Set Covering Machines and Reference-Free Genome Comparisons Uncover Predictive Biomarkers of Antibiotic Resistance

Alexandre Drouin

Department of Computer Science and Software Engineering, Université Laval
Department of Molecular Medicine, Université Laval

CCBC/GLBIO 2016
Toronto, Canada
May 17, 2016
Introduction

Biomarker: a **measurable characteristic** that is **predictive** of a phenotype

- Better understand the biological processes involved
- Develop diagnostic tests, new therapies and drug treatments
Formalization as a Supervised Learning Problem

Data Set

A data set is a collection of genomes and associated labels:

\[ S \stackrel{\text{def}}{=} \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \sim D^m \]

- \( x \in \mathcal{X} \stackrel{\text{def}}{=} \{A, C, G, T\}^* \) is a genome
- \( y \in \{0, 1\} \) is a label (control or case)
- \( D \) is a data generating distribution

Objective

1. Define a suitable (vectorial) representation for genomes \( \phi : \mathcal{X} \rightarrow \mathbb{R}^d \)
2. Find a predictor \( h : \mathbb{R}^d \rightarrow \{0, 1\} \) that has a good generalization performance, i.e. that minimizes:

\[ R(h) \stackrel{\text{def}}{=} \Pr_{(x,y) \sim D} [h(\phi(x)) \neq y] \]

3. The predictor must be interpretable!
Formalization as a Supervised Learning Problem

Data Set

A data set is a collection of genomes and associated labels:

\[ S \overset{\text{def}}{=} \{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \} \sim D^m \]

\( x \in \mathcal{X} \overset{\text{def}}{=} \{ A, C, G, T \}^* \) is a genome
\( y \in \{ 0, 1 \} \) is a label (control or case)
\( D \) is a data generating distribution

Objective

1. Define a suitable (vectorial) representation for genomes \( \phi : \mathcal{X} \rightarrow \mathbb{R}^d \)
2. Find a predictor \( h : \mathbb{R}^d \rightarrow \{ 0, 1 \} \) that has a good generalization performance, i.e. that minimizes:

\[ R(h) \overset{\text{def}}{=} \Pr_{(x, y) \sim D} [ h(\phi(x)) \neq y ] \]

3. The predictor must be interpretable!
Formalization as a Supervised Learning Problem

Data Set

A data set is a collection of genomes and associated labels:

\[ S \overset{\text{def}}{=} \left\{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \right\} \sim D^m \]

- \( x \in \mathcal{X} \overset{\text{def}}{=} \{ A, C, G, T \}^* \) is a genome
- \( y \in \{ 0, 1 \} \) is a label (control or case)
- \( D \) is a data generating distribution

Objective

1. Define a suitable (vectorial) representation for genomes \( \phi : \mathcal{X} \to \mathbb{R}^d \)
2. Find a predictor \( h : \mathbb{R}^d \to \{ 0, 1 \} \) that has a good generalization performance, i.e. that minimizes:

\[
R(h) \overset{\text{def}}{=} \Pr_{(x, y) \sim D} [h(\phi(x)) \neq y]
\]

3. The predictor must be interpretable!
Formalization as a Supervised Learning Problem

Data Set

A data set is a collection of genomes and associated labels:

\[ S \overset{\text{def}}{=} \{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \} \sim D^m \]

- \( x \in \mathcal{X} \overset{\text{def}}{=} \{ A, C, G, T \}^* \) is a genome
- \( y \in \{ 0, 1 \} \) is a label (control or case)
- \( D \) is a data generating distribution

Objective

1. Define a suitable (vectorial) representation for genomes \( \phi : \mathcal{X} \rightarrow \mathbb{R}^d \)
2. Find a predictor \( h : \mathbb{R}^d \rightarrow \{ 0, 1 \} \) that has a good generalization performance, i.e. that minimizes:

\[ R(h) \overset{\text{def}}{=} \Pr_{(x,y) \sim D} [ h(\phi(x)) \neq y ] \]

