

# Enterprise Platform and Integration Concepts

Winter Term 2015/2016



Trends in Bioinformatics: Kick-off meeting

*Cindy Perscheid, Mariana Neves, Milena Kraus, Harry Cruz*

*October 20th, 2015*

# Bioinformatics



(<https://nsaunders.wordpress.com/2008/08/23/what-if-journal-current-contents-were-tag-clouds/>)

# Overview

- Seminar organization
- Seminar topics

# Overview

- Seminar organization
- Seminar topics

## Seminar Organization Setup

- Supervisors:
  - Cindy Perscheid,
  - Milena Kraus
  - Dr. Mariana Neves,
  - Harry Cruz
- Location: HPI Campus II, Room D.E-9/10 (former SNB)
- When: Tuesdays 9:15-10:45 a.m. (s.t.)
- Periods: 2 SWS (3 graded ECTS)
- Enrollment: Due Fri October 23, 2015 (HPI deadline)
- <http://hpi.de/plattner/teaching/winter-term-201516/trends-in-bioinformatics.html>

## What you can expect from us

- Broaden your horizon in the fields of
  - Bioinformatics,
  - Life sciences, and
  - Your selected seminar topic
- Enhance your skills in English presentation, scientific working, and writing



## What we expect from you

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

# Grading

- The grading of the seminar works as follows (aka “Leistungserfassungsprozess”):
  - 40% Seminar presentation
  - 40% Research article
  - 20% individual commitment
- **All individual parts have to be passed** to pass the seminar



## Submission of a paper (optional)

- Publication of the survey
  - Journal
    - Briefings in Bioinformatics
  - Workshops and Conferences
    - Poster in the BioCuration'16



## Seminar Trends in Bioinformatics

- Mail with **3 top choices** to „cindy.perscheid@hpi.de“ until **Thursday, October 22nd at 11:59 PM.**
  - 1 (very high adherence): ...
  - 2 (high adherence): ...
  - 3 (medium adherence): ...
  
- Receive your topic assignment until **Friday, October 23rd at 12:00 (noon).**

# Schedule

- Final presentation
  - One session per person/team
  - 1h30min, at least 30 minutes presentation
  - Dates to be decided
  - One-page abstract one week prior the presentation
- Introduction on scientific writing (Cindy)
  - End of the lecture time
- Scientific report
  - End of the semester

# Overview



- Seminar organization
- Seminar topics

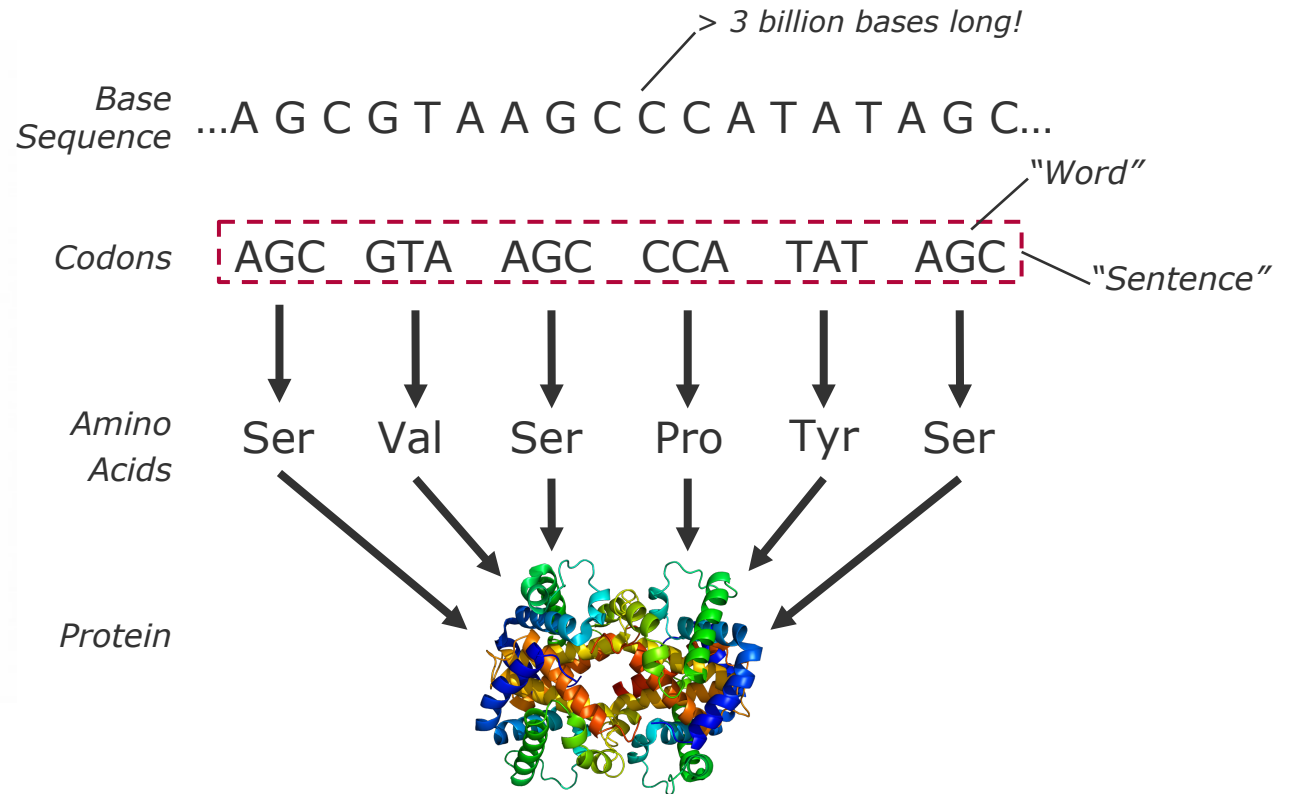
# Topics

1. Preprocessing of Raw Genome Data
2. Gene Prediction in Genome Data
3. Pattern Detection in Microarrays
4. *Ab initio* Protein Structure Prediction
5. Question Answering
6. Text Mining Tools for Biocuration
7. Clinical Decision Support Systems

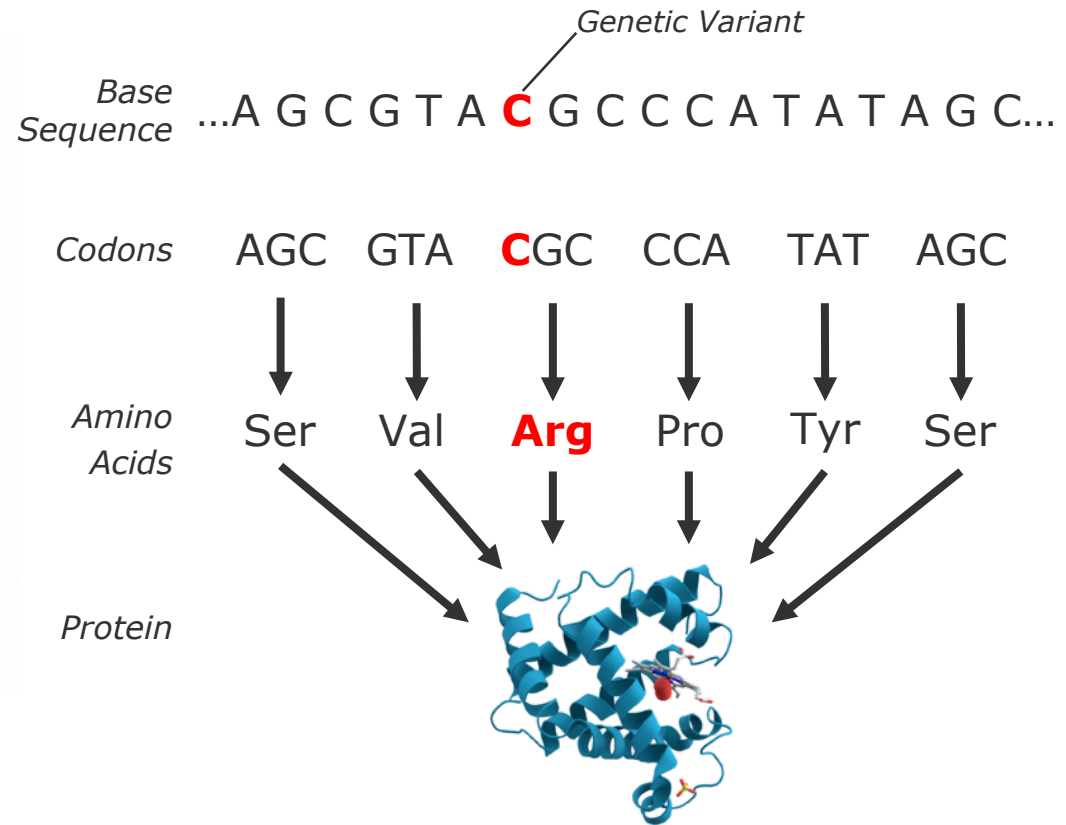
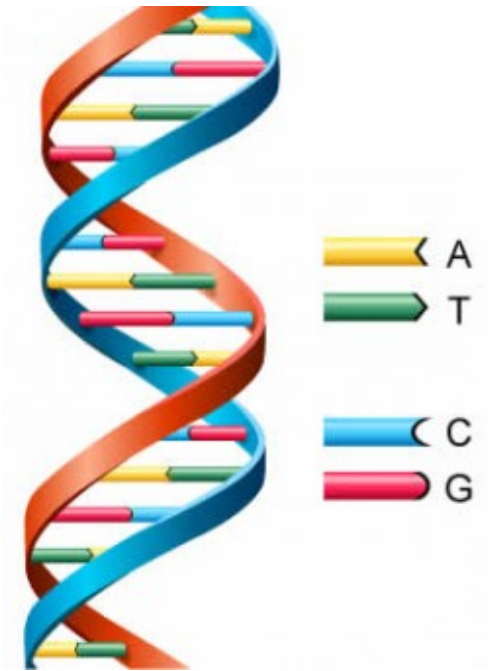
# Crashcourse: The Human Genome



