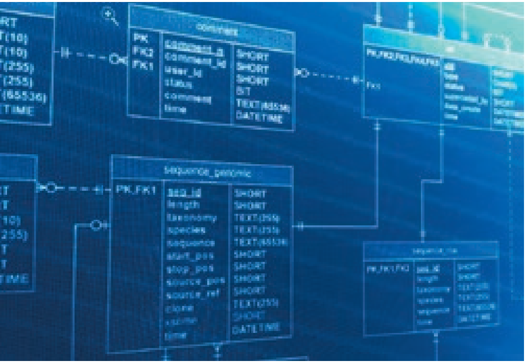
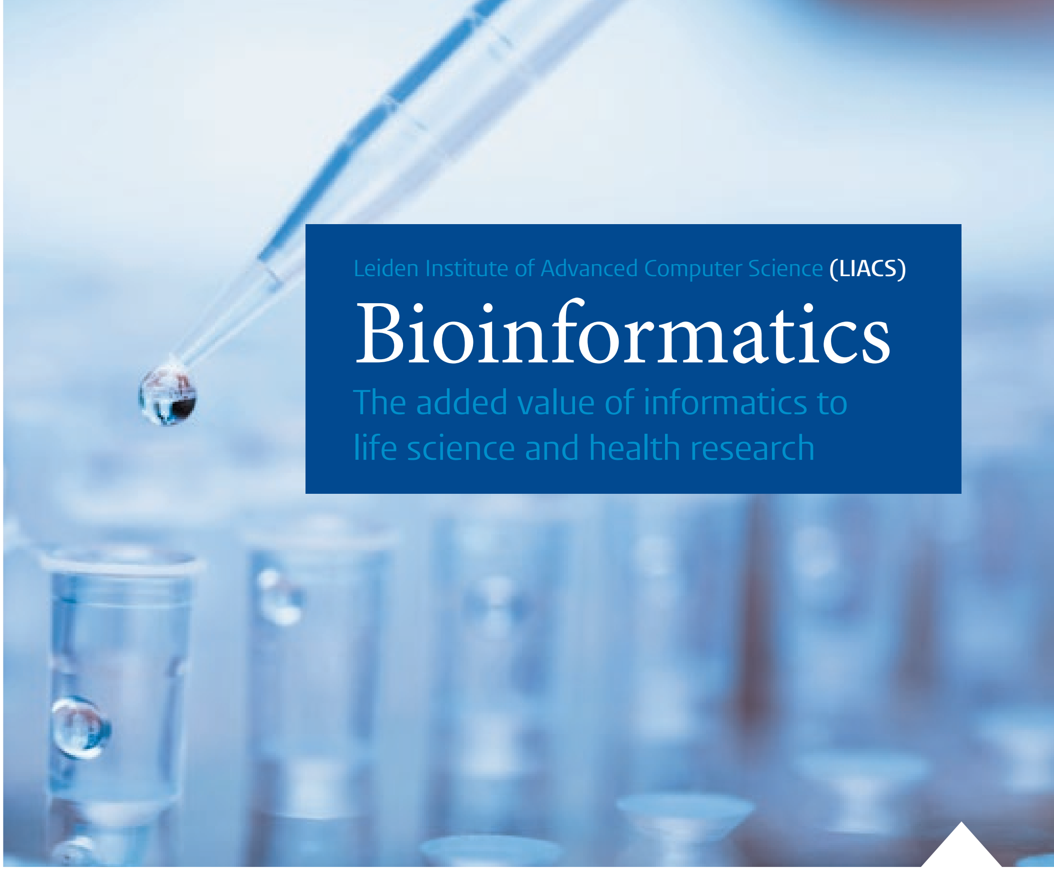




Leiden Institute of Advanced Computer Science (LIACS)

Bioinformatics

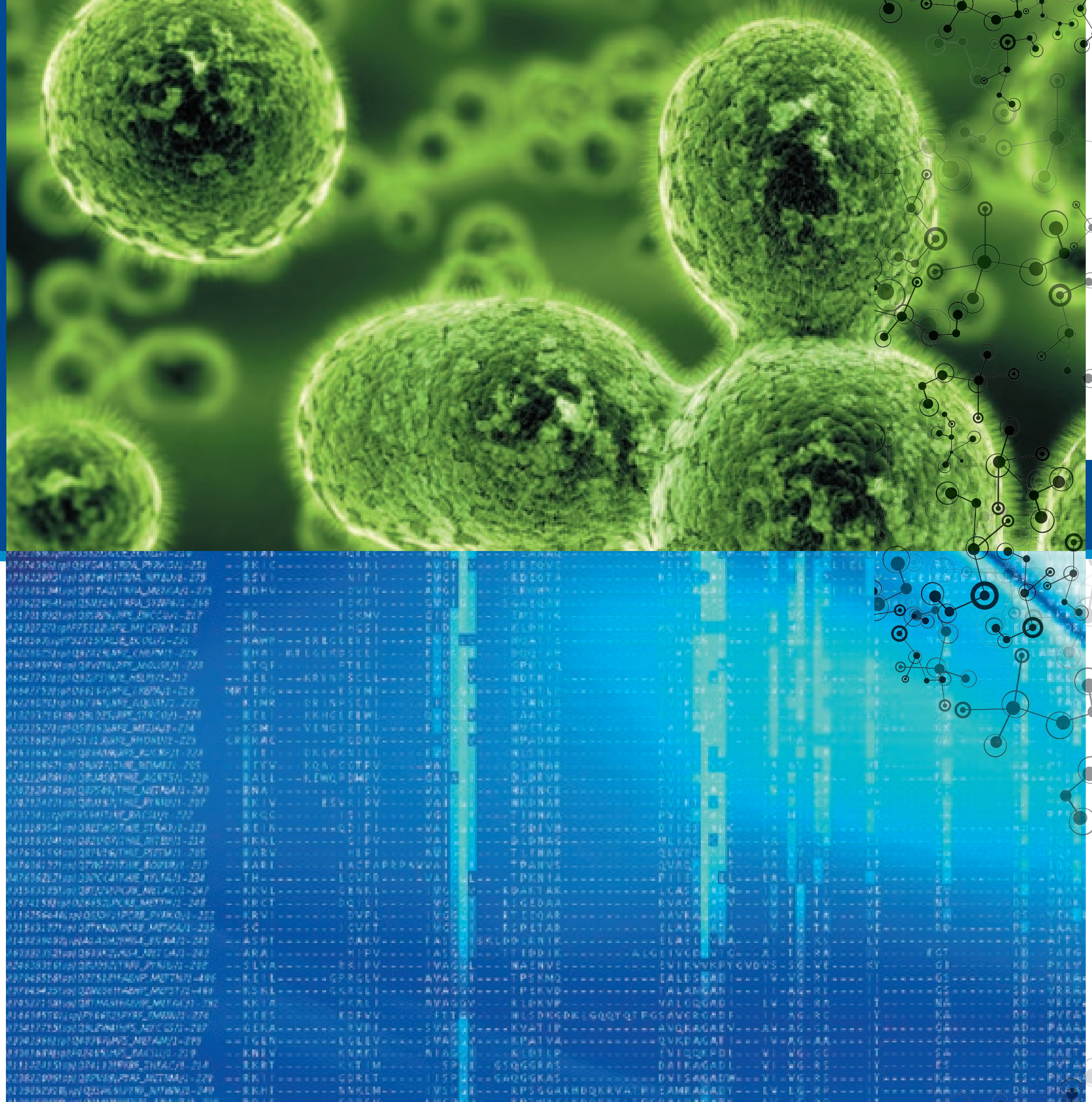
The added value of informatics to life science and health research

