

Learning Automata on Protein Sequences

Jobim 2006

François Coste and Goulven Kerbellec



10 juillet 2006

Characterization of Protein Families

Zinc finger example

Set of nucleic acid binding proteins :

...YLGPLNCKSCWQKFDSFSKCHDHYLCRHCLNLLL...

...ILMCFICKLSIGNVKSFSLHANTEHRLNL...

...HKCEICLLSFPKESQFQRHMRDHE...

...

Characterization of Protein Families

Zinc finger example

Set of nucleic acid binding proteins :

...YLGPLNCKSCWQKFDSEFSKCHDHYLCRHCLNLLL...

...ILMCFICKLSIGNVKSFSLHANTEHRLNL...

...HKCEICLLSFPKESQFQRHMRDHE...

...

Prosite's C2H2 Pattern :

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

Characterization of Protein Families

Zinc finger example

Set of nucleic acid binding proteins :

...YLGPLNCKSCWQKFDSFSKCHDHYLCRHCNLLL...

...ILMCFICKLSIGNVKSFSLHANTEHRLNL...

...HKCEICLLSFPKESQFQRHMRDHE...

...

Prosite's C2H2 Pattern :

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

Characterization of Protein Families

Zinc finger example

Set of nucleic acid binding proteins :

...YLGPLNCKSCWQKFDSFSKCHDHYLCRHCNLLL...

...ILMCFICKLSIGNVKSFSLHANTEHRLNL...

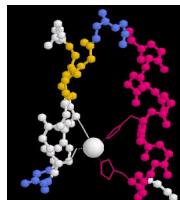
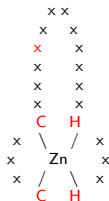
...HKCEICLLSFPKESQFQRHMRDHE...

...

Prosite's C2H2 Pattern :

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

'Zinc finger' domain :



Pattern Discovery on Protein Sequences

- Pratt
- Teiresias
- Emotif

Pattern Discovery on Protein Sequences

- Pratt
- Teiresias
- Emotif

Mostly used :

Pattern Discovery on Protein Sequences

- Pratt
- Teiresias
- Emotif

Mostly used :

- ClustalW !

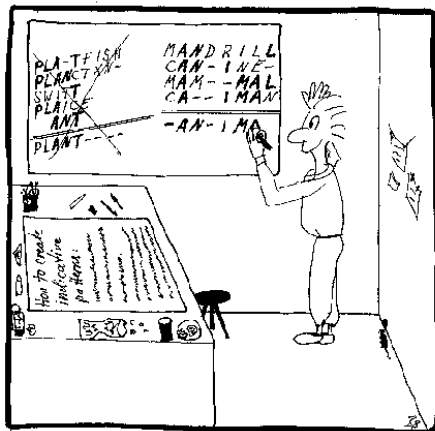
Pattern Discovery on Protein Sequences

- Pratt
- Teiresias
- Emotif

Mostly used :

- ClustalW !

How we develop Prosite patterns!



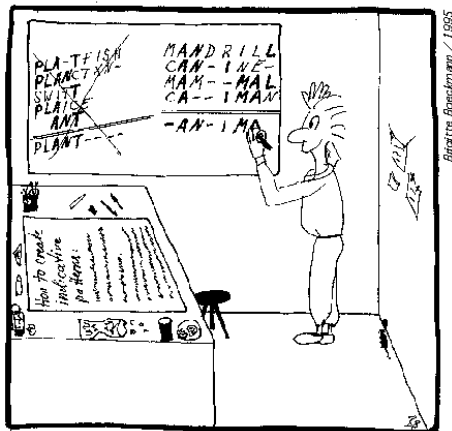
Pattern Discovery on Protein Sequences

- Pratt
- Teiresias
- Emotif

Mostly used :

- ClustalW !

How we develop Prosite patterns!



Position specific characterizations (even the profiles...)

