-a	
B	D-R-C-C-x(2)-H-D-x-C	
C	G-G-G-T-F-D-[ILV]-[ST]-[ILV]	
D	V-x-P-x(2)-[RQ]-x(4)-G-x(2)-L-[LM]	
E	G-C-x(1,3)-C-P-x(8,10)-C-C	
F	C-x(2,4)-C-x(3)-[ILVFYC]-x(8)-H-x(3,5)-H	**(Prosite, Pratt)**
G	G-G-G-T-F-D-*- D-R-C-C-P	
H	G-G-G-T-F-[DE]-*- D-R-C-[PAR]-C	
I	G-G-G-x(2,5)-T-F-[DE]-*-D-x(0,1)-C-[PAR]-C	

- Caractères anonymes fixe ou variable — x(2), x(2,4)
- Regroupement de caractères — [ILV]
- Association de pattern — P1-*-P2

Caractérisation de séquences

Motifs :

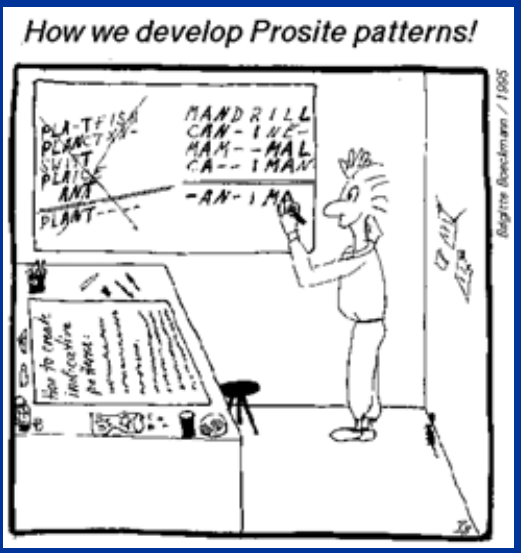
Classe	Exemple
A	t-c-t-t-g-a
B	D-R-C-C-x(2)-H-D-x-C
C	G-G-G-T-F-D-[ILV]-[ST]-[ILV]
D	V-x-P-x(2)-[RQ]-x(4)-G-x(2)-L-[LM]
E	G-C-x(1,3)-C-P-x(8,10)-C-C
F	C-x(2,4)-C-x(3)-[ILVFYC]-x(8)-H-x(3,5)-H (Prosite, Pratt)
G	G-G-G-T-F-D-*- D-R-C-C-P
H	G-G-G-T-F-[DE]-*- D-R-C-[PAR]-C
I	G-G-G-x(2,5)-T-F-[DE]-*-D-x(0,1)-C-[PAR]-C
J	Expression régulière/Grammaire régulière/Automate
K	Grammaire algébrique
M	Grammaire contextuelle
N	Grammaire à structure de phrase

Prosite vs Régulier

Prosite ~ position :

- « local »
  x(2,5)
  pas M1-*-M2
- Pas de corrélation possible
  [AC]-[AC]
  si A alors ensuite C sinon …?

Découverte de motifs : méthode PROSITE



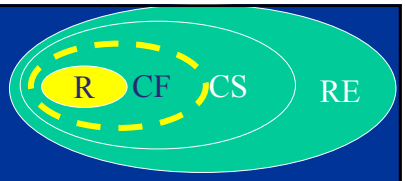
Découverte de motifs : méthodes automatiques

1) Pratt (I. Jonassen 1996)
   http://www.ii.uib.no/~inge/Pratt.html
   http://www2.ebi.ac.uk/pratt/
2) Splash (A. Califano 2000), Teiresias (Rigoutsos 1998)
   http://www.research.ibm.com/splash
   http://www.research.ibm.com/bioinformatics/teiresias
3) Gibbs Motif Sampler (C. Lawrence 1993)
   http://bayesweb.wadsworth.org/gibbs/gibbs.html
4) Meme et MetaMeme (T. Bailey 1995)
   http://meme.sdsc.edu/meme/website/meme.html
   http://metameme.sdsc.edu/cgi-bin/submit-verify.cgi

