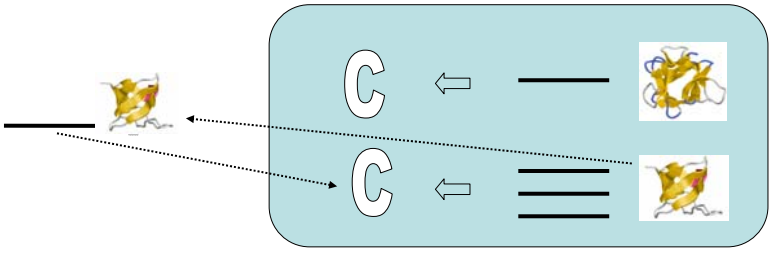
Application de l'approche par fusion de fragments significativement similaires à la caractérisation de repliements ?

François Coste, Symbiose IRISA,
Réunion ACI Genoto3D du 24 février 2005

Caractérisation de repliements



Apprentissage de la caractérisation de repliements

- Nombre de repliements limité
- De plus en plus de séquences pour un même repliement

⇨ Apprentissage de signatures de repliements à partir des séquences

Application de l'approche par fusion de fragments significativement similaires ™ IRISA implémentée dans Protomata-L ?

Protomata-L

1. Caractérisation
   1. Identification d'un ensemble de paires de fragments significativement similaires
   2. Ordonnancement des paires de fragments / score
2. Généralisation
   1. Fusion des fragments dans l'ordre
   2. Conservation des zones caractéristiques (élimination des positions non caractéristiques/quorum)
   3. Identification des positions importantes (acides aminés ou propriétés physico-chimiques conservés)

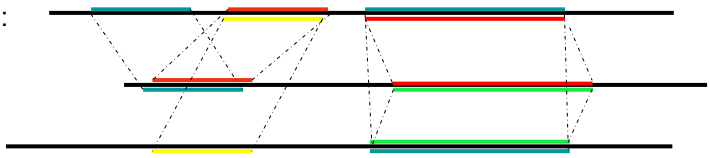
Caractérisation

Paires de fragments significativement similaires

- Fragments significativement similaires
  - Conservation par sélection naturelle
  - Zones « importantes »

Séquences :

Dialign:



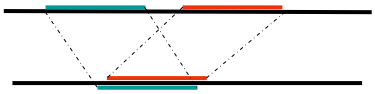
Paires de fragments (f1,f2) tq similarité(f1,f2) >> similarité « aléatoire »

Blossum62

Seuil à choisir

Ordonnancement des paires de fragments

- Problème :

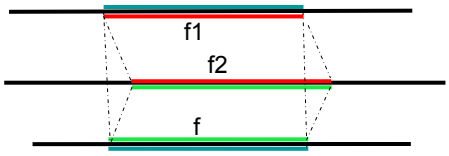


Quelles paires de fragment choisir ?

- Proposition : ordonner les paires de fragments
- Dialign fournit un score de significativité de la similarité : s(f1,f2) « indépendant » de la longueur
  - Ne prend pas en compte la représentativité de la paire de fragments dans la famille de séquences …

Représentativité des paires de fragments

- Support d'une paire de fragments:
  (f1,f2) est supporté par f si s(f1,f)+s(f2,f) >= s(f1,f2)
  (inégalité triangulaire)



→ n+ : nombre de séquences '+' supportant (f1,f2)
(→ n- : nombre de séquences '-' supportant (f1,f2))

- Scores :
  - Support dans S+
  - Indice d'implication (S+, S- ou S?)
  - …

Généralisation

1. Fusion des fragments
2. Suppression des états non caractéristiques
3. Détection des propriétés physico-chimiques et expansion

