Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Trends in Bioinformatics

M. Kraus, H. F. da Cruz, M. Neves

Seminar Kick-Off

Oct 19, 2016

Agenda

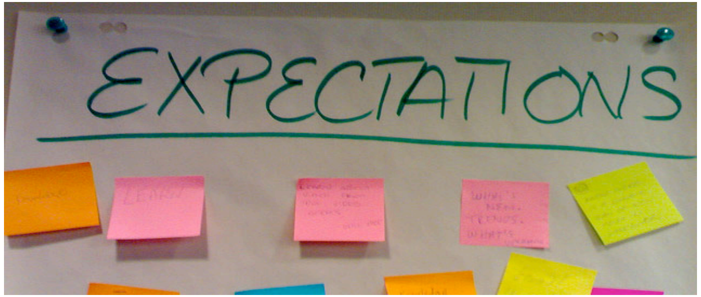
- Seminar Organization
- Seminar Topics

Seminar Organization
Setup

- Supervisors: Milena Kraus, Harry Freitas da Cruz, Dr. Mariana Neves, Dr. Matthias Uflacker
- ~~Location: HPI Campus II, Room D.E-9/10 (former SNB), Tuesdays 9:15-10:45 a.m. (s.t.)~~ → individual appointments with your supervisor
- Periods: 2 SWS (3 graded ECTS)
- Enrollment:
  - Prioritized topic wish list via e-mail to milena.kraus@hpi.de
  - Due Thu Oct 27, 2016
  - Sign up for the course until Fri Oct 28, 2016
- https://hpi.de/plattner/teaching/winter-term-201617/trends-in-bioinformatics.html

Seminar Organization
What you can expect from us

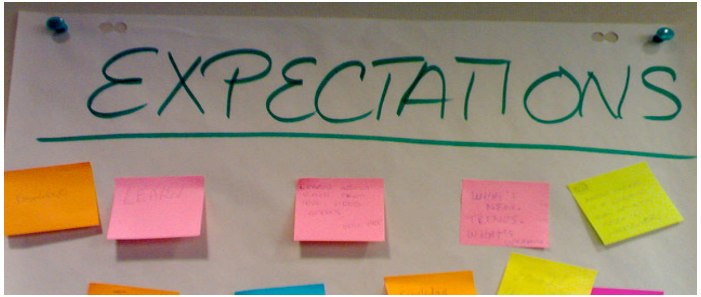
- Broaden your horizon in the fields of
  - Bioinformatics,
  - Life sciences, and
  - Your selected seminar topic
- Get in touch and work with real-world data
- Get experienced in collaborative project work
- Enhance your skills in English presentation, scientific working, and writing



http://i.kinja-img.com/gawker-media/image/upload/s--cREIB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
What we expect from you

- Commitment on your selected seminar topic
- Perform autonomously research to acquire knowledge about your selected seminar topic
- Hands-on experiments of selected tools on benchmarking
- Participate in every seminar meeting
- Contribute with your expertise also to your colleagues / other teams
- Update supervisors regularly on your progress / issues



http://i.kinja-img.com/gawker-media/image/upload/s--cREIB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
Grading

- The grading of the seminar works as follows (aka "Leistungserfassungsprozess"):
  - 40% seminar presentation and abstract
  - 40% scientific research article
  - 20% individual commitment

- **All individual parts have to be passed** to pass the complete seminar

http://www.hpi.uni-potsdam.de/fileadmin/hpi/presse/Fotos/campus_und_gebaeude/20111017_HPI_Hoersaal.jpg

Submission of a paper (optional)

- Publication of the article at
  - □ A journal
    - – Briefings in Bioinformatics
  - □ Workshops and Conferences
    - – Poster in the BioCuration’17

ISB International Society for Biocuration

Next Steps
Enrollment for Seminar Topics

How to apply for a topic?

- Send prioritized list of top 3 topics to Milena Kraus (milena.kraus@hpi.de)
  - 1st choice: …
  - 2nd choice: …
  - 3rd choice: …
- Deadline: **Thu Oct 27, 2016 12pm (noon)**
- HPI Deadline: **Fri Oct 28, 2016**

Schedule

- Final presentation
  - One session per person/team
  - 1h30min, at least 30 minutes presentation
  - Dates tbd
  - One-page abstract one week prior to the presentation
- Introduction to scientific writing
  - End of lecture time
- Scientific report
  - End of semester
- Excursions (optional)
- OpenHPI course “Code of Life” (starting in Nov, optional)

Excursions (optional)

- Max-Delbrück-Center in Berlin-Buch
- Max-Planck-Institute in Berlin-Dahlem
- Fraunhofer Institute in Potsdam-Golm
- Gläsernes Labor in Berlin-Buch
  - □ Hand-on wet lab session, e.g. genetic fingerprint
  - □ Costs: 12 € p.P. will be provided by the HPI, if at least 5 students sign up
  - □ Please sign up in this doodle for your preferred date (http://doodle.com/poll/kxznayv2syciq8qr)
- All other excursions will be free of charge and will be organized as soon as we get positive feedback from you.

