WHY DO WE NEED AN AUTOMATED NMR ANNOTATION TOOL
Metabolomic workflow

Control mice → Urine, plasma, tissues...
Exposed mice → Analytical analysis (NMR, MS...)

Metabolites Identification
- Glucose
- Taurine
- Glycine
- Glutamate
- Lactate
- Lysine
- Leucine

NMR and MS spectra

Multivariate Statistical Analysis

Data Reduction

Urine, plasma, tissues...
$^1$H NMR spectrum of reference compound

Glycine

- Each group of protons gives one signal
- A singlet for Glycine
$^1$H NMR spectrum of reference compound

- 1 signal is observed for each type of proton
- 18 different signals for this compound
- Complex spectrum
The two forms $\alpha$ et $\beta$ give different signals
- 14 different signals for this compound
No separation of compounds before NMR analysis
A metabolite can have many signals at different chemical shifts
Many signals overlap

Identification of metabolites in mixtures is really complex
Metabolite identification (manually)

- Based on 1H chemical shift, coupling pattern and coupling constants
- Comparison with annotated spectra of similar matrices in literature
- Spiking: a reference compound is added in the sample
- Comparison with NMR spectra of reference compounds
  - database in-house
  - Chenomx (commercial)
  - HMDB (freeware)
- Multi-dimensional NMR
2D NMR experiments mostly used

Information is split in two dimensions

- $^1\text{H}-^1\text{H}$ COSY (Correlation Spectroscopy) experiment: correlations $^1\text{H}-^1\text{H}$ via 3 bonds

- $^1\text{H}-^1\text{H}$ TOCSY (TOtal Correlation Spectroscopy) experiment: correlations between all protons within a given spin system

- $^1\text{H}-^{13}\text{C}$ HSQC (Heteronuclear Single Quantum Coherence) experiment: correlations $^1\text{H}-^{13}\text{C}$ via 1 bond

- $^1\text{H}-^{13}\text{C}$ HMBC (Heteronuclear Multiple Bond Coherence) experiment: correlations $^1\text{H}-^{13}\text{C}$ via 2, 3 or 4 bonds

Information on hydrocarbon skeletal
Example: $^1$H NMR spectrum of aqueous pup extract (PND2)
2D $^1$H-$^1$H COSY NMR spectrum of aqueous pup extract (PND2)

CH$_3$-CH(OH)COOH

- 4.11 ppm threonine
- 1.33 ppm lactate
taurine
- 4.11 ppm alanine
- 3.78 ppm glutamine
- 1.33 ppm glutathion
- 1.33 ppm glutamate
- 1.33 ppm leucine
- 1.33 ppm lysine
- 1.33 ppm valine
- 1.33 ppm glucose
- 1.33 ppm choline
- 1.33 ppm glucose
2D $^1$H-$^{13}$C HSQC NMR spectrum of aqueous pup extract (PND2)
Comparison with NMR spectra of reference compounds:
Human Metabolome database

www.hmdb.ca/
Comparison with NMR spectra of reference compounds: Human Metabolome database

Many responses (299) : the expert must check all the spectra to find the right metabolite.
Comparison with NMR spectra of reference compounds: Human Metabolome database
Comparison with NMR spectra of reference compounds: Human Metabolome database

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Structure</th>
<th>Library Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Lactic acid (HMDB01311)</td>
<td>H2C4O3</td>
<td><img src="image1" alt="Structure" /></td>
<td>1/2</td>
</tr>
<tr>
<td>10226-41-7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Octanol (HMDB00092)</td>
<td>C8H16O3</td>
<td><img src="image2" alt="Structure" /></td>
<td>1/7</td>
</tr>
<tr>
<td>1871-07-6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a-Androstan-3b,17b-diol (HMDB000403)</td>
<td>C27H44O2</td>
<td><img src="image3" alt="Structure" /></td>
<td>1/25</td>
</tr>
<tr>
<td>571-20-6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a-Hole (iso-1,9-decalactone) (HMDB000871)</td>
<td>C23H38O6</td>
<td><img src="image4" alt="Structure" /></td>
<td>1/29</td>
</tr>
<tr>
<td>566-89-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

13 results
Comparison with NMR spectra of reference compounds:
- Database in-house

- Glutamate
- Valine
- Alanine
- Taurine
- Creatine
- Lactate
- Aqueous pup extract
Metabolite identification (manually)

