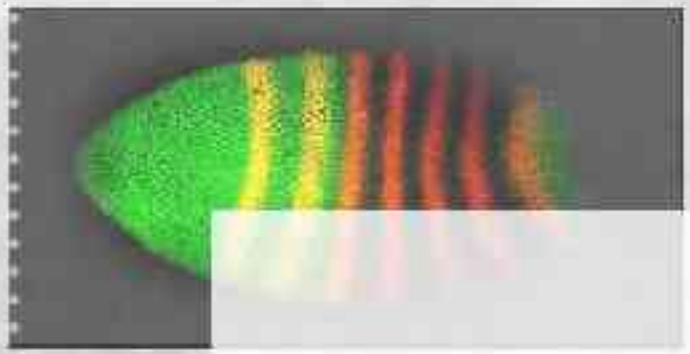


On joint research projects

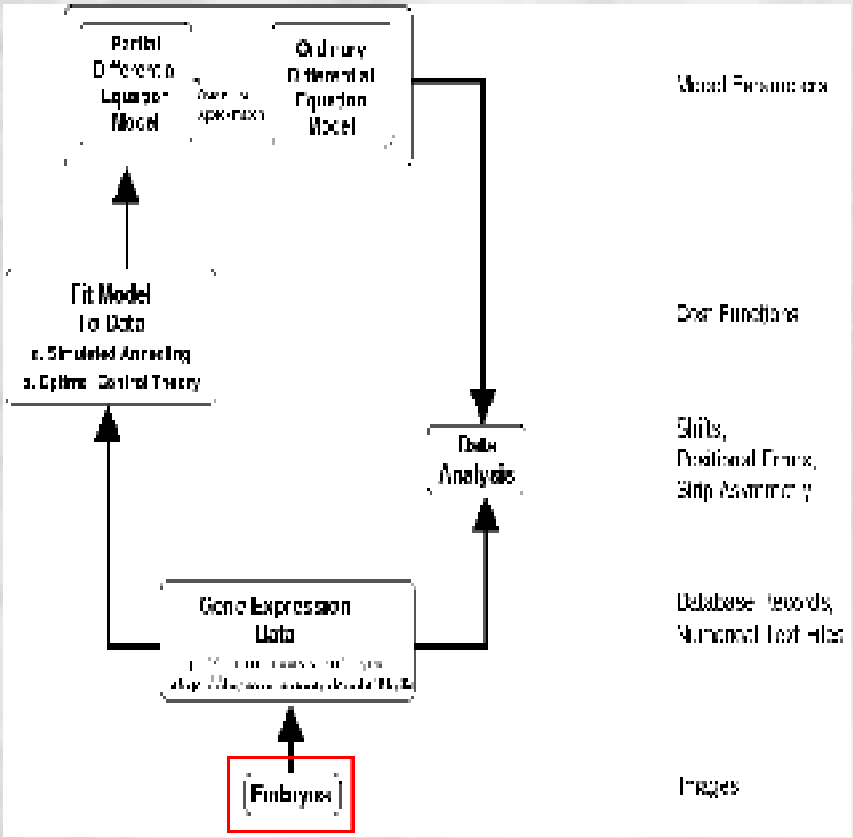
Maria Samsonova

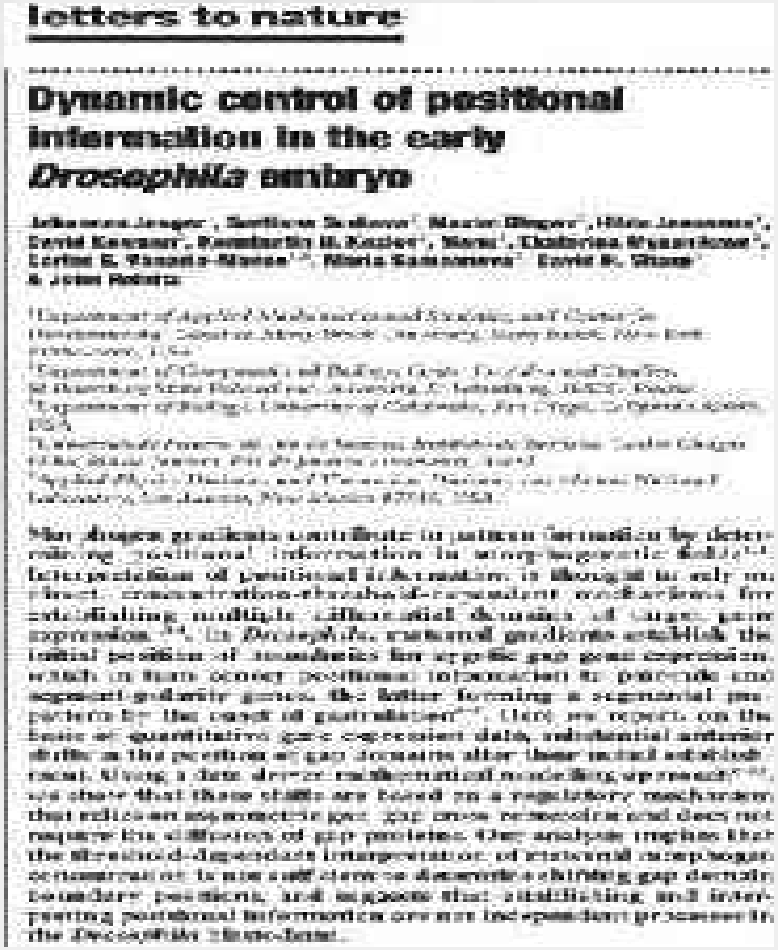
St.Petersburg State Polytechnical University, Russia

Research focus

- methods and tools for acquisition of high-precision data;
- mathematical methods for data mining and system modeling;
- methods for data integration and information extraction.