 A  
 T  
 C  
 G

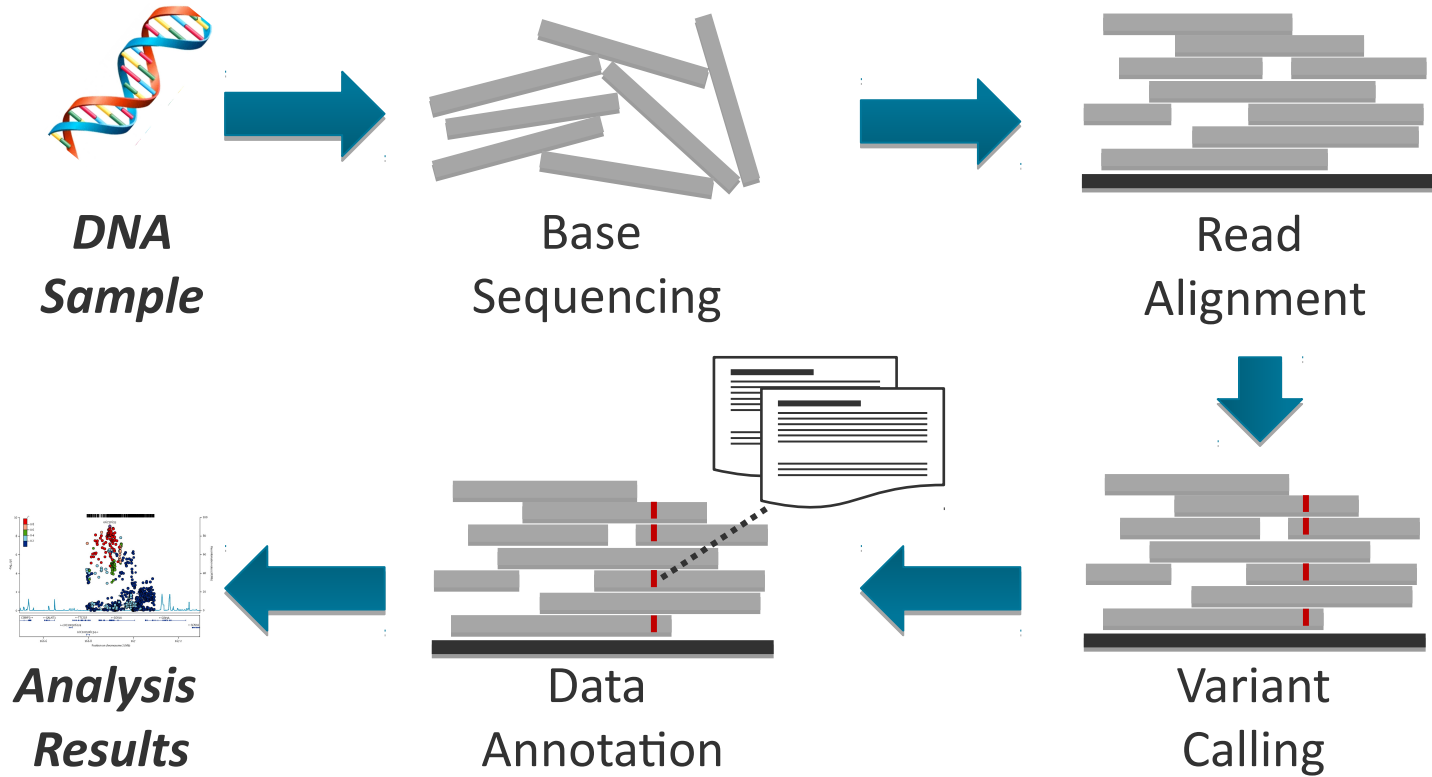


# Crashcourse: The Human Genome



# Topic 1: Preprocessing of Raw Genome Data

## Genome Data Analysis Process





# Topic 1: Preprocessing of Raw Genome Data

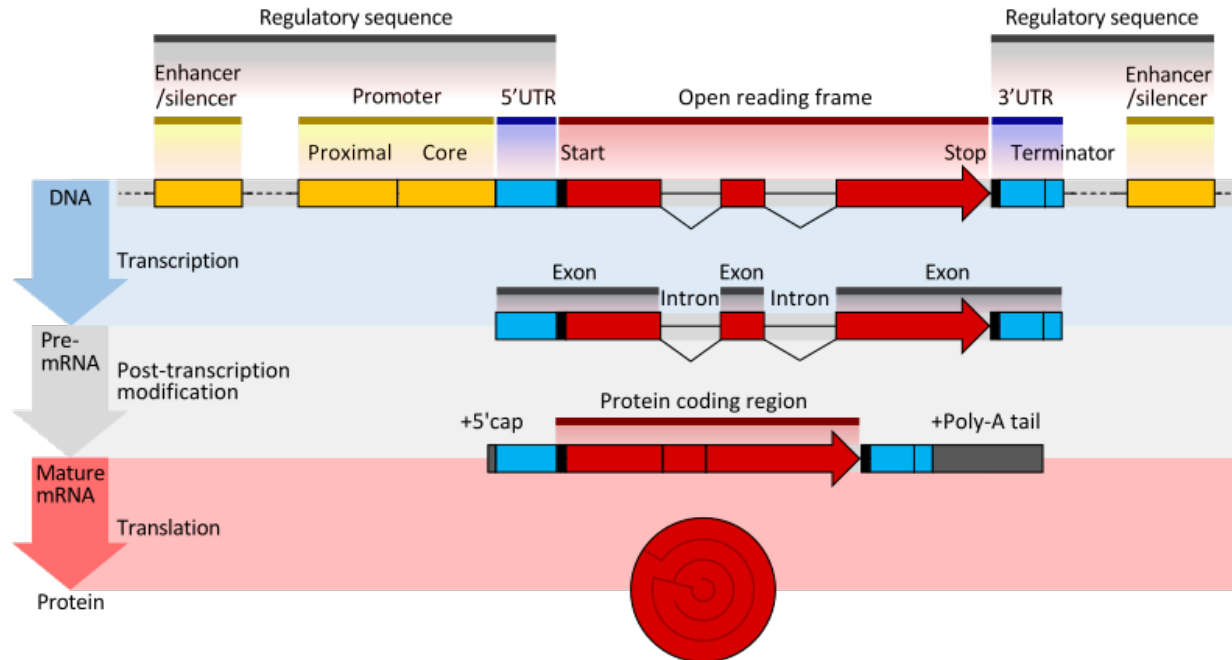
## Your task

- Understand:
  - Research on the distinct steps of one particular pipeline type for Genome Data Analysis, e.g. for single reads
  - Examine the data artifacts produced, e.g. FASTQ, SAM/BAM, VCF
- Specialize: Become an expert on one of the algorithms that can be applied, e.g. Bayesian or Haplotype Variant Callers
- Try out: Run evaluation experiments with selected tools
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss benefits and drawbacks of the approach

# Topic 2: Gene Prediction in Genome Data

## Protein Coding Regions

- Not all protein coding regions (= genes) identified yet in human genome
- Machine learning to get a clue where genes are to find



