V5 – peak detection

Detecting peaks in observed data is a common task in many fields.

Program for today:

- Principles of peak detection
- Peak detection in biomedical 1D-data
  - ChIP-seq data
  - MS data
- Peak detection in biomedical 2D-data
  - breathomics

Today, in lecture #5, we will discuss the issue of identifying peaks in a series of data points.

This is a typical problem in diverse areas of bioinformatics and in data analysis in general.

Of course, there exist many different solutions.

Which one is most suitable for a particular problem depends a lot on the kind of data.

Peak detection - basics

Computer scientists
(-> Cormen book)
are mostly interested in devising methods to determine peaks most efficiently
-> Divide & Conquer strategy

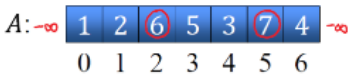
Noise is often irrelevant to computer scientists.

Instead, **bioinformaticians** must detect peaks in noisy data most precisely.

This an algorithm from the idealized world of CS …

1D Peak Finding

- Given an array $A[0..n-1]$:



- $A[i]$ is a **peak** if it is not smaller than its neighbor(s):
  $$A[i-1] \leq A[i] \geq A[i+1]$$
  where we imagine
  $$A[-1] = A[n] = -\infty$$
- Goal: Find *any* peak

https://courses.csail.mit.edu/6.006/spring11/lectures/lec02.pdf

In computer science, one typically deals with very accurate data.

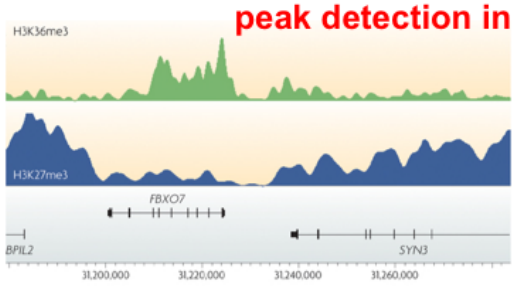
In the 1D example shown on the right, one can easily see that the red-circled entries in fields 2 and 5 are local peaks.

They fulfil the simple requirement that they shouldn‘t be smaller than their left and right neighbors.

Algorithms for finding peaks in such perfect data are described e.g. in the classic book by Cormen et al. with the title „Introduction to Algorithms“.

In contrast, bioinformaticians must detect peaks in inherently „noisy“ data = data that is subject to sizeable fluctuations due to biological and technical variation.

peak detection in ChIP-seq data



Regulation of gene expression is one of the fundamental means by which cells adapt to internal and external environ-ments.
Many **regulatory mechanisms** rely on **modifying the DNA** either through covalent modification or by intermolecular interactions.

H3K36me3: Lys36 of histone 3 is tri-methylated

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) data provide readouts of

- the location and frequency of binding of a transcription factor (TF) or
- the distribution of histone modifications that are used by the cell to establish or maintain specialized chromatin domains.

Park J, Nature Reviews Genetics, 10, 669 (2009)
Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

As first example, we will discuss the case of histone modifications.

These are an important type of epigenetic marks and consist of posttranslational modifications (methylation, acetylation, phosphorylation …) of lysine and other amino acids in the N-terminal flexible tails of histone proteins.

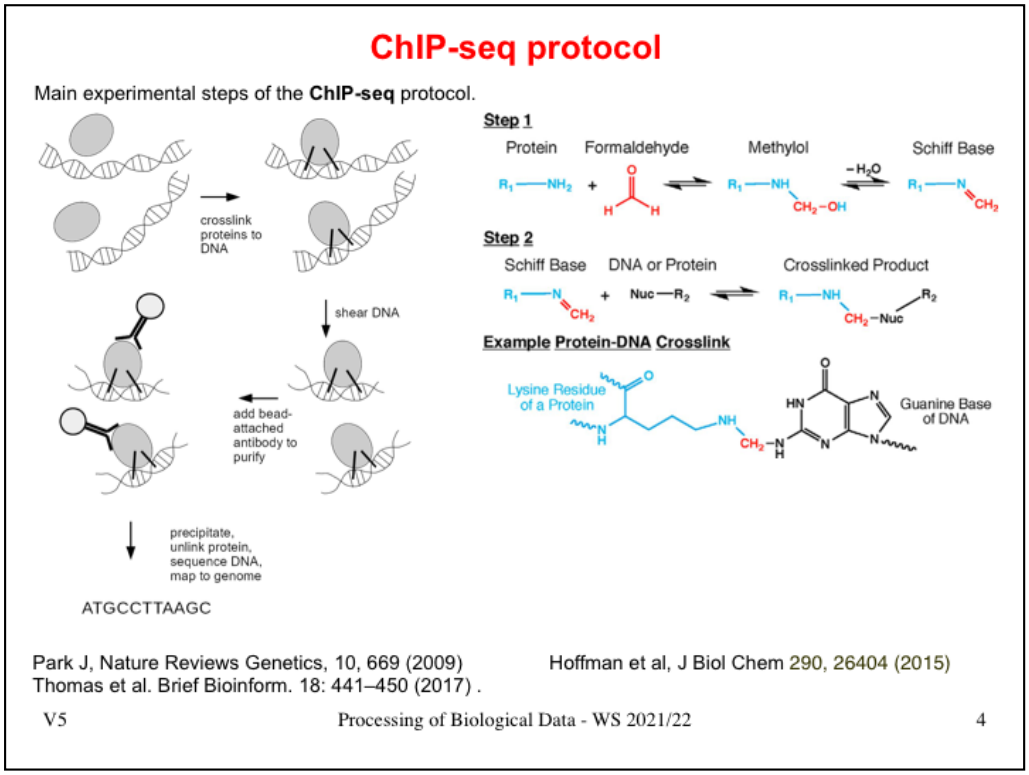
Shown in the figure are the two marks H3K36 me3 (tri-methylation of lysine36 of histone #3) and H3K27me3 along the genome sequence.

Also marked are the exons of two genes, FBXO7 and SYN3. The vertical lines or bars indicate the position of exons.

H3K36me3 is typically enriched in the gene body region (inside the gene, not in its promoter or enhancer regions) and associated with active gene transcription.

H3K27me3 is typically a repressive histone modification of nearby genes.

Histone marks can be detected by the ChIP-seq method that will be explained on the next slide.



Experimentally, histone marks are nowadays ususally detected by the ChIP-seq method (Chromatin Immuno Precipitation followed by sequencing) that is illustrated on the left.

First, DNA is **crosslinked** to bound proteins e.g. by applying **formaldehyde**, see right figure.

Formaldehyde crosslinking is routinely employed for detection and quantification of protein-DNA interactions, interactions between chromatin proteins, and interactions between distal segments of the chromatin fiber.

The DNA-protein mixture is then sheared into ~500 bp DNA fragments by **sonication** (application of ultrasound, induces DNA vibrations) or by digesting the free DNA ends with the enzyme **DNA nuclease**.

Then, an antibody is added to the mixture that is attached to a bead that can later be used to „fish“ the antibodies from the sample. One selects for this a particular antibody that binds selectively e.g. to a histone protein carrying a particular histone mark.´The antibodies are then „fished“ from the solution. Subsequently, the protein-DNA crosslinks are broken up and the DNA is sequenced.

One assumes that all DNA prepared in this way was bound e.g. to the histone protein carrying the particular histone mark.

This experimental strategy is quite labor intensive and costly.

Every histone mark needs to be detected in a separate experiment using a different special antibody.

peak detection in ChIP-seq data

Data for ChIP-seq peak calling: stacks of **aligned reads** across a genome.

Some of these stacks correspond to the **signal of interest**.

Many other stacks are regarded as experimental noise.