Systems biology of segmentation in *Drosophila*

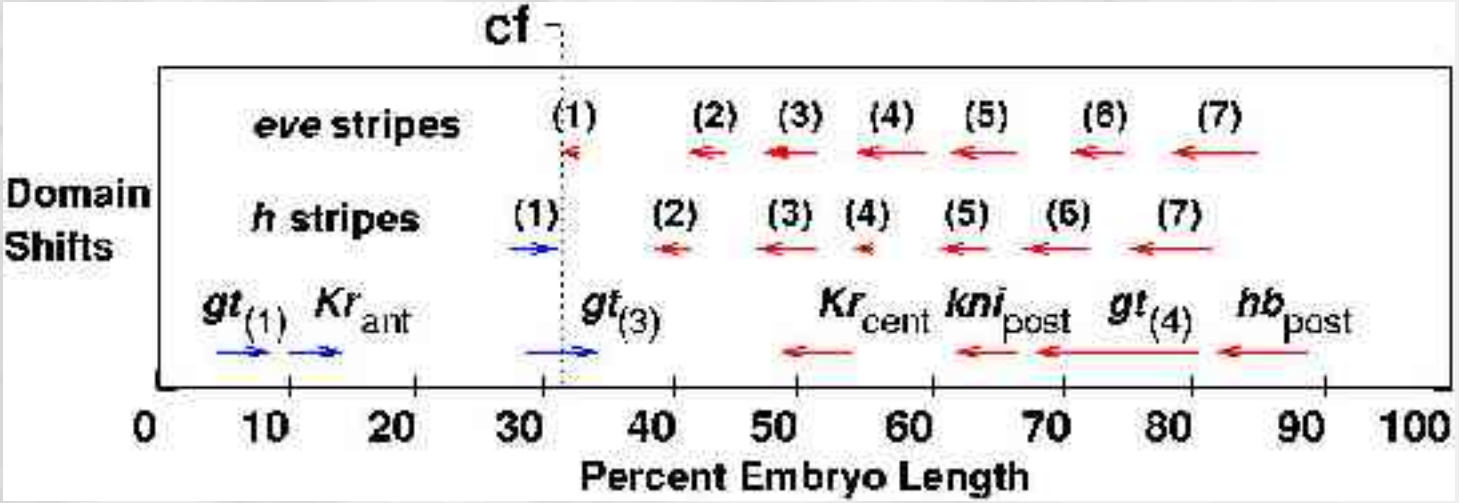



letters to nature

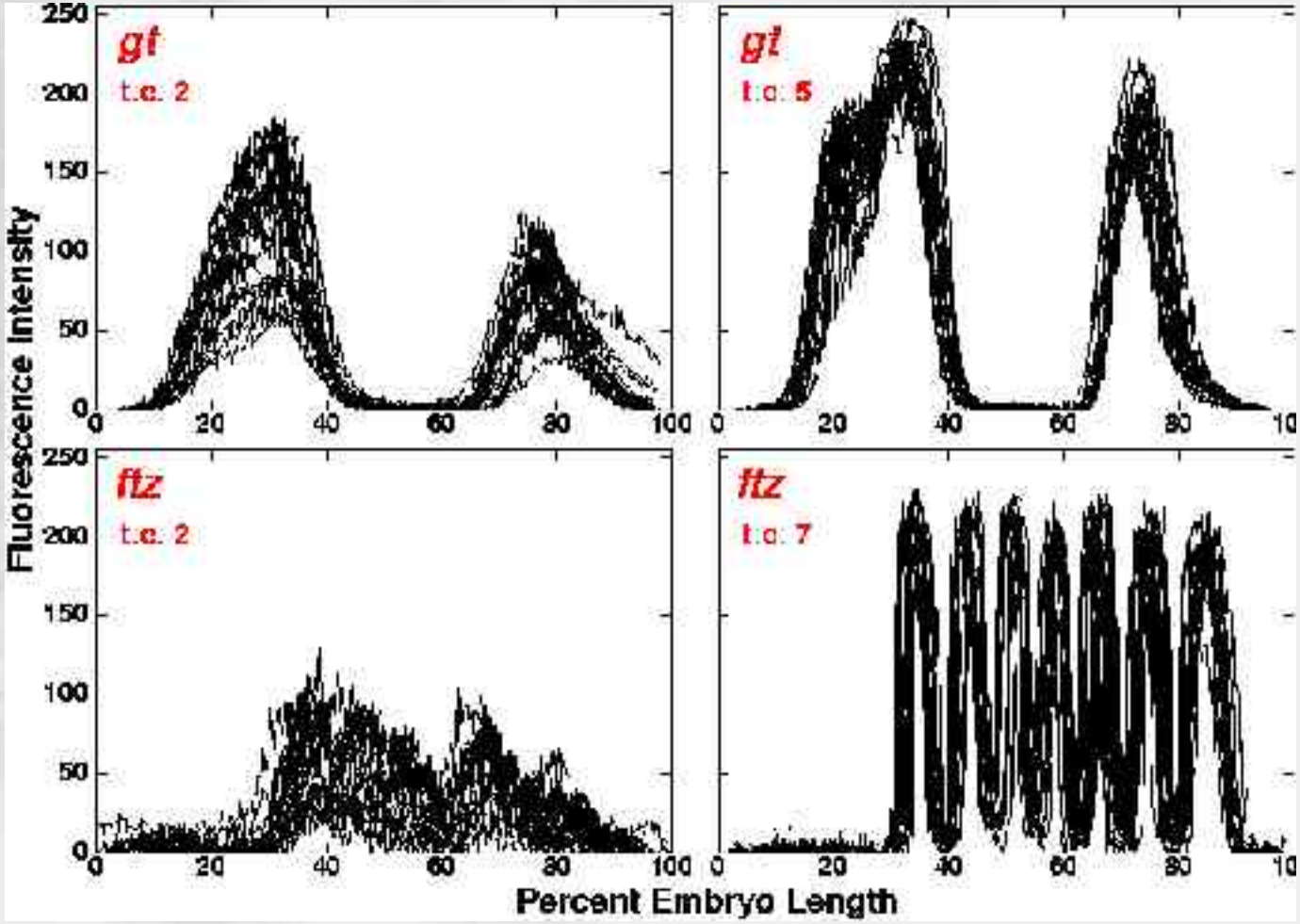
Dynamic control of positional information in the early *Drosophila* embryo

Nature, v 430, 15 July 2004: 368-371

Domains of gap and pair-rule gene expression in head and trunk regions move in opposite directions with time



