



Computational Life Sciences

April 17, 2007

Preface

This booklet contains a summary of all ongoing projects with a link to bioinformatics, biomathematics and computational life sciences within the following research programmes:

- Computational Life Sciences (Netherlands Organisation for Scientific Research): pages 2 to 23
- Biomolecular Informatics (Netherlands Organisation for Scientific Research): pages 24 to 29
- Horizon (Netherlands Genomics Initiative): pages 30 to 36
- BioRange (Netherlands Bioinformatics Centre): pages 37 to 49

A portrait of each project is given by the title, the name and email address of the principal investigator/ coordinator, a short summary or (if available) first results and the link to the project website.

This booklet therefore presents a valuable overview of ongoing research in these fields in the Netherlands.

A new round of the Computational Life Sciences programme is on its way offering an opportunity to continue and integrate the research efforts presented.

Computational Life Sciences (CLS)

The Computational Life Sciences programme operates at the interface between the computational sciences and the life sciences. This programme is a joint initiative by the NWO councils Physical Sciences, Medical Sciences (ZonMW), Earth and Life Sciences as well as the NWO Governing board and Netherlands National Computing Facilities foundation.

The scientific challenge of the CLS research programme is to understand the behaviour of linked dynamic models which take account of the effects on a micro, meso and macro scale and the relationships between these. The computational methods and techniques being developed for this purpose lie in the field of the physical sciences. In 2003 with a budget of 5.5 million euro sixteen projects were funded where computer scientists, mathematicians and life scientists work together to solve a life science problem.

A new round of this programme, albeit with modified research goals is on its way starting spring 2007. This new round is financially supported by the Netherlands Genomics Initiative (NGI) as well as the Netherlands Bioinformatics Centre (NBIC).

Further information on the CLS programme can be found at www.cls.nl or www.nwo.nl/cls.

2 **Metabolism, gene regulation and evolution: mutual interaction across time-scales**

Prof. dr. P. Hogeweg, p.hogeweg@bio.uu.nl, www-binf.bio.uu.nl/CLS/

In this project we develop and use computational techniques to solve fundamental question in biology, in particular with respect to the evolution of complex biological systems. Here we focus on interaction between different scales, ranging from metabolic time-scales to evolutionary timescales, and ranging from intracellular processes to whole organisms to populations.

The central general themes of this project are:

- *from a biological perspective*
 1. multilevel evolution: interaction between timescales
 2. evolvable genotype-phenotype mapping
 3. evolution of evolvability
- *from a computational perspective*
 4. modeling multilevel processes, space and timescales
 5. combining dynamic bioinformatic modeling and data-analysis

We work on it in a multidisciplinary team: Stan Marée, Nobuto Takeuchi, Paulien Hogeweg (theoretical biology/bioinformatics), Otto Cordero, Anton Crombach (computational science) and Milan van Hoek (physics).

Results

– *Genome structuring and adaptability*

Short term evolution of yeast to glucose-poor environments is very efficient and often involves gross chromosomal rearrangements (GCR). We have shown in a multilevel model which includes transposon dynamics, that long term evolution to a variable environments leads to genome structuring such that short term evolution becomes very efficient and involves such GCR's. (themes 1 – 4)

– *Bi-stability in the lac operon*

The operon structure of the Lac-operon suggests the possibility of hysteresis and bi-stability. Using an artificial inducer, this has indeed be demonstrated experimentally. Using a multilevel modeling approach (including metabolic up-to evolutionary timescales) we show that evolution leads to loss of bi-stability, favoring a gradual shift rather than a switch for the natural inducer. The evolved operons, do however show bistability for artificial (non-metabolizable) inducers. This research demonstrates the need for multi-scale models even to solve single scale problems, and the usefulness of evolutionary models to alleviate parameter uncertainty. (themes 1 – 4)

– *Stochastic gene regulation*

We extended the previous model to include stochasticity of protein expression. The model reproduced experimental results very well. In our evolutionary model we show that the role of stochasticity on gene expression is minor and is even further reduced through evolution. (themes 1 – 4)

– *Evolution of gene regulation I*

Gene regulation networks of yeast and bacteria show a power law scaling of connectivity and an over-representation of certain small scale network motifs (e.g. the feed forward loop). We show that these features, as well as details of how they occur, can be explained as side-effect of gene duplication and deletion mutations. (themes: 2, 3, 5)

– *Evolution of gene regulation ii*

Super linear scaling laws between the number of transcription regulators and the genome size had been observed. Analyzing the set of fully sequences bacterial genomes, in comparison with a simulation model of genome and regulome expansion, we have shown that this is due to sudden “bursts” or contractions of the number of regulators, which occur at the root of major bacterial lineages, whereas within most lineages genome size and regulome size scale almost linearly. (themes 2, 3, 5)

– *Genotype-phenotype mapping and evolvability*

A fundamental problem for the evolution of complexity is the information threshold, which shows that the information which can be accumulated in evolution is limited, given a certain mutation rate. We have quantified to what extend this problem is alleviated by non-linear genotypephenotype mapping; and we have shown that lethal mutants do not alleviate the problem as previously suggested. (themes 2, 3)

– *Multilevel modelling*

We further developed a very powerfully approach for multilevel modelling of biological cells, called the

“cellular Pots model CPM” (or alternatively by some people the GGH, Glazier, Graner Hogeweg model). This formalism allows us to model the morphogenesis, spanning sub-cellular to macroscopic scales, as well as the evolution of morphogenic mechanisms. A major recent extension is the coupling of the model to internal polarization cytoskeleton dynamics, and the development of tools to quantify the forces acting on the model cells. (themes 2, 4)

Data-driven multi-level models of infectious diseases

Prof. dr. O. Diekmann, diekmann@math.uu.nl, www.math.uu.nl/people/boldin/nwo/

4 The treatment of hospital patients suffering from bacterial infections is increasingly hampered by antibiotic resistance. When the means for curing infections diminish, the prevention of infection gains importance. Thus, for example, Scandinavian countries and the Netherlands implemented already in the 1980's a “search-and-destroy” policy to prevent the rise of MRSA prevalence. If we want to ascertain the effectiveness of various potential control measures, we need to know whether cross-transmission or opportunistic growth is the predominant colonization route (in the first case, hand washing and other traditional infection prevention methods make sense, whereas in the second case a prudent use of the arsenal of antibiotics should be pursued). But how to assess the relative importance of these two routes to colonization? One way is laborious and costly: it combines epidemiological surveillance with genotyping. A good thing is that it provides us with a standard to judge how well other methods perform (some people call it the “gold” standard, although “gilded” may be better suited, since the results are not cast-iron). A central aim of the project is to develop a reliable tool for assessing the relative importance on the basis of readily available longitudinal data. Once developed, such a tool can both be used to ascertain a current situation as well as to determine the efficacy of control measures (by establishing the change in the relative importance). In order to develop this tool, medical doctors, microbiologists and applied mathematicians need to cooperate. Right from the start, a substantial investment of time and energy of all involved is required. Yet success is not guaranteed and language barriers as well as culture differences are substantial. The CLS program has acted as a catalyser of a somewhat improbable reaction and now, halfway, we can report with contentment that several mathematicians frequent the hospital and that a medical doctor is regularly coming to a mathematics institute. Joint publications have appeared and are in the process of being written. A rather unique collaboration is gathering steam. A major methodological issue is not the presence of multiple scales, but the mix of certainty (data about admission, discharge, results of cultures) and uncertainty (did colonization happen? when? how?). Moreover, note that the possibility of cross-transmission makes that patients are NOT independent from one another. Our approach is to exploit the full extent of what we know for sure and to model cross-transmission mechanistically. The consequence is that we need to develop a very flexible model formalism (glued

Markov chains) and to build algorithms for data analysis in that unconventional framework. A first paper in which this approach is elaborated and, as a case study, applied to analyse data concerning cephalosporin resistant Enterobacteriaceae, is submitted to the American Journal of Epidemiology. The reported results make us confident that the essential steps in the development of the tool have been taken. The problem of scale does arise in another aspect of antibiotic resistance, viz., the feedback to the hospital via the community at large (residence times in the ICU are in the order of days and in the hospital at most weeks, but carriage prevalence in the population changes at the time scale of years or even decennia). The influence of this feedback was analysed by devising a judicious mix of analytical considerations and simulations. Our conclusions pertain to MRSA and show what contribution the various components of the search-and-destroy policy make to its success. In a submitted paper we ventured to predict how a new variant, Community Associated MRSA, will fare. In an accepted paper a model formalism is developed to deal with the possibility of colonization at multiple sites (like skin, lungs, intestines). The formalism is then used to study, in carefully designed thought experiments, the relative contributions of within-patient versus between-patient transmission. Our findings represent a firm theoretical argument against the routine use of topical antibiotics.

Evolutionary approach to modelling the plant cytoskeleton

Prof. dr. B.M. Mulder, mulder@amolf.nl, www.amolf.nl/research/theory_of_biomolecular_matter/

The project is an integral part of a multidisciplinary collaboration between experimental cell biologists from Wageningen University (dr. Jan Vos, drs. Jelmer Lindeboom and master student Antonios Lioutas) and theoretical and computational biophysicists from AMOLF (prof. Dr. Bela Mulder, dr. Rhoda Hawkins and drs. Simon Tindemans, the latter funded by the CLS program). In the first phase of the project, the focus was strongly on the development of a base model to deal with the complexity of the inherently multiscale problem posed by understanding the dynamical architecture of the cortical plant microtubule. Individual microtubules, long filamentous polymers, are dynamical systems in their own right, as they can switch stochastically between growing and shrinking states. In the plant cortex they are moreover constrained to lie in an effectively 2-dimensional geometry, allowing them to interact through collision angle dependent dynamical events, causing segments of the microtubule either to reorient or to start shrinking. Moreover the overall state is also influenced by the presence of protein complexes that nucleate new microtubules, proteins that are able to sever microtubules and the fact that the dynamics takes place on curved, partially closed surface. On a global scale the cortical microtubules show a remarkable degree of ordering, as well as intriguing dynamics during the progression of the cell cycle. Our overall aim is to be able to describe these ordering processes, and nail down the microscopic factors that determine them. To that end we are employing a two-pronged approach. On the one hand, we have developed a particle-based simulation

tool for simulating the cortical microtubules, including all known effects. This program is currently coded in Matlab, but will be ported to C++ for efficiency in the near future. On the other hand we have set up a mean-field theory for the cortical dynamics, which also incorporates most of the dynamical effects. In principle, these two approaches are complementary, the simulations allowing for a detailed investigation of a necessarily limited number of specific parameter choices, while the analytical approach is able to resolve for instance linear stability issues as function of the parameters in a more global manner. Searching for specific solutions, corresponding to in vivo observed patterns, in both these systems, the simulation and the mean-field theory, is an inherently multiparameter optimization problem, with little or no a priori insight in what represent good trial solutions. In the second phase of the project the focus will therefore shift towards the development of tailor-made evolutionary algorithms to efficiently solve this problem.