Typically, there are 3 – 5 data sets of replicates.

Regions are scored by the number of tags in a window of a given size.

Then they are assessed by **enrichment** over control.

Different applications of ChIP-seq produce different types of peaks.

Most current tools are designed to detect **sharp peaks** (TF binding, histone modifications at regulatory elements)

Alternative tools exist to detect **broader peaks** (expressed/repressed domains).

Park J, Nature Reviews Genetics, 10, 669 (2009)
Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

Now we discuss the output of the final sequencing step of a ChIP-seq experiment.

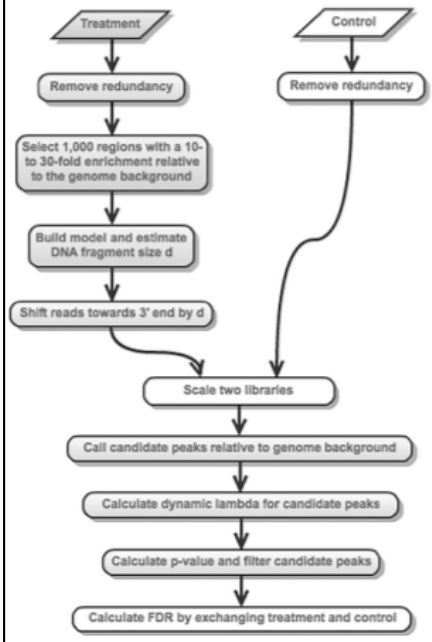
One obtains sequencing reads that belong to the DNA sequences that were „protected“ by the protein of interest (e.g. a histone protein) against digestion by DNA nuclease or against DNA breakage during sonication.

Thus, one can assume that these DNA sequences bind specifically to the protein of interest. Of course, these regions will not only consist of the DNA stretch that makes physical contacts with the protein. The regions will extend a bit further. The sequencing reads may also contain further regions that are included by accident (experimental noise or unspecific binding events).

Some of this noise can be suppressed by performing several replicate experiments.

One checks which regions show a higher coverage (enrichment) over the background of the full genome.

MACS: popular for detecting peaks in ChIP-seq data



MACS slides a window of 2× sonication size across the genome to identify regions that are significantly enriched relative to the genome background.

MACS models the number of reads from a genomic region as a **Poisson distribution** with dynamic parameter λ_{local}.

$$f(n, \lambda) = (\lambda^n e^{-\lambda})/n!$$

Based on λ_{local}, MACS assigns every candidate region an enrichment p-value. Those regions passing a user-defined threshold (default 10^{-5}) are reported as the final **peaks**.

Zhang et al. Genome Biol. (2008) 9, R137
Feng et al. Nature Prot 7, 1728 (2012)

V5 Processing of Biological Data - WS 2021/22 6

MACS is a very popular tool to detect peaks in ChIP data.

It considers the average read coverage in a window relative to the background.

The Poisson distribution (compare V4) is a statistical distribution that is often used to model stochastic processes.

Here, one assumes that obtaining NGS reads from a genomic sample is such a stochastic process.

Regions in the upper tail of the distribution (default 10^{-5}) are reported as peaks.

Needed for this is an estimate of the lambda parameter.

MACS does not use a uniform lambda for the full sample, but a local lambda for the local segment.

Features of ChIP-seq peak detection methods

Table 1. Features of peak calling methods

	GEM	BCP (TF)	BCP (Histone)	MUSIC	MACS2	ZINBA	TM
Locating the potential peaks							
High resolution	Yes	Yes	No	Yes	Yes	No	Yes
ChIP and input sample signals combined	No	No	No	No	No	Yes	Yes
Multiple alternate window sizes	Yes	Yes	Yes	Yes	No	No	No
Use of variability of local signal	Yes	Yes	Yes	No	Yes	Yes	No
Ranking of peaks							
Binomial test	Yes	No	No	Yes	No	No	No
Poisson test	No	Yes	No	No	Yes	No	No
Normalized difference score	No	No	No	No	No	No	Yes
Use of underlying genome sequence	Yes	No	No	No	No	No	No
Posterior probability of enrichment	No	No	Yes	No	No	Yes	No

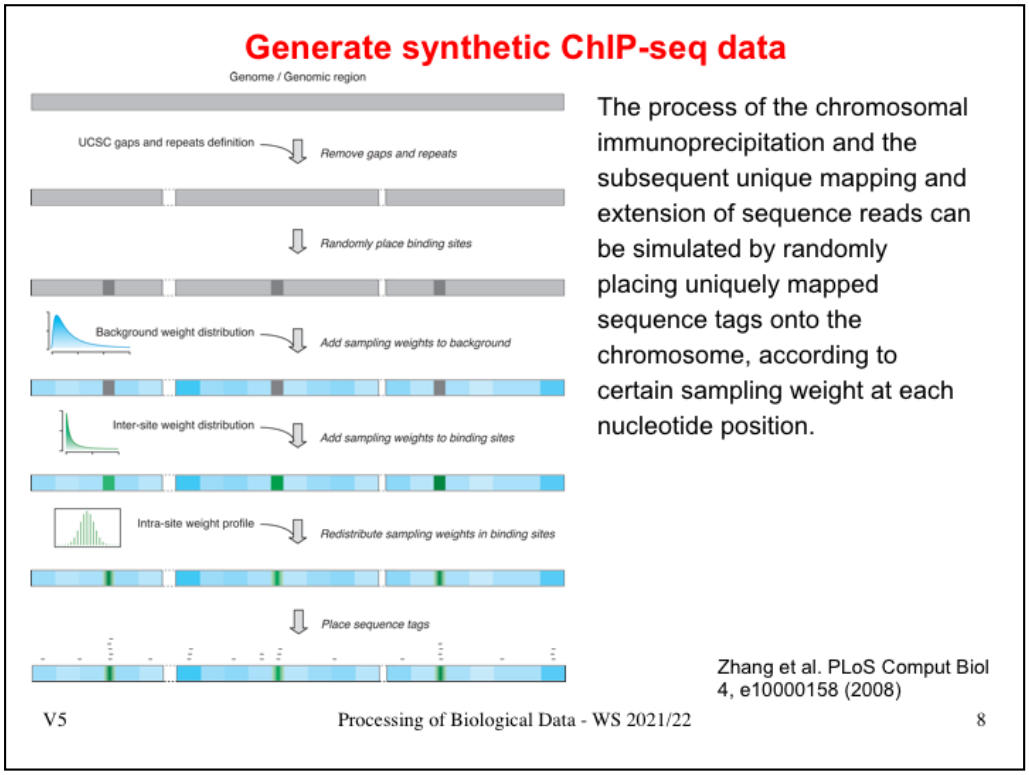
Representative selection from over 30 existing tools.

Park J, Nature Reviews Genetics, 10, 669 (2009)
Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

Thomas et al.: https://www.ncbi.nlm.nih.gov/pubmed/27169896

This is a comparison of several tools that are used to identify ChIP-seq peaks.

GEM is a 2-step method. In the second step, GEM also considers the motif content of the analyzed sequences (red circle).



Presented here is a protocol to generate synthetic ChIP data.

Link to Zhang et al. paper: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000158

Gap regions in the UCSC genome assembly are excluded. Also, repetitive regions are excluded (row 2).

In row 3 row, we place synthetic transcription factor binding sites that should be detected by the ChIP-seq protocol.

In row 4, we select a suitable (blue colored) probability distribution for the expected read coverage (looks like a Poisson distribution) of the background and assign a coverage to each sequence region. Based on this distribution, many regions will get an average (low) coverage. Few regions will get a high coverage (darker blue).