3. The predictor must be interpretable!
# Genome Representation

## Definition

- **k-mer**: a sequence of *k* nucleotides
- **\( \mathcal{K} \)**: the set of all *k*-mers that are in at least one genome of *S*

Each genome is represented by its **k-mer profile**, which is a binary vector indicating the presence (1) or absence (0) of each *k*-mer.

\[
\mathcal{K} = \{ \text{CAGATA, GATAGA, GAACAG, CGATGA, AGATAG, AGAACA, ATAGAA, CCGGCT, AACAGC, TAGAAC, TTTCGG, AAATAC} \}
\]

\[
\mathbf{x} = \text{CAGATAGAACAGC}
\]

\[
\phi(\mathbf{x}) = \begin{bmatrix}
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\]
The Set Covering Machine (Marchand and Shawe-Taylor, 2002)

- **Input:** A set of *boolean-valued rules* (presence/absence of $k$-mers) and a *data set* (examples, labels)

- **Objective:** Find the *shortest* conjunction (logical-AND) or disjunction (logical-OR) of rules that *most accurately predicts* the labels
The Set Covering Machine (Marchand and Shawe-Taylor, 2002)

- **Input:** A set of boolean-valued rules (presence/absence of $k$-mers) and a data set (examples, labels)

- **Objective:** Find the shortest conjunction (logical-AND) or disjunction (logical-OR) of rules that most accurately predicts the labels
The Set Covering Machine (Marchand and Shawe-Taylor, 2002)

- **Input:** A set of **boolean-valued rules** (presence/absence of k-mers) and a **data set** (examples, labels)

- **Objective:** Find the **shortest** conjunction (logical-AND) or disjunction (logical-OR) of rules that **most accurately predicts** the labels

```
\phi(x) \rightarrow \text{Rule} \rightarrow \text{True} \quad \phi(x) \rightarrow \text{Rule} \rightarrow \text{False}
```

```
\text{Candidate Rules} \quad \text{Data Set}
\begin{array}{llll}
\text{Rule} & \text{Rule} & \cdots & \text{Rule} \\
\text{x1} & \text{y1} & \cdots & \text{xm} \\
\text{y2} & \text{y3} & \cdots & \text{ym}
\end{array}
```

```
\phi(x) \rightarrow \text{Rule} \rightarrow \text{AND} \rightarrow \text{Rule} \rightarrow \text{AND} \rightarrow \text{Rule}
```

```
\text{Model (h)}
\begin{array}{llll}
\text{Rule} & \text{AND} & \text{Rule} & \text{AND} \\
\text{y = 1} & \text{True} \rightarrow \text{Rule} \\
\text{y = 0} & \text{False \rightarrow Rule}
\end{array}
```
What is special about this learning algorithm?

- Feature selection is **not required**
- **Out-of-core** implementation: data is accessed in small blocks
- **Theoretical performance guarantees** that are backed by empirical results
What is special about this learning algorithm?

- Feature selection is **not required**
- **Out-of-core** implementation: data is accessed in small blocks
- **Theoretical performance guarantees** that are backed by empirical results
What is special about this learning algorithm?

- Feature selection is **not required**
- **Out-of-core** implementation: data is accessed in small blocks
- **Theoretical performance guarantees** that are backed by empirical results
The genomes were assembled and subsequently split into 31-mers.

Between 8,058,479 and 123,466,989 \( k \)-mers.
Data sets

The genomes were assembled and subsequently split into 31-mers. Between 8,058,479 and 123,466,989 k-mers.