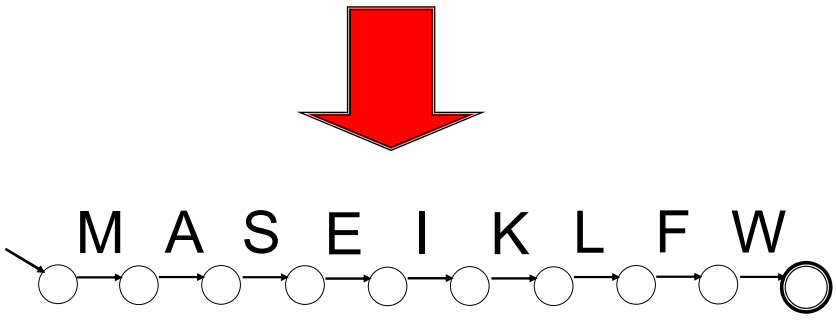
Fusion des fragments

- Tri/scores
- Fusion des paires de fragments

  cf fusion d'états utilisée en apprentissage d'automates…

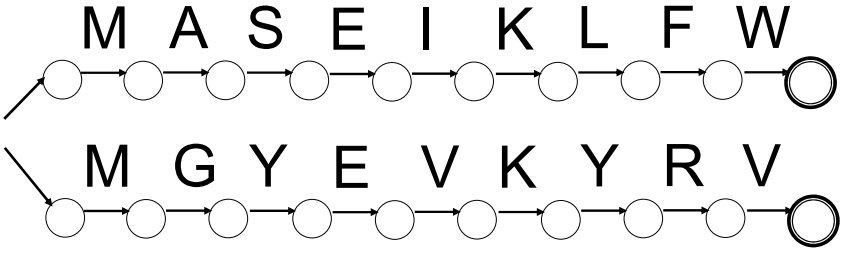
Des protéines aux automates

MASEIKLFW



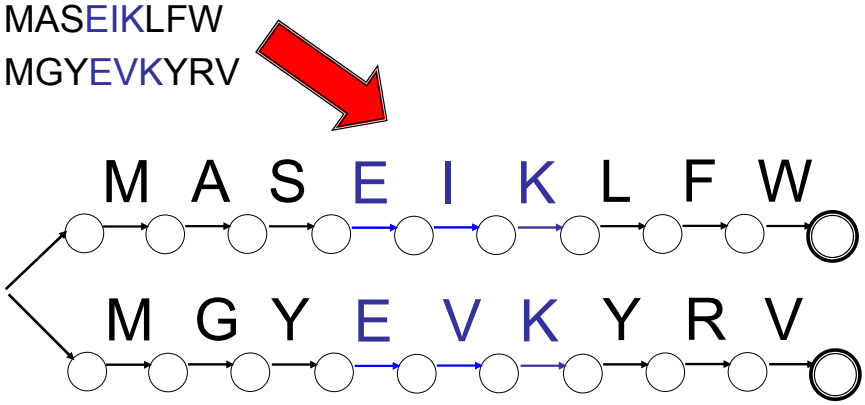
Des protéines aux automates

MASEIKLFW
MGYEVKYRV



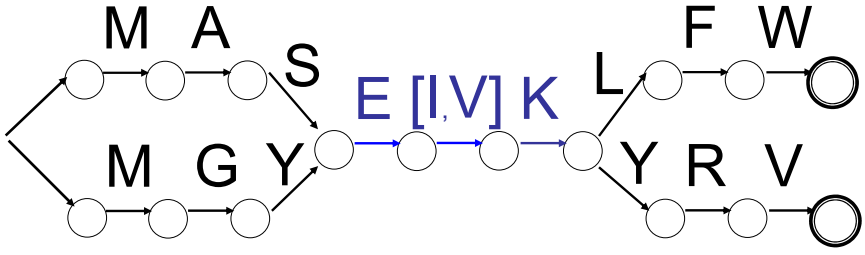
Fragments similaires

MASEIKLFW
MGYEVKYRV



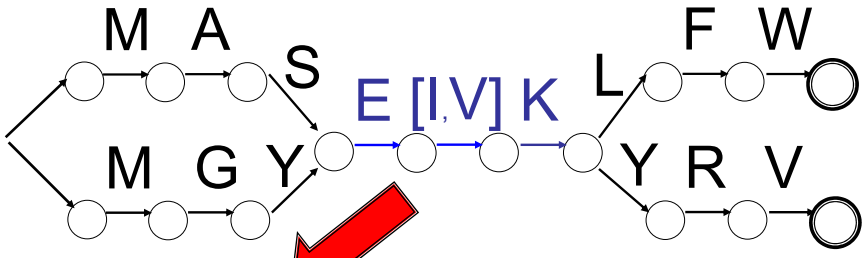
Fusion des fragments

MASEIKLFW
MGYEVKYRV



Généralisation

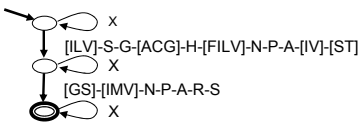
MASEIKLFW
MGYEVKYRV



MASEVKLFM	MGYEIKYRV
MASEIKYRV	MGYEVKLFW
MASEVKYRV	MGYEIKLFW

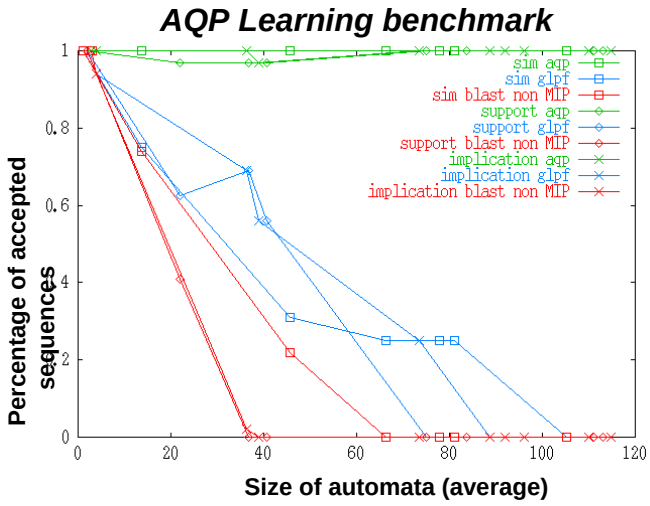
Fusion des fragments