For the binding sites, we use a different (green) probability distribution for their coverage (row 5).

In row 6, the coverage of each binding site is adjusted to follow somehow a Gaussian profile.

Finally, in the bottom row, we generate synthetic sequence reads. Their coverage matches the previously assigned coverage values.

Comparison of actual and synthetic ChIP-seq data

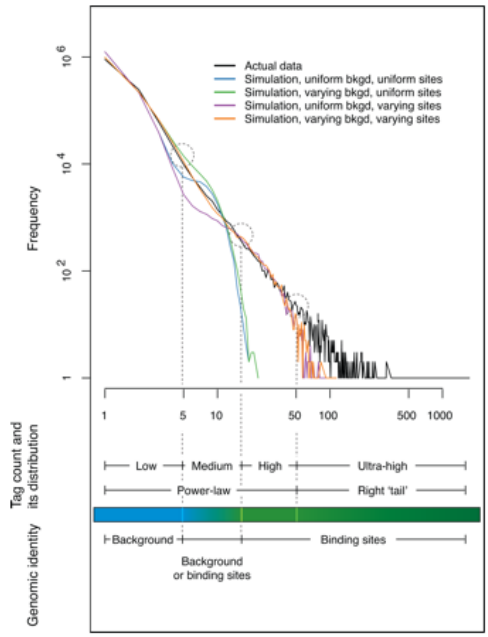


Figure shows the actual tag count distribution and the simulated ones generated under different background and site models.

Actual data contains more sites with ultra-high tag counts (right trail of distribution).

Zhang et al. PLoS Comput Biol 4, e1000158 (2008)

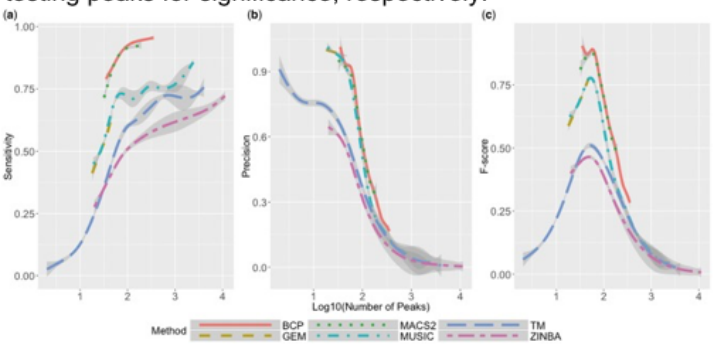
It makes quite a difference whether one assumes a uniform background or varying backgrounds. For a uniform background, every nucleotide position in the background is given one as its sampling weight. For a varying background, every adjacent 1-kb block in the background is given a random weight drawn from a pre-specified underlying distribution and all nucleotide positions in a block are assigned the same weight.

The authors distinguished 4 regions of varying tag counts: low / medium / high / ultra-high. Tag clusters with low and high (including ultrahigh) tag counts are almost certain to be background and binding sites, respectively. Because there is a mixture of signals, the true identities of the clusters with medium tag counts are much less certain, and thus some form of thresholding is necessary.

Benchmarking of ChIP-seq peak calling

Abstract the peak calling problem into two sub-problems:

- identifying peaks and
- testing peaks for significance, respectively.



BCP and MACS2 perform best.

Sensitivity (**a**), Precision (**b**) and F-score (**c**) as a function of the $\log_{10}$ of the number of called peaks for 6 peak calling methods on 100 simulated transcription factor ChIP-seq data sets.

Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

(left) The **sensitivity** is also called **true positive rate** (TRP) or **recall**. TRP = TP / P = TP / (TP + FN).

The more peaks exist (x-axis from the left to right: 10^1, 10^2, 10^3, 10^4), the better all methods perform in terms of sensitivity.

(Middle) Precision PPV is also called positive predictive value. PPV = TP / (TP + FP)

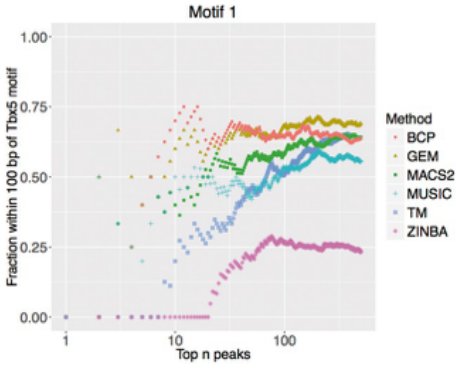
Precision measures how many identified peaks are correct. Here, the performance decreases steadily from left to right. The more peaks exist, the more difficult it is to detect the correct ones.

(Right) The F1-score is a measure that combines sensitivity and precision. It is the harmonic mean of precision and sensitivity F1-score = 2 · PPV · TPR /(PPV + TPR)

Consequently, it shows an optimal performace near 10^2 reads.

This comparison was done based on a simulated data set for which the correct answer is known.

Performance on real data from Tbx5 ChIP-seq experiment



Fraction of top *n* peaks within 100 bp of the Tbx5 motif 1 for the 6 methods.

BCP and GEM perform particularly well = high fraction with 100 bp.

Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

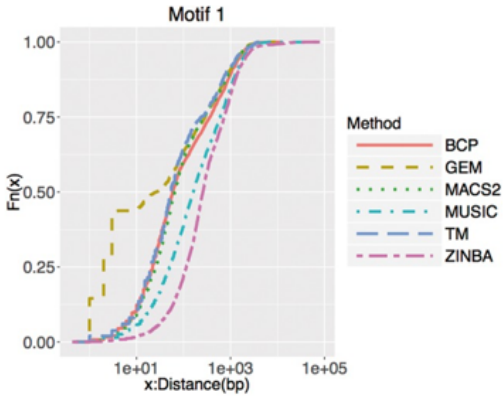
Here, ChIP-seq was used to identify binding positions of the transcription factor Tbx5 on the genomic DNA. This example shows real data.

The precise binding motif where a transcription factor binds to DNA is known for many transcription factors including Tbx5.

One can identify such motifs e.g. with the MEME tool by checking for often occurring DNA strings in the ChIP-data for this transcription factors.

Here, several methods can identify the precise location of about half the Tbx5 binding positions to about 10 bp and even more to about 100 bp.

Performance on real data from Tbx5 ChIP-seq experiment



Empirical distribution of the shortest distance to the Tbx5 motif 1 of the significant peaks called by the 6 methods.

GEM peaks are closer to a Tbx5 motif than any other method.

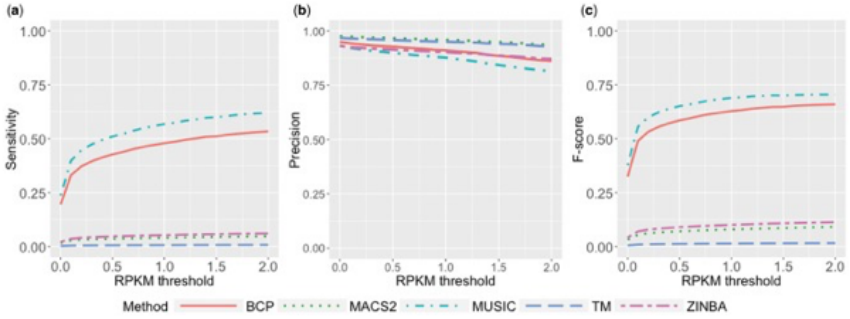
Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

This is the cumulative distribution of the plot on the previous slide.
About 90% of the regions are detected within 1000 bp.

Benchmarking of ChIP-seq peak calling

Histones typically have wider peaks than TFs.

Test how well H3K36me3 peaks overlap genes that are actively transcribed.