## Topic 2: Gene Prediction in Genome Data

### Your task

- Understand:
  - Find out the rough mechanisms of producing proteins
  - Research on Machine Learning approaches for identifying gene regions
  - Examine the data artifacts produced
- Specialize: Become an expert on one of the algorithms that can be applied, e.g. signal or content sensors, combined approaches
- Try out: Run evaluation experiments with selected tools
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss benefits and drawbacks of the approach

# Topic 3: Pattern Detection in Microarrays

## Gene Expression

- Gene expression: Cell process where the protein is built from gene information in DNA
  - Transcription
  - Translation
- Protein can be expressed multiple times = Expression level of a gene

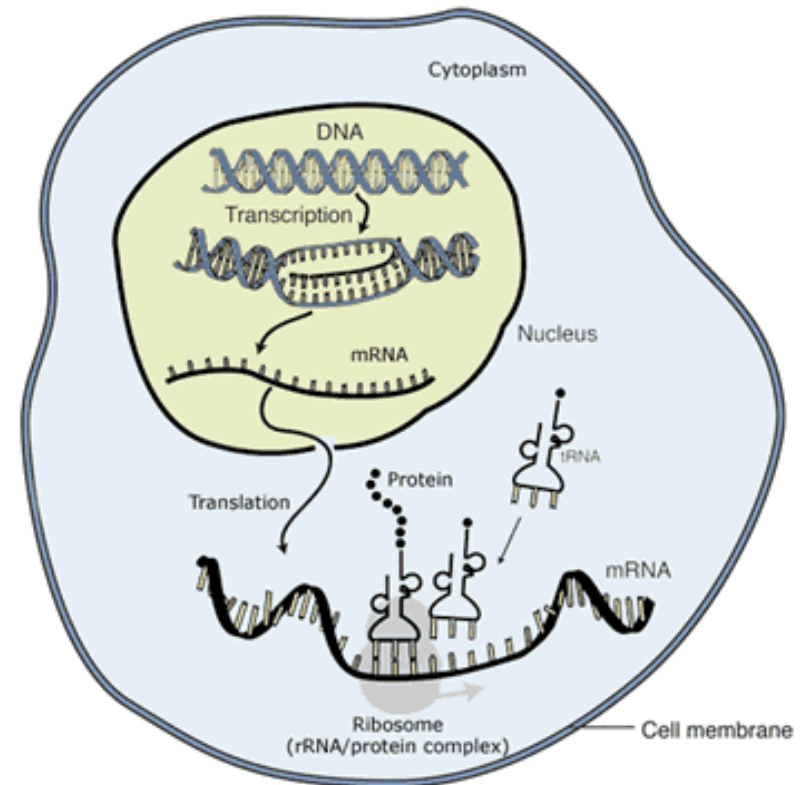
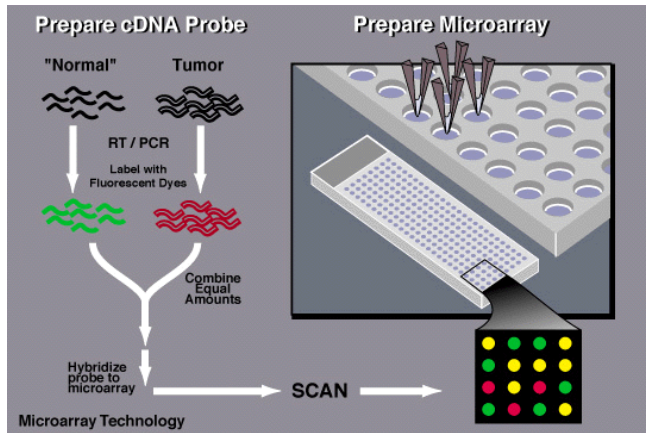


Image adapted from: National Human Genome Research Institute.

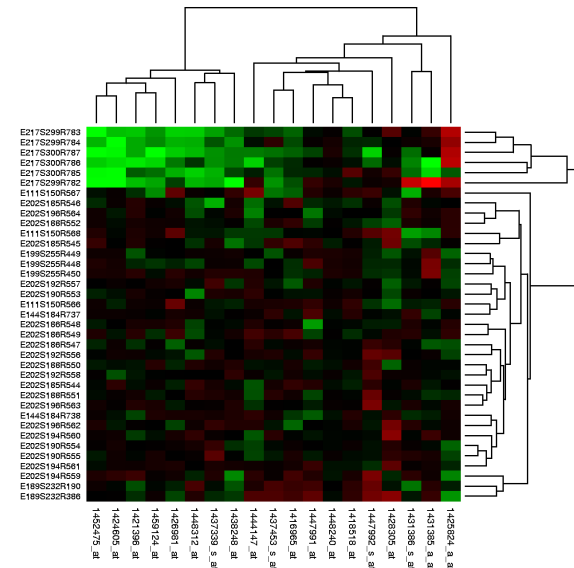
# Topic 3: Pattern Detection in Microarrays

## Microarray Analysis

- Microarrays: Measure expression levels of genes simultaneously for many samples
- Idea: Similar expression levels mean similar cell processes
- Computational analysis necessary to identify patterns in data



**Machine Learning Techniques**

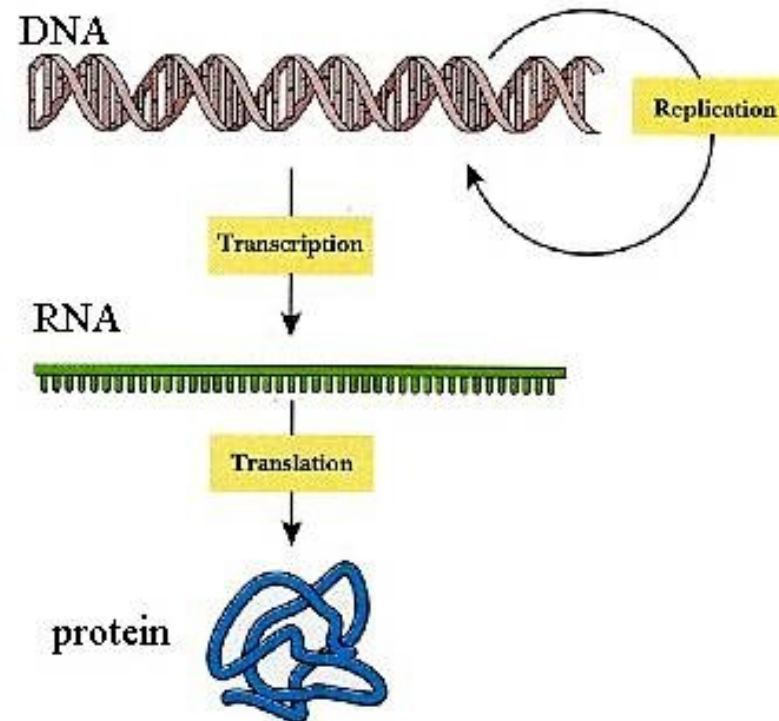


## Topic 3: Pattern Detection in Microarrays

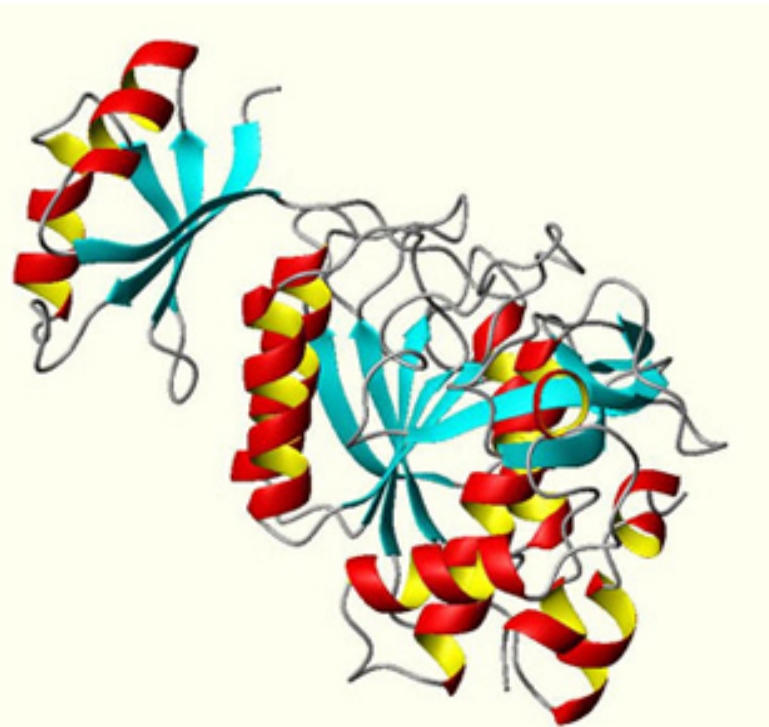
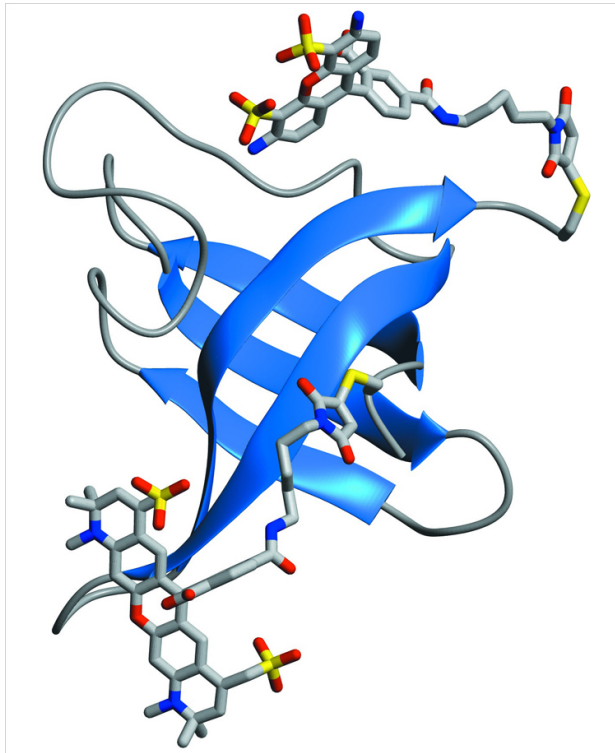
### Your task