Our project also has an interesting spin-off in the direction of a more general question: how can one design (bio)molecular systems for a specific purpose. This question belongs to the nascent field of biomimetics, where one tries to employ both the materials and the techniques from molecular biology to create novel functional materials often on a nanometer scale. The design process often involves the prediction of properties of complex multicomponent systems, similar to the cytoskeletal structures at the heart of our project. The ideas developed here can thus also be transplanted to other areas. We are working on a trial application that involves the design specific spatial patterns with using DNA-coated colloids that interact through direct or linker particle mediated, base-pair ligation. By choosing the composition of the DNA coat on the particles, one can create particles that can interact both specifically and in a completely tunable way with other particles. We are focusing on the question of how to choose these interactions in order to achieve specific patterns, and what limits there are to the type of patterns that can be designed in this manner. This again involves the type of efficient searches through large phase spaces, which we aim to tackle using our genetic algorithms. It is becoming widely recognized that understanding living systems at the level of the single cell requires more than just a compilation of the component parts (proteins, genes) from which they are built. Cells are dynamical systems that operate in space and time. The cytoskeleton of higher organisms, like plant, is a paradigmatic example. The various ordering processes that apparent in these structures emerge from the collective behaviour of their components, which arguably can only be understood by modelling their dynamics, which in turn requires the input from quantitative observations. Our project is close to achieving this synthesis for the specific case of the interphase cortical microtubule array in plant cells in a format which we believe will become an example of “best-practice” in cell biology.

6

Apart from the expected output in terms of publications, we will make our simulation package freely available to other interested parties. Already a number of biological groups working on the cytoskeleton have expressed their interest, as it is clear that it will be able serve as an in silico laboratory for testing hypotheses in this field.

Development of a computationally efficient model of the human heart

Dr. A.V. Panfilov, A.V.Panfilov@bio.uu.nl, www.binf.bio.uu.nl/~panfilov/CLS.html

Our project includes multidisciplinary cooperation between various groups from computational and life sciences. The planned cooperation included work with Dr. Nash from New Zealand. Working with him we obtained access to results of detailed clinical measurements of ventricular fibrillation performed in Great Britain and processed in his group. These results were used for validation of our whole ventricles model against clinical data. In addition we cooperated with Prof. J. Sneyd, from University of Auckland in development of a new version of our model for human ventricular cells. Prof. J. Sneyd is an expert on intracellular calcium dynamics modeling. For verification of our model of the Purkinje system we cooperated with Prof. T. Oostendorp from RUN who is specialist in modelling of electrocardiogram for wave propagation pattern.

We switched from OpenMP parallelization on shared memory supercomputers to MPI parallelization on beowulf clusters for reasons of both accessibility and generalizability. We developed the MPI code ourselves. We developed a mathematically reduced version of our action potential model for human ventricular cells. We obtained this model by removing intracellular ion dynamics and by using steady state assumptions for fast changing variables, reducing the number of variables in our model from 19 to 9. The new model is over four times more computationally efficient than the full model and therefore very suited for performing whole heart simulations, yet still detailed enough to study the effect of single ion channel mutations on arrhythmia dynamics and organization. As a consequence the new model is a helpful tool in spanning spatial scales from the subcellular to the whole organ level. We discussed possibilities of comparison of efficiency of different numerical techniques for whole heart modelling with a Simula Research Laboratory, Oslo, Norway. They also made for us a tetrahedron mesh of our ventricular model. We extended our previously developed electrophysiological model of human ventricular cells. The new model incorporates a more extensive description of intracellular calcium dynamics, which is assumed to play an important role in arrhythmogenesis, and is capable of reproducing clinically measured action potential duration restitution slopes, another property assumed to play an important role in arrhythmogenesis. The model is used to study the conditions for a particular type of arrhythmogenesis: spiral breakup. Our model is one of the few models available for human cardiac cells, it is widely used by groups over the world, available on our own website (www.binf.bio.uu.nl/khwjtuss/), and on a general repository for cell model codes (www.cellml.org).

Computational Analysis of Spatiotemporal Patterns of Activity in Neuronal Networks

Dr. J. van Pelt, jaap.van.pelt@falw.vu.nl, www.neurodynamics.nl

The CASPAN project aims at increasing our understanding of structure-function relationships in neuronal networks. The focus is on spatio-temporal patterns of electrical activity in neuronal networks, as they are considered to correspond essentially with cognitive brain functioning. We believe that this understanding helps bridging the gap between the areas of cognition and neurosciences. The project follows a computational modelling / mathematical analysis methodology, embedded in an experimental neuroscientific environment which provides the experimental data for analysis and validation. There are three subprojects:

1. Development of a macroscopic neuronal network model with realistic functional and structural connectivity to simulate neuronal activity in cortical brain slices and cultured neuronal networks.
2. Development of statistical methods for analysis and comparison of experimentally observed spatiotemporal activity patterns.
3. Development of a neuronal microcircuit model composed of neurons with full morphological complexity to investigate how the fine structure of synaptic connectivity contributes to the dynamics of neuronal activity.

8

From the beginning on the CASPAN group was strengthened by the collaboration with the statistician Dr. Fabio Rigat from the statistics group of Prof. Mathisca de Gunst. His interest in the biological questions of the CASPAN project has resulted in new Bayesian techniques for modeling firing activity in cultured neuronal networks. The progress in the CASPAN project has recently attracted the interests of two Master students Cognitive Sciences (Peter van Hees, Betty Tijms) who are now conducting stage projects in the CASPAN group.

Modelling tools are essential to investigate biological phenomena as originating from complex interactions between a multitude of mechanisms. With the NETMORPH network generator we will for the first time be able to investigate how the spatial overlap between axonal and dendritic branching patterns of neurons in a network determine the spatial distribution of potential synaptic connections, and to lay down the basis for synaptic connectivity. In a follow up phase, the neurons in NETMORH will be given detailed electrical properties in order to conduct systematic studies of how these structural characteristics determine network activity patterns. The detailed microcircuit models have made it possible to interpret experimentally observed changes in oscillatory activity in the brain in terms of ion channel kinetics in well-defined experiments with mutant mice.

Computational Views on Membranes and Ciliates

Dr. H.J. Hoogeboom, hoogeboo@liacs.nl, views.liacs.nl/

The aim of the project is to study basic aspects of life processes. The novelty of the research is the principal point of view that a number of basic life processes can be considered as computations. This point of view turned out to be very succesful in research in biology. Especially, in the area of Natural Computing, which has two “faces”: computing going on in nature, and human designed computing inspired by nature. This interdisciplinary research involves three areas of science: biology, computer science and mathematics. In order to focus the research we have chosen two research topics which are representative for the two faces of the Natural Computing research: (i) Gene Assembly in Ciliates and (ii) Membrane Computing. The research on the computational nature of gene assembly has turned out to be already succesful in shedding light on the basic nature of this process, which is one of the most fascinating examples of basic life processes. The research on the second topic has led to very interesting models of computation inspired by the role of membranes in living cells. Typically for research on “life processes as computation” point of view, the major mathematical tools used here come from discrete mathematics (rewriting systems, combinatorics of words, graph theory).

One of the goals of our project is to investigate how this research can be enriched by applying more mathematical methods and techniques from probability and nonlinear analysis. For example, current research might be enriched by applying state-of-the-art methods from dynamical systems. We believe that new developments in the analysis of dynamical systems that possess a discrete spatial structure, i.e., cellular automata, coupled map lattices and lattice differential equations, can also be applied in the present setting. An essential part of the project aims at getting a deeper and more complete biological insight into the above two research topics. In particular that means to understand the biomolecular implementation of various computational aspects of gene assembly (in other words, learning about the biological hardware - bioware - used by the ciliates to implement computation going on during gene assembly). Moreover for membrane computing we plan to redirect some research in into the realm of original biological motivation: the role and the functioning of biological membranes. As a matter of fact through this project we want to strengthen (and to get a sound understanding of) the mutual feedback between biological, dynamical and computational research of basic life processes. In particular, the study of the qualitative behaviour of these processes leads to fundamental questions in the theory of dynamical systems and probability theory, which can only be solved in close interaction with computational science. In order to better understand the mechanisms involved in the gene rearrangement in ciliates, the “reality and desire” diagram (or breakpoint graph) classically known from the field of computational molecular biology that studies the differences between gene orders in genomes, has been introduced into the Ciliate context. It turns out that

the structure of this graph predicts which loop recombination operations are involved in the process of transforming the original genome in the new one.

In the past the main focus of research in this area has been from the viewpoint of formal languages. Membrane computing was studied as abstract computational model, and its power was compared to that of classical devices. As planned the VIEWS project enlarged this narrow toolbox that was used to analyse these computational processes, modelled after natural phenomena. New techniques from molecular dynamics gave new insight in the configuration spaces of membrane systems. The authors of the paper involve researchers from Leiden, Amsterdam and Milano (Italy). Another explicit goal of the project is to “bring membrane computing back to biology”, which means not to study the topic as a pure computational device, but to use it to model and analyse processes in nature. One of the key publications in this direction studies the biological phenomenon of quorum sensing in bacteria using the framework and techniques from membrane systems.

Mathematics and Computation for the System Biology of Cells

Prof. dr. M.A. Peletier, mpeletie@win.tue.nl, www.siliconcell.net/sica/NWO-CLS/CellMath/

10

The approach of the project is a combined thrust of biology and mathematics. E.g., the aim of WP1 is to apply system reduction methods on models of biochemical networks, primarily metabolic networks. This work gains to show the usefulness and need of model reduction for this type of systems. The innovative character of WP1 is to investigate to what extent the theory of system reduction (an area within mathematical system theory) can be useful for learning to handle the size and complexity of mathematical models of biochemical systems, since those fields have not yet been extensively combined. The size of the models of small parts of the metabolic networks is of the order of 10 to 50 state variables while this type of models is getting larger. Models of complete cells might be in the order of 20,000 to 100,000 variables in the future. The aim of the project is to investigate what type of model reduction methods are suitable for those large biochemical systems. One approach will be to partition such networks into modules, and to apply system reduction algorithms to these modules. The mathematics of reducing rational positive systems, the mathematical models concerned, into models of the same type, does not yet exist. Theory for related classes of systems is being adapted to the class of rational positive systems. The available theory in the area of dynamical systems is neither applicable nor appropriate for the problems considered. In WP2 a particle-based spatial stochastic method (GMP) has been developed to simulate biochemical networks in space, including fluctuations from the diffusion of particles and reactions. Gradients emerging from membrane reactions can be resolved as well. GMP combines in an operator-splitting way the multiparticle method for diffusion (Chopard et al., *Int. J. Modern Physics C*, 1994) and the kinetic Monte Carlo method

(Gillespie, *J. Phys. Chem.*, 1977). The GMP-code has been made publicly available. The coupling between particle-based models and PDE-based models in different parts of the domain has been studied in WP3. Also a number of (publicly) available mesoscopic methods for solving reaction-diffusion processes in the cell has been compared to find the regimes of applicability of the underlying models and their computational implementations in biochemical networks involving a relatively small number of molecules. Both direct computations of the biological process and a reduced system can be used to analyse the dynamic behavior of the systems which then leads to conclusions on the reaction of the cell to particular stimuli. The interpretations can then be used to test hypotheses on the biological behavior of a cell. Both can also be used for network-based drug design, the determination of chemical substances which either block or stimulate particular chemical reactions in the cell so as to negatively influence viruses or bacteria.