- Tri/scores des paires de fragment
- Fusion des fragments dans l'ordre en préservant les fragments déjà fusionnés
- Paramètre : nombre de fragments à fusionner
  - Peu de paires de fragments
    → discrimination ≈Prosite



  - Beaucoup de paires de fragments
    → caractérisation (jusqu'au consensus/alignement multiple)

Validation curves

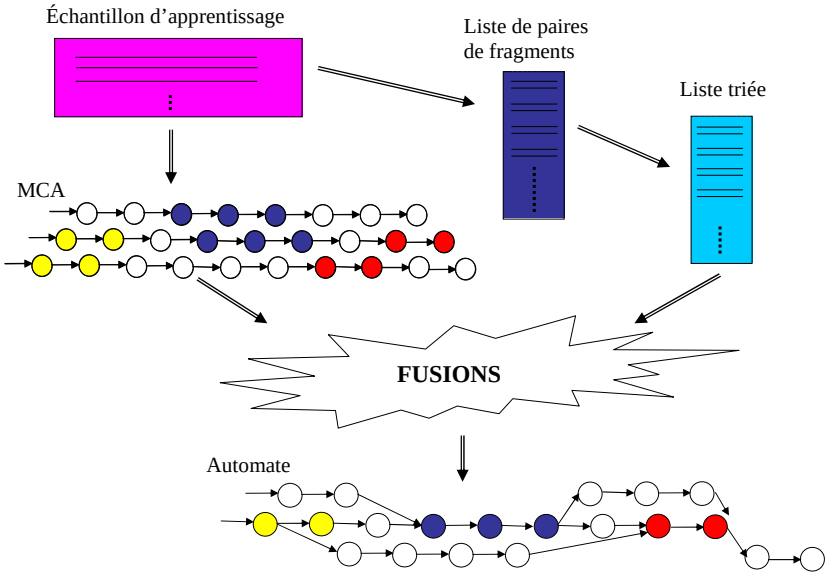


Procedure **fma**(S,ω)

S : échantillon d'apprentissage
ω : seuil de similarité

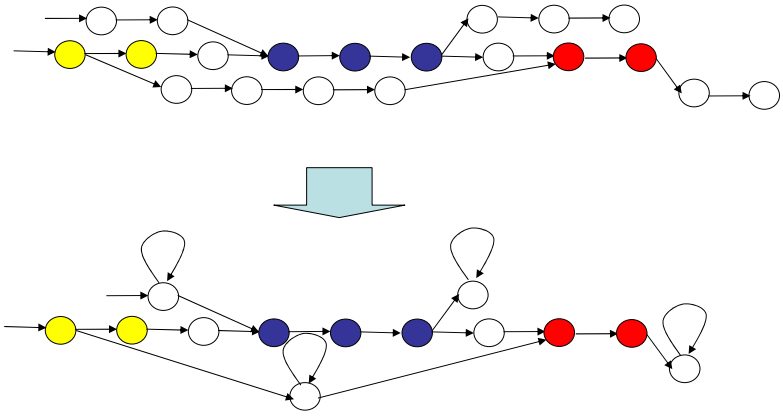
1. A ← MCA(S)
2. F ← fragments_significatifs(S,ω)
3. F ← trier(F)
4. pour tout p=<f1,f2> dans F faire
5. si fusion_fragment_acceptable(A,f1,f2)
6. alors A ← fusion_fragment(A,f1,f2)
7. fsi
8. fpour
9. retourne A