- Time-consuming
- Complicated
- Some resonances are not identified
- An automated NMR annotation tool will be helpful:
  - Package R Batman: very complex and not user-friendly
  - BAYESIL (http://bayesil.ca/): web interface (limited to serum, plasma, and CSF biofluids)


We developed a new tool for the annotation of NMR spectra
Objectives

- Automated identification of compounds in NMR spectra of complex mixtures
- Estimation of the proportion of the metabolites in the mixture
Difficulties

- No separation of compounds before NMR analysis
- A metabolite can have many signals at different chemical shifts
- Many signals overlap

Identification of metabolites in mixtures is really complex
Modeling

Mixture Y spectrum

Library Z1, ..., Zn

Problems: noise and peaks warping
Noise Modeling

Several NMR spectra of glucose allow to model the noise
Warping problems

- Metabolite concentration is different in pure compound spectrum and in mixture spectrum: peaks do not have necessary the same shape.

- Experimental conditions are different between pure compound and mixture (pH, ionic strength, analytical variability…): peaks can be shifted.
Warping modeling

Two NMR spectra of a same mixture obtained in different conditions
Warping modeling

Before modeling

After modeling

NMR spectra of choline pure or in the mixture

We used a warping function: $\Phi$
Warping

- Shift depend on pH: shift is not the same for all compounds
- Deformations are localized
- Peaks can be shifted, but also expanded or contracted
- Maximum deformation given by the expert
- NMR spectra are recorded at pH 7 to limit the shifts
Metabolite spectrum modeling

- $f_1(t)$: NMR spectrum of pure compound non observed
- NMR spectrum observed with noise

\[ Z_1(t) = f_1(\phi(t)) + \varepsilon_1(t)\sqrt{f_1(\phi(t))} + \varepsilon_2(t) \]

- Library $\{Z_1, \ldots, Z_{200}\}$
Mixture spectrum modeling

- A mixture $g$ is modeled as a positive function $g(t)$:
  \[ g(t) = \alpha_1 f_1(t) + \cdots + \alpha_{200} f_{200}(t) \]

- Signal is noised and warped:
  \[ Y(t) = g(\phi(t)) + \varepsilon_1(t) \sqrt{g(\phi(t))} + \varepsilon_2(t) \]
  
  - Déformation
  - Bruit multiplicatif
  - Bruit additif

- $\alpha$ is the relative proportion of metabolite in the mixture spectrum: area under the curve of the metabolite divided by the total area of the spectrum

- Determine non-zero proportions
- Estimate the proportions: quantification
Method

- A false detection is a metabolite identified but this metabolite is not present in the mixture.
- The detection threshold is the proportion from which we are sure that the metabolite is identified.
- Compromise between low probability of false detection and low threshold.
- In this method:
  - Control the probability to obtain one or several false detections.
  - Give the detection threshold of each metabolite.
  - Minimize the detection threshold.
Real data set

A mixture of 6 metabolites:

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Glucose</td>
<td>5.8 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>21.0 %</td>
</tr>
<tr>
<td>L-Phenylalanine</td>
<td>3.3 %</td>
</tr>
<tr>
<td>L-Proline</td>
<td>6.7 %</td>
</tr>
<tr>
<td>Benzoic acid</td>
<td>8.4 %</td>
</tr>
<tr>
<td>Choline chloride</td>
<td>54.6 %</td>
</tr>
</tbody>
</table>

Library of 36 metabolites
Automatic cleaning

Black: mixture spectrum; Red: hydroxybenzoic spectrum

Elimination of 18 metabolites
<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Identification</th>
<th>Detection Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzoic acid</td>
<td>Yes</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Choline</td>
<td>Yes</td>
<td>2.3 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>Yes</td>
<td>1.1 %</td>
</tr>
<tr>
<td>Glucose</td>
<td>Yes</td>
<td>3.1 %</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Yes</td>
<td>2.1 %</td>
</tr>
<tr>
<td>Proline</td>
<td>Yes</td>
<td>2.6 %</td>
</tr>
<tr>
<td>Acetic acid</td>
<td>No</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Galactose</td>
<td>No</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Mannitol</td>
<td>No</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Mannose</td>
<td>No</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Autres</td>
<td>No</td>
<td>&lt; 2.5 %</td>
</tr>
</tbody>
</table>
## Quantification