Inférence Grammaticale

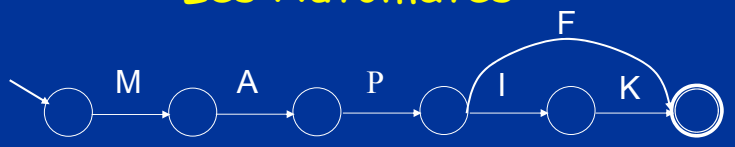
Apprentissage automatique
de modèles « grammaticaux »
à partir de séquences

Apprenabilité



Languages	*Automaton*	*Grammar*	*Recognition*	*Dependency*	*Biology*
Recursively Enumerable	Turing Machine	Unrestricted Baa → A	Undecidable	Arbitrary	Unknown
Context-Sensitive	Linear-Bounded	Context-Sensitive At → aA	NP-Complete	Crossing	Pseudoknots, etc.
Context-Free	Pushdown (stack)	Context-Free S → gSc	Polynomial	Nested	Orthodox 2^{o} Structure
Regular	Finite-State Machine	Regular A → cA	Linear	Strictly Local	Central Dogma

Les Automates



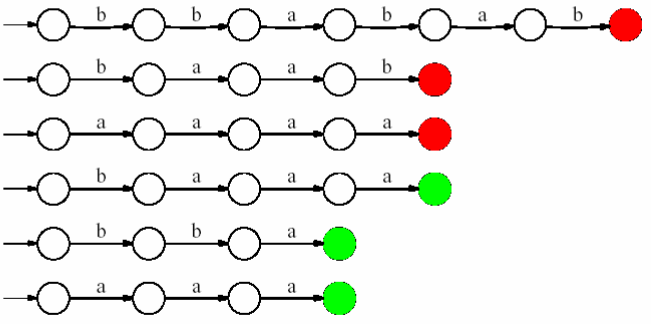
- Un automates est défini par :
  - ✓ un ensemble d 'états
  - ✓ un ensemble de transitions
  - ✓ des états initiaux
  - ✓ des états finaux
  - ✓ un alphabet auquel appartient chaque transition
- « Structure d'un HMM »

Algorithmes par fusions d'états

Maximal Canonical Automaton

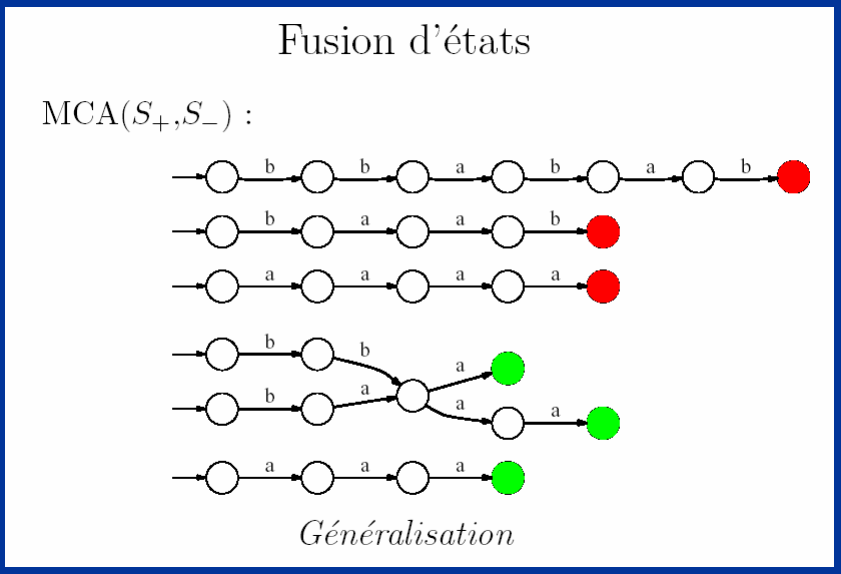
$S_+ = \{aaa, bba, baaa\}$; $S_- = \{aaaa, baab, bbabab\}$

MCA(S_+,S_-) :