Seminar Topics

A. Inferring Genetic Variants from RNAseq Data

B. Explorative Analysis of RNAseq Data

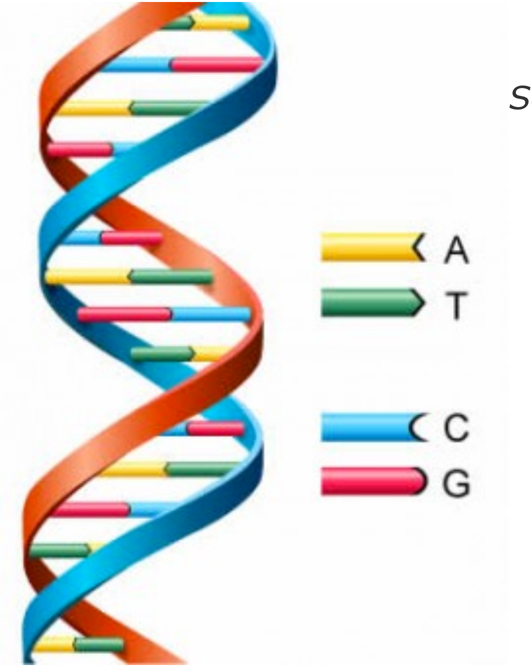
C. Automatic Summaries for Cancer Research

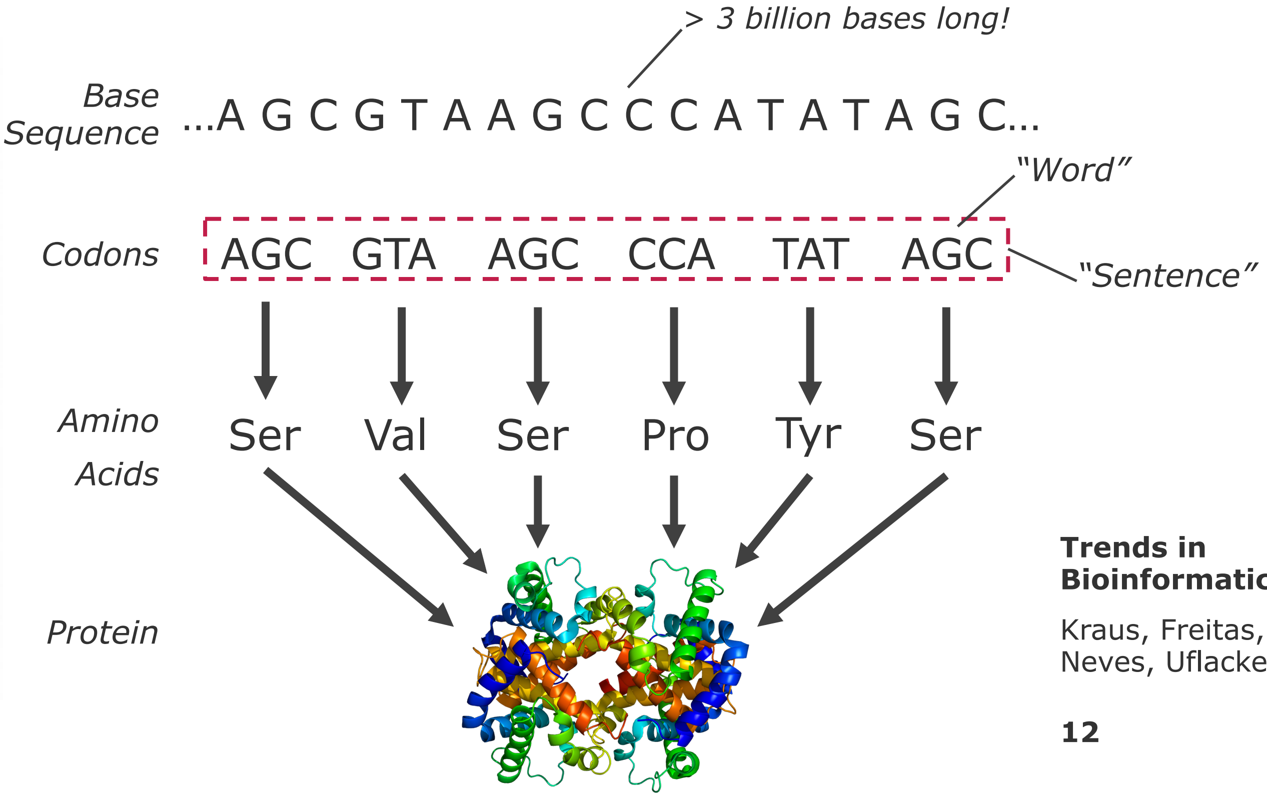
D. Document Retrieval to Support Clinical Decision

E. Information Extraction for Data Curation

F. Prediction of Dialysis Length

Crash Course:
The Human Genome





> 3 billion bases long!

Base Sequence ...A G C G T A A G C C C A T A T A G C...

