Welcome!

Mass Spectrometry meets Cheminformatics

Tobias Kind and Julie Leary
UC Davis

Course 9: Prediction and simulation of mass spectra

Class website: CHE 241 - Spring 2008 - CRN 16583
Slides: http://fiehnlab.ucdavis.edu/staff/kind/Teaching/

PPT is hyperlinked – please change to Slide Show Mode

History of artificial intelligence and mass spectrometry

Dendral project at Stanford University (USA)
Started in 1960s
Pioneered approaches in artificial intelligence (AI)

Aim:
Prediction of isomer structures from mass spectra
Idea: Self-learning or intelligent algorithm

Participants:
Lederberg, Sutherland, Buchanan, Feigenbaum, Duffield, Djerassi, Smith, Rindfleisch, many others…

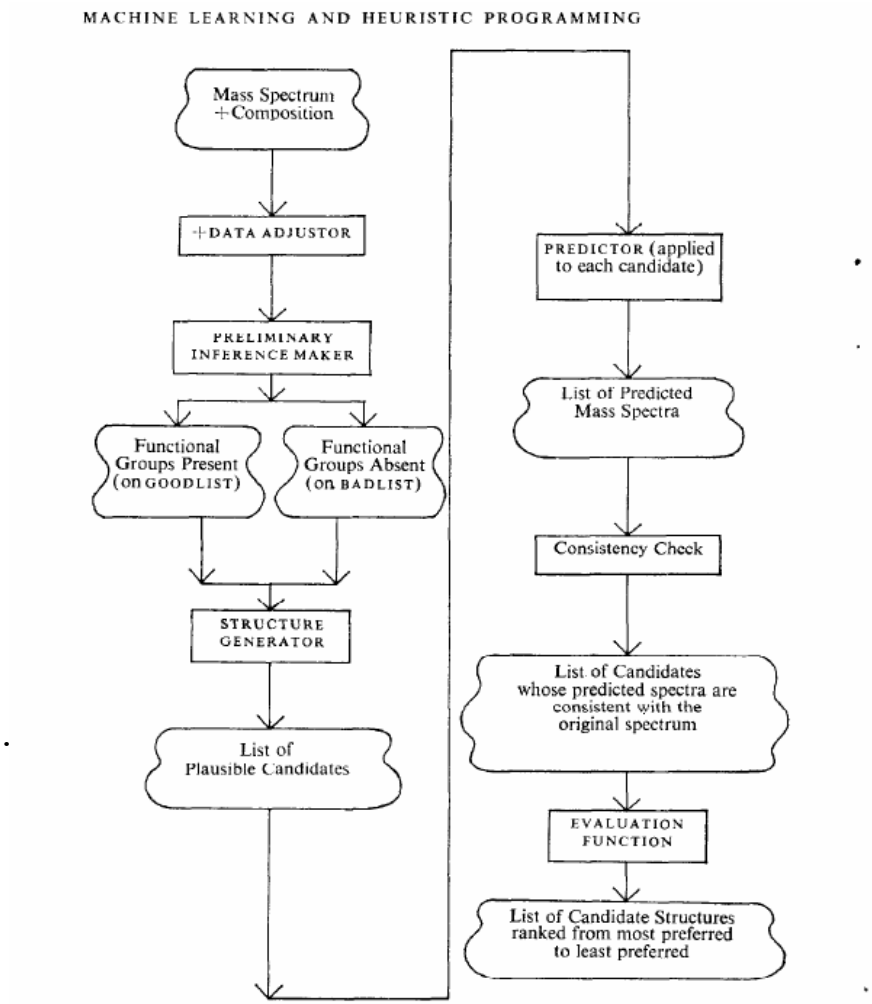
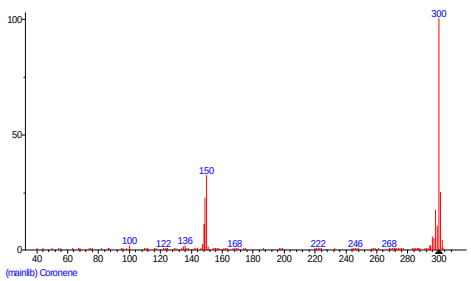


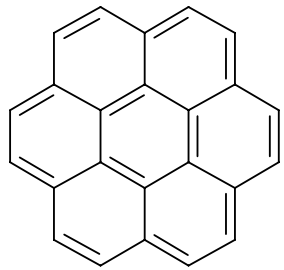
Figure: Heuristic DENDRAL:
A Program for Generating Explanatory Hypotheses in Organic Chemistry

[Dendral PDF]

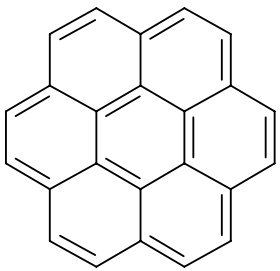
Prediction and simulation of mass spectra

