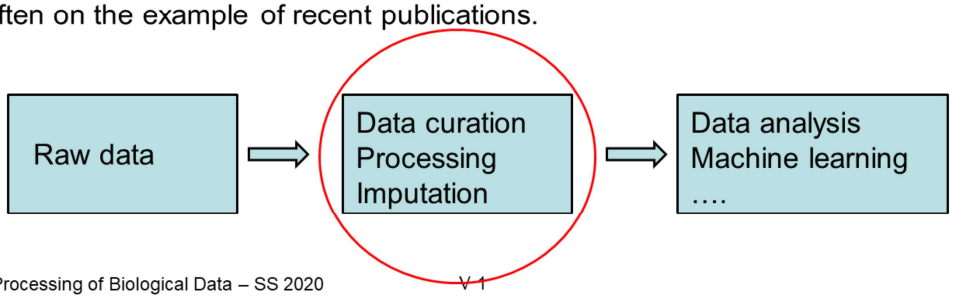
V1 Processing of Biological Data

Leistungspunkte/Credit points: 5 (V2/Ü1)
This course is taught in English language.

The material (from books and original literature) are provided online at the course website:
https://www-cbi.cs.uni-saarland.de/teaching/ss-2020/special-topic-lecture-bioinformatics-processing-of-biological-data/

Topics to be covered:
This course will discuss the handling of different sorts of biological data, often on the example of recent publications.



Processing of Biological Data – SS 2020 V 1 1

Welcome to the lecture „Processing of Biological Data“ in summer 2020.

Due to the special conditions during the Corona pandemic, this lecture will be taught by video conferencing.

I recommend that you should first read the content of each slide and then either read the text in this comment block or listen to the recorded Audio.

At the bottom, I visualized the typical flow of a bioinformatics project from raw data over several preprocessing steps listed in the middle box to the data analysis/machine learning block on the right.

Obviously, the last block is expected to reveal the biological or biomedical insight that may be contained in the provided raw data.

Often, answering a biological question relies on selecting 2 suitable groups of samples and comparing them.

So the left and right blocks are obviously most interesting.

However, as I will point out in this lecture, the middle block is equally important in reality as the other two.

Tutorial

We will handout 6 **bi-weekly assignments**.
Groups of up to two students can hand in a solved assignment.

Send your **solutions** by e-mail to the responsible tutors
until the time+date indicated on the assignment sheet.

The **bi-weekly tutorial** on Tuesday 12.45 am – 2.15 pm (time is negotiable) will discuss the assignment solutions.

On demand, the tutors may also give some advice for solving the new assignments.

Assignments will be connected to the content of the lecture and will deal with typical tasks of a bioinformatician who is processing biological data.

Some assignments will contain programming tasks, others can be solved either with available software or even by hand.

Schein conditions

The successful participation in the lecture course („Schein“) will be certified upon fulfilling

- Schein condition 1 : ≥ 50% of the points for the assignments
- Schein condition 2 : pass **final oral exam** at end of semester (late July). Each student takes an individual exam.

The **grade** on your „Schein“ will equal that of your final exam.

Those who failed or missed the final exam can take a **oral re-exam** at the beginning of WS21.

Note that this is different from our standard regulations (e.g. bioinformatics III) where normally everybody can take the written re-exam.

The final exams will be conducted as oral exams of around 20 minute duration.

Depending on how the Corona epidemic develops, oral exams may either be conducted via video conferencing or in person.

As it is very time-consuming to conduct oral exams, we will not offer the chance for a re-exam to those who have passed the first exam.

Planned lecture - overview

V1: bacterial data (*S. aureus*): clustering / PCA

V2: bacterial data/DNA methylation: prediction of missing values (BEclear)

V3: differential gene expression, detection of outliers

V4: MS proteomic data, imputation, normalization, protein arrays

V5: peak detection, breathomics

V6: shape detection, processing of kidney tumor MRI scans

V7: genomic sequences, SNPs

V8: functional GO annotations

V9: curve fitting, data smoothing (AKSmooth …)

V10: protein X-ray structures: titration states, hydration sites, multiple side chain and ligand conformations, superposition … protein-protein complexes: crystal contacts, interfaces, …

V11: analysis of MD simulation trajectories: correlation of snapshots, remove CMS motion

V12: multi-variate analysis

V13: integrative analysis of multidimensional data sets

Processing of Biological Data – SS 2020 V 1 4

In the summer term 2020, due to the Corona epidemic, the lecture will likely contain only 11 instead of 13 lectures.

Depending on how things go and what assignments will be scheduled, we may e.g. skip the normal lectures 10 and 11 on analysis of protein structures and on data from molecular dynamics simulations.

Data preprocessing

Data preprocessing is one of the most critical steps in data mining.

Data preprocessing methods are divided into 4 categories:

- Data cleaning
- Data integration
- Data transformation
- Data reduction

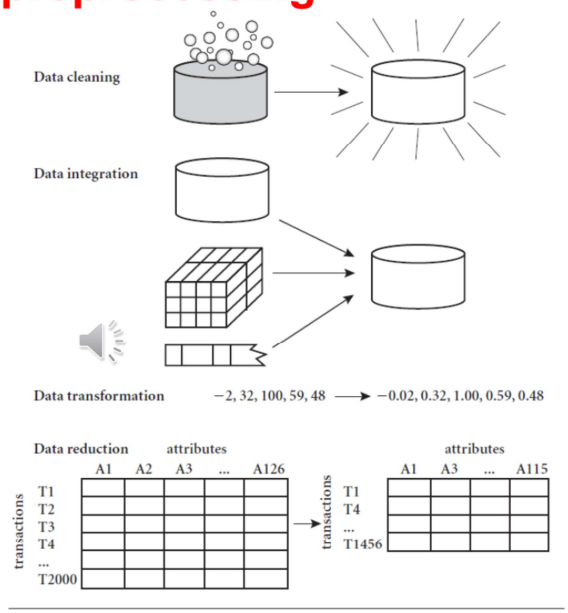


Figure 2.1 Forms of data preprocessing.

Data Mining: Know It All by Ian H. Witten et al. Publisher: Morgan Kaufmann (*2008)*

Although this may not be very obvious to you right now, data preprocessing is a very crucial step of data processing.

If we do not remove problematic data points from the data set at the beginning and if we do not apply proper normalization in the next step, then all downstream processing becomes highly problematic and possibly misleading.

Listed here are 4 categories of data preprocessing methods.

In this lecture, we will discuss examples from all 4 listed categories.

Data preprocessing

- **Data cleaning**: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- **Data integration**: using multiple databases, data cubes, or files.
- **Data transformation**: normalization and aggregation.
- **Data reduction**: reducing the volume but producing the same or similar analytical results.
  - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data Mining: Know It All by Ian H. Witten et al. Publisher: Morgan Kaufmann (*2008)*

Listed here are some typical tasks for all 4 categories of data preprocessing methods.

Note that I took these examples from the text book of Ian Witten listed at the bottom.

In fact, these tasks are typical to any data mining field.