Talk by Frédéric Raymond, Wednesday at 4:40 pm
The models are accurate and sparse

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.970</td>
<td>3.3</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.927</td>
<td>2.6</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.980</td>
<td>3.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.979</td>
<td>1.4</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.980</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.821</td>
<td>1.4</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.979</td>
<td>1.0</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.682</td>
<td>3.1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.969</td>
<td>1.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.950</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.825</td>
<td>4.9</td>
</tr>
<tr>
<td>Doripenem</td>
<td>0.730</td>
<td>1.4</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td>0.927</td>
<td>1.2</td>
</tr>
<tr>
<td>Meropenem</td>
<td>0.733</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.987</td>
<td>1.1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.963</td>
<td>2.0</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.969</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- The average number of $k$-mers contributing to the model is less than 5.
- For 12/17 data sets the error rate is smaller than 10%.
- The least accurate models were obtained for *M. tuberculosis* and *P. aeruginosa*.

▶ Is the genome enough to predict antibiotic resistance?
The models are accurate and sparse

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.970</td>
<td>3.3</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.927</td>
<td>2.6</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.989</td>
<td>3.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.979</td>
<td>1.4</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.980</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.821</td>
<td>1.4</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.979</td>
<td>1.0</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.682</td>
<td>3.1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.969</td>
<td>1.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.950</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.825</td>
<td>4.9</td>
</tr>
<tr>
<td>Doripenem</td>
<td>0.730</td>
<td>1.4</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td>0.927</td>
<td>1.2</td>
</tr>
<tr>
<td>Meropenem</td>
<td>0.733</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.987</td>
<td>1.1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.963</td>
<td>2.0</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.969</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- The average number of $k$-mers contributing to the model is less than 5.
- For 12/17 data sets the error rate is smaller than 10%.
- The least accurate models were obtained for *M. tuberculosis* and *P. aeruginosa*.

▶ Is the genome enough to predict antibiotic resistance?
The models are accurate and sparse

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.970</td>
<td>3.3</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.927</td>
<td>2.6</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.989</td>
<td>3.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.979</td>
<td>1.4</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.980</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.821</td>
<td>1.4</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.979</td>
<td>1.0</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.682</td>
<td>3.1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.969</td>
<td>1.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.950</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.825</td>
<td>4.9</td>
</tr>
<tr>
<td>Doripenem</td>
<td>0.730</td>
<td>1.4</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td>0.927</td>
<td>1.2</td>
</tr>
<tr>
<td>Meropenem</td>
<td>0.733</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.987</td>
<td>1.1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.963</td>
<td>2.0</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.969</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- The average number of $k$-mers contributing to the model is less than 5.
- For 12/17 datasets the error rate is smaller than 10%.
- The least accurate models were obtained for *M. tuberculosis* and *P. aeruginosa*.

▶ Is the genome enough to predict antibiotic resistance?
The models are accurate and sparse

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.970</td>
<td>3.3</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.927</td>
<td>2.6</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.989</td>
<td>3.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.979</td>
<td>1.4</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.980</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.821</td>
<td>1.4</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.979</td>
<td>1.0</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.682</td>
<td>3.1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.969</td>
<td>1.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.950</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.825</td>
<td>4.9</td>
</tr>
<tr>
<td>Doripenem</td>
<td>0.730</td>
<td>1.4</td>
</tr>
<tr>
<td>Levofloxacil</td>
<td>0.927</td>
<td>1.2</td>
</tr>
<tr>
<td>Meropenem</td>
<td>0.733</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.987</td>
<td>1.1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.963</td>
<td>2.0</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.969</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- The average number of $k$-mers contributing to the model is less than 5.
- For 12/17 data sets the error rate is smaller than 10%.
- The least accurate models were obtained for **M. tuberculosis** and **P. aeruginosa**.

▶ Is the genome enough to predict antibiotic resistance?
The models are accurate and sparse

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.970</td>
<td>3.3</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.927</td>
<td>2.6</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.989</td>
<td>3.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.979</td>
<td>1.4</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.980</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.821</td>
<td>1.4</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.979</td>
<td>1.0</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.682</td>
<td>3.1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.969</td>
<td>1.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.950</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.825</td>
<td>4.9</td>
</tr>
<tr>
<td>Doripenem</td>
<td>0.730</td>
<td>1.4</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td>0.927</td>
<td>1.2</td>
</tr>
<tr>
<td>Meropenem</td>
<td>0.733</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.987</td>
<td>1.1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.963</td>
<td>2.0</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.969</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- The average number of $k$-mers contributing to the model is less than 5.