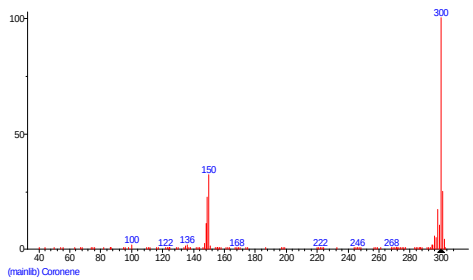
A) Prediction of the isomer structure or substructures from a given mass spectrum
The structure is directly deduced from the mass spectrum or generated by
a molecular isomer generator or existing structures can be found in a structure database



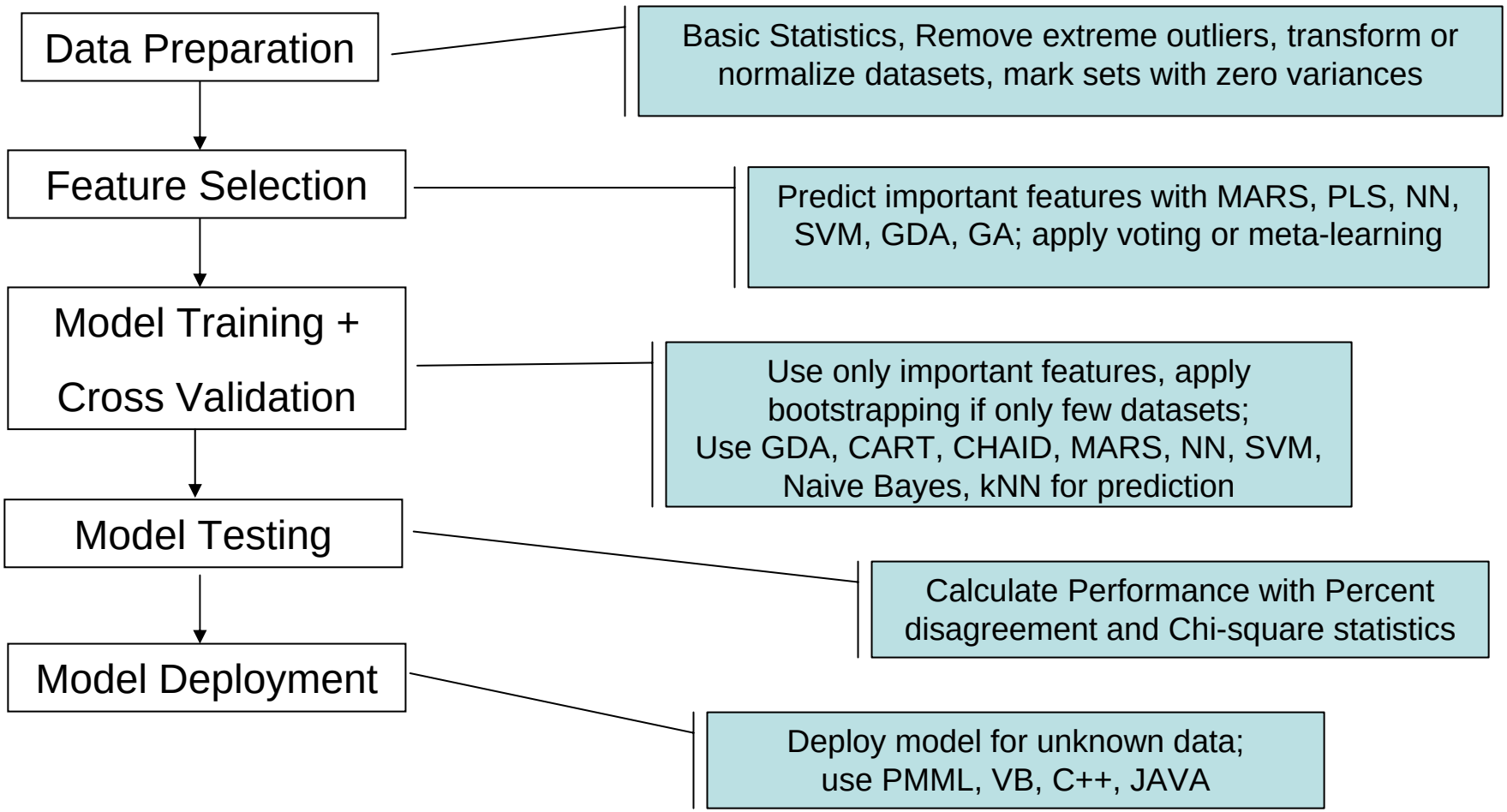


B) Simulation of a mass spectrum from a given isomer structure
The mass spectral peaks and abundances are generated by a machine learning algorithm
The structures can be obtained from a isomer database (PubChem, LipidMaps)
or a sequence database (Swiss-Prot, NCBI) in case of proteins



