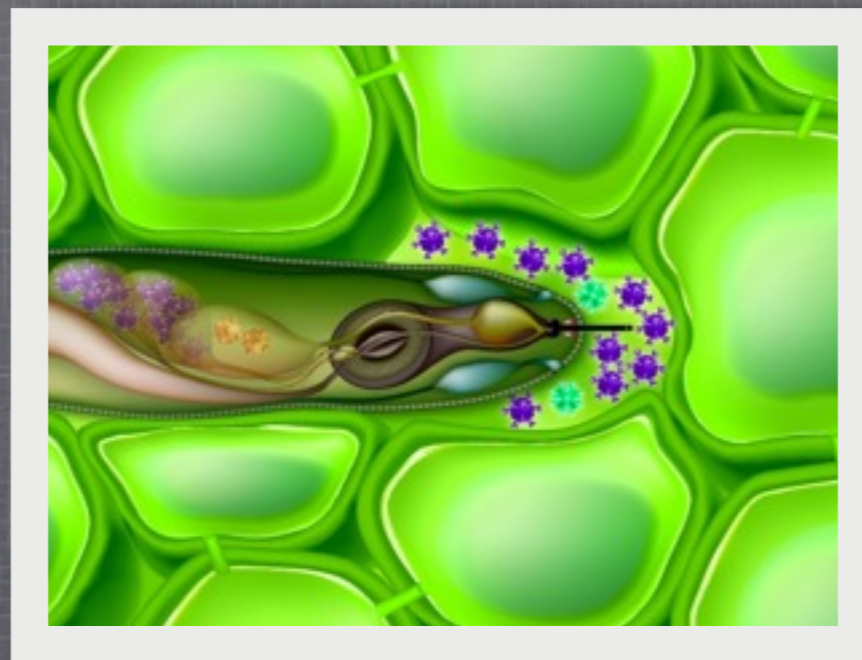


IDENTIFYING PROTEINS INVOLVED IN PARASITISM BY DISCOVERING DEGENERATED MOTIFS



Celine Vens^{1,2}, Etienne Danchin², Marie-Noëlle Rosso²

¹ Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium

² Institut National de la Recherche Agronomique, Sophia-Antipolis, France

CONTENT

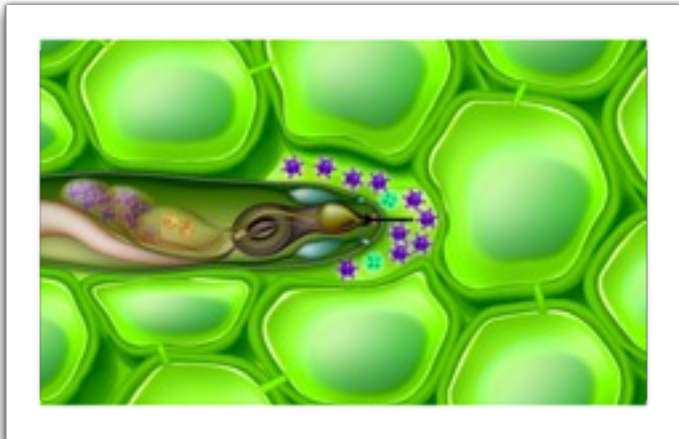
- **Introduction**
- Method
- Results
- Conclusion

CONTEXT

- *Meloidogyne Incognita*
 - Plant-parasitic nematode
 - Major crop devastator
 - Annotated genome sequence



CONTEXT

- Sophisticated interactions with plants
 - penetration of root tissue
 - establishment of a feeding site
 - Set of effector proteins is crucial for these processes
- A microscopic image showing a nematode (a small, worm-like animal) feeding on plant tissue. The nematode is positioned in the center, with its head and mouthparts visible. It is surrounded by green, rounded plant cells. Small purple and blue dots are scattered around the nematode, representing secreted effectors or nutrients.
- **Goal:** identifying complete set of secreted effectors
 - Common conserved motif(s)?
 - Emerging motifs, positive and negative set needed

