Raman-Spec(k)troscopy
Sequencing with Las0rz

Kevin Lamkiewicz

13.02.2018
33rd TBI Winterseminar in Bled
Raman, Speck – what’s that all about?
Do you know Raman spectroscopy?

- Yes
- Also yes, but in orange
Raman Spectroscopy

TERS – Tip-Enhanced Raman Spectroscopy

- proposed in 1895, first experiments in 2000
- usually silver or gold tips

N. Kumar et al. (2015): Tip-enhanced Raman spectroscopy: principles and applications
TERS – Tip-enhanced Raman-spectroscopy

- proposed in 1895, first experiments in 2000
- usually silver or gold tips
- provides higher resolution
- enhances the reflected raman signal
- expands the wavelengths ranges

N. Kumar et al. (2015): Tip-enhanced Raman spectroscopy: principles and applications
TERS AND SEQUENCING

TERS AND SEQUENCING

Okay, fancy. Show me the sequences!
**RAW DATA = RAMAN SPECTRA**

<table>
<thead>
<tr>
<th>Raman Intensity</th>
<th>wave number</th>
</tr>
</thead>
<tbody>
<tr>
<td>4895</td>
<td>330.793</td>
</tr>
<tr>
<td>4963</td>
<td>334.587</td>
</tr>
<tr>
<td>4978</td>
<td>338.380</td>
</tr>
<tr>
<td>4948</td>
<td>342.171</td>
</tr>
<tr>
<td>4941</td>
<td>345.960</td>
</tr>
<tr>
<td>5002</td>
<td>349.748</td>
</tr>
<tr>
<td>4841</td>
<td>353.534</td>
</tr>
<tr>
<td>4936</td>
<td>357.318</td>
</tr>
<tr>
<td>4861</td>
<td>361.101</td>
</tr>
<tr>
<td>4947</td>
<td>364.882</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- for each tip-position one table
- wave numbers are identical for each tip-position
### Raw Data = Raman Spectra

<table>
<thead>
<tr>
<th>Raman Intensity</th>
<th>wave number</th>
</tr>
</thead>
<tbody>
<tr>
<td>4895</td>
<td>330.793</td>
</tr>
<tr>
<td>4963</td>
<td>334.587</td>
</tr>
<tr>
<td>4978</td>
<td>338.380</td>
</tr>
<tr>
<td>4948</td>
<td>342.171</td>
</tr>
<tr>
<td>4941</td>
<td>345.960</td>
</tr>
<tr>
<td>5002</td>
<td>349.748</td>
</tr>
<tr>
<td>4841</td>
<td>353.534</td>
</tr>
<tr>
<td>4936</td>
<td>357.318</td>
</tr>
<tr>
<td>4861</td>
<td>361.101</td>
</tr>
<tr>
<td>4947</td>
<td>364.882</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- for each tip-position one table
- wave numbers are identical for each tip-position

**RAW DATA = RAMAN SPECTRA**

<table>
<thead>
<tr>
<th>Raman Intensity</th>
<th>wave number</th>
</tr>
</thead>
<tbody>
<tr>
<td>4895</td>
<td>330.793</td>
</tr>
<tr>
<td>4963</td>
<td>334.587</td>
</tr>
<tr>
<td>4978</td>
<td>338.380</td>
</tr>
<tr>
<td>4948</td>
<td>342.171</td>
</tr>
<tr>
<td>4941</td>
<td>345.960</td>
</tr>
<tr>
<td>5002</td>
<td>349.748</td>
</tr>
<tr>
<td>4841</td>
<td>353.534</td>
</tr>
<tr>
<td>4936</td>
<td>357.318</td>
</tr>
<tr>
<td>4861</td>
<td>361.101</td>
</tr>
<tr>
<td>4947</td>
<td>364.882</td>
</tr>
</tbody>
</table>

▶ for each tip-position one table
▶ wave numbers are identical for each tip-position

---

And what am I doing now?
CURRENT EXPERIMENT

X. Lin, T. Deckert-Gaudig et al. (2016): Direct Base-to-Base Transitions in ssDNA Revealed by Tip-Enhanced Raman Scattering
DERIVING NUCLEOTIDES FROM SPECTRA

Adenine

Cytosine
DERIVING NUCLEOTIDES FROM SPECTRA

Adenine

Cytosine

Spectrum 1

wavenumber
intensity

wavelength
intensity
DERIVING NUCLEOTIDES FROM SPECTRA

Adenine

Cytosine

Spectrum 2
DERIVING NUCLEOTIDES FROM SPECTRA

Adenine

Cytosine

Spectrum 3
DERIVING NUCLEOTIDES FROM SPECTRA

Spectrum 4
**TEACHING MACHINES HOW TO INTERPRET SPECTRA**

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>⋯</th>
<th>Sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>read count 1</td>
<td>read count 2</td>
<td>⋯</td>
<td>read count n</td>
</tr>
<tr>
<td>Gene 2</td>
<td>read count 1</td>
<td>read count 2</td>
<td>⋯</td>
<td>read count n</td>
</tr>
<tr>
<td></td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>Gene n</td>
<td>read count 1</td>
<td>read count 2</td>
<td>⋯</td>
<td>read count n</td>
</tr>
</tbody>
</table>
**Teaching Machines How to Interpret Spectra**