Sensitivity (**A**), precision (**B**) and F-score (**C**) of the overlap of the called significant peak regions with active gene bodies for H3K36me3 data.

The threshold for defining active genes was varied from 0 to 2 RPKM.

MUSIC and BCP perform best.

Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

H3K36me3 is a mark that is characteristic for actively transcribed genes.

Benchmarking of ChIP-seq peak calling: key points

Peak calling using Chip-seq data consists of 2 sub-problems: identifying candidate peaks and testing candidate peaks for statistical significance.

Methods that explicitly combine the signals from ChIP and input samples to define candidate peaks are less powerful than methods that do not.

Methods that use windows of different sizes to scan the genome for potential peaks are more powerful than ones that do not.

Methods that use a Poisson test to rank their candidate peaks are more powerful than those that use a Binomial test (not shown).

Thomas et al. Brief Bioinform. 18: 441–450 (2017) .

Summary by Thomas et al.

Basics of mass spectroscopy

3 key stages of a basic mass spectrometer (no high-end instrument):

1. Ionization.

Molecules in a sample may be vaporized by heating. Then, an electron beam bombards the vapors, which converts the vapors to ions.

Because mass spectroscopy measures the mass of charged particles, only ions will be detected. Neutral molecules will not be seen.

Ions are created by either adding electrons to a molecule (yields negatively charged ion) or abstracting electrons from a molecule (yields positively charged ion).

2. Acceleration and Deflection.

Next, the ions are sorted according to their mass in 2 stages.

Acceleration is simple Coulombic attraction. The positive ions created in the ionization stage accelerate towards negative plates at a speed dependent on their mass. Lighter molecules move quicker than heavier ones.

Deflection: the ions are then deflected by a magnetic field. The extent of deflection is again dependent on mass.

https://bitesizebio.com/6016/how-does-mass-spec-work/

In the second example of this lecture, we will discuss the task of identifying peaks in mass spectroscopy data.

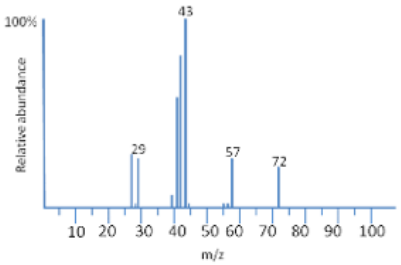
We have already introduced the basic principle of MS in lecture V2.

This is a quick reminder of the main principles.

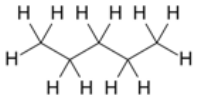
Basics of mass spectroscopy

3. Detection.

Ions of increasing mass eventually reach the detector one after another. This yields a spectrum as shown in the figure.



Simplified mass spectrum of **pentane** produced by a mass spectrometer.



https://bitesizebio.com/6016/how-does-mass-spec-work/

Shown here is the MS spectrum of the simple alkane molecule pentane shown at the bottom.

A carbon atom has mass 12 Da, a hydrogen has mass 1 Da.

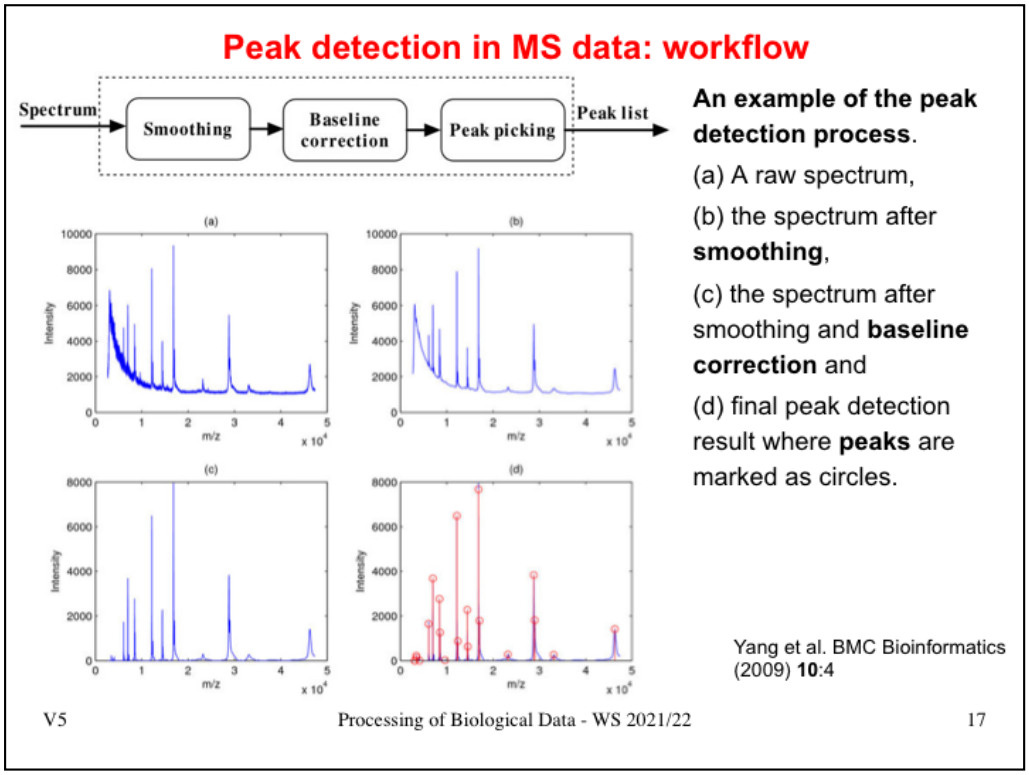
Hence, the mass of an intact pentane molecule with 5 carbon atoms and 12 hydrogens is 5 x 12 + 12 x 1 = 72 Daltons.

This is the right-most peak in the upper spectrum. Apparently, this molecule was detected with charge z = 1, giving a m/z ratio of 72.

Also detected are peaks at 57 Da (4 carbons with 9 hydrogens – meaning that one of the terminal carbon atoms has 3 hydrogens attached to it, the other one has 2 hydrogens),

43 Da (3 carbons with 7 hydrogens), and at 29 Da (2 carbons with 5 hydrogens).

The peak at 43 Da is highest showing that ionization of pentane mostly produces fragments with 3 carbon atoms.

