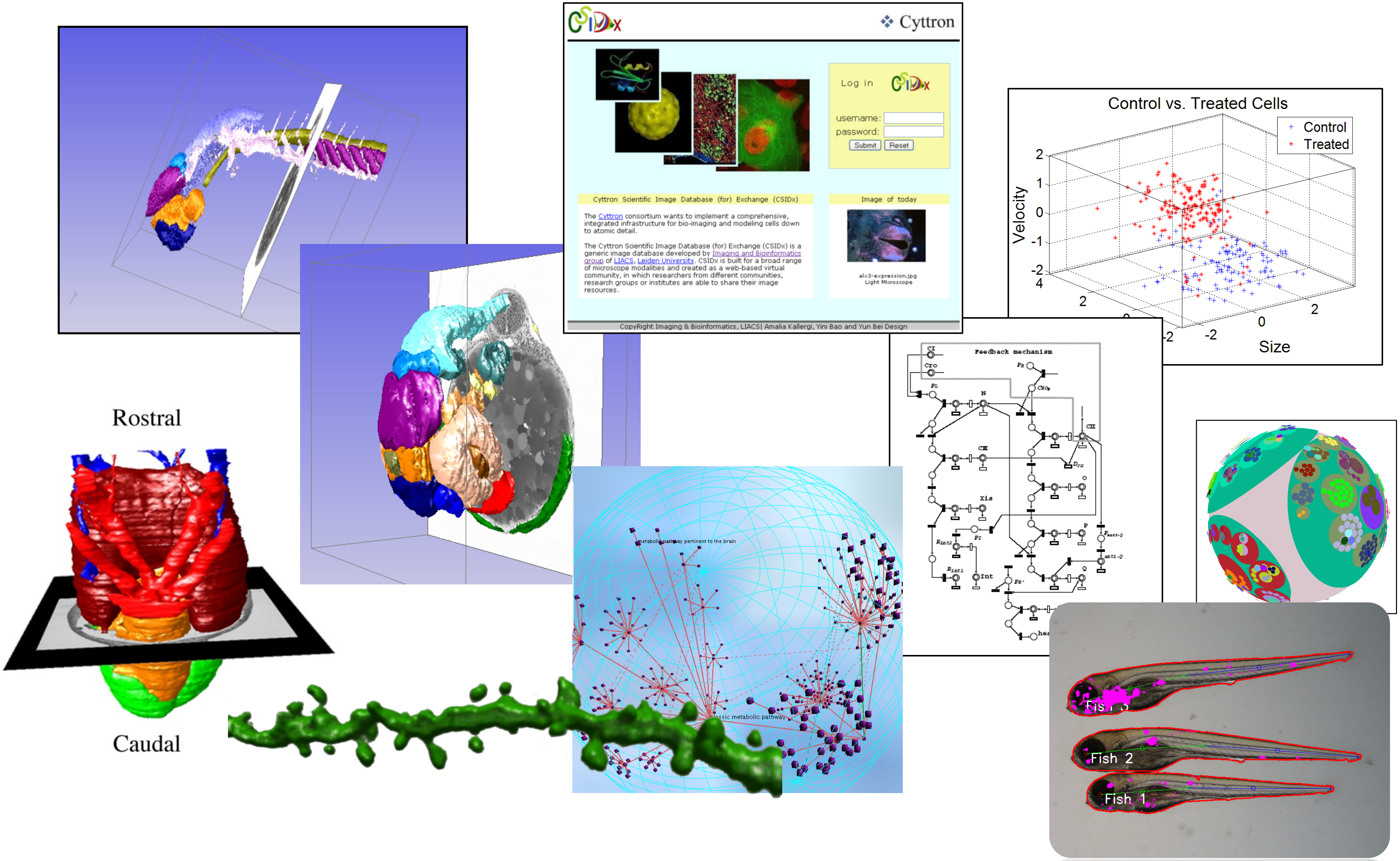
# Imaging & BioInformatics

FJ. Verbeek, K. Wolstencroft, A. Gultyaev,
J Slob, R Carvalho, L. Cao,
C. Fuyu, E Larios, M. Tleis, Z. Xiong

# Imaging & BioInformatics @ LIACS



Cyttron Scientific Image Database (for) Exchange (CSIDx)

The Cyttron consortium wants to implement a comprehensive, integrated infrastructure for bio-imaging and modeling cells down to atomic detail.

The Cyttron Scientific Image Database (for) Exchange (CSIDx) is a generic image database developed by Imaging and Bioinformatics group of LIACS, Leiden University. CSIDx is built for a broad range of microscope modalities and created as a web-based virtual community, in which researchers from different communities, research groups or institutes are able to share their image resources.

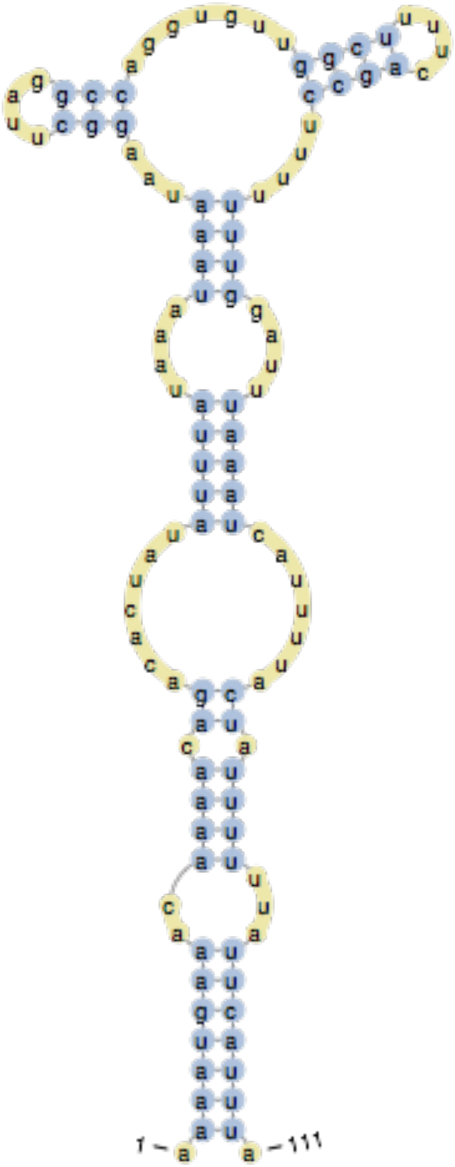
Image of today

alx3-expression.jpg
Light Microscope

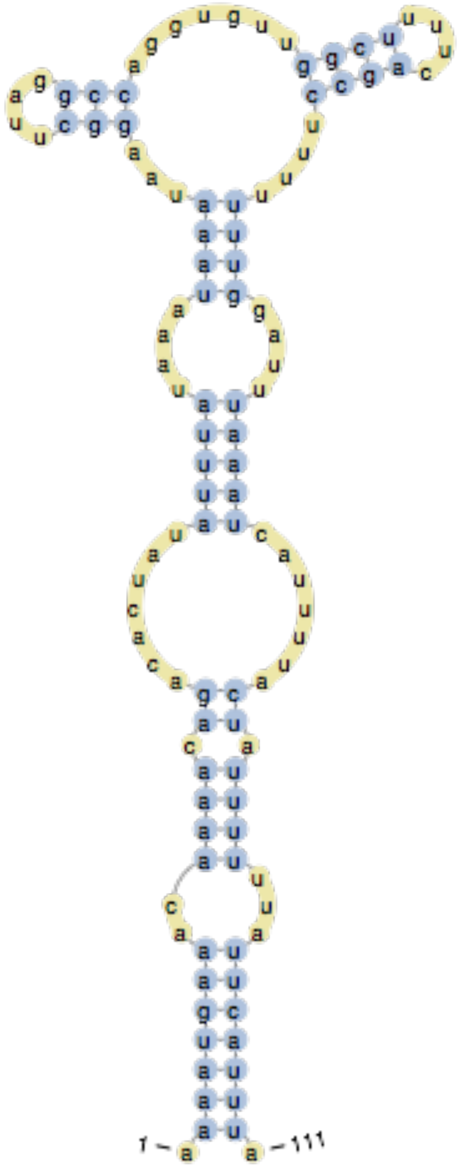
CopyRight Imaging & Bioinformatics, LIACS| Amalia Kallergi, Yini Bao and Yun Bei Design

