Development of a software platform for the analysis of nonribosomal peptides

Ségoîlène Caboche
segolene.caboche@lifl.fr

LIFL (UMR CNRS/Université Lille 1) – INRIA – ProBioGEM

8 September, 2009
Outline

Work context
Nonribosomal peptide synthesis and products
Issue and tools

Tools for comparing NRPs
NRP and pattern modeling
Method for pattern matching

Database dedicated to NRPs
NORINE
Web interface

Mining Norine data
Statistical study of NORINE data
Activity prediction tool
Central dogma

DNA \xrightarrow{transcription} mRNA \xrightarrow{translation} protein

[Crick, 1970]
Central dogma

DNA \rightarrow \textit{transcription} \rightarrow \text{mRNA} \rightarrow \textit{translation} \rightarrow \text{protein}

- synthetases
- nonribosomal peptide synthesis
- nonribosomal peptides (2-50 aa)

[LiPmann et al., 1971]
Why being interested in nonribosomal peptides?

- a lot of important industrial applications (pharmacology, biotechnology, agriculture...)
- a lot of discovered molecules
  - directly from culture supernatant
  - from genomes (PCR, metagenomic, ...)
- possibility to modify the peptides in order to improve or change their biological activities
  - genetic approaches
  - chemo-enzymatic approaches
Nonribosomal peptide synthesis

- described for the first time in 1970’s, [Lipmann et al., 1971]
- insignificant in 1970’s, becoming more and more present in scientific literature (rapid growth of the number of publications dedicated to this pathway)

- in this pathway, the peptides (NRP) are synthesized by huge enzymatic complexes called synthetases or NRPS (NonRibosomal Peptide Synthetases)
genes are organized in clusters
synthetases are organized in modules
each module is responsible for the incorporation of a specific amino acid
modules are organized in domains
each domain harbors a specific enzymatic function
Synthetase organization

principa l domains

- Adenyl ation domain (A) : selection and acti vation of aa
- Thiolation (T) or Peptidyl carrier protein (PCP) domain : covalent fixation on synthetase
- Condensation domain (C) : peptidic bond between two adjacent aa
- Thioesterase domain (Te) : release of the peptide

secondary domains (E, M, Cy, ...) : modification of the incorporated aa
Different types of biosynthesis

- 3 types of biosynthesis [Mootz et al., 2002]
Different types of biosynthesis

- 3 types of biosynthesis [Mootz et al., 2002]
- these types of biosynthesis lead to possibly non-linear primary structures and a great structure diversity among nonribosomal peptides
Structure diversity

linear
D-Val — Cys — Aad — ACV-tripeptide

branched
DMOG — NSPD — DMOG
      |           
      DiOH-Bz  vibriobactin

partial cyclic
FA — Trp — D-Asn — Asp — Thr
     |   |   |   |
     Gly D-Ala Asp

bi-cyclic
Pro — NMe-Gly
     /  
    Pro — NMe-Gly

complex
Van
D-Glc
bOH-Cl-Tyr
Hpg

bOH-Cl-Tyr
Hpg
Asn

vancomycin

NMe-Leu

D-Val
NMe-Val
D-Val
NMe-Val

Thr
Chr

actinomycin D

cyclosporin A
Composition diversity

- use of nonproteogenic amino acids
Composition diversity

- use of nonproteogenic amino acids
- secondary domains (epimerisation, methylation, heterocyclization, ...) modify the incorporated amino acid
Composition diversity

▶ use of nonproteogenic amino acids
▶ secondary domains (epimerisation, methylation, heterocyclization, ...) modify the incorporated amino acid
▶ possible incorporation of fatty acids or carbohydrates
Composition diversity

- use of nonproteogenic amino acids
- secondary domains (epimerisation, methylation, heterocyclization, ...) modify the incorporated amino acid
- possible incorporation of fatty acids or carbohydrates
- possible incorporation of PKS (Polyketide synthase) part (NRPS/PKS hybrids)
- use of the term monomer rather than amino acid
Biological activity diversity

- **antibiotics** : *ACV tripeptide* (penicillin precursor), *Daptomycin* (Cubicin, treatment of infections caused by Gram positive bacteria)

- **immuno-modulating** : *cyclosporin* (immunosuppressant drug used in organ transplant)

- **anti-tumoral** : *bleomycin* (used in the treatment of some cancers)

- **siderophores** : *pyoverdines* (iron chelating molecules in response to iron limitation in the environment)

- **toxins** : *microcystins* (toxic for plants and animals including humans, may cause serious damage to the liver)

- ...  