- For 12/17 data sets the error rate is smaller than 10%.

- The least accurate models were obtained for *M. tuberculosis* and *P. aeruginosa*.
  - Is the genome enough to predict antibiotic resistance?
The models are interpretable

Within hours of computation and without prior knowledge, known antibiotic resistance mechanisms have been recovered.
The models are interpretable.

---

**M. tuberculosis**

- Isoniazid
  - 1.00
  - 15
- Rifampicin
  - 0.95
  - 6
  - 1

**S. pneumoniae**

- Erythromycin
  - 0.85
  - 31
  - 3

**C. difficile**

- Azithromycin
  - 0.79
  - 2
- Clarithromycin
  - 0.83
  - 17
- Clindamycin
  - 0.23
  - 616
  - 5

---


**Isoniazid-resistance conferring mutations in Mycobacterium tuberculosis KatG: catalase, peroxidase, and INH-NADH adduct formation activities.**

Cade CE¹, Dlouhy AC, Medzihradszky KF, Salas-Castillo SP, Ghiladi RA.
The models are interpretable

\[ \begin{align*}
M.\text{ tuberculosis} & \quad S.\text{ pneumoniae} \\
\text{Isoniazid} & \quad \text{Erythromycin} \\
1.00 & \quad 0.85 \\
15 & \quad 31 \\
\text{Rifampicin} & \quad 0.07 \\
6 & \quad 19 \\
& \quad 19 \\
\text{Azithromycin} & \quad \text{Clarithromycin} \\
2 & \quad 17 \\
0.79 & \quad 0.83 \\
0.79 & \quad 0.83 \\
0.23 & \quad 0.23 \\
\text{Clindamycin} & \quad \text{Clindamycin} \\
3 & \quad 5 \\
0.85 & \quad 0.19
\end{align*} \]

\[ \begin{align*}
\text{katG} & \quad \text{Catalase-Peroxydase} \\
\text{rpoB} & \quad \text{RNA Polymerase }\beta\text{-Subunit} \\
\text{mel} & \quad \text{ABC Transporter ATPase Subunit} \\
\text{metE} & \quad \text{Methionine Synthase} \\
\text{ermB} & \quad \text{rRNA Adenine N-6-Methyltransferase}
\end{align*} \]

\[ \begin{align*}
\text{Penicillin-Binding Protein} \\
\text{Tn6194-like Conjugative Transposon} \\
\text{Tn6110 Transposon}
\end{align*} \]

---


**Rifampicin resistance and mutation of the rpoB gene in Mycobacterium tuberculosis.**

Taniguchi H\(^1\), Aramaki H, Nikaido Y, Mizuguchi Y, Nakamura M, Koga T, Yoshida S.
The models are interpretable

**M. tuberculosis**
- Isoniazid
  - 1.00
  - 15
  - 0.95
  - 6

**S. pneumoniae**
- Erythromycin
  - 0.85
  - 31

**C. difficile**
- Azithromycin
  - 0.79
  - 2

**Genomic Biomarker Discovery**


**Macrolide efflux in Streptococcus pneumoniae is mediated by a dual efflux pump (mel and mef) and is erythromycin inducible.**

*Ambrose KD*, *Nisbet R*, *Stephens DS*. 

---

**katG** - Catalase-Peroxidase

**rpoB** - RNA Polymerase β-Subunit

**mel** - ABC Transporter ATPase Subunit

**metE** - Methionine Synthase

**ermB** - rRNA Adenine N-6-Methyltransferase

**Penicillin-Binding Protein**

**Tn6194-like Conjugative Transposon**

**Tn6110 Transposon**
The models are interpretable

**M. tuberculosis**
- Isoniazid
- Rifampicin

**S. pneumoniae**
- Erythromycin

**C. difficile**
- Azithromycin
- Clindamycin

The effects of methionine acquisition and synthesis on *Streptococcus pneumoniae* growth and virulence.