Modeling the gene network underlying neuronal outgrowth

Prof. dr. M.C.M. de Gunst, degunst@cs.vu.nl, www.cs.vu.nl/~degunst/neurogene.html

This CLS project is a joint research project of the department of Mathematics and the Department of Molecular and Cellular Neurobiology of the Vrije Universiteit Amsterdam. The ultimate purpose of the research is to gain insight in the transcriptional network that underlies neuronal outgrowth. Thus, the main goal is to identify transcription factors that are key regulators of the response that drives gene expression after nerve injury, and the targets that are regulated by these factors. The specific aim of the project is to develop statistical and computational methods to help reach this goal. A two fold-strategy is used: i) building a statistical model that is used for the inference of gene regulatory relationships from time series of gene expression data *in vivo* on both successful and abortive neuronal outgrowth; ii) identification of DNA binding sites of outgrowth associated transcription factors in control regions of genes of interest using probabilistic DNA search algorithms. Our initial results will be used as input for new series of biological experiments with several different techniques (gene expression microarrays, ChIP-chip, RNAi) so that the new experimental data in turn can be used to refine the model. From the beginning of the project the research team has been working in close and continuous collaboration via regular work discussion meetings. Moreover, the postdoc Nicola Armstrong and the PhD student Geert Geeven, who had/has her/his base in the mathematics department, also took/takes part in the lab meetings of the biology group. During the first year of the project postdoc Dr. Nicola Armstrong has performed a literature and data base search for transcription factor binding sites (TFBS), and she has investigated the use of several methods for finding TFBS in the context of the project as planned. The main problems that she met were of a biological rather than a statistical or computational nature: the number of known TFBS for rats was too small to start with our preferred HMM approach, since this approach needs a sufficiently large training set to be able to infer new binding sites. Moreover, it turned out that data in the data bases were either very sparse

or unreliable for our biological model system, the rat. During the second year Dr. Armstrong performed several pilot studies of genomic sequence analysis based on small subsets of the genes of which the data for the upstream regions seemed to be reliable. Unfortunately, Dr. Armstrong left the team after 1 year and 9 months, but these studies served as a basis on which Drs. Geert Geeven, who joined the project after about 1.5 years after its start, could continue this part of the project. Fortunately, over the past two years, the information on the rat genome in the data bases has grown and improved. Dr. Armstrong also has worked on the preparation of the data sets for input in the model analysis to be used by Drs. Geeven in the project's other part on model building.

PhD student drs. Geert Geeven has designed a basic Bayesian network algorithm and tested it on selected simulated and experimental data sets. He also studied and tested other methods for network building, such as Graphical Gaussian modeling, on these selected data sets. A start was made of the verification of sequence search results by combining these with our extended gene expression data set. In the figures below an example of our first results to model a small part of the gene network underlying neuronal outgrowth in rat are shown. The genes are color-coded depending on the specific combination of over-represented transcription factor binding sites they are predicted to have. The genes are represented by nodes of different sizes depending on the degree of over-representation and edges between genes represent strong correlation in expression.

12

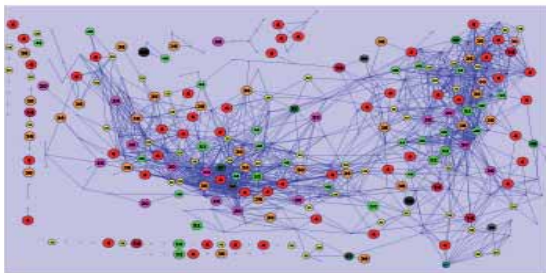


Fig. 1. A cluster of up-regulated genes. Explanation see text.

The outcome of the network modelling and sequence search will consist of probabilistic relationships between transcription factors and genes. The biologists will use the results to verify predicted relationships experimentally. Already now we have obtained some interesting preliminary results in the form of candidate transcription factors that are likely to regulate other genes in our study (see figure). Some findings confirm known facts, but some others are new.

Understanding the Organic Carbon Pump in meso-scale ocean flows

Prof. dr. S.A.L.M. Kooijman, bas@bio.vu.nl, www.bio.vu.nl/thbl/research/project/cpump/

Three groups with very different backgrounds collaborate in this innovative project: theoretical biology (VU, Amsterdam), dynamical oceanography (IMAU, Utrecht), numerical mathematics (CWI, Amsterdam); apart from frequent direct contacts, the group as a whole meets 4 till 5 times per year. This collaboration is essential for contributing to the complex problem of quantifying the carbon flux from the atmosphere to the ocean. Realistic transport in oceans requires a high spatial resolution, which does not combine well with a complex specification of the behaviour of the biota and of the chemical compounds. The time integration of the dynamic equations requires special features such as to accurately represent the behaviour of the state variables, to preserve certain physical quantities, and to avoid negative (and hence unrealistic) values for amounts. Apart from the design of new and more efficient numerical schemes, we invested in the development of ecosystem models that takes biodiversity into account, but avoids the evaluation of all species separately; realistic formulations for autotrophic and heterotrophic activity are key to this project. We found useful formulations for the water column, and now try to reduce the complexity further for implementation in circulation models. We also studied the effect of transport in the spatial structure that is created by eddies on primary production, and found that growth is large at the edges of eddies. We think we now understand this phenomenon in depth. Last but not least we tried to estimate carbon fluxes on the basis of rather crude assumptions on mass balances and quasi steady states on transport without dealing with all the complexities on transport and ecosystem dynamics. This gives us a reference for comparison with our more elaborate study on much smaller scales in space and time. We think that the simultaneous understanding of the process at small and large space-time scales is essential. A three-dimensional ocean solver coupled to biological advection-diffusion-reaction equations has been implemented. The advection term in the tracer-equations has been discretised with a third-order upwind Discontinue Galerkin method for advection. The code was tested with three-dimensional analytic solutions.

13

Simulation of developmental regulatory networks

Dr. J.A. Kaandorp, jaapk@science.uva.nl, www.science.uva.nl/research/scs/3D-RegNet/

In this project we are developing models for simulating regulatory networks that are capable of quantitatively reproducing spatial and temporal expression patterns in developmental processes. The model is a generalization of the standard connectionist model used for modelling genetic interactions. The parameters in the model are inferred from spatio-temporal gene expression data. This type of data can be obtained using a technique known as in situ hybridization (see Figs 1 and 2a for examples). By comparing the quantitative expression of genes in the actual data set to the modelled data, the parameter values in

the model can be estimated in an optimization process. The model will be coupled with a biomechanical model of cell aggregates and used to study the formation of spatial and temporal expression patterns of gene products during development in cellular systems. As a case study we are currently using the body plan formation in early development of *Drosophila* and in a relatively simple multi-cellular organism (sponges and scleractinian corals).

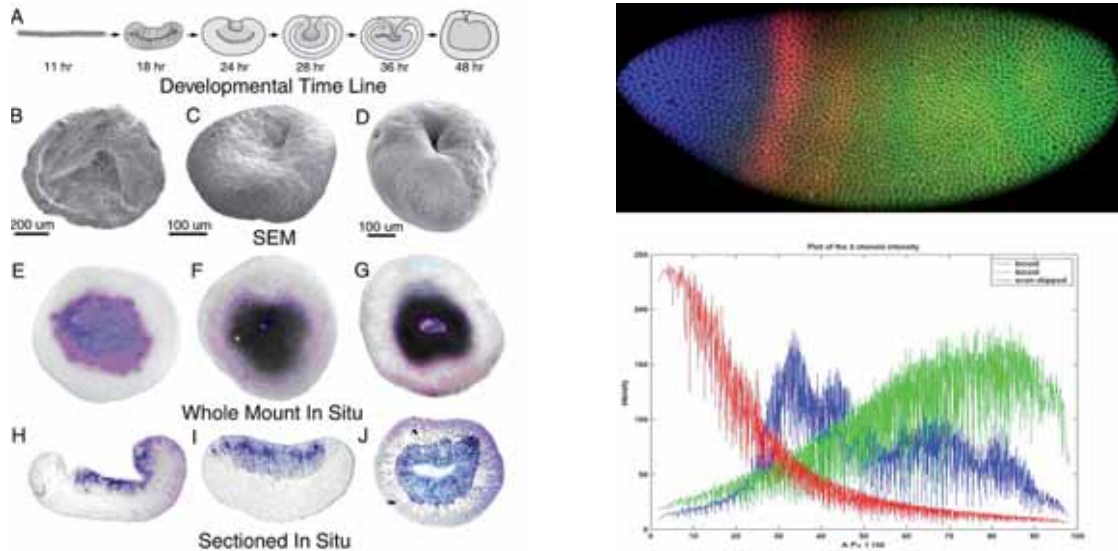


Fig. 1 (left). In situ hybridizations of early development in the scleractinian coral *Acropora millepora* (E.E. Ball, D.C. Hayward, R. Saint and D.J. Miller, Nature Genetics, 2004, 5, 567-577)

Fig. 2 (a: right, top). In situ-hybridisation of three genes in early development of *Drosophila*.

(b: right, bottom) Quantitative expression of the three genes in (A) along the x-axis

In the project we have a close collaboration with Prof. D.J. Miller (James Cook University Townsville, Australia) and Prof. W.E.G. Müller (Johannes Gutenberg-Universität Mainz, Germany). Through this collaboration we could have access to in situ hybridizations of early developmental stages in the scleractinian coral *Acropora millepora* (a typical example of a set of in situ hybridizations of *Acropora millepora* is shown in Fig. 1) and in situ hybridizations of genes involved in development in sponges. The quality of the data set is crucial in order to infer regulatory networks using optimization techniques, it is required that there are sufficient in situ's available for every gene involved in the regulatory network and there are pictures available of successive developmental stages. We have developed a number of new methods and techniques within the project:

1. A cell-based technique for modelling spatio-temporal gene expression in an aggregate of cells (M. Ashyraliyev, J.G. Blom and J.A. Kaandorp). Currently we are working on a paper on this discrete-continuum model to describe reaction diffusion processes during development. A numerical method to solve this model has been developed, analyzed and implemented.
2. A new technique has been developed for inferring model parameters from actual spatial temporal gene expression patterns. The method is based on an evolutionary strategy and is significantly more efficient compared to optimization strategies as for example simulated annealing.
3. In collaboration with Prof. C. Klaassen and Dr. N. Lalam a new method has been developed for the statistical comparison between actual and simulated data sets based on pseudo maximum likelihood estimators.

Spatio-temporal modelling infochemicals in a food web context

Dr. L. Hemerik, Lia.Hemerik@wur.nl, www.biometris.wur.nl/UK/Staff/Lia+hemerik/

For the development of the spatio-temporal food web model researchers from Biometris and from the Laboratory of Entomology cooperate intensively. Within Biometris the PhD Ir. M. Lof is supervised by Dr M. de Gee and Dr L. Hemerik. Prof. J. Powell and Dr R.S. Etienne are mathematical advisors at a distance. During the second year M. Lof stayed for a month in Utah, where she closely cooperated with Dr. J. Powell. Every two months M. Lof has a meeting with Prof. Dr. M. Dicke, where entomological details of the model are discussed. Dr B. Wertheim is frequently contacted by e-mail to share her biological knowledge on the system with our group. The project greatly benefits from input from the entomologists on the one hand and from the mathematicians on the other hand. The mathematical investigators develop a good intuition for what is biologically relevant, whereas the entomologists gain some knowledge on what is mathematically feasible or possible.