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Leiden Institute
of Advanced
Computer Science

Authors: Rolf Spil, Irene Nooren and LIACS staff members
Design: Creja Ontwerpen
Photos: Arthur Koppejan, stock and personal photos/illustrations
Printer: Drukkerij Deltabach
Copyright: Faculty of Science, Leiden University



With the development and improvement of new technologies to analyse biological material and model biology, life science and health research is becoming increasingly data intensive and complex. Technology application areas such as next-generation sequencing, microarray technology, mass spectrometry, imaging, biobanking and systems biology modelling deal with ever-growing amounts of data that need to be stored, managed, processed, analysed and interpreted.

LIACS: connecting life science and informatics

Collaboration between life science and informatics research organisations, such as the Leiden Institute of Advanced Computer Science (LIACS) helps in dealing with the phenomenon of big and complex data. LIACS has both the expertise and the infrastructure to face these challenges.

This brochure presents examples of the research and support that LIACS provides in the area of bioinformatics.

Contact:
requests@liacs.leidenuniv.nl

LIACS Bioinformatics Partnership

Research

LIACS participates in national and international research projects in order to augment the effectiveness of life science research in The Netherlands. These projects are not limited to universities and university hospitals but also embrace industrial companies and non-profit organisations involved in life science research, both in the Netherlands and in other European countries. Most of these research projects, like Commit, DDMoRe, NeCEN and Cyttron, are funded or supported by the European Union or by the Dutch funding agencies such as NWO and STW. Partners have relied on LIACS to find precious information hidden in large datasets, to simulate experiments in silico before performing costly time-consuming experiments, to give direction to subsequent research or to formulate paradigms and solutions that, without the aid of LIACS, would have never been applied or even been considered. Besides life science companies, automotive manufacturers like BMW and Honda benefit from LIACS expertise to overcome developmental issues using techniques derived from genetic evolutionary algorithms.

LIACS can act as a valuable research partner throughout life science projects, or provide information technology (IT) consultancy and support during a specific phase of a project. Partners in bioscientific research are aware that extensive cooperation between scientific institutions, commercial companies and government is indispensable to reach higher levels.

Education

LIACS is not only a source of knowledge for her partners, but is also a resource for skilled computer scientists and bioinformaticians. LIACS offers a Bioinformatics Master track with specialisations in several Bioinformatics areas, as well as PhD programs. Many graduated students have found their way to industrial companies and healthcare institutions. Graduation theses typically involve about nine months of interdisciplinary research, combining computer science and biology. Commonly, a thesis involves the application of bioinformatics methods and tools, such as data mining applied to corresponding data sources, building a database of related biological data, modelling biological systems, or applying medical image analysis tools.

BioIT team

The LIACS BioIT team provides support and expertise to biologists who require help with the analysis and interpretation of their big and complex data, such as those derived from technologies like microarray and next-generation sequencing. The team consists of bioinformatics experts and closely collaborates with the Leiden University Medical Centre (LUMC) and the University of Amsterdam Bioinformatics support teams. LIACS typically runs short projects of less then 3 months to help biologists on their way with the right tools and expertise.

Our expertise includes the design of omics experimentation (i.e. collective characterization and quantification of pools of biological molecules, e.g. genomics, proteomics), statistics, data management, data and image analysis, visualisation and interpretation, as well as developing new methods and tools for life science research applications. Furthermore, LIACS supports the building of data analysis pipelines, and the use of IT infrastructures, with high-performance computing and storage, to enable big data life science research projects. The expertise can be applied to the biomedical domain as well as biodiversity informatics, foods and plant breeding.

The BioIT team was formed from the successful e-BioGrid project (www.e-biogrid.nl), which was initiated to build the Dutch e-science infrastructure for life science research. The BioIT team is nationally recognized as a Dutch Techcentre for Life sciences (DTL) hotel and actively participates in this and other national life science and bioinformatics platforms such as Dutch Masters of Life Sciences Health and Medical Delta.

The LIACS BioIT team collaborates extensively with other institutes within the Faculty of Science, supporting biologists in bioinformatics.

LIACS Bioinformatics expertise

The added value of informatics to life science and health research

1 Extracting meaning from large quantities of biological data

In the life sciences enormous amounts of data are collected that contain valuable information. In order to unravel the patterns that lie enclosed in the biological data, in-depth knowledge of the development and application of algorithms is required.

LIACS is experienced in discovering these interesting new patterns. This expertise, called data mining, operates at the intersection of computer science and statistics. By automatic or semi-automatic analysis of data, previously unknown patterns are extracted and transformed into an understandable structure for further use and interpretation. Since data mining is a computationally very intensive process, computers with enormous calculating power are required, and these are available at LIACS.

2 Understanding biology by modelling its processes

The various steps involved in biological processes, e.g. protein activations or metabolic reactions, can be described by mathematical models. This field, known as systems biology, is an emerging field that aims to build virtual models of dynamic biological systems. The models enable the simulation of biological processes in bacteria, animals or the human body, allowing the investigation of molecular or regulatory mechanisms. For example, models can be used to understand how a system would react to impulses, stimulants or changes in the environment. LIACS builds models to understand and investigate biological processes using evolutionary algorithms, Petri Nets and other methods.

3 Enhancing speed and quality of data analysis by IT infrastructure

In the health care domain, genomics and imaging are playing increasingly important roles in both diagnostics and research. In practically all fields of health care research large amounts of genomics and image data are created. These have to be processed, stored, managed and, analysed, in order to extract meaningful and legitimate conclusions. It is vital that these analyses are performed efficiently and with high quality.

Multiple aspects have to be considered, as illustrated by LIACS achievements, in joint projects with academic hospitals and other research partners, in the field of brain research and genetics. LIACS makes use of a high performance data and compute IT infrastructure on site, to speed up and improve the quality of data analyses.

4 Standardisation, integration and visualisation of research data

The increasing heterogeneity and volume of data in the life sciences means that techniques for data integration and the use of community standards are increasingly important in order to ensure we get the most value from data created, and to enable the sharing and reuse of data.

The visualisation of data enables researchers to interpret data. Also, making complex biological structures visible in 3D can be a huge advantage for scientists to see their subject of investigation in greater detail and from all angles to e.g. study docking properties of target receptors. Augmented reality is also an example that allows scientists to test for these characteristics.

5 Building and applying tools for life science big data analysis

With the growing amounts of biological data life science and health researchers depend more and more on expertise from the informatics domain. Tasks including data science and stewardship, data management, bioinformatics analyses, all rely on informatics. Running a big data research project requires a well-thought-out design of the experiment, data integration, processing and analysis.