Variability in expression of zygotic segmentation genes



Synthesis

$$\frac{dV_i}{dt} = \sum_{j=1}^{N} \left(\text{[illegible]} \right)$$

Transport

$$\text{[illegible]}$$

Decay

$$-\lambda_a V_i$$

Synthesis

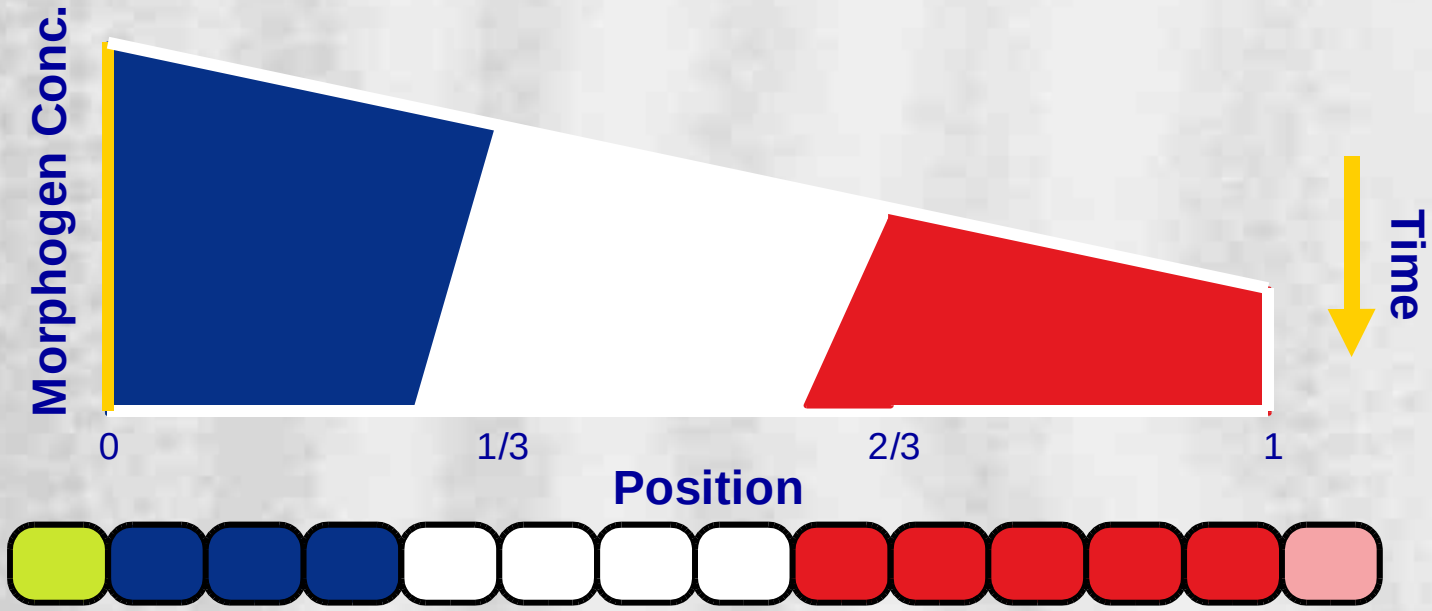
Dynamic dissection:

We can look at individual parts of this sum to ‘dissect’ the various regulatory contributions on a specific gene.

For example:

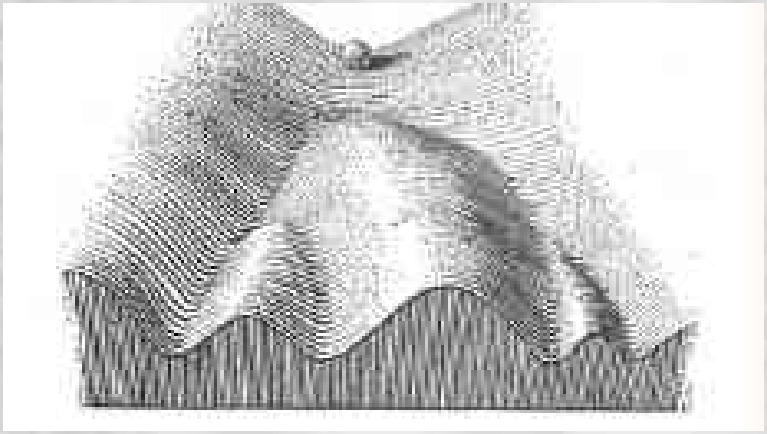
$T^{Kr->hb}\ v^{Kr}$ represents Kr’s regulatory input on *hb*

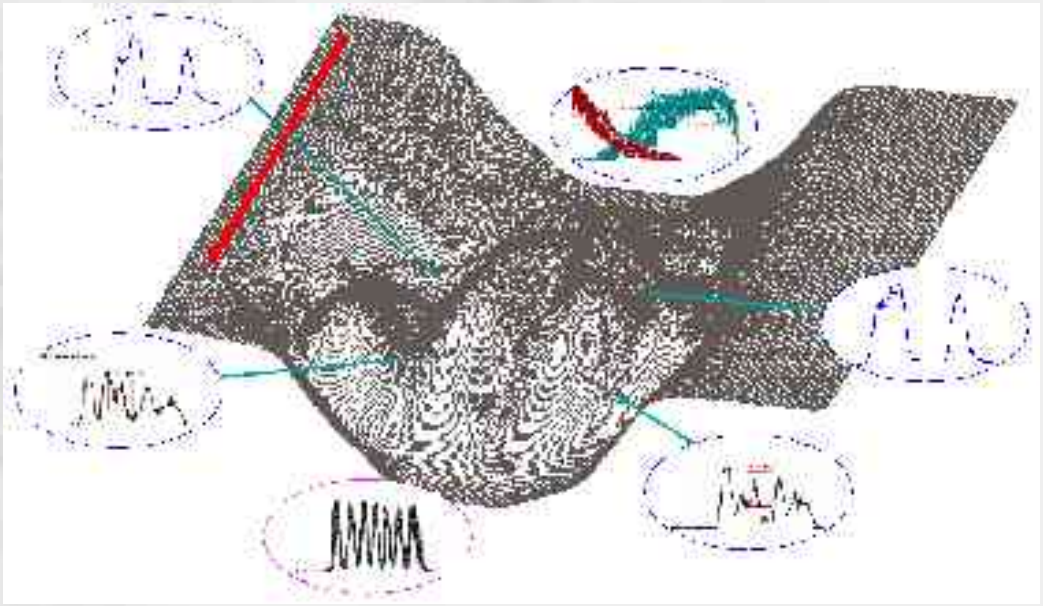
Conclusions: The French Flag Revisited



- **Posterior gap domains shift because of regulative cross-interactions.**
- **Positional information in early embryo is dynamic and can no longer be seen as a static coordinate system imposed on embryo by maternal genes.**

Conclusions:
Waddington's concept of epigenetic landscape can be adopted to explain the pattern formation phenomenon (Cont.)