- Understand:
  - Find out the rough mechanisms of gene expression and Microarrays
  - Research on the current ML procedures for pattern matching
  - Examine the data artifacts produced, e.g. from Microarray experiments
- Specialize: Become an expert on one aspect of the analysis; e.g. Feature Selection, Hierarchical Clustering, Validation
- Try out: Run evaluation experiments with selected tools
- Write:
  - Describe your algorithm and experiments in a scientific paper
  - Discuss benefits and drawbacks of the approach

## Topic 4: *Ab initio* protein structure prediction



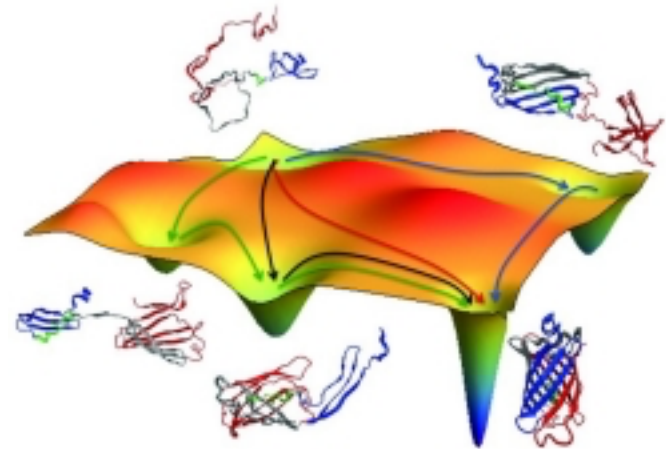
## Topic 4: *Ab initio* protein structure prediction





## Topic 4: *Ab initio* protein structure prediction

- A protein, consisting of 100 amino acids could fold into 3198 different 3D structures.
- Take 1 picosecond for every calculation of a structure and spend about  $9 * 10^{74}$  years to find the right one.
- Folding of a protein in the cell takes at max. 1 millisecond!



(age of the universe:  $13 * 10^9$  years)

## Topic 4: *Ab initio* protein structure prediction

- Computational approaches to predict 3D protein structures use multiple heuristics to be faster than  $9 * 10^{74}$  years.
- Different approaches: Hidden Markov Models, Threading, Monte-Carlo-Simulations
- Big amounts of computational resources are needed (super computers, public sharing of resources)
- Challenge on prediction algorithms every two years (CASP)
- Folding@home/Foldit

## Topic 4: *Ab initio* protein structure prediction

- Task:
  - Use the I-TASSER tool (server or local installation) to predict the 3D structure of the K-Ras protein and compare your results to the structure found in the crystallography experiment.
- Protein Database (PDB) entry for KRAS:
  - <http://www.rcsb.org/pdb/explore/explore.do?structureId=4L8G>
- I-TASSER service:
  - <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>



What disease is mirtazapine predominantly used for?

↓  
major depression  
↓

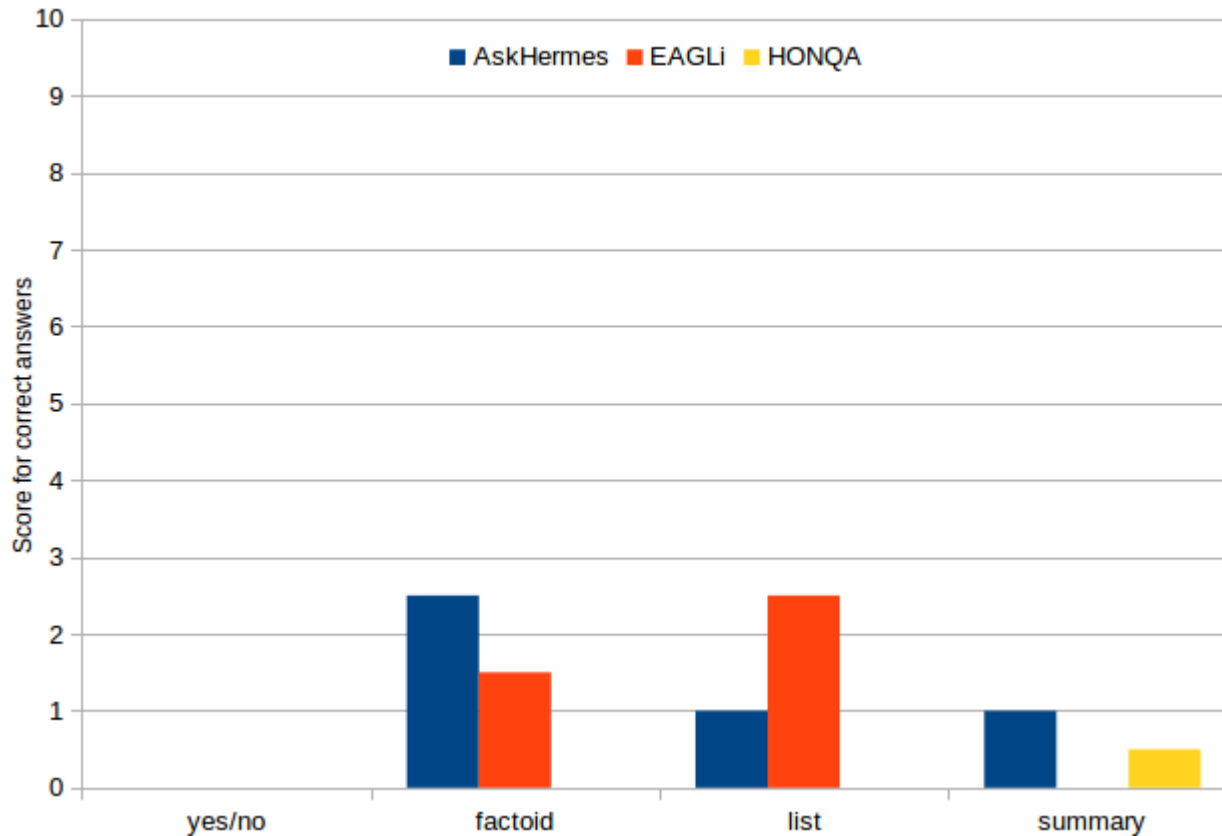
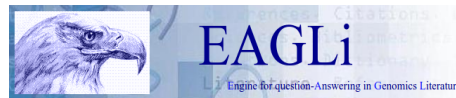
“The 10 most commonly prescribed antidepressant drugs (citalopram hydrobromide (selective serotonin reuptake inhibitor), fluoxetine hydrochloride (selective serotonin reuptake inhibitor), amitriptyline hydrochloride (tricyclic antidepressant), dosulepin hydrochloride (tricyclic antidepressant), paroxetine hydrochloride (selective serotonin reuptake inhibitor), venlafaxine hydrochloride (other), sertraline hydrochloride (selective serotonin reuptake inhibitor), mirtazapine (other), lofepramine (tricyclic antidepressant), and escitalopram (selective serotonin reuptake inhibitor)) comprised 93.6% (n=1) of all antidepressant prescriptions.” (PMID 21810886)