LIACS can offer expertise to consult on setting up large scale and high-throughput experiments, build databases to manage large amounts of related biological data, analyse data, build tools to visualise data, or support on selecting and using the right software or tools to analyse and visualise data.

On the following pages examples of research and IT support activities at LIACS are presented. Each of the examples describes the added value of informatics to life science and health research.

Deriving valuable medical information from heart failure data

Gathering and analysing medical data is an expensive and time-consuming activity. It is a burdensome for both the patient and the medical institution. But not all data are of equal importance. Some data provide results that have already been obtained from other sources, other data don't seem to contain relevant information. LIACS developed algorithms to process all data and provide directions for medical examiners for further research.

Data collected from patients have to be compared to data collected on a control group. Since these data have been gathered using different protocols, they cannot be compared directly. Whereas these data in the past would have been considered incomparable, thus worthless, LIACS is developing algorithms that integrate these data to make data comparable, thus helping in early diagnosis and treatment and even preventive actions.

In a particular example, machine learning techniques were applied on three cohort heart failure studies, for which heterogeneous data was available including SNP data and metabolic data from free fatty acid studies and NMR measurements.

This work is part of the COMMIT project (www.commit.eu) which involves many academic and commercial partners.

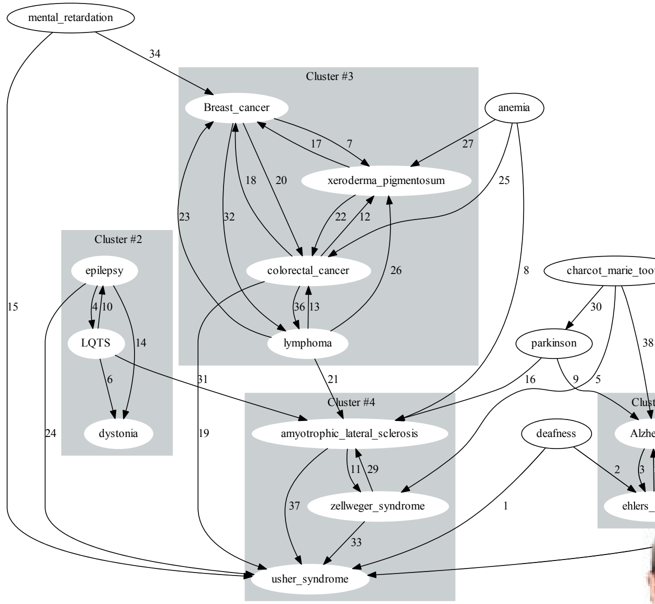
Drs. Jonathan Vis
data mining expert

COMMIT/ is a public-private research community solving grand challenges in information and communication science shaping tomorrow's society

Identifying protein-protein interactions related to human diseases

Protein-protein Interaction (PPI) networks are widely used for the analysis of diseases. LIACS developed a novel method for discovering correlations, such as, similarity in the origin of the disease, or proteins involved among different diseases. These correlations are retracted from a PPI network and background information, and represented in a human disease network (see Figure). The human disease network is then used for making predictions about the involvement of genes in diseases. Evaluation has shown that this approach unveils relations between proteins and diseases, as well as between diseases, that were missed by earlier approaches.

This Annotated graph mining project was part of a NWO VIDI and was done in collaboration with the Leiden Academic Center for Drug Research (LACDR).



Dr. Hendrik Blockeel
machine learning expert

Mining experimental data, and enriching the results

Modern high-throughput analysis techniques offer the biologist many new ways of measuring various aspects of life on a large scale. The wealth of data generated potentially reveals interesting biological details, but the sheer size of the data is a challenge for existing statistical techniques. The LIACS Data Mining group has developed techniques for finding up-regulated genes, relevant proteins and so on, in experimental datasets of many variables, both under conditions of few samples measured, as well as of many. Specifically, this analysis can be performed with multiple datasets across several biological domains. The techniques have been applied, through various co-operations with (medical) biologists, to domains such as embryonal tumours, leukaemia and the metabolic syndrome.

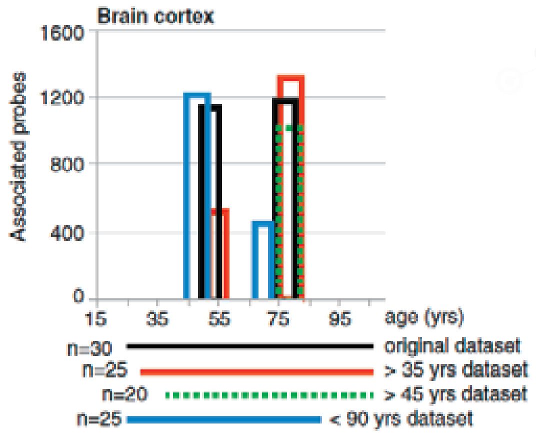
Furthermore, LIACS developed an enrichment method capable of translating detailed findings in experimental data into high-level concepts, capturing the larger phenomena at play in the biological system. LIACS developed software to integrate knowledge from a variety of domain-crossing sources, like the entire literature, the online knowledgebases GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes), as well as custom-made databases. This literature-based enrichment method proved to produce more meaningful results in comparison to existing GO-based enrichment methods. LIACS also demonstrated that the new method is capable of producing relevant results for biological entities other than genes, which is a clear advantage over other enrichment methods.

Research was funded from various sources, including the EU and the Netherlands Consortium for Systems Biology (NCSB), which involves various national and international research partners.

Dr. Arno Knobbe
data mining expert

Identifying genes involved in the aging process

Aging is characterised by progressive tissue degeneration and is associated with genome-wide changes of mRNA expression profiles. Therefore, it is expected that age-regulated changes in expression profiles could describe the process of aging. In cooperation with the Leiden University Medical Center (LUMC) LIACS has developed a robust and unbiased methodology to identify age-regulated expression trends in cross-sectional data sets from healthy donors. The method combines state-of-the-art methods from nonlinear regression and cluster analysis and allowed to identify significant patterns in gene expression time series.



In contradiction to existing beliefs it was discovered that aging is not a linear process. Gene expression analysis of four different tissues, including brain cortex and kidney, of healthy donors indicate that the aging process is progressive at two age bends, around the age of 50 and 70.

This work was done in collaboration with the Department of Human Genetics, Leiden University Medical Center (LUMC).