"Word"

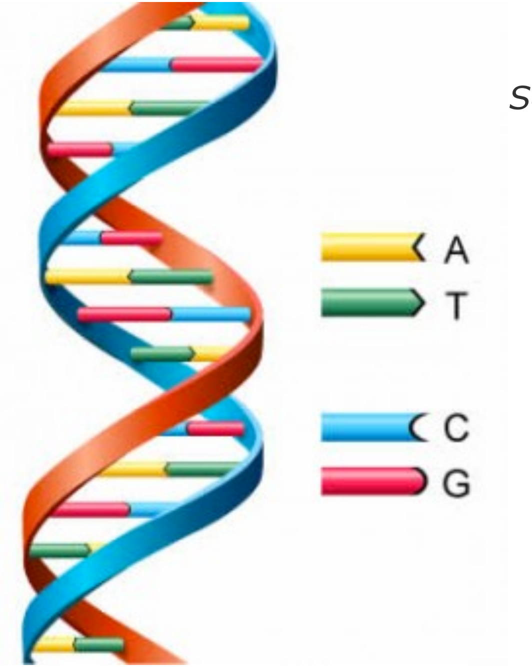
Codons AGC GTA AGC CCA TAT AGC

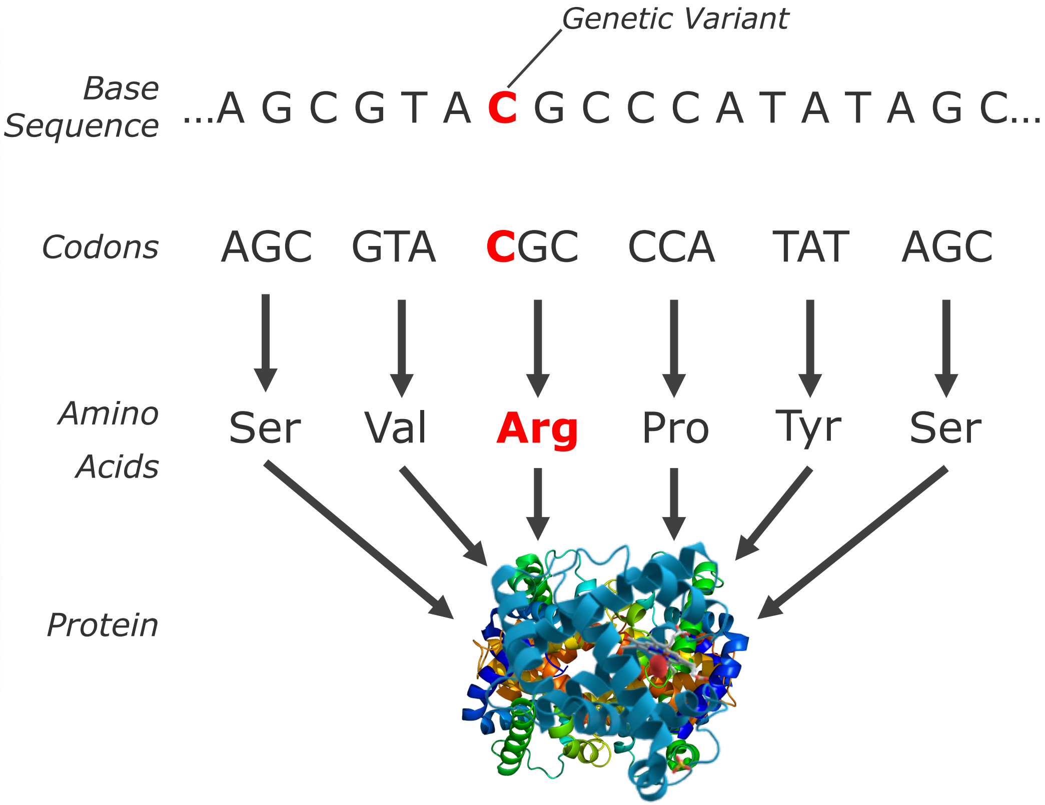
"Sentence"

Amino Acids Ser Val Ser Pro Tyr Ser

Protein

Crash Course:
The Human Genome





Genetic Variant

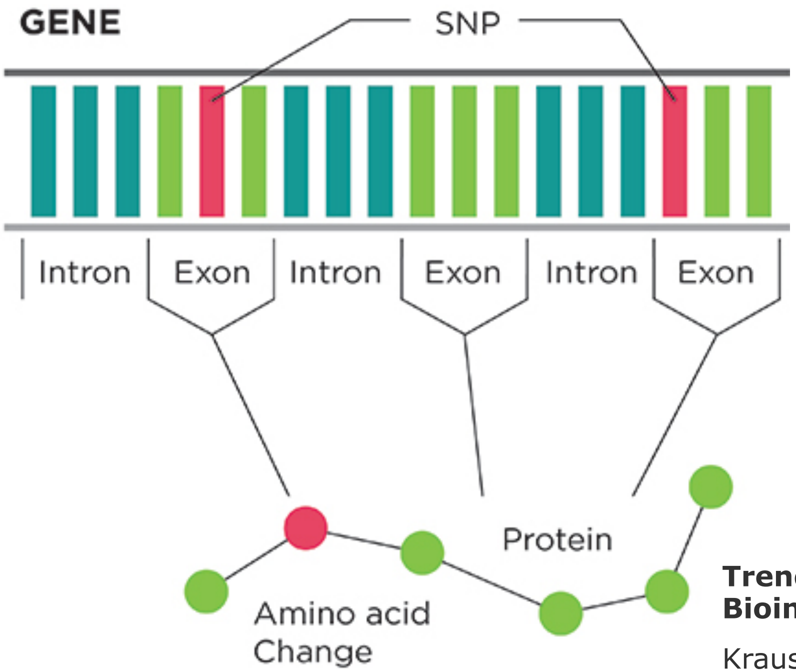
Base Sequence ...A G C G T A **C** G C C C A T A T A G C...