<table>
<thead>
<tr>
<th>Position 1</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>...</th>
<th>Sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
<tr>
<td>Position 2</td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Position n</td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
</tbody>
</table>
### Teaching machines how to interpret spectra

<table>
<thead>
<tr>
<th></th>
<th>wave number 1</th>
<th>wave number 2</th>
<th>...</th>
<th>wave number n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
<tr>
<td>Position 2</td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Position n</td>
<td>read count 1</td>
<td>read count 2</td>
<td>...</td>
<td>read count n</td>
</tr>
</tbody>
</table>
**Teaching Machines How to Interpret Spectra**

<table>
<thead>
<tr>
<th>Position</th>
<th>wave number 1</th>
<th>wave number 2</th>
<th>...</th>
<th>wave number n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>spectrum 1</td>
<td>spectrum 2</td>
<td>...</td>
<td>spectrum n</td>
</tr>
<tr>
<td>2</td>
<td>spectrum 1</td>
<td>spectrum 2</td>
<td>...</td>
<td>spectrum n</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>spectrum 1</td>
<td>spectrum 2</td>
<td>...</td>
<td>spectrum n</td>
</tr>
</tbody>
</table>
CLUSTERING THE DATA

PCA of first TERS dataset: grid7
MACHINE LEARNING?

PCA of first TERS dataset: grid7

PC1: 40% variance
PC2: 17% variance

Nucleotide
-  
A  
C  
M  

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
But... we have sequencing methods! Why should anyone use TERS?
RAMAN SPECTROSCOPY CHANCES

- sequencing of DNA/RNA modifications
- direct aminoacid sequencing
RAMAN SPECTROSCOPY CHANCES

- sequencing of DNA/RNA modifications
- direct aminoacid sequencing
- viral fingerprint
CHALLENGES & DRAWBACKS

- no high-throughput method
- oxidation of silver tip
- thermal drift
- no protocol for RNA/aminoacid sequences
CHALLENGES & DRAWBACKS

- no high-throughput method
- oxidation of silver tip
- thermal drift
- no protocol for RNA/aminoacid sequences
Acknowledgements:
RNA Bioinformatics Group Jena
Deckert Gruppe

Funding:
BMBF InfectControl 2020
Project 03ZZ0820A

Thank you for your attention!
A short update of my viral research
ViMiFi – Viral microRNA Finder

For each sequence:

- RNAfold

Single sequence mode

- RNAalifold

Alignment mode

Input genomes

MAFFT

MSA

Found candidates in Tombusviruses using different training sets

A. Th. precursor

O. Sa. precursor

viral precursor

Viral Update
ViMiFi – VIRAL MICRORNA FINDER

ViMiFi

Input genomes

MAFFT

MSA

For each sequence:

RNAfold

Single sequence mode

Alignment mode

RNAalifold

Feature extraction

Triplets

nAn nGn nAn nUn ...

Structure

MFE, loop length, number of bulges, base-pair types,
...

Sequence

Length, %GC-content, GpC-islands

Training

k-fold cv

supervised learning

model comparison

Classification

potential viral pre-miRNAs

Found candidates in Tombusviruses using different training sets

A. Th. precursor
O. Sa. precursor
viral precursor

Viral Update
**ViMiFi – Viral MicroRNA Finder**

- **Input genomes**
- **MAFFT**
- **MSA**
- **Triplets**
- **nAn nGn nAn nUn** ...
- **Structure**
- **MFE, loop length, number of bulges, base-pair types** ...
- **Sequence**
- **Length, %GC-content, GpC-islands**
- For each sequence:
- **RNAfold**
- **Single sequence mode**
- **RNAalifold**
- **Alignment mode**
- **Feature extraction**
- **Training**
- **Classification**
- **k-fold cv**
- **Supervised learning**
- **Model comparison**
- **Potential viral pre-miRNAs**
- **Viral precursor**
- **Viral CDS**

- Found candidates in Tombusviruses using different training sets
**ViMiFi – Viral MicroRNA Finder**

Found candidates in Tombusviruses using different training sets

- A. Th. precursor
- O. Sa. precursor

Venn diagram showing the overlap of found candidates among different training sets.
ViMiFi – Viral MicroRNA Finder

Found candidates in EBV using different training sets

H. Sa. precursor 21
viral precursor 2 96

Found candidates in Tombusviruses using different training sets

A. Th. precursor 12
O. Sa. precursor 43
viral precursor 3 3 15

LOCATION – LONG COMPENSATORY MUTATIONs
LOCATION – LONG COMPENSATORY MUTATIONS
LOCATION – LONG COMPENSATORY MUTATIONS

short read data

long read data

?
LOCATION – LONG COMPENSATORY MUTATIONs
LOCATION – LONG COMPENSATORY MUTATIONs
LOCATION – LONG COMPENSATORY MUTATIONS
LOCATION – LONG COMPENSATORY MUTATIONS