DATA

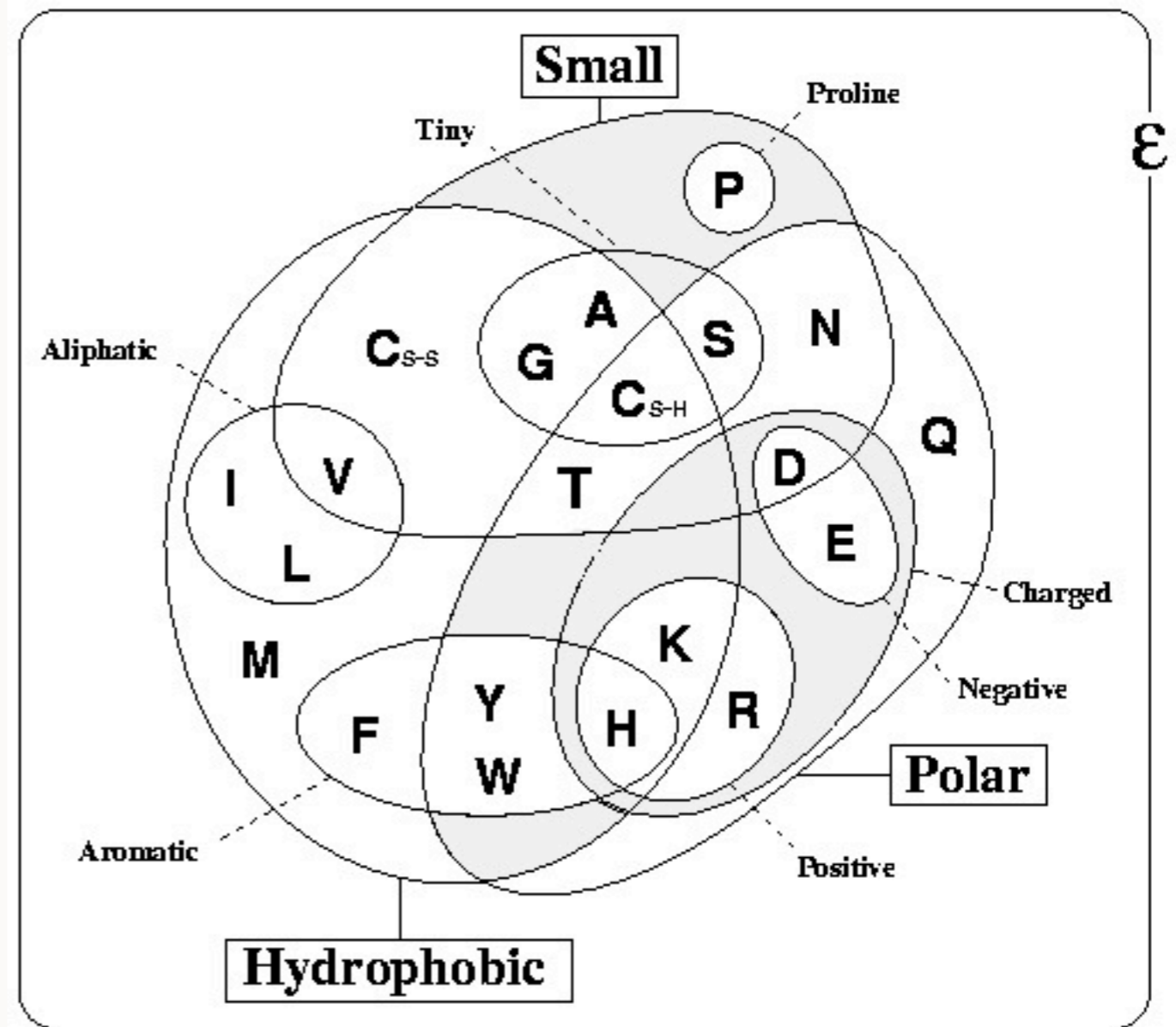
- 100 “positive” proteins
 - 59 with expression in secretory glands
 - 38 with identification in secretome
 - 3 translated EST contigs identified with mass-spectroscopy
- 459 “negative” proteins
 - 7 proteomes: *M. incognita*, *M. hapla*, *B. malayi*, *P. pacificus*, *C. elegans*, *C. briggsae*, *D. 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 >

PHYSICO-CHEMICAL PROPERTIES

- There exist several classifications of amino acids



Betts&Russell(2003)