Codons	AGC	GTA	**C**GC	CCA	TAT	AGC
Amino Acids	Ser	Val	**Arg**	Pro	Tyr	Ser

Protein

Crash Course:
Genome vs. Exome

- Genome provides more information
- Exome is cheaper to sequence
- Both contain redundant information



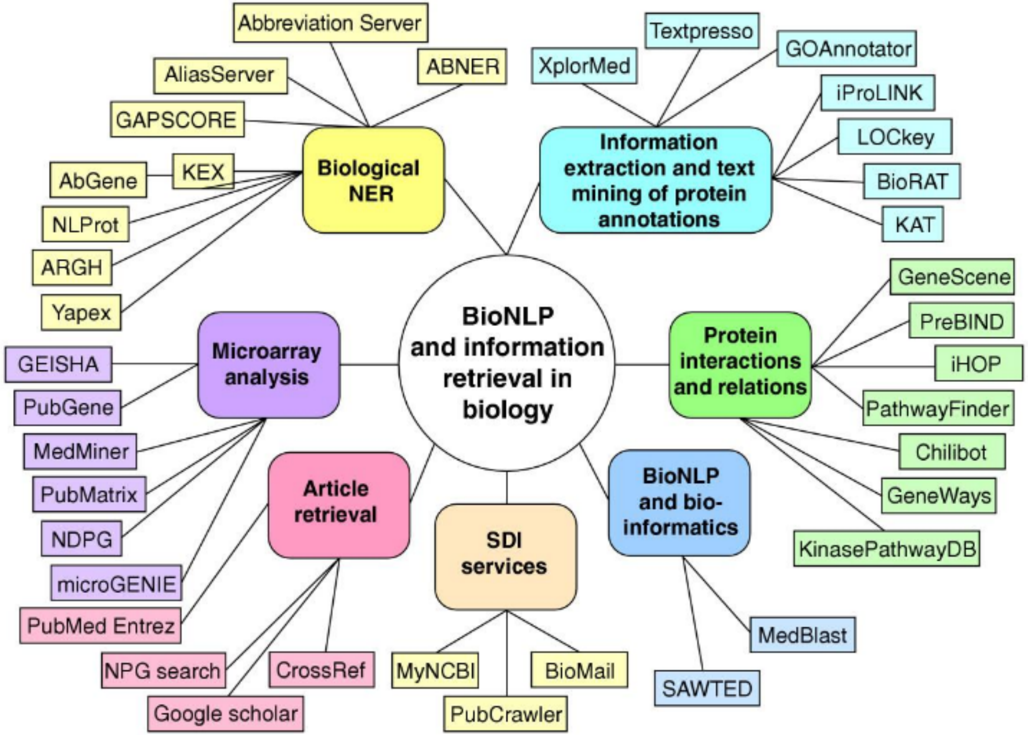
A. Inferring genetic variants from RNAseq data

- Understand:
  - Difference between exome & genome
  - How variant calling works
  - Examine the data artifacts produced, e.g. FASTQ, SAM/BAM, VCF
- Try out:
  - Find already known variants related to heart failure
  - Run two or more variant calling algorithms on a provided set of RNAseq data
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss benefits and drawbacks of the approach

B. Explorative analysis of RNAseq data

- Understand:
  - General characteristics of RNAseq expression data
  - Unsupervised machine learning algorithms
  - Variability of processing results
- Try out:
  - Two or more unsupervised ML algorithms to examine RNA expressions
  - Compare own results with published results
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss benefits and drawbacks of the approach

Natural Language Processing for Biomedicine



(https://genomebiology.biomedcentral.com/articles/10.1186/gb-2005-6-7-224)

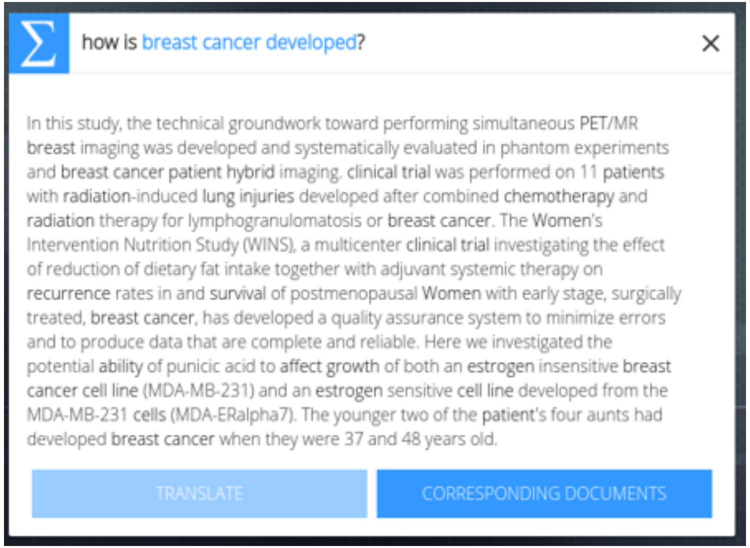
C: Automatic Summaries for Cancer Research

Issue:

Automatically-generated summaries are useful for an overview on a topic or for summarizing various publications.