- intense research to obtain new active products
Existing databanks and databases

- **generalist**:  
  - nucleic sequences (GenBank, EMBL-Bank, DDBJ) → cluster of synthetases  
  - proteic sequences (UniProt) → synthetases  
  - 3D structures (wwPDB) → modules, domains and peptides  
  - proteic domains (Pfam) → synthetase domains  
  - molecules (PubChem) → peptides

- **fragmentary information about NRPs diluted in a lot of data**

- **specialized**:  
  - peptaibol database : dedicated to peptaibols (300 entries)  
  - NRPS-PKS database : 17 NRPS clusters  
  - ClustScan database : 5 PKS clusters

- **no database explicitly listing nonribosomal peptides**
Existing bioinformatic tools

- NRPS/PKS analysis web site, [Conti et al., 1997]
  - input : NRPS proteic sequence
  - output : synthetase organization, adenylation domain specificity

- NRPS-PKS database, [Ansari et al., 2004]
  - input : NRPS proteic sequence
  - output : synthetase organization, adenylation domain specificity

- NRPSpredictor, [Rausch et al., 2005]
  - input : NRPS proteic sequence
  - output : adenylation domain specificity (tSVMs)
Existing bioinformatic tools

► **ClustScan**, [Starcevic et al., 2008]
  → *input*: genome
  → *output*: modular biosynthesis cluster, prediction of chemical structures

► **Clusean**, [Weber et al., 2009]
  → *input*: bacterial genome
  → *output*: secondary metabolite biosynthetic gene clusters

► no tool for peptide analysis
Goals of the work

Context:

NRPs are of great interest and show a lot of specificities:

- several hundreds of different monomers incorporated vs 20 for ribosomal proteins
- structures possibly with cycles or branchings vs linear structure for ribosomal proteins
- small size (7-8 in average) vs large size (several hundreds)

→ impossible to use tools developed for ribosomal proteins

- no database and no tools dedicated to NRPs

→ lack of a global view on the diversity makes the NRP analysis difficult

Goals:

- development of a database dedicated to NRPs
- development of tools for NRP analysis
Outline

Work context
Nonribosomal peptide synthesis and products
Issue and tools

Tools for comparing NRPs
NRP and pattern modeling
Method for pattern matching

Database dedicated to NRPs
Norine
Web interface

Mining Norine data
Statistical study of Norine data
Activity prediction tool
NRP modeling

we propose to represent NRPs by their monomeric structure

manual conversion

devolved formula

monomeric structure

NRP structures are modeled by non-oriented labeled graphs

non-oriented justified by the existence of nonpeptidic bonds
Pattern matching

- identifies a pattern of size $k$ in all the graphs contained in the database
- need of an efficient method for pattern matching
Pattern modeling

- non-oriented labeled graphs
- additional symbols

- X represents any monomer
- / several possible monomers at a given position
Compatibility Graph

- method widely used in chemoinformatics, [Raymond and Willett, 2002] is based on the compatibility graph (CG) (encoding potential mappings between two graphs)
- search for a $k$-clique in CG, corresponding to a common substructure of graphs $P$ and $G$
Compatibility Graph

Pattern P

Graph G

compatibility graph
Compatibility Graph

Pattern P

Graph G

compatibility graph
Compatibility Graph

Pattern P

Graph G

compatibility graph
Compatibility Graph

Pattern P

Graph G

compatibility graph
Compatibility Graph

Pattern P

Graph G

compatibility graph
Compatibility Graph

Pattern P

Graph G

compatibility graph

Graph G

Pattern P
Compatibility graph: problems

Problems:

- a lot of edges due to the case not-adjacent/not-adjacent
  → big CGs, search for a $k$-clique takes a lot of time

<table>
<thead>
<tr>
<th>pattern</th>
<th>peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear pattern of 19 X</td>
<td>alamethicin (19 monomers)</td>
</tr>
</tbody>
</table>

compatibility graph
380 nodes
53,010 edges

- a linear pattern cannot be found in a cyclic form
redefining CG building rules

- based on elementary paths (EPs) = paths from one node to another without loops, [Caboche et al., 2009]
redefining CG building rules