The integration of the chemical behavioural and ecological modelling approach requires the connection of processes on different spatial and temporal time scales. The temporal development of odour plumes has been intensively studied at large spatio-temporal scales for investigating how environmental pollution is distributed in the atmosphere and subsequently deposited on land through rain showers. This kind of modelling is therefore linked to large spatial and time scales of at least 100 metres, respectively several minutes. For the development of an odour plume arising from one rotting apple in an orchard, the currently available technique is unfortunately not directly applicable, because the temporal and spatial scales are too small (less than 1 minute, 1-10 metres respectively). For studying the influence of infochemicals in a food web we started with comparing a Gaussian plume model based on time-averaged values with a turbulent filamentous plume model based on the release of odour puffs at different regular intervals. Although we presumed that the same chemotactical reaction of the virtual fruit flies to both

these plumes models should result in approximately the same distribution of fruit flies, we have found that it makes a big difference (Lof et al, 2007). Chemotactical reaction of the fruit flies to a filamentous plume model gave a far more realistic spatial distribution over apples than such a reaction to the time-averaged Gaussian plume model. In the near future we compare some other plume models with the two already investigated. Based on the results (with respect to realism and computation time) we choose which one to use in our future research. In the preliminary model published by Etienne et al. (2002) the establishment of a *Drosophila* population was investigated. This model is refined and extended by incorporating chemotaxis and a more realistic model for the patch leaving by adult *Drosophila* fruit flies in two papers recently submitted to the Bulletin of Mathematical Biology (Lof et al. (submitted) and de Gee et al. (submitted)). The resulting model is used to investigate the effect of chemotaxis on overcoming the Allee effect and competition for the survival of the larval population. Three different situations are compared: (1) adult fruit flies disperse randomly, (2) they react to food odours only or (3) they react to a combination of food odours and the aggregation pheromone produced by adult *Drosophila* fruit flies. When fruit flies react to the combination of odours, the mortality due to the Allee effect greatly diminished. The effect on the competition was less clear. This result suggests that the ability to react upon aggregation pheromone has mainly evolved to enhance the ability to re-colonize an area after the winter, because *Drosophila melanogaster* cannot survive real winter temperatures outside: each year they have to make a restart from the cities, where they manage to survive the winter within houses.

16

New results have been reported on how yeast and detrimental fungi interact on apples. The growth of a yeast population provides food for the developing *Drosophila* larvae, while the fungi compete with these larvae for space on the apple and might harm the larvae. Dr Marko Rohlfs is specialised in this field. Therefore, we plan to invite him for exchanging information on the nature of these biological relationships, before we start to model this subpart of the food web model.

Amyloidogenesis: a computer simulation study

Prof. dr. S.W. de Leeuw, s.w.deleeuw@tnw.tudelft.nl, www.tudelft.nl/live/pagina.jsp?id=7bd9034f-8f91-4998-9170-a785251ad3f1&lang=en

Amyloidogenesis can be described as the aggregation of misfolded proteins to form amyloid fibrils. The process is associated with a host of diseases, including Alzheimer's ($A\beta_{1-42}$) and Huntington's disease. These fibrils exist for many proteins, independent of protein architecture, and they show many similar features. The proteins adopt a β -strand conformation, which interact inside the fibrils to form protofibrils consisting of long intermolecular β -sheets. Fully matured fibrils consist of several intertwined protofibrils. While the length of the fibrils varies considerably their diameter is remarkably uniform. In this project we study the process of amyloid formation through modelling and computer simulation. A statistical-mechanical model,

which combines the conformational transitions with the theory of self-assembly, has been developed. This model can be solved exactly and it gives an adequate description of fibril growth along the aggregate axis. It requires four free-energy parameters. We consider monomers, which are in a more or less disordered state, filaments, which are linear assemblies of protein molecules, in which these molecules can have a β -strand character or a disordered one, and fibrils, which consist of several laterally interacting filaments:

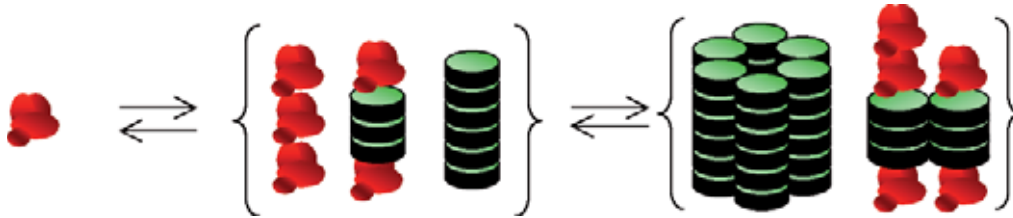


Fig. 1. Schematic reaction path for amyloidosis. From left to right: protein molecule, filaments and fibrils.

This is shown schematically in Fig. 1. Here, blobs indicate molecules in a disordered (non- β) conformation, whereas disks represent monomers in a β -strand state. Applying our model, we can determine, exactly and analytically, the mean fibril and filament lengths, and the fractions of monomer, fibril and filament in (dilute) solution. We find that there appear to be two regimes as a function of concentration and binding energy: one where fibrils dominate, and one where smaller aggregates do. The transition between these regimes can be quite sharp and becomes sharper if we allow more filaments to be present per fibril. We further find that thinner fibrils than those containing the maximum allowed number of filaments are repressed. The model developed for the aggregation of proteins into amyloid fibrils requires four free energy parameters. In addition prediction of the pitch requires the value of Young's modulus of elasticity. Computer simulation techniques are developed to compute these quantities from atomistic models. The Gromos package has been used to carry out the molecular dynamics simulations. Free energies of dimerization and elongation have been computed for small proteins using the method of overlapping distributions. To our knowledge this technique is applied for the first time to protein aggregation. In Fig. 2 we show a set of snapshots of the dissociation of the hydrophobic part AB_{17-21} (LVAFF) from a filament in solution. Only the protein framework is shown. The LVAFF protein is kept in a β -strand conformation until fully removed from the filament. Finally the β -strand constraint is released and the LVAFF protein is allowed to go to its native state in solution.

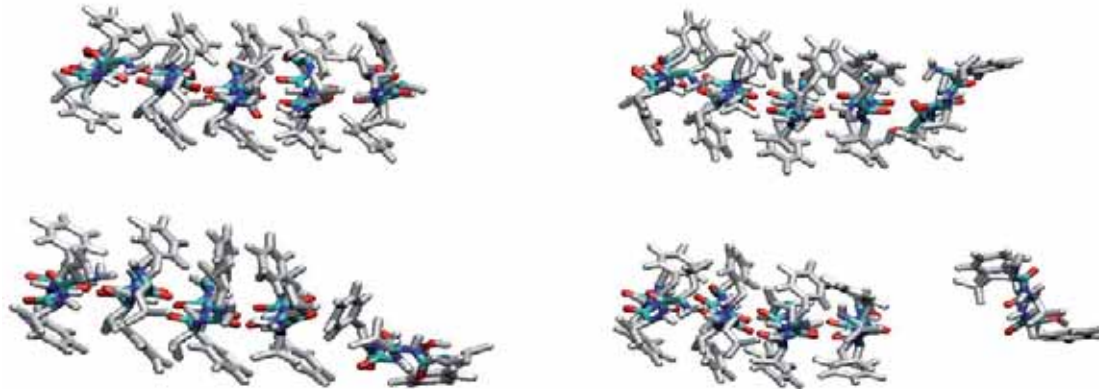


Fig. 2. Dissociation of $A\beta_{17-21}$ (LVAFF) from a filament in solution. Only the protein framework is shown.

The method gives reasonable values for the free energies of dissociation ($\sim 15 k_B T$). Unfortunately it becomes rapidly very time- and memory-consuming. It is therefore only feasible for small proteins, typically containing up to 8 or 9 residues. To study the elongation of the full $A\beta_{1-42}$ filaments techniques have to be developed to accelerate the rate of partial (un)folding and association. The formation and dissociation of hydrogen bonds is one of the rate-limiting steps in these processes. An extended version of Targeted Coordinate Dynamics (TCD) has been developed, in which hydrogen bonds are selectively broken and reformed at an accelerated rate. This leads indeed to an increased folding rate. Below the formation of a β -hairpin in G-protein is shown using this technique. Experimentally it takes $6 \mu s$; molecular dynamics simulations yield $4.7 \mu s$ for β -hairpin formation in G-protein. Using TCD the formation of a β -hairpin takes only 20-30 ns, a more than hundredfold increase in the rate of formation!

18

The technique is presently applied to the $A\beta_{1-42}$ amyloid protein the formation of $A\beta_{1-42}$ oligomers and elongation of the $A\beta_{1-42}$ filaments. In collaboration with the group of Dr. Ten Wolde from the University of Amsterdam the technique is combined with replica exchange molecular dynamics to compute the free energies of these processes.

Numerical bifurcation analysis of community models including the study of global bifurcations

Dr. ir. B.W. Kooi, kooi@bio.vu.nl, www.bio.vu.nl/thb/research/project/globif/

Ecosystem models are often formulated as continuous-time dynamical systems described by (autonomous) ordinary differential equations. This makes the application possible of bifurcation theory, where the dynamics of the system is studied as function of its parameters. At bifurcation points the qualitative long-

term dynamical behavior changes under parameter variation. For the study of ecosystems especially two local bifurcations are important. The so-called transcritical bifurcation (one eigenvalue of the Jacobian matrix equals zero) marks boundaries of regimes where a top-predator can invade an existing ecosystem. Hopf bifurcation (two conjugated imaginary eigenvalues) separates regions where the system possesses a stable equilibrium and shows cyclic behavior. In cooperation with the group Complex Systems, Oldenburg University, Germany, these local bifurcations are studied in terms of system's stability.