The models are interpretable

**M. tuberculosis**
- Isoniazid
  - 1.00
  - 15
- Rifampicin
  - 0.95
  - 6

**S. pneumoniae**
- Erythromycin
  - 0.85
  - 31

**C. difficile**
- Azithromycin
  - 0.79
  - 2
- Clarithromycin
  - 0.79
  - 17
- Clindamycin
  - 0.83
  - 616

Research:


**The macrolide-lincosamide-streptogramin B resistance determinant from Clostridium difficile 630 contains two erm(B) genes.**

Farrow KA¹, Lyras D, Rood JI.
Going beyond \( k \)-mers... with \( k \)-mers

- Known resistance conferring mutations were retrieved (S315G, S315I, S315N and S315T)

- The model captures the absence of the wild-type sequence, efficiently including the presence of all these variants.
Known resistance conferring mutations were retrieved (S315G, S315I, S315N and S315T).

The model captures the absence of the wild-type sequence, efficiently including the presence of all these variants.
Conclusion

- **Scalable:** out-of-core machine learning-based method
- **Accurate:** models of antibiotic resistance with low error rates
- **Interpretable:** models highlight the importance of a small set of genomic loci
- **Wide applicability:** applicable to other phenotypes, organisms
- **Theoretical performance guarantees** support the applicability of the method in this challenging setting
Conclusion

- **Scalable:** out-of-core machine learning-based method

- **Accurate:** models of antibiotic resistance with low error rates

- **Interpretable:** models highlight the importance of a small set of genomic loci

- **Wide applicability:** applicable to other phenotypes, organisms

- **Theoretical performance guarantees** support the applicability of the method in this challenging setting
Conclusion

- **Scalable:** out-of-core machine learning-based method
- **Accurate:** models of antibiotic resistance with low error rates
- **Interpretable:** models highlight the importance of a small set of genomic loci
- **Wide applicability:** applicable to other phenotypes, organisms
- **Theoretical performance guarantees** support the applicability of the method in this challenging setting
Conclusion

- **Scalable:** out-of-core machine learning-based method
- **Accurate:** models of antibiotic resistance with low error rates
- **Interpretable:** models highlight the importance of a small set of genomic loci
- **Wide applicability:** applicable to other phenotypes, organisms
- **Theoretical performance guarantees** support the applicability of the method in this challenging setting
Conclusion

- **Scalable:** out-of-core machine learning-based method
- **Accurate:** models of antibiotic resistance with low error rates
- **Interpretable:** models highlight the importance of a small set of genomic loci
- **Wide applicability:** applicable to other phenotypes, organisms
- **Theoretical performance guarantees** support the applicability of the method in this challenging setting
Future Works

- *k*-mer abundances instead of presence/absence
- Phylogeny aware Set Covering Machine (population structure)
- Integrate more data (transcriptomics, epigenetics)
- Application to human genomes
Thank you!
(Poster 168, tomorrow)

Sébastien Giguère
Maxime Déraspe
François Laviolette
Mario Marchand
Jacques Corbeil

KOVER
http://github.com/aldro61/kover

Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons
Alexandre Drouin, Sébastien Giguère, Maxime Déraspe, Mario Marchand, Michael Tyers, Vivian G Loo, Anne-Marie Bourgault, François Laviolette, Jacques Corbeil
doi: http://dx.doi.org/10.1101/045153