- redefining CG building rules
- based on elementary paths (EP)

![Graph showing redefined CG building rules]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>{0}</td>
<td>{1}</td>
<td>{2}</td>
<td>{3}</td>
<td>{4}</td>
<td>{4}</td>
<td>{3}</td>
</tr>
<tr>
<td>1</td>
<td>{0}</td>
<td>{1}</td>
<td>{2}</td>
<td>{3,4}</td>
<td>{3,4}</td>
<td>{2}</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>{0}</td>
<td>{1,4}</td>
<td>{2,3}</td>
<td>{2,3}</td>
<td>{1,4}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>{0}</td>
<td>{1,4}</td>
<td>{2,3}</td>
<td>{2,3}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>{0}</td>
<td>{1,4}</td>
<td>{2,3}</td>
<td>{2,3}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>{0}</td>
<td>{1,4}</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>{0}</td>
<td></td>
</tr>
</tbody>
</table>
Elementary path method

pattern

D–Tyr  Ser  Ala

compatibility graph

D–Tyr

graph

Ala  Ser

\( \text{a) } 0,0 \)

\( \text{b) } 1,1 \)

\( \text{c) } 2,2 \)
Elementary path method

pattern

D−Tyr  Ser  Ala

{1}

compatibility graph

D−Tyr

{1,2}

graph

Ala  Ser

0,0  1,1  2,2
Elementary path method

Pattern:
- D-Tyr
- Ser
- Ala

Compatibility graph:
- Nodes labeled 0, 1, 2
- Edges connecting nodes:
  - 0,0
  - 1,1
  - 2,2

Graph:
- Nodes labeled 0, 1
- Edges connecting nodes:
  - 0,0
  - 1,1
  - 2,2
Elementary path method

pattern

D-Tyr  Ser  Ala

compatibility graph

D-Tyr  Ser  Ala

3-clique

a

0,0

b

1,1

c

2,2
Elementary path method

pattern

D-Tyr  Ser  Ala

graph

D-Tyr  0

Ala  2

Ser  1

compatibility graph

no 3–clique with classical method
Comparison of the number of edges

- CG method
- EP method
Comparison of the number of edges

- CG method
- EP method

<table>
<thead>
<tr>
<th>Number of Edges</th>
<th>Number of Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>53 010</td>
<td>380</td>
</tr>
<tr>
<td>3 948</td>
<td>500</td>
</tr>
</tbody>
</table>
### Performance comparison

<table>
<thead>
<tr>
<th>pattern</th>
<th>number of results</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CG</td>
<td>EP</td>
</tr>
<tr>
<td>X__X</td>
<td>698</td>
<td>711</td>
</tr>
<tr>
<td>X__X_5__X</td>
<td>332</td>
<td>511</td>
</tr>
<tr>
<td>X__X_9__X</td>
<td>113</td>
<td>175</td>
</tr>
<tr>
<td>X__X_12__X</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>X__X_16__X</td>
<td>?</td>
<td>24</td>
</tr>
<tr>
<td>X__X_47__X</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X__X_14__X__X</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>
## Performance comparison

<table>
<thead>
<tr>
<th>pattern</th>
<th>number of results</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CG</td>
<td>EP</td>
</tr>
<tr>
<td>X_X</td>
<td>698</td>
<td>711</td>
</tr>
<tr>
<td>X_X5_X</td>
<td>332</td>
<td>511</td>
</tr>
<tr>
<td>X_X9_X</td>
<td>113</td>
<td>175</td>
</tr>
<tr>
<td>X_X12_X</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>X_X16_X</td>
<td>?</td>
<td>24</td>
</tr>
<tr>
<td>X_X47_X</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X_X14_X_X / X</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>
## Performance comparison

<table>
<thead>
<tr>
<th>pattern</th>
<th>number of results</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CG</td>
<td>EP</td>
</tr>
<tr>
<td>X_X</td>
<td>698</td>
<td>711</td>
</tr>
<tr>
<td>X_X5_X</td>
<td>332</td>
<td>511</td>
</tr>
<tr>
<td>X_X9_X</td>
<td>113</td>
<td>175</td>
</tr>
<tr>
<td>X_X12_X</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>X_X16_X</td>
<td>?</td>
<td>24</td>
</tr>
<tr>
<td>X_X47_X</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X_X14_X_X</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