Idea:

- Research the current solutions for automatic summarization.
- Evaluation on summaries for cancer research.
- Project in collaboration with the DKG.



how is breast cancer developed?

In this study, the technical groundwork toward performing simultaneous PET/MR breast imaging was developed and systematically evaluated in phantom experiments and breast cancer patient hybrid imaging. clinical trial was performed on 11 patients with radiation-induced lung injuries developed after combined chemotherapy and radiation therapy for lymphogranulomatosis or breast cancer. The Women's Intervention Nutrition Study (WINS), a multicenter clinical trial investigating the effect of reduction of dietary fat intake together with adjuvant systemic therapy on recurrence rates in and survival of postmenopausal Women with early stage, surgically treated, breast cancer, has developed a quality assurance system to minimize errors and to produce data that are complete and reliable. Here we investigated the potential ability of punicic acid to affect growth of both an estrogen insensitive breast cancer cell line (MDA-MB-231) and an estrogen sensitive cell line developed from the MDA-MB-231 cells (MDA-ERalpha7). The younger two of the patient's four aunts had developed breast cancer when they were 37 and 48 years old.

DKG KREBSGESELLSCHAFT

D: Document Retrieval to Support Clinical Decision

Issue:

Physicians frequently need to screen many publications to search for answers for clinical cases.

What is the patient's diagnosis?

What tests should the patient receive?

How should the patient be treated?

Idea:

- Research systems for document retrieval.
- Retrieve relevant publications to support answering these questions.
- Evaluation on the TREC'2016 benchmark.

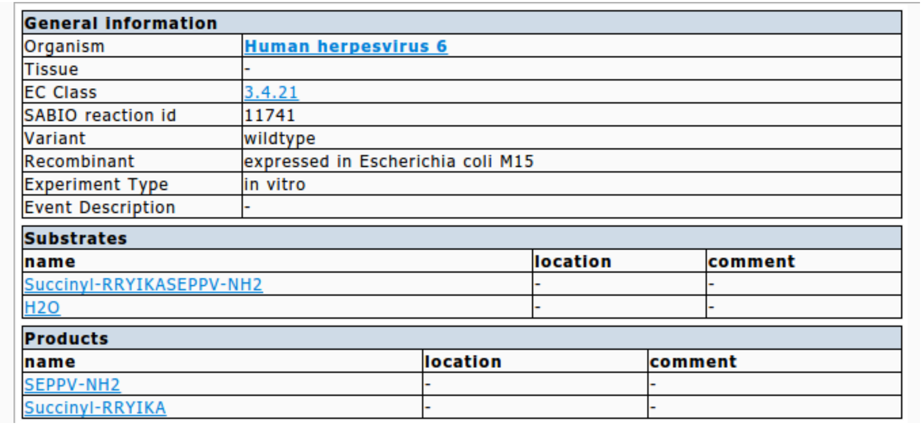
E: Information Extraction for Data Curation

Issue:

Researchers need to screen many documents to extract specific data to feed biological databases.

Idea:

- Research systems for information extraction of biomedical data.
- Evaluation on the existing (curated) data.
- Project in collaboration with SABIO-RK database.



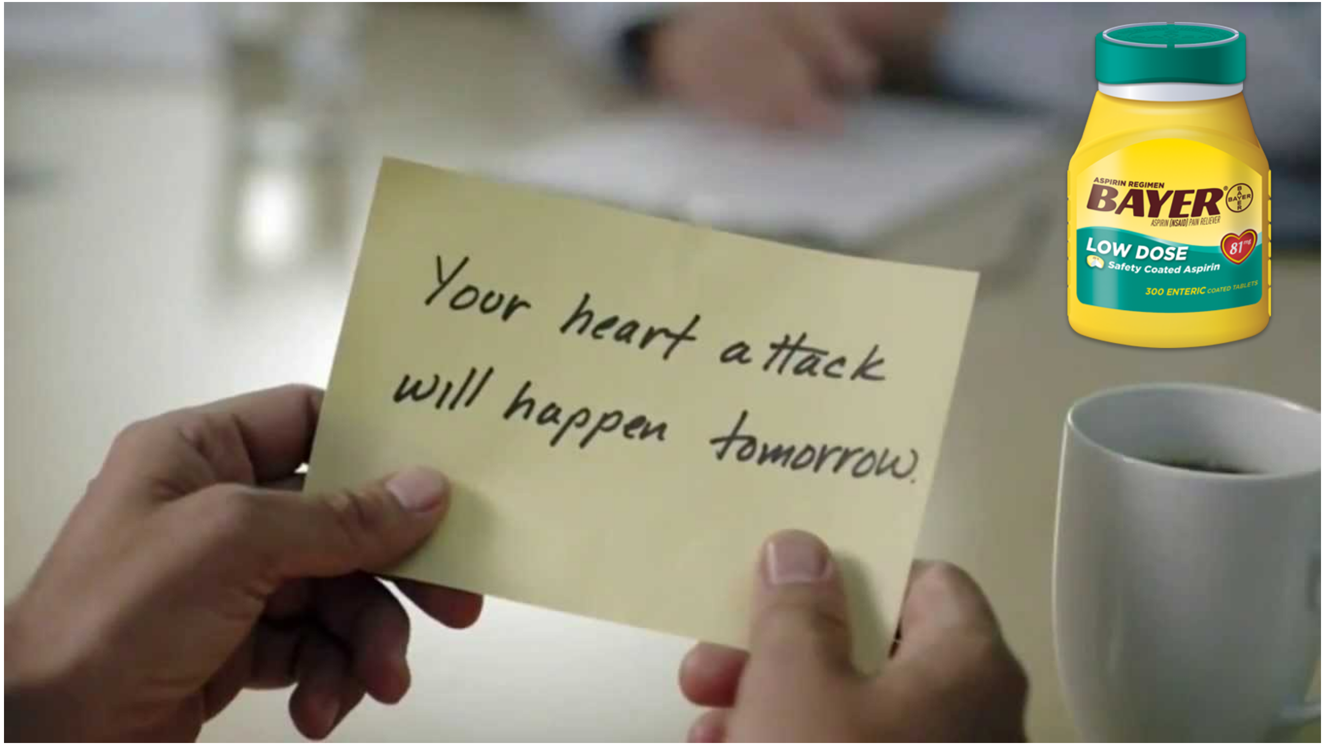
General information	
Organism	Human herpesvirus 6
Tissue	-
EC Class	3.4.21
SABIO reaction id	11741
Variant	wildtype
Recombinant	expressed in Escherichia coli M15
Experiment Type	in vitro
Event Description	-