So after you have taken this lecture, you will be ready for data mining☺

ZBI ZENTRUM FÜR BIOINFORMATIK

UNIVERSITÄT DES SAARLANDES

Whole Genome Sequence Typing and Microarray Profiling of Methicillin-Resistant *Staphylococcus aureus* isolates

(1) Classification of MSSA / MRSA *S. aureus* strains in Saarland (PLoS ONE 2012)

(2) DFG Germany-Africa project (J. Clin. Microbiol. 2016; Sci. Reports 2017)

Co-workers

(1) Ruslan Akulenko, Ulla Ruffing, Mathias Herrmann, Lutz von Müller,

(2) StaphNet Consortium led by Mathias Herrmann, funded by **DFG**

Often in this lecture, we will discuss examples from our past and current research projects.

This is how I came into contact with the various tasks of data preprocessing.

You will often read very little about the data preprocessing steps in the methods section of publications.

But be assured, most if not all bioinformatics projects involve significant amount of data preprocessing.

In the example of today's first lecture, we will look at genomic data from a project related to bacterial resistance.

Together with Prof. Mathias Herrmann and Prof. Lutz von Müller from the medical department of Saarland University in Homburg, we first started with a pilot study on S. aureus samples that was published in PLoS ONE.

The results from this pilot study then helped us to acquire funding for a large scale multi-center study involving a number of groups from Germany and Africa.

Pilot study: classification of resistant *Staphylococcus aureus* strains

Table 1. Risk factors of MRSA and matched MSSA control group isolates.

Risk factors	MRSA, n (%)	MSSA, n (%)	p-value
Male	18 (39.13%)	18 (39.13%)	#
Female	28 (60.87%)	28 (60.87%)	#
<70 years	24 (52.17%)	24 (52.17%)	#
≥70 years	22 (47.83%)	22 (47.83%)	#
Hospitalisations <6 months	21 (45.65%)	21 (45.65%)	#
Inter-hospital transfer	5 (10.64%)	1 (2.17%)	ns
Previous MRSA colonization	3 (6.52%)	1 (2.17%)	ns
MRSA contacts	8 (17.39%)	4 (8.70%)	ns
Long-term care	11 (23.91%)	2 (4.26%)	0.014
Retirement home	3 (6.52%)	0 (0.00%)	ns
Diabetes mellitus	9 (19.57%)	8 (17.39%)	ns
Antibiotic therapy	21 (45.65%)	8 (17.39%)	0.007
Dialysis	3 (6.52%)	0 (0.00%)	ns
Medical devices	8 (17.39%)	0 (0.00%)	0.006
Skin lesions	6 (13.04%)	2 (4.26%)	ns

#statistical analysis was not performed for clinical criteria applied for selection of matched MSSA cases, ns = not significant.

OPEN ACCESS Freely available online PLOS | ONE

Matched-Cohort DNA Microarray Diversity Analysis of Methicillin Sensitive and Methicillin Resistant *Staphylococcus aureus* Isolates from Hospital Admission Patients

Ulla Ruffing[1], Ruslan Akulenko[2], Markus Bischoff[1], Volkhard Helms[2], Mathias Herrmann[1], Lutz von Müller[1]*

1 Institute of Medical Microbiology and Hygiene, Saarland University Medical Center, Homburg/Saar, Germany, 2 Center for Bioinformatics, Saarland University, Saarbrücken, Germany

December 2012 | Volume 7 | Issue 12 | e52487

Aim: classify MRSA / MSSA according to gene repertoire

At some point in 2011, all patients who were admitted to the university hospital during a period of 1 month

were screened for the presence of methicillin sensitive or methicillin resistant S. aureus strains.

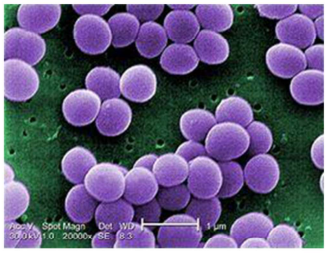
As shown in table 1, we selected 46 MRSA isolates and 46 MSSA colonized patients.

The two groups were matched for gender, age and diverse types of predisposition and exposition.

The aim of the study was to identify the clonal lineage distribution of MSSA and MRSA isolates and to detect differences in the accessory gene equipment of MRSA and MSSA isolates.

Methycillin sensitive/resistant *Staphylococcus aureus* (MSSA/MRSA)

MSSA

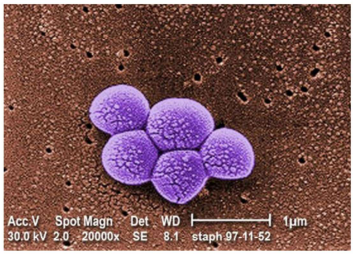


anaerobic Gram-positive coccal bacterium,

frequently part of the normal skin flora,

60% of population are carriers

MRSA



any strain of *S. aureus* with **resistance** to beta-lactam antibiotics:

- penicillins;
- cephalosporins;

Need to classify MRSA strains to detect infections, prevent transmission

MSSA S.aureus strains are „good strains“. They are sensitive to the antibiotic named methycillin.

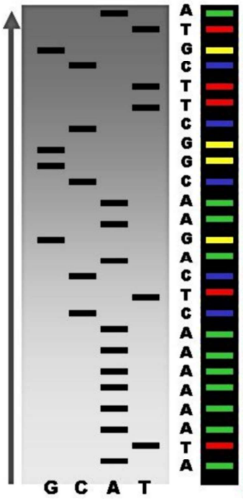
More than half of the human population carry MSSA strains in their nose.

MRSA S.aureus strains are „evil strains“ that are multi-resistant to several classes of antibiotics.

It is very important to detect MRSA early on to avoid treating patients with useless therapies and to disrupt transmission chains.

routine: Characterize MRSA by Spa-typing

- DNA preparation of polymorphic X-region of ***staphylococcus* protein A** from *S. aureus* (Spa)
  - amplify by PCR
- sequencing assignment using Ridom StaphType software



Ridom Spa Server

Spa-types:	Repeats:	Total strains:	Strain records:	Strain countries:
17897	762	398228	165914	135

In 2012, the typical detection of MRSA relied on so-called Spa-typing.

Spa stands for a protein termed protein A from S. aureus.

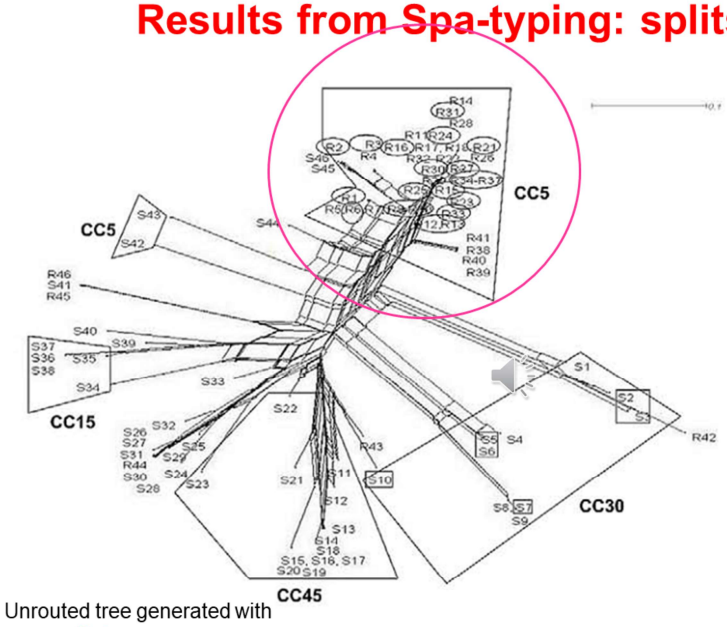
It contains a highly variable polymorphic region.