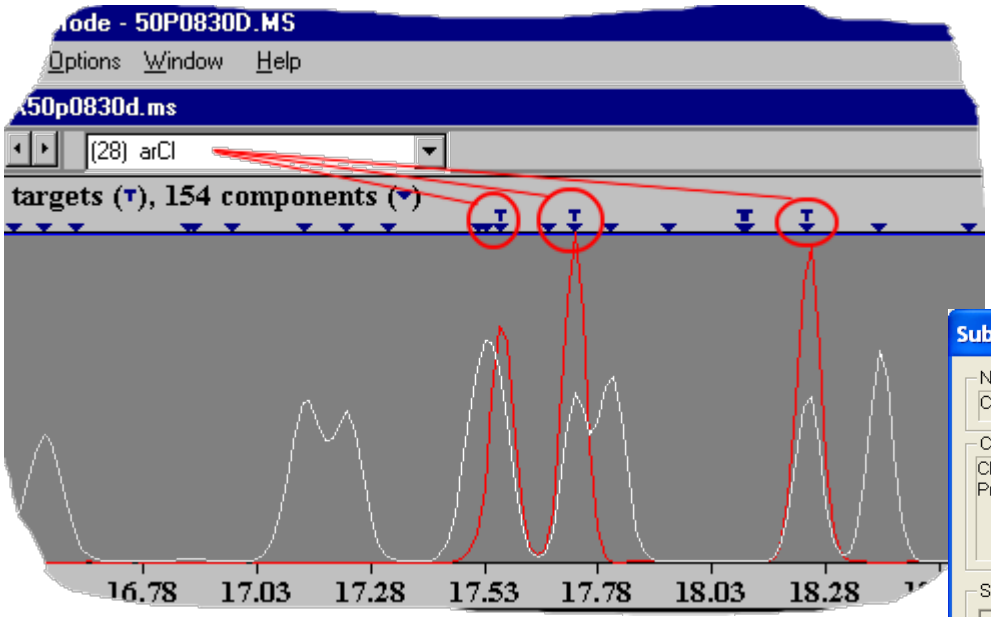


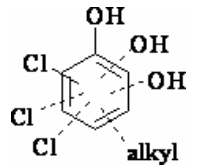
Application of machine learning for detection of substructures from mass spectra

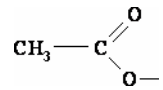


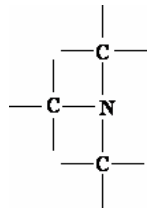
What is machine learning?

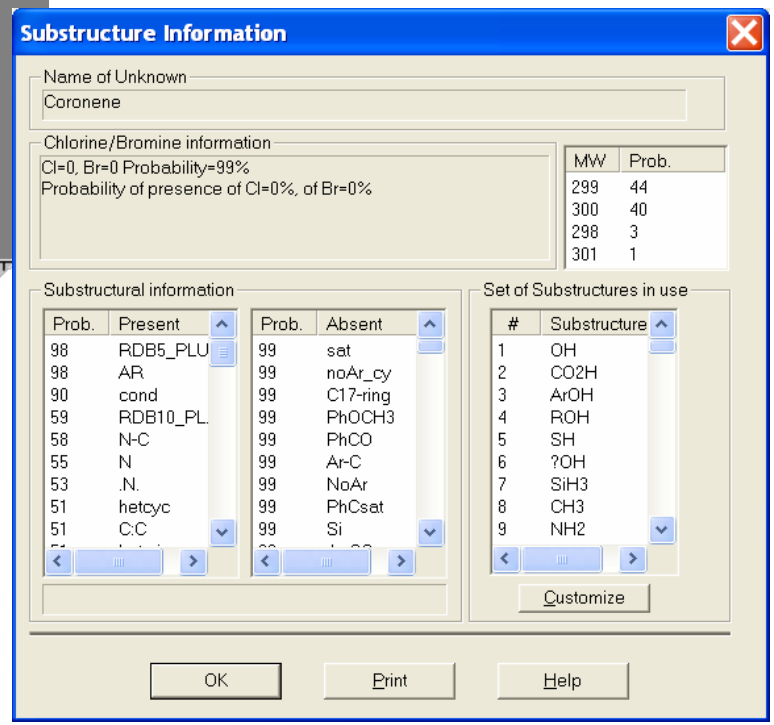
Prediction of substructures from mass spectra











Working examples for EI mass spectra:
Varmuza classifiers in **AMDIS** and **MOLGEN-MS**

Substructure algorithm (Stein S.E.)
Implemented in NIST-MS search program

Mass spectral classifiers for supporting systematic structure elucidation
Varmuza K., Werther W., J. Chem. Inf. Comput. Sci., 36, 323-333 (1996).
Chemical Substructure Identification by Mass Spectral Library Searching
S.E. Stein, J. Am. Soc. Mass Spectrom., 1995, 6, (644-655)

Picture source: amdis.net

Substructures deduced from mass spectra for generation of isomer structures

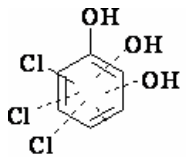
1) **Molecular formula** must be known - can be detected from molecular ion and isotopic pattern
2) **Good-list** (substructure exists) and **bad-list** (substructure not existent) approach
3) Sub-structures are combined in **deterministic** or **stochastic** (random) manner
4) **Database** or **molecular isomer generator** (combinatorial, graph theory) approach for generating or finding possible structure candidates