Contact: alexandre.drouin.8@ulaval.ca
Choosing the $k$-mer Length

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SCM-31</th>
<th>SCM-CV</th>
<th>$k$-mer length</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.030 (3.3)</td>
<td>0.032 (3.2)</td>
<td>40.0 ± 23.0</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>0.073 (2.6)</td>
<td>0.085 (3.0)</td>
<td>41.0 ± 31.3</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>0.011 (3.0)</td>
<td>0.018 (2.9)</td>
<td>26.0 ± 12.0</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.021 (1.4)</td>
<td><strong>0.011</strong> (1.9)</td>
<td>56.0 ± 21.6</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td><strong>0.020</strong> (1.0)</td>
<td><strong>0.020</strong> (1.0)</td>
<td>39.0 ± 16.0</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.179 (1.4)</td>
<td><strong>0.164</strong> (1.2)</td>
<td>49.0 ± 22.7</td>
</tr>
<tr>
<td>Isoniazid</td>
<td><strong>0.021</strong> (1.0)</td>
<td><strong>0.021</strong> (1.0)</td>
<td>24.0 ± 4.6</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.318 (3.1)</td>
<td>0.353 (2.3)</td>
<td>35.0 ± 24.6</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.031 (1.4)</td>
<td><strong>0.006</strong> (1.0)</td>
<td>49.0 ± 6.0</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.050 (1.0)</td>
<td><strong>0.048</strong> (1.0)</td>
<td>25.0 ± 4.9</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td><strong>0.175</strong> (4.9)</td>
<td>0.188 (4.9)</td>
<td>43.0 ± 28.2</td>
</tr>
<tr>
<td>Doripenem</td>
<td><strong>0.270</strong> (1.4)</td>
<td><strong>0.270</strong> (1.8)</td>
<td>39.0 ± 28.6</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td><strong>0.072</strong> (1.2)</td>
<td>0.073 (1.4)</td>
<td>34.0 ± 12.7</td>
</tr>
<tr>
<td>Meropenem</td>
<td><strong>0.267</strong> (1.6)</td>
<td>0.276 (2.0)</td>
<td>32.0 ± 14.5</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td>0.013 (1.1)</td>
<td><strong>0.011</strong> (1.0)</td>
<td>50.0 ± 28.8</td>
</tr>
<tr>
<td>Erythromycin</td>
<td><strong>0.037</strong> (2.0)</td>
<td>0.038 (1.9)</td>
<td>39.0 ± 29.6</td>
</tr>
<tr>
<td>Tetracycline</td>
<td><strong>0.031</strong> (1.1)</td>
<td><strong>0.031</strong> (1.1)</td>
<td>55.0 ± 24.6</td>
</tr>
</tbody>
</table>
Can we expect good generalization?

We can **bound the risk** of a predictor based on its **performance on the training set**. The following term bounds the risk of every conjunction $h$ of rules in $\mathcal{R}$ with probability $\geq 1 - \delta$.

**Occam’s Razor Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - r} \left[ \ln \left( \binom{m}{r} \right) + \ln \left( 2 \cdot 4^k \right) - \ln (\zeta(r) \cdot \zeta(|h|) \cdot \delta) \right],
$$

where $r$ is the number of errors on the training set,

$|h|$ is the number of rules in the conjunction,

$\zeta$ is any function such that $\sum_{b \in \mathbb{N}} \zeta(b) \leq 1$

- The combinatorial term dominates the bound even for classifiers that make few errors
- The bound seems to indicate bad generalization performance.
Can we expect good generalization?

We can **bound the risk** of a predictor based on its **performance on the training set**. The following term bounds the risk of every conjunction $h$ of rules in $\mathcal{R}$ with probability $\geq 1 - \delta$.

**Occam’s Razor Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - r} \left[ \ln \left( \binom{m}{r} \right) + \ln \left( \frac{2 \cdot 4^k}{|h|} \right) - \ln(\zeta(r) \cdot \zeta(|h|) \cdot \delta) \right],
$$

where $r$ is the number of errors on the training set, $|h|$ is the number of rules in the conjunction, $\zeta$ is any function such that $\sum_{b \in \mathbb{N}} \zeta(b) \leq 1$.