# RNOMICS

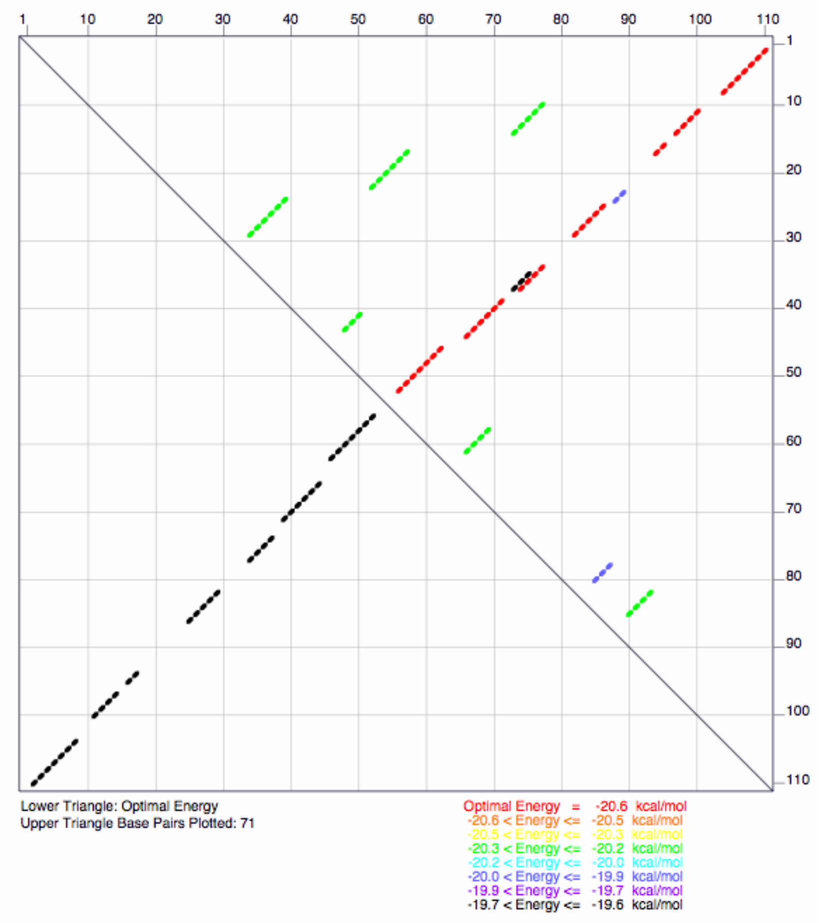
# RNomics: identification of functional RNAs encoded in genomes



# RNA secondary structure prediction



Dynamic programming algorithm can compute both optimal and suboptimal structures that can be shown in various ways in e.g. “dot-plots”:



# Comparative RNA structure analysis

*A powerful approach in RNA structure prediction, in particular, due to RNA-specific patterns of variation, nucleotide covariations.*

An example of two covariations in three related RNA’s:

```
 n n          n n          n n
n   n        n   n        n   n
 n-n          n-n          n-n
 G-C          U-A          A-U
 n-n          n-n          n-n
 n-n          n-n          n-n
 A-U          G-C          C-G

RNA 1        RNA 2        RNA 3
```

**A**nn**G**nnnnnn**C**nn**U** RNA 1
**G**nn**U**nnnnnn**A**nn**C** RNA 2
**C**nn**A**nnnnnn**U**nn**G** RNA 3
(((((....))))) consensus "bracket view"

# Detecting conserved structures in related RNAs

*(prediction of “consensus” structures)*

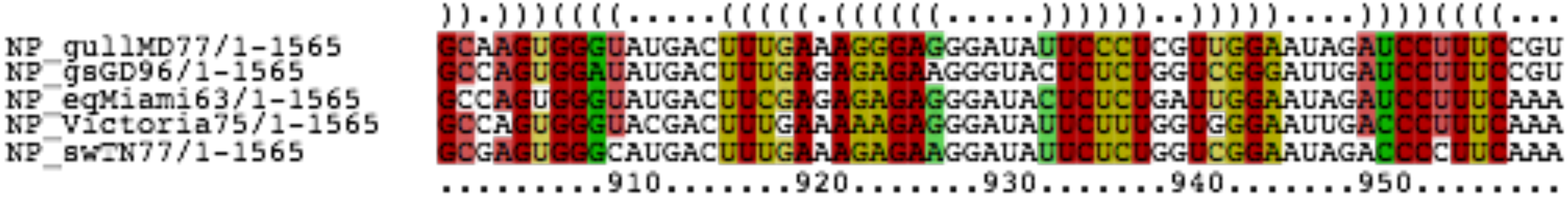
Consensus structures can be computed from sequence alignments using information from suboptimal structures, base probabilities and covariation patterns

**Input:** Sequence alignment

**Calculation:** suboptimal structures/partition functions/base probabilities for individual sequences; detection of common patterns and their scoring

**Output:** The “consensus” structure, (ideally) conserved in all sequences of the dataset.

For instance, a fragment of the output of RNAalifold algorithm:



```
                         )).)))((((.....(((((.((((((.....))))))..)))))....))))((((...
NP_gullMD77/1-1565       GCAAGUGGGUAUGACUUUGAAAGGGAGGGAUAUUCCCUCGUUGGAAUAGAUCCUUUCCGU
NP_gsGD96/1-1565         GCCAGUGGAUAUGACUUUGAGAGAGAAGGGUACUCUCUGGUCGGGAUUGAUCCUUUCCGU
NP_eqMiami63/1-1565      GCCAGUGGGUAUGACUUCGAGAGAGAGGGAUACUCUCUGAUUGGAAUAGAUCCUUUCAAA
NP_Victoria75/1-1565     GCCAGUGGGUACGACUUUGAAAAAGAGGGAUAUUCUUUGGUGGGAAUUGACCCUUUCAAA
NP_swTN77/1-1565         GCGAGUGGGCAUGACUUUGAAAGAGAAGGAUAUUCUCUGGUCGGAAUAGACCCCUUCAAA
                         .........910.......920.......930.......940.......950........
```

# Detecting conserved structures in related RNAs

*(prediction of "consensus" structures)*

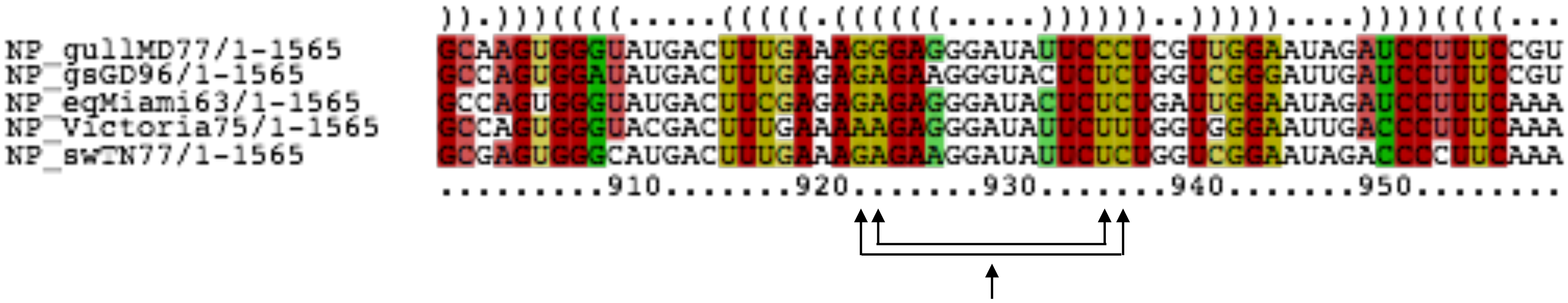
Consensus structures can be computed from sequence alignments using information from suboptimal structures, base probabilities and covariation patterns

**Input:** Sequence alignment

**Calculation:** suboptimal structures/partition functions/base probabilities for individual sequences; detection of common patterns and their scoring

**Output:** The "consensus" structure, (ideally) conserved in all sequences of the dataset.

For instance, a fragment of the output of RNAalifold algorithm:



```
                        )).)))((((.....(((((.((((((.....))))))..)))))....))))((((...
NP_gullMD77/1-1565      GCAAGUGGGUAUGACUUUGAAAGGGAGGGAUAUUCCCUCGUUGGAAUAGAUCCUUUCCGU
NP_gsGD96/1-1565        GCCAGUGGAUAUGACUUUGAGAGAGAAGGGUACUCUCUGGUCGGGAUUGAUCCUUUCCGU
NP_eqMiami63/1-1565     GCCAGUGGGUAUGACUUCGAGAGAGAGGGAUACUCUCUGAUUGGAAUAGAUCCUUUCAAA
NP_Victoria75/1-1565    GCCAGUGGGUACGACUUUGAAAAAGAGGGAUAUUCUUUGGUGGGAAUUGACCCUUUCAAA
NP_swTN77/1-1565        GCGAGUGGGCAUGACUUUGAAAGAGAAGGAUAUUCUCUGGUCGGAAUAGACCCCUUCAAA
                        .........910.......920.......930.......940.......950........
```

Such structure-annotated alignments allow one to identify covariations.