Substrates		
name	location	comment
Succinyl-RRYIKASEPPV-NH2	-	-
H2O	-	-

Products		
name	location	comment
SEPPV-NH2	-	-
Succinyl-RRYIKA	-	-

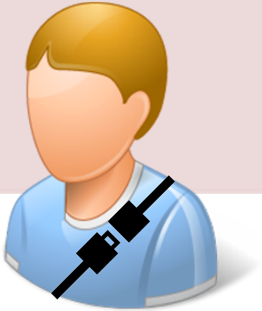
SABIO-RK
Biochemical Reaction Kinetics Database

Patient-level predictive analytics
What is it?


Your heart attack will happen tomorrow.
ASPIRIN REGIMEN
BAYER
ASPIRIN (NSAID) PAIN RELIEVER
LOW DOSE
81 mg
Safety Coated Aspirin
300 ENTERIC COATED TABLETS

Patient-level predictive analytics
Applications

Patient safety



Precision medicine

Patient outcomes

F: Prediction of Dialysis Length

- **Dialysis in Germany**[1]
  - □ 70,000 patients / 2.5 Mio. EUR p.a.
  - □ 100,000 patients by 2020
  - □ High risk of mortality / high costs
- **Tasks:**
  - □ Predict length of dialysis using:
    - – Support Vector Machines (SVM)
    - – Logistic Regression (LR)
  - □ Dissect the SVM and LR algorithms
  - □ Write up a research paper

Source: Anna Frodesiak, CC0

[1] http://www.aerzteblatt.de/nachrichten/41258/Zahl-der-Dialysepatienten-steigt

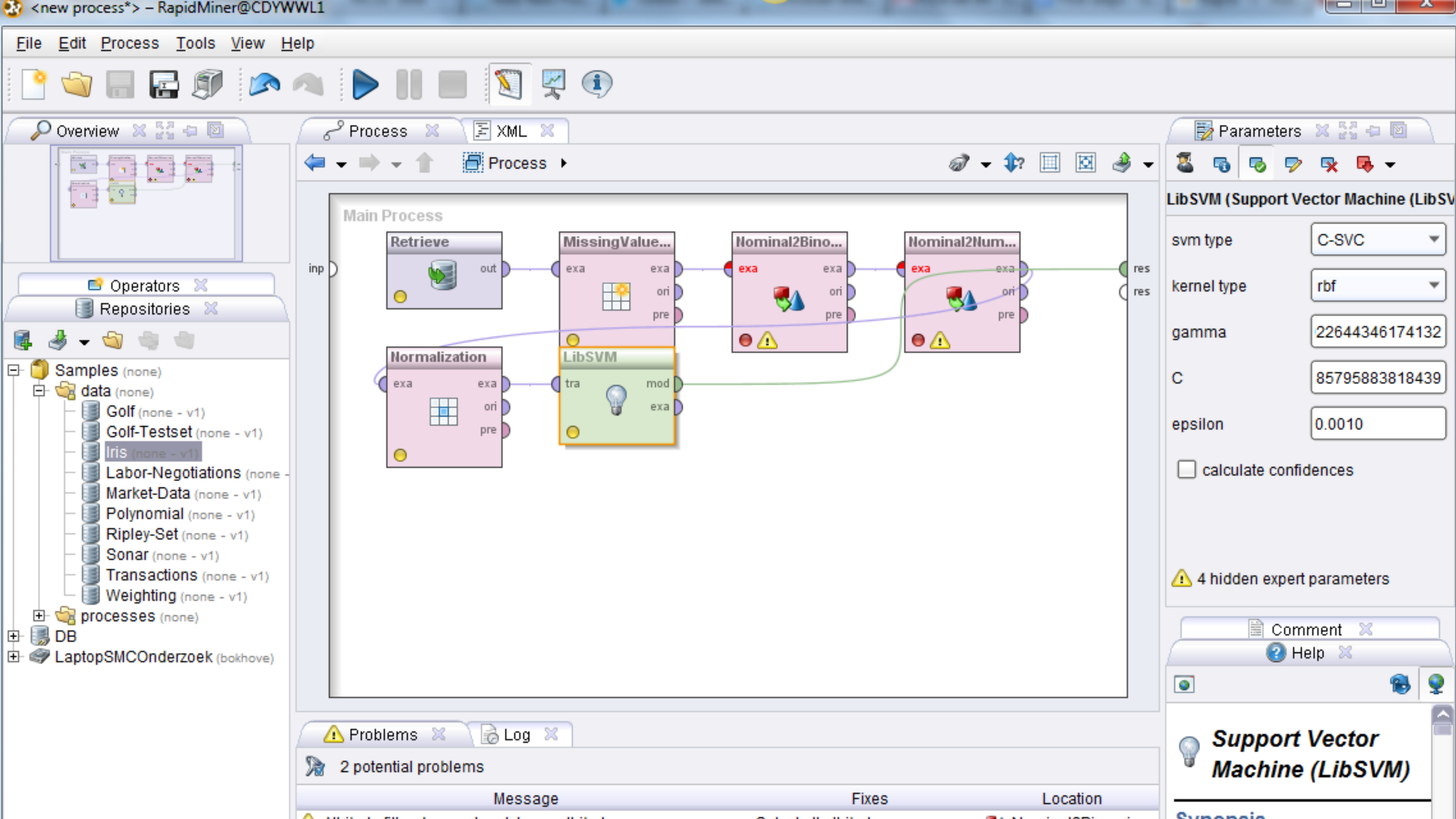
F: Prediction of dialysis length (cont.)