- ***All kinds of variability inherent to expression patterns of segmentation genes are significantly decreased by gastrulation.***

Mathematical modeling and statistical analysis of data

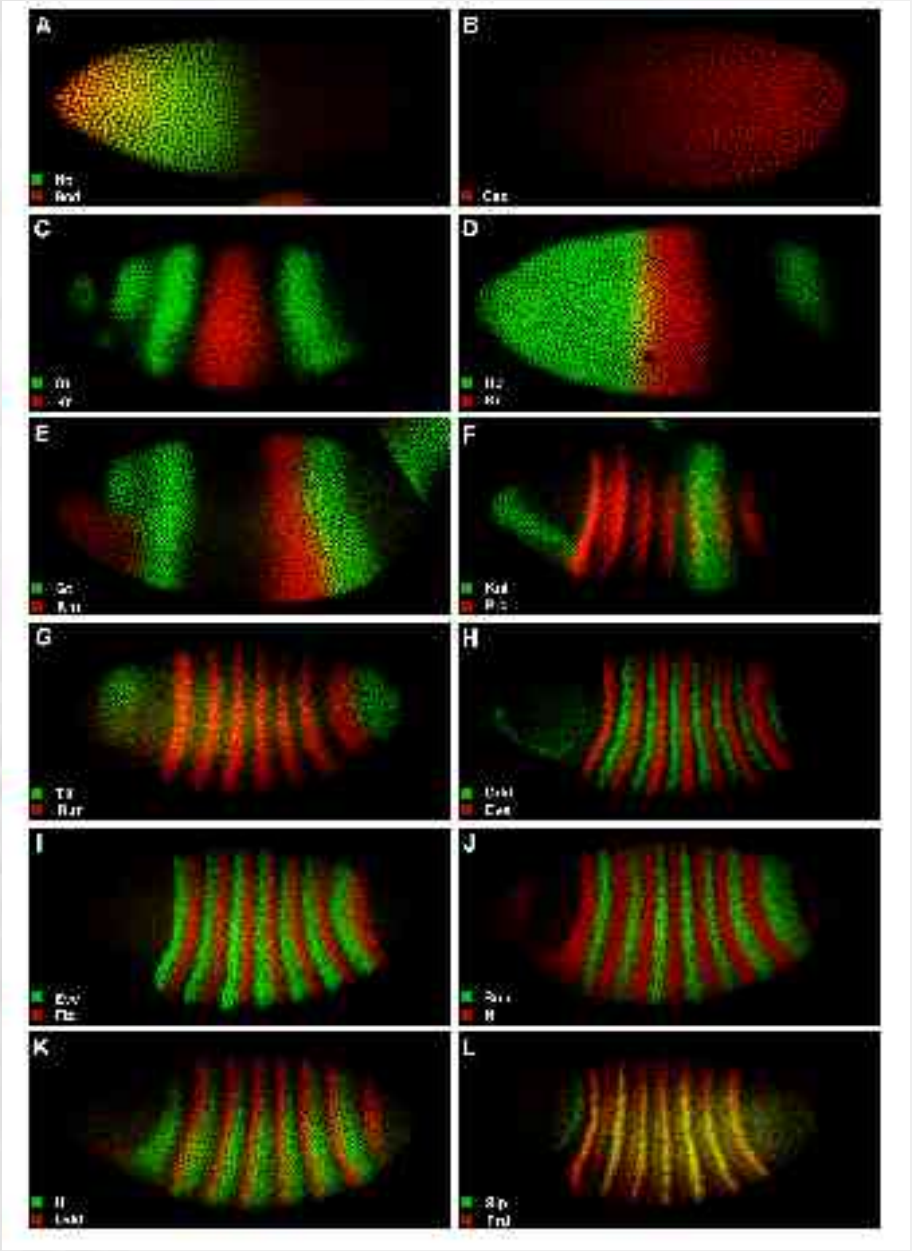
- Various statistical and machine learning methods are in use for data mining;
- Models based on ordinary and partial differential reaction-diffusion equations;
- Different optimization methods are available to find the system parameters:
  - Simulated Annealing
  - Optimal Steepest Descent Algorithm
  - Tunneling
  - Parallel Differential Evolution.

Open questions

- Construction of models of biological systems and processes which operate on different time scales and/or complexity levels
- Approaches towards validation of existing models;
- Adequate representation of diffusion in models;
- Logics behind selection of relevant models to uncover the structure and dynamics of a particular biological system.

Acquisition of high-precision quantitative data: confocal scans of gene expression patterns

Myasnikova et al (2001), Bioinformatics 17:3-12

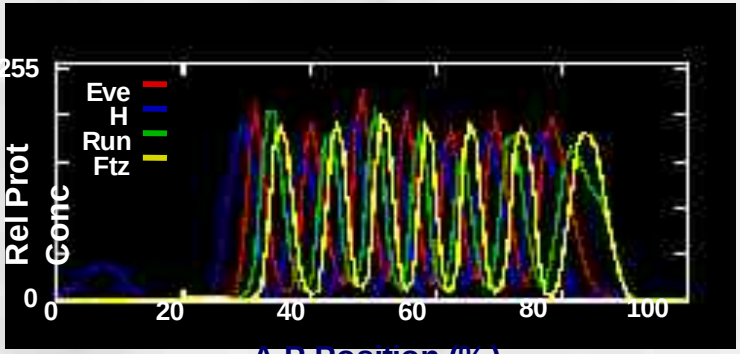


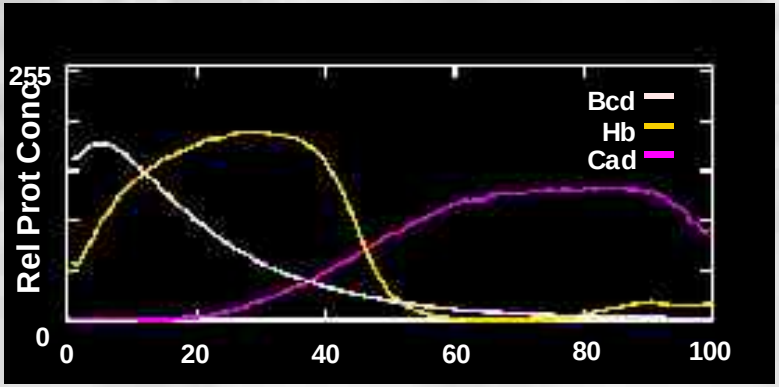
Data Pipeline

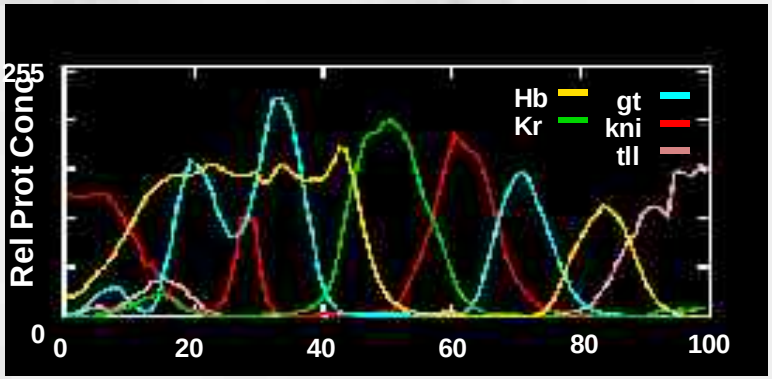
1. Image segmentation.
2. Remove the background.
3. Temporal characterization.
4. Register the data.
5. Average the data.