“mirtazapine will make it the first-choice drug in depressive patients with gastric ulcers.” (PMID 19034656)

"If antidepressants are used to treat insomnia, sedating ones should be preferred over activating agents such as serotonin reuptake inhibitors. In general, drugs lacking strong cholinergic activity should be preferred. Drugs blocking serotonin 5-HT<sub>2A</sub> or 5-HT<sub>2C</sub> receptors should be preferred over those whose sedative property is caused by histamine receptor blockade only. The dose should be as low as possible (e.g. as an initial dose: doxepin 25 mg, mirtazapine 15 mg, trazodone 50 mg, trimipramine 25 mg). Regarding the lack of substantial data allowing for evidence-based recommendations, we are facing a clear need for well designed, long-term, comparative studies to further define the role of antidepressants versus other agents in the management of insomnia." (PMID 19016570)

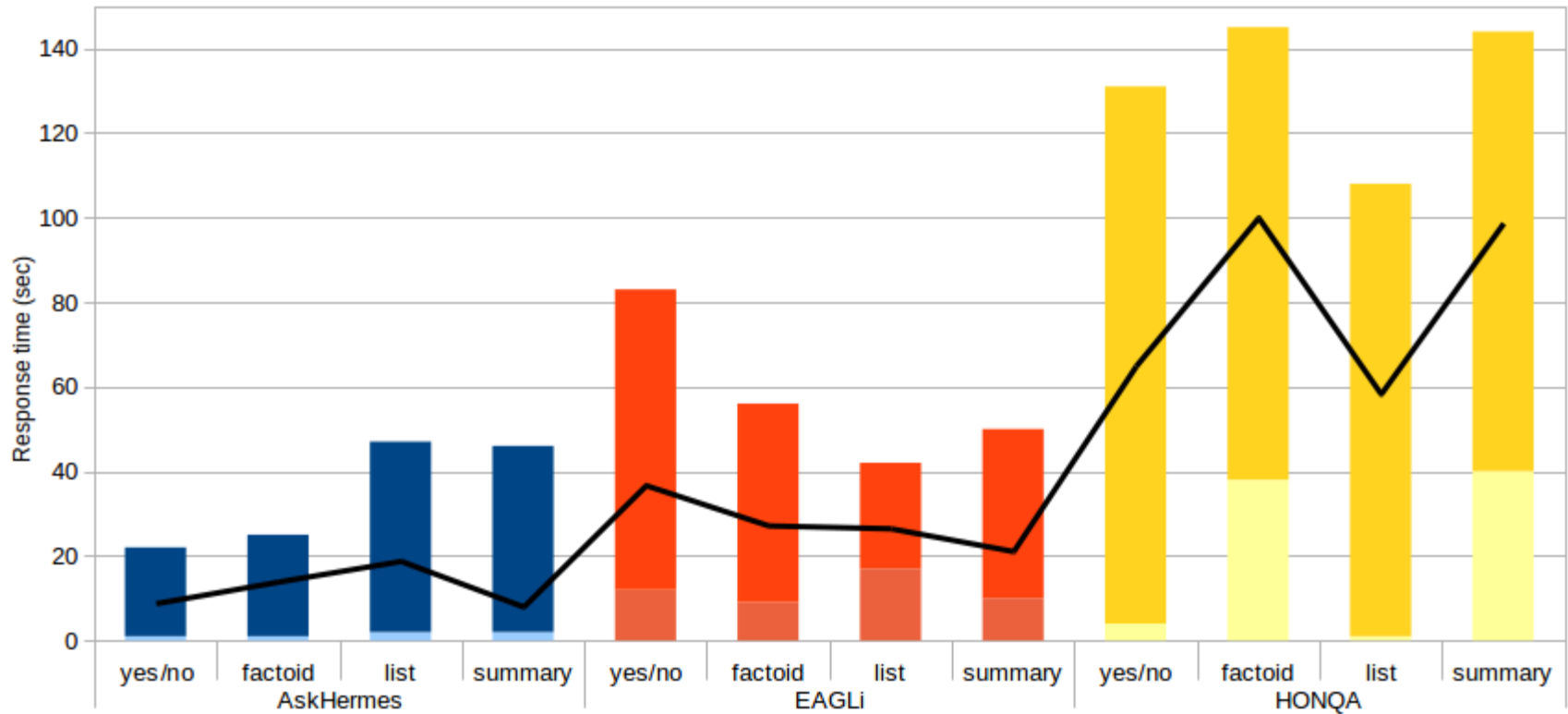
# Topic 5: Question answering

## State-of-the art



# Topic 5: Question answering

## State-of-the art



# Topic 5: Question answering

## ss2015: Master project „Ask your database“

### BioMedical Question Answering System

VM (166,133 documents)

What do you want to know?

What disease is mirtazapine predominantly used for?

ASK

Show analysis details

**METHODS :** Using a case control design, we found nine female patients with AN who had been treated with mirtazapine for depression or anxiety during hospitalization in our department. **RESULTS :** The overall analysis using the convention that a patient is at risk if the HAMD suicide item score is  $>$  or 3, and excluding patients at risk at baseline, demonstrated a statistically significantly lower risk for mirtazapine - compared to placebo treated patients on the HAMD ( odds ratio mirtazapine versus placebo 0.38 ; 95% confidence interval 0.21 0.66 ; P 0.0008 ). However, the use of mirtazapine could be useful in the treatment of AN in adolescence. To study the tumor inhibition effect of mirtazapine, a drug for patients with depression, CT26 luc colon carcinoma bearing animal model was used. Positive effects of mirtazapine treatment on early insomnia were suggested by an item analysis of the HAMD. **CONCLUSIONS AND SCIENTIFIC SIGNIFICANCE :** The results of this study suggest that mirtazapine is superior to placebo in improving sleep in patients with comorbid depression and cocaine dependence, but is not more effective than placebo in reducing cocaine use.

#### Relevant Documents:

- Mirtazapine: a review of its use in major depression.
- Suppressive effect of mirtazapine on the HPA system in acutely depressed women seems to be transient and not related to antidepressant action.
- Mirtazapine in essential tremor: a double-blind, placebo-controlled pilot study.
- Mirtazapine to reduce methamphetamine use: a randomized controlled trial.

## Topic 5: Question answering

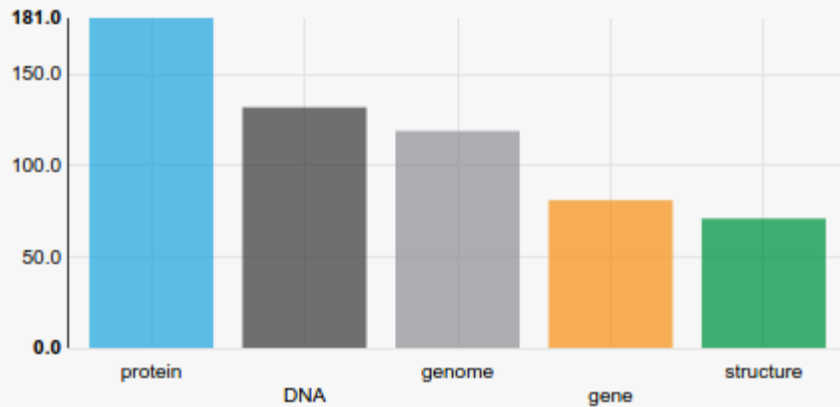
- Understanding how a question answering system works
- Compare and evaluate four tools
  - AskHermes, EAGLi, HONQA and our HPI QA system
- Use a sample of the BioASQ dataset for evaluation
- Evaluate
  - Usability
  - Correctness of the answers (also summaries)
  - Response time
- Describe experiments and evaluation in a **scientific** paper



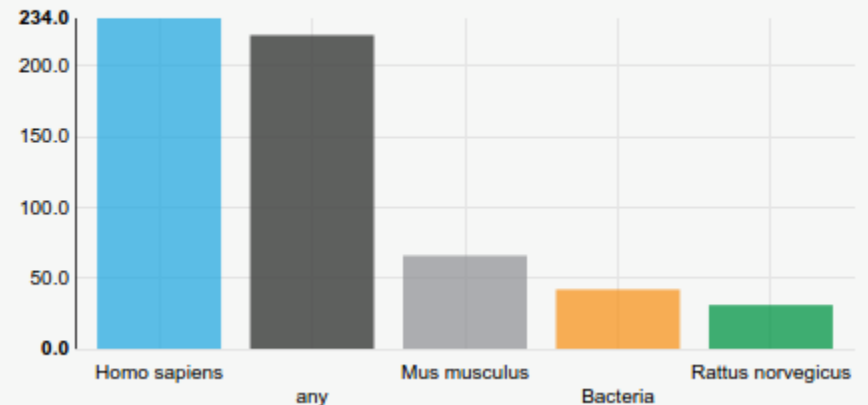
# Topic 6: Text mining tools for biocuration