Learning Automata on Protein Sequences

A Grammatical Inference Approach

- Grammatical inference : learn the grammar (syntax) of a sequence family from a sample
 - Explicit models : we want to predict if a new sequence belongs to the family, but also *why*
 - *Global* characterization of the sequences rather than *local* motif characterizations
- Automata :
 - Escape from position specific characterizations
 - “Sufficient” expressivity for proteins
 - Structure of HMM

Learning Automata on Protein Sequences

It is easy to represent a set of protein sequences by an automaton :

Learning Automata on Protein Sequences

It is easy to represent a set of protein sequences by an automaton :

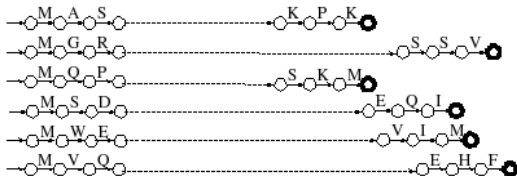
```
>AQP1_BOVIN
MASEFKKKLFWRVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGSFKQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Learning Automata on Protein Sequences

It is easy to represent a set of protein sequences by an automaton :

```
>AQP1_BOVIN
MASEFKKKLFWRAVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGKSFQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWRGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Maximal Canonical Automaton

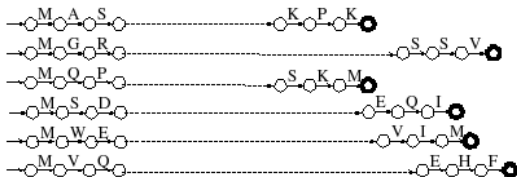


Learning Automata on Protein Sequences

It is easy to represent a set of protein sequences by an automaton :

```
>AQP1_BOVIN
MASEFKKKLFWRAVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGKSFQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Maximal Canonical Automaton



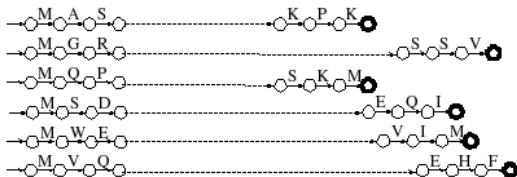
Rote learning !

Learning Automata on Protein Sequences

It is easy to represent a set of protein sequences by an automaton :

```
>AQP1_BOVIN
MASEFKKKLFWRAVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGKSFQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Maximal Canonical Automaton



Rote learning! An inductive leap is needed...

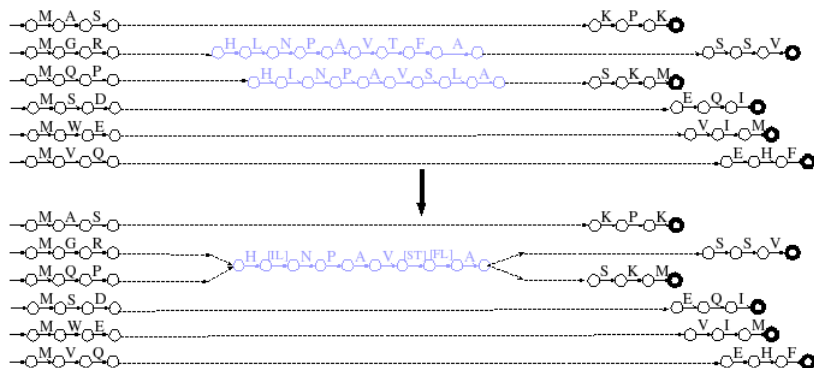
Generalization

- State merging scheme from grammatical inference

Generalization

- State merging scheme from grammatical inference
- Merging similar fragment pairs

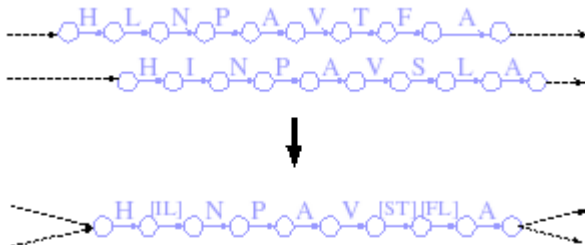
Maximal Canonical Automaton



Generalization

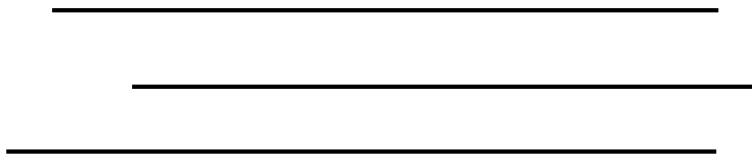
- State merging scheme from grammatical inference
- Merging similar fragment pairs

a closer view on merging 2 fragments :



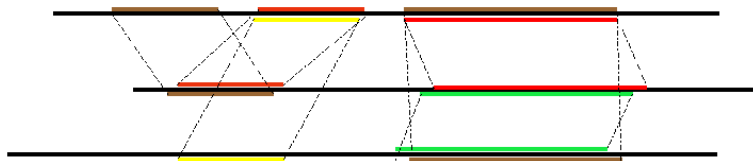
Similar Fragment Pairs ?