- **You will learn:**
  - In and outs of SVM and LR
  - How to set up a ML experiment
  - Basic data structures in a Hospital Information System
- **Data source:**
  - MIMIC (Multiparameter Intelligent Monitoring in Intensive Care)
  - Intensive Care Unit Data from MIT
- **Tool:**
  - RapidMiner
  - Visual modelling

Source: Anna Frodesiak, CC0

Trends in Bioinformatics

Kraus, Freitas, Neves, Uflacker



LibSVM (Support Vector Machine (LibSV

svm type: C-SVC

kernel type: rbf

gamma: 22644346174132

C: 85795883818439

epsilon: 0.0010

□ calculate confidences

⚠ 4 hidden expert parameters

2 potential problems

Support Vector Machine (LibSVM)

Thank you for your attention!

ToDos:

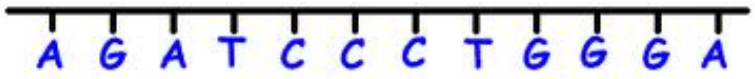
- Please sign up in the doodle asap (http://doodle.com/poll/kxznayv2syciq8qr)!
- Choose your favorite topics and tell us about it!

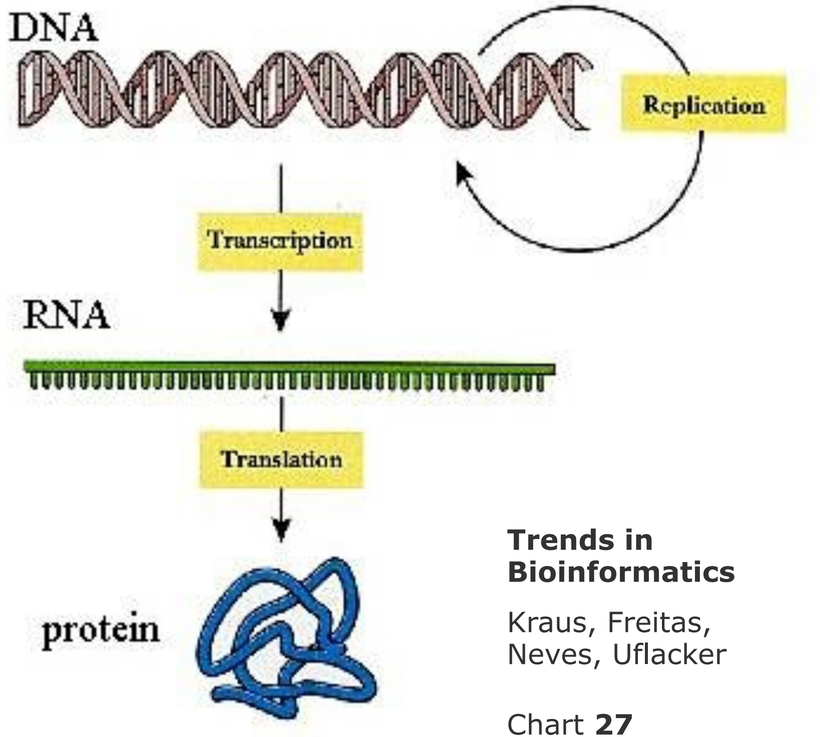
What is gene expression?