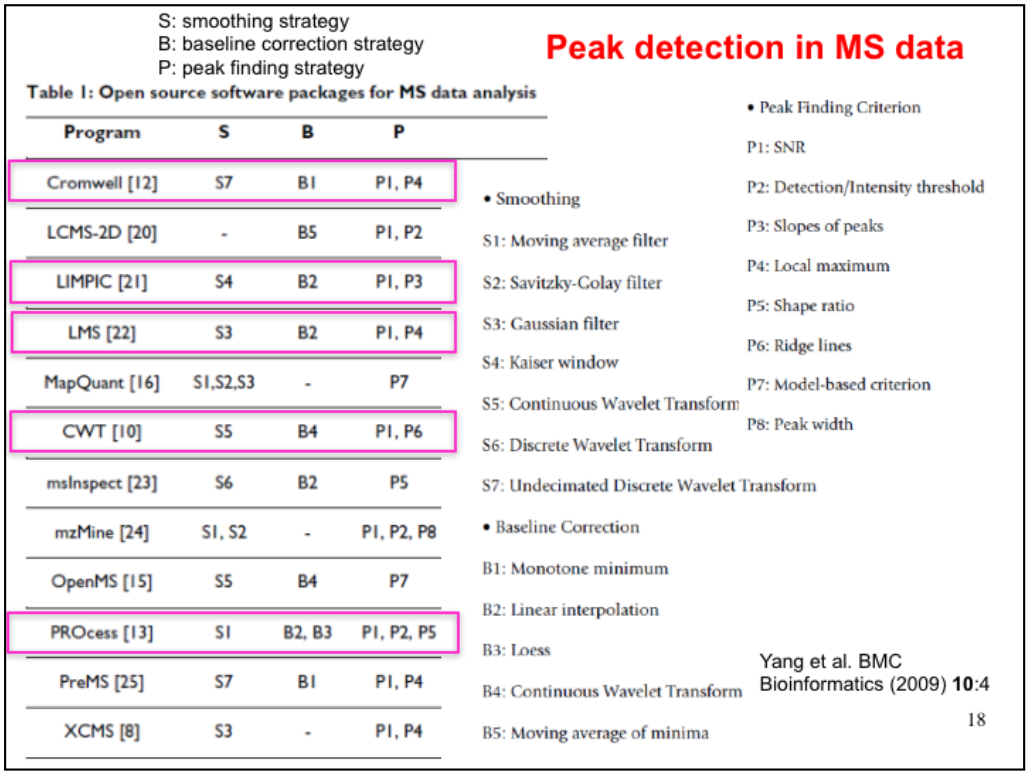


This is the main protocol for processing of raw MS m/z data and identification of peaks.

First, the raw data is smoothened (a -> b). This suppresses many small intensity peaks.

Then, (b -> c) a baseline signal is removed (this is high (4000 to 6000 intensities) at small m/z values, and converges to an intensity of around 1000 for large m/z.

This step makes sure that one can identify peaks against a uniform background intensity of 0.



In this benchmark, the authors compared 12 tools that use various strategies for smoothing (S), baseline correction (B) and for peak finding (P).

Peak detection in MS data: smoothing

Aim: remove high-frequency (likely unimportant) variations from the data

Approach: replace current value *y(n)* by an average taken over its neighbor points.

Moving average filter

$$y[n] = x[n] * w[n] = \frac{1}{2k+1} \sum_{i=-k}^{k} x[n-i]$$

2k +1 is the **filter width**

$$w[n] = \frac{1}{2k+1}\ ,\ -k \le n \le k$$

* stands for "convolution"

Gaussian filter

$$y(t) = x(t) * w(t) = \int_{-\infty}^{+\infty} x(\tau) w(t-\tau) d\tau$$

$$w(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}$$

Yang et al. BMC Bioinformatics (2009) **10**:4

A typical approach for smoothing of the raw data is to replace actual values y(n) or y(t) by averages taken over a local region.

The simplest approach is a „**moving average filter**". Here, one simply adds the values of the *k* values to the left and the *k* values to the right to the central value and divides the sum by *2k* +1.

This average is then assigned as smoothened value to the central data point.

An alternative is applying a **Gaussian filter** that takes into account essentially all data points from –infinitiy to +infinity, but weights the contribution of each point by the negative exponential of its quadratic distance *t* to the central point (as in the Gaussian distribution). Again, this weighted average is assigned as smoothened value to the central data point.

Peak detection in MS data: continuous wavelet transform

CWT

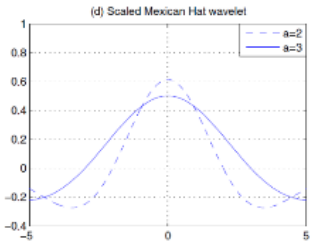
$$y(t) = x(t) * w(t) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x(\tau)\psi\left(\frac{t-\tau}{a}\right)d\tau$$

$$w(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t}{a}\right)$$

ψ(t) is a **wavelet function**,

e.g. a **Mexican-hat wavelet**
(an inverted parabola, that is squeezed (in the middle) and flattened (at the sides) by multiplication with an exponential function)

$$\psi(t) = \frac{2}{\sqrt{3}\pi^{1/4}}(1-t^2)e^{-t^2/2}$$



Yang et al. BMC Bioinformatics (2009) **10**:4

Another smoothing method is to weight neighboring data points by a so-called Mexican-hat wavelet, see figure.

This belongs to the so-called continuous wavelet transforms (CWT).

Peak detection in MS data: peak identification

Signal-to-noise ratio (SNR)

Different methods define noise differently. E.g. noise may be estimated as:

- 95-percentage quantile of absolute continuous wavelet transform (CWT) coefficients of scale one within a local window.
- the median of the absolute deviation (MAD) of points within a window.

Slopes of peaks

This criterion uses the shape of peaks to remove false peak candidates.

- A peak candidate is discarded if both **left slope** and **right slope** are smaller than a threshold.
- This threshold may e.g. taken as half of the local noise level

Yang et al. BMC Bioinformatics (2009) **10**:4

Now we introduce different methods for identifying peaks in the smoothened data.

The SNR method tries to identify peaks as „signals“ relative to the normal fluctuation („noise“) of the data.

The noise is identified e.g. as the area including most (95%) of the data points or as MAD (see lecture 4, slide 22).

The „Slopes of peaks“ method inspects the shape of any peak.

Left slope and right slope need to be steeper (i.e. the first derivative of the signal) than a certain threshold.

This criterion was likely developed to prevent detection of very broad and slowly rising mountains.

Peak detection in MS data: peak identification

Local maximum

A peak is a local maximum of *N* neighboring points.

Shape ratio

A "peak area" is computed as the area under the curve within a small distance of a peak candidate.

A "shape ratio" is then computed as the peak area divided by the maximum of all peak areas.

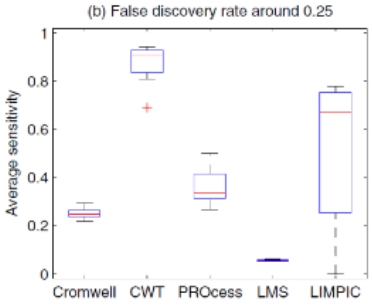
The shape ratio of a **peak** must be larger than a threshold.

Yang et al. BMC Bioinformatics (2009) **10**:4

A local maximum is simply the largest data point among all its neighbors.

The shape ratio requires that the peak area should exceed a certain threshold. This excludes peaks that appear like sharp needles.

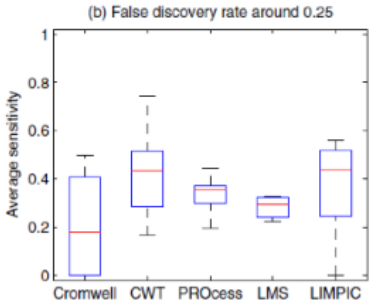
Peak detection in MS data: continuous wavelet transform



Performance on simulated data that was generated using a model that incorporates some characteristics of real MALDI-TOF mass spectrometers.

CWT performed best in this comparison.

The reason is likely that its **shape** matches best the shape of experimental MS peaks.



Aurum Dataset is a high resolution data set, which contains spectra from 246 known, individually purified and trypsin-digested protein samples taken with an ABI 4700 MALDI TOF/TOF mass spectrometer.

Yang et al. BMC Bioinformatics (2009) **10**:4

The left example tests how well different peak detection methods can identify peaks in synthetically generated data.

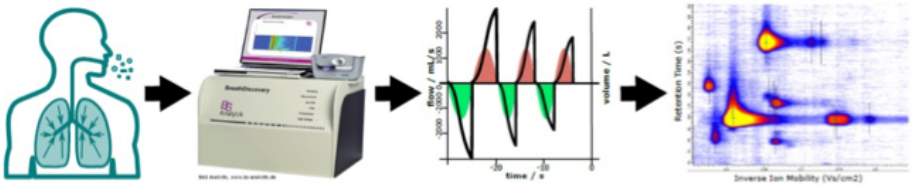
The right example is an experimental benchmark data set of 246 given proteins that have been digested by trypsin.