Sequencing of this region was demonstrated to be a rapid and accurate method to discriminate

S. aureus outbreak isolates from those deemed epidemiologically unrelated.

The Spa sequences can be submitted to a Webserver that classifies the submitted strain.

Results from Spa-typing: splits graph



For MSSA, *spa*-typing allowed for good discrimination of patient isolates.

However, the majority of MRSA isolates clustered into clonal complex **CC5/t003.**

This hampers sub-classification by *spa*-typing

Unrouted tree generated with www.splitstree.org

MSSA strains labeled S__
MRSA strains labeled R__

Shown here is something like a phylogeny of the 96 samples from this pilot study based on their Spa-sequences.

The labels of MSSA samples start with the letter S, those of MRSA samples with the letter R.

It turns out that MSSA samples can be well separated into different clusters, so-called clonal complexes.

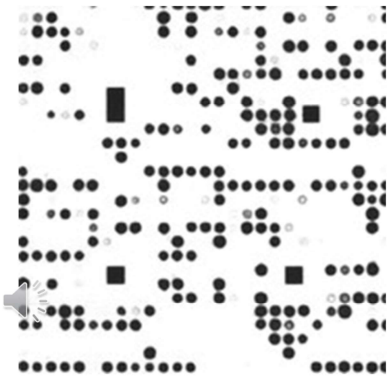
The terminology „clonal complex“ will be explained on one of the following slides.

However, most MRSA samples fall into one big cluster that belongs to CC5.

We concluded that Spa-typing alone was not able to properly resolve the MRSA samples.

Therefore, we looked for an alternative method that would characterize information about many more genes.

DNA microarray (IdentiBAC – Alere)



Microarray contains 334 DNA probes for genes/regions that are clinically relevant and/or relevant for clonal typing

alere-technologies.com

The bacterial DNA isolated from the patients is loaded on a manufactured DNA microarray.

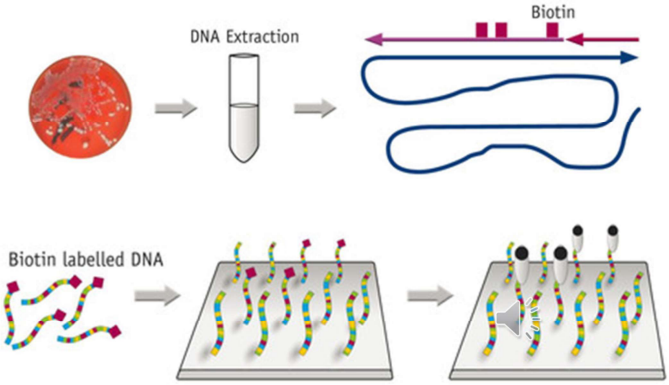
Each of its wells contains many copies of a particular DNA probe for one out of 334 regions from the S. aureus genome.

Some of these genomic regions belong to known virulence or resistance genes.

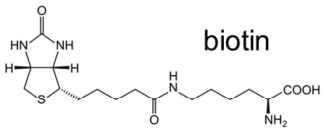
Other regions are relevant to determine to which clonal complex the bacterial strain belongs to.

In the picture on the right, black circles indicate positive hybridization meaning that this genomic region was detected in the sample.

DNA microarray principle



The extracted RNA free genomic DNA from the bacterial overnight culture is internally biotin labelled through a set of antisense primers.



The resulting single stranded and biotin labelled amplicons are hybridized to a set of discriminative probes that are covalently bound onto the microarrays.

The biotin labelled DNA bound to the probes on the array is subsequently stained.

alere-technologies.com, www.wikipedia.de

In the top row, this slide illustrates some of the technical steps involved in the preparation of bacterial DNA.

The presence of any of the 334 genetic probes in the bacterial sample is detected using antisense primers and subsequent PCR amplification.

The PCR products are termed amplicons. They are then labeled with biotin, a small molecule that is also known as vitamin B7.

The second row shows how the biotin-labeled DNA stretches are applied to the microarray.

Then, the bound PCR products are detected using a horse-radish peroxidase – streptavidin conjugate. Streptavidin binds tightly to biotin.

Then, a substrate (seramun green) is applied to the probe that is converted by the enzyme peroxidase into a dark-colored precipitate.

The colored spots are then read out automatically by the image reader.

Process microarray data (334 probes)

StaphyType Test Report

Operator	
Sample ID	2192119
Experiment ID	2192119 - {4083AD2C-7D42-4FB9-82D5-E50CC0FD6206}
Date of Result	Thu Apr 14 10:46:01 2011
Assay Name	StaphyType
Assay ID	10248
Well Position	01 (01-A)
Software Version	2009-07-09
Device	04a0022

Internal Controls

Data Quality	passed

Genetic markers for S. aureus / MRSA / PVL

Taxonomy	Species Marker (*S.aureus*) positive
MRSA (mecA)	positive
PVL	negative

Resistance Genotype

Hybridisation (Gene)	Result	Expected Resistance
mecA	positive	Methicillin, Oxacillin and all Beta-Lactams, defining MRSA
blaZ	negative	Beta-Laktamase
ermA	positive	Macrolide, Lincosamide, Streptogramin
ermB	negative	Macrolide, Lincosamide, Streptogramin
ermC	negative	Macrolide, Lincosamide, Streptogramin
linA	negative	Lincosamides

	11	46	10	33	28
MRSA (mecA)	0	0	0	0	0
PVL	0	0	0	0	0
23S-rRNA	1	1	1	1	1
gapA	1	1	1	1	1
katA	1	1	1	1	1
coA	1	0	1	1	1
Protein A	1	1	1	1	1
sbi	1	1	1	1	1
nuc	1	1	1	1	1
fnbA	1	1	1	1	1
vraS	1	1	1	1	1
sarA	1	1	1	1	1
eno	1	1	1	1	1
saeS	1	1	1	1	1
mecA	0	0	0	0	0
blaZ	0	1	0	0	0
blaI	0	1	0	0	0
blaR	0	1	0	0	0
ermA	0	0	0	0	0
ermB	0	0	0	0	0
ermC	0	0	0	0	0
linA	0	0	0	0	0

Simple idea: Compute **Euclidian distance** between samples

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Other distances are possible, also weighted distances, where some probes get higher weights.

The left picture shows the header of the output that we obtained as a PDF file from the image reader.

With a small piece of code we extracted the presence and absence of the 334 probes in each sample.

In the right plot, 0 and 1 entries denote absence and presence of about 20 genetic probes in 5 samples labeled from 11 to 28.

For example, the gene *mecA* encodes penicillin-binding protein 2A, which makes S.aureus resistant against penicillin-like antibiotics.

The task was now to express the degree of similarity between samples in a numerical way.

For this, we computed the Euclidian distance between any two columns of this matrix.

Further distance measures

An **edit distance** is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.

Edit distance variant 1: The **Levenshtein distance** allows deletions, insertions and substitutions.

Edit distance variant 2: The **Hamming distance** allows only substitutions. Hence, it only applies to strings of the same length and counts the number of positions at which the corresponding symbols are different.

Example: The Hamming distance between: "**karolin**" and "**kathrin**" is 3.
1011101 and **1001001** is 2.

The **Mahalanobis distance** is a measure of the distance between a point P and a distribution D (P. C. Mahalanobis, 1936).
It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

https://en.wikipedia.org/wiki/Category:Similarity_and_distance_measures

There exist many further distance measures that are used in diverse fields of data mining.