# Plant enod40 gene

**- One of the most intriguing genes involved in the regulation of symbiotic interaction between plants and bacteria or fungi;**

**- Initially identified at the early stages of formation of nitrogen-fixing root nodules of legume plants;**

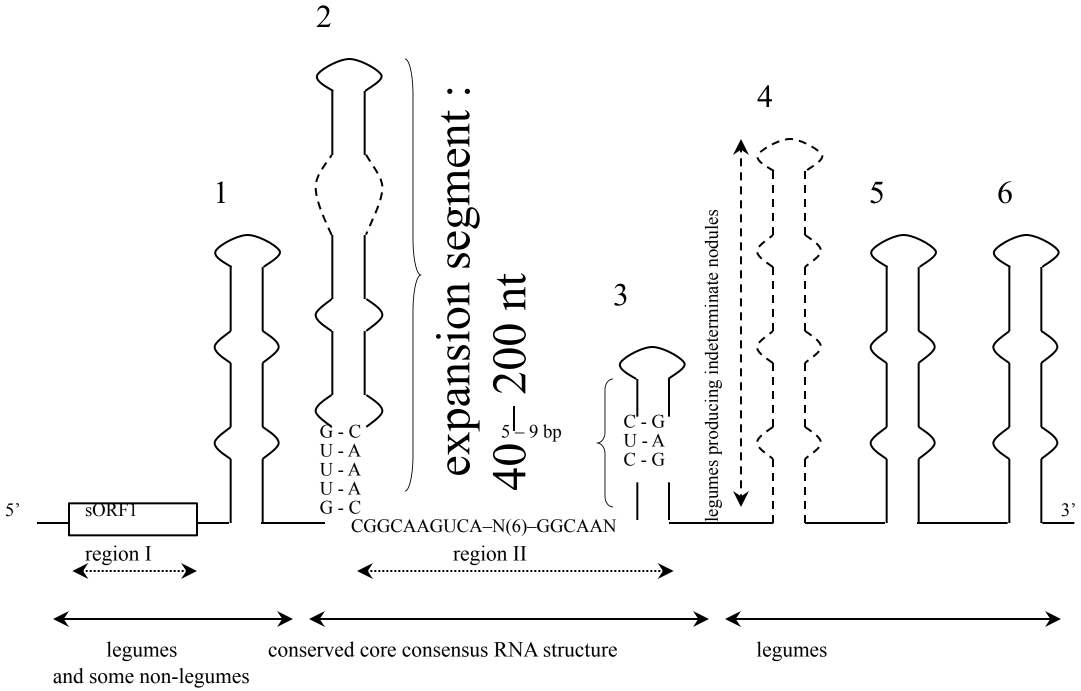
**- Also found in non-legume species;**

**- Apparently has a dual function:**

*- encodes small conserved peptides*

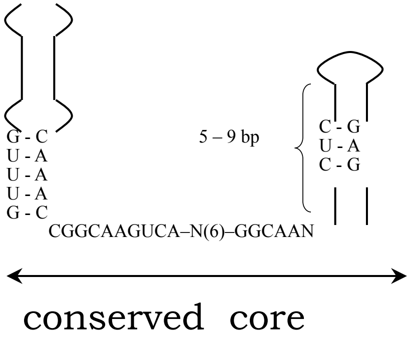
*- contains conserved RNA structural elements*

- **Enod40 RNA is highly structured in legumes.**
- **The consensus core structure turns out to be very conserved in other plants as well.**



(Gultyaev & Roussis, 2007)

- **Enod40 RNA is highly structured in legumes.**
- **The consensus core structure turns out to be very conserved in other plants as well.**
- **This consensus can be used for the search of new enod40 genes in the nucleotide databases (GenBank, EST, WGS).**



A search protocol:

-Using the conserved core sequence of a known enod40 RNA as a query, BLAST the databases.

- Analyze the structures flanking the putative enod40 RNA cores in the BLAST hits (including those with rather high E-values). The presence of a structure consistent with consensus is an evidence for enod40 motif.

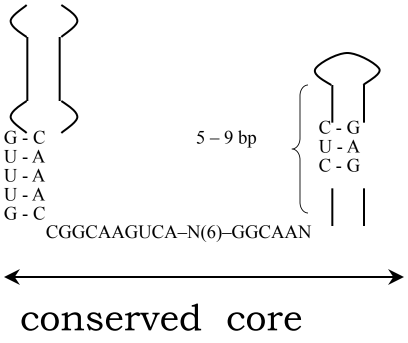
- New (putative) enod40 sequences are used again as BLAST queries to find enod40s in other species.

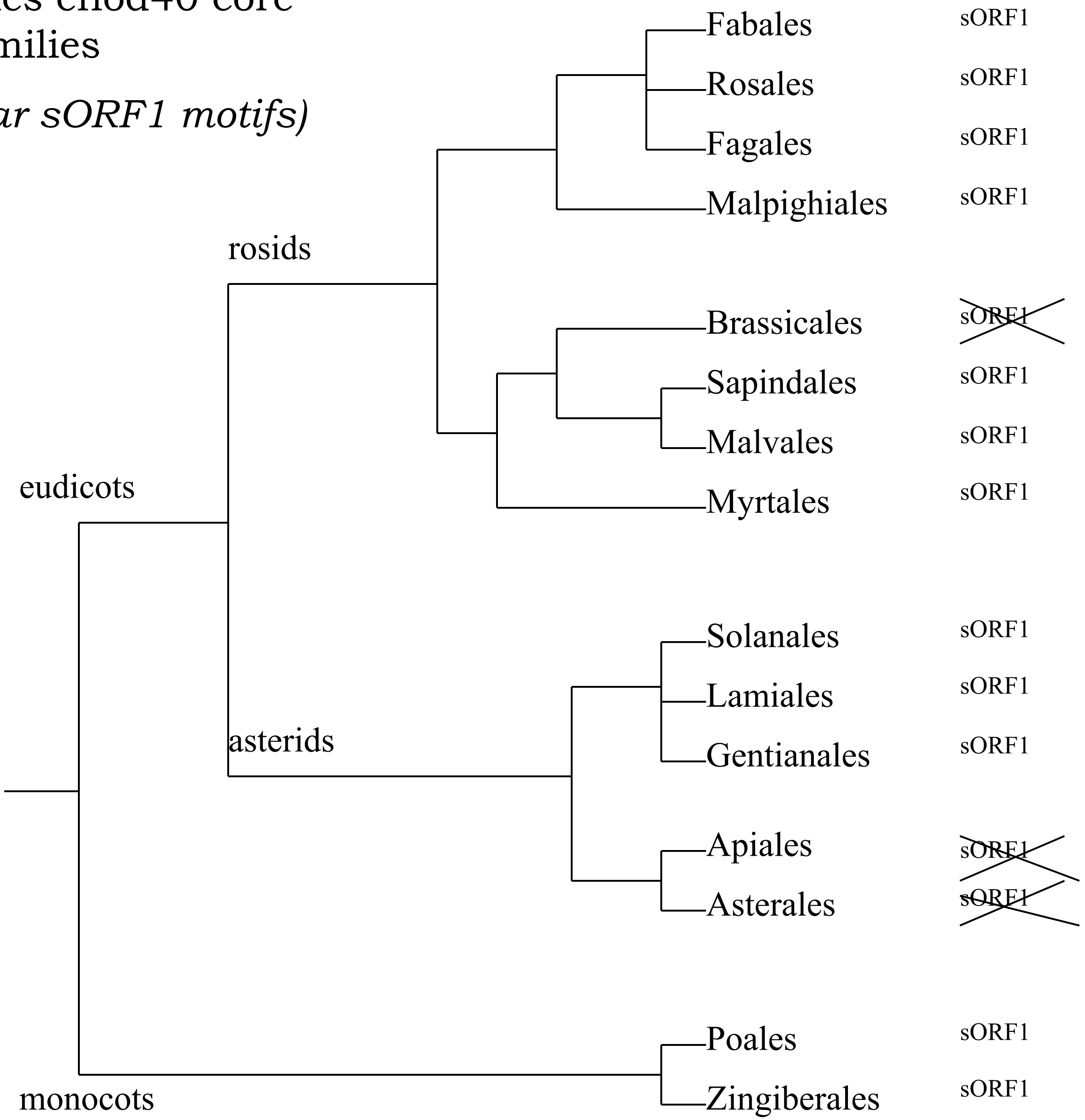
- NB. Due to low sequence similarities, a straightforward BLAST is not efficient (reliable BLAST hits with low E-values are yielded only for close relatives, while many hits with high E-values are false-positive results).