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

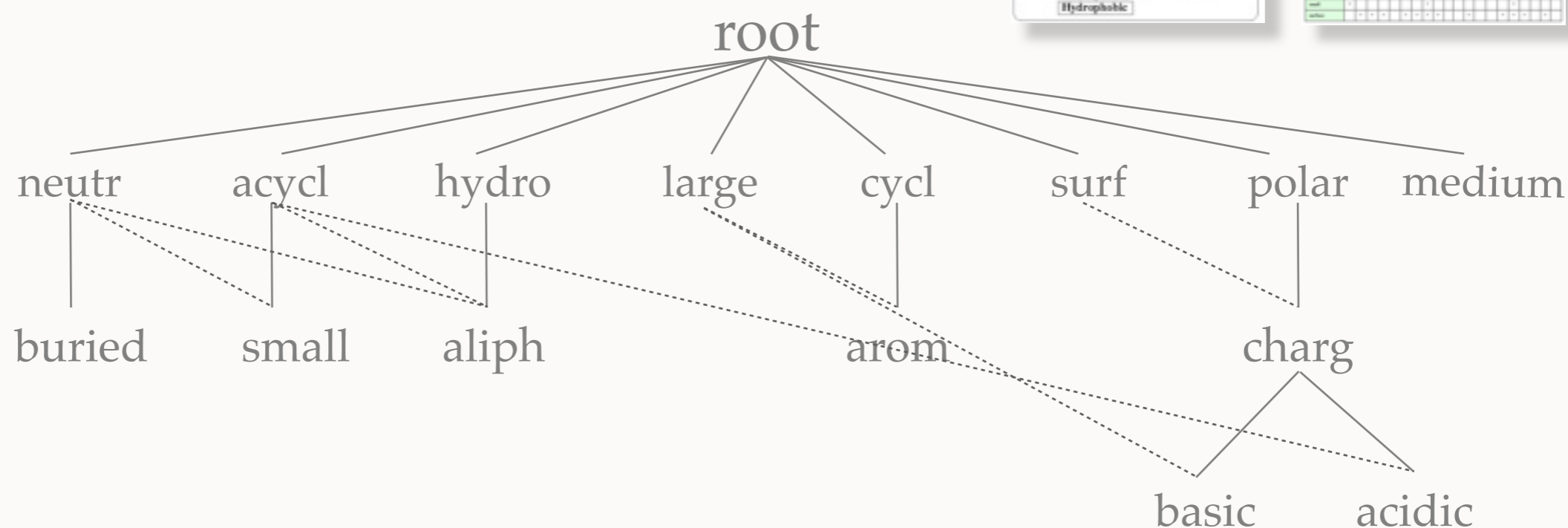
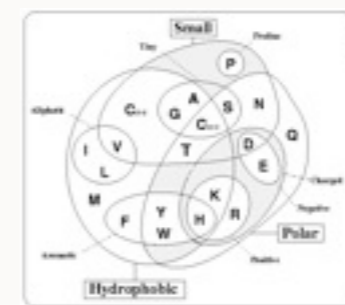
- Introduction
- **Method**
- Results
- Conclusion

MERCI

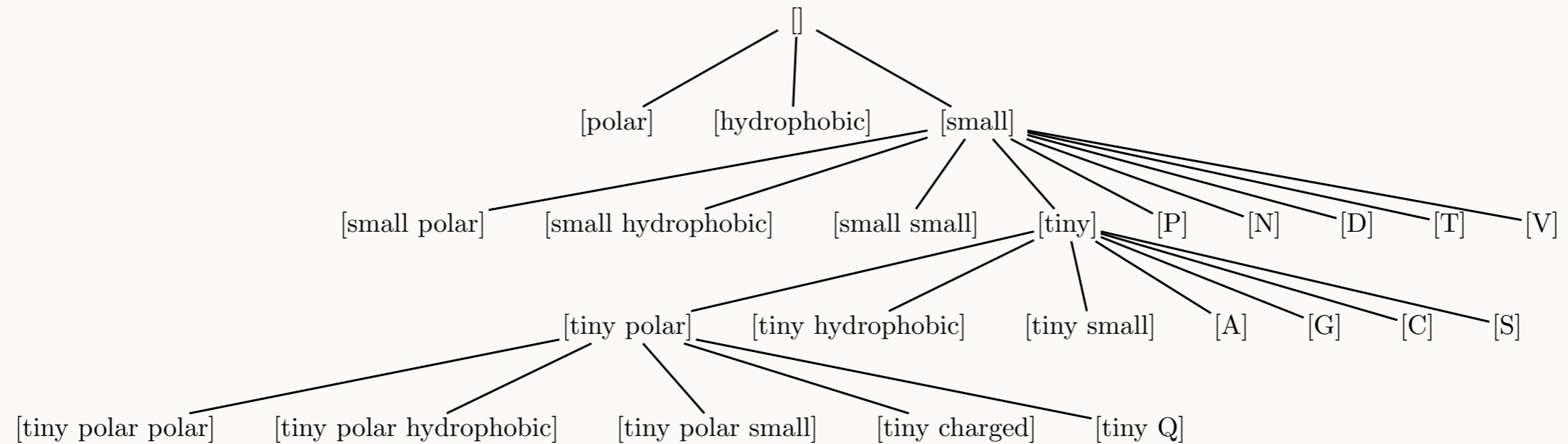
- MERCI: Motif - EmeRging and with Classes - Identification
- Generate-and-test approach
 - Look for frequent motifs and meanwhile check their infrequency
 - Structure all possible motifs using a “more general than” relation
 - Traverse the structure from general to specific, such that
 - each relevant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

GENERALITY ORDER

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



CANDIDATE GENERATION



- 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 $\text{freq}(M, \text{pos}) \leq f_{\text{pos}}$, then prune