Dr. Michael Emmerich
biosystems expert

Modelling and simulating biology

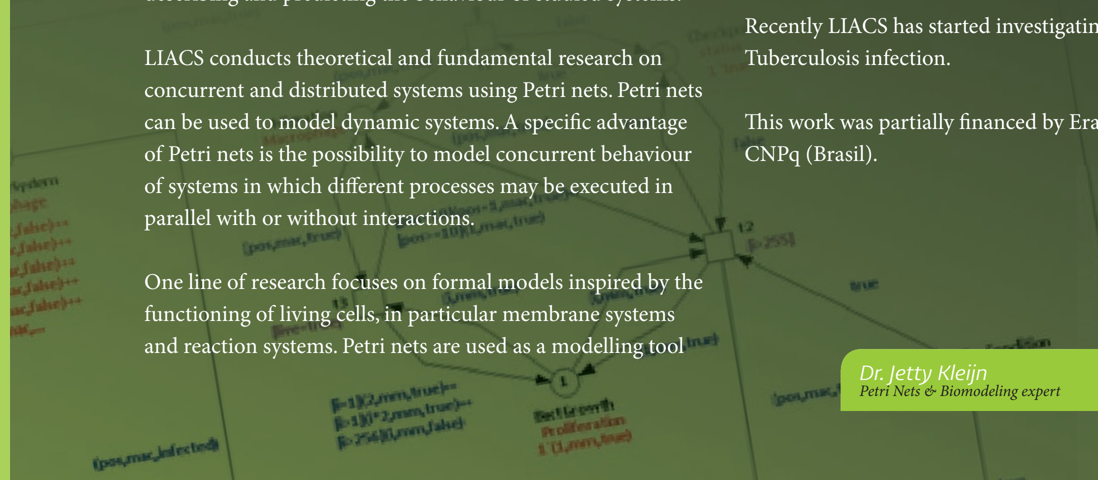
Bioscience modelling and simulating (M&S) are two important techniques to predict outcomes, thus giving direction to subsequent scientific research. M&S provide the basis for informed, quantitative decision-making. M&S enable scientists to enhance their hypotheses before conducting in vivo experiments, thus saving time, effort and money. M&S also facilitate the continuous integration of available information related to for instance a drug or disease into constantly evolving mathematical models capable of describing and predicting the behaviour of studied systems.

LIACS conducts theoretical and fundamental research on concurrent and distributed systems using Petri nets. Petri nets can be used to model dynamic systems. A specific advantage of Petri nets is the possibility to model concurrent behaviour of systems in which different processes may be executed in parallel with or without interactions.

One line of research focuses on formal models inspired by the functioning of living cells, in particular membrane systems and reaction systems. Petri nets are used as a modelling tool to describe complex processes in developmental biology. They are dynamic and thus provide an operational approach suitable for the simulation and visualization of events in a system. Petri nets have been used successfully in the modelling of biochemistry and systems biology, but rarely at a higher level. Initially motivated by the development of the body axis in embryos of Zebrafish, *Danio rerio*, a general and abstract Petri net model for the formation of a molecular gradient has been developed and validated by LIACS.

Recently LIACS has started investigating and the modelling of Tuberculosis infection.

This work was partially financed by Erasmus Mundus and CNPq (Brasil).



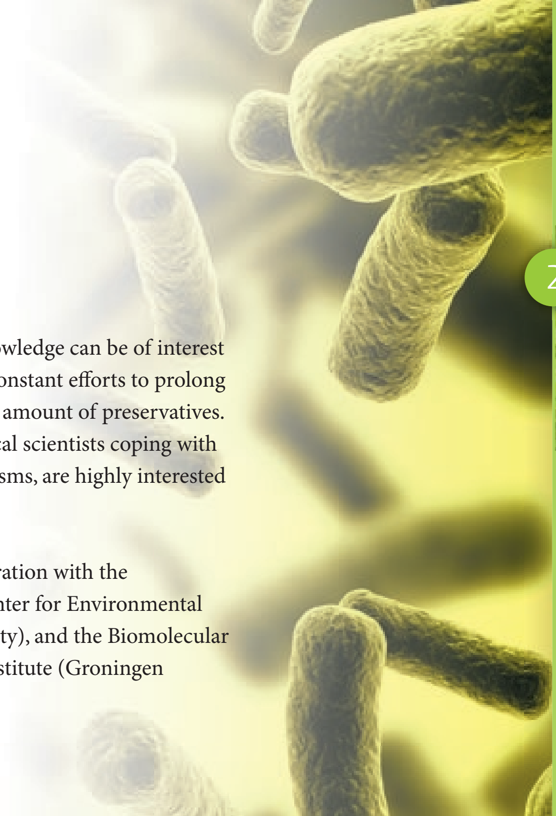
Dr. Jetty Kleijn
Petri Nets & Biomodeling expert

The evolution of bacteria under fluctuating nutrient conditions

Bacteria are highly adaptive to fluctuating environmental conditions. LIACS is investigating the evolutionary dynamics of bacteria in environments with a randomly fluctuating availability of nutrients. Soil bacteria are modelled in different time scales (gene regulation, evolutionary) and efficient stochastic simulation algorithms are provided and calibrated with data from the Biomolecular Sciences and Biotechnology Institute (Groningen University). LIACS enhances the effectiveness and accuracy of these stochastic simulation algorithms and develops methods for identifying model parameters based on the available data.

LIACS aims to obtain a deeper understanding of the interplay between randomness, evolution, and gene regulation in bacteria. This knowledge can be of interest for the food industry in their constant efforts to prolong sustainability with a minimum amount of preservatives. Also other scientists, e.g. medical scientists coping with drug resistance in microorganisms, are highly interested in these processes.

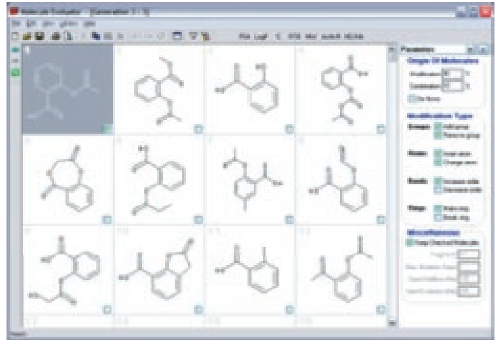
This work was done in collaboration with the Mathematical Institute, the Center for Environmental Sciences (both Leiden University), and the Biomolecular Sciences and Biotechnology Institute (Groningen University).



Drs. Alexander Nezhinsky
biosystems expert

Optimising the search for potential drug molecules

Drug-like molecules must adhere to a number of chemical properties, such as not exceeding a maximal size, containing a specific group, having a certain polarity, staying in the body for a certain period and being soluble in blood. Apart from these "convenience reasons" a drug molecule must, of course, be effective, have no or little side effects and must be synthesisable. The amount of potentially useful molecules is estimated to be of the unimaginable size 10^{60}.



Evolutionary algorithms for the simultaneous optimization of multiple drug criteria have been developed. They can explore the space of potential molecules and identify trade-off efficient sets of potentially useful molecules, therewith enhancing the success rate of drug effectiveness enormously. As opposed to earlier methods, these techniques enable the exploration of trade-offs between multiple design criteria and support interactive search. With the use of these algorithms effectiveness can be simulated and docking characteristics of molecules to the targeted receptors can be evaluated.

This work is done in collaboration with the Leiden Academic Center for Drug Research (LACDR), Leiden University.

Drs. Iryna Yevseyeva
chemoinformatics expert

Imaging MS-based molecular histology for clinical investigations

Imaging Mass Spectrometry (MS) enables the distribution of hundreds of peptides and proteins to be determined directly from tissue samples. The application of multivariate methods enables tissues to be annotated on the basis of their peptide and protein profiles. Such molecular histology can provide novel diagnostic capabilities and help identify which patients will respond to chemotherapy. For the application of these techniques, increased computational power is required. Large tissue series are needed for clinical investigations, each tissue consisting of 10-50k profiles and 100-1000 variables per profile. In cooperation with the Bio-molecular Mass Spectrometry Unit of the Parasitology Department at LUMC LIACS performs research on the development of automated design and programming of embedded systems on chips, such as Graphical Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs). They are considered the most cost-effective solution for providing the necessary computational power.