(Gultyaev & Roussis, 2007)

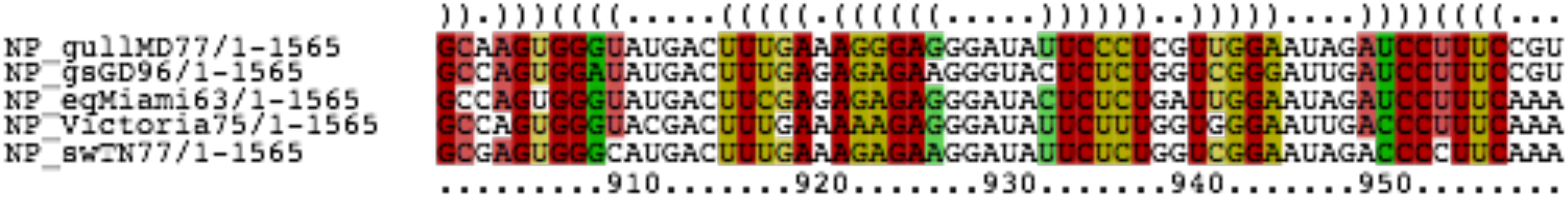
# A database search identifies enod40 core motifs in various plant families

*(also in the absence of clear sORF1 motifs)*

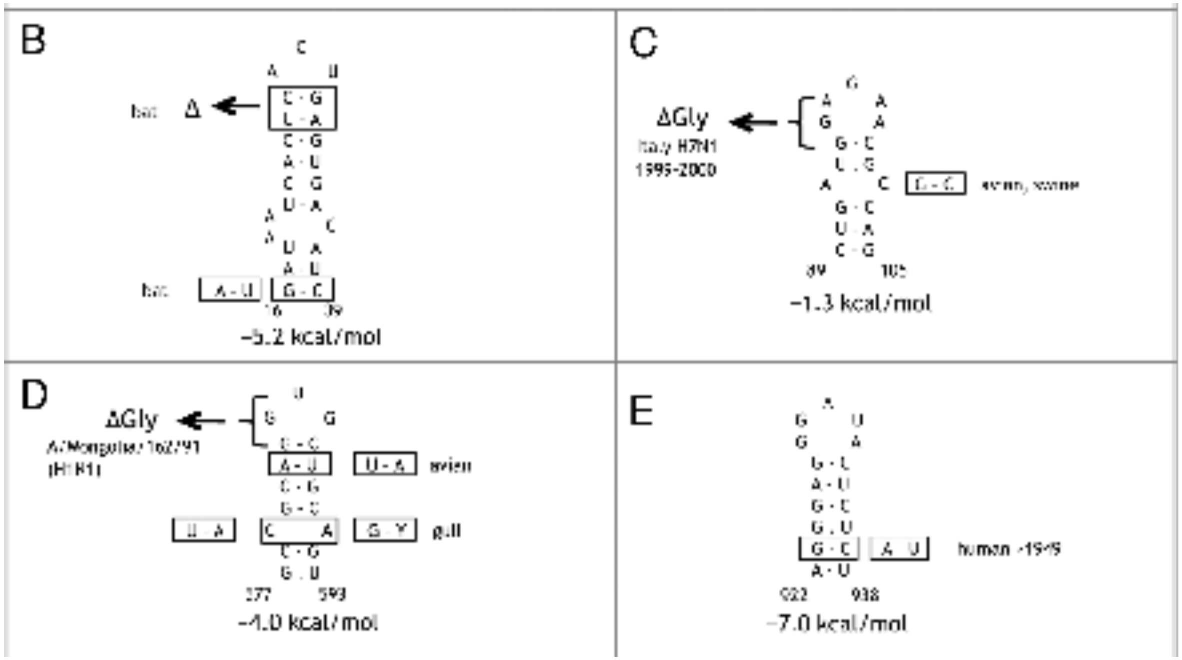




# Detecting conserved structures in the influenza virus genome



```
                            )).))))((((.....(((((.((((((.....))))))..))))).....))))((((...
NP_gullMD77/1-1565          GCAAGUGGGUAUGACUUUGAAAGGGAGGGAUAUUCCCUCGUUGGAAUAGAUCCUUUCCGU
NP_gsGD96/1-1565            GCCAGUGGAUAUGACUUUGAGAGAGAAGGGUACUCUCUGGUCGGGAUUGAUCCUUUCCGU
NP_eqMiami63/1-1565         GCCAGUGGGUAUGACUUCGAGAGAGAGGGAUACUCUCUGAUUGGAAUAGAUCCUUUCAAA
NP_Victoria75/1-1565        GCCAGUGGGUACGACUUUGAAAAAGAGGGAUAUUCUUUGGUGGGAAUUGACCCUUUCAAA
NP_swTN77/1-1565            GCGAGUGGGCAUGACUUUGAAAGAGAAGGAUAUUCUCUGGUCGGAAUAGACCCCUUCAAA
                            .........910.......920.......930.......940.......950........
```



(Gultyaev et al., 2014)

Important similarities and differences between virus strains can be identified

# WORKFLOWS & SEMANTIC INTEGRATION

# Data and Knowledge Integration

## Scientific workflows

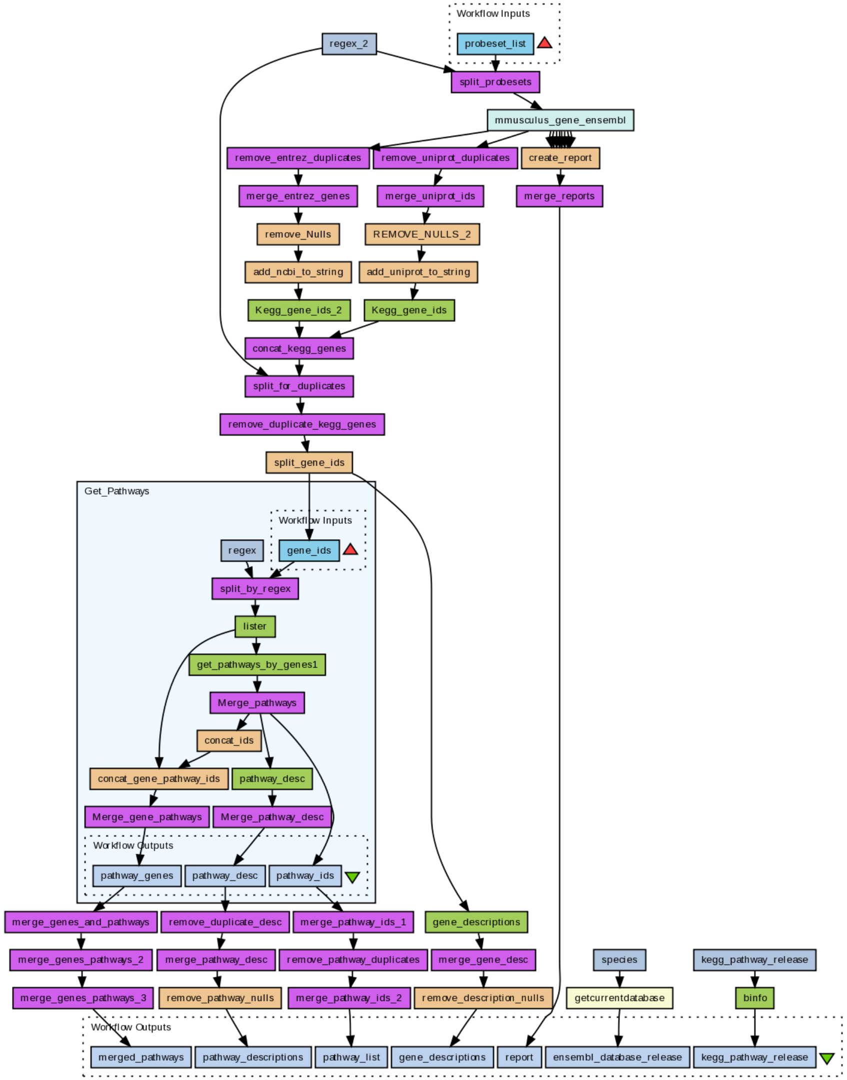
- Distributed computing
- Experimental reproducibility
- Analysing and processing high-throughput (bio)informatics data

## Semantic integration

- Using semantic annotation to explore and understand complex, heterogeneous biological data
- Make new connections and inferences from existing data

# Taverna Scientific Workflows

- Sophisticated analysis pipelines
- A set of services to analyse or manage data (either local or remote Web services)
- Automation of data flow
- Iteration over data sets
- Control of service invocation
- Provenance collection
- Experimental protocols



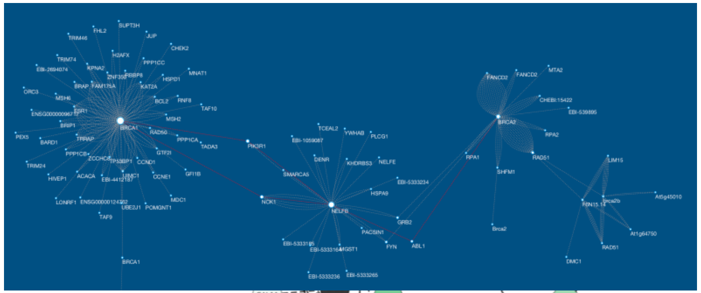
# Taverna Database Integration

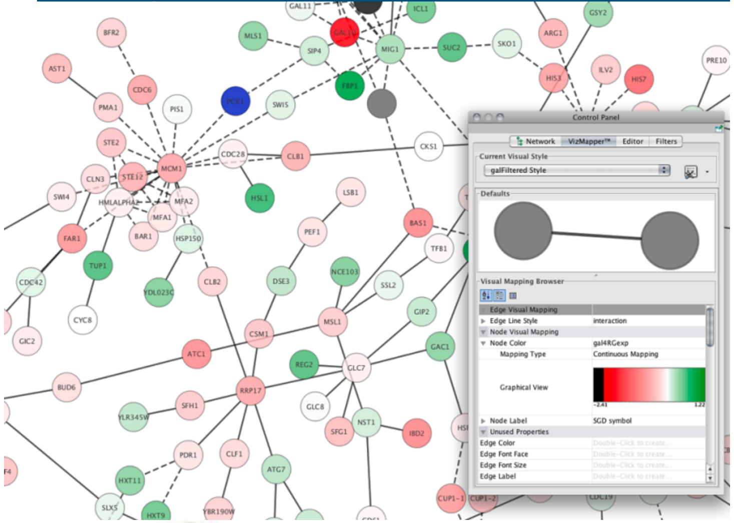
- Data processing is by strings and lists
  - Data transformation services required
  - What we often require are tables
- Develop a plugin to enable database population from Taverna workflows
  - local desktop
  - Server, cloud/grid
- Evaluation with case studies in high throughput bioinformatics analysis