Example:
Molecular formula C_6ClH_5O;
calculated from molecular ion

Database (Chemspider): 25 hits
(including all possible existing structures)

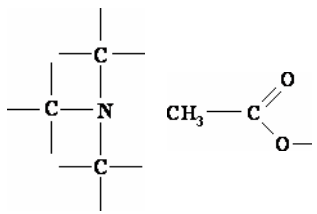
MOLGEN Demo:
All constructed isomers: 8372

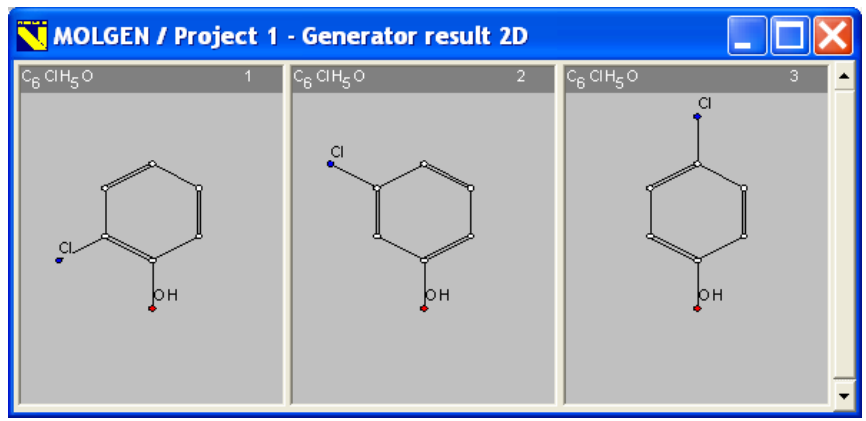
Goodlist:



-benzene
-hydroxy
-chlorine

Badlist:





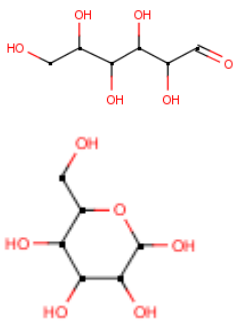
Total: 3 possible results

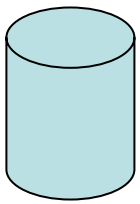
Picture source: amdis.net

Simulation of mass spectra

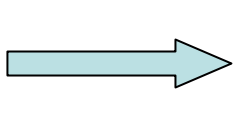
Why is simulation of mass spectral fragmentation important?

Imagine – you have a **structure database** of all molecules
Imagine – you can **simulate mass spectra** for all these molecules
Imagine – you can **match** your **experimental spectra** against a database of **calculated spectra**

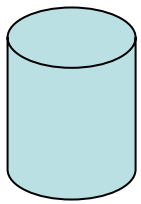


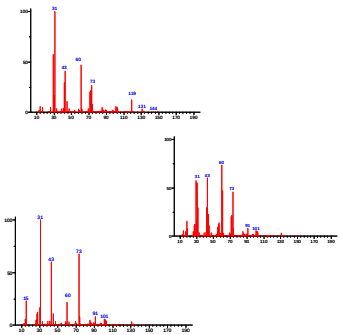


Isomer DB



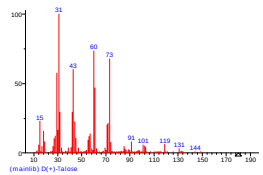
Generate mass spectra
with
Machine Learning Algorithm





MS DB
of theoretical spectra

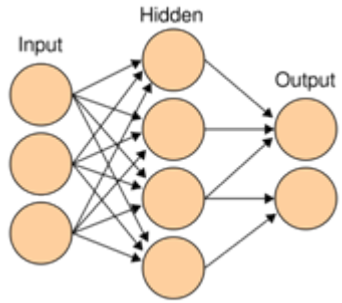
Compare
MS(calc) vs. MS(exp)



Experimental mass spectrum

If the calculation is simple the database is not needed;
In-silico MS fragments can be calculated on-the-fly

Simulation of alkane mass spectra (I)



Source: WIKI

Approach

Use of artificial neural networks (ANN) (machine learning)
Electron impact spectra 70 eV
Substructure descriptors were used for calculation
Selection of 44 m/z positions – training was performed for correct intensity

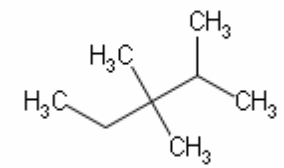
117 noncyclic alkanes and 145 noncyclic alkenes
training set: 236 molecules
prediction set: 26 compounds