Apprentissage par cœur

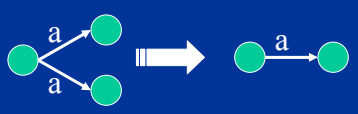
Algorithmes par fusions d'états



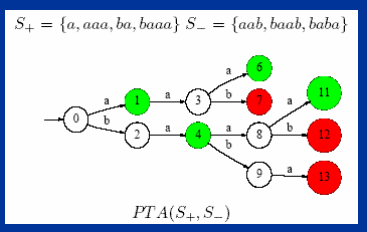
Heuristique EDSM

Abbadingo 1998, http://abbadingo.cs.unm.edu/

- Automates déterministes
  - ✓ Fusion pour déterminisation



  - ✓ MCA(S+,S-)-> PTA(S+,S-)



- Heuristique choix de (p,q) : EDSM [Price, Lang]
  Score de fusion de (p,q) = Δ #{état final}
  - ✓ Fusions les plus sûres d'abord
  - ✓ Ressemblance des 2 sous-arbres (attestée par les états finaux)

EDSM sur les protéines

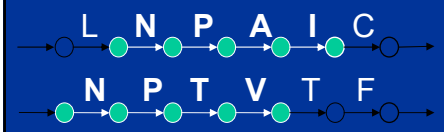
Ça marche pas…

- Séquences longues, taille alphabet…
- **Notion d'état final ?**
- **Déterminisme ?**

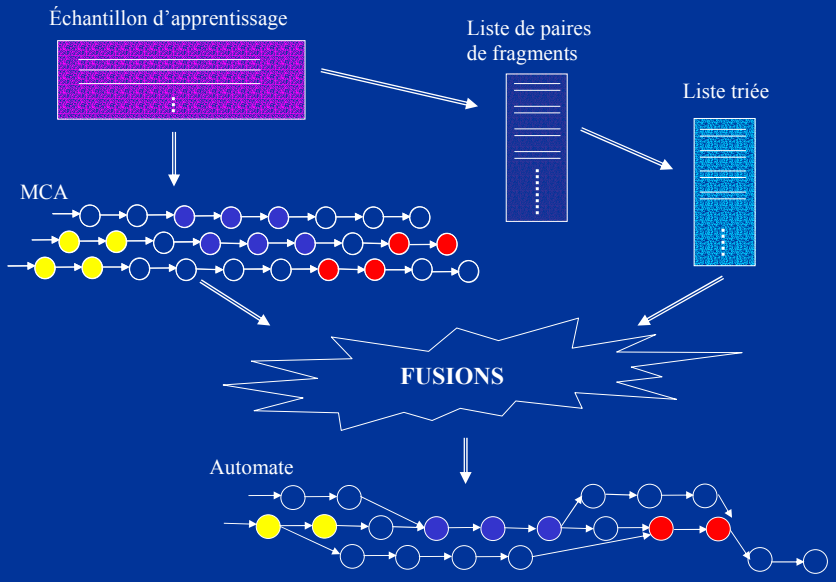
Garder idées d'évidence et de ressemblance
Enlever état final et déterminisme…

Approche par fusion de fragments significativement similaires

- Fragments significativement similaires
  - ✓ Conservation par sélection naturelle
  - ✓ Zones « importantes »
- Paires de fragments significativement similaires
  - ✓ DIALIGN2
    - → Ensemble de paires de fragments similaires (Blossum) avec indice de la significativité de la similarité
- Fusion de paires de fragments



Approche par fusion de fragments



Procedure **fma**(S,ω)

S : échantillon d'apprentissage

ω : seuil de similarité

1. A ← MCA(S)
2. F ← fragments_significatifs(S,ω)
3. F ← trier(F)
4. pour tout p=<f1,f2> dans F faire
5. si fusion_fragment_acceptable(A,f1,f2)
6. alors A ← fusion_fragment(A,f1,f2)
7. fsi
8. fpour
9. retourne A