# Taverna and Cytoscape Integration

Cytoscape Network Visualisation tool

- Execute Taverna workflows through a cytoscape plugin to allow ingestion and integration of multiple data sources (‘biological layers’)
- Explore plugin in relation to systems biology case studies





Control Panel

Network | VizMapper™ | Editor | Filters

Current Visual Style: galFiltered Style

Defaults

Visual Mapping Browser

| Edge Visual Mapping | |
|---|---|
| Edge Line Style | interaction |
| Node Visual Mapping | |
| Node Color | gal4RGexp |
| Mapping Type | Continuous Mapping |
| Graphical View | -2.41 … 1.22 |
| Node Label | SGD symbol |
| Unused Properties | |
| Edge Color | Double-Click to create... |
| Edge Font Face | Double-Click to create... |
| Edge Font Size | Double-Click to create... |
| Edge Label | Double-Click to create... |

Cytoscape

# Semantic Annotation and Matching of CellML Models

CellML is an XML standard for describing mathematical models of biological processes (e.g. cardiovascular circulation, metabolism, neurobiology).

- Over 1000 published models
- Models have structured metadata, but biological entities have no semantic descriptions.
  - Difficult to find relevant models and identify where models may overlap
  - Difficult to find experimental data to compare to model simulations

**Objectives:**

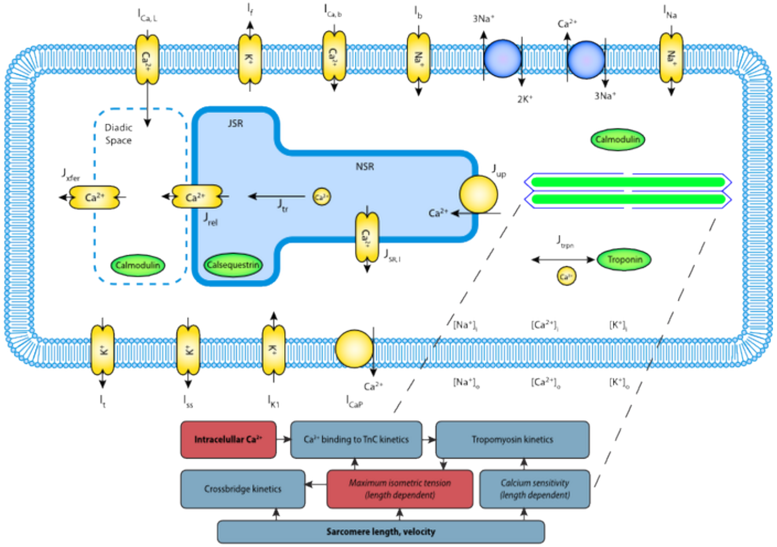
- Develop an automated approach to semantically annotating CellML models with terms from biological ontologies
- Perform semantic matching and similarity measures between CellML models
- Link models to relevant data, using scientific workflows

**Prerequisites** (helpful, not mandatory):

- Databases
- Ontologies
- Classification and semantic similarity
- An interest in bioinformatics

**Advisor:** Dr Katy Wolstencroft

**Contact:** k.j.wolstencroft@liacs.leidenuniv.nl



# COMPUTATION

# Leiden Life Sciences Cluster (LLSC)

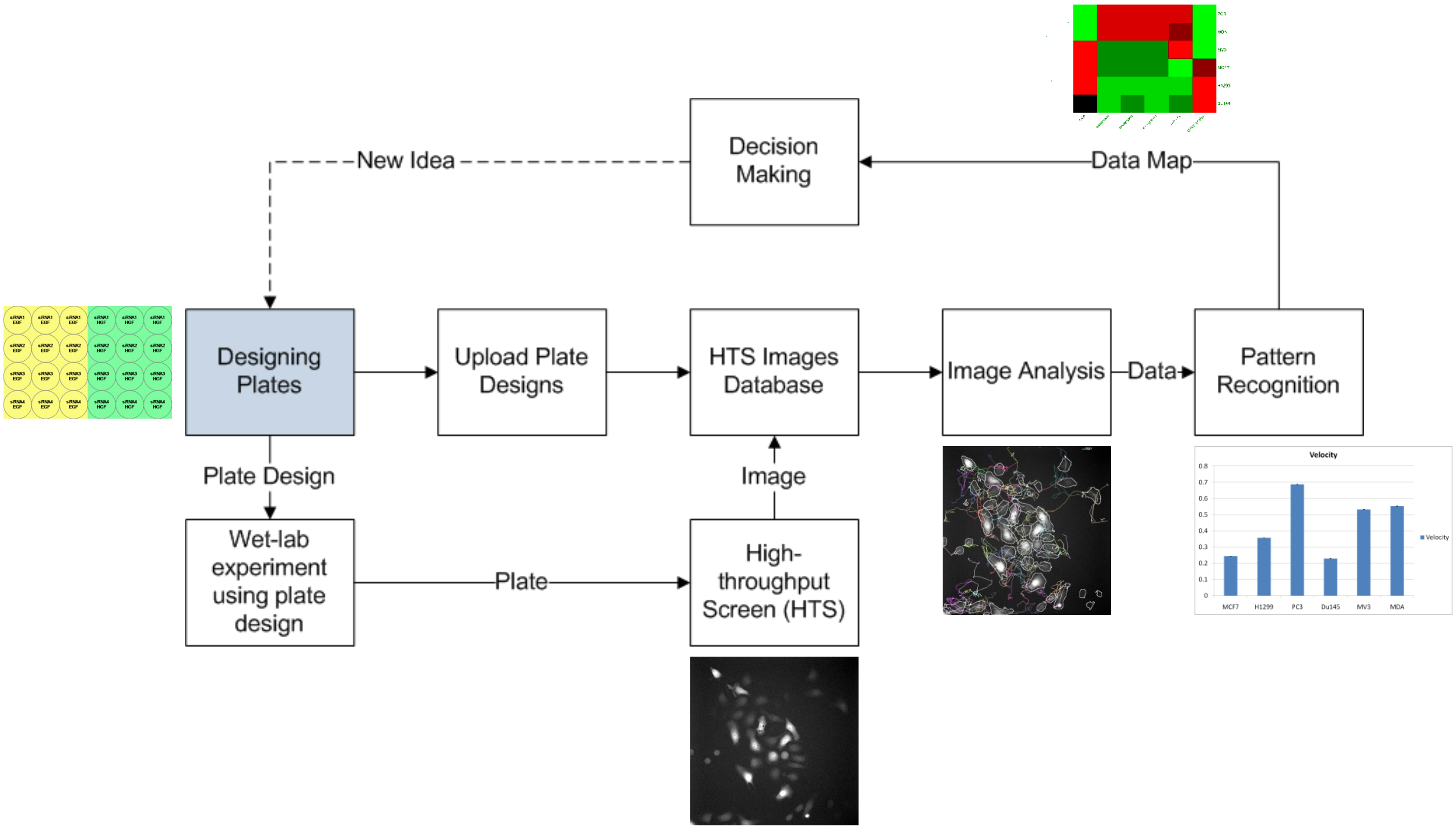
# LLSC: Scientific Cluster

- 48 Dell 2u 8x cores
- Cluster
  - Head Node
  - Main Node
  - Worker Nodes
- Map computational intensive programs to Cluster
  - image processing/analysis
  - pattern recognition
  - bioinformatics jobs from workflows
- RESEARCH
  - Develop strategy/template for web-services
  - Develop strategy for Parallelization

# Cluster Computation

- NeCEN
  - – Netherlands Centre for Electron Nanoscopy
- Cell Observatory
  - – Collaboration within Faculty of Science
  - – IBL, LIC, LION, LIACS
- High-Throughput imaging
  - – 3D zebrafish imaging

