Learning Automata on Protein Sequences

François Coste    Goulven Kerbellec

Symbiose project
IRISA/INRIA
Rennes

June 12, 2006
Bioinformatic problem

- Biological question:

- Computer science answer:
Bioinformatic problem

- Biological question:
  How to define signatures of known protein families?
- Computer science answer:
Bioinformatic problem

- Biological question:
  How to define signatures of known protein families?

- Computer science answer:
  Using machine learning algorithms!
Protein families

- Amino acid alphabet
Protein families

- Amino acid alphabet
- Protein sequence
  \[ \text{AQP1\_bovin MASEFKKKLFWRAVVAEFLAMILFI-} \]
  \[ \text{FISIGSALGFHYPIKSNQTTGA} \]
  \[ \text{VQDNVKVSLAFG} \]
  \[ \text{LSI...} \]
Protein families

- Amino acid alphabet
- Protein sequence
  \[ \text{AQP1}\_\text{bovin} \quad \text{MASEFKKKLFWRAVVAEFLAMILFI-FISIGSALGFHYPIKSNQTTGAVQDNVKVS} \ldots \]
- Protein data set
  \[ \text{AQP1}\_\text{bovin} \quad \text{MASEFKKKLFWRAVVAEFLAMILFI-FISIGSALGFHYPIKSNQTTGAVQDNVKVS} \ldots \]
  \[ \text{AQP2}\_\text{rat} \quad \text{MWELRSIAFSRAVLAEFLATLLFVF-FGLGSALQWASSPPSVLQIAVAFGLGIGILVQALGH} \ldots \]
  \[ \text{AQP3}\_\text{mouse} \quad \text{MGRQKELMNRCGEMLHIRYRLL-RQALAECLGLTLILVMFGCGSVAQVVLS} \ldots \]
Protein family & Natural selection properties
Protein family & Natural selection properties

- Common function
Protein family & Natural selection properties

- Common topology (3D structure)
Protein family & Natural selection properties

➢ Common signature
Protein family & Natural selection properties

- Common function
- Common topology (3D structure)
- Common signature
Pattern of the zinc finger protein family

ZBT11 ...Csi..CgrtLpklys1riHmlk..H...
ZBT10 ...Cdi..CgklFtrrehvkrHslv..H...
ZBT34 ...Ckf..CgkkYtrkdqleyHirg..H...

Zinc Finger Pattern

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
### Expressivity classes of patterns

<table>
<thead>
<tr>
<th>Class</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>T-C-T-T-G-A</td>
</tr>
<tr>
<td>B</td>
<td>D-R-C-C-x(2)-H-D-x-C</td>
</tr>
<tr>
<td>C</td>
<td>G-G-G-T-F-[ILV]-[ST]-[ILV]</td>
</tr>
<tr>
<td>D</td>
<td>V-x-P-x(2)-[RQ]-x(4)-G-x(2)-L-[LM]</td>
</tr>
<tr>
<td>E</td>
<td>G-C-x(1,3)-C-P-x(8,10)-C-C</td>
</tr>
<tr>
<td>F</td>
<td>C-x(2,4)-C-x(3)-[ILVFYC]-x(8)-H-x(3,5)-H</td>
</tr>
<tr>
<td>H</td>
<td>D-T-A-G-[NQ]-*L-V-G-N-[KEH]</td>
</tr>
<tr>
<td>I</td>
<td>D-T-A-x(2,5)-G-[NQ]-*L-V-G-N-[KEH]</td>
</tr>
<tr>
<td>J</td>
<td>Regular Expression / Automaton</td>
</tr>
</tbody>
</table>
Outline

Characterization
   PLMAs
   SFPs
   CLIQUEs

Generalization
   Building Protomata
   Identification of Physico-chemical properties

Experiments
   MIP Family
   TNF Family

Conclusion
Outline

Characterization
- PLMAs
- SFPs
- CLIQUEs

Generalization
- Building Protomata
- Identification of Physico-chemical properties

Experiments
- MIP Family
- TNF Family

Conclusion
Outline

Characterization

PLMAs
SFPs
CLIQUEs

Generalization

Building Protomata
Identification of Physico-chemical properties

Experiments

MIP Family
TNF Family

Conclusion
PLMAs

Definition

The term of Partial Local Multiple Alignment designate strongly conserved regions in protein sequences.
Overview

**Sequences:**

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________

_________________________________________________________________
Overview

PLMAs :
Overview

SFPs:

F. Coste, G. Kerbellec
Learning Automata on Protein Sequences
Overview

**CLIQUEs:**

[Diagram of CLIQUEs]
Overview

*Protomata:*
Outline

Characterization
  PLMAs
  SFPs
  CLIQUEs

Generalization
  Building Protomata
  Identification of Physico-chemical properties

Experiments
  MIP Family
  TNF Family

Conclusion
SFPs

Definition

SFP (Significantly Similar Fragment Pair): important characterization area conserved by the natural selection.