- The **combinatorial term** dominates the bound even for classifiers that make few errors.
- The bound seems to indicate **bad generalization performance**.
Can we expect good generalization?

We can **bound the risk** of a predictor based on its **performance on the training set**. The following term bounds the risk of every conjunction $h$ of rules in $R$ with probability $\geq 1 - \delta$.

**Occam’s Razor Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - r} \left[ \ln \binom{m}{r} + \ln \left( \frac{2 \cdot 4^k}{|h|} \right) - \ln(\zeta(r) \cdot \zeta(|h|) \cdot \delta) \right],
$$

where $r$ is the number of errors on the training set,

$|h|$ is the number of rules in the conjunction,

$\zeta$ is any function such that $\sum_{b \in \mathbb{N}} \zeta(b) \leq 1$

- The **combinatorial term** dominates the bound even for classifiers that make few errors
- The bound seems to indicate **bad generalization performance**.
Can we expect good generalization?

We can bound the risk of a predictor based on its performance on the training set. The following term bounds the risk of every conjunction \( h \) of rules in \( \mathcal{R} \) with probability \( \geq 1 - \delta \).

**Occam’s Razor Bound**

\[
\epsilon \overset{\text{def}}{=} \frac{1}{m - r} \left[ \ln \binom{m}{r} + \ln \left( 2 \cdot 4^k \right) - \ln(\zeta(r) \cdot \zeta(|h|) \cdot \delta) \right],
\]

where \( r \) is the number of errors on the training set,

\( |h| \) is the number of rules in the conjunction,

\( \zeta \) is any function such that \( \sum_{b \in \mathbb{N}} \zeta(b) \leq 1 \).

- The combinatorial term dominates the bound even for classifiers that make few errors.
- The bound seems to indicate bad generalization performance.
Can we expect good generalization?

In the sample compression framework, the predictor $h$ is specified using a small set of training examples ($\mathcal{Z}_i$):

\[
\epsilon \overset{\text{def}}{=} \frac{1}{m - |h| - r} \left[ \ln \left( \frac{m}{|h|} \right) + \ln \left( \frac{m - |h|}{r} \right) \right]
\]

\[
+ \sum_{x \in \mathcal{Z}_i} \ln(2 \cdot |x|) - \ln(\zeta(|h|) \cdot \zeta(r) \cdot \delta)
\]

where $r$ is the number of errors made on $S \setminus \mathcal{Z}_i$.

- The bound does not depend on $k$ anymore.
- We can consider exponentially more complex feature spaces without any penalty on the generalization error.
Can we expect good generalization?

In the sample compression framework, the predictor $h$ is specified using a **small** set of training examples ($\mathcal{Z}_i$):

**Sample Compression Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - |h| - r} \left[ \ln \left( \frac{m}{|h|} \right) + \ln \left( \frac{m - |h|}{r} \right) \right. \\
+ \left. \sum_{x \in \mathcal{Z}_i} \ln(2 \cdot |x|) - \ln(\zeta(|h|) \cdot \zeta(r) \cdot \delta) \right],
$$

where $r$ is the number of errors made on $S \setminus \mathcal{Z}_i$.

- The bound does not depend on $k$ anymore.
- We can consider **exponentially more complex** feature spaces without any penalty on the generalization error.
Can we expect good generalization?

In the sample compression framework, the predictor $h$ is specified using a small set of training examples ($Z_i$):

**Sample Compression Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - |h| - r} \left[ \ln \left( \frac{m}{|h|} \right) + \ln \left( \frac{m - |h|}{r} \right) + \sum_{x \in Z_i} \ln(2 \cdot |x|) - \ln(\zeta(|h|) \cdot \zeta(r) \cdot \delta) \right],
$$

where $r$ is the number of errors made on $S \setminus Z_i$.

- The bound does not depend on $k$ anymore
- We can consider exponentially more complex feature spaces without any penalty on the generalization error
Can we expect good generalization?