# Systems for High-Throughput



Decision Making

New Idea

Data Map

Designing Plates

Upload Plate Designs

HTS Images Database

Image Analysis

Data

Pattern Recognition

Plate Design

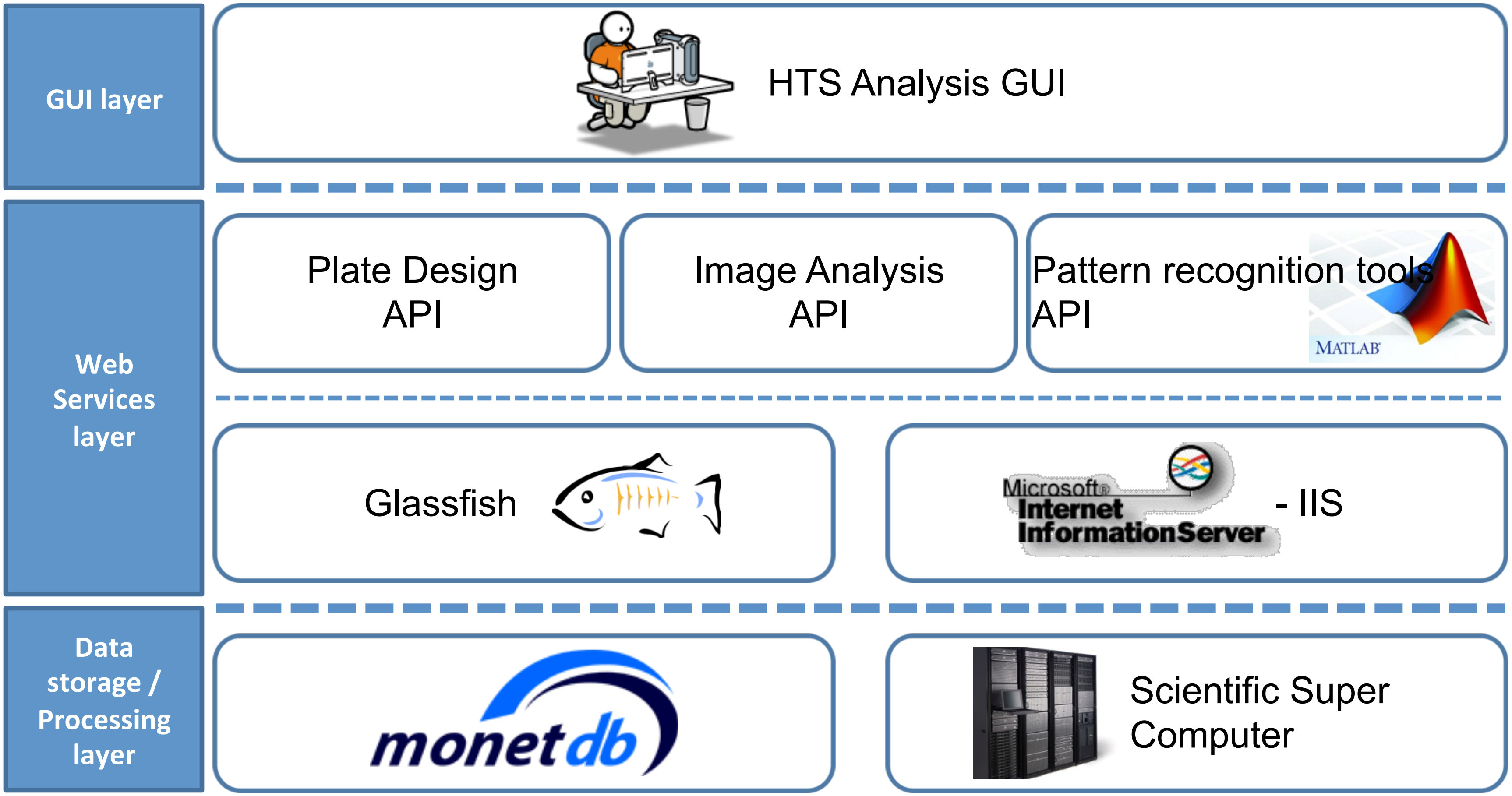
Image

Wet-lab experiment using plate design

Plate

High-throughput Screen (HTS)

# Description of Solution



GUI layer

HTS Analysis GUI

Web Services layer

Plate Design API

Image Analysis API

Pattern recognition tools API

Glassfish

- IIS

Data storage / Processing layer

Scientific Super Computer

Redes Corporativas y NSM

Licencias de SW y HW

# Specific Projects

- API from database to Cluster
  - API Architecture; Requirements
  - Technical Solutions, builds on previous results
  - What is the architecture required from MonetDB to HTS-processing
- Image Processing Software
  - Datastructures / Code optimization
  - Better fit it for cluster/parallel computation
  - Where and how can we get computational benefit

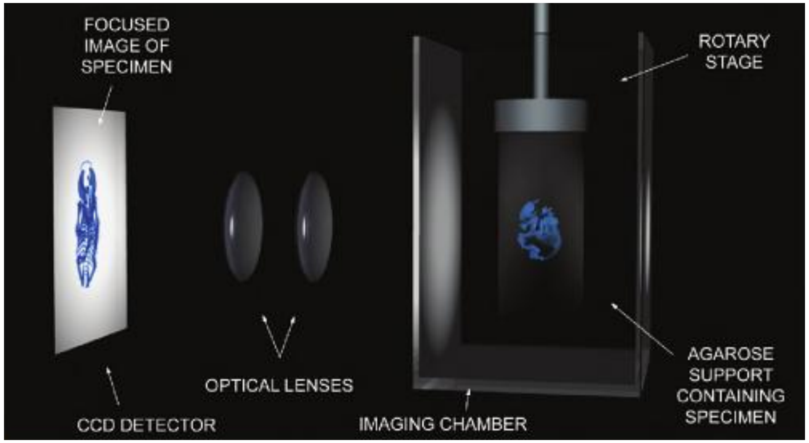
# Project Summary

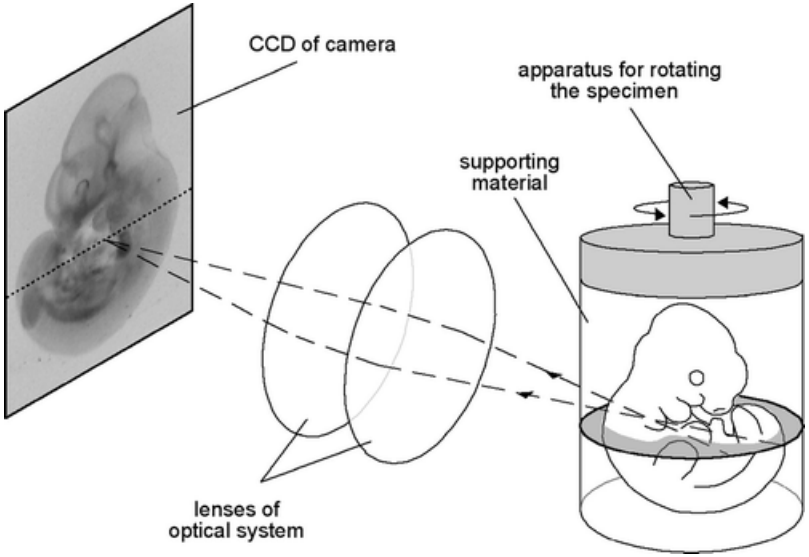
- Analyze software to be ported to the Cluster
- Develop mapping to the cluster architecture
- Apply the results to a dataset
- Generalize
- Collaboration with Kris Rietveld

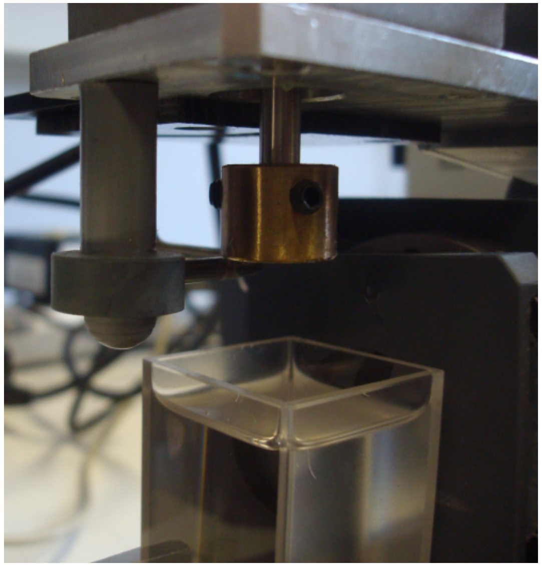
| | |
|---|---|
| – NeCEN: | Java, C, Python |
| – Cell Observatory: | Java |
| – High Troughput: | Java, Python, SQL |