Data set $D$:
Ordering the SFPs

Problem:

Sequence 1

Sequence 2

A

B

1

2

3

→ 3 different scoring functions

\[ S(f_1, f_2) =? \]
Ordering the SFPs

Problem:

\[ S(f_1, f_2) = ? \]

Solution ordering the SFP by scoring each SFP

\[ 3 \text{ different scoring functions} \]

- dialign \( S_d \)
- support \( S_s \)
- implication \( S_i \)
Dialign Score

\[ S_d(f_1, f_2) = -\log P(L, Sim) \]

- \( Sim \) = Sum of the individual similarity values
- \( L = |f_1| = |f_2| \)
- \( P = \) Probability that a random SFP of the same L has the same S

Coste, G. Kerbellec
IRISA
Learning Automata on Protein Sequences
Support Score

- \( S_s(f_1, f_2, D) = \) Number of sequences supporting \( < f_1, f_2 > \)

\[ \rightarrow \] Taking into account the representativeness of SFP

\( <f_1,f_2> \) is supported by \( f \) with respect the triangular inequality:

\[ Sd(f,f1) + Sd(f,f2) \geq Sd(f_1,f_2) \]
Implication Score

\[ S_i(f_1, f_2, D, N) = \frac{-P(Ss(f_1, f_2, N)) + P(Ss(f_1, f_2, D)) \times P(N)}{\sqrt{P(Ss(f_1, f_2, D)) \times |N|}} \]

\[ \text{avec } P(X) = \frac{|X|}{|D| + |N|} \]

- Taking into account a counter-example set N
- Discriminative fragments
Outline

Characterization
- PLMAs
- SFPs
- CLIQUEs

Generalization
- Building Protomata
- Identification of Physico-chemical properties

Experiments
- MIP Family
- TNF Family

Conclusion
CLIQUES

- SFPs and Transitivity problem
CLIQUEs

- SFPs and Transitivity problem

- A Clique of fragments is a PLMA with a significant similarity between each pair of fragment
CLIQUEs algo

Require: a set $P$ of SFPs, a set $S$ of sequences.
Result: ordered list $L$ of PLMAs s.t. each one is a clique of SFPs.

$z \leftarrow |S|$  \hspace{1cm} ▷ target size of the clique

while $z > 1$ do

$C \leftarrow \{ C = \{f_1, \ldots, f_z\} \mid \forall (f_i, f_j) \in C, (f_i, f_j) \in P\}\$

for each $C \in C$ do

compute $\text{SCORE}(C)$

\hspace{1cm} ▷ classically: $\Sigma_{p \in C} w(p)$

while $C \neq \emptyset$ do

$C \leftarrow \text{BEST\_SCORE\_CLIQUES}(C)$

$L.\text{APPEND}(C)$

$l_1 \leftarrow \{ p \in P \mid \exists q \in C, p \text{ incompatible with } q\}$

$l_2 \leftarrow \{(f_a, f_b) \in P \mid \exists f_i, f_j \in C, f_a \subset f_i, f_b \subset f_j\}$

$l_3 \leftarrow \{(f_a, f_b) \in P \mid \exists f_i \in C, f_a \cap f_i \neq \emptyset,
\forall f_j \in C, f_b \cap f_j = \emptyset\}$

$C \leftarrow C \setminus \{\{C\} \cup \{C' \in C \mid C' \cap (l_1 \cup l_2 \cup l_3) \neq \emptyset\}\}$

$z \leftarrow z - 1$

return $L$
Feasibility functions
Feasibility functions

- Incompatible
Feasibility functions

- Incompatible
- Included
Feasibility functions

- Incompatible
- Included
- Interfering
Outline

Characterization
- PLMAs
- SFPs
- CLIQUEs

Generalization
- Building Protomata
- Identification of Physico-chemical properties

Experiments
- MIP Family
- TNF Family

Conclusion
From protein data sets to automata

MASEIKLFW

F. Coste, G. Kerbellec
Learning Automata on Protein Sequences
From protein data sets to automata

MASEIKLFW
MGYEVKYRV

Learning Automata on Protein Sequences
Merging

MAS EIKLFW
MGYEVKYRV
Merging

MAS EIKLFW
MGYEVKYRV

学习自动机在蛋白质序列上的应用
Merging

MASEIKLFW
MGYEVKYRV

MAS
AS
MS
MGY
MY

E [I, V] K

F W
L
Y R V

MASEVKLFM  MGYEIKYRV
MASEIKYRV  MGYEVKLFW
MASEVYRV   MGYEIKLFW
Building protomata by merging PLMAs

Protein Sequence Data Set

List of PLMAs

MCA

Ordered List of PLMAs

MERGING

Automaton / Regular Expression
Gap Generalization

- Merging on themself non-representative transitions
- Treat them as gaps
Exceptions