The Wikipedia page cited at the bottom of the slide contains links to 26 distance measures including e.g. cosine similarity, Jaccard index or overlap coefficient.

On this slide, we explain the simple Hamming distance that belongs to the class of edit distances and also mention the Mahalanobis distance.

Hierarchical agglomerative clustering based on MA data

Hierarchical clustering:

(1) Calculate pairwise distance matrix for all samples to be clustered based on their **Euclidian distances**.

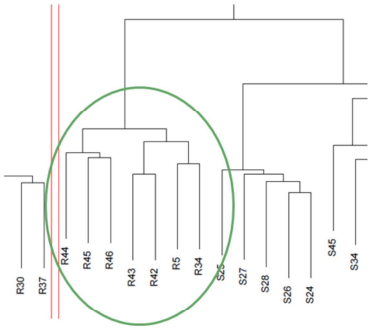
(2) Search distance matrix for two most similar samples or clusters (initially each cluster consists of a single sample).

If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives.

(3) The two selected clusters are merged to produce a new cluster that now contains at least two objects.

(4) The distances are calculated between this new cluster and all other clusters.

(5) Repeat steps 2–4 until all objects are in one cluster.



Clustering based on Euclidian distance yields almost perfect separation between MSSA/MRSA

except the encircled resistant samples

Processing of Biological Data – SS 2020 V 1 16

Based on their pairwise Euclidian distances, the bacterial samples were now hierarchically clustered. This method ís a type of agglomerative clustering and is explained on the left side.

This yielded an almost perfect separation of MRSA and MSSA samples, except for 7 resistant samples (enclosed by a green circle) that are clustered together with MSSA samples.

S. aureus in Germany vs. Africa: StaphNet

6 study sites each collected 100 isolates of healthy volunteers and 100 of blood culture or clinical infection sites

→ 1200 isolates

Aim

microbiological and molecular characterization of African *S. aureus* isolates

by DNA microarray analysis including clonal complex analysis

supplemented by Whole Genome Sequencing

Processing of Biological Data – SS 2020 V 1 17

In a follow-up project of the first pilot study, we became partners of the international StaphNet consortium that was led by Prof. Mathias Herrmann from the medical department of Saarland University.

The consortium included partners from three African countries, Mozambique, Tanzania and Gabun and from three German cities, Münster, Freiburg and Homburg.

Each site collected isolates from the nose of healthy individuals and isolates from the blood of infected patients.

The objective of this study was to compare the molecular-epidemiologic profiles of *S. aureus* isolates from Sub-Saharan Africa and Germany.

What does the microarray measure?

Naively, one can interpret the microarray result as

1 : gene is present in the strain

0 : gene is not present in the strain

However, **false negative** non-detections of particular targets may occur due to **non-binding** of the sample amplicon to the microarray's probe or primer oligonucleotide due to **polymorphisms** in the respective target gene.

On the other hand, **false positive results** may occur between highly similar probe and amplicon sequences, e. g. between agrI and agrIV.

→ check MA results by whole genome sequencing

Strauss et al. J Clin Microbiol (2016)

The samples were again processed in the same way as in the pilot study by the DNA microarray.

Ideally, one could interpret the output of the image reader as presence and absence of genes in a bacterial genome.

However, you should realize that the microarray actually relies on a PCR protocol and on the hybridization of amplicons to the probe sequences on the chip.

If a particular resistance gene of this sample contains several SNP mutations with respect to the reference genome of S. aureus, the PCR product reflecting this gene may show poor hybridization to the probe.

One would then conclude that the gene is absent, although it is in fact present, but simply contains one or more mutations.

This would be an example of a false negative testing result.

On the other hand, one can also imagine false positive results that may occur, for example, by cross-hybridization.

An amplicon sequence representing a different gene may by accident also hybridize to a similar probe that actually stands for another gene.

MA assignment to CCs confirmed by whole-genome sequencing

154 *S. aureus* isolates (182 target genes) from Germany-vs-Africa study

Result Category		Result caused by		Functional Category of genes: Identification	Regulation	Resistance	Virulence	Total	% Total
Concordant	**Positive**	Microarray and WGS (*de novo*)		829	990	1,060	8,495	11,374	40.6%
n=27,119	**Negative**	Microarray and WGS (*de novo*)		0	1,159	8,100	6,486	15,745	56.2%
(96.8 %)									
Discrepant	**False Positive**	Microarray	Mishybridizations	0	78	21	103	202	0.7%
n=909 (3.2 %)									
	False Negative	Microarray	Polymorphisms	0	3	14	140	157	0.6%
		WGS	Assembly error	88	42	16	164	310	1.1%
			Cropped contig	1	12	15	28	56	0.2%
			Not sequenced or aberrant allele	6	9	8	100	123	0.4%
	Unknown			0	0	4	24	28	0.1%
		Total number of typing results		924	2,310	9,235	15,554	28,028	100%

→ **97% agreement of MA and WGS**

Strauss et al. J Clin Microbiol (2016)

As a control, our partners from Münster therefore sequenced 154 bacterial isolates also by next generation sequencing which

can be assumed to provide more extensive and also more accurate information.

The point of the comparison was to validate the results of the DNA microarray experiments.

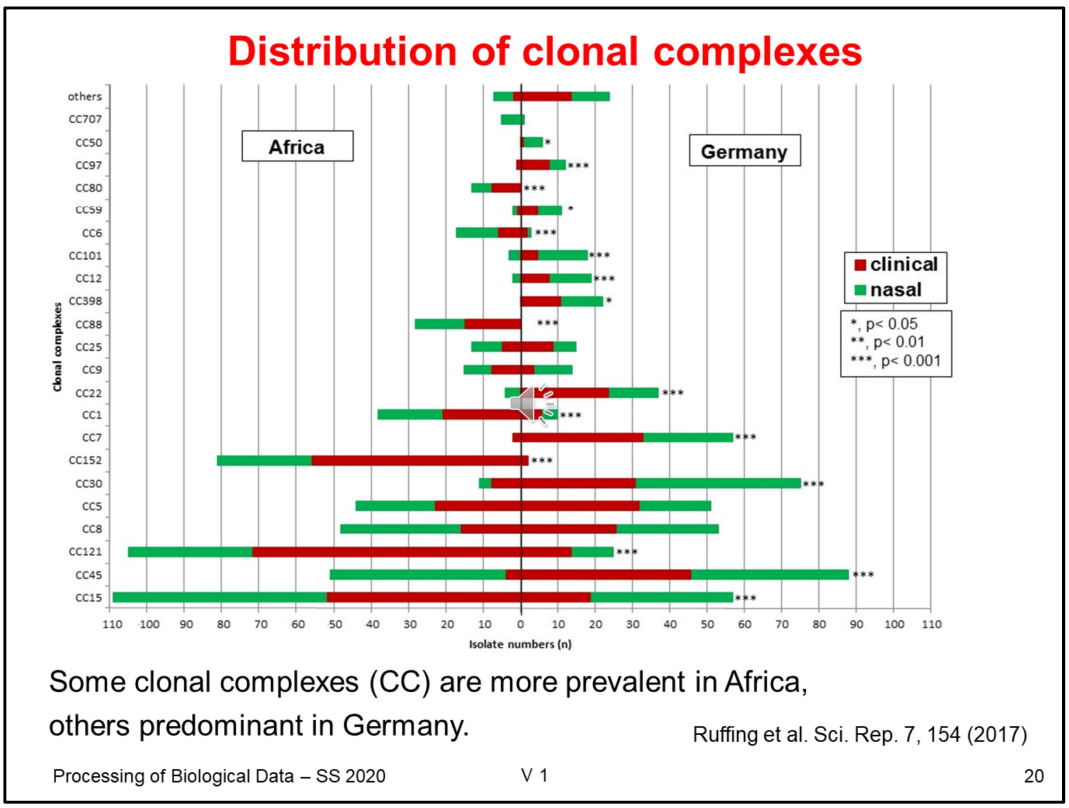
The comparison was restricted to 182 unique genes that are present on the microarray.