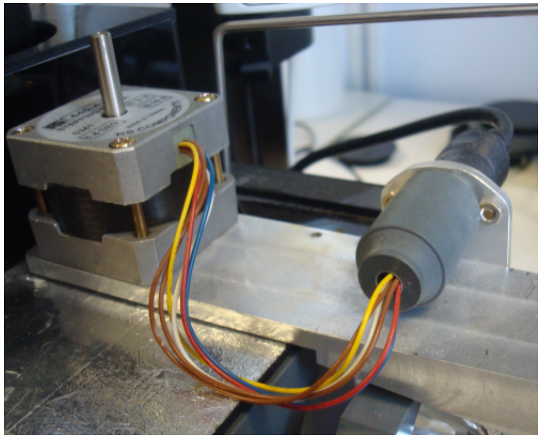
# DATA ACQUISITION

# OPT microscopy


FOCUSED IMAGE OF SPECIMEN
ROTARY STAGE
OPTICAL LENSES
AGAROSE SUPPORT CONTAINING SPECIMEN
CCD DETECTOR
IMAGING CHAMBER

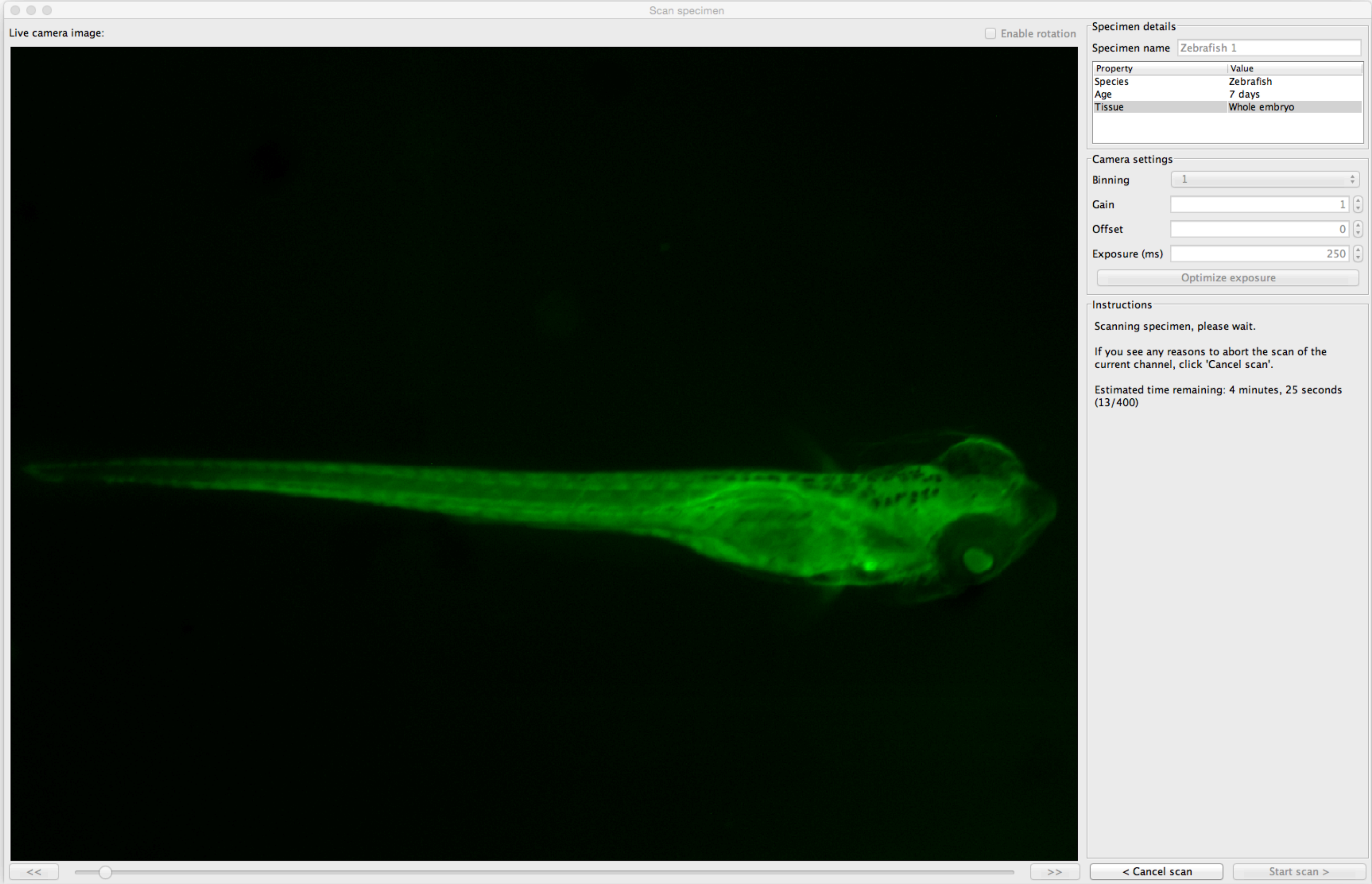






- Object Clearance

# OPT Software - Method



Specimen details

Specimen name: Zebrafish 1

| Property | Value |
|---|---|
| Species | Zebrafish |
| Age | 7 days |
| Tissue | Whole embryo |

Camera settings

| | |
|---|---|
| Binning | 1 |
| Gain | 1 |
| Offset | 0 |
| Exposure (ms) | 250 |

Instructions

Scanning specimen, please wait.

If you see any reasons to abort the scan of the current channel, click 'Cancel scan'.

Estimated time remaining: 4 minutes, 25 seconds (13/400)

# Project Summary

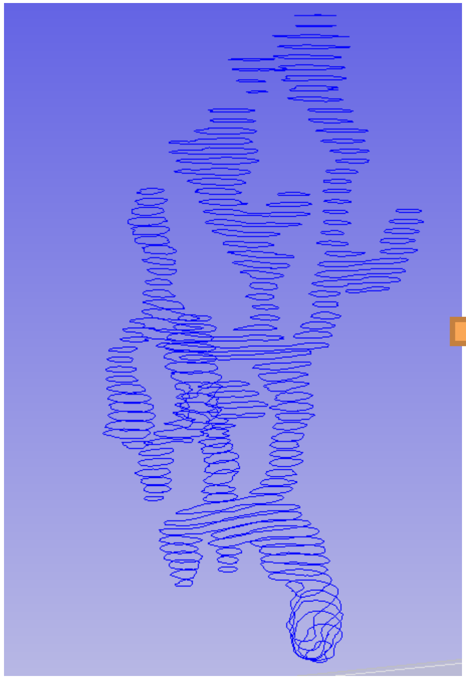
- Calibration system optimization
- Develop Back Projection of tomogram to 3D volume = inverse Radon Transform
- Investigate parallelisation
- Implement on Cluster
- Workflow optimazation for the imaging

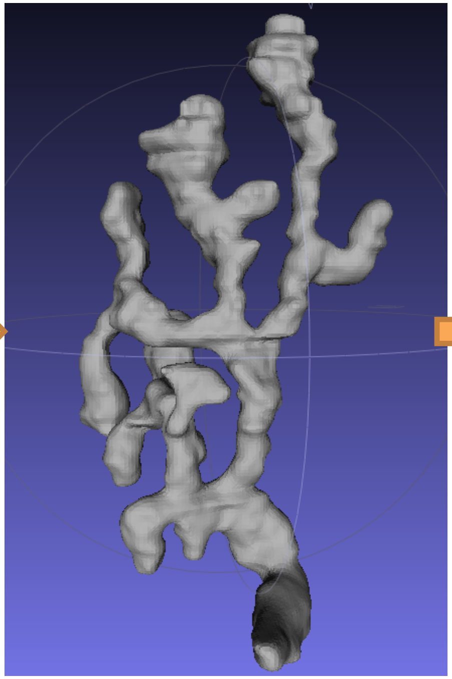
- Project with Sander Hille (Mathematics)
- Dubble Bachelor team Math-CS