Building Protomata

Characteristic path

Exception path

F. Coste, G. Kerbellec
IRISA
Learning Automata on Protein Sequences
Outline

Characterization
  PLMAs
  SFPs
  CLIQUEs

Generalization
  Building Protomata

Experiments
  MIP Family
  TNF Family

Conclusion

F. Coste, G. Kerbellec
Learning Automata on Protein Sequences
Identification of physico-chemical properties

- Similar Fragments $\rightarrow$ Potential function area
- Amino acids share out the same position
- Physicochemical property at play
- Generalization from a group (of amino acids) to a Taylor group

---

**Identifying Similar Fragments**

- Aliphatic
- No information

**Diagram**

- [I,V] $\rightarrow$ [I,L,V]
- C $\rightarrow$ C
Likelihood ratio test

- To decide if the multi-set $P$ has been generated according to a physico-chemical group $G$ or not by a likelihood ratio test $LR_{G/P} = \frac{L_G}{L_P} = \left(\frac{\sum_{a \in P} p_a}{\sum_{a \in G} p_a}\right)^n$.

- Given a threshold $\lambda$, we test the expansion of $P$ to $\Sigma$ by rejecting it when $LR_{\Sigma/P} = \frac{L_G}{L_P} = \left(\sum_{a \in \Sigma} p_a\right)^n < \lambda_{\Sigma}$.
Detailed Overview
Identification of Physico-chemical properties

Detailed Overview

MIP Sample

> AQP1_BOVIN
MASEFKKKLFWRRAVVAEFL...KPK
> AQP3_MOUSE
MGQRKELNRGCGE...SSV
> AQP9_HUMAN
MQPEGAEKGKSFQRLVKKSLA...SKM
> AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EIQ
> AQP2_RAT
MWELRSIASRVAFLATE...VIM
> AQP7_HUMAN
MVQAGHRSTRGSKMVWSVP...EHF

Maximal Canonical Automaton

List of ordered PLMAs
Detailed Overview

Maximal Canonical Automaton

(a)

(b)

(c)

F. Coste, G. Kerbellec
Learning Automata on Protein Sequences
Detailed Overview
Identification of Physico-chemical properties

Detailed Overview

MIP Sample

Maximal Canonical Automaton

List of ordered PIMAs

F. Coste, G. Kerbellec

Learning Automata on Protein Sequences
Outline

Characterization
  PLMAs
  SFPs
  CLIQUEs

Generalization
  Building Protomata
  Identification of Physico-chemical properties

Experiments
  MIP Family
  TNF Family

Conclusion
MIP : the Major Intrinsic Protein Family

Family
MIP
Subfamilies
AQP, Glpf, Gla
**Data sets**

- **Set « T »** (159 seq)
- **Set « E »** (79 seq)
- **Set « M »** (44 seq), identity < 90%
- **Water-specific**
  - Set « W+» (24 seq)
  - Set « W-» (16 seq)
- **Set « C »** (49 seq)
  - Blast(1<e<100) not MIP

**UNIPROT**

F. Coste, G. Kerbellec

Learning Automata on Protein Sequences
Protomata-PL, $W^+$ vs $W^-$ with a LVO experiment
Outline

Characterization
- PLMAs
- SFPs
- CLIQUEs

Generalization
- Building Protomata
- Identification of Physico-chemical properties

Experiments
- MIP Family
- TNF Family

Conclusion
TNF : Tumor Necrosis Factor
TNF : Tumor Necrosis Factor
Data sets

- The TNF family is included in the cytokine super-family.
- The sequence divergence in the family is very high.
- The positive set is made of the 18 human sequences:
  - The average percentage of identity in the positive set is 33.6% (minimum of 0% and maximum of 71%).
- The negative test set contains the 4 false positive hits of the Prosite pattern plus 16 cytokines members known to be outside of 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 to other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strict Parsing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Prosite</td>
<td>0.75</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>Teiresias</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pratt</td>
<td>0.85</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>Protomata-PL Q=17</td>
<td>0.88</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Threshold Parsing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pratt</td>
<td>0.86</td>
<td>1</td>
<td>0.92</td>
</tr>
<tr>
<td>Protomata-CL Q=7</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Protomata-CL Q=6</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Protomata-CL Q=5</td>
<td>1</td>
<td>0.94</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Leave-one-out test. Q is the minimum value of the quorum.
Impact of quorum on F-measure for Protomata-CL

![Graph showing the impact of quorum on F-measure for Protomata-CL.](image)

- **Recall**
- **Precision**
- **F-measure**

**Quorum**

**F. Coste, G. Kerbellec**

Learning Automata on Protein Sequences
Conclusion

- Good characterization of protein family using automata
- No need of a multiple alignment
- Greedy data-driven algorithm
  - Important subparts localization
  - Physico-chemical identification and generalization
- Counter example sets
- Bringing of knowledge is possible in automata