## Biological databases

Top 5 Domains



Top 5 Species



(<https://www.biosharing.org/summary-statistics/?collection=all>)

# Topic 6: Text mining tools for biocuration

## Biological curation

Arginase (L-arginine urea amidino hydrolase, EC 3.5.3.1) catalyses the hydrolysis of arginine to ornithine and urea and requires a bivalent metal ion, specially  $Mn^{2+}$ , for catalytic activity [1], [2], [3], [4], [5] and [6] and structural stabilization [4], [6] and [7]. Manganese ions are thought to activate a metal-bound water molecule, generating the hydroxide ion that nucleophilically attack the scissile guanidinium carbon of arginine [8] and [9]. An specially interesting aspect of the studies reported to date has been the detection of a  $Mn^{2+}$ - $Mn^{2+}$  cluster in the active site of fully activated arginases from rat liver and *Bacillus caldovelox* [10] and [11]. One of the  $Mn^{2+}$ , designated  $Mn^{2+}_A$  in the case of rat liver arginase, is more weakly bound than the other,  $Mn^{2+}_B$  [12].

General information		
Organism	<a href="#">Homo sapiens</a>	
Tissue	<a href="#">liver</a> ↗	
EC Class	<a href="#">3.5.3.1</a>	
SABIO reaction id	574	
Variant	mutant H101N activated	
Recombinant	expressed in Escherichia coli	
Experiment Type	in vitro	
Pathways	<a href="#">Arginine and Proline metabolism</a> <a href="#">Insulin signaling pathway</a> <a href="#">Urea cycle</a>	
Event Description	-	
Substrates		
name	location	comment
<a href="#">H2O</a>	-	-
<a href="#">L-Arginine</a>	-	-
Products		
name	location	comment
<a href="#">L-Ornithine</a>	-	-
<a href="#">Urea</a>	-	-

(<http://sabio.villa-bosch.de/>)

# Topic 6: Text mining tools for biocuration

## Tools to support biocuration

www.tagtog.net

peter/public/ebola

Settings Corpus Learning Downloads

Upload

pool gold

Save

### Ebolavirus Is Internalized into Host Cells via Macropinocytosis in a Viral Glycoprotein-Dependent Manner

#### Abstract

Ebolavirus (EBOV) is an enveloped, single-stranded, negative-sense RNA virus that causes severe hemorrhagic fever with mortality rates of up to 90% in humans and nonhuman primates. Previous studies suggest roles for clathrin- or caveolae-mediated endocytosis in EBOV entry; however, ebolavirus virions are long, filamentous particles that are larger than the plasma membrane invaginations that characterize clathrin- or caveolae-mediated endocytosis. The mechanism of EBOV entry remains, therefore, poorly understood. To better understand Ebolavirus entry, we carried out internalization studies with fluorescently labeled, biologically contained Ebolavirus and Ebolavirus-like particles (Ebola VLPs), both of which resemble authentic Ebolavirus in their morphology. We examined the mechanism of Ebolavirus internalization by real-time analysis of these fluorescently labeled Ebolavirus particles and found that their internalization was independent of clathrin- or caveolae-mediated endocytosis, but that they co-localized with sorting nexin (SNX) 5, a marker of macropinocytosis-specific endosomes (macropinosomes). Moreover, the internalization of Ebolavirus virions accelerated the uptake of a macropinocytosis-specific cargo, was associated with plasma membrane ruffling, and was dependent on cellular GTPases and kinases involved in macropinocytosis. A pseudotyped vesicular stomatitis virus possessing the Ebolavirus glycoprotein (GP) also co-localized with SNX5 and its internalization and infectivity were affected by macropinocytosis inhibitors. Taken together, our data suggest that Ebolavirus is internalized into cells by stimulating macropinocytosis in a GP-dependent manner. These findings provide new insights into the lifecycle of Ebolavirus and may aid in the development of therapeutics for Ebolavirus infection.

#### Author Summary

#### Meta Information

? on\_ebola  
? pinocytosis

#### Entity Tally

unique 4 total 83

Ebolavirus	18/18
hemorrhagic fever	3/3
caveolae	19/19
Ebola VLPs	42/42

protein	0
organism	1
place	0
disease	1
time	0
localization	2
drug	1

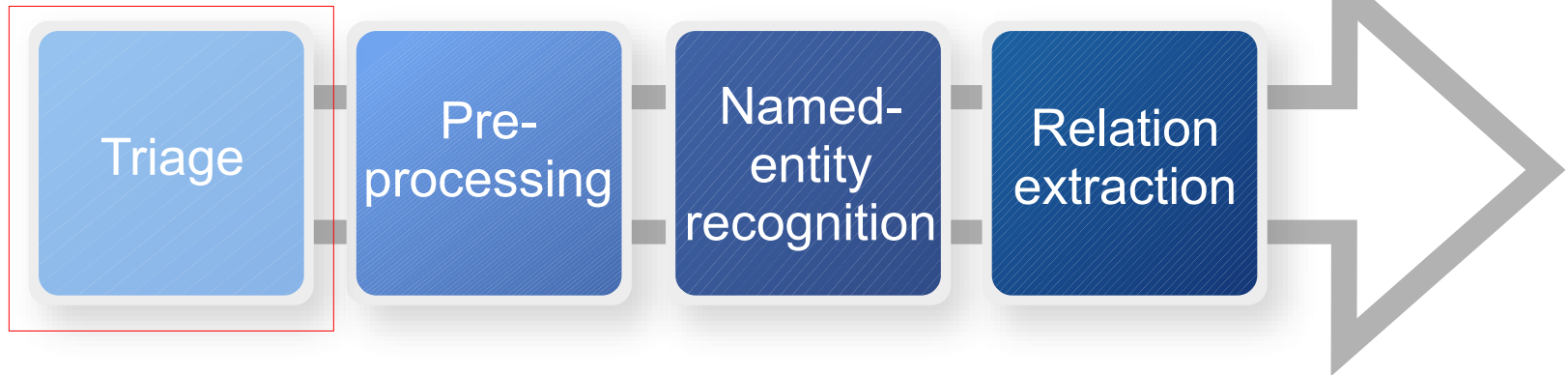
#### Relation Tally

total 0

To create a new Relation:  
1. Click on an entity and then click on Add Relation

# Topic 6: Text mining tools for biocuration

## Text mining workflow



Information retrieval

Text classification

[23030233](#): Safety and proof-of-concept efficacy of inhaled drug loaded nano- and immunonanoparticles in a c-Raf transgenic lung cancer model.

[22429766](#): Non-steroidal anti-inflammatory drug use and the risk of benign prostatic hyperplasia-related outcomes and nocturia in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial.

[19759520](#): Dual targeting of EGFR can overcome a major drug resistance mutation in mouse models of EGFR mutant lung cancer.

[19349852](#): Computed tomography assessment of lung density in patients with lung cancer treated with accelerated hypofractionated radio-chemotherapy supported with amifostine.

[22985911](#): The synergistic effect of EGFR tyrosine kinase inhibitor gefitinib in combination with aromatase inhibitor anastrozole in non-small cell lung cancer cell lines.

[23074402](#): Epidermal Growth Factor Receptor Mutation (EGFR) Testing for Prediction of Response to EGFR-Targeting Tyrosine Kinase Inhibitor (TKI) Drugs in Patients with Advanced Non-Small-Cell Lung Cancer: An Evidence-Based Analysis.

[21713759](#): Human immunodeficiency virus-associated lung cancer in the era of highly active antiretroviral therapy.