EP method is **efficient** and **appropriate** to identify linear pattern in cyclic form.
method extension

- extension to exact entire structure matching (even faster)
- extension to approximate comparison (similarity search) via maximal clique computation. Definition of a similarity distance.
Outline

Work context
Nonribosomal peptide synthesis and products
Issue and tools

Tools for comparing NRPs
NRP and pattern modeling
Method for pattern matching

Database dedicated to NRPs
NORINE
Web interface

Mining Norine data
Statistical study of NORINE data
Activity prediction tool
**Norine**: a database dedicated to NRPs

- Database contains more than 1,000 peptides, [Caboche et al., 2008], first version of Norine (250 peptides) available since 2006
- Manually curated annotations from scientific literature
- Freely available: [http://bioinfo.lifl.fr/norine](http://bioinfo.lifl.fr/norine)

- Web interface: easy search for peptides corresponding to various criteria
  - General search
  - Structure search
When several fields are selected, the results must match each of them.

- **Basic search**

  - **by name ID** (ex: NCP00681):
  - **by name** (general or specific):
  - **by status**:
  - **by activity**:
  - **by class**:
  - **by structure type**:
  - **by molecular weight**:
    - ranging between _______ and _______
    - peptides containing a **number of monomers**: _______
    - peptides containing the **monomer**:
    - peptides containing a derivative of the **monomer**: _______

  * search for the **monomer code**, the underscore symbol '_' replace one character

- **Bibliography reference search**

  - author _______
  - author _______

- **Organism search**

  - **bacillus subtilis**
  - by bacteria type: **Gram** None

[Reset] [Submit]
- list of peptides corresponding to search criteria
- description of each peptide
fengycin A

- **Peptide**
  - **Name Id:** NOP00126
  - **general:** fengycin
  - **activity:** antibiotic; surfactant
  - **class:** peptide; lipopeptide
  - **formula:** C72H110O1220
  - **molecular weight:** 1443.7123 g/mol
  - **comments:** The lipid moiety of both analogs is more variable, as fatty acids have been identified as anteiso-pentadecanoyl acid (i-C15), iso-hexadecanoyl acid (i-C16), and there is evidence for further saturated and unsaturated residues up to C18.

- **Entry Information**
  - **status:** curated
  - **last modification date:** 2008-06-04
    - Home Team, UML (UMFRGS, USTL), INRA, France, ProbiotGEM (UPR/ES EA 1025 USTL), France.
  - **View all entry history**

- **Structure**
  - **type:** cyclic
  - **number of monomers:** 11
  - **monomeric composition:** C16:0-0H(3), Glu,D-OH, Tyr,D-aThr,Glu,D-Ala, Pro,Glu,D-Tyr,Le
  - **linear representation:** C16:0-0H(3), Glu,D-OH, Tyr,D-aThr,Glu,D-Ala, Pro,Glu,D-Tyr,Le
  - **graph representation:** C16:0-0H(3), Glu,D-OH, Tyr,D-aThr,Glu,D-Ala, Pro,Glu,D-Tyr,Le
  - **Visualization:**
    - [Click](#)

- **Organisms**
  - **Bacillus subtilis**
    - **taxonomy:** cellular organisms; Bacteria; Firmicutes; Bacillae; Bacillaceae; Bacillus
    - **Gram positive**
    - **synonyms:** Bacillus amyloliquefaciens, Bacillus natto, Bacillus globigii, Vibrio subtilis
    - **taxid:** 1423

- **References**

- **Links**
  - [UniProt](https://www.uniprot.org/uniprot/0389980)
fengycin A

- **Peptide**
  - **Name ID:** HQR0228
  - **general:** lipopeptide
  - **activity:** antibiotic; surfactant
  - **class:** peptide, lipopeptide
  - **formula:** C72H110N12O20
  - **molecular weight:** 1493.7123 g/mol
  - **comment:** The lipopeptide of both analogs is more variable, as fatty acids have been identified as anteiso-pentadecanoic acid (ai-C15), iso-hexadecanoic acid (i-C16), n-hexadecanoic acid (n-C16), and there is evidence for further saturated and unsaturated residues up to C18.