Imaging MS datasets of large patient tissue series can be processed up to 13x faster on Embedded GPUs. In addition, speed improvements of up to 200x for wavelet decomposition of mass spectra using Embedded FPGAs are expected. Such speed improvements would enable the cross-validated analysis of 100 patient tissues to be performed in 20 minutes instead of 7 hrs.

This work was financed by LUMC via an NWO research grant.

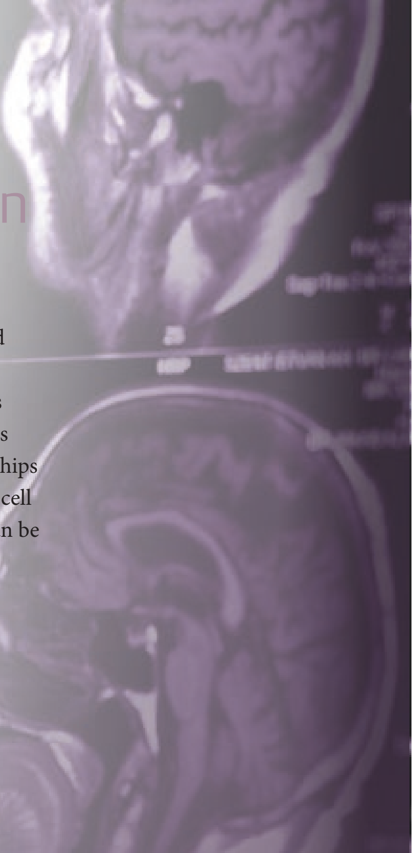
Dr. Ana Balevic
embedded computing expert

Acceleration of medical image registration

Image registration is an important task in medical image processing. Applications such as image-guided surgery in the brain, where low quality intra-operative ultrasound scans need to be registered and matched to high quality pre-operative computer tomography (CT) or magnetic resonance (MR) scans, require the registration to be performed within a minute or less.

In cooperation with the Radiology Department at the Leiden University Medical Centre (LUMC), research at LIACS aims to accelerate publicly available software for medical image registration. These research results and LIACS expertise in Embedded Software acceleration have been used to accelerate the Medical Image Registration software by means of parallel processing. As a result this software runs very fast on heterogeneous embedded system on chips (SoCs) such as graphics processing unit (GPU) or cell microprocessors. Acceleration of up to 30 times can be achieved with this technology.

This work was financed by LUMC via an NWO research grant.

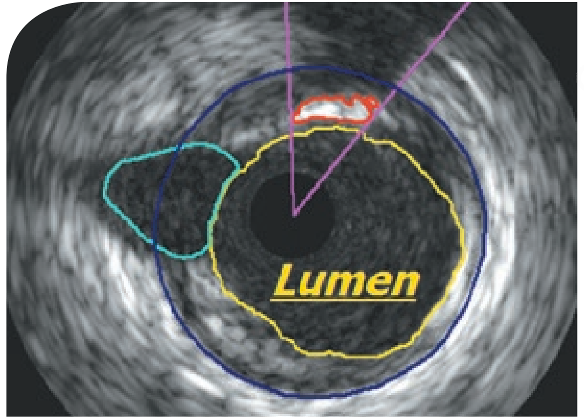


Dr. Hristo Nikolov
embedded computer expert

Optimising image processing software in cardiology

The detection and measurement of calcification in blood vessels is a difficult, yet extremely important, problem in medicine. Intubation of blood vessels is a very burdensome process for the patient. It should therefore be done quickly and smoothly, yet accurately. It takes a well-trained expert considerable time and effort to interpret all the images the scanner makes on its journey through the vessel. A complicating factor is that well-trained experts for the annotation of Computed Tomography Angiogram (CTA) or ultrasound movies are rare. Reliable systems that can assist cardiologists in identifying plagues are therefore highly desirable.

Existing algorithms for analysing data from body scans are often manually calibrated, which, due to the large number of control parameters often leads to suboptimal performance. Together with the Leiden University Medical Center, LIACS has developed parameter learning algorithms, Mixed Integer Evolution Strategy (MIES) for automatically learning optimal settings. This helped develop high performing software detectors for measuring the position and size of calcified plaque. The detectors have been incorporated into commercial software.



This work was financed through NWO and was performed in collaboration with the Division of Image Processing (LKEB) of the Leiden University Medical Center.

Prof. dr. Thomas Bäck
natural computing expert

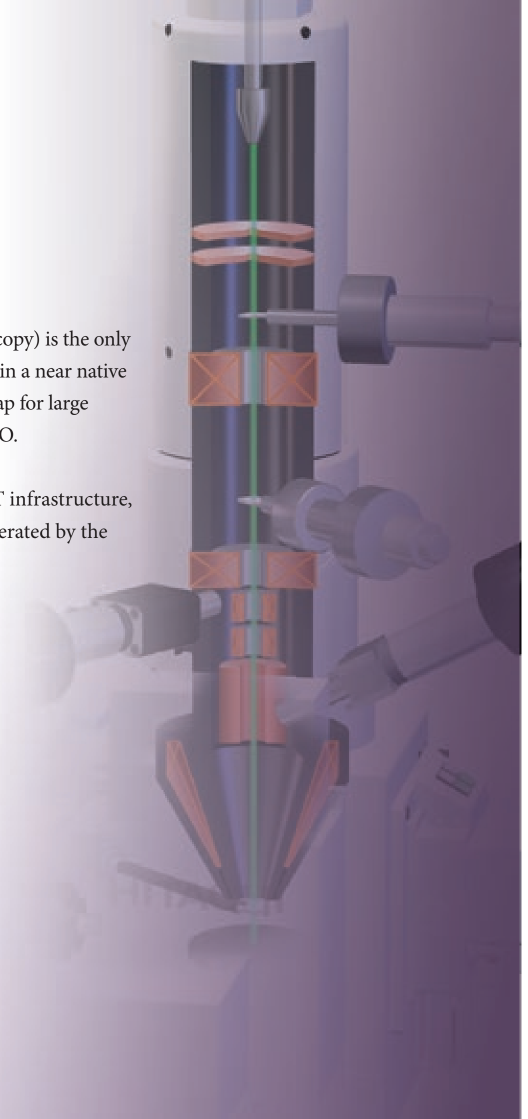
NeCEN microscopy IT infrastructure

In 2011 The Netherlands Centre for Electron Nanoscopy (NeCEN, www.necen.nl), a joint effort from Leiden University and 10 partners, and funded by NWO and de European Union, made access to two of the most advanced cryo-transmission electron microscopes available to research institutes and companies. The two NeCEN microscopes are specifically designed to explore complex structures inside cells with great detail. At present, cryo-TEM (transmission electron microscopy) is the only way to image macromolecular complexes in a near native state. NeCEN was included in the Roadmap for large infrastructures, as defined in 2008 by NWO.