Protein sequences :



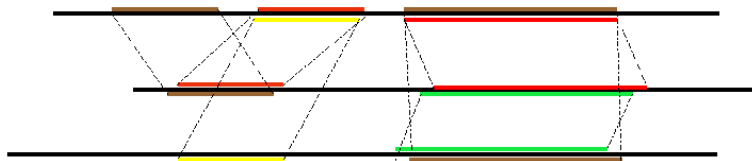
Similar Fragment Pairs ?

Significantly Similar Fragment Pair (SFP) :



Similar Fragment Pairs ?

Significantly Similar Fragment Pair (SFP) :



SFP : (f_1, f_2) s.t. $w(f_1, f_2) > 0$ DIALIGN [Morgenstern et al...]

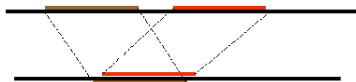
$w(f_1, f_2)$: weight of a fragment pair = $-\log P(s, l)$

$P(s, l)$: probability for a random fragment pair of length l to have a similarity greater than s

s : similarity of (f_1, f_2) , l : length of f_1 and f_2

Merging all the SFPs is not a solution !

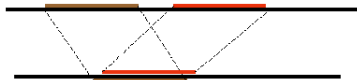
- Problem : incompatible SFPs



Consistency or preservation constraints

Merging all the SFPs is not a solution !

- Problem : incompatible SFPs



Consistency or preservation constraints

- Solution : ordering the SFPs according to a score
 - 3 different scoring functions
 - dialign weight (local)
 - support (family)
 - implication (discrimination wrt counter-example set)

Between two incompatible SFPs, choose the first.

Cliques mode

- Idea : avoid single linkage of SFPs

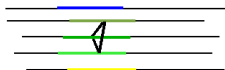


Cliques mode

- Idea : avoid single linkage of SFPs



- Clique of fragments : any pair of fragments in the clique is a SFP



Cliques mode

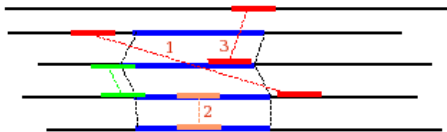
- Idea : avoid single linkage of SFPs



- Clique of fragments : any pair of fragments in the clique is a SFP



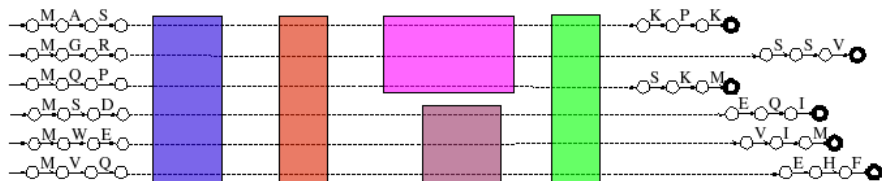
- Best clique search algorithm for decreasing size
(clique score : $\sum_{(f_1, f_2) \in C} w(f_1, f_2)$)
- For each best clique, discard incompatible ⁽¹⁾, included ⁽²⁾ and interfering ⁽³⁾ SFPs before searching for next best clique...



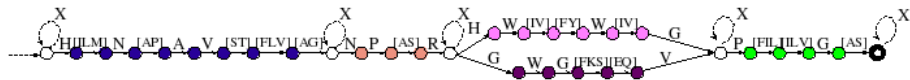
Fragments merging

Characterization stage → Partial Local Multiple Alignments (PLMAs) :

Maximal Canonical Automaton

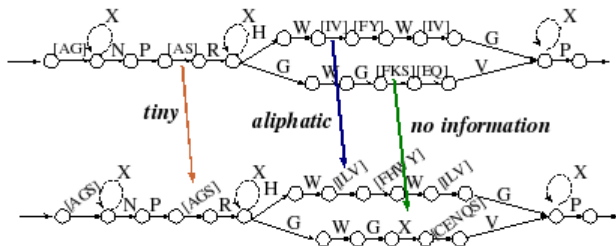
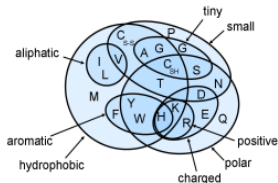


Merging PLMAs :



Identification of physico-chemical properties

- PLMA : average conservation of subsequences (BLOSUM)
- Identification of actual important physico-chemical properties

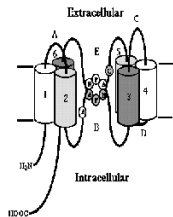


Likelihood ratio test

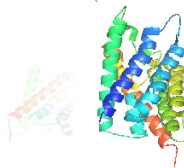
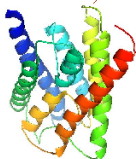
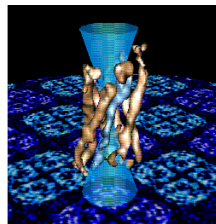
First validations