- **entry information**
  - status: curated
  - last modification date: 2008-08-04
  - Yves Demont, LIFL (UMR8022 CNRS/USTL)-INRIA, France, ProBioSEM (UPR6 EA1026 USTL), France.
  - View all entry history

- **Structure**
  - **type:** partial cyclic
  - **number of monomers:** 11
  - **monomeric composition:** C16-D-OH, Glu-D-OH, Tyr-D-oThr, Glu-D-Ala, Pro, Glu-D-Tyr,3e
  - **linear representation:** C19-OH(3)-Glu-D-OH(1)-Tyr-D-oThr-Glu-D-Ala-Pro-Glu-D-Tyr,3e
  - **graph representation:** C16-D-OH(3)-Glu-D-OH(1)-Tyr-D-oThr-Glu-D-Ala-Pro-Glu-D-Tyr,3e

- **Organisms**
  - **Bacillus subtilis**
    - **taxonomy:** cellular organisms; Bacteria; Firmicutes; Bacilli; Bacillus subtilis; Bacilliaceae; Bacillus
    - **Gram:** positive
    - **synonyms:** Bacillus unifilis, Bacillus unifilis, Bacillus subtilis
    - **txid:** 4122

- **References**
  - **Fengycin, a novel antifungal lipopeptide antibiotic produced by Bacillus subtilis F-29-3.**
  - **Neurobiological synthesis of fengycin on an enzyme complex formed by fengycin synthetases.**

- **Links**
  - PubChem [443591]
  - UniProt [CD00880]
Orn

- **complete name**: Ornithine
- **synonym(s)**: L-Orn, 5-amino-
- **type**: NRPS
- **molecular formula**: C5H12N2O2
- **molecular weight**: 132.16098
- **IUPAC**: (2S)-2,5-diaminopentanoic acid
- **SMILES**: C\((\text{C}(=\text{O})\text{O})\text{N})\text{CN}
- **isomeric SMILES**: C\((\text{C}(=\text{O})\text{O})\text{N})\text{CN}
**Peptide**

- **Non-IID:** I0P00126
- **general:** fengycin
- **activity:** antibiotic; surfactant
- **class:** peptide; lipopeptide
- **formula:** C72H110N12O20
- **molecular weight:** 1463.7123 g/mol
- **comment:** The lipid moiety of both analogs is more variable, as fatty acids have been identified as antisorso-pentadecanoic acid (a-C15), iso-octadecanoic acid (i-C18), and there is evidence for further saturated and unsaturated residues up to C18.

**Entry Information**

- status: curated
- last modification date: 2008-06-04
- Home Team, UMR CNRS USTL-CHIMIE, Laboratoire de Chimie des Biomolecules, UPR 9005, USTL, France.
- View all entry history

**Structure**

- **type:** cyclic
- **number of monomers:** 11
- **monomeric composition:** C16, 0-CH(O)-Glu-D-OMe-Tyr-D-aThr-Glu-D-Ala-Pro-Glu-D-Tyr-ile
- **linear representation:** C16:0-CH(O)-Glu-D-OMe-Tyr-D-aThr-Glu-D-Ala-Pro-Glu-D-Tyr-ile
- **graph representation:** C16:0-CH(O)-Glu-D-OMe-Tyr-D-aThr-Glu-D-Ala-Pro-Glu-D-Tyr-ile
- **Visualization:**

![Visualization](image)

**Organisms**

- **Bacillus subtilis**
  - **taxonomy:** cellular organisms; Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus
  - **Gram positive**
  - **synonyms:** Bacillus amyloliquefaciens, Bacillus natto, Bacillus gliroagri, Vibrio subtilis
  - **taxid:** 1423

**References**


**Links**

- [UniProt](https://www.uniprot.org/uniprot/039980)
Example of use: analysis of predicted peptides

- illustrate the structure search interface
- genome of *Pseudomonas entomophila L48* (2006): search for NRPS clusters (search for genes annotated as *putative NRPS* and by BLAST)
- identification of 4 NRPS clusters
Example of use: analysis of predicted peptides

- use of NRPS/PKS analysis web site, NRPS/PKS database and NRPSpredictor to obtain the synthetase organization and adenylation domain specificity
  - **cluster 1**: 5 monomers
    predicted monomeric chain: Arg_Pro-Thz_X_Ile_Pro
  - **cluster 2**: 5 monomers
    predicted monomeric chain: Gly_Thr_Ile_X_Glu
  - **cluster 3**: 12 monomers
    predicted monomeric chain: X_Val_Leu_X_Val_Leu_X_Ser_Val_Leu_Ser_Leu/Ile
  - **cluster 4**: 10 monomers
    predicted monomeric chain: D-Ala_X_X_D-Asp_Gly_Gly_Ser_Thr_D-Ser_X
Example of use: analysis of predicted peptides