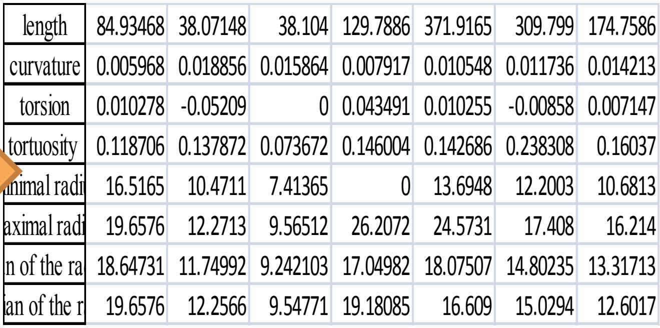
# ANALYSIS & VISUALIZATION

# 3D Model optimization

- We start from 3D models obtained from various sources
- Measurement

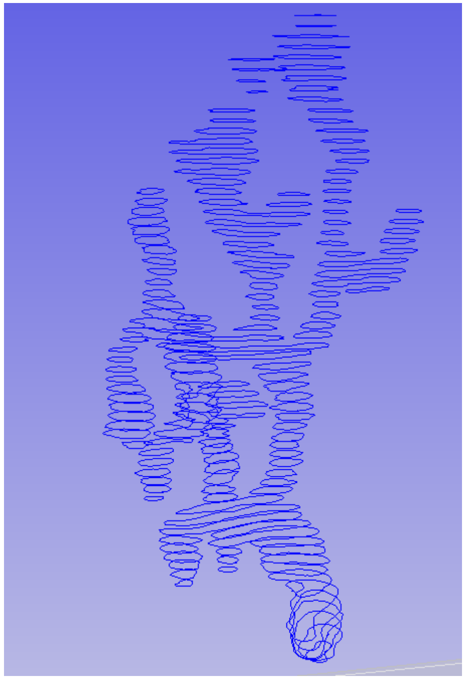






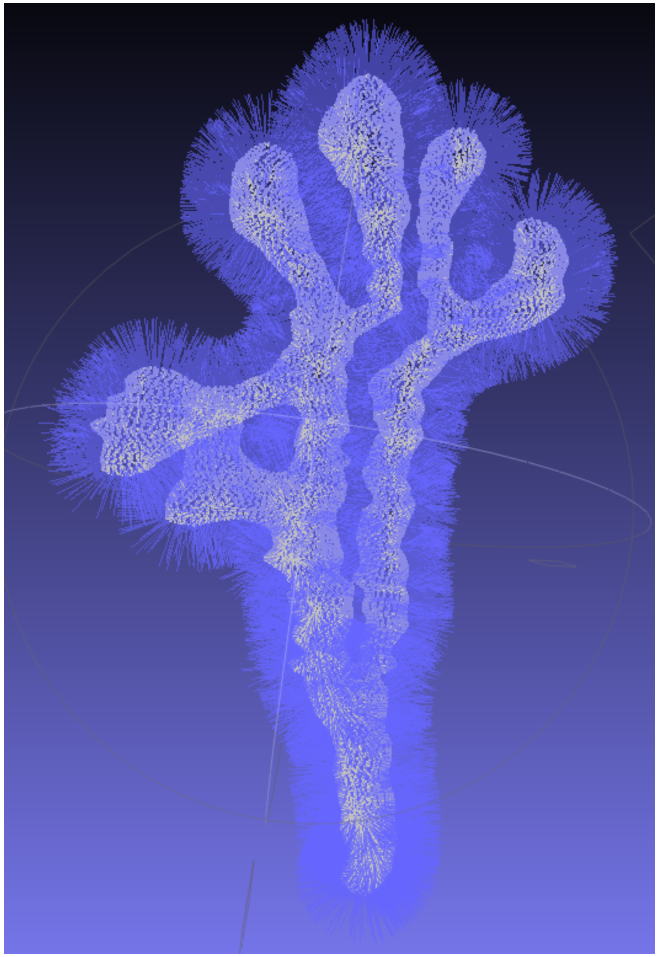
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| length | 84.93468 | 38.07148 | 38.104 | 129.7886 | 371.9165 | 309.799 | 174.7586 |
| curvature | 0.005968 | 0.018856 | 0.015864 | 0.007917 | 0.010548 | 0.011736 | 0.014213 |
| torsion | 0.010278 | -0.05209 | 0 | 0.043491 | 0.010255 | -0.00858 | 0.007147 |
| tortuosity | 0.118706 | 0.137872 | 0.073672 | 0.146004 | 0.142686 | 0.238308 | 0.16037 |
| nimal radi | 16.5165 | 10.4711 | 7.41365 | 0 | 13.6948 | 12.2003 | 10.6813 |
| aximal radi | 19.6576 | 12.2713 | 9.56512 | 26.2072 | 24.5731 | 17.408 | 16.214 |
| n of the ra | 18.64731 | 11.74992 | 9.242103 | 17.04982 | 18.07507 | 14.80235 | 13.31713 |
| an of the r | 19.6576 | 12.2566 | 9.54771 | 19.18085 | 16.609 | 15.0294 | 12.6017 |

# Surface Reconstruction

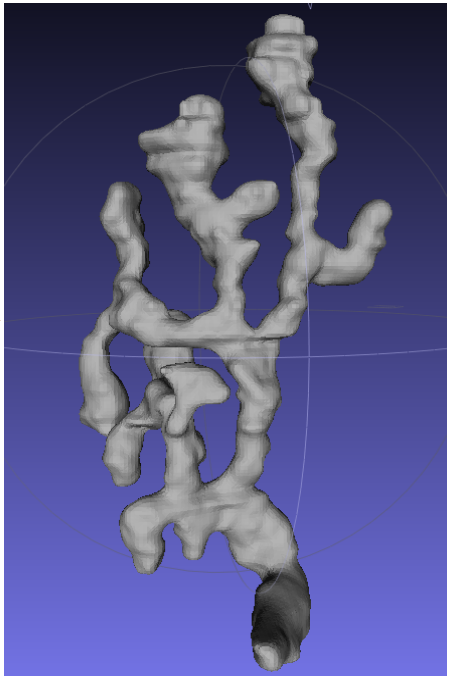


Contour representation

Poisson Reconstruction



Point Cloud representation



Surface  representation

# Connecting Components

- Read files
- Reconstruction & Optimization
  - Poisson reconstruction
  - (L.Cao & FJ Verbeek, Electronic Imaging 2013)
- Visualization
  - VTK (visualization toolkit)
  - Geometrical data-structures
- GUI components

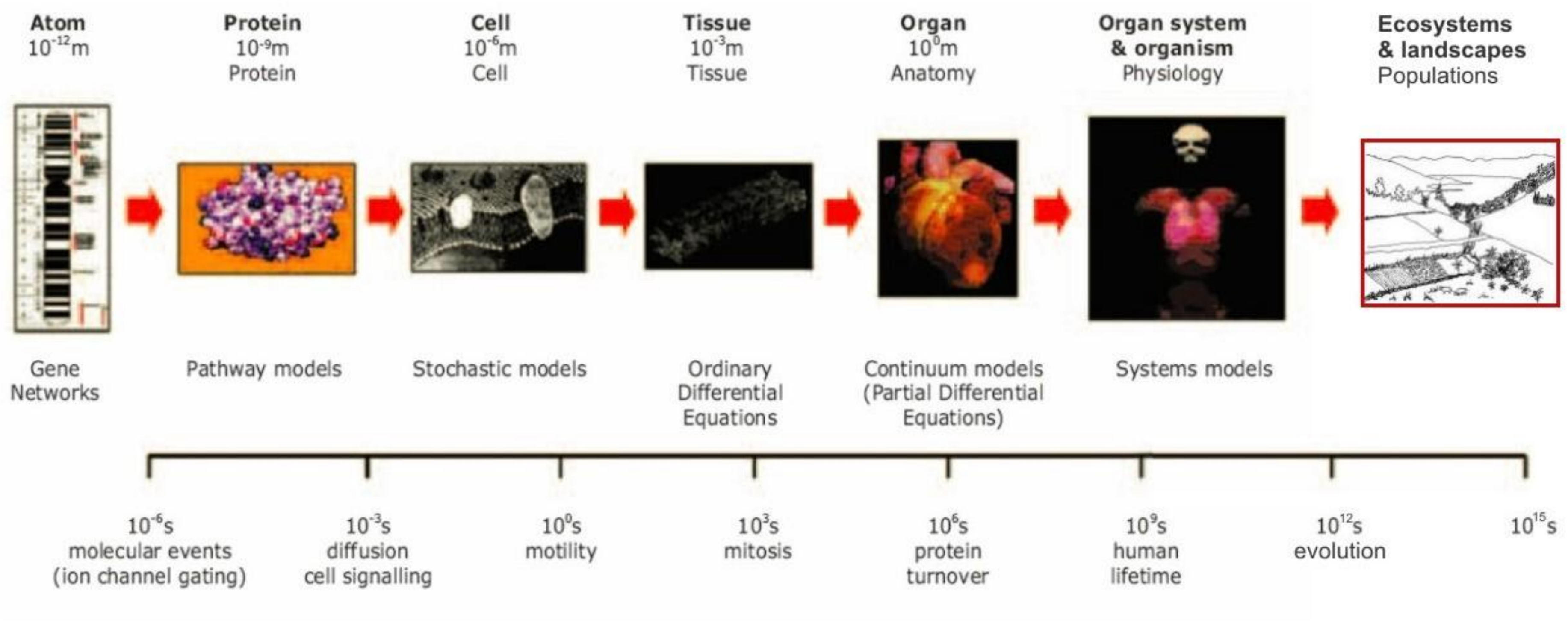
# Project Summary

- Analyze components
- Develop infrastructure that fits workflow for 3D modelling
- Develop interface
- Connect components

- C, C++, QT, VTK

# SOFTWARE AGENTS

# Modeling Different Levels of Biology



# Intelligent Software Agents

- Images are found on different scales in biology
- These scales need to be connected

- Force the connection with software agent
- Annotate large collections
- Integrate in existing database (Cyttron)

- Test and develop agent
- From prototype increase intelligence.

# Contacts

- Fons Verbeek
  - f.j.verbeek@liacs.leidenuniv.nl

- Katy Wolstencroft
  - k.j.wolstencroft@liacs.leidenuniv.nl

- Sacha Gultyaev
  - A.P.Gultyaev@liacs.leidenuniv.nl

URL

http://bio-imaging.liacs.nl/projects