On both examples, CWT (detecting a Mexican hat profile) worked best.

Case study: peak detection in breathomics

MCC/IMS: Ion mobility spectrometry (IMS), coupled with multi-capillary columns (MCCs) is gaining importance for biotechnological and medical applications.

With MCC/IMS, one can e.g. measure the presence and concentration of volatile organic compounds in the air or in **exhaled breath** with high sensitivity.



Kopczynski, Rahmann,
Algorithms for Molecular Biology
(2015) **10**:17
PhD thesis Ann-Christin Hauschild,
Saarland University (2016)

Now we will discuss a related example, detected peaks in 2D data from MS.

Precisely, the field of breathomics attempts to identify organic compounds in exhaled breath.

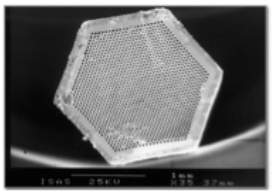
The aim is – as can be expected – to use this method as early detection for diseases of the individual.

Shown here is how the exhaled breath is analyzed by a MS instrument and then processed in several steps of data analysis.

MCC/IMS experiments: output

In an MCC/IMS experiment, a mixture of several unknown volatile organic compounds is separated in two dimensions:

(1) By the **retention time** *r* in the capillary column (the time required for a particular compound to pass through a multi-capillary column (MCC), see right fig.). The retention time is proportional to the substance's **affinity** for the stationary phase.



PhD thesis Mario Wachowiak, UdS (2017)

(2) By the **drift time** *d* through the ion mobility spectrometer.

Instead of the drift time itself, one uses a quantity that is normalized for pressure and temperature called the **inverse reduced mobility** (IRM) *t*.

This allows comparing spectra taken under different or changing conditions.

These two separation steps are carried out sequentially. Compounds leaving the column at different times are separately analyzed by MS.

Kopczynski, Rahmann, Algorithms for Molecular Biology (2015) **10**:17

If the sample contains many different species, their MS signals could largely overlap if we try to analyze them only in a 1D m/z spectrum.

Therefore, breathomics separates the data in two dimensions.

Along the y-axis, we plot the retention time how fast a substance passes a capillary column. One uses a 17 cm long, 3mm diameter column that contains about 1000 thin capillaries. This architecture largely increases the surface of the capillary walls. The walls are coated with a thin „stationary phase“, often a silica polymer.

Along the x-axis, we plot a kinetic property measured by the mass spectrometer.

MCC/IMS experiments: inversed reduced mobility

the reduced mobility of an ion drifting through a buffer gas in an electric field is given by

$$K = (3q/16N)(2\pi/\mu kT)^{1/2}(1/\Omega_D) \qquad (1)$$

where q is the charge of the ion and m its mass, N is the density of the neutral molecules and M their mass, μ is the reduced mass $\mu = mM/(m + M)$, k is the Boltzmann constant, T is the effective temperature, and Ω_D is the collision cross section.

From K, one derives the reduced (normalized) ion mobility:

$$Ko = K(273/T)(P/760)$$

and the **inversed reduced ion mobility** (after some rearrangement)

$$Ko^{-1} = 1.697 \times 10^{-4}(\mu T)^{1/2}\Omega_D$$

Karpas et al. JACS 111, 6015 (1989)

The reduced mobility *K* of an ion drifting through a buffer gas is related to the square root of the charge over mass ratio, see eq. (1).

Instead of the mass of the ion, one considers the „reduced mass“ that is combined from the ion mass and the mass of the gas molecules in the buffer gas inside the mass spectrometer.

The details of converting *K* into the inversed reduced ion mobility are not relevant for us here.

IM spectrum-chromatogram

r : set of (equidistant) **retention time** points

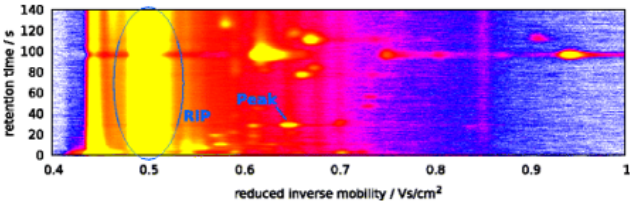
t : set of (equidistant) **IRMs** where a measurement is made,

e.g. 12500 time points taken every 4 x 10^{-6} s -> 50 ms in total)

Then the data is an $|r| \times |t|$ matrix of measured ion intensities,

which we call an *IM spectrum-chromatogram* (IMSC).

The matrix can be visualized as a **heat map**.



An IM spectrometer uses an ionized **carrier gas**. These ions are present in every spectrum in addition to the analyte ions, and they create the **reactant ion peak (RIP)**.

The reduced inverse ion mobility (x-axis) is proportional to the drift time.
The colors reflect the signal height:
[white (low) < blue < purple < red < yellow (high signal)].

Kopczynski, Rahmann, Algorithms for Molecular Biology (2015) **10**:17

This figure shows the raw data of an IM spectrum-chromogram from which we want to identify the peaks of individual organic molecules.

Remember, plotted on the y-axis is the retention time through the MCC capillary column in seconds. Compounds that pass quickly, will show up at the bottom (short retention times).

Plotted on the x-axis are signals with different reduced inverse mobilities. The MS measurements are carried out sequentially for different retention time points.

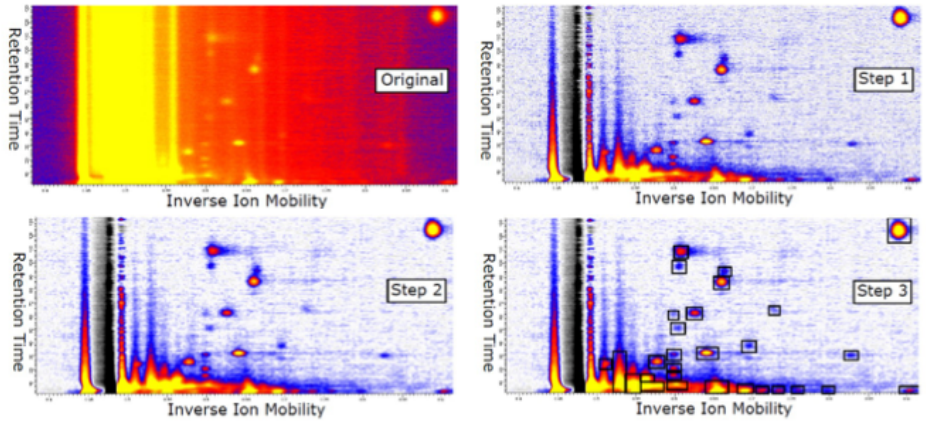
This spectrum is provided to us as an r x t matrix.

The brightest peak of the spectrum (colored in yellow) is a peak at x = 0.5 that is present at all retention times.

This RIP peak belongs to the ions of the carrier gas in the MS spectrometer and is not relevant for us.

The other yellow and red peaks shown right of the RIP are only present at one retention time.

breathomics



Example of a processing strategy of MCC/IMS data involving
(Step 1) RIP-detailing (removal of RIP peak)
(Step 2) denoising and baseline correction
(Step 3) peak picking.