- Gene expression = synthesis of a protein with the help of genetic information

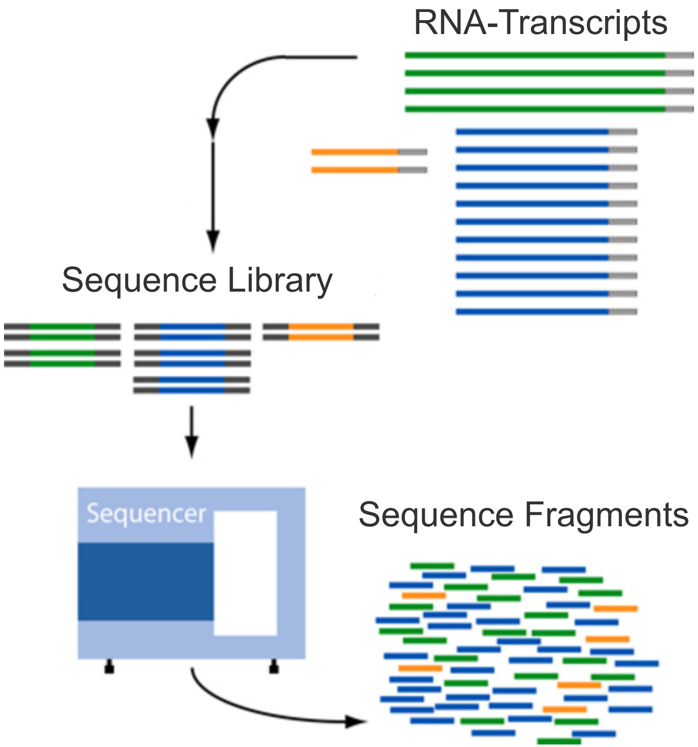
Most important facts for your task:

- A cell of a failing heart expresses other genes than a healthy heart cell → expression profile
- The number of found RNAs of one gene gives you the quantity of the corresponding protein
- RNA consists of the letters A, T(U), C, and G





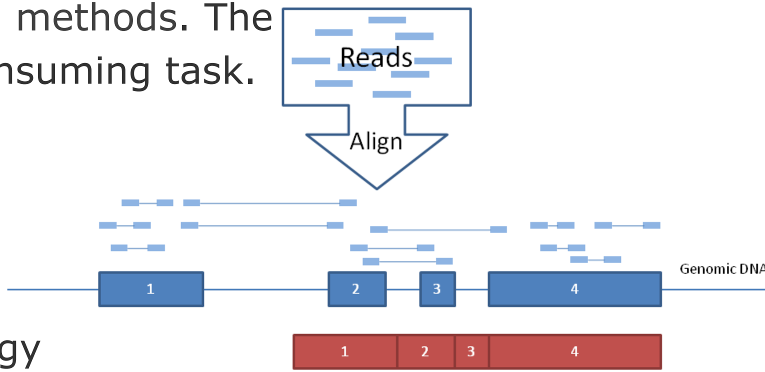
Creating the transcriptome out of raw experimental sequencing data



- RNA transcripts are broken into smaller (puzzle) pieces of short sequence reads
- Reads need to be “sorted” and aligned to a reference genome
- Aligned Reads are counted to give the respective RNA quantity
- Differences between conditions (ill, healthy) are computed through statistical methods and visualized accordingly

A: Processing of large RNAseq data sets to elucidate causes of heart failure

Issue: The transcriptome of a patient provides rich information for the elucidation of causes of heart failure. It needs to be build from raw RNAseq data with computationally and algorithmically challenging methods. The recreation of the transcriptome is a complex and time-consuming task.



Idea:

- Familiarize with processing of RNAseq data
- Evaluate means of optimization through IMDB technology
- Implement different processing pipelines in an IMDB
- Benchmark and evaluate your pipeline(s) with real patient data and compare to existing solutions

B: Statistical analysis of the transcriptome and differentially expressed genes

Issue: Methods for statistical analysis and visual exploration of RNAseq processing pipeline outputs exist and need to be implemented in our system. Inherent capabilities of the IMDB (PAL, Lumira) and R can be used to meet the requirements of our partner researchers.

Idea:

- Familiarize with the output RNAseq preprocessing
- Explore possibilities of statistical analysis in our IMDB and in R and also the IMDB-Rserv interface in particular
- Explore visualization capabilities of Lumira and R
- Choose and implement the best options for statistical analysis

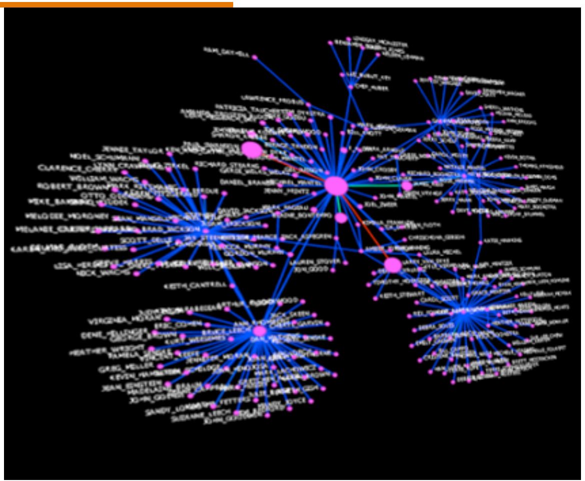
C: Integration and Harmonization of Medical Data

Issue:

Clinical data is acquired in heterogeneous data formats in distributed data silos. Combining existing data sets for analysis is a manual task, which prevents efficient exploration of existing knowledge.

Idea:

- Explore existing data silos
- Define an integrated database model for harmonization
- Use existing analysis tools to test analysis capabilities of your data model
- Work in interdisciplinary teams with our cooperation partner



http://www.programmableweb.com/wp-content/FirstGiving_2.jpg