Ordonnancement des paires de fragments

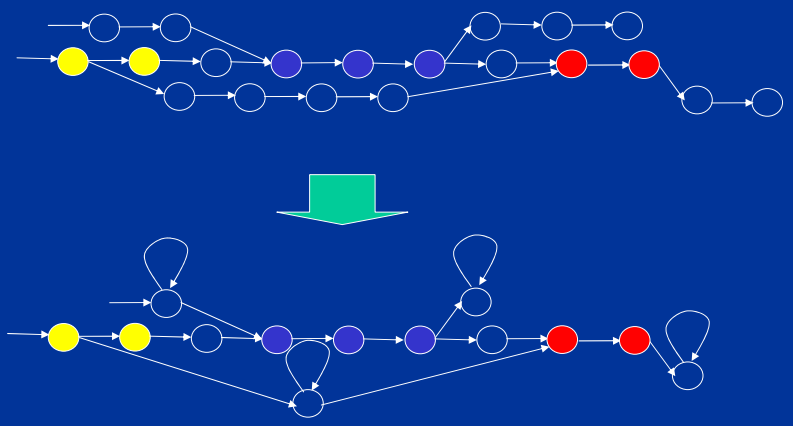
- Préservation des fragments
- Ordre important
- Réordonnancement des paires de fragments :
  - ✓ Significativité (score Dialign) — S+
  - ✓ Support — S+
    (dans les autres séquences de la paire de fragments)
    pf = <f1,f2> est supportée par f si w(f,f1) + w(f,f2) ≥ w(f1,f2)
  - ✓ Indice d'implication I.C. Lerman — S+,S-
    (non présence dans des contre-exemples)
  - ✓ Entropie… — S1,S2

Généralisation

1. Suppression des états non caractéristiques
2. Détection des propriétés physico-chimiques et extension

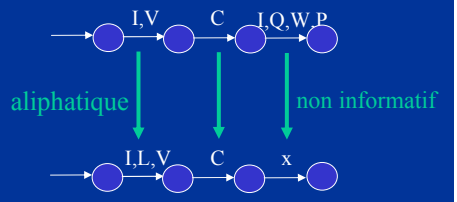
Etape de généralisation 1

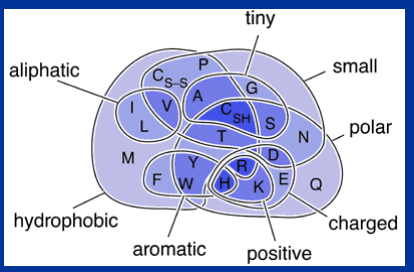
- Fusion sur elles-mêmes des parties non caractéristiques (« gaps »)



Etape de généralisation 2

- Fragments similaires ~ localisation d'une sélection / fonction
- Positionne en vis à vis des aa
- Propriétés physico-chimiques en jeu ?
- Généralisation à groupe de Taylor (ou autres…)





1ères expérimentations sur les MIPS

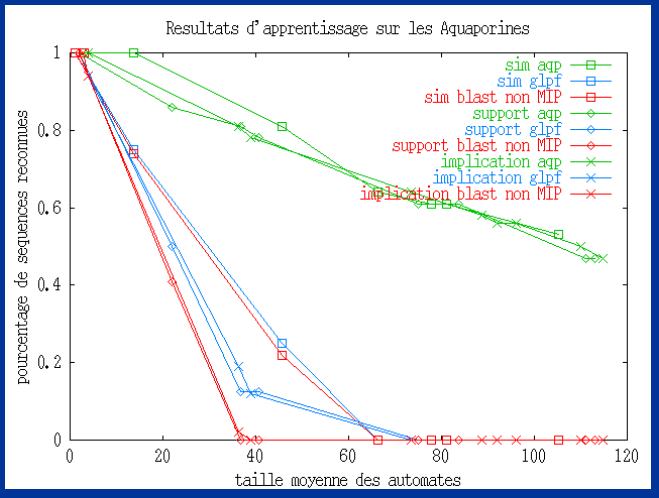
- 36 AQP
- 16 Glpf

(Moins de 90% de similarité)