Fusion des fragments



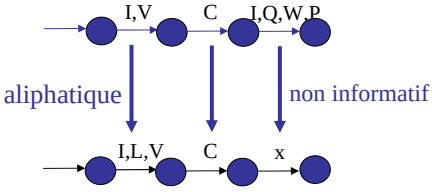
Etape de généralisation 2

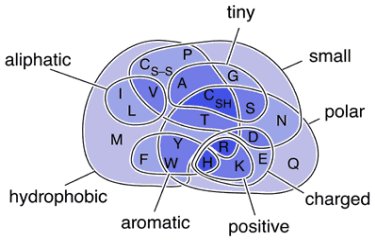
- Fusion sur elles-mêmes des parties non caractéristiques (« gaps »)

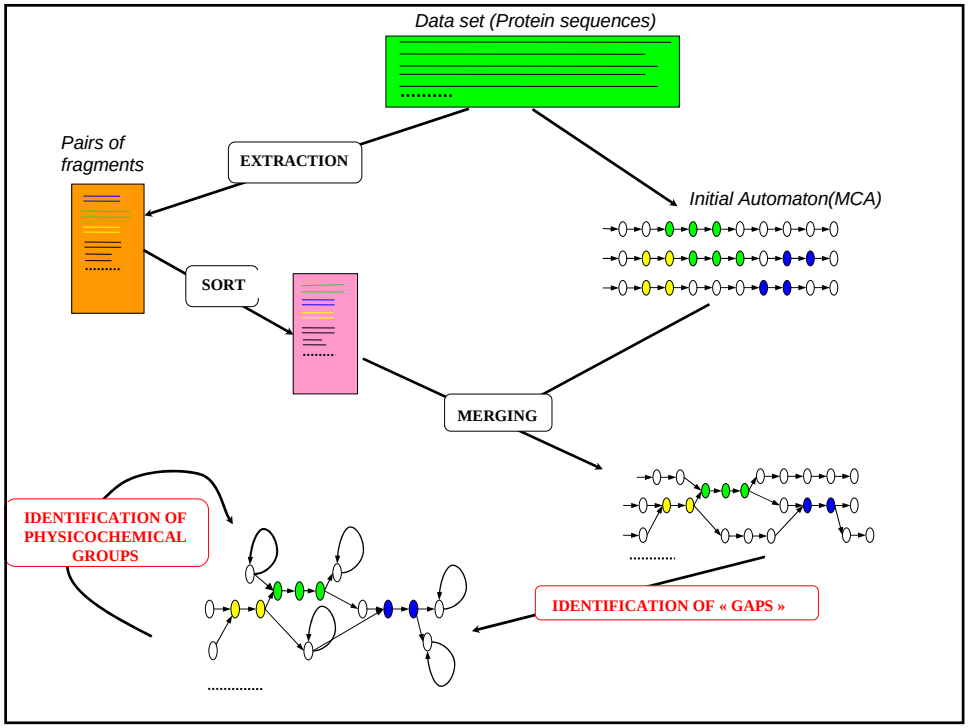


Etape de généralisation 3

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







Récapitulatif Protomata-L

- Caractérisation de familles de protéines par automates
  (structures de HMM…)
- Ne nécessite pas d'alignement multiple
- Algorithme glouton *data-driven*
  - Similarité de fragment → localisation de zones intéressantes / famille (scores)
  - Identification des propriétés physico-chimique
- Choix du degré d'apprentissage
  De la classification (Occam) à la caractérisation fine (« MDL ») au consensus
- Utilisation possible d'exemples négatifs (ou non étiquetés, bruit supporté)
- Introduction possible de connaissances (définition de zones à différencier)

Application/extension à la caractérisation de repliements ?

Stage Marie Lahaye (Master 2)

- Pertinence de l'approche par fragments similaires ?
- Que caractériser ?
  Choix des familles, niveau de caractérisation (les repliements, les domaines, … ?)
- Utilisation des informations structurales pour l'apprentissage ?
  (repliements, fragments en contact, points de contact,…)
- Apprendre des motifs structuraux ?
  (Modèle 1D ? Intégration d'informations 3D dans les automates ? dans le score ?)
- …