Ecosystem models consisting of more than two populations can display not only equilibria and limit cycles, but also more complex long-term dynamic behavior, including chaotic. It is known that in such dynamics global bifurcations play an important role, especially in the so-called crises where under parameter variation chaotic behavior disappears abruptly. Global bifurcations thus mark regions in parameter space where chaos is to be expected. In contrast to local bifurcations, for the analysis of global bifurcations information around the equilibrium points is not sufficient and the calculation of nontrivial orbits is required. The calculation of a global connection and the continuation thereof is, therefore, much more complicated than the calculation and continuation of local bifurcations. For local bifurcation analysis computer packages such as content and AUTO can be used. For global bifurcations, only one computer package HomCont, implemented in AUTO, can be used to study various types of orbits connecting hyperbolic and non-hyperbolic equilibria (i.e. point-to-point connections). Global bifurcation associated with point-to-point connections occur in some predator-prey system models. The Allee effect leads to bistability properties in the models. In a study in cooperation with Biometris Wageningen University it has been found that in Allee models point-to-point heteroclinic connections exist that lead to a phenomenon termed overharvesting. When the prey is consumed by a predator, our analysis shows that there is only coexistence when the mortality rate of the predator is in an intermediate range, while surprisingly with low mortalities a total collapse of the system occurs. The global bifurcation forms the boundary in the parameter space where this overharvesting occurs. These global bifurcations were also found in models for a toxicity stressed predator-prey system. Physiological parameters such as cost for growth or maintenance depend on the toxicant concentration in the organisms that make up the populations. The DEBtox approach gives expressions for this dependency. In addition to the model for the populations, a mass balance model formulation is used to describe the transport of the toxicant and the uptake and elimination into and out of the organisms. In this integrated approach effects due to toxicity stressors, environmental stresses (temperature) and ecological stressors (predation pressure) are studied simultaneously. The bifurcation analysis of the model based on the Dynamics Energy Budget theory showed global bifurcations associated with complex dynamics. Computer packages for the calculation of point-to-cycle and cycle-to-cycle connections are not available yet. An important aim of this project is to develop numerical algorithms for these connections and implement them in AUTO. Currently, techniques for the continuation of the heteroclinic point-to-cycle connections have been developed. Mathematically such a connection is a solution

of a boundary value problem. The connection orbit itself is described by the set of ODE's augmented by the boundary conditions at both ends. These boundary conditions are formulated in terms of the periodic orbits and their eigenfunctions and are included in the boundary value problem. To allow numerical treatment, the problem is truncated into a finite time interval and suitable boundary conditions are obtained by using linear approximations near equilibrium points and/or limit cycles. A homotopy method is used to find appropriate starting points to find a connection from which the continuation process starts. Files with scripts and a manual to run examples (Lorenz system and a tri-trophic food chain model) are downloadable from the project website: www.bio.vu.nl/thb/research/project/globif/ and a manuscript describing the results of the developed techniques has been submitted.

Bone-cell based computer simulation of skeletal morphogenesis and osteoporosis

Prof. dr. ir. H.W.J. Huiskes, h.w.j.huiskes@tue.nl, www.bmt.tue.nl/research/

We present a simple bone remodeling theory: strain-induced osteocyte signals inhibit osteoclasts and stimulate osteoblasts. In computer simulations, using Finite Element Method for strain calculation and Glazier & Graner method to simulate cell behavior, we show that this theory explains several aspects of bone remodeling and bone morphology:

1. orientation of osteons
2. orientation of trabeculae
3. the direction of resorption in the BMU
4. the coupling of formation in the BMU
5. bone repair: why osteoclasts target microfractures / dead osteocytes
6. why osteons are wider in bone regions that experience low strain magnitude
7. formation curve: why bone formation slows down as osteoblasts fill the BMU cavity

The theory shows that remodeling in cortical and cancellous bone are governed by the same mechanism. Thus, it makes clear that cortical and cancellous bone are not really different tissues, but rather different morphologies of the same tissue, resulting from a difference in strain magnitude).

Further, we resolve a paradox: osteoclastic resorption appears to target both unloaded and microdamaged (result of excessive loading) bone. In our model both cause a lack in the osteocyte signal that inhibits resorption:

- no loading, no strain-induced osteocyte signal
- microdamage is believed to cause osteocyte death and/or disrupt osteocyte signalling. Thus, both resorption "targets" are simply explained by one mechanism. The combination of cellular-automata (CA) models and finite-element (FE) based bone-adaptation models.

Mining factors of Celiac

Prof. dr. A.P.J.M. Siebes, siebes@cs.uu.nl, www.cs.uu.nl/groups/LDD/

The goal of the project is to develop data mining algorithms that help solving Life Sciences problems by mining the vast collections of available Life Sciences data. In other words, the interface between the two disciplines in the project is formed by the data analysis problems and the associated databases.

A concrete problem in genomic research in general and in our joint project in particular is the discovery of regulatory networks: how do genes interact? There are two primary types of data sources that can be used to answer this question, viz.,

1. Protein interaction databases, this are databases in which both laboratory verified and predicted interactions are stored
2. Micro-array experiment results

The first type is clearly directly relevant, if an interaction is verified in the lab, one knows that the genes interact. Unfortunately, if a potential interaction is not present in such a database, it doesn't mean that no such interaction exists. It may simply mean that no-one has tested whether or not the proteins interact. Moreover, there are no databases that store pairs of proteins that have been proven not to interact. Therefore, the second type of databases is also very important. Micro-array experiments show, e.g., co-regulation of genes and in this way they shed indirectly light on the question how genes interact. Given the relevance of this problem for the discovery of the genetic factors of Celiac disease, the analysis of both types of data plays an important part in the collaboration in this project. To date, this has resulted in two joint publications of Life Scientists and Computer Scientists.

The analysis of gene interaction discussed above has given rise to interesting new developments. If gene A interacts with gene B, while gene B interacts with gene C, one might measure an interaction between A and C, while this is only indirect and should not be modelled as a direct interaction. By analysing, e.g., the mutual information between the three pairs AB, BC, and AC, one can exploit a triangle inequality to filter out AC. The development of such techniques is an essential part of our research. The lack of negative examples, makes testing the methods on real data rather difficult. The second line of development is targeted at a higher level problem: Life Science data doesn't come in a single table. Rather, the data to be analysed comes from various data sources that have no unique translation to the standard data analysis format of single table data. Already for a few years such data analysis problems have been studied, both in the data mining community in general and in our group, under the heading of relational data mining. In this approach, there is most often a central table to be analysed, together with a collection of associated tables that provide further information that might help in the task at hand. Some promising algorithms have been developed, but they only aggregate the information in the associated tables. In the context of this project, we develop methods that allow for the use of more general patterns in the associated

tables. A key problem for this approach is that the number of patterns in these associated tables tends to “explode”, i.e., to be very high. It is not uncommon that the number of patterns is orders of magnitude larger than the number of rows in the table. Simply using all these patterns in the mining process would make the search space far too large. In other words, one should select the informative patterns only. Many syntactic solutions, such as closed patterns, have been developed for this, but while these are often far fewer than the collection of all patterns, there still tend to be far too many. We have developed another approach based on the MDL principle: the set of patterns that compress the database best is the most informative set of patterns. In experiments, this approach showed to reduce the number of patterns by many orders of magnitude. The MDL principle implies that this small set of patterns is the most informative. Still, we decided to test this in an alternative way, using classification. In a subsequent paper we have shown that a simple classifier based on the small set of patterns performs as well as top-of-the-line special purpose classifiers. In other words, the small set of patterns form a good characterization of the data in the database. Direct tests on Life Science data of this approach will be performed and published in 2007/2008.

The emergence of biocomplexity: a steady state between physical and biotic evolution?

Prof. dr. H. Olf, h.olf@biol.rug.nl, www.rug.nl/biologie/onderzoek/overonderzoek/onderzoekgroepen/cocon/projects/biomerger/

22

The understanding of the structure of ecological communities is essential in the light of the global current loss of biodiversity, and the potential impact of that loss on the functioning of critical ecosystems. Ecological communities consist of many components (individuals, populations, species) that interact non-linearly on different spatial and temporal scales, often with strong legacies of history. Furthermore, biota in ecological communities are often limited by, but in turn also actively affect, a large number of physical factors in their environment, such as water, light and nutrients. Most current models and simulations in ecology ignore this complexity by focussing on just a few species, a few resources, and little spatial structure, mostly because more realistic models quickly encounter computational and conceptual limits. However, recent insight in complex adaptive systems, mostly coming from physics and chemistry, and new computational tools may overcome these limitations. These new insights and methods allow complex systems to be summarized in emergent properties, such as scaling laws, which allows the mechanistic prediction of complex higher order phenomena from lower order processes. Specifically, we hypothesize that ecological communities maintain a thermodynamic steady state between the forces of physical evolution (increasing entropy) and biotic evolution (decreasing entropy), and that complex dynamics and pattern in such systems can be captured through emergent properties at different levels of organisation. In the proposed project, we will test this idea for systems with increasing compositional (more species) and patial complexity (more pattern) through modelling and simulation. This will involve powerful computational methods, such as parallel computing,

to deal with numerical complexity, the use of new 3-D visualisation techniques and a broad spectrum of advanced simulation algorithms (cellular automata, kinetic Monte Carlo, partial differential equation modelling).

Biomolecular Informatics (BMI)

The Biomolecular Informatics Programme (BMI) is a joint research program funded by the NWO councils Chemical Sciences, Medical Sciences (ZonMW), Earth and Life Sciences and Physical Sciences. The starting budget in 2001 amounted to 9 million euro.

The central aim of the BMI programme is the enforcement of BMI in the Netherlands, by:

- stimulating the formation of specialized BMI research groups;
- the supply of BMI expertise to BMI research groups.

Instruments to achieve these aims were the provision of project grants, fellowship grants and the support of the Bio-ASP (Application Service Provider), a central BMI facility in the Netherlands. The 11 projects below with a total budget of 4.9 million euro were granted in 2001. By the end of 2002, the BMI Programme and the remaining budget went up in the Netherlands Genomics Initiative (NGI). Fellowships were granted by NGI. The Bio-ASP was made part of NBIC. The last part of the budget is being applied for the new round of the Computational Life Sciences Programme.

24

A sequence-based human transcriptome map: identification and computational analysis of landmarks in the human genome

Dr. A.H.C. van Kampen, a.h.vankampen@amc.uva.nl

The project aims at the establishment of an integrated bioinformatics, computational and genomic research infrastructure. We will apply bioinformatics tools to integrate the human genome sequence with high throughput mRNA analysis data to generate a complete expression profile of the human genome. This will be used to identify long-range transcriptional domains. We will analyze the DNA sequences that define the transcriptional domains and control long-range transcriptional activity. A similar transcriptome map for the mouse will be constructed to complement these studies. Establishment of new approaches for the implementation of database applications and the development of computational analysis methods for large DNA data sets are integral part of these aims.

Evolution of the antigen presentation system and pathogenic evasion

Prof. dr. P. Hogeweg, p.hogeweg@bio.uu.nl

The aim of the project is two-fold. First, we study the function and the evolution of antigen presentation and processing in the adaptive immune system of higher vertebrates. Second, we identify the selection

pressure that antigen presentation imposes on pathogens. Our objective is to develop bioinformatic tools to:

1. evaluate the selection pressure caused by rapidly evolving pathogens had an impact on the evolution of the antigen presentation system.
2. identify which steps of antigen presentation and processing impose the strongest selection pressure on pathogens.
3. identify the mechanisms that facilitate the presented self antigens to be as different as possible from the presented non-self antigens.

Besides addressing these evolutionary questions, the tools developed in this project will be useful for other immunologists that work on identifying immunogenic regions in pathogens or in the human genome for autoimmune diseases.

Relevance for BMI

In the last decade genomic information of several pathogens has emerged. However, these data by themselves are not of much use for immunologists. Tools are needed to identify the immunogenicity, and to understand pathogenic evasion. The tools that will be developed to study the evolution of the antigen presentation system will be made publicly available, and they will be very useful for other immunologists that design vaccines, or T cell based therapies. By identifying the step in antigen presentation system that is most often exploited by pathogens for immune evasion, therapies to prevent escape mutants in many persistent infections can be designed.