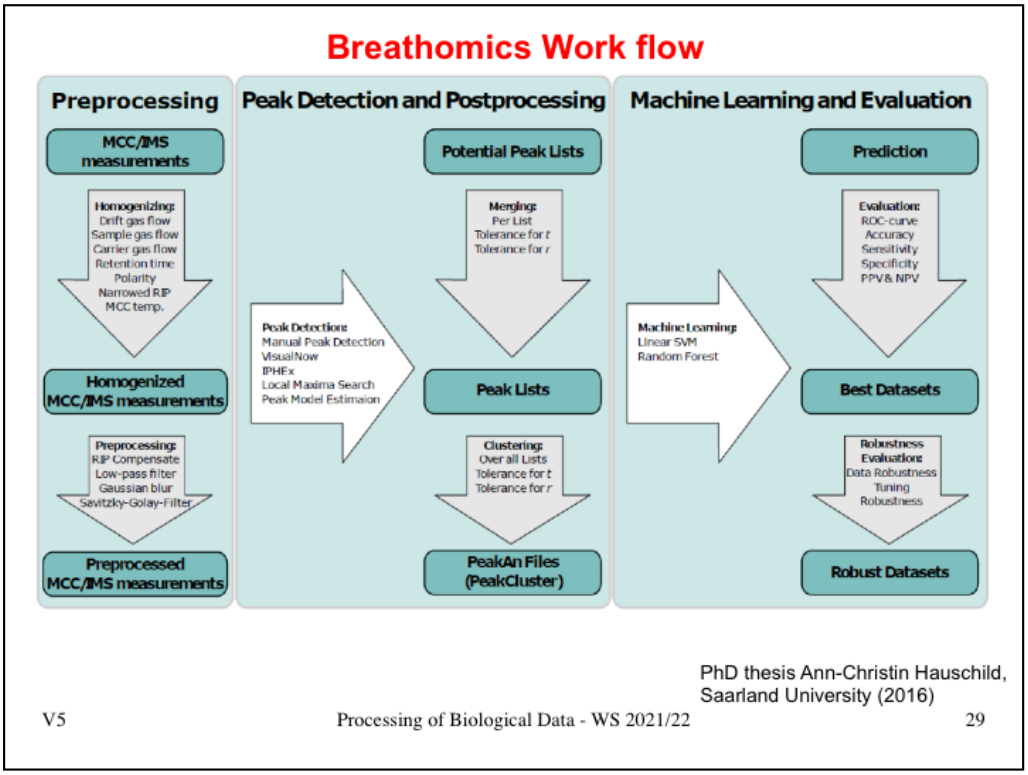
PhD thesis Ann-Christin Hauschild, Saarland University (2016)

These are different steps of breathomics analysis.

In step 1, the RIP peak is removed from the spectrum.

In step 2, the signal is denoised (smoothed) and the baseline is subtracted.

In step 3, the peaks of interest are identified, here marked by boxes.



This is a flowchart presented in the PhD thesis of Dr. Ann-Christin Hauschild who worked on this topic in the group of Dr. Jan Baumbach.

Jan Baumbach was previously a young group leader at CBI and is now a full professor at TU Munich.

Manual Peak detection

The easiest and most intuitive way of peak detection is **manual evaluation** of a visualization of the measurement.

The human eye and visual cortex is optimized for pattern recognition in 3D.

Therefore one can immediately spot most of the peaks in the measurement.

There are several **drawbacks** of the manual approach:

- it is **time consuming** and therefore inappropriate in a high-throughput context,
- the results depend on a **subjective** assessment, and are therefore hardly reproducible.

Nevertheless, manual evaluation is still the state of the art for the evaluation of smaller MCC/IMS data sets.

Manually created peak lists are used as "**gold standard**" in MCC/IMS studies.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

Humans are best able to identify the most interesting peaks in such a complicated spectrum.

Local maxima search

According to this criterion, a point is a **local maximum** if all 8 neighbors in the matrix have a lower intensity than the intensity at the central point.

We call the neighborhood of a point “significant" if

- its own intensity,
- the intensity of its 8 neighbors, and
- that of A additional adjacent points (e.g. A = 2),

lie above a given intensity threshold I.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

Dr. Hauschild compared different algorithms and their ability to precisely identify peaks.

A simple „local maximum search“ identifies central points as peaks with higher intensity than that of all 8 neighboring points.

Even very tiny differences would then be reported as local maximum.

Therefore in a second step, „significant“ maxima are identified as those points that are higher at least by a given minimal intensity threshold than their neighbors.

Merged peak cluster localization (MPCL)

The MPCL consists of two phases: (1) clustering and (2) merging.

(1) each data point in the chromatogram is assigned to one of 2 classes, either **peak** or **non-peak**.
For this, one uses a clustering method that is based e.g. on the Euclidean distance metric of the intensity values.

(2) neighboring data points that are both labeled as **peak** can be assumed to belong to the same peak and are **merged together**.

(3) each peak of the analyzed measurement is characterized by its **centroid point**, i.e. the data point, which has the smallest mean distance to all other points in this peak region.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)
PhD dissertation Sabine Bader Dortmund University (2008)

Also clustering can be used to identify peaks.

Watershed algorithm

Here, the IMS chromatogram is treated like a **landscape** including hills and valleys.

The algorithm starts with a water level above the highest intensity followed by a continuous lowering of the level while uncovering more and more of the local maxima.

At each step, the new uncovered data points are annotated by the label of adjacent labeled neighbors. Those data points that remain unlabeled are identified as a new peak and receive a new label.

The highest data point among a set of new labeled positions denotes the **peak** coordinate.

The algorithm stops if all data points are labeled or the level drops below a given threshold.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

The watershed algorithm is a widely used algorithm in image processing: https://en.wikipedia.org/wiki/Watershed_(image_processing).

This is an overview of the algorithm when it is applied for peak detection.

Watershed algorithm: implementation

The watershed algorithm can be implemented as a **priority queue** where all data points (having 2D coordinates) are sorted by magnitude.

(1) The largest data point is extracted and labeled first.

(2 - n) This is followed by the next largest point in the queue and so on.

- Each point drawn out of the queue is compared with its neighbors in space.
- If the neighbors are of equal or larger value, the extracted point is given the same label as its largest neighbor.

(comment: if of equal value, neighbor has not necessarily been labeled …)

- In contrast, if the data point is larger than its neighbors (i.e. the neighbors have not been labelled sofar), the data point is given a new label to indicate that it is part of another peak.

(n + 1) This procedure is repeated until the queue is empty.

Latha et al. Journal of Chromatography A, 1218 (2011) 6792– 6798

The Watershed algorithm was adapted for 2D chromatographic peak detection by S. Reichenbach, M. Ni, V.V.A. Kottapalli, Chemom. Intell. Lab. Syst. 71 (2004) 107.

Peak model estimation

In the PME method, the expectation maximization (EM) algorithm is used to optimize the parameters of a mixture model from a given set of starting values.

The algorithm requires a given set of “seed" coordinates for each peak to be modeled.

In general, any peak detection method is suitable to provide these initial " seeds". However, the quality of the results strongly depends on the chosen seed-ing approach.

Utilizing the EM algorithm, each peak is described by a model function consisting of two shifted Gaussian distributions and an additional peak volume parameter.

Finally, the set of model functions plus a noise component describe the whole MCC/IMS measurement.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

The PME method will not be explained in detail here.

LMS : Automated **local maxima search**

MPCL : Automated peak detection via **merged peak cluster localization** supported by VisualNow

WST : Automated peak detection via **water shed transformation** implemented in IPHEx,

PME : **Peak model estimation** approach by the PeaX tool.

breathomics

Table 6.1: Number of peaks detected by all methods. Number of peak clusters after merging the peak lists (postprocessing).

Method	# Peaks	# Peak Clusters
Manual (VisualNow)	1661	41
LMS (PeaX)	1477	69
MPCL (VisualNow)	4292	88
WST (IPHEx)	5697	420
PME (PeaX)	1358	69

Table 6.2: Overlap of the five peak detection methods. The overlap of the peak list A (row) and peak list B (column) is defined as the number of peaks in V that can be mapped to at least one peak in W. Note that the resulting mapping count table is not symmetric.

