IDENTIFYING PROTEINS INVOLVED IN PARASITISM BY DISCOVERING DEGENERATED MOTIFS

Celine Vens\textsuperscript{1,2}, Etienne Danchin\textsuperscript{2}, Marie-Noëlle Rosso\textsuperscript{2}

\textsuperscript{1} Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
\textsuperscript{2} Institut National de la Recherche Agronomique, Sophia-Antipolis, France
CONTENT

- Introduction
- Method
- Results
- Conclusion
Meloidogyne Incognita
- Plant-parasitic nematode
- Major crop devastator
- Annotated genome sequence
Sophisticated interactions with plants
- penetration of root tissue
- establishment of a feeding site

Set of effector proteins is crucial for these processes

**Goal:** identifying complete set of secreted effectors
- Common conserved motif(s)?
- Emerging motifs, positive and negative set needed
DATA

- 100 “positive” proteins
  - 59 with expression in secretory glands
  - 38 with identification in secretome
  - 3 translated EST contigs identified with mass-spectroscopy

- 459 “negative” proteins
  - 7 proteomes: M. incognita, M. hapla, B. malayi, P. pacificus, C. elegans, C. briggsae, D. melanogaster
  - take proteins that have orthologs in all 7 organisms, and are present as a single copy in each of them (OrthoMCL)
MOTIF DISCOVERY

- Identifying motifs in protein sequences important challenge
- Can identify proteins involved in the same biological process
- All existing methods search motifs at the amino acid level
- Conservation of physico-chemical properties more important than conservation of amino acids
- Motifs that include properties: < L I small D D acidic >
There exist several classifications of amino acids

# Physico-Chemical Properties

| Residues:      | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Atom Set      | A   | R   | N   | D   | C   | E   | Q   | G   | H   | I   | L   | K   | M   | F   | P   | S   | T   | W   | Y   | V   |
| Acidic        | *   |     |     |     |     |     |     |     |     | *   |     |     |     |     |     |     |     |     |     |     |
| Acyclic       |     | *   |     |     |     |     |     |     |     |     | *   | *   |     |     |     |     |     |     |     |     |
| Aliphatic     |     |     | *   | *   | *   |     |     |     |     |     |     |     | *   |     |     |     |     |     |     |     |
| Aromatic      |     |     |     |     |     |     | *   | *   |     |     |     |     |     | *   |     |     |     |     |     |     |
| Basic         |     |     |     |     |     |     |     |     |     | *   |     |     |     |     |     |     |     |     |     |     |
| Buried        |     | *   | *   | *   | *   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Charged       |     | *   |     |     |     |     |     |     |     |     | *   |     |     |     |     |     |     |     |     |     |
| Cyclic        |     |     | *   |     |     |     |     |     |     |     |     |     |     |     | *   | *   | *   |     |     |     |
| Hydrophobic   |     | *   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Large         |     |     |     | *   | *   | *   |     | *   |     |     |     |     |     |     |     |     |     |     |     |     |
| Medium        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | *   |     |     |
| Negative      |     |     | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| Neutral       |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Polar         |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Positive      |     | *   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Small         |     |     | *   | *   | *   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Surface       |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

Rasmol
TASK DESCRIPTION

- **Given**
  - a set of positive proteins, and a set of negative proteins
  - frequency thresholds $f_{pos}$, $f_{neg}$
  - a classification of amino acids

- **Find** all motifs
  - that are frequent in the positives (frequency $\geq f_{pos}$)
  - that are infrequent in the negatives (frequency $\leq f_{neg}$)
  - using specific amino acids and properties/classes
CONTENT

- Introduction
- Method
- Results
- Conclusion
MERCI

- MERCI: Motif - EmeRging and with Classes - Identification

- Generate-and-test approach
  - Look for frequent motifs and meanwhile check their infrequency
  - Structure all possible motifs using a “more general than” relation
  - Traverse the structure from general to specific, such that
    - each relevant motif is considered
    - no motif is considered more than once
GENERALITY ORDER

- $< \text{A C D} >$ more general than $< \text{A C D E} >$
- $< \text{small C D} >$ more general than $< \text{A C D} >$
- Generality ordering between classes

```
neutr  acycl  hydro  large  cycl  surf  polar  medium
  |    |    |   |    |    |    |
buried  small  aliph  arom  charg
  |    |    |   |    |
  basic  acidic
```
 Candidate generation

- add top-level element of the property DAG to the end of the pattern
- minimally specialize the last element of the pattern

Depth first traversal of lattice
CANDIDATE PRUNING AND TESTING

- Exploit anti-monotonicity:
  - If \( \text{freq}(M, \text{pos}) \leq f_{pos} \), then prune
  - If \( \text{freq}(M, \text{neg}) \leq f_{neg} \), then no need to test children

- Checking frequency in positive set:
  - only check sequences containing parents (vertical id-list format)
  - only if all parents are frequent

- Checking infrequency in negative set:
  - stop counting when \( f_{neg} \) is exceeded
CONTENT

- Introduction
- Method
- Results
- Conclusion
SOME RESULTS

- Search motifs specific for M. incognita effectors
- Parameter $f_{neg}$ was set to 0 in all experiments
  - motifs specific for positive proteins
- Parameter $f_{pos}$ was adapted to get manageable set of motifs
SOME RESULTS

- Without properties
  - \( fpos = 5 \)
  - 6 motifs
  - coverage of 21 positive sequences

- Betts & Russel classification
  - \( fpos = 12 \)
  - 76 motifs
  - top 2 motifs: coverage of 22 positive sequences
RASMOL CLASSIFICATION

- Signals that control destination of proteins at N-terminal region
- Search in first 30 positions
- Maximal motif length of 15
- \( fpos = 35 \)
- 97 motifs, covering 68 positive sequences
- Corresponds with SP
66 of 97 motifs specific for SP positives

- They cover
  - 56 of 57 SP positives
  - 0 non-SP positives
  - 0 negatives (includes SP sequences)

Subset of 4 motifs with same coverage

<table>
<thead>
<tr>
<th>Motif</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;large hydrophobic neutral buried neutral neutral buried buried neutral acyclic acyclic hydrophobic neutral acyclic acyclic acyclic&gt;</td>
<td>35</td>
</tr>
<tr>
<td>&lt;neutral buried neutral large buried neutral neutral hydrophobic hydrophobic neutral acyclic acyclic acyclic acyclic buried&gt;</td>
<td>38</td>
</tr>
<tr>
<td>&lt;hydrophobic neutral buried acyclic neutral neutral neutral buried neutral large neutral acyclic neutral polar acyclic&gt;</td>
<td>35</td>
</tr>
<tr>
<td>&lt;neutral neutral L buried hydrophobic buried neutral hydrophobic neutral neutral acyclic neutral&gt;</td>
<td>35</td>
</tr>
</tbody>
</table>
12% of proteome of M. Incognita covered (2,579 of 20,359 proteins)

80% of them have predicted SP (only 17% of proteome has predicted SP)

7 of 8 positive control PCWD proteins covered

Covered proteins contain 21 of 26 additional candidate PCWD proteins
CONTENT

- Introduction
- Method
- Results
- Conclusion
CONCLUSION

- Finding conserved motifs in biological sequences
- Biological system of interest: protein secretion of plant parasitic nematode
- MERCI: find discriminative motifs that use physico-chemical properties
- Works with user defined classifications
- Contribution to goal of identifying the whole set of effectors in M. incognita
MERCI :-)  

http://dtai.cs.kuleuven.be/ml/systems/merci