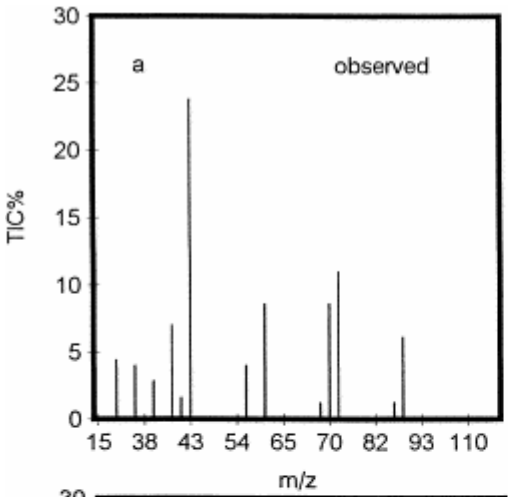
Problems

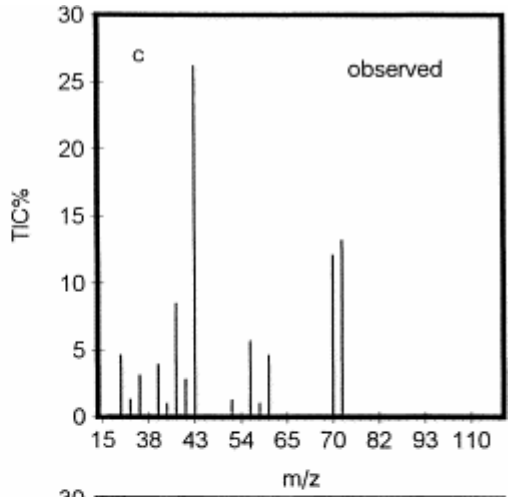
Prediction or validation set very small (should be 30%)
Prediction of molecular ion (usually very low abundant)
Overfitting possible, works only for selected substance classes

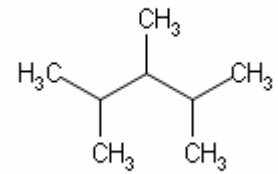
Source: Jalali-Heravi M. and Fatemi M. H.; Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network

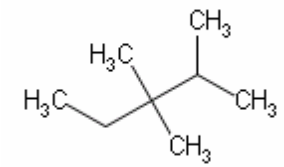
Simulation of alkane mass spectra (II)

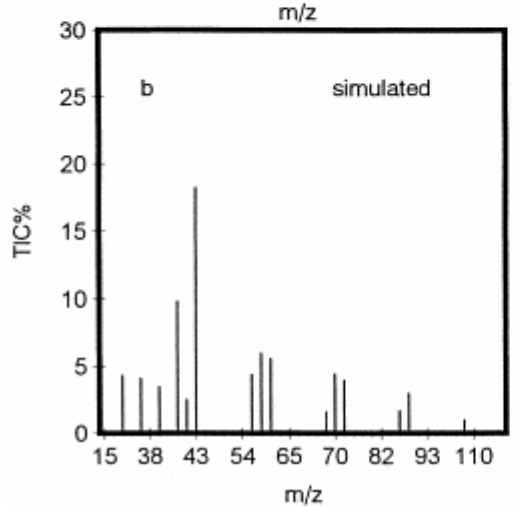


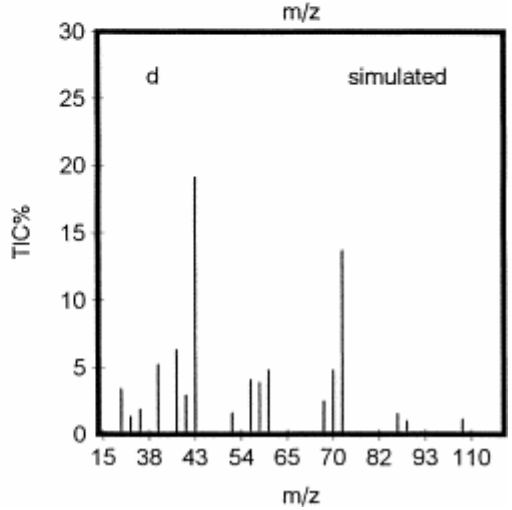


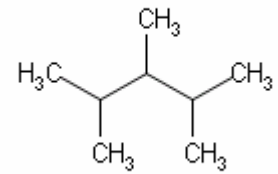












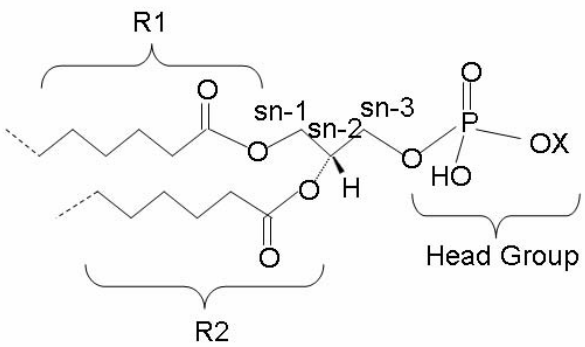
2,3,3-trimethylpentane (a and b) and 2,3,4-trimethylpentane (c and d).