Method	Manual	LMS	MPCL	WST	PME	Software
Manual	1661	911	1522	1184	791	VisualNow
LMS	868	1477	1096	1074	1128	PeaX
MPCL	2667	2233	4292	2341	2082	VisualNow
WST	1112	1009	1157	5697	912	IPHEx
PME	737	1086	983	926	1358	PeaX

PhD thesis Ann-Christin Hauschild, Saarland University (2016) p.95

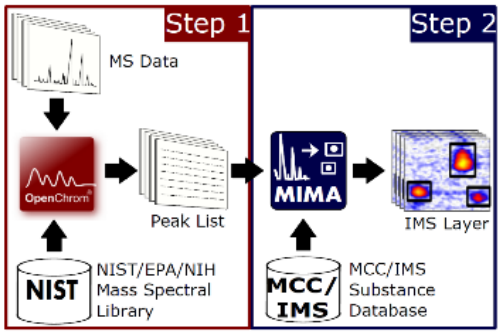
Thesis: https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/26718

Table 6.1: The number of identified clusters varies between 41 and 88 except for the Watershed algorithm WST

Table 6.2: The overlap between the peaks identified by different methods is quite reasonable.

Automated metabolite detection

Aim: annotate peaks to chemicals (not only detecting peaks)



Collect **reference IMS data** for compound library

Run IMS experiment on sample of interest - compare against reference data

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

Testing of the peak annotation was performed using samples containing known reference molecules.

This is similar to the spike-in protocol presented in lecture #4, slide 38.

Proof of principle

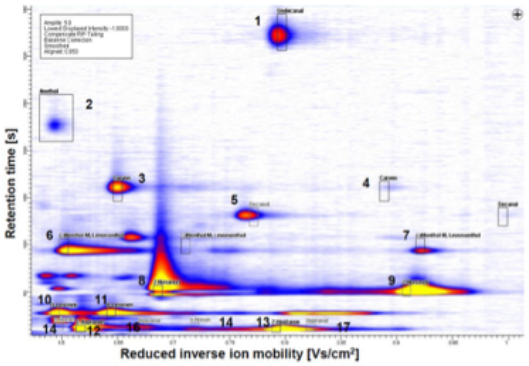


Table 7.1: Automatically identified signals

No.	CAS	compound
1	112-44-7	undecanal
2	104-46-1	anethol (trans-anethol)
3	6485-40-1	carvon (monomer)
4	6485-40-1	carvon (dimer)
5	112-31-2	decanal
6	2216-51-5	(-)menthol (monomer)
7	2216-51-5	(-)menthol (trimer)
8	821-55-6	2-nonanon (monomer)
9	821-55-6	2-nonanon (dimer)
10	5989-27-5	D-limonen (monomer)
11	5989-27-5	D-limonen (dimer)
12	110-43-0	2-heptanon (monomer)
13	110-43-0	2-heptanon (dimer)
14	111-84-2	n-nonan (monomer)
15	111-84-2	n-nonan (dimer)
16	111-71-7	heptanal (monomer)
17	111-71-7	heptanal (dimer)

Test on a mixture of 7 reference compounds

17 signals in the measurement could be matched: 12 of them originate from the reference compounds (including dimers and trimers)

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

Signals #5 and #14 - #17 were not part of the reference analyte mixture, but could be clearly identified as decanal, n-nonan and heptanal.

They are components in many fragrances and could have entered the IMS from the room air.

Application: can one detect COPD in exhaled breath?

Chronic obstructive pulmonary disease (COPD) is an umbrella term used to describe chronic lung diseases that cause a permanent blockage of airflow from the lungs, which is not fully reversible (WHO).

The most prominent symptoms are

- breathlessness,
- a chronic cough, and
- excessive sputum production.

Airways and lungs react to noxious particles or gases, like smoke from cigarettes or fuel, with an increased inflammatory response.

The World Health Organization (WHO) reported COPD as one of the four most frequent causes of death.

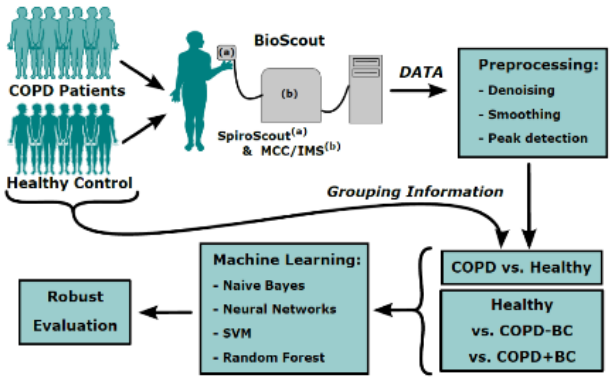
PhD thesis Ann-Christin Hauschild, Saarland University (2016)

It would be great if one could use breathomics for detection of complicated diseases.

Obvious candidates that may affect the composition of exhaled breath are lung diseases.

Application: can one detect COPD in exhaled breath?

Westhoff et al. (2011) took MCC/IMS breath probes of 42 COPD patients and of 35 healthy volunteers (HC).



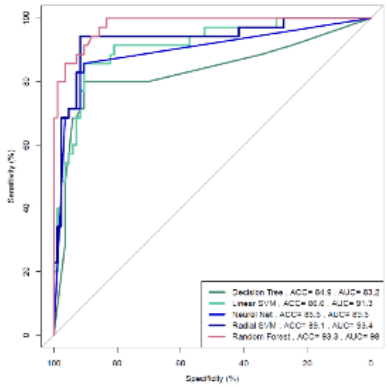
PhD thesis Ann-Christin Hauschild, Saarland University (2016)

The software was tested on a public MCC/IMS dataset of COPD patients and healthy controls.

Application: can one detect COPD in exhaled breath?

Table 5.1: Results of the two-class-classification problem, evaluating the differences between COPD and the HC.

Method	AUC	Accuracy	Sensitivity	Specificity
Decision Tree	81	85	91	71
Linear SVM	83	87	92	74
Naive Bayes	79	82	87	71
Neural Net	86	89	93	80
Radial SVM	87	89	92	83
Random Forest	92	94	98	86



Distinguishing COPD patients from healthy controls based on IMS spectra of exhaled air works really well!

Distinguishing COPD patients from patients that also have breast cancer did not work equally well.

PhD thesis Ann-Christin Hauschild, Saarland University (2016)

V5 Processing of Biological Data - WS 2021/22 41

This study is described in https://www.mdpi.com/2218-1989/5/2/344/htm.

In the spectra, characteristic peaks of 120 volatile organic compounds were identified that are present in at least three of the patients' measurements.

Then, the 120 metabolites were clustered by hierarchical agglomerative clustering (HAC) and Pearson correlation.

By a suitable clustering threshold, the set of metabolites was split into 40 subsets, one for each cluster of correlating metabolites.

All clusters with less than three compounds were excluded, yielding a total of 14 metabolite sets.

Using this data, COPD could be separated from healthy samples with good success (85-95% success).

A Random Forest classifier achieved the highest accuracy.

Summary

Peak detection is a frequent task in diverse areas of biology.

The challenge is posed by the noisy nature of biological data and the irregular shape of peaks.

Testing and benchmarking of methods is typically done with synthetic (artificially generated) data.

Peak detection and judging their significance are equally important tasks.

Today, we discussed examples ranging from identification of the peaks of certain histone marks over 1D mass spectroscopy to 2D MCC/IMS-based breathomics analysis.

These examples illustrated that one needs to adapt various peak identification methods to the data type and problem being studied.