- use of **Norine** tools in order to obtain new information on predicted structures
  - pattern search (search for an incomplete structure with undefined monomers)
  - composition search (search for a list of monomers, useful when the exact structure is not known)
  - similarity search (search for a structure with possible monomer substitutions)
P. entomophila Cluster 3

- **Structure-based search** [?]

  peptide(s) which match exactly the structure: 

  Editor
  
  [Reset] [submit]

  peptide(s) containing the structural pattern: 

  ○ containing the complete pattern
  ○ containing the pattern substructure with at least [ ] monomers

  Editor
  
  [Reset] [submit]
P. entomophila Cluster 3

- **Structure-based search**

  peptide(s) which match exactly the structure: 

  Editor

  [Reset] [submit]

  peptide(s) containing the structural pattern: 

  ○ containing the complete pattern

  ○ containing the pattern substructure with at least [ ] monomers

  Editor

  [Reset] [submit]

---

**Non Ribosomal Peptide Editor**

[Help]

[add Monomer] [create Monomer] [add Links] [delete Monomer] [Go] [reset All] [open]

P. entomophila Cluster 3

- **Structure-based search**

  peptide(s) which match exactly the structure: [Input]
  
  **Editor**
  Reset  
  submit

  peptide(s) containing the structural pattern: X,Val,Leu,X,Val,Leu,X,Ser,Val,Leu,Ser,Leu/Ile@1@0,2@1,3@2,4@3
  containing the complete pattern
  containing the pattern substructure with at least [Input] monomers
  
  **Editor**
  Reset  
  submit

  *sorry, no peptide contains the pattern:*

  X,Val,Leu,X,Val,Leu,X,Ser,Val,Leu,Ser,Leu/Ile@1@0,2@1,3@2,4@3,5@4,6@5,7@6,8@7,9@8,10@9,11@10

  The putative NRPS products appear in **green color**.
P. entomophila Cluster 3

- **Structure-based search**

peptide(s) which match exactly the structure: 

Editor

Reset  submit

peptide(s) containing the structural pattern:

X, Val, Leu, X, Val, Leu, X, Ser, Val, Leu, Ser, Ile/Leu@1@0, 2@1, 3@2, 4@5

- containing the complete pattern
- containing the pattern substructure with at least 4 monomers

Editor

Reset  submit

There are 25 results with a pattern substructure of at least 4/12 monomers:

Select all | Deselect all

Download

- syringafactin A
- syringafactin D
- bergofungin B
- putisolvin I
- trichobrachin A4
- trichobrachin A2
- cyclolinopeptide A
- phakellistatin 8
- cyclosporin U

lipopeptide antibiotic, surfactant
P. entomophila Cluster 3

- **Composition-based search [?]**

  peptide(s) containing the monomers* (separated by a comma):
  
  Val,Leu,Val,Leu,Ser,Val,Leu,Ser

  and a maximum number of errors of [2] monomer(s)

  [Reset] [submit]

  * search for the monomer code

  there are 7 results

Select all | Deselect all

Download

1 peptide containing the monomeric composition with 1 error(s)

- putisolvin I  lipopeptide (antibiotic, surfactant)

6 peptides containing the monomeric composition with 2 error(s)

- putisolvin II
- putisolvin III
- syringopeptin SC 1
- syringopeptin 508B
- syringopeptin SC 2
- syringopeptin 508A

lipopeptide (antibiotic, surfactant)

lipopeptide (antibiotic, toxin)

The putative NRPS products appear in **green color**.
P. entomophila Cluster 3

- **Similarity-based search**

  peptide(s) similar with: X,Val,Leu,X,Val,Leu,X,Ser,Val,Leu,Ser,Ile/Leu@1@0,2@1,3@2,4@3

  - no clustering
  - clustering 1
  - clustering 2

  Editor

  ![Submit Button]