Quantitative data on segmentation gene expression with cellular resolution in space and temporal resolution of 6.5 minutes of development

Nuclear Coordinates		Fluorescence intensities		
0	3.67346	41.5401	53.6951	25.939
1	4.05099	45.3917	53.8919	27.0811
2	4.00196	37.3767	54.7	22.85
3	5.01298	39.8973	59.4688	26.5625
4	5.07766	52.1475	58.8267	29.04
5	5.44521	43.4772	62.25	30.5125
6	5.52642	47.3609	56.7215	28.9494
7	5.78564	36.9171	56.8689	26.2623
8	6.26299	50.786	64.3521	30.6056
9	6.62735	40.1422	59.3086	32.0123
10	6.69946	54.7947	60.0795	29.5227







FlyEx database http://urchin.spbcas.ru/flyex; h**ttp://flyex.ams.sunysb.edu/flyex**

Why the effective image processing packages are necessary to develop?

- we need high-precision quantitative data with high spatial resolution;
- different sophisticated microscope techniques are currently available;
- software packages provided by microscope manufactures are designed to produce a high-quality image. Usually do not support effective processing and analysis of images in batch.

Pro-stack package

("Prostack" means "a straight man" in Russian ;-))

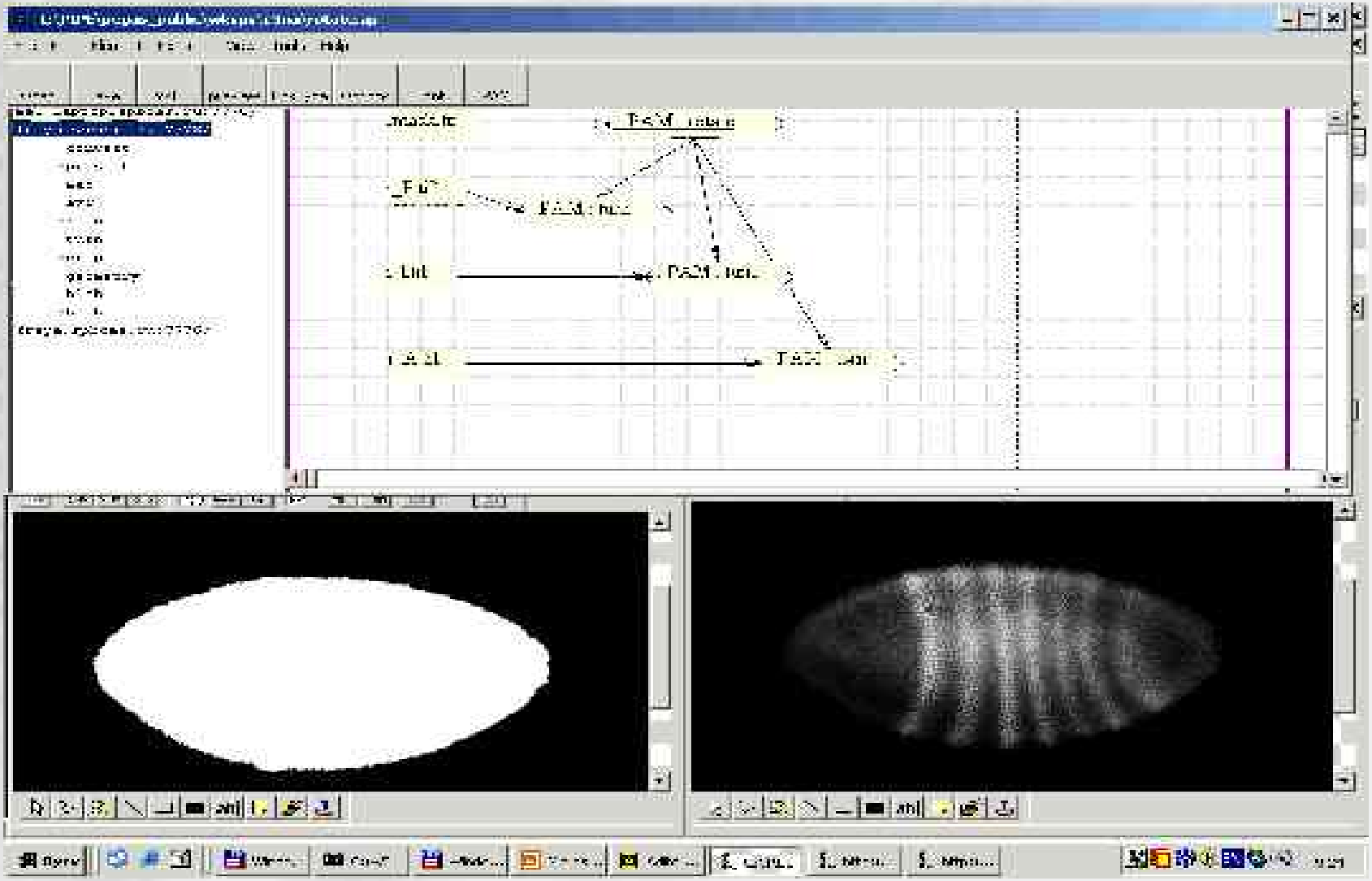
Functionallity:

- ***Pro***cessing of ***stack***s of images,
- More than 40 image processing methods, from thresholding to the calculation of object characteristics.
- The number of methods is growing.

Technical Data:

- Implemented in ANSI C as methods library libparus and command line interface prostack.
- All methods are available in distributed computing environment iSIMBioS via wrappers written in Perl.
- Can work with images in TIFF format with 8 bits per pixel using libtiff library for I/O operations.

Image processing scenario



Data integration

- We have developed a method for integration of databases with common subject domain (http://urchin.spbcas.ru/NLP.htm).
- Now we propose to extend our approach to design an information management system for collaboration within distributed working environment.