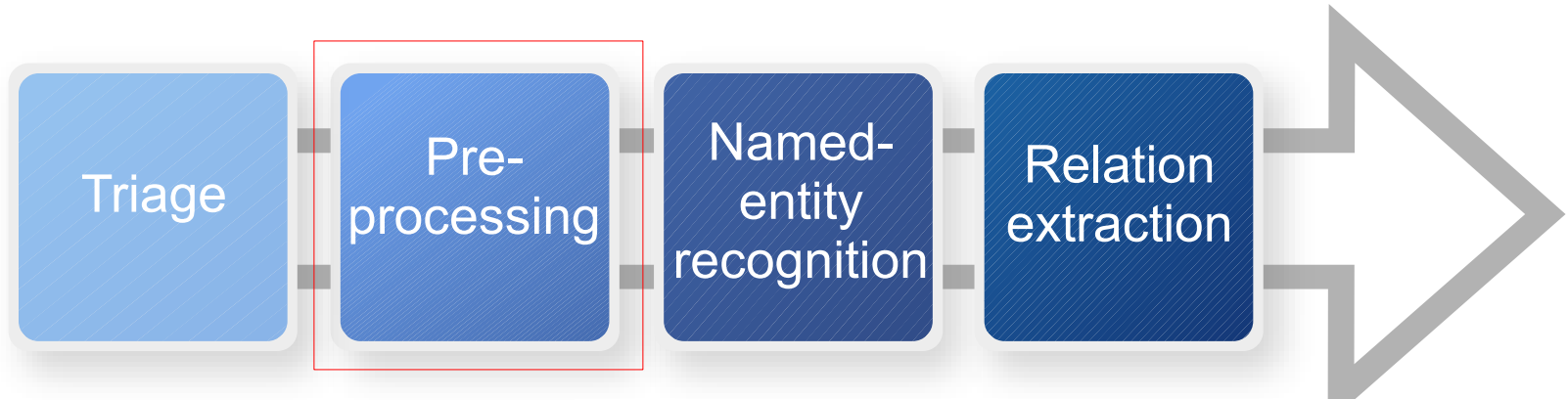
[22253467](#): Recommendations for clinical trials of off-label drugs used to treat advanced-stage cancer.

[22923670](#): The use of metformin and the incidence of lung cancer in patients with type 2 diabetes.

[20026433](#): Structural and mechanistic underpinnings of the differential drug sensitivity of EGFR mutations in non-small cell lung cancer.

# Topic 6: Text mining tools for biocuration

## Text mining workflow



Sentence splitting

Tokenization

Part-of-speech tagging

Syntact parsing

..

```

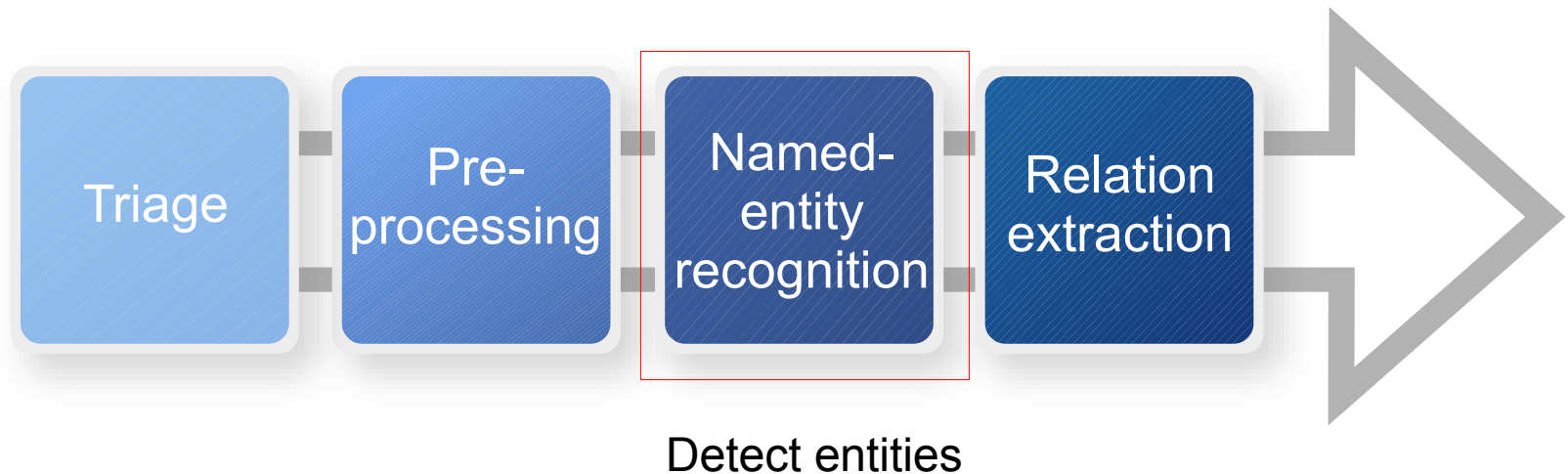
(ROOT
 (S
  (NP (DT Both) (NNS patients))
  (VP
   (VP (VBD responded)
        (PP (TO to)
            (NP (NN treatment))))
        (PP (IN with)
            (NP (NN erlotinib))))
   (CC and)
   (VP (VBD displayed)
        (NP (DT a) (JJ mutated) (NN EGFR))))
  (. .)))
  
```

```

(ROOT
 (S
  (NP
   (NP (NNP Treatment))
   (PP (IN with)
        (NP (NN erlotinib))))
  (VP (MD may)
       (VP (VB induce)
            (NP (JJ complete) (NN response))
            (PP (IN in)
                 (NP
                  (NP (NNS patients))
                  (PP (IN with)
                       (NP (JJ metastatic) (JJ non-small) (NN cell) (NN lung) (NN cancer)))))))
  (. .)))
  
```

## Topic 6: Text mining tools for biocuration

### Text mining workflow

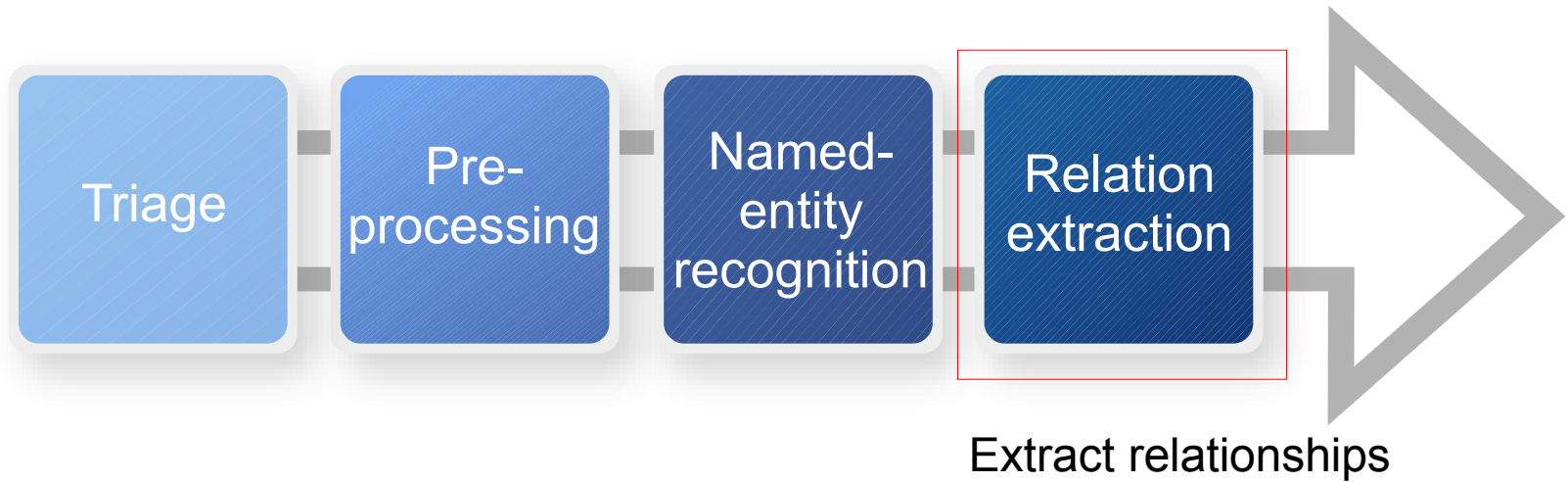


Both patients responded to treatment with **erlotinib** and displayed a mutated **EGFR**.

Treatment with **erlotinib** may induce complete response in patients with metastatic **non-small cell lung cancer**.

# Topic 6: Text mining tools for biocuration

## Text mining workflow



Both patients responded to treatment with **erlotinib** and displayed a mutated **EGFR**.

Treatment with **erlotinib** may induce complete response in patients with metastatic **non-small cell lung cancer**.

## Topic 6: Text mining tools for biocuration

- Understand a text mining workflow for data curation
- Compare and evaluate selected tools
  - Argo, tagtog, PubTator, ...
  - Use case: SABIO-RK database
- Evaluate
  - Usability (easy to use)
  - Flexibility (configurable)
  - Quality of the predictions (entities)
- Describe experiments and evaluation in a **scientific** paper



## Topic 7: Clinical Decision support systems

- Huge variety of clinical (increasingly genomic) data
- Questions to be answered1:
  - What disease does this patient have?
  - Should this patient be treated?
  - Should testing be done?
- Accurate, complete, relevant data
- Reliance on pattern recognition and customary practices
- Evidence-based medicine (clinical guidelines)
- Medical errors

## Topic 7: Clinical Decision support systems

- Can IT help physicians make better decisions?