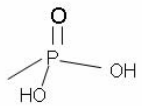
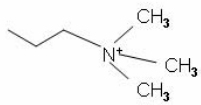
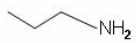
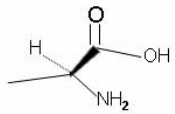
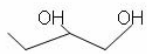
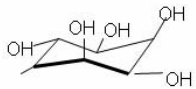
OKVWYBALHQFVFP-UHFFFAOYAT RLPGDEORIPLBNF-UHFFFAOYAR

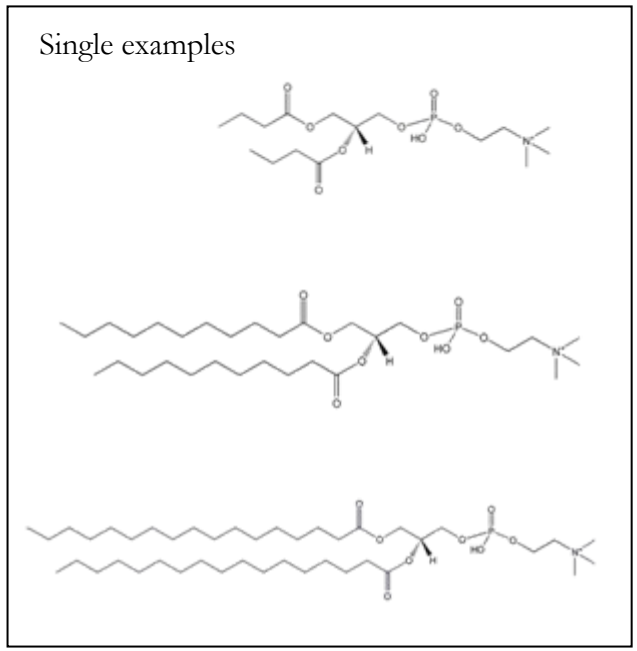
Source: Jalali-Heravi M. and Fatemi M. H.; Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network
Analytica Chimica Acta; Elsevier permission use for coursepack/classroom material

Structures: Chemspider

Simulation of lipid tandem mass spectra (I)



X	Head Group
-H	PA
	PPA
	PC
	PE
	PS
	PG
	PI



Similar structures; plus CH2 in side chains sn1 and sn2; double bonds possible
Similar and almost constant fragmentation rules
Loss of head group (diagnostic ion in MS and MS/MS spectrum)
Loss of rest one (R1) and rest two (R2) can be observed in MS/MS spectrum

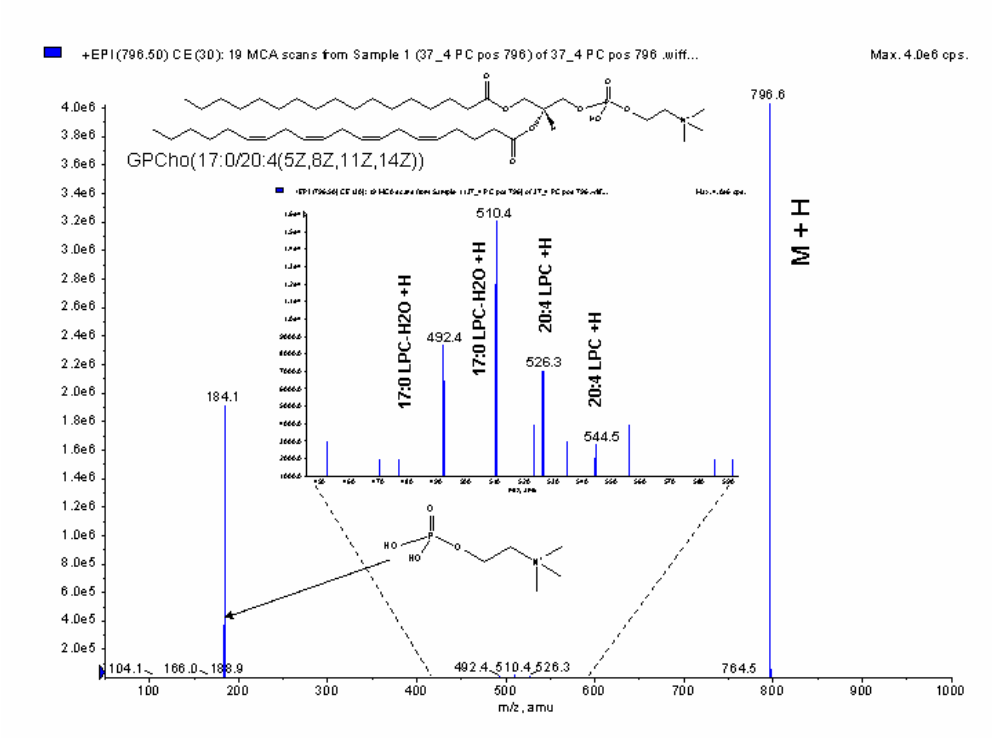
Picture: Thanks to Yetukuri *et al. BMC Systems Biology* 2007 **1**:12 doi:10.1186/1752-0509-1-12

Simulation of lipid tandem mass spectra (II)