25

Wageningen Phytoinformatics: the added value from plants

Prof. dr. W.J. Stiekema, willem.stiekema@CBSG.NL

The NWO-BMI programme 'Wageningen Phytoinformatics, the added value from Plants' will integrate informatics and statistics with plant genomics and breeding in co-operation with the Bio-ASP (Application Service Provider) facilities and existing worldwide data. In a combination of interrelated and multidisciplinary projects the program aims to

- (1) Establish a Dutch Expertise Centre for BMI research into plants, plant-related and plant-dedicated data;
- (2) Advance the BMI implementation in plant research,
- (3) Capitalise on genomics facilities and data of Wageningen UR;
- (4) Create a research infrastructure for education and training of young researchers in the field of plant biomolecular informatics.

It has as key objectives to develop a prototype reference information model for linking, querying and access-control of plant-dedicated databases (PlantNet). This system will be combined with a user-friendly plant-dedicated jumpstation (PlantMine) to facilitate access to and optimal functional interpretation of public plant-dedicated datasets. This set-up will contribute to the functional annotation and interpretation of the non-coding parts of the genome of *Arabidopsis thaliana* (Ara-mining) and other plants. For *Arabidopsis* and the crop potato, the further linking of phenotype with genotype will be accomplished by implementing the tools of statistical genetics (notably QTL mapping tools and alternatives) to data generated by genomics and breeding (genetical genomics).

Flexworks: A Framework for Flexible Data Integration to Support System Biology

Prof. dr. W.J. Stiekema, willem.stiekema@CBSG.NL

The overall aim of the BMI *Flexwork Project* – Framework for Flexible Data Integration to Support System Biology, has been to combine data from different bio-molecular techniques on a genome-wide scale in a flexible manner. The idea was that the project would allow users to define their own data analysis/gathering by linking bio-informatics tools in a Virtual Laboratory (VL) environment that could be approached via GRID.

26

Unraveling regulatory circuitry by combinatorial analysis of genome-scale data

Prof. dr. F.C.P. Holstege, f.c.p.holstege@med.uu.nl

The overall aim of this project is to investigate genome-wide regulatory circuitry (transcription and signal transduction), by concurrently developing and performing bioinformatic analyses of various genome-scale data sets: expression microarray data; chromosome localisation microarray data; protein-protein interaction data; high throughput phenotype data; sequence data as well as data currently stored in literature and WWW databases. This requires work on algorithms, their comparative, statistical and biological validation as well as research into how database management systems should be designed to hold and query different kinds of biomedical data. Data will be analysed from (our) previously published and ongoing studies, all concerned with understanding cellular regulation.

(The key (biological) objectives are listed in the subheadings below.

1. Identification of regulatory factors and regulatory mechanisms through DNA motif pattern recognition.
2. Identification of related function of genes by pattern recognition of expression profiles of mutants.
3. Identification of regulatory function and mechanisms by pattern recognition of causal changes within (physiological) datasets.

4. Combining different types of genome-scale data to answer the same biological question.
5. A database management system (DBMS) specifically designed to hold various types of biomedical data, allowing implementation of 1-4 in a robust, efficient and user-friendly manner. This system will also be designed to make novel “off the cuff” queries relatively easy and will take into account computationally intensive, reiterative queries.
6. Tools will be validated by wet lab experiments whenever feasible.
7. Besides using previously published data for developing tools and for biomedical discovery, it is a goal of this program to analyse new data generated by ourselves and by the advisory and collaborating groups.)

Computational genomics of prokaryotes

Prof. dr. R.J. Siezen, R.Siezen@cmbi.kun.nl

This project aims to reconstruct the cellular processes, metabolic potential (metabolome) and regulatory networks of selected gram-positive Bacteria and Archaea, by in silico analysis of all proteins encoded by their chromosome. Initially the “in-house confidential” and publicly available genome sequences will be used, and later this analysis will be extended to genomes that become available during the project.

The results will be incorporated into “virtual cell” databases, consisting of modules that include the majority of genes and predicted encoded proteins, signalling and information pathways, transport systems, regulatory networks, metabolic routes and their possible interactions, but also the corresponding substrates, intermediates and products. This will provide a basis for the translation of the genotype into specific phenotypic traits of several lactic acid bacteria (*Lactococcus*, *Lactobacillus*, *Streptococcus*), bacilli, clostridia and Archaea (*Pyrococcus*, *Sulfolobus*). Subsequently, in silico comparative genomics of different prokaryotic microbes will provide a detailed picture of the distribution of the overall gene-pool among these organisms, the architecture of the metabolome and an inventory of both shared and unique genes and encoded properties of each species. In general, this should lead to important advances in the understanding of prokaryote evolution, and will contribute to the prediction of their metabolic functions. As a model case, special attention will be paid to the regulatory networks operating in *Bacillus subtilis*. Ongoing transcriptome analysis using DNA-microarrays of this bacterium provides a wealth of transcriptome data, for which bioinformatics tools will be used and developed to be able to identify and visualize in a dynamic way the underlying regulatory networks. The resulting system will then be applied for the determination of gene networks in other target prokaryotes.

Integrating homology and genomic association for pathway prediction from genome sequences

Prof. dr. M.A. Huynen, huynen@cmbi.ru.nl

Some of the most exciting developments in genomics are the new methods for the prediction of protein function and pathways. By having complete genomes we can not only interpret the function of a protein within its proteomic context, we can also predict new protein functions and pathways. The latter is based on the observation that genes that are associated with each other in genomes (e.g. in operons) encode proteins that functionally interact (e.g. they are part of the same metabolic pathway). The goal of the project is to combine the various new types of genomic-association based protein-function prediction with each other and with the existing homology-based types of function prediction. The work involves both large-scale “genomics” analyses and method development, as well as detailed studies of the variation and evolution of specific pathways.

G protein-coupled receptors: from sequence to function

Prof. dr. A.P. IJzerman, ijzerman@lacdr.leidenuniv.nl

28

G protein-coupled receptors (GPCRS) are the target of the majority of today's medicines. These receptors are membrane-bound proteins that serve as anchor points for hormones and neurotransmitters, such as the classical biogenic amines (nor)adrenaline, dopamine and serotonin. Given their importance it is remarkable that we know relatively little about their actual 3D structure. This is due to their overall lipophilic nature and membrane-bound localization, which makes crystallization notoriously difficult. This knowledge gap frustrates the traditional triad in protein research, i.e., that sequence determines structure and structure determines function.

The aim of this proposal is to introduce new technologies from computer science, developed for intelligent data analysis and efficient search processes in vast search spaces - such as the human genome - to the important field of drug target validation. In particular, the work proposed here, aims at using and combining these methods for directly linking receptor sequence with receptor function, thereby circumventing the difficult structural element mentioned above.

The key objectives in the current project are geared towards addressing novel and formidable questions and challenges, e.g.:

- can the massive flood of novel receptor sequences be stored and mined in a user-friendly and intuitive way?
- can distinctions be made between functional receptors and (products of) pseudogenes?
- can ‘orphan’ receptors be reliably classified in terms of ligand binding and signalling cascade?

- can the (type of) endogenous ligand(s) for such ‘orphan’ receptors be speculated on?
- are these receptors validated targets for novel drugs?

Linking and mining spatio-temporal databases of gene expression patterns of developmental model systems

Dr. F.J. Verbeek, fverbeek@liacs.nl

The overall aim of the research project is to produce a mechanism for storage and retrieval of patterns of gene expression using their spatio-temporal (in situ) characteristics and making this mechanism available to the research community as a novel tool (Verbeek et al., 1999). This tool, a database, enables to analyse patterns of gene expression in the coherence with the spatio-temporal developmental concept of the organism. In practice this means information management of research results in which we take genes as the pivot and develop data mining strategies that retrieve information from data resources of different animal model systems. Patterns of gene expression will be related to a standardised spatio-temporal model, i.e. a digital atlas (Verbeek et al., 2000), of the animal model system.

29

Molecular evolution of polyglutamine and other amino acid repeats

Prof. dr. W.W. de Jong, w.dejong@ncmls.ru.nl

Expansion of polyglutamine repeats in several unrelated proteins causes a variety of neurodegenerative disorders, amongst which Huntington’s disease. The repeats occur in shorter form in healthy people, but become pathogenic above a length of 30-40 glutamines, forming intranuclear inclusions. Our aim is to understand the origins and consequences of polyglutamine sequences in terms of: types or regions of genes that are prone to develop the underlying unstable CAG repeats; whether such repeats are found in orthologous genes in different species; the causes of CAG triplet expansions; whether and why these are read as Gln rather than Ser or Ala repeats; and the structural and functional aspects of the resulting polyglutamine sequences. Polyglutamine repeats being just one of the many types of simple sequence repeats, the evolution of CAG repeats will be studied in the broader context of amino acid reiterations in mammalian proteins.

HORIZON projects linked to bioinformatics

The Horizon Programme encourages groundbreaking genomics and bioinformatics research. The programme offers a breeding ground for the creation of new research themes, programmes and groups. The objective of Horizon Programme is twofold, on the one hand it aims to ensure the continuation of top-level genomics research in the Netherlands and on the other it strives to stimulate and nourish research talent by allowing them to pursue fresh and creative ideas.

The total programme budget amounts to 12.3 million euro for the period 2003 - 2007. This is divided over the two components of the programme: Horizon Projects (running) and Horizon Breakthrough Projects (completed). The purpose of the Horizon Breakthrough Projects is to assess the feasibility of an entirely new and principally innovative idea in genomics, bioinformatics or on their interface. The projects have a limited time-span (1 - 1.5 years) and provide scientists the opportunity to establish proof-of-concept on an issue that has so far not been researched extensively.

Following the Horizon (breakthrough) projects linked to bioinformatics are described.

Exploitation of genomics data to identify new disease genes

30

Prof. dr. M.A. Huynen, huynen@cmbi.kun.nl

In the genomics era, rather than knowing the functions of individual genes, we would also like to know how they interact with each other in protein complexes and pathways. Bioinformatics supplies us with a growing number of so-called genomic context methods that predict such interactions, based on the fact that functionally interacting proteins tend to also associate with each other in genomics data. They are e.g. encoded in the same genomes, are co-expressed with each other in multiple species. In this proposal I will take such genomic and functional associations between proteins to a higher level: that of the phenotype as reflected in multifactorial diseases. I will determine which are the genomic associations presented between genes involved in the same genetic disease. Subsequently I will identify new genes having the same genomics patterns and test their implication in similar genetic diseases. The experimental validation step will focus on the cobblestone lissencephaly group of disorders. Walker-Warburg Syndrome (WWS), Muscle-Eye-Brain Disease (MEB), and Fukuyama Congenital Muscular Dystrophy (FCMD), are the three major entities of this group that is characterized by severe brain malformations, eye defects, and congenital muscular dystrophy. Three genes have been identified to underlie some of the cases. These genes code for three glycosyltransferases that cooperate in the same glycan synthesis pathway and share a common target, (-dystroglycan. This proposal thus contains both method development to make specific predictions about new disease genes, as well as testing those predictions by analysis of linkage data and the finding of disease causing mutations, to obtain proof of principle.