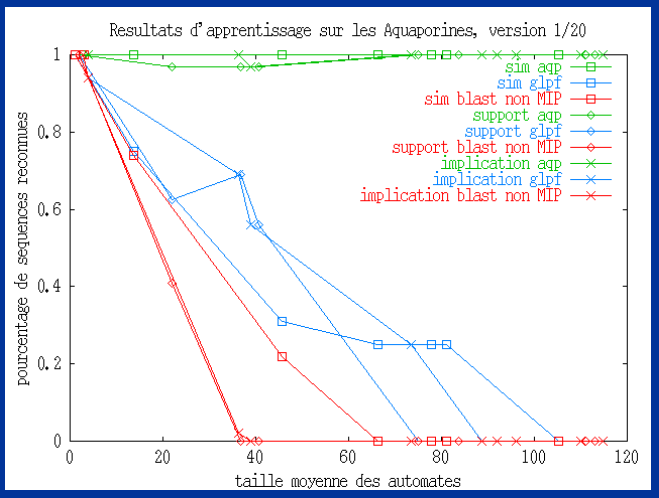
- 41 séquences non MIPS (obtenue par blast)

- Apprentissages « leave one out »
- Pour plusieurs nombres de fragments fixés pour comparer heuristiques

1ères expérimentations sur les MIPS

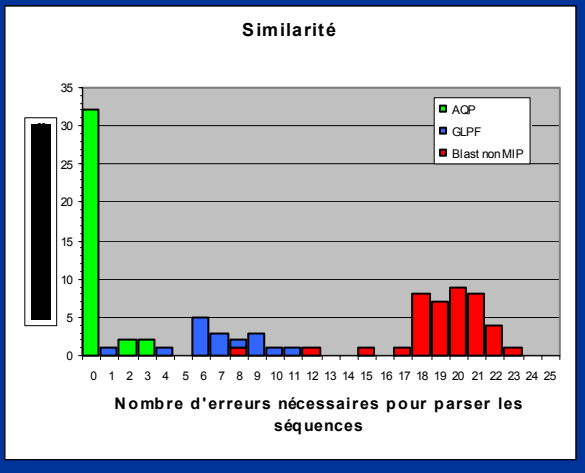


1ères expérimentations sur les MIPS



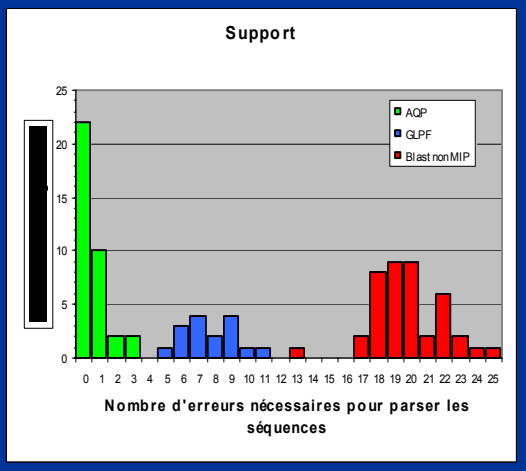
Influence tri

Apprentissage AQP, automate taille 80



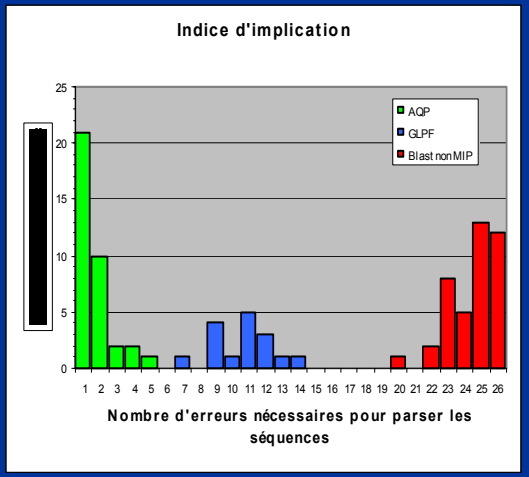
Influence tri

Apprentissage AQP, automate taille 80



Influence tri

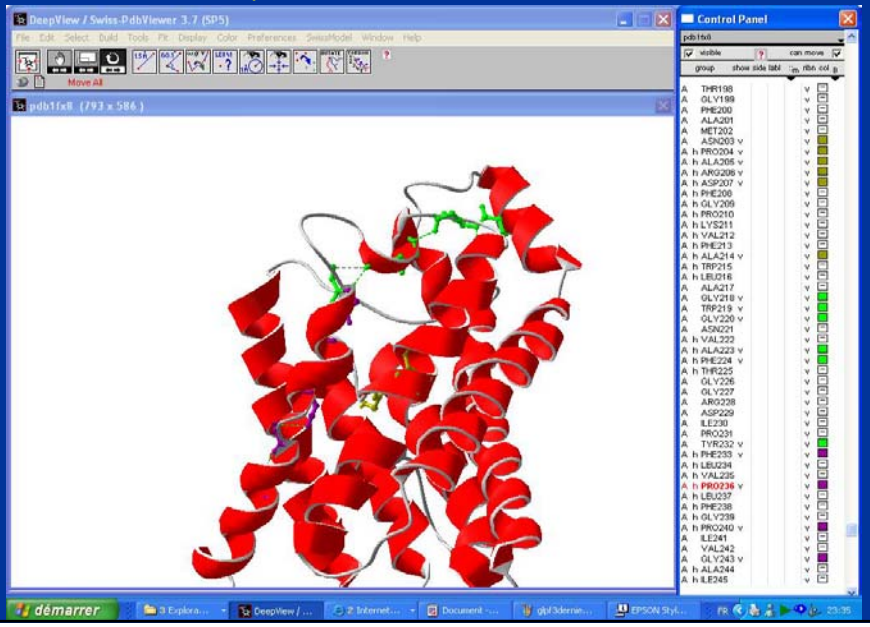
Apprentissage AQP, automate taille 80



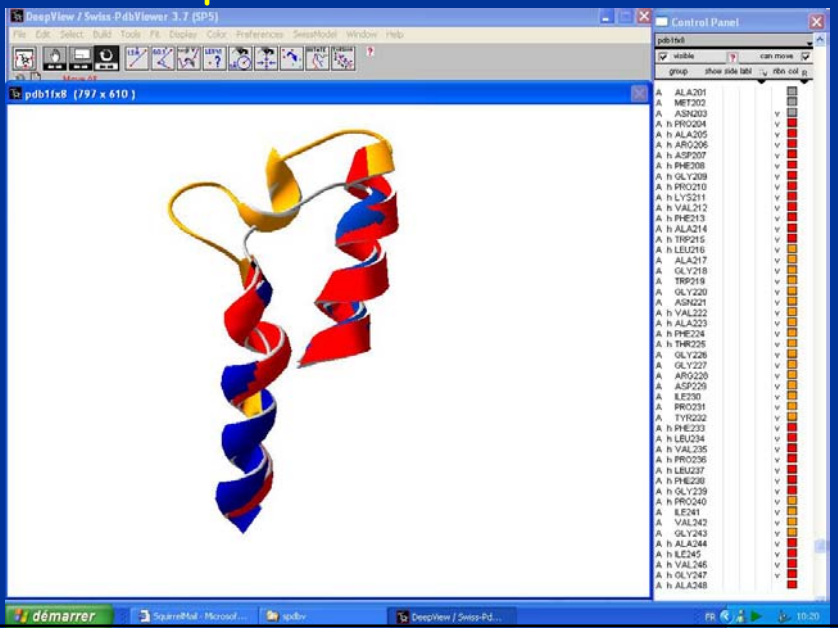
Approche par fusion de fragments similaires

- Apprentissage d'automates
- Sans alignement multiple
- Approche souple (paires de fragments)
  - ✓ Différents scores pour caractérisation, caractérisation avec contre-exemples, discrimination…
  - ✓ Taille de l'automate variable du motif discriminant « à la Prosite » au « consensus »…
- Identification et localisation de propriétés physico-chimiques
- Aspects structuraux ?

Aspects structuraux ?



Aspects structuraux ?



A développer…

Apprentissage d'automates « pertinents »…
… à systématiser
… à vérifier sur d'autres familles de protéines

- Influence paramètres à étudier :
  - ✓ nombre de fragments
  - ✓ %séquences pour généralisation gap (sous-familles?)
  - ✓ choix des groupes physico-chimiques pertinents…
- Taille optimale automate : MDL ?
- Approches discrimination (score « entropie ») ?
- Quand généraliser à un groupe de Taylor (estimer le bon « remplissage » d'un groupe) ?
- Intégrer les covariations d'acides aminés ?
- Algorithmique…
- Probabilités…