Simulation of tandem mass spectra or MS/MS fragment data from LipidMaps

Experimental Mass spectrum

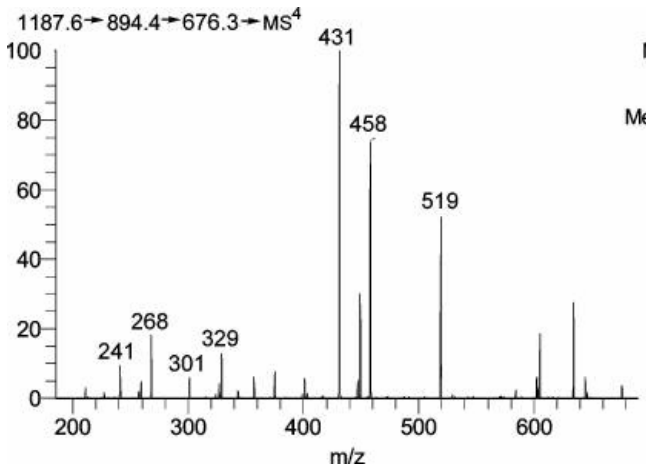
In-silico prediction of MS/MS mass spectral fragments

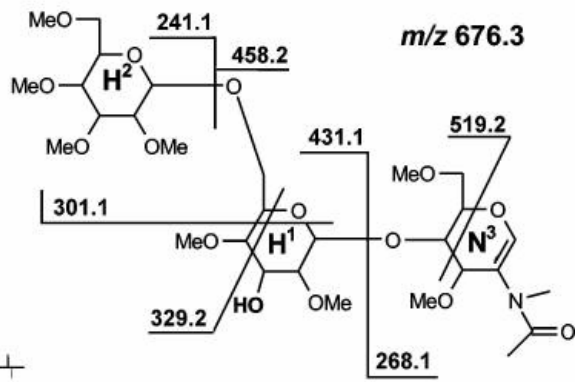


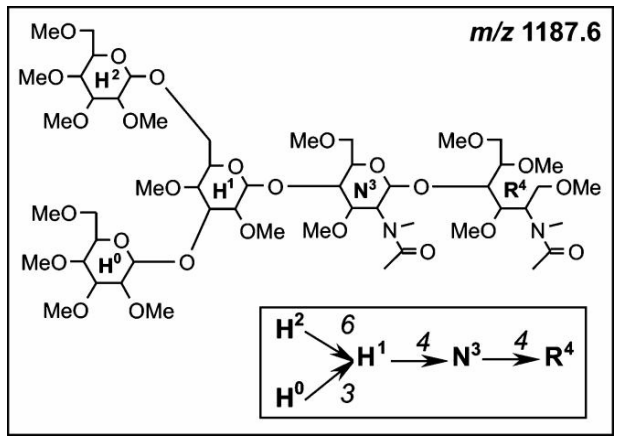
Mass	C	DB	Abbrev.	M-sn1+H	M-sn1-H2O+H	M-sn2+H	M-sn2-H2O+H	sn1 acid(-)	sn2 acid(-)	HG	Formula
797.5180	31	0	14:0/17:0	587.3196	569.309	545.2727	527.2621	227.2011	269.2481	GPIns	$C_{40}H_{77}O_{13}P$
797.5180	31	0	17:0/14:0	545.2727	527.2621	587.3196	569.309	269.2481	227.2011	GPIns	$C_{40}H_{77}O_{13}P$
796.5128	37	5	17:0/20:5(5Z,8Z,11Z,14Z,17Z)	544.2675	526.2569	512.2988	494.2882	269.2481	301.2168	GPSer	$C_{43}H_{74}NO_{10}P$
796.5128	37	5	20:5(5Z,8Z,11Z,14Z,17Z)/17:0	512.2988	494.2882	544.2675	526.2569	301.2168	269.2481	GPSer	$C_{43}H_{74}NO_{10}P$
796.5856	37	4	17:0/20:4(5Z,8Z,11Z,14Z)	544.3403	526.3297	510.3559	492.3453	269.2481	303.2324	GPCho	$C_{45}H_{82}NO_8P$
796.5856	37	4	20:4(5Z,8Z,11Z,14Z)/17:0	510.3559	492.3453	544.3403	526.3297	303.2324	269.2481	GPCho	$C_{45}H_{82}NO_8P$

Spectrum Source:Lipidmaps.org

Simulation or prediction of oligosaccharide spectra (carbohydrate sequencing)