Key Components of the method for integration of databases with common subject domain

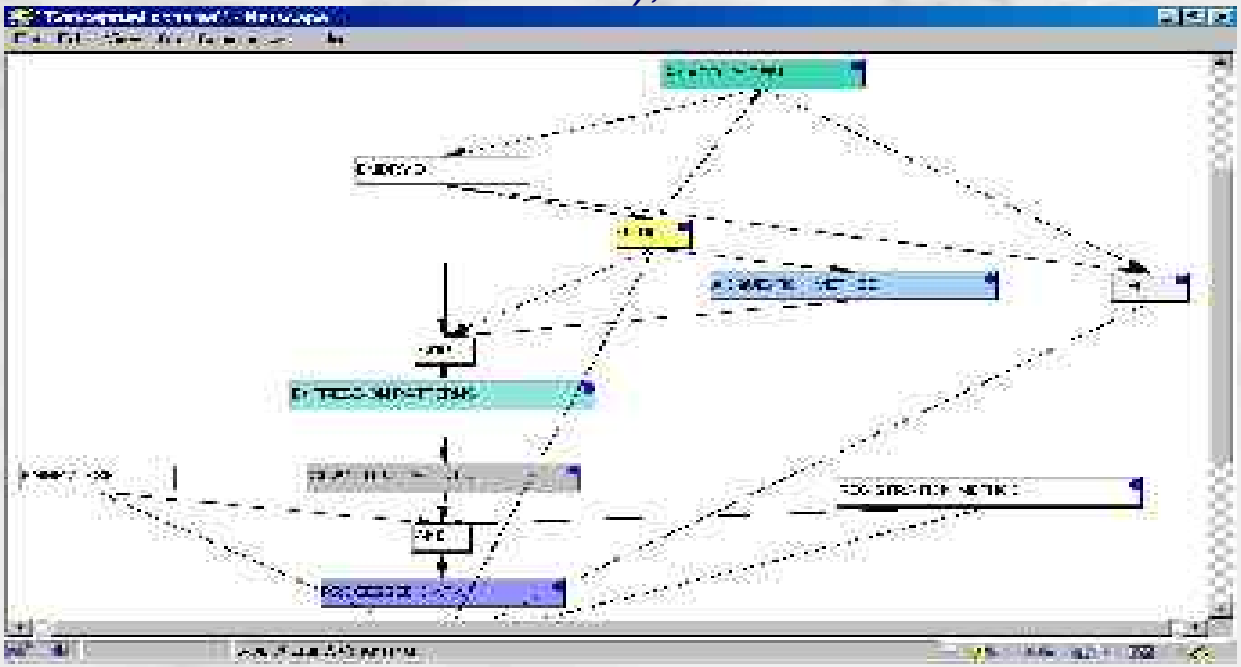
1. Conceptual scheme of knowledge domain and domain oriented dictionaries.

3. Processor of natural language queries to a database.

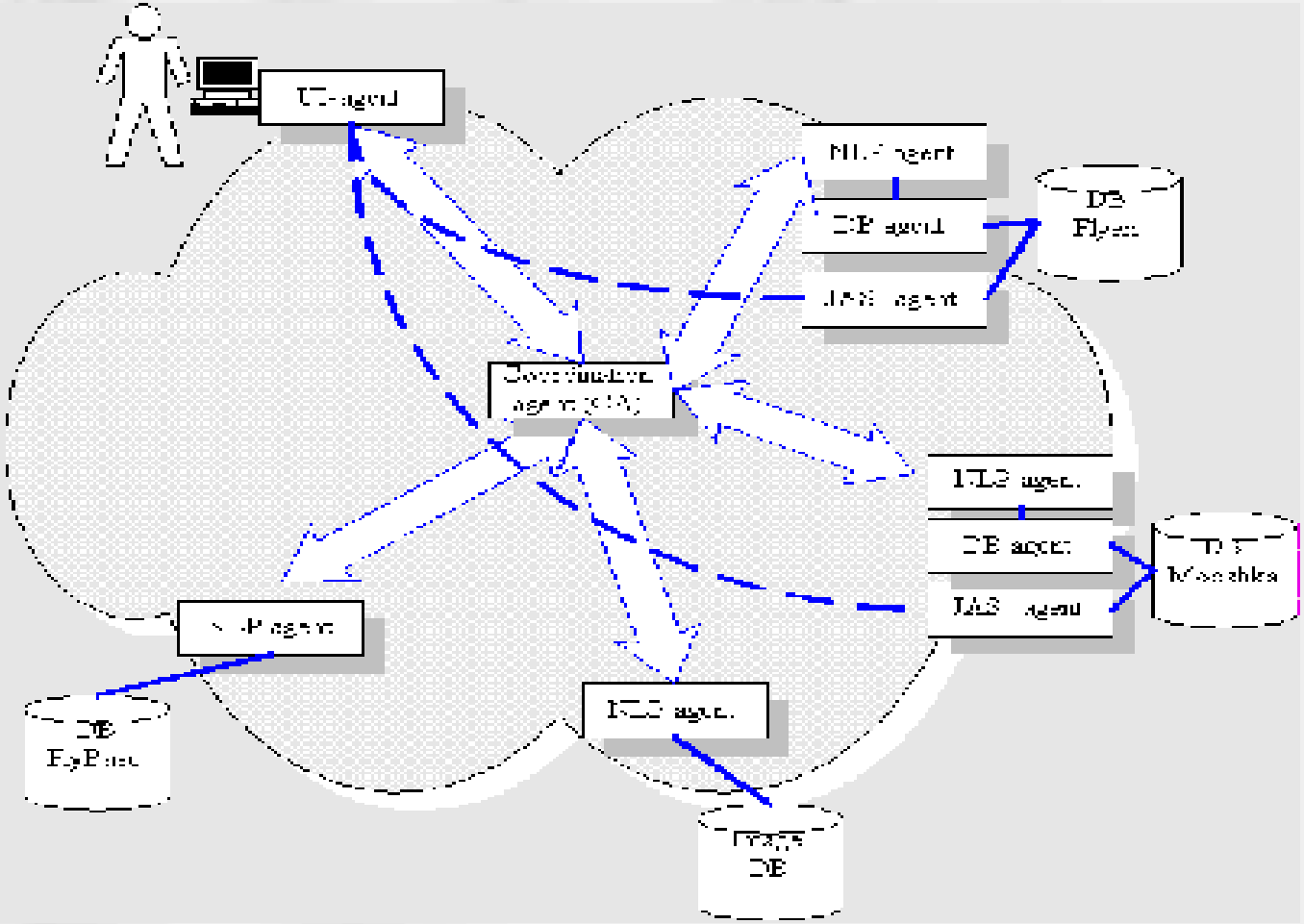
5. Multiagent architecture to integrate results of information retrieval from different databases.

Conceptual Scheme

"is characterized by"("EMBRYO","DEVELOPMENTAL TIME")
"is an instance of"("TIME FROM ONSET OF CYCLE 14A", "DEVELOPMENTAL TIME")
"is derived from"("TIME FROM ONSET OF CYCLE 14A", "INVAGINATING MEMBRANE"), etc.



System architecture



iSIMBioS

- ✓ fast and convenient access to data;
- ✓ data can be processed and analyzed by combining programs and services (modules) into workflows;
- ✓ workflow modules and data can be distributed over network;
- ✓ workflows can be constructed visually;
- ✓ simultaneous access of multiple users to shared data and methods;
- ✓ extendable, scalable and flexible in specification and modification of analysis methods;
- ✓ failure resistant, portable;
- ✓ provides access through firewall and proxy servers;
- ✓ dissemination of data and programs.