In the sample compression framework, the predictor $h$ is specified using a small set of training examples ($\mathcal{Z}_i$):

**Sample Compression Bound**

$$
\epsilon \overset{\text{def}}{=} \frac{1}{m - |h| - r} \left[ \ln \left( \frac{m}{|h|} \right) + \ln \left( \frac{m - |h|}{r} \right) 
+ \sum_{x \in \mathcal{Z}_i} \ln(2 \cdot |x|) - \ln(\zeta(|h|) \cdot \zeta(r) \cdot \delta) \right],
$$

where $r$ is the number of errors made on $S \setminus \mathcal{Z}_i$.

- The bound does not depend on $k$ anymore
- We can consider exponentially more complex feature spaces without any penalty on the generalization error
## Replacing Cross-Validation by Bound-Selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cross-validation</th>
<th>Bound Selection</th>
<th>Bound Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. difficile</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azithromycin</td>
<td>0.030 (3.3)</td>
<td><strong>0.025</strong> (2.6)</td>
<td>0.251</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td><strong>0.073</strong> (2.6)</td>
<td>0.089 (1.4)</td>
<td>0.379</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td><strong>0.011</strong> (3.0)</td>
<td>0.026 (2.6)</td>
<td>0.256</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>0.021 (1.4)</td>
<td><strong>0.018</strong> (1.8)</td>
<td>0.181</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td><strong>0.020</strong> (1.0)</td>
<td><strong>0.020</strong> (1.0)</td>
<td>0.174</td>
</tr>
<tr>
<td><strong>M. tuberculosis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.179 (1.4)</td>
<td><strong>0.172</strong> (1.1)</td>
<td>0.564</td>
</tr>
<tr>
<td>Isoniazid</td>
<td><strong>0.021</strong> (1.0)</td>
<td><strong>0.021</strong> (1.0)</td>
<td>0.327</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td><strong>0.318</strong> (3.1)</td>
<td>0.392 (1.4)</td>
<td>0.695</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>0.031 (1.4)</td>
<td><strong>0.027</strong> (1.0)</td>
<td>0.358</td>
</tr>
<tr>
<td>Streptomycin</td>
<td><strong>0.050</strong> (1.0)</td>
<td><strong>0.050</strong> (1.0)</td>
<td>0.393</td>
</tr>
<tr>
<td><strong>P. aeruginosa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amikacin</td>
<td><strong>0.175</strong> (4.9)</td>
<td>0.177 (2.7)</td>
<td>0.493</td>
</tr>
<tr>
<td>Doripenem</td>
<td><strong>0.270</strong> (1.4)</td>
<td>0.275 (1.3)</td>
<td>0.562</td>
</tr>
<tr>
<td>Levofloxacin</td>
<td><strong>0.072</strong> (1.2)</td>
<td><strong>0.072</strong> (1.0)</td>
<td>0.324</td>
</tr>
<tr>
<td>Meropenem</td>
<td><strong>0.267</strong> (1.6)</td>
<td>0.284 (1.7)</td>
<td>0.596</td>
</tr>
<tr>
<td><strong>S. pneumoniae</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzylpenicillin</td>
<td><strong>0.013</strong> (1.1)</td>
<td>0.011 (1.0)</td>
<td>0.142</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>0.037 (2.0)</td>
<td><strong>0.035</strong> (2.0)</td>
<td>0.210</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.031 (1.1)</td>
<td><strong>0.027</strong> (1.0)</td>
<td>0.164</td>
</tr>
</tbody>
</table>
Spurious Correlations can be Overcome

a) Ethambutol Isoniazid Pyrazinamide Rifampicin Streptomycin

1.00 1.00 1.00 1.00 1.00
Round 1 (0.050) Round 2 (0.059) Round 3 (0.059) Round 4 (0.061) Round 5 (0.061) Round 6 (0.072)

katG - Catalase-Peroxidase
rpoB - RNA Polymerase β-Subunit
rpsL - 30S Ribosomal Protein S12

Alexandre Drouin (Université Laval)