As shown here, the results of NGS and microarray are highly consistent or concordant.

In 40.6% of the cases, NGS and microarray jointly detected a gene, in 56.2% of the cases both methods agreed that it is absent.

This makes 96.8% agreement.

Both methods show an error rate of 1-2% due to various reasons that are listed here.



This is an overview of the samples collected in this project.

The y-axis displays different clonal complexes of Staphylococcus aureus. They are named CC followed by a number.

For S. aureus, a clonal complex contains a group of sequence types that share at least five of seven identical alleles with at least one other sequence type in the group.

Shown on the x-axis are the number of bacterial isolates of a particular clonal complex found either in Africa or in Germany.

The data is colored according to the origin, nasal isolates are colored green, clinical isolates red.

No members of CC80 and CC88 were found in Germany. No members of CC50 and CC398 in Africa.

All other CCs with at least six isolates were found both in Africa as well as in Germany.

For about half of the detected CCs, significant geographic distribution differences were found.

The statistical imbalance is marked by asterisks.

Activititv of individual probes for CCs

S2 Excel Book - Microsoft Excel

Gene distribution in African vs German *S. aureus* isolates of the 10 predominant CCs

		Groups	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German
		Numbers (n)	600	600	109	57	51	88	105	25	48	53	44	51	11	75	83	2	11	48
		Clonal complex (CC)	all CCs		CC15		CC45		CC121		CC8		CC5		CC30		CC152		CC7	
	SPECIES MARKERS	rrnD1..S..aureus.	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		gapA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		katA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		coA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		nuc1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		spa	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		sbi	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	REGULATORY GENES	sarA	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		saeS	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		vraS	100%	100%	100%	[illegible]	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		agrI..total.	35%	55%	0%	[illegible]	51%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%
		agrB.I	54%	60%	0%	0%	100%	100%	84%	92%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%
		agrC.I	57%	59%	0%	2%	96%	92%	99%	100%	100%	100%	0%	0%	0%	0%	100%	100%	100%	98%
		agrD.I	35%	55%	0%	0%	41%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%
		agrII..total.	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%
		agrB.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	98%	100%	0%	0%	0%	0%	0%	0%
		agrC.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%
		agrD.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%
		agrIII..total.	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%
		agrB.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%
		agrC.III	15%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%	97%	0%	0%	0%	0%
		agrD.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%
		agrIV..total.	37%	6%	0%	0%	59%	1%	100%	100%	6%	2%	0%	0%	0%	0%	100%	100%	0%	0%
		agrB.IV	53%	41%	0%	0%	59%	1%	100%	100%	96%	98%	0%	0%	0%	0%	100%	100%	100%	98%
		agrC.IV	23%	5%	0%	0%	59%	1%	100%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		hld	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	METHICILLIN RESISTANCE AND SCCmec TYPING	mecA	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%
		delta_mecR	2%	3%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%
		ugpQ	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%
		ccrA.1	1%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
		ccrB.1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		pIsSCC..COL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Q9XB68.dcs	1%	3%	0%	0%	0%	2%	0%	0%	10%	0%	5%	12%	0%	0%	0%	0%	0%	0%
		ccrA.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%
		ccrB.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%
		kdpA	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
		kdpB	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
		kdpC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
		kdpD.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
		kdpE.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
		mecI	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%

Shown is here the frequency of finding the probe genes of the DNA microarray in a group of bacterial isolates.

The third column contains all isolates, columns to the right contain particular clonal complexes.

The top group of genes are characteristic marker genes of S.aureus that are detected in all isolates.

The bottom group are methicillin resistance genes. Very few isolates contain them.

The largest differences are observed in the middle group of regulatory genes of the agr family.

The accessory gene regulator (*agr*) locus of *Staphylococcus aureus* encodes a two-component signal transduction system that affects the expression of surface proteins and secreted proteins and also influences the expression of many virulence genes.

Antibiotic resistance

Table S2: Rates of *in vitro* antibiotic resistance of *Staphylococcus aureus* from colonization and infection in Africa and Germany

Source	Antimicrobial agent	Resistant isolates, % (n)		p value
		Africa (n=300)	Germany (n=300)	
Colonization	Cefoxitin	2.3% (7)	0.7% (2)	ns
	Tetracycline	35.6% (107)	8% (24)	<0.001
	Erythromycin	20.3% (61)	15.7% (47)	ns
	Clindamycin	4.7% (14)	12.7% (38)	0.005
	Gentamicin	5% (15)	0.3% (1)	0.006
	Trimethoprim-sulfamethoxazole	18.3% (55)	0.3% (1)	<0.001
Infection	Cefoxitin	3.3% (10)	7.3% (22)	ns
	Tetracycline	49.7% (149)	5.7% (17)	<0.001
	Erythromycin	18.7% (56)	19.7% (59)	ns
	Clindamycin	3.7% (11)	14.3% (43)	<0.001
	Gentamicin	1% (3)	2.6% (8)	ns
	Trimethoprim-sulfamethoxazole	19.2% (58)	1.3% (4)	<0.001

NS=not statistically significant

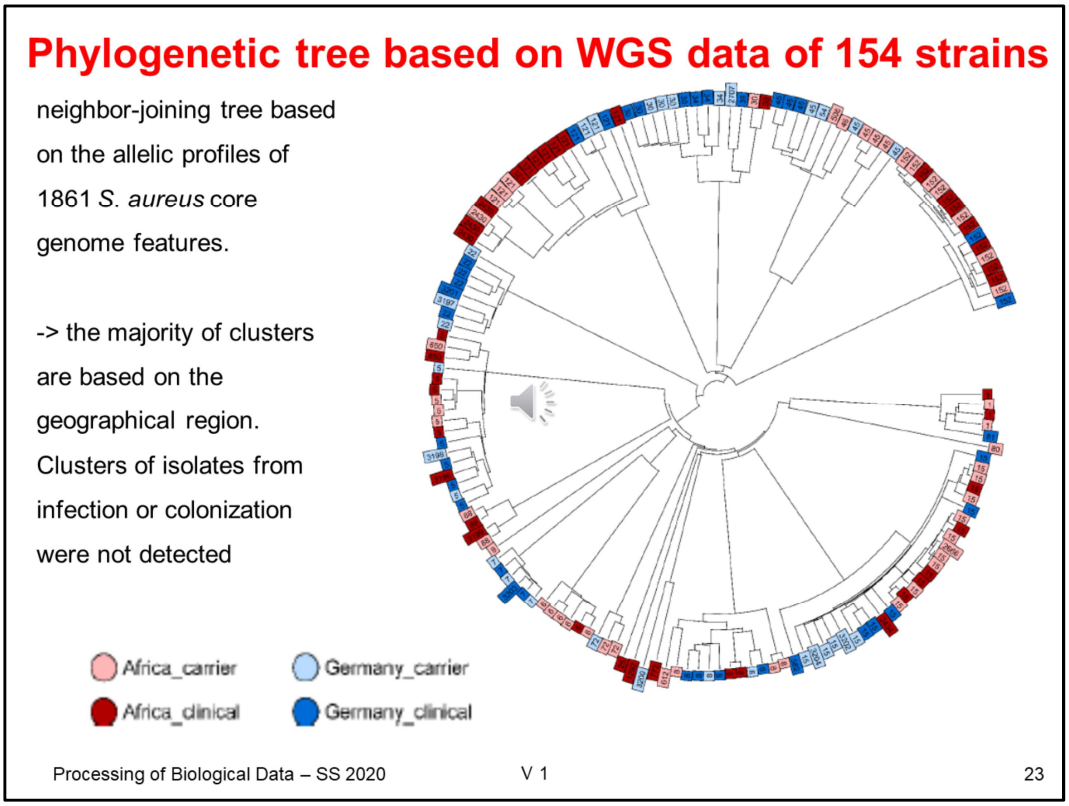
The majority of resistance genes were equally distributed among isolates from Africa and Germany. Striking differences in phenotypic resistance could be observed for tetracycline and trimethoprim-sulfamethoxazole with a larger proportion of resistant isolates in the African population, and clindamycin, with resistance more prevalent among German isolates