## Topic 7: Clinical Decision support systems

- The best answer is: **maybe**

Both commercially and locally developed CDSSs are effective at improving health care process measures across diverse settings, but **evidence** for clinical, economic, workload, and efficiency outcomes **remains sparse**. (Bright et al. 2012)

**Few studies have found any benefits on patient outcomes**, though many of these have been too small in sample size or too short in time to reveal clinically important effects. (Jaspers et al. 2011)

## Topic 7: Clinical Decision support systems

- What are CDSS?
  - Definition:

“Clinical decision support systems (CDSS) provide clinicians, staff, patients, and other individuals with **knowledge** and **person-specific information**, intelligently filtered and presented at **appropriate times**, to enhance health and health care” Berner (2009)

## Topic 7: Clinical Decision support systems

- They may provide:

1) Contextual retrieval of highly relevant information (infobuttons)



2) Patient-specific reminders and recommendations (direct actions)



3) Organization and presentation of information (dashboards)



# Topic 7: Clinical Decision support systems

- Contextual retrieval of highly relevant information

**Patient has been diagnosed with Angina.**


**One click of a button looks up the evidence-based treatment and patient management information relevant to this patient. No searching or log-in needed!**

Source: <http://www.practicefusion.com/>

# Topic 7: Clinical Decision support systems

- Patient-specific reminders and recommendations

Discern:

 **Medication Alert - Capecitabine**

**Patient has DPD deficiency**

This patient has deficiency of dihydropyrimidine dehydrogenase (DPD) activity.

Rarely, unexpected, severe toxicity (e.g. stomatitis, diarrhea, neutropenia and neurotoxicity) associated with 5-fluorouracil has been attributed to DPD deficiency.

A link between decreased levels of DPD and increased, potential fatal toxic effects of 5-fluorouracil therefore cannot be excluded.

*[[Information derived from FDA drug label]]*

**Alert Action**

Cancel Order

Override Alert

Modify Order

EVIDENCE OK

Discern: (1 of 1)

 **PHARMACOGENOMICS ALERT**

**WARNING:** Patient carries a genetic variant that influences clopidogrel (Plavix) metabolism, resulting in impaired responsiveness.

- Consider prasugrel (Effient) or other alternative therapy
- For CPIC dosing guidelines, click the Guidelines button below, or
- Contact a clinical pharmacist for more information (598-6347)

**MEDICATION ORDER:** clopidogrel

**GENETIC RESULT:** NEXT Exome CYP2C19 Result (Sendout) Clopidogrel, Impaired Responsiveness October 23, 2013 17:26:00 PDT

**NOTE:** This is an experimental pharmacogenomics alert created for patients in the NEXT01 Exome sequencing study.

- You may receive an e-mail asking for your feedback on this alert.

**Alert Action**

Cancel Order

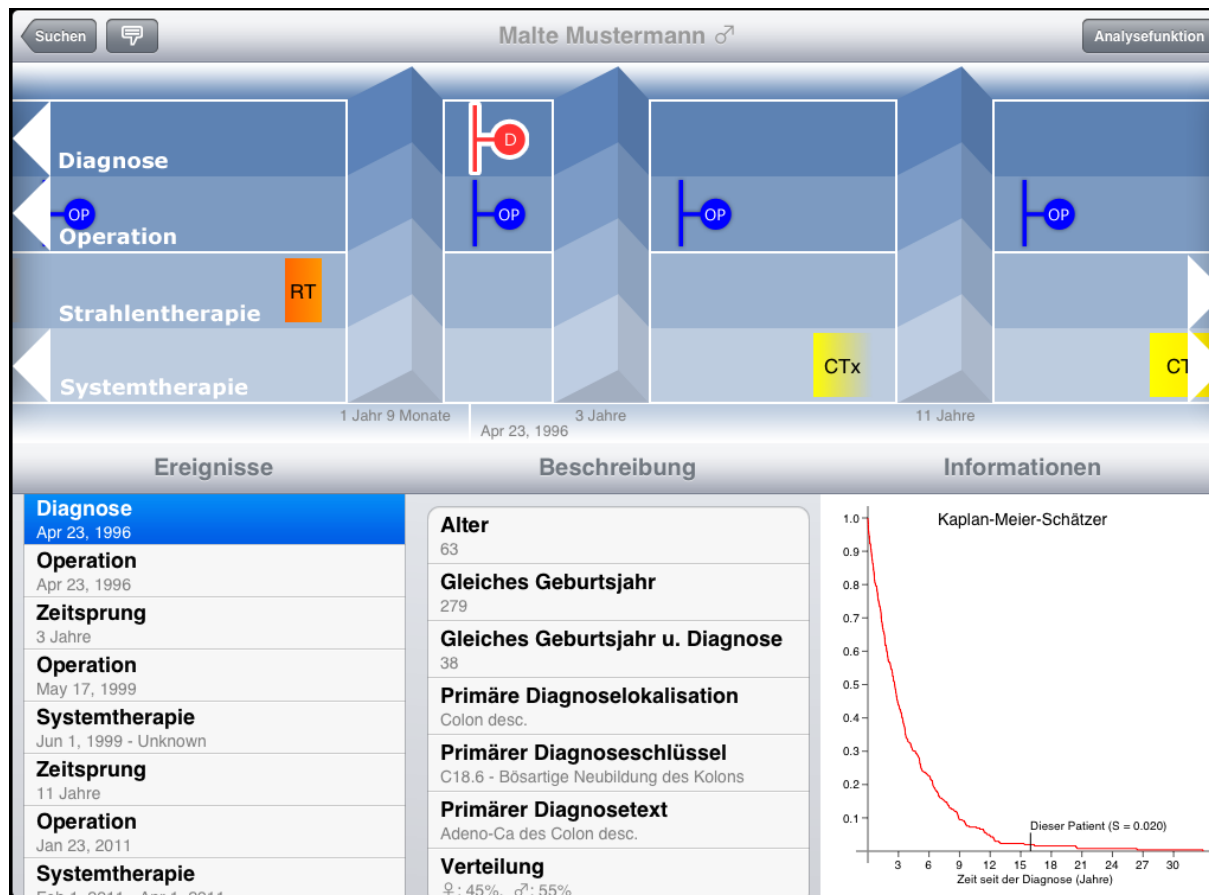
Override Alert

Modify Order

Guidelines OK

# Topic 7: Clinical Decision support systems

- Organization and presentation of information





## Topic 7: Clinical Decision support systems

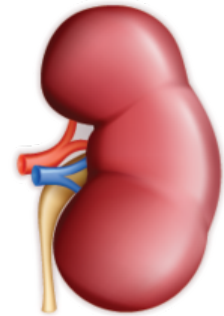
- The scenario
  - Acute and chronic kidney disease
- Currently in Germany<sup>1</sup>
  - 70.000 patients / 2,5 Mio. EUR p.a.
  - 100.000 patients by 2020
- Kidney disease is asymptomatic (silent)
- Severe implications for patients
- Higher risk of mortality
- Very high medical costs for dialysis



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

## Topic 7: Clinical Decision support systems

- Task: develop a Bayesian network for Acute Kidney Disease diagnosis
- Drastic drop in kidney function (7 days)
- Increase in potassium levels
- Risk factors: diabetes, hypertension, CVD, CKD history, and age > 60 years<sup>1</sup>
- Scoring guidelines: RIFLE and AKIN criteria, CKD stages, KDIGO<sup>2</sup>
- Glomerular Filtration Rate (GFR), creatinine (over time), urine protein<sup>3</sup>



Source: peterdobias.com

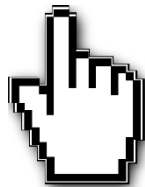
- [1] Levey & Coresh (2013)
- [2] Kidney International (2007)
- [3] Vassalotti et al. (2007)
- [4] Tawadrous et al. (2011)
- [5] Sawhney et al. (2015)

## Topic 7: Clinical Decision support systems

- What tools shall be used? Weka and GeNIe



[weka.sourceforge.net](http://weka.sourceforge.net)



[dslpitt.org/genie/](http://dslpitt.org/genie/)



# Seminar Trends in Bioinformatics

- Contact: V0.01 (Villa), HPI Campus II



Cindy Perscheid

Cindy.Perscheid@hpi.de



Dr. Mariana Neves

Mariana.Neves@hpi.de



Milena Kraus

Milena.Kraus@hpi.de



Harry Cruz

Harry.FreitasDaCruz@hpi.de