- Leave-one-out validation : assert quality of automata by its prediction performances on *unknown* sequences
- Two difficult families
 - MIP : Major Intrinsic Protein (homologous family), water specific proteins vs glycerol facilitator
 - TNF : Tumor Necrosis Factor (divergent family)
- Comparison of performances with :
 - Pratt, Teiresias
 - (Prosite)

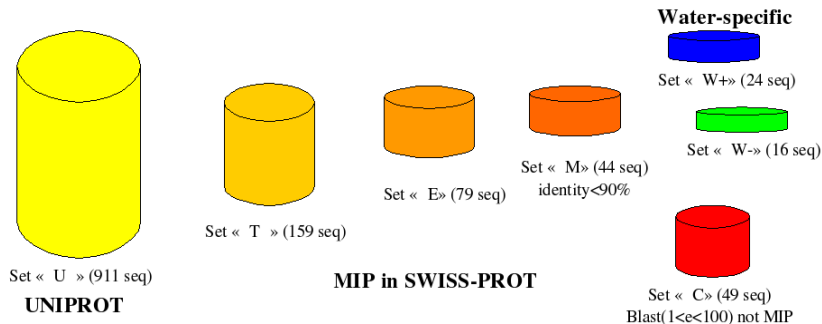
MIP : the Major Intrinsic Protein Family



Family
MIP
Subfamilies
AQP, Glpf, Gla



MIP data sets



Protomata-PL, MIP

For comparison with pattern discovery algorithms :

First PLMA involving all the sequences (support score)

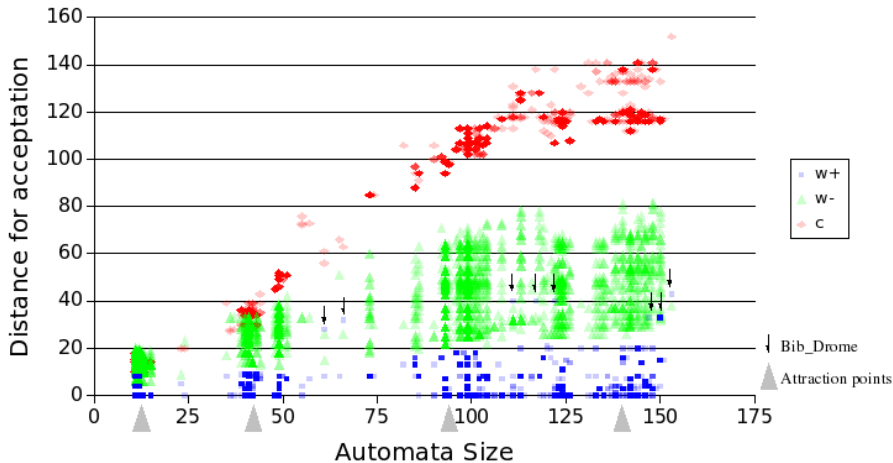
Training sample M (44 seq.), test sample T (159 seq.)

Method	Precision	Recall	F-meas.
Prosite (reference)	0.95	0.91	0.93
Pratt	0.90	0.78	0.83
Teiresias	0.23	0.89	0.37
Protomata-PL	1	0.87	0.93

Precision : $\frac{|Truepositive|}{|Retrievedsequences|}$, Recall : $\frac{|Truepositive|}{|Familysequences|}$, F-Measure : $\frac{2 \times Precision \times Recall}{Precision + Recall}$.