To appear in U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 204–215, 2006.

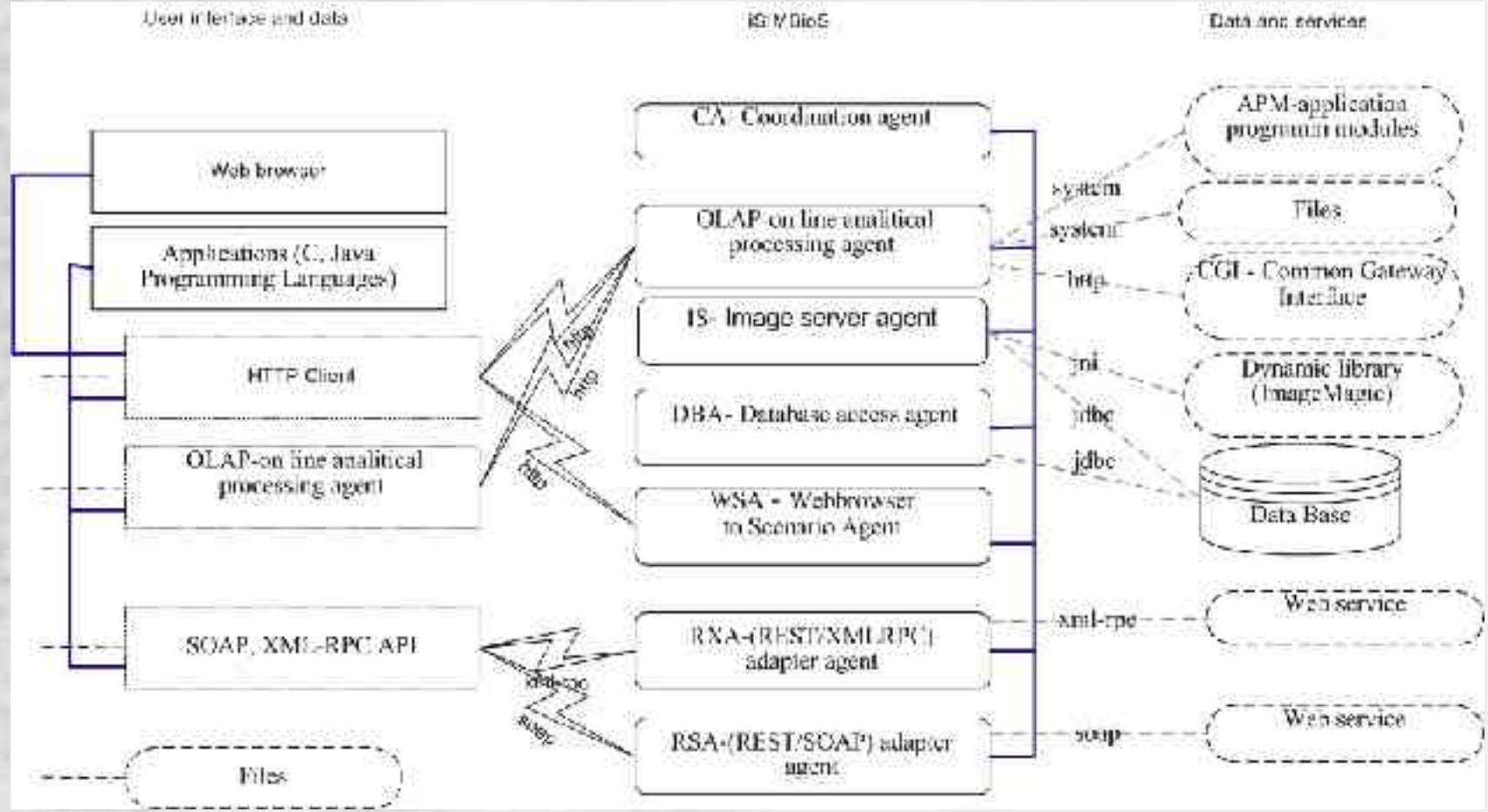
iSIMBioS store

- All data types.
- Information about
  - modules, workflows, workflow enactments,
  - system architecture,
  - distribution of data and programs among different servers,
  - users and user groups.

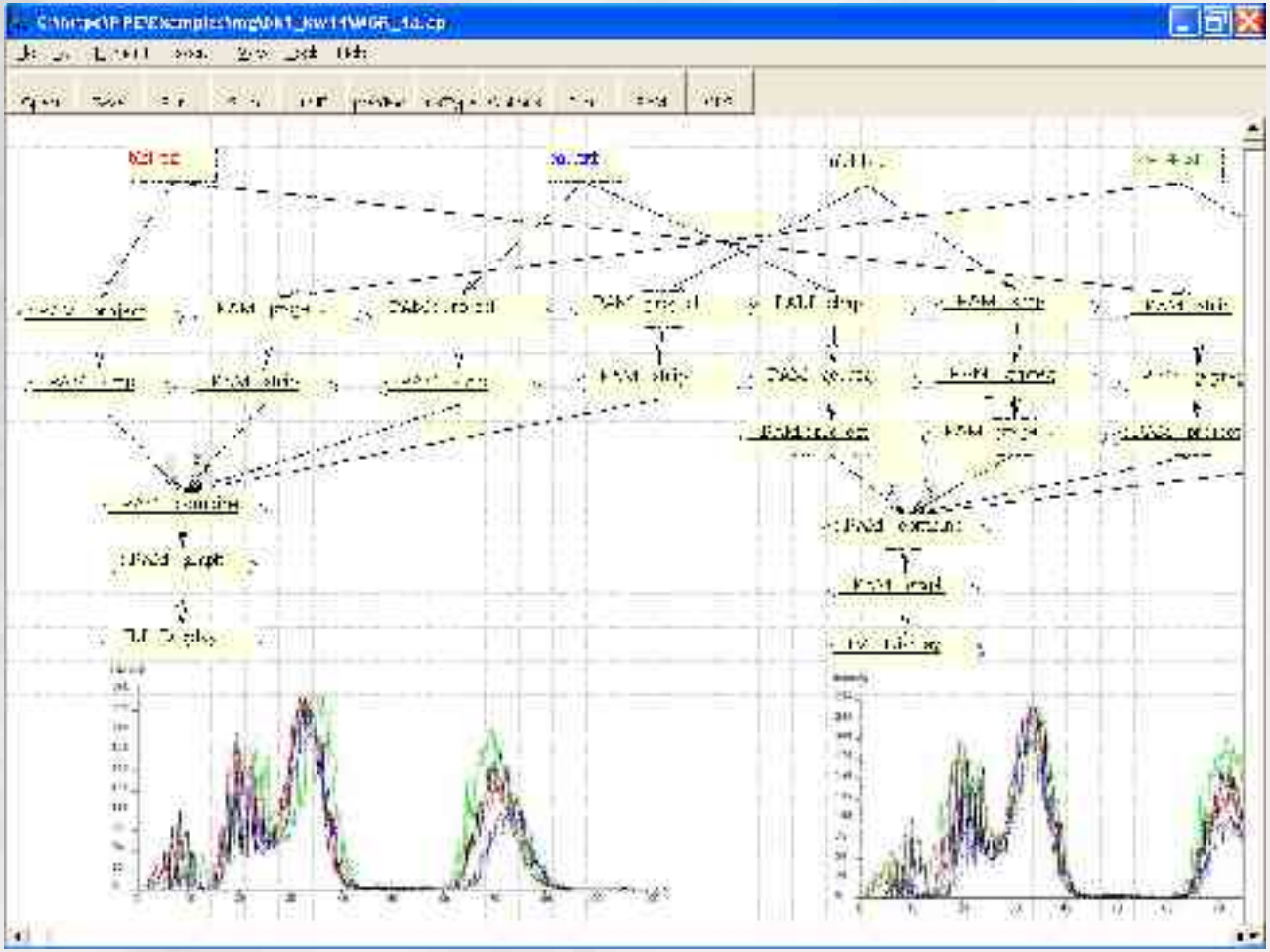
User policy - restricted access to private data.