Ruffing et al. Sci. Rep. 7, 154 (2017)

As shown on the previous slide, the majority of resistance genes were equally distributed among isolates from Africa and Germany.

These findings correspond well to the phenotypic resistance profile against certain antibiotics which are shown on this slide.

The most striking differences are marked by red boxes.



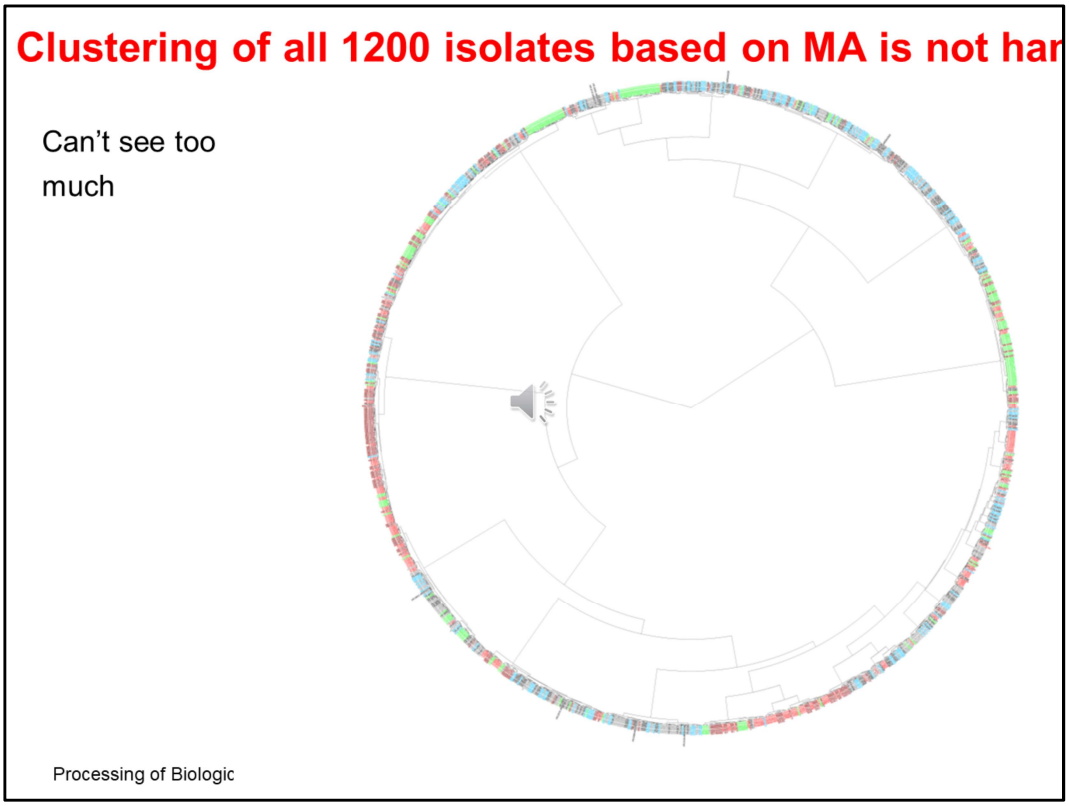
Shown is a phylogeny of 154 strains based on data from whole genome sequencing (WGS).

Reference genomes of S. aureus at NCBI contain around 2800 genes.

A so-called core genome contains 1861 genes that are detected in practically all S. aureus isolates.

The phylogeny was constructed from the sequence variations found between these 1861 genes.

Most clusters detected in this way contain samples either from Africa or from Germany.



Here, we tried to cluster all 1200 isolates based on the DNA microarray data and using the Euclidian distances as done before.

One can merely see some red, green or blue clusters, but actually the image is very messy.

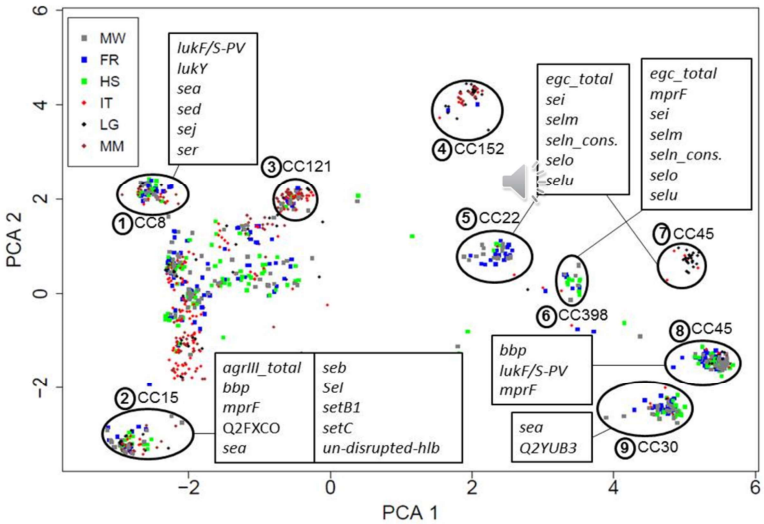
The question is whether and how one can present this data in a somehow condensed fashion that would better reveal the differences between samples.

Principle component analysis of 1200 strains

Input data: binary matrix of MA data; dimension 1200 x 334 probes

PCA identifies local gene clusters that are characteristic

for particular clonal complexes



Color code:

6 different sites

Marked in boxes:

Characteristic genes present in this cluster.

Shown here are the results of visualizing the same data – the gene inventory of 1200 bacterial samples – by a so-called principal component analysis.

The x-axis represents the projection of each sample along the first principal component vector termed PC1.

The y-axis those along PC2.

The color coding respresents the geographical origin of the probes.

Red or warm colors are used for African samples, green/blue or cold colors for German samples.

Many samples from the same clonal complexes cluster together because they share the same genes.

This is captured by the microarray data.

Listed in text boxes are genes that are enriched in the samples in a circle with respect to the background of all other probes.

So altogether, this PCA analysis looks quite successful.