Protomata-PL, W_+ vs W_-

Increasing number of ordered SFP (implication score), quorum = 100%

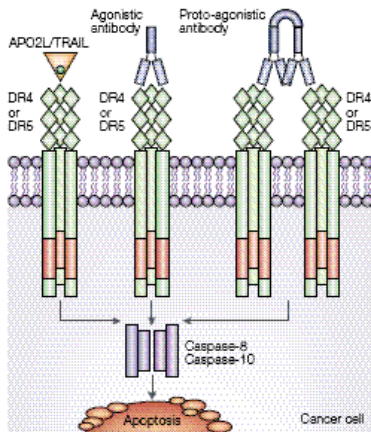


Protomata-PL, W_+ vs W_-

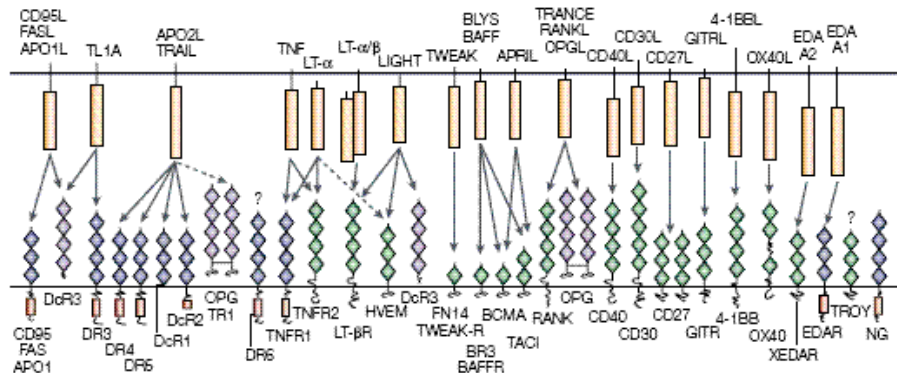
At attraction points :

Automata Size	Strict Parsing			Threshold Parsing		
	Prec.	Recall	F-meas.	Prec.	Recall	F-meas.
10	1	0.92	0.96	1	0.96	0.98
40	1	0.71	0.83	1	1	1
100	1	0.54	0.70	1	1	1
130	1	0.42	0.59	1	0.96	0.98

TNF : Tumor Necrosis Factor



TNF : Tumor Necrosis Factor



TNF data sets

- TNF family is included in the cytokine super-family
- Sequence divergence in the family is very high
- Positive set : 18 human TNF sequences
 - The average percentage of identity in the positive set is 33,6% (minimum of 0% and maximum of 71%)
- Negative test set : 4 false positive hits of the Prosite pattern plus 16 cytokines members not in the TNF family
 - The average percentage of identity between positive and negative sequences is 28,56% (minimum of 0% and a maximum of 81%)

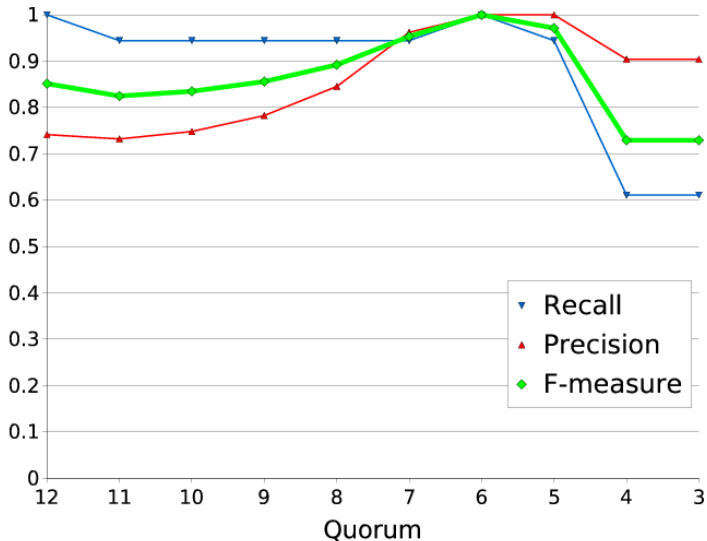
Comparison of Protomata-CL with other methods

Method	Precision	Recall	F-measure
Strict Parsing			
MCA	0	0	0
Prosite	0.75	0.67	0.71
Teiresias	0	1	0
Pratt	0.85	0.94	0.89
Protomata-PL Q=17	0.88	0.89	0.88
Threshold Parsing			
Pratt	0.86	1	0.92
Protomata-CL Q=7	0.96	0.94	0.95
Protomata-CL Q=6	1	1	1
Protomata-CL Q=5	1	0.94	0.97

Leave-one-out experimentations.

Q : quorum (minimal number of sequences covered by each PLMAs).

Impact of quorum for Protomata-CL



Conclusion

- New sequence-SFP driven algorithm
 - Learning Automata
 - Pattern Discovery
- Identification of **Partial** Local Multiple Alignments (significant “vertical” and “horizontal” conservations)
- Fragment merging generalization
- Refinement by identification of physico-chemical properties
- Validation by leave-one-out experiments

- Experimentations . . .
- Formalization of characterization stage
- Extension to more expressive grammars
- Application to structural families

Acknowledgments

- Christian Delamarche (MIP)
- Thierry Guillaudeau (TNF)
- Boris Idmont and Daniel Fredouille (CS students)