<table>
<thead>
<tr>
<th>distance</th>
<th>common monomers</th>
<th>peptide</th>
<th>download</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin C</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin B</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin II</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin A</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin D</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin E</td>
<td>□</td>
</tr>
<tr>
<td>0.631578947368421</td>
<td>7</td>
<td>tolaasin F</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>callynomine 8</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>massetolide F</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>massetolide E</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>viscosin</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>cyclosporin 27</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>lVal7surfactin</td>
<td>□</td>
</tr>
<tr>
<td>0.6999999999999999</td>
<td>4</td>
<td>cyclosporin 8</td>
<td>□</td>
</tr>
</tbody>
</table>
P. entomophila Cluster 3

Results

distance: 0.6666666666666667
number of common monomers: 4
clustering: clustering1
*The common monomers between the two peptides appear in red.*

- **Query structure:**

  total number of monomers: 12
  X,Val,Leu,X,Val,Leu,X,Ser,Val,Leu,Ser,Ile/Leu@1@0,2@1,3@2,4@3,5@4,6@5,7@6,8@7,9@8,10@9,11@10

  View structure:
  
  Click

- **Matching Structure:**

  corresponding peptide: White Line Inducing Principle
  total number of monomers: 10
  FA,Leu,Glu,aThr,Val,Leu,Ser,Leu,Ser,Ile@1@0,2@1,3@2,4,9@3,5@4,6@5,7@6,8@7,9@3,8

  View structure:
  
  Click
P. entomophila Cluster 3

- Cluster 3 seems to produce a lipopeptide
- probably also sufactant and antibiotic
P. entomophila Cluster 4

- same approach for cluster 4
- Cluster 4 seems to produce a new pyoverdine
- pyoverdines are siderophores (chelating iron) produced by *Pseudomonas* species
- pyoverdines are fluorescent molecules
P. entomophila Cluster 4

a) B. subtilis 168, negative control b) P. entomophila L48 c) P. aeruginosa 7NSK2, positive control

▶ similar results reported in a recent study, [Matthijs et al., 2009]
Outline

Work context
Nonribosomal peptide synthesis and products
Issue and tools

Tools for comparing NRPs
NRP and pattern modeling
Method for pattern matching

Database dedicated to NRPs
Norine
Web interface

Mining Norine data
Statistical study of Norine data
Activity prediction tool
Statistical study

- difficult to comprehend the nonribosomal peptide diversity (structures, monomers, ...)

- first large scale statistical analyses of about a thousand peptides that represents a total coverage of more than 10,000 monomers, revealing the presence of as many as 500 different monomers
Structure type distribution

- Less than 30% of NRPS peptides are linear
- Comprehend the structure diversity
Study of producing organisms

- synthetase genes have been observed in bacteria and fungi but not in sponges or other metazoa

- hypothesis:
  - peptides synthesized by bacteria harbor features different from those synthesized by fungi
  - the peptides isolated from sponges are in reality synthesized by symbiotic bacteria rather than the sponges themselves,

[Piel, 2009]
Study of producing organisms

- Study of monomer distribution in peptides produced by bacteria, fungi and metazoa (mainly sponges)

- Compute the correlation coefficient (CC) between monomeric distributions (more the CC is close to 0 more the monomeric distributions are unrelated)

<table>
<thead>
<tr>
<th>organism 1</th>
<th>organism 2</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>bacteria</td>
<td>fungi</td>
<td>0.273</td>
</tr>
<tr>
<td>bacteria</td>
<td>metazoa</td>
<td>0.566</td>
</tr>
<tr>
<td>fungi</td>
<td>metazoa</td>
<td>0.370</td>
</tr>
</tbody>
</table>

- monomers used by bacteria are different to those used by fungi
- monomers used by bacteria are similar to those incorporated in peptides isolated from metazoa
Study of producing organisms

- peptide size distribution:
  - are different between bacteria and fungi
  - are similar between bacteria and metazoa

these results support the hypothesis
Study of biological activities

- Toxin (142)
- Antibiotic (488)
- Immuno-modulator (30)
- Surfactant (130)
- Anti-tumor (31)
- Siderophore (88)
Study of biological activities

▶ **hypothesis**:
  ▶ monomers are specific to a class of activities
  ▶ study of monomer distribution for different classes of activities
Study of biological activities

- monomer distribution in siderophores
Study of biological activities

- monomers are specific to a class of activities
- Idea: monomeric composition can help to predict the biological activities of a peptide
Activity prediction tool