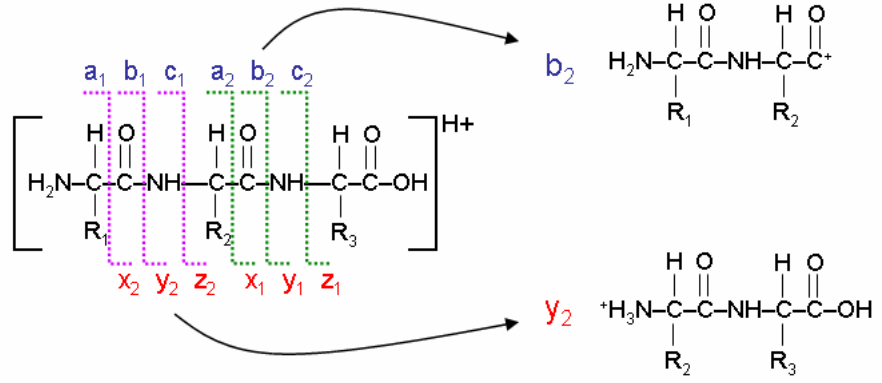


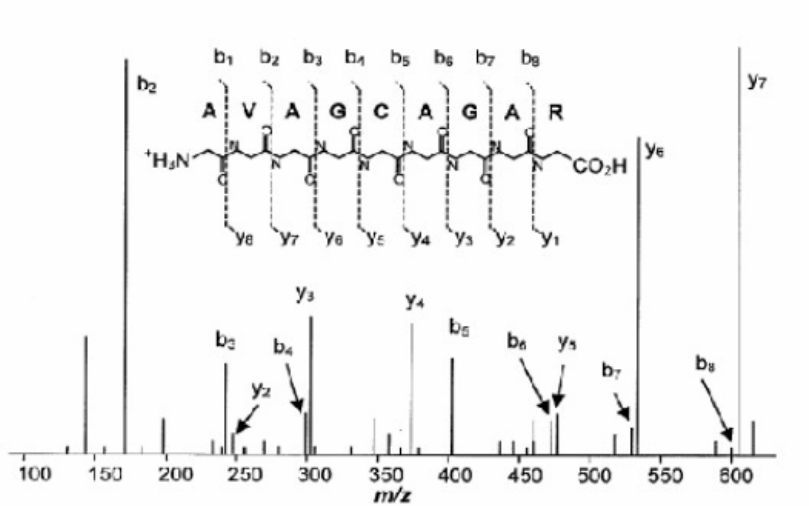
Consistent building blocks (sugars)
Consistent fragmentation allows in-silico fragment prediction
Pre-calculated fragments from known structures can be stored in database (use NIST-MS-Search)
Algorithm works also on-the-fly without database
De-novo algorithms work for truly unknown structures

See Oscar and FragLib
See GlySpy

Source: Congruent Strategies for Carbohydrate Sequencing.
3. OSCAR: An Algorithm for Assigning Oligosaccharide Topology from MSn Data
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1435829

Simulation of peptide fragmentations (De-novo sequencing of peptides)





Principle:

De-novo sequencing of peptides (determine amino acid sequences)

De-novo algorithms can perform permutations and combinatorial calculations from all 20 amino acids (superior if the sequence is not found in a database)

Highly dependent on good mass accuracy (less than 1 ppm) of precursor ion and MS/MS fragments

Generate match score by matching in-silico fragments against experimental MS/MS spectrum

Problems:

Leucine and isoleucine have same mass

Post translational modifications (PMTs)

Missing fragment peaks

Picture source: MWTWIN help file2 (Monroe/PNNL)

Picture 2 source: Tandem mass spectrometry data quality assessment by self-convolution Keng Wah Choo and Wai Mun Tham http://www.biomedcentral.com/1471-2105/8/352

The Last Page - What is important to remember:

Fragmentation and rearrangement rules and ion physics can be programmed into algorithms
→ Abundance calculations are problematic

Prediction of isomer substructures from mass spectra is possible
→ Works for reproducible mass spectra

A simplified simulation of mass spectra and simulation of fragmentation pattern
is only possible for certain molecule classes
→ Works only for peptides, lipids, oligosaccharides, alkanes
→ Does not work for all other molecules
→ Does not work with complex (side chain) modifications

Machine Learning Methods for simulation and prediction of mass spectra
require a large pool of diverse experimental mass spectra and MS^n spectra for training

Tasks (42 min):

Download one of the following tools:
MOLGEN, MOLGEN-MS, AMDIS, OMMSA, OSCAR or any free/commercial/demo program for in-silico peptide fragment determination or de-novo sequencing.
Report on use.

Literature (36 min):

Mathematical tools in analytical mass spectrometry [DOI]
Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells [DOI]
Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry [PDF]
Mass Analysis Peptide Sequence Prediction [LINK]

Links:

Used for research: (right click – open hyperlink)

http://scholar.google.com/scholar?hl=en&q=%22Simulation+of+mass+spectra
http://scholar.google.com/scholar?num=100&hl=en&lr=&safe=off&q=+Simulation+of+%22mass+spectral+fragmentation
http://www.google.com/search?num=100&hl=en&safe=off&q=in-silico+prediction+tandem+mass+spectra&btnG=Search
http://www.aseanbiotechnology.info/Abstract/21020883.pdf
http://www.google.com/search?hl=en&q=GNU+polyxmass%2C&btnG=Google+Search
http://www.google.com/search?hl=en&q=C41H76N2O15&btnG=Google+Search
http://www.google.com/search?num=100&hl=en&safe=off&q=MOLGEN+MS&btnG=Search
http://www.google.com/search?hl=en&q=G.+L.+Sutherland&btnG=Google+Search

GlySpy and the Oligosaccharide Subtree Constraint Algorithm (OSCAR)
See Mass Frontier for further discussion
MOLGEN-MS [LINK]

Of general importance for this course:

http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Structure_Elucidation/