PCA- intro

PCA is the most popular multivariate statistical technique.
It is used by almost all scientific disciplines.

It is likely also the oldest multivariate technique.

Its origin can be traced back to Pearson, Cauchy, Jordan, Cayley etc

This part of the lecture is based on the article
"Principal component analysis" by Herve Abdi & Lynne J. Williams in
WIREs Computational Statistics, 2, 433-459 (2010)

We will now discuss in detail how PCA works.

PCA- intro

PCA analyzes a data table **X** representing observations described by several dependent variables, which are, in general, inter-correlated.

Q: What is the difference of dependent and independent variables?

The goal of PCA is to extract the important information from the data table and express this information as a set of new orthogonal variables called **principal components** that capture the directions of **largest variance** in the data.

We will consider a data table **X** of I observations and J variables.

The elements are x_{ij}.

The matrix **X** has rank L where $L \leq \min [I,J]$

The two main variables in an experiment are the independent and dependent variable.
An independent variable is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable.
A dependent variable is the variable being tested and measured in a scientific experiment.

In our case, the dependent variable is the binary output of the DNA microarray experiment.

The independent variable could be the count of clonal complexes CC1, CC2, CC3 …. or the country of origin or the age of the individuals or whether they have diabetes or are co-infected by HIV.
The question would then be whether the presence/absence of genes that is detected by the microarray is a function of such independent variables.

PCA- preprocessing data entries

In general, the data table will be **preprocessed** before the analysis.

The columns of **X** are **centered** so that the **mean** of each column is equal to 0.

$$x_{ij} \rightarrow x_{ij} - \mu_j$$

If in addition, each element of **X** is divided by $\sqrt{I}$ or $\sqrt{I-1}$ (# of observations: I) the matrix Σ = $\mathbf{X}^T\mathbf{X}$ that we will later analyze is a covariance matrix,

$$\Sigma = \left[(\mathbf{X} - \mu)^T(\mathbf{X} - \mu) \quad \right]$$

and the analysis is referred to as **covariance PCA**.

If you don‘t center the data, the result would differ and its interpretation becomes more difficult.

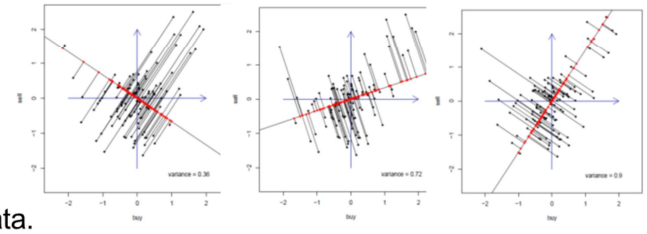
PCA- preprocessing data entries

In addition to centering, when the variables are measured with different units, it is customary to **standardize** each variable to **unit norm**.

This is obtained by dividing each variable by its norm (i.e. the square root of the sum of all squared elements of this variable) $\sqrt{\sum_i (x_i)^2}$, which is equivalent to dividing it by its standard deviation (except dividing by *n* vs *n*-1).

In this case, the analysis is referred to as a **correlation PCA** because, then, then matrix $\mathbf{X}^T\mathbf{X}$ is a correlation matrix.

One way of computing PC vectors is by **geometric construction** of the set of orthogonal vectors describing the largest variances in the data.



http://www.stefan-evert.de/PUB/Handout_LA_Trento_3.pdf

Processing of Biological Data – SS 2020 V 1 29

Standardizing or normalizing the data is important if one uses variables with different units
or variables which measure incomparable „things“.

Remember that the PCA algorithm tries to find PC vectors that capture the largest variance in the data.

Let us assume we measure the size of different cars in three dimensions: length, width and height.
If we used units of metre for length and width and units of centimetre for the height, the coordinates for the height would be a factor of 100 larger than the coordinates of length and width.
PC1 would definitely be oriented along the height axis because the PCA algorithm thinks that this is the axis where cars differ most.

PCA- preprocessing data entries

Another way of deriving a PCA uses the fact that the data matrix **X** has a **singular value decomposition (SVD)**
that decomposes a rectangular matrix **X** into three simple matrices:
two orthogonal matrices **P** and **Q** and one diagonal matrix Δ.

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

What is a SVD?

We will now review some basics from linear algebra that take us to the singular value decomposition.

Insert: review of eigenvalues

A vector **u** that satisfies $\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u}$

or $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{u} = 0$

is an **eigenvector** of this matrix **A**.

The scalar value λ is the **eigenvalue** associated with this eigenvector.

For example, the matrix $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$ has the eigenvectors

$u_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ with eigenvalue $\lambda_1 = 4$. Test: $\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix} \overset{?}{=} 4 \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

Test $2 \cdot 3 + 3 \cdot 2 = 4 \cdot 3$; $2 \cdot 3 + 1 \cdot 2 = 4 \cdot 2$

and

$u_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ with eigenvalue $\lambda_1 = -1$. Test: $\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} \overset{?}{=} -1 \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Test $2 \cdot (-1) + 3 \cdot 1 = (-1) \cdot (-1)$; $2 \cdot (-1) + 1 \cdot 1 = (-1) \cdot 1$

You probably know these things very well. No need to add any explanations here.

Insert: review of eigenvalues

For most applications we normalize the eigenvectors so that their length is equal to 1, i.e.

$$\mathbf{u}^T\mathbf{u} = 1$$

Traditionally, we put the set of eigenvectors of **A** in a matrix denoted by **U**.

Then, each column of **U** contains an eigenvector of **A**.

The eigenvalues are stored as diagonal elements of a diagonal matrix Λ .

Then we can write $\mathbf{A}\,\mathbf{U} = \mathbf{U}\,\Lambda$ or: $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{U}^{-1}$ (if we multiply with $\mathbf{U}^{-1}$)

This is the **eigendecomposition** of this matrix. Not all matrices have a EDC.

Only diagonalizable matrices can be factorized as an eigendecomposition.
We will leave the details to the mathematicians.

Insert: positive (semi-) definite matrices

A type of matrices used often in statistics are called **positive semi-definite** (PSD)

The eigen-decomposition of such matrices always exists, and has a particularly convenient form.

A matrix **A** is positive (semi-)definite, if there exists a real-valued matrix **X** and

$$\mathbf{A} = \mathbf{X}\,\mathbf{X}^T$$

Correlation matrices, covariance, and cross-product matrices are all semi-definite matrices.

The eigenvalues of PSD matrices are always positive or null.

The eigenvectors of PSD are pairwise orthogonal when their eigenvalues are different.

Also here, we will skip the mathematical details.

Insert: positive (semi-) definite matrices

This implies $\mathbf{U}^{-1} = \mathbf{U}^T$

Then we can express **A** as $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ with $\mathbf{U}^T\mathbf{U} = 1$

where **U** is the matrix storing the normalized eigenvectors.

E.g. $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1} = \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4\sqrt{\frac{1}{2}} & 4\sqrt{\frac{1}{2}} \\ 2\sqrt{\frac{1}{2}} & -2\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 2+1 & 2-1 \\ 2-1 & 2+1 \end{bmatrix}$$

with $\begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}+\frac{1}{2} & \frac{1}{2}-\frac{1}{2} \\ \frac{1}{2}-\frac{1}{2} & \frac{1}{2}+\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

showing that the 2 eigenvectors $\begin{bmatrix} \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{bmatrix}$ and $\begin{bmatrix} \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \end{bmatrix}$ are orthonormal.