- **purpose**: help to predict the biological activities of a peptide from its monomeric composition

- six major classes in **Norine**: antibiotic (488), anti-tumor (31), immuno (30), siderophore (88), surfactant (130) and toxin (142)

- the thesis manuscript version can be improved:
  - by refining the prediction significance: From which threshold the predicted activity is significant?
  - by returning several activities
  - by minimizing the impact of rare monomers

- method extension
Method extension

- first step: normalize number of monomers in order to have the same number of monomers in each class

- a score $S$ of a monomer $M$ in a class $C$ is calculated as follows:

  $$ S^C_M = \text{Freq}(M_C) \cdot \frac{1}{-\log(\text{Freq}(M_{data}))} $$

- the prediction score $S_P$ of a peptide of length $N$ in the class $C$ is calculated as follows:

  $$ S^C_P = \frac{\sum_{i=1}^{N} S^C_{M_i}}{N} $$
Method extension

![Scatter plot showing the relationship between score siderophore and score antibiotic]
Method extension

- generating random monomeric compositions
- normality test: the score distribution of random peptides follows the normal law
- compute $P(X > S_{obs})$ and get a p-value
Results

- data sets of Norine peptides (the same as in the thesis manuscript)

- prediction of several activities, i.e. marinobactin B annotated as siderophore and surfactant and predicted to have siderophore and surfactant activities

- complete annotation, i.e. actinomycin C2, annotated as antibiotic, the extended method predicts this peptide to have antitumoral activity that was confirmed in scientific literature → can help to identify new activities of a known peptide
Test on NRPs not in Norine

- **trichotoxins**, antibiotics synthesized by *Trichoderma asperellum* (fungi)

<table>
<thead>
<tr>
<th>activities</th>
<th>first version</th>
<th>extended version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>score</td>
<td>score</td>
</tr>
<tr>
<td>antibiotic</td>
<td>0.89</td>
<td>0.42</td>
</tr>
<tr>
<td>antitumor</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>immuno</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>siderophore</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>surfactant</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>toxin</td>
<td>0.06</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- property confirmed
Test on NRPs not in Norine

- **fuscachelins**, siderophores synthesized by *Thermobifida fusca* (bacteria)

<table>
<thead>
<tr>
<th>activities</th>
<th>first version</th>
<th>extended version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>score</td>
<td>score</td>
</tr>
<tr>
<td>antibiotic</td>
<td>0.37</td>
<td>0.12</td>
</tr>
<tr>
<td>antitumor</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>immuno</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>siderophore</td>
<td><strong>0.40</strong></td>
<td><strong>0.25</strong></td>
</tr>
<tr>
<td>surfactant</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>toxin</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- Removes any doubt
Conclusion

- Development of **Norine**, the first public resource dedicated to nonribosomal peptides
- **Norine** became the reference resource for nonribosomal peptides (more than 200 connections/month)

Geographic distribution of **Norine** users
Conclusion

- Development of efficient tools for the analysis of nonribosomal peptides
  - pattern matching
  - biological activity prediction
  - ...

- Improvement of nonribosomal peptide knowledge
  - number of NRPs (more than 1,000)
  - structure diversity
  - monomer diversity (more than 500 different monomers)
  - ...
Perspectives

- improvement of similarity search (nodes insertion/deletion)
- continue the development of biological activity prediction
- Gene Ontology annotations (pathway and target of a peptide) in order to better understand the structure/function relationships
- close collaboration with wwPDB (post-doctoral position)
- collaboration with Daslav Hranueli team (University of Zagreb, Croatia) on the development of synthetase database linked to Norine
- collaboration with Pavel Pevzner team (University of California San Diego, USA) on integration of mass spectrometry data
<table>
<thead>
<tr>
<th>Work context</th>
<th>Tools for comparing NRPs</th>
<th>Database dedicated to NRPs</th>
<th>Mining Norine data</th>
<th>Perspectives</th>
</tr>
</thead>
</table>

Thank you for your attention
<table>
<thead>
<tr>
<th>Work context</th>
<th>Tools for comparing NRPs</th>
<th>Database dedicated to NRPs</th>
<th>Mining Norine data</th>
<th>Perspectives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work context</td>
<td>Tools for comparing NRPs</td>
<td>Database dedicated to NRPs</td>
<td>Mining Norine data</td>
<td>Perspectives</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-------------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>