Apart from consulting on the required IT infrastructure, LIACS offers support to analyse data generated by the NeCEN TEM's.

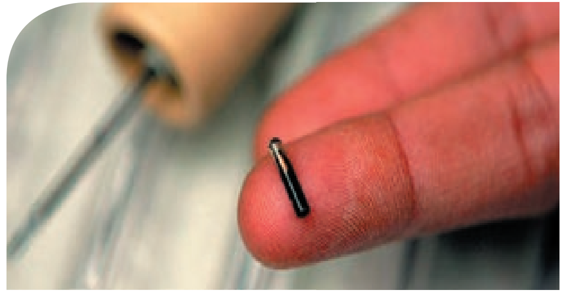
NANOSCOPY

High performance microscopes are becoming increasingly powerful, but also increasingly expensive. The NECEN facilityoffersaccessto two of the worlds most advanced cryo transmission electron microscopes for organisations involved in life science research. Life science nanoscopy research involves large datasets of images. Organisations face storage challenges as well as data management and processing challenges that LIACS can advise on.



Energy efficient computer-brain Interaction

In order to avoid the use of an Electroencephalography (EEG) headset on the skull, which stigmatises the patient, subcutaneous implanted sensors are developed. Exploiting such sensors allow the use of a baseball cap or hair-band, to enable EEG monitoring. The monitoring device, hidden in the baseball cap, should provide energy to the subcutaneous sensors and enable read out of a much better quality EEG signals, in comparison to signals sensed on the skull.



The EEG monitoring device, as developed at LIACS, is an energy-autonomous Medical Embedded System (MES). Energy autonomous systems either at start-up have sufficient energy capacity for the whole operational life of the system, or they need energy scavenging mechanisms to harvest energy during their operational lifetime. Furthermore, the processing and communication of the collected medical data during monitoring needs to be aware of the availability of energy. It is apparent that all medical data processed and communicated in such energy-autonomous MES need to be performed at the lowest possible energy cost.

LIACS aim to develop generic methods and tools for the design and programming of energy-autonomous MES.

Energy aware algorithms need to be developed where the amount of medical data processed, communicated and stored depends on the amount of available energy. E.g. send less data or less accurate data in case of insufficient energy availability; decide whether data should be stored locally to maintain a certain level of off-line monitoring.

This work was financed by STW and performed in cooperation with TU Eindhoven and the University of Twente.

Dr. Todor Stefanov
embedded computing expert

e-BioGrid: enabling life science research and technology

LIACS has been involved in the e-BioGrid project (2010-2013, www.e-biogrid.nl), building the e-infrastructure for life science research, in collaboration with SURFsara and the Netherlands Bioinformatics Centre (NBIC). The e-BioGrid project was part of the NWO BiG Grid project, including partner Nikhef, to facilitate scientific research through access to advanced ICT research infrastructure.

The e-BioGrid project provides support in e-science problem-solving environment development and facilitates access to advanced hardware facilities with supporting manpower for bioscientific research. Within the e-BioGrid project several simple to complex e-science problem-solving environments are developed that can connect to cloud computing, compute clusters, graphic processing units, grid networks and other distributed or parallel computing solutions. The e-BioGrid project focussed on big data research projects involved with these technology areas: biobanking, mass spectrometry, microarray technology, nanoscopy and imaging, next generation sequencing, NMR Spectroscopy, medical imaging.

The e-BioGrid project involved small (< 3 months) and larger (1-2 years) subprojects. All received support on demand. Examples include building an analysis pipeline on Grid facilities for the 750 human genome project in collaboration with Biobanking and Biomolecular Research Infrastructure (BBMRI), establishing an e-infrastructure for the microscope image analysis at NeCEN, and optimizing the analysis pipeline for the Virtual Lab of Plant Breeding (VLPB).

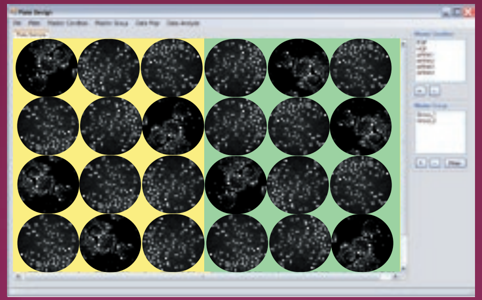
Prof. dr. Joost Kok
data mining expert

SURF SARA nbic

Image and data analysis for high-throughput screening

In modern biology screening for effects of agents or compounds on biological material is applied to obtain insight into cause-effect relations as well as to develop new treatments and medicine.

High-throughput screening (HTS) that is applied on cells is referred to as Cytomics. The read-out is accomplished through an automated microscope setup and results in a large volume of image data, often as time-series. From the images typical features need to be extracted. This should be realized in an efficient manner and once features are available, patterns need be identified in these features.



Moreover, it should also be possible to derive patterns from a collection of screens. This requires a dedicated database, suitable for such data mining tasks. The Imaging & BioInformatics group of LIACS has developed an accurate and efficient automated pipeline for processing of HTS data. This pipeline is actively maintained and updated with new computational tools.

On another level, that of the organism, HTS is applied on the zebrafish model organism to study the effect of agents and/or compounds. Zebrafish screens use different imaging techniques and also the feature extraction and data processing are quite specific. For zebrafish HTS a unique processing pipeline has been successfully developed. Examples of these pipelines are available on bio-imaging.liacs.nl.

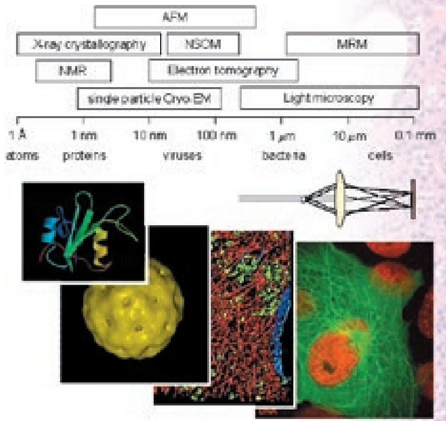
This work was supported by NBIC Biorange, Smartmix, and Erasmus Mundus

A virtual microscope supporting image integration and knowledge discovery

As a member of Cyttron (cyttron.org), a consortium of 14 academic and industrial partners that aims to develop and integrate bio-imaging technologies, LIACS pursues the integration of images of different modalities, i.e. X-ray Crystallography, NMR, single particle cryo-EM, Scanning Force Microscopy, Electron Tomography, Light Microscopy and Magnetic Resonance. The images may depict phenomena at different levels of detail. These imaging modalities produce results in specific formats. The goal is to have a system that can acquire images from organ level all the way down to molecule level, allowing continuous scaling over the entire resolution scale. The images are enriched by metadata, describing the image and structures in the images. This will allow the discovery of relations between structures and events.

This objective has led to the design and implementation of the Cyttron Imaging Platform. The platform embodies the Cyttron Scientific Image Database for Exchange, where images can be shared and connected; and the Cyttron Visualisation Platform, for the visualisation of these images and relations.