Database structures can be extended dynamically, this makes storage independent from knowledge domain, i.e. allows to store new, not known in advance data.

iSIMBioS architecture



iSIMBioS workflow: estimation of the registration accuracy



Information extraction

- We have developed a method for processing of natural language queries to a relational database (*Samsonova et al. (2003) Bioinformatics 19, suppl. 1, i241-249*).

- Now we propose to use NLP for information extraction from biomedical literature.

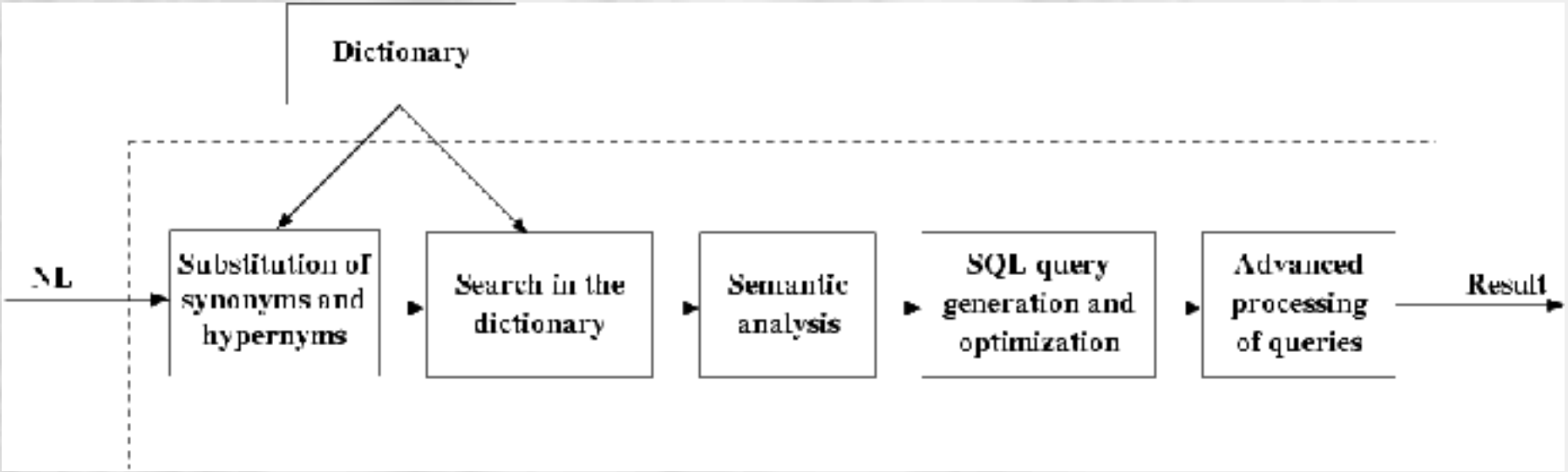
FlyEx database

Contains

- Images of segmentation gene expression in individual embryos.
- Quantitative data on segmentation gene expression in each nucleus of an individual embryo.
- Averaged data on expression of each segmentation gene at each time point and at cellular resolution.

Natural language interface is available at
http://urchin.spbcas.ru/NLP/NLP.html

Main Steps of Processing of a Query in NL



Russian Version of the Query Form

Query can be formulated as a List of Keywords

`embryo Kr gt '* or *`Which embryos were scanned for expression of Kruppel and giant?'

FlyEx Database

Natural language query (words used to retrieve the information from the database are shown in red color)

embryo Kr gt

Query result

Embryo name	Cleavage cycle	Temporal class	Gene 1	Gene 2	Gene 3
bk1	14	3.00	even-skipped	Kruppel	giant
bk10	14	7.00	even-skipped	Kruppel	giant
bk11	14	8.00	even-skipped	Kruppel	giant
bk12	14	9.00	even-skipped	Kruppel	giant
bk13	14	10.00	even-skipped	Kruppel	giant
bk14	14	4.00	even-skipped	Kruppel	giant
bk15	14	5.00	even-skipped	Kruppel	giant
bk2	14	4.00	even-skipped	Kruppel	giant

Information extraction

- In collaboration with Prof. Rubashkin (St.Petersburg State University) and Prof. Kolchanov (Institute for Cytology and Genetics of the RAS) groups.
- System prototype to automatically extract information about gene expression in *Arabidopsis thaliana*
  - text analysis in both automatic and interactive modes;
  - interactive training of the system by an expert;
  - use of text recognition and analysis algorithms;
  - relational database to store extracted information.

Acknowledgments

St.Petersburg

Alexander Samsonov
Vitaly Gursky
Konstantin Kozlov
Ekaterina Myasnikova
Andrei Pisarev
Ekaterina Poustelnikova
Svetlana Surkova

Stony Brook

John Reinitz
Jean Cadet
King-Wai Chu
Yuefan Deng
Hilde Janssens
Johannes Jaeger
Manu

San Diego

Dave Kosman

Los Alamos

Dave Sharp

Bielefeld

Ralf Hoffestadt

Novosibirsk

Nikolai Kolchanov.
Nadezhda Omelyanchuk

http://urchin.spbcas.ru/flyex
http://flyex.ams.sunysb.edu/flyex