This is a brief review of some linear algebra.

Singular Value Decomposition (SVD)

SVD is a generalization of the eigen-decomposition.

SVD decomposes a rectangular matrix **A** into three simple matrices:
two orthogonal matrices **P** and **Q** and one diagonal matrix Δ.

$$\mathbf{A} = \mathbf{P}\Delta\mathbf{Q}^T$$

P : contains the normalized eigenvectors of the matrix $\mathbf{A}\,\mathbf{A}^T$. (i.e. $\mathbf{P}^T\mathbf{P} = \mathbf{1}$)
The columns of **P** are called *left singular vectors* of **A**.

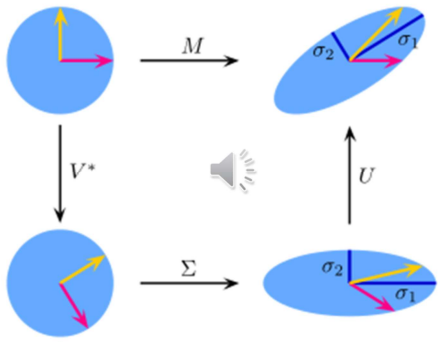
Q : contains the normalized eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$. (i.e. $\mathbf{Q}^T\mathbf{Q} = \mathbf{1}$)
The columns of **Q** are called *right singular vectors* of **A**.

Δ : the diagonal matrix of the *singular values.* They are the square root values of the eigenvalues of matrix $\mathbf{A}\,\mathbf{A}^T$ (they are the same as those of $\mathbf{A}^T\mathbf{A}$).

The rows of an orthogonal matrix are an orthonormal basis.
That is, each row has length one, and are mutually perpendicular.

Interpretation of SVD

In the special, yet common, case when **M** is an $m \times m$ real square matrix with positive determinant, **U**, **V***, and **Σ** are real $m \times m$ matrices as well. **Σ** can be regarded as a scaling matrix, and **U**, **V*** can be viewed as rotation matrices.



$$M = U \cdot \Sigma \cdot V^*$$

www.wikipedia.org

This is an intuitive illustration how the combined action of matrix M on the top-left data points can be decomposed as sequential application of the rotation matrix V*, a compression by the diagonal matrix sigma which scales the two coordinate axes, and of the second rotation matrix U.

Goals of PCA

(1) Extract the most important information from the data table

→ PC1 should describe the direction along which the data contains the largest variance;

PC2 is orthogonal to PC1 and describes the direction of the largest remaining variance etc

(1) Compress the size of the data set by keeping only this important information

(2) Simplify the description of the data set

(3) Analyze the structure of the observation and the variables.

In order to achieve these goals, PCA computes new variables called principal components (PCs) as linear combinations of the original variables.

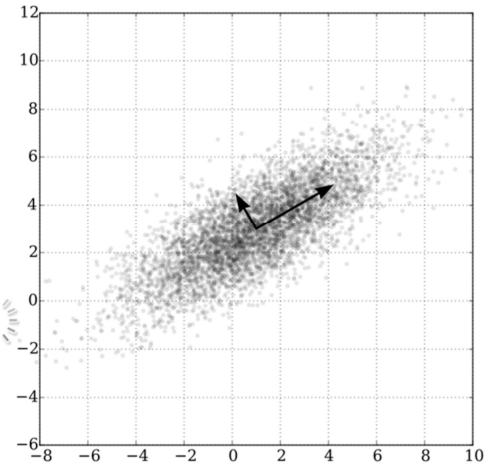
PC1 is the eigenvector of $\mathbf{X}^T\mathbf{X}$ with largest eigenvalue etc.

These are again the goals of PCA.

PCA example

PCA of a multivariate Gaussian distribution $\mathbf{X}$ centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction.

The two PCA vectors shown are the eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$ scaled by the square root of the corresponding eigenvalue, and shifted so that their tails are at the mean.



Note that shown here is the data along the original coordinates.
In a PCA plot, the data is projected onto two PCs, usually PC1 and PC2.

www.wikipedia.org

The PCA of this data set was likely performed on centered data.
The two PC vectors were then shifted back to the mean of the original data and rotated with respect to the original variables according to their loadings.

Deriving the components

The principal components are obtained from the SVD of **X**,

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

Q contains the principal components (normalized eigenvectors of $\mathbf{X}^T\mathbf{X}$).

The *I* x *L* matrix of **factor scores**, denoted **F**, is obtained as

$$\mathbf{F} = \mathbf{P}\Delta = \mathbf{P}\Delta\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q}$$

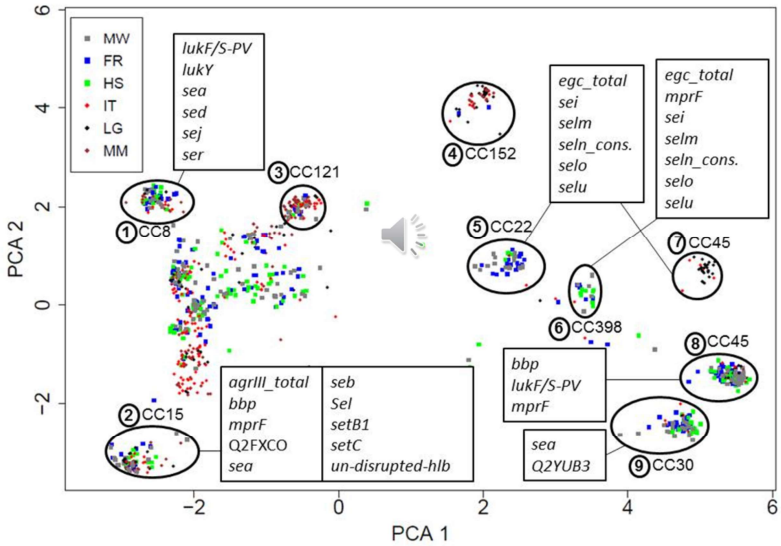
Thus, **F** can be interpreted as a **projection matrix**
because multiplying **X** with **Q** gives the values
of the projections of the observations **X** on the principal components **Q**.

The results of a PCA are usually discussed in terms of *component scores* (or *factor scores)* and *loadings.*
The *factor scores* are the projections of the original data points onto the principal component vectors.

The so-called *loadings* describe the relationship between the PCs and the original variables.

PCA of MA hybridization data (again)

PCA identifies local clusters that are characteristic for particular clonal complexes



Projection (factor score) of data points on PC1

This is the same plot that was shown earlier.

Shown are projections (factor scores) of the original data points onto the two first PC vectors.

Summary

What we have covered **today**:

- Detection of DNA probes by DNA microarray
- Euclidian distance of 1/0 signals as distance measure
- Clustering of MA data
- PCA analysis of MA data

Next lecture:
- Reconstruct missing (ambiguous) data values with BEclear

Today‘s lecture was a typical mix of looking into a real-life task from a past research project

and an introduction of some helpful mathematical techniques that we used in this project.

Often your experimental collaborators will completely depend on you.

They will tell you „You are the bioinformatician. You know what needs to be done.“

In some cases, you may actually know what to do.

In the other cases, you need to refresh your math or to pick up new skills.

In such a collaboration, the data analysis part is your job!

Luckily you are usually not the first one to solve such a problem.

So you should read a lot and talk to other people how they have solved the same problem before.