This work was done with a large number of academic and industrial research partners within the Cyttron consortium.

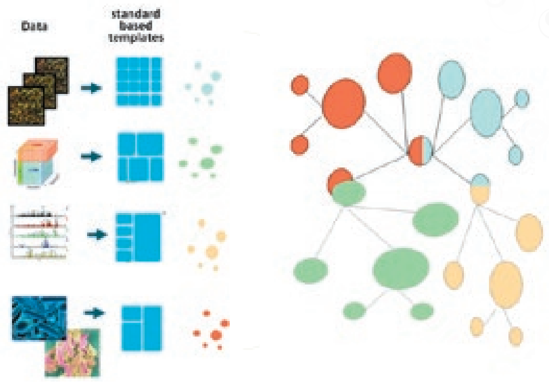


Cyttron

Semantic Integration of Systems Biology Data and Models

Systems Biology studies the dynamic and complex interactions in systems of biological components. Modelling the behaviour of groups of interacting components, using quantitative measurements from technologies like genomics, proteomics and metabolomics, allows the description and prediction of dynamic processes. Consequently, Systems Biology is a multi-disciplinary science that requires large-scale data integration and close links between experimental and modelling activities.

The SEEK Platform (www.seek4science.org) is a web-based environment designed to enable the sharing of systems biology data and models, in the context of the experiments that produced them. Embedded tools and resources enable linking between data and models, model simulation and semantic annotation. This provides a rich platform for systems biologists to store, analyse and explore the results of their research.



SEEK was developed in the SysMO-DB project for the SysMO consortium (Wolstencroft, University of Manchester), but it is open source and has since been adopted by many other European Systems Biology consortia. Work on the SEEK is now the foundation for research on semantic data and knowledge integration at LIACS and it is being evaluated for use in other model-driven disciplines.

Dr. Katy Wolstencroft
semantic data integration expert

Augmented reality of biology

Research funding institutions worldwide are compiling large databases of molecular data, such as drug or DNA information. These molecular structures are being used by scientists to advance the frontiers of our understanding of the fundamental machinery in the cell and the way in which molecules interact. This is of great help in designing custom drugs.

LIACS is developing new paradigms for visualising and exploring these molecular databases. Our technology, the Leiden Augmented Reality System (LARS) can turn any computer into a virtual microscope. Unlike other Augmented Reality systems one does not need a special marker to project the image on. Common objects in the user's environment, a book, a soda can or even the user's hand will do.

One can manipulate the 3D image by moving, turning, spinning the object around. This allows you to dynamically visualize interactions between multiple molecules. You can align them, look for ways to recombine them, explore their docking abilities etc. This allows chemists, searching for a suitable drug, to match molecules with receptors, thus preselecting molecules as possibly effective drugs.

LARS is designed to work with low memory devices. It has sufficiently low computational requirements that you can project the image at all times at all places, provided you have a notebook and a web cam.

This project is a collaboration between members of the Leiden University (Massive) Multimedia Information Retrieval Center (LUMIR), on advanced visualisation techniques for biological analysis and education. It includes LIACS, the Leiden University Medical Center (LUMC) and the Leiden Institute of Chemistry (LIC)).

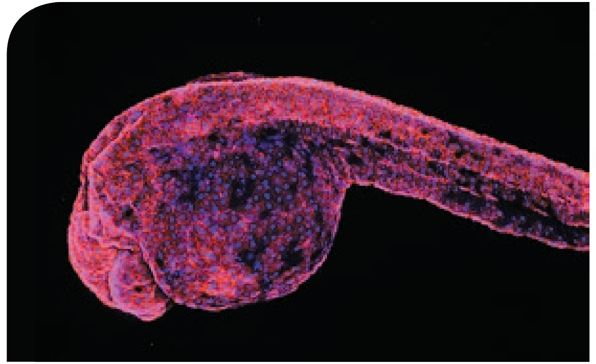
Dr. Michael Lew
search & visualisation expert

3D visualisation of embryonic development of Zebrafish

Zebrafish (Daniorerio) has become an important model organism in scientific research because of its importance in understanding vertebrate development and its relevance in human disease studies. LIACS has developed two important information systems to manage and visualise data.

The 3D atlas of Zebrafish development is a digital representation of Zebrafish embryo anatomy. The atlas contains 3D representations reconstructed from serials sections obtained from histological studies as well as models obtained from confocal microscopy. It includes graphical annotations and text-based descriptions of anatomical domains in images. Users can access 3D models through a web application and query the 3D atlas without prior knowledge of the exact anatomical terms.

In complement, the Gene Expression Management System (GEMS) is a database for 3D spatio-temporal gene expression patterns in Zebrafish. GEMS includes mechanisms for linking and mining data. GEMS is an integrative visualisation platform that addresses an important challenge of developmental biology: how to integrate genetic (e.g. expression) data that underpin morphological changes during embryogenesis. GEMS provides integration with other resources such as the Zebrafish Information Network (ZFIN).



For effective data integration annotation with common terminology is required. To achieve this, LIACS developed the Developmental Anatomy Ontology of Zebrafish (DAOZ).

This work was partially supported by NWO, SURF and performed in collaboration with partners of the Institute of Biology Leiden (IBL).

Dr. Fons Verbeek
semantic data integration expert

BioMaps, a graphical depiction of human physiology

Clinicians and scientists in drug development operate at varying scales of human anatomy and physiology, ranging from molecules, to cells, to organs and complete physiological systems. In the course of these efforts, biomedical professionals generate considerable volumes of data resources. The productive collaboration between biomedical professionals requires an effective infrastructure to share such data resources. For evident reasons of patient privacy, legal constraints and commercial competition, it is unrealistic to expect that such data and model resources are to be openly shared in the public domain.

To reconcile and bridge these opposing requirements, pre-competitive communities in bioengineering and pharmaceutical industries have developed knowledge management approaches to share semantic metadata about biomedical resources, rather than the resources themselves. Methodologies depend on the application of publicly available ontologies to convey explicit biomedical meaning to resource metadata. However, this type of metadata infrastructure still requires considerable computational expertise to manage and, as such, is beyond the effective reach of clinicians and biomedical scientists.

The solution is for biomedical professionals to visually navigate and interact with graphical depictions of physiology circuit-board schematics that are, in practice, the automatically generated embodiment of multi-scale ontology knowledge.
LIACS develops a web-based google-maps like graphical tool for the visual management of ontologies and metadata.

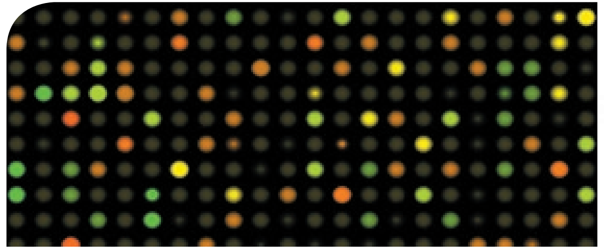
This work had received support from the DDMoRe project (www.ddmore.eu), which involves a great number of academic and commercial partners.