Functional genomics of secondary metabolite biosynthesis: the Arabidopsis glucosinolate model

Dr. ir. M.G.M. Aarts, mark.aarts@wur.nl

Plants produce a myriad of secondary metabolite compounds that are beneficial for humans. High-throughput methods to study the genes behind the biosynthetic pathway of secondary metabolites in plants have not yet been used systematically. Among these metabolites glucosinolates are of special interest, mainly due to their anti-cancer and food flavour properties. Plants of the Brassicaceae family, which includes cabbages, oil seed rape, radish and mustard, among others, produce glucosinolates. From the Brassicaceae family, the genome of the model plant species *Arabidopsis thaliana* has been sequenced and impressive resources including good quality whole-genome microarrays, genetically amenable segregating populations and many bioinformatics tools are available. This project aims to exploit the knowledge and resources available for *Arabidopsis* to study the glucosinolate biosynthetic pathway by combining metabolic analysis, transcript profiling and QTL analysis. This study is expected to provide proof of concept of a robust high-throughput method to perform comparable analyses of metabolites biosynthesis pathways in a wide range of species.

Understanding the Eukaryotic genome: unravelling the functional role of RIDGES

Dr. A.H.C. van Kampen, a.h.vankampen@amc.uva.nl

Transcriptome maps link gene expression profiles to the genome sequence and thereby reveal expression activity along the genome. The human, mouse and drosophila maps revealed pronounced gene expression domains with significant increased or reduced expression as compared to the average genome. These domains were named RIDGES (Regions of Increased Gene Expression). We aim to develop a bioinformatics approach that allows a functional analysis of these expression domains. This approach is a first systematic step towards understanding the role of RIDGES in different (normal and diseased) tissue, developmental stages and biological processes. This will contribute to our knowledge of gene regulation in eukaryotic genomes. We take three steps. First we refine and improve our current statistical method to allow a more reliable and exact definition of RIDGE boundaries. Secondly, we will analyze the genes and gene classes within a RIDGE. As part of this analysis we will perform a comparative analysis of several genomes. Finally, we analyze the dynamics of RIDGES across different tissues, developmental stages and biological processes. This project will provide new research lines for bioinformatics and experimental approaches that aim towards the full elucidation of RIDGES.

Sequential pattern mining of protein sequences. The case of G protein-coupled receptors

Prof. dr. A.P. IJzerman, ijzerman@lacdr.leidenuniv.nl

Many bioinformatics tools and methods are available to assist in predicting the nature and function of a new protein sequence. Comparison with known sequences is essential, as is the case with sequence aligning, clustering, similarity searching, motif and/or profile matching. The starting point is mostly multiple sequence alignment, but this method, useful as it is, has important disadvantages, however. It cannot do without parameterization with its inherent ambiguities, is computer-intensive, often requires manual curation, and can be particularly difficult for a set of deviating sequences. To overcome such drawbacks of alignment we propose to develop and apply a new, efficient method to discover motifs from unaligned protein sequences, viz. sequential pattern mining. It is a data mining technique for the detection of sequential patterns (consisting of succeeding items) that are frequent in a given dataset. The technique has been successfully used for large data sets (from e.g., insurance companies), due to the increased efficiency that the current algorithms display. Sequential pattern mining has not been applied to research topics in the life sciences, e.g., protein analysis. Therefore, we aim to do just that in this proposal, i.e. to introduce sequential pattern mining into the world of bioinformatics. For proof-of-concept we will focus on G protein-coupled receptors, a large family of biologically and pharmaceutically relevant proteins that are excellent drug targets. Our longstanding experience with this protein family enables us to assess the biological relevance of the sequence patterns we derive with this new technique.

32

Novel prediction approach for plant microRNA genes: the proof of a miRNA is in the interaction.

Dr. ir. J.P.H. Nap, janpeter.nap@wur.nl

Micro(mi)RNAs are now known to be an important novel part of the regulatory circuits of cells. The proper identification of miRNAs and their targets from an organism is of fundamental importance for understanding gene regulation in future applications. Computational approaches towards miRNA identification may so far have been suffered from a pars pro toto fallacy: mistaking part for the whole. Predictions are biased by structural constraints suggested by the first miRNA sequences isolated and by the assumption of evolutionary conservation of miRNA sequences in divergent plant species. We here propose a novel approach based on the assumption that in different species non-homologous miRNAs can regulate orthologous genes. All orthologous genes of Arabidopsis and rice will be analyzed in small windows for the presence of complementary sequences elsewhere in the respective genomes that are able to adopt the relaxed hairpin structure characteristic for precursor miRNA molecules. Non-homologous miRNAs that are

predicted to interact with an orthologous target site should be considered potential miRNA candidates. The proposed research should result in a comprehensive list of predicted miRNA:target combinations. As proof of concept, all currently known miRNA genes and miRNA/mRNA interactions should be included in this data set. We expect that this novel bioinformatics approach will significantly enlarge the known set of potential miRNAs and miRNA/mRNA interactions.

Soft atoms: Computational protein structure refinement through unphysical sampling

Dr. B.C. Oostenbrink, c.oostenbrink@few.vu.nl

In order to fully understand biochemical systems and processes, the determination of three-dimensional protein structures is crucial. Both in experimental structure determination and computational approaches such as homology modelling, the structure of loop regions is most difficult to obtain. In this project, a new, efficient tool to optimize conformations of loop regions in protein structure elucidation based on molecular dynamics simulations is proposed. This method will provide insight into the ensemble of structures that is available to the protein and allow for an identification of low energy conformations. Rather than crossing potential energy barriers in the conformational space through an increase of the temperature (kinetic energy) for the whole protein, we propose to remove specific, localized, potential energy barriers in conformational space through the use of so-called, "soft" atoms. Simulated annealing and replica exchange methods that make use of localized softness parameters will be used to obtain a broad ensemble of conformations from a physically meaningful simulation. All relevant degrees of freedom, including an explicit treatment of solvent will be taken into account. The proposed algorithms will be implemented in the GROMOS simulation package that is used by academia and industry all over the world. As a proof of concept, conformations of loop regions in well-characterized high-resolution X-ray structures will be reproduced, after which the method will be applied to two types of biologically relevant protein structures. In future applications the proposed methods can also be applied to ab initio structure prediction (protein folding) of peptides and short proteins.

33

Discovering interactions and functional groups of oncogenic lesions

Dr. L.F.A. Wessels, L.F.A.Wessels@its.tudelft.nl

Retroviral insertion mutagenesis is a powerful method to identify oncogenes and tumour suppressors from tumours induced in mice. We recently developed a platform for identifying retroviral insertion sites through the combination of high-throughput amplification of sequences flanking the proviral insertions and robotized sequencing. This high-throughput approach enables us to screen a large number of tumours

in a very efficient manner resulting in a higher number of retrieved mutations per tumour. Development of novel bioinformatics to analyze this large dataset is required and will be the main focus of this project. Optimal identification of oncogenic events found in these tumours will be dependent on the development of robust methods to determine whether mutations in specific genomic loci are statistically significant. Furthermore, we would determine whether specific combinations of oncogenic lesions are found and whether oncogenic lesions correlate with the genetic backgrounds of the mice carrying the tumours. Moreover, combining information from different types of data regarding gene function or oncogenic lesions, we would search for any bias of insertions towards genes involved in particular biological processes or pathways. Previously, it has been shown that genes implicated in human cancers, for example, Bmi-1, Myc and p53 are found frequently mutated in retroviral insertion mutagenesis screens. Our screen together with appropriate bioinformatics should allow identification of novel oncogenes and/or tumour suppressors involved in tumorigenesis. Moreover, combination of our dataset with other sources of information will increase the efficiency of this process and identify pathways, rather than isolated genes.

Accelerated and automated QTL analysis for genetical genomics

Dr. ir. M.J.M. Rutten, marc.rutten@wur.nl

34

Genetical genomics aims to provide new insight into the genetic basis of transcript variation by linking transcript data of genetically different individuals to marker data of these individuals. This is achieved by QTL analysis (QTL = quantitative trait loci) that is, since long, the workhorse of quantitative genetics. Some QTL influence only transcripts from near the QTL locus (cis-acting), while other trans-acting QTL influence (many) transcripts from other parts of the genome. This information helps to unravel gene networks and pathways. Current QTL analysis either uses over-simplistic models (e.g. without epistasis) or ad-hoc methods that require human interaction or require excessive computing time. These limitations become severe when thousands of RNA, proteins and metabolites all require QTL analysis. It is thus essential for the success of genetical genomics to speed up and automate the QTL analysis.

Informatics Tool for Proteomic Biomarker Detection using large-scale nanoLC-FT Mass Spectrometry Data

Dr. E. Marchiori, elena@cs.vu.nl

Efficiently identifying and quantifying disease- or treatment-related changes in the abundance of proteins in easy accessible body fluids such as serum is an important area of research for biomarker discovery. Currently, cancer diagnosis and management are hampered by lack of discriminatory and easy

obtainable biomarkers. In order to improve disease management more sensitive and specific biomarkers need to be identified. In this light, the simultaneous detection and identification of multiple biomarkers ('molecular signatures') may be more accurate than single marker detection. Therefore great promise holds in combining global profiling methods, such as proteomics with powerful bioinformatics tools that allow for marker identification. The additional dimension of separation provided by coupling nanoliquid chromatography to high-performance Fourier transform mass spectrometry (nanoLC-FTMS) allows for profiling large numbers of peptides and proteins in complex biological samples at great resolution, sensitivity and dynamic range. NanoLC-FTMS applied to a sample yields a large matrix of (time, m/z intensity) measurements, each indicating that, at a particular time (at about one second interval), an ion with a particular mass-to-charge (m/z) ratio was detected with a particular intensity. Analysis and interpretability of these large scale datasets requires application of automated methods applied to the complete datasets, thus going beyond detection and comparison of peak areas or features. The aim of the present proposal is to introduce a novel fully automatic approach for the comparative analysis of large-scale nanoLC-FTMS data sets based on ensemble machine learning that will allow for classification of complex serum samples of cancer patients, speeding discovery of diagnostic, prognostic and drug-response biomarkers. The final product of this project will be a first prototype of a user-friendly tool incorporating advanced bioinformatic techniques for automated analysis of multiple large-scale LC-MS datasets. The combination of cutting edge technology and an integrated advanced bioinformatics tool to be developed during this project will no doubt boost biomarker discovery and will yield applications with high clinical relevance.

35

Mitochondria: from evolution to function

Prof. dr. M.A. Huynen, huynen@cmbi.ru.nl

Mitochondria are an essential and universal organelle of eukaryotic cells. Their transformation from free-living bacteria to highly integrated systems, and the evolutionary tinkering that went with it, have never been systematically addressed. Nor has the potential of such an analysis for the prediction of protein function been exploited. The current sequencing of large numbers of eukaryotic genomes, including ones with highly specialized mitochondria like hydrogenosomes and mitosomes, in conjunction with the determination of mitochondrial proteomes and other mitochondrial genomics data offers a unique opportunity to predict the functions of all mitochondrial proteins and their links to each other in complexes and pathways.