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Estimated proportions</th>
<th>Real proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzoic acid</td>
<td>5.7 %</td>
<td>8.4 %</td>
</tr>
<tr>
<td>Choline</td>
<td>55.8 %</td>
<td>54.6 %</td>
</tr>
<tr>
<td>Creatinine</td>
<td>13.3 %</td>
<td>21 %</td>
</tr>
<tr>
<td>Glucose</td>
<td>5.5 %</td>
<td>5.8 %</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>3.9 %</td>
<td>3.3 %</td>
</tr>
<tr>
<td>Proline</td>
<td>5.7 %</td>
<td>5.9 %</td>
</tr>
</tbody>
</table>

Detection thresholds are too high
# Improvement of detection thresholds

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Identification</th>
<th>New thresholds</th>
<th>Previous thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetic acid</td>
<td>No</td>
<td>0.16 %</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Galactose</td>
<td>No</td>
<td>0.19 %</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Mannitol</td>
<td>No</td>
<td>0.24 %</td>
<td>1.6 %</td>
</tr>
<tr>
<td>Mannose</td>
<td>No</td>
<td>0.17 %</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Others</td>
<td>No</td>
<td>&lt; 0.68%</td>
<td>&lt; 2.5 %</td>
</tr>
</tbody>
</table>
Synthesis Urine

- Mix in 500 ml of water:
  - 3.8 g of potassium chloride
  - 8.5 g of sodium chloride
  - 24.5 g of urea
  - 1.03 g of citric acid
  - 0.34 g of ascorbic acid
  - 1.18 g of potassium dihydrogenophosphate
  - 1.4 g of creatinine
  - 0.64 g of sodium hydroxide
  - 0.47 g of sodium bicarbonate
  - 0.28 ml of phosphoric acid
Synthesis Urine

- Addition of metabolites:
  - Hippuric acid: 8.2 mM
  - TMAO: 7 mM
  - Acetic acid: 0.8 mM
  - Alanine: 1 mM
  - Betaine: 1 mM
  - Carnitine: 0.7 mM
  - Dimethylamine: 1 mM
  - Ethanolamine: 1.5 mM
  - Formate: 0.4 mM
  - Glucose: 1 mM
  - Glutamine: 2 mM
  - Glycine: 0.6 mM
  - Guanidinoacetate: 0.8 mM
  - Lactate: 0.6 mM
  - Lysine: 1 mM
  - Malonate: 1.7 mM
  - Trigonelline: 0.6 mM
  - Tyrosine: 0.3 mM

21 metabolites detectables in NMR
$^1$H NMR spectrum
Conclusion

• The program is powerful for detecting the presence of compounds in mixture
• The program is not very powerful for estimating the true proportion, but we are working on this problem
• At the moment, pH of compounds in the library is 7, but we can import spectra at different pH
• At the moment, we have 175 compounds in the library but we will record more compounds (MetaboHub)
• More tests are needed (other mixtures)
HOW TO DO WITH GALAXY?
Name of the « zip » file containing Bruker directories
Name of the « zip » file containing Bruker directories