Dr. Natallia Kokash
tool development expert

Innovative tools for semantic data integration and usage

To enable novel discoveries from sequencing, micro array and other data intensive experiments innovative tools for semantic data integration and usage are essential to organize, search and fully exploit these huge biological datasets.
In a joint effort with Leiden University Medical Center (LUMC) and Vrije Universiteit Medical Center (VUMC) we addressed some of these challenges and designed and implemented new generic tools and databases for micro array database applications, allowing storage, processing and integration of data, independent of the software platform. In a follow-up project these tools have been extended for copy number analysis, high-density SNP arrays, de-novo sequencing data, and proteomics data.

A general Gene Name Normalization Framework containing a novel Name Entity Recognition (NER) tool has been designed and implemented. It has been studied and evaluated against existing NER tools like Mallet, Abner and LingPipe on BioCreative data sets to ensure applicability for richly annotated Bio-data.

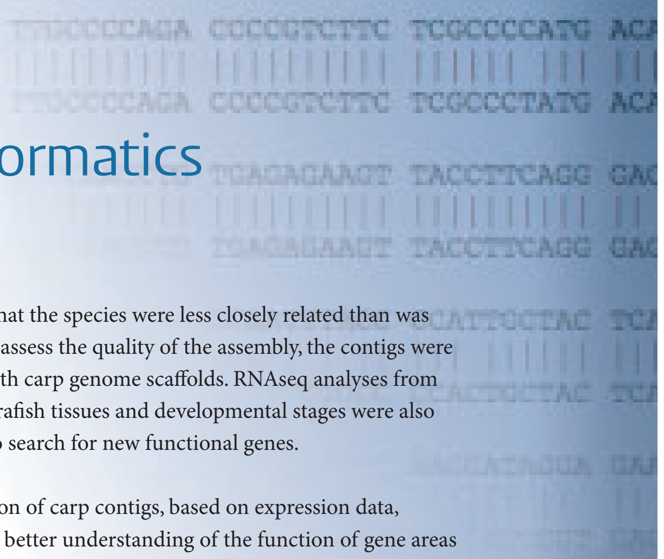


Currently, a Semantic Biobase Toolkit is being developed using the Gene Name Normalization Framework for the construction of semantic indexes that enable several semantic ranking schemes for information on (candidate) gene name lists. The toolkit will provide very powerful functionality for example in population based genotyping phenotype processing pipelines for detecting genetic markers, and semantic searching and - exploitation of biodata in general.

This work was part of the CMSB project (www.cmsb.nl).

CENTRE FOR
Medical Systems Biology

Dr. Erwin Bakker
content based indexing expert



Insight into genomes by bioinformatics tools and support

The presence of large and complex data sets creates an enormous need for bioinformatics expertise and infrastructure in contemporary biology for advanced bioinformatics analyses. The LIACS BioIT team helped perform these bioinformatics analyses.

A number of small projects in collaboration with the Institute of Biology Leiden (IBL), resulted in new insight in zebrafish research. The carp transcriptome has been sequenced. A comprehensive comparison of carp and zebrafish transcript sequences investigated similarities between the species concluding that the species were less closely related than was expected. To assess the quality of the assembly, the contigs were compared with carp genome scaffolds. RNAseq analyses from different zebrafish tissues and developmental stages were also performed to search for new functional genes.

The annotation of carp contigs, based on expression data, allowed for a better understanding of the function of gene areas in the carp genome. Also, tools were developed to search for DNA motifs with location constraints.

NEXT-GENERATION SEQUENCING

Next-generation sequencing technology produces large amounts of data. Depending on the research question, the experiment is designed, and results analysed. This requires the development of advanced bioinformatics analysis methods, bioinformatics expertise and infrastructure. Topics of interest cover expertise areas like: design for omics experimentation, statistics, data-analysis, visualization and experiment interpretation. Important infrastructural elements are data storage, data-analysis pipelines, specific bioinformatics tools, databases and high-performance computing

Dr. Genevieve Girard
Bioinformatics expert

Facts & figures

The added value of informatics to life science and health research

1 **Extracting meaning from large quantities of biological data**
Deriving valuable medical information from heart failure data
Identifying protein-protein interactions related to human diseases
Mining experimental data, and enriching the results
Identifying genes involved in the aging process

2 **Understanding biology by modelling its processes**
Modelling and simulating biology
The evolution of bacteria under fluctuating nutrient conditions
Optimising the search for potential drug molecules

3 **Enhancing speed and quality of data analysis by IT infrastructure**
Imaging MS-based molecular histology for clinical investigations
Acceleration of medical image registration
Optimising image processing software in cardiology
NeCEN microscopy IT infrastructure
Energy efficient computer-brain interaction
e-BioGrid: enabling life science research and technology

4 **Standardisation, integration and visualisation of research data**
Image and data analysis for high-throughput screening
A virtual microscope supporting image integration and knowledge
Semantic integration of systems biology data and models
Augmented reality of biology
3D visualisation of embryonic development of Zebrafish

5 **Building and applying tools for life science big data analysis**
BioMaps, a graphical depiction of human physiology
Innovative tools for semantic data integration and usage
Insight into genomes by bioinformatics tools and support

For more information see www.liacs.nl/bioinformatics. Contact LIACS for joint Bioinformatics research and educational projects or support in Bioinformatics by the BioIT team: *requests@liacs.leidenuniv.nl*

A digital version of this brochure is available on
http://www.liacs.nl/bioinformatics/brochure

Survey of partners

Erasmus Medical Centre Leiden University Medical Centre Trigion Beveiliging B.V. Vrije Universiteit Medisch Centrum Academisch Medisch Centrum, Universiteit van Amsterdam TNO Uppsala University European Molecular Biology Laboratory Mango Business Solution Universita' degliStudi di Pavia Interface Europe Serpo B.V. Institut National de Recherche en Informatique et en Automatique Simcyp Limited Delft University of Technology Cyprotex Discovery Limited Martin-Luther-Universitaet Halle-Wittenberg University of Navarra Université Paris Diderot-Paris Optimata Ltd Galapagos N.V. Novartis Pharma AG AstraZeneca AB GlaxoSmithKline Research & Development Ltd Eli Lilly and Company Limited Merck Novo Nordisk A/S F. Hoffman-La Roche AG Institut de Recherches Internationales Servier UCB Pharma SA Netherlands Institute for Systems Biology Cancer Genomics Centre Centre for BioSystems Genomics Centre for Medical Systems Biology Kluyver Centre for Genomics of Industrial Fermentation Netherlands Bioinformatics Centre Top Institute Food and Nutrition Top Institute Pharma Philips Electronics Nederland B.V. FEI Electron Optics B.V. DSM Innovative Synthesis B.V. Pepscan Therapeutics B.V. Nikon Intruments Europe B.V. Virtual Proteins B.V. Science and Technology Facilities Council NCB Naturalis Maastricht University University Utrecht IMI Joint Undertaking Pfizer Almende B.V. DEAI Services B.V. Hubrecht Laboratorium ZF-Screens B.V. Arthrogen B.V. Nijmegen University Progentix B.V. Fytagoras B.V. GlaxoSmithKline Ltd. Agendia NWO STW Agentschap NL

Leiden Institute
of Advanced
Computer Science

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University