- If $\text{freq}(M, \text{neg}) \leq f_{\text{neg}}$, then no need to test children

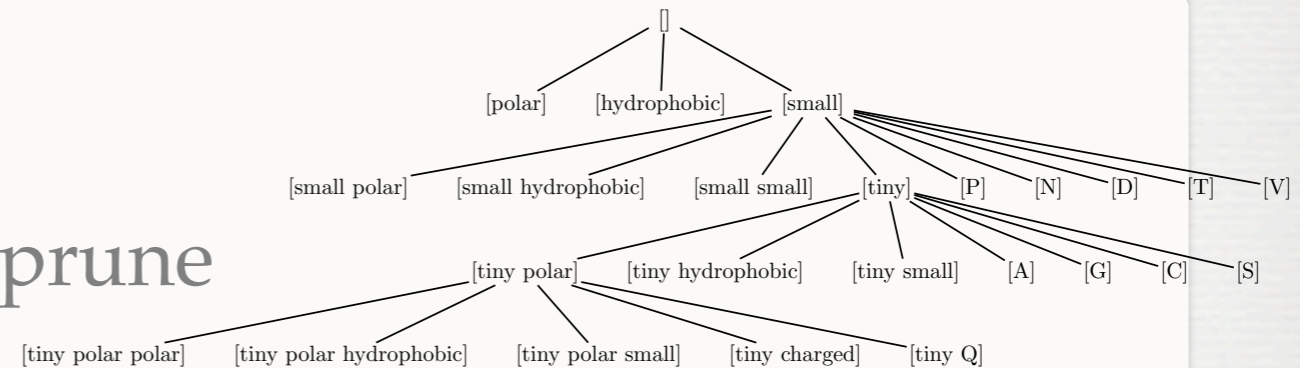
- Checking frequency in positive set:

- 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 *fneg* was set to 0 in all experiments
 - motifs specific for positive proteins
- Parameter *fpos* was adapted to get manageable set of motifs

SOME RESULTS

- Without properties

- $fpos = 5$

- 6 motifs

- coverage of 21 positive sequences

Classific.	Motif	freq(Motif,P)
None	<A E G D>	5
	<A S K Y>	5
	<E G A G>	6
	<L L I I S>	8
	<T L L I I>	5
	<T L L I I S>	5

- Betts & Russel classification

- $fpos = 12$

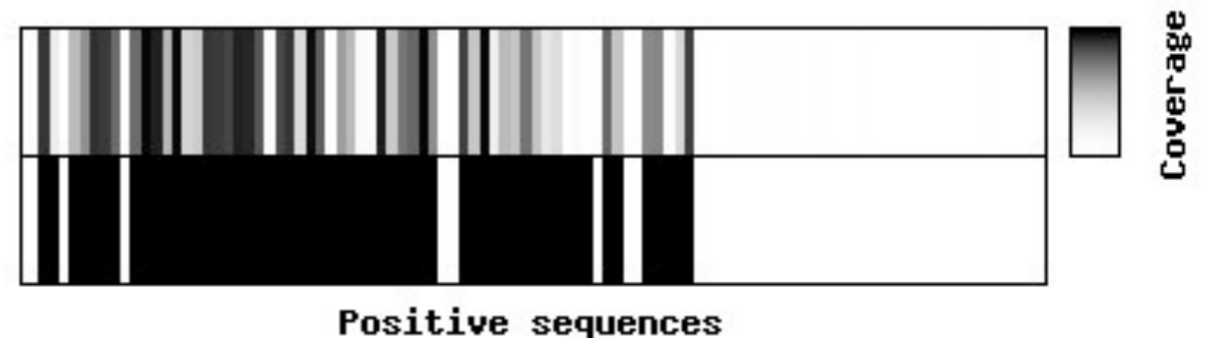
- 76 motifs

- top 2 motifs: coverage of 22 positive sequences

Betts and Russell	<hydro small hydro hydro polar hydro small tiny small charged polar small polar>	15
	<small small polar polar small small small hydro small small small small polar hydro polar hydro small>	14

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



RASMOL CLASSIFICATION

- 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

<large hydrophobic neutral buried neutral neutral buried buried neutral acyclic acyclic hydrophobic neutral acyclic acyclic>	35
<neutral buried neutral large buried neutral neutral neutral hydrophobic hydrophobic neutral acyclic acyclic acyclic buried>	38
<hydrophobic neutral buried acyclic neutral neutral neutral buried neutral large neutral acyclic neutral polar acyclic>	35
<neutral neutral L buried hydrophobic buried neutral hydrophobic neutral neutral acyclic neutral>	35

EVALUATING THE 4 MOTIFS

- 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>