Region(s) to exclude

Maximum chemical shift
MATRIX RESULT: IDENTIFIED METABOLITES

 Identified metabolites

 Estimated proportion
MATRIX RESULT: NON IDENTIFIED METABOLITES

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Estimated proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cacedaric acid</td>
<td>0.0000999402005144</td>
</tr>
<tr>
<td>L-Carnitine</td>
<td>0.0051228298370066</td>
</tr>
<tr>
<td>THMA</td>
<td>0.0032086354454842</td>
</tr>
<tr>
<td>D-tartaric acid</td>
<td>0.0392453312006026</td>
</tr>
<tr>
<td>D-Maltose</td>
<td>0.0002145436107036</td>
</tr>
<tr>
<td>L-Arginine</td>
<td>0.0600108787440417</td>
</tr>
<tr>
<td>7-Methylanthranol</td>
<td>0.4078312267202097</td>
</tr>
<tr>
<td>N-(2-Furaldehyde)</td>
<td>0.0246865275199445</td>
</tr>
<tr>
<td>L-Tyrosine</td>
<td>0.065825169171122</td>
</tr>
<tr>
<td>D-Glucosamine</td>
<td>0.03757221328725</td>
</tr>
<tr>
<td>2-desoxyglycine</td>
<td>0.0113994519024879</td>
</tr>
<tr>
<td>Argininosuccinic acid</td>
<td>0.0144450427481172</td>
</tr>
<tr>
<td>L-Tryptophane</td>
<td>0.0018138351508713</td>
</tr>
<tr>
<td>Guanidinoacetate acid</td>
<td>0.459015840173987</td>
</tr>
<tr>
<td>Glycerol</td>
<td>0.0000229924092498</td>
</tr>
<tr>
<td>D-xylose</td>
<td>0.0159705090192758</td>
</tr>
<tr>
<td>Myo-Inositol</td>
<td>0.0182050102807666</td>
</tr>
<tr>
<td>Xylitol</td>
<td>0.0073871537408381</td>
</tr>
<tr>
<td>D-Fructose</td>
<td>0.0095408297025226</td>
</tr>
<tr>
<td>Threonine</td>
<td>0.018206758006007</td>
</tr>
<tr>
<td>L-Arginine</td>
<td>0.010848576859088</td>
</tr>
<tr>
<td>D-Glucose</td>
<td>0.0066434420659341</td>
</tr>
<tr>
<td>Glycophosphocholine</td>
<td>0.007324871817315</td>
</tr>
<tr>
<td>L-Glutamine-reduced</td>
<td>0.0008909890982714</td>
</tr>
<tr>
<td>Acetate acid</td>
<td>0.0626090830359325</td>
</tr>
<tr>
<td>L-Carnitine</td>
<td>0.006121239553821</td>
</tr>
<tr>
<td>UDPG</td>
<td>0.00288950642816789</td>
</tr>
<tr>
<td>2-desoxyguanosine</td>
<td>0.01445005213688</td>
</tr>
<tr>
<td>CDP</td>
<td>0.0092020603240214</td>
</tr>
<tr>
<td>L-Glutamylcarnitine</td>
<td>0.0154701404035155</td>
</tr>
<tr>
<td>Glyceric acid</td>
<td>0.0249241644325159</td>
</tr>
<tr>
<td>2-desoxyadenosine</td>
<td>0.00034440047939087</td>
</tr>
<tr>
<td>2-desoxyadenosine-nucleoside</td>
<td>0.0189791960209989</td>
</tr>
<tr>
<td>L-Phenylalanine</td>
<td>0.0071652170782873</td>
</tr>
<tr>
<td>L-Citrulline</td>
<td>0.0118231668073267</td>
</tr>
<tr>
<td>D-Fucose</td>
<td>0.0047880584433412</td>
</tr>
<tr>
<td>3-Methyl-L-Histidine</td>
<td>0.0005201567508571</td>
</tr>
<tr>
<td>Dimethylglycine</td>
<td>0.0061577229818722</td>
</tr>
<tr>
<td>L-Histidine</td>
<td>0.0152146835782277</td>
</tr>
</tbody>
</table>
GRAPH RESULT
Example: Synthesis Urine

- Upload the $^1$H NMR spectrum of synthesis urine
- Region to exclude: 6.5-4.5 ppm for urine
- Determine the metabolites present in this sample
- Display the spectra
- Compare the results with the list of metabolites present in the mixture
Example : Synthesis Urine

- List of metabolites present in the sample:
  - Creatinine (24.8 mM; 0.378)
  - Citric acid (10.7 mM; 0.163)
  - Hippuric acid (8.2 mM; 0.125)
  - TMAO (7 mM; 0.107)
  - Glutamine (2 mM; 0.030)
  - Malonate (1.7 mM; 0.026)
  - Ethanolamine (1.5 mM; 0.023)
  - Alanine (1 mM; 0.015)
  - Betaine (1 mM; 0.015)
  - Dimethylamine (1 mM; 0.015)
  - D-Glucose (1 mM; 0.015)
  - Lysine (1 mM; 0.015)
  - Acetic acid (0.8 mM; 0.012)
  - Guanidinoacetate (0.8 mM; 0.012)
  - Carnitine (0.7 mM; 0.011)
  - Glycine (0.6 mM; 0.009)
  - Lactate (0.6 mM; 0.009)
  - Trigonelline (0.6 mM; 0.009)
  - Formate (0.4 mM; 0.006)
  - Tyrosine (0.3 mM; 0.004)