We will trace the evolution of the mitochondrial proteome from its alpha-proteobacterial ancestor to current day mitochondria in mammals, fungi and plants. We will combine this analysis with other genomics data to predict and subsequently experimentally verify the function of proteins involved in key cellular processes like the metabolic communication between mitochondria and the cytosol, mitochondrial translation & protein complex assembly, and mitochondrial dynamics & calcium regulation. The project is bioinformatics driven: by comparing sequenced genomes and mitochondrial proteomics data we will map the evolution of the mitochondrial proteome. We use the correlated evolutionary behaviour of genes, e.g. in terms of their presence in genomes, in combination with other genomics data to predict their functions and interactions. Predicted functions will be tested in the project and by our international collaborators, using techniques appropriate for the type of function predicted.

BioRange

BioRange is the bioinformatics research programme in the Netherlands carried out by the BioRange consortium. The programme has started in 2005 and runs until 2010. The BioRange consortium is a (formalized) collaboration of 11 universities, four academic research institutes together with six academic hospitals, three private research institutes (NIZO, Plant Research International, TNO), the computing centre SARA, one industrial party (Organon) and one research “top instituut” (WCFS), with Foundation NBIC as coordinating party (penvoerder). The Dutch government funds 50% of the BioRange programme through the Netherlands Genomics Initiative (NGI) under BSIK grant 03013. The partners in BioRange also fund 50% of the programme (matching). Further detailed information on BioRange, the BioRange consortium and the involved research groups as well as the Netherlands Bioinformatics Centre (NBIC) can be found at www.nbic.nl.

The projects in BioRange are categorized in the following sub programmes:

Subprogramme 1: Bioinformatics for micro array technology

Subprogramme 2: Bioinformatics for proteomics and metabolomics

Subprogramme 3: Integrative bioinformatics

Subprogramme 4: Virtual Laboratory for e-Science for bioinformatics

Subprogramme X: Theme-bridging research

37

Each sub programme knows a number of themes addressed within the projects.

BioRange SP 1: Bioinformatics for micro array technology

Experimental design, sample size determination and data analysis

Dr. L.A. Gilhuijs-Pederson, l.a.pederson@amc.uva.nl

This study aims to optimize the power of discovery in micro array studies, so that differentially expressed genes can be identified in a more cost effective, trustworthy, and rapid manner. Micro arrays are increasingly used in life sciences, both to investigate molecular processes and for molecular diagnostics. Their power is dependent upon such things as platform choice, experimental design, the type of biological material under study, the anticipated magnitude of differential expression, and methods for normalization and statistics. At several stages experimental noise, systematic bias, and variability can be introduced. To reduce all these experimental effects, experimental design choices and data analysis methods must be well considered and complementary. This project aims not only to suggest an optimal sample size for any given

experimental design but also to determine optimal choices for data processing. Although primarily aimed at expression micro arrays, these guidelines and methods can be applied to other types of arrays such as SNP, CGH or even protein arrays, as used by other collaborators. The first results included a new multivariate method for combining results from independent micro array studies that interrogate similar hypotheses, a better characterization of the SF-ANOVA analysis method, and also some comparisons of different gene expression micro array platforms.

Unifying framework for data-driven pathway discovery

Prof. dr. ir. M.J.T. Reinders, M.J.T.Reinders@EWI.TUDeft.NL

This project aims to develop generic tools that enable the discovery of (evolutionary conserved) pathways by exploiting the various -omics data types. Besides integrating these different data, the project strives towards capturing pathways beyond the level of pairwise comparisons in order to find more complex relationships such as linear paths of interacting proteins (modeling signal transduction pathways) or dense clusters of interactions (modeling protein complexes). The first results include a statistical prediction model that treats the measurement noise differently per feature, per feature value as well as per sample, a computational method to predict fluxes (routes) in a (genome-wide) metabolic network solely based on measured gene expression data, a new framework to perform the statistical analysis of common insertion sites in retroviral insertional mutagenesis screens, a computational approach to detect common co-occurring insertions (within a tumour) as well as gene-gene family interactions, and a high-throughput protein interaction map based on Bayesian statistics (in collaboration with UCSF). An article was published about the evolutionary origin of the Peroxisomal Proteome and its links with the Endoplasmic Reticulum.

38

Design and analysis of genetical genomics experiments

Prof. dr. R.C. Jansen, r.c.jansen@rug.nl

Our research will develop mathematical models, methods, algorithms and software for modern large-scale genetics experiments. The focus is on creating new approaches that are relevant for answering biological questions. For example, we want to reconstruct cellular signaling pathways from large amounts of quantitative data. In this case, we might obtain global measurements of gene activity in genetically diverse animals or plants, measuring the amount of messenger RNA for every single gene in the organism. Based on the slight differences in gene expression in each individual, we are then able to infer how genes influence each other by direct or indirect regulatory connections. Specific objectives are to contribute to the optimal design of metagenome micro array probes and to build databases that help to manage

the accumulated information. Also, the effects of alternative experimental designs and the influence of experimental parameters and errors will be investigated. Finally, we intend to generate comprehensive statistical methods and software for discovering significant trait/genotype correlations that indicate causal relations.

Understanding and reconstructing biological pathways

Dr. ir. P.D. Moerland, p.d.moerland@amc.uva.nl

DNA micro arrays are used to elucidate biological pathways. To visualize the expression data in a pathway context, the data can be linked to a pathway database (e.g., KEGG or GenMapp). These data can give even more insights in the biological processes if more advanced and sophisticated databases and tools would be available. The development of such pathway databases, as well as a more effective integration with analysis and visualization tools, is the focus of this project. Towards this goal, the project will develop a knowledge base for the (semi-)automatic interpretation of micro array data in the context of biological pathways. This knowledge base will use pathway databases and an ontology-based representation of the data, and will add domain knowledge that helps the biologist to fully understand his data. Integration of other types of data will aid further understanding, as well as reconstruction, of biological pathways. New methodologies may identify missing genes in pathways and may reconstruct larger parts of the pathways. The project uses graph-based representation of time-series data to progressively unravel regulatory pathways. Also, ranking methods and kernel-based methods will be used to integrate micro array data with other types of genomics data to unravel networks. The project has already realised a beta version of a Cytoscape-based application, as well as an alpha version of a new pathway analysis tool GMML-visio (could be thought of as GenMAPP3). This was presented to the GenMAPP group in San Francisco.

39

Quantifying the association between multidimensional phenotypes and multidimensional genetic/genomic data

Prof. dr. A.H. Zwinderman, a.h.zwinderman@amc.uva.nl

This project aims to develop and test bioinformatics tools to (quantitatively) study the association between phenotypic variables, and genetic, genomic, or proteomic measurements. Existing multivariate statistical techniques need to be generalized to handle data with different measurement scales, because a multivariate normal distribution can not always be assumed with these data. Also, modeling the joint distribution through a chain of conditional distributions will be attempted. The second research line addresses the overfit-problem: having much more variables than individuals causes many spurious relations

to be found. The third research line concerns nonparametric smoothing and data visualisation by exploiting the intercorrelations between neighboring genes. A generalized penalized quantile regression will enable a nonparametrical estimation of median trends, and will visualize association. The results of this project will be tested and demonstrated in collaboration with experimental groups. For example, phenotypic variation in expression of cardiovascular disease will be related to large sets of single nucleotide markers; variation in gene-expression will be related to variation in gene-copy numbers in colon cancer samples, to variation in potato-, and tomato-growth, and to changes in the fatty acid metabolism.

Already, correlations have been found between variations in response to treatment and genotypic SNP variations. Significant progress was made toward the overfit-problem. Preliminary results in a study, using integrated analysis of two data sources, identify known cancer-related genes as well as new candidate genes.

Developing clinical predictors based on high-dimensional genomics data, pathway information and directed experimentation

Dr. L.F.A. Wessels, L.F.A.Wessels@EWI.TUdelft.nl

40 Genomics is increasingly used for diagnosis and prognosis in human disease. Within a particular patient population, several clinically relevant outcome groups can be defined based on parameters such as response to therapy or survival time. Data-driven approaches can be employed to construct predictors of these clinical parameters. Finding a good predictor based on an optimal number of genes (or other features) is still a major challenge. The gene sets that are associated with currently employed predictors are typically derived in a purely data-driven fashion, performance is the sole determinant of the composition of the gene set. The gene sets derived in this fashion provide only partial information about the biological processes that underlie the observed outcome. Typically, indirect ('downstream') effects of the primary cause (e.g. a mutation in a key regulator) are the most dominant, and primarily determine the constitution of the gene set employed in the predictor. In addition, redundancy in observed downstream effects can cause a limited set of affected genes to be included in the final predictor gene set. In both cases, the predictive performance could be close to optimal but the gene set only partially reflects the underlying biology. In addition, it is also expected that this could lead to a degradation in the robustness of the derived predictor.

In this project we will develop statistical techniques which modifies the gene selection strategy such that it takes particular known relationships between genes, defining e.g. a particular pathway, into account. This will be augmented with approaches that 1) derive expression fingerprints (biologically interesting gene sets with a with particular expression pattern) from general data obtained from compendiums of cancer samples and specific data from model systems subjected to directed perturbations and 2) employ

these fingerprints to identify subtypes in breast, colon and melanoma tumor series which have a clinically relevant predictive value. Finally, the integration of similar gene expression data from different systems (patients and model systems) will be augmented with prognostic predictors derived from proteomics data originating from the same patient series. These approaches promise not only to improve the performance of the predictors, and allow better treatment choices, but also to greatly advance our understanding of the biology of cancer.

A modular storage and analysis pipeline for quality control, analysis and MAGE-NL export of micro array data

Prof. dr. F.C.P. Holstege, f.c.p.holstege@med.uu.nl

A rate-limiting step in applying micro arrays is data handling. This includes extraction of raw signal intensities, quality control, experiment annotation, up to date array annotation, storage of other relevant information, analysis at several levels and in multiple ways, visualization, publication and exchange of data. Standards for micro array data are under development but only a few groups have experience implementing these standards. Dissemination and implementation of micro array data standards at the national level is one important goal of this project. As part of this dissemination, the project foresees in developing a pipeline for all aspects of micro array handling, which will be comprehensive, flexible, efficient and include incorporation of micro array data standards for publication and exchange. The modular nature will allow separate parts to be implemented elsewhere as well as incorporation of contributions from other projects. The pipeline will serve a wide body of micro array users and will also be a training platform for micro array data analysis. Scripts for the integration of software that extracts micro array features have already been completed, and a user meeting was held at UMC Utrecht.

41

BioRange SP 2: Bioinformatics for proteomics and metabolomics

Software development for improved analysis and data-handling of high-throughput MS data

Dr. J. Krijgsveld, j.krijgsveld@chem.uu.nl

Although the NPC is developing bioinformatic approaches for proteomics in the Netherlands, several Dutch institutes are also setting up proteomics facilities. Two labs, Stunnenberg at the RU and Krijgsveld/Heck at the NPC, are clearly ahead of the pack. This project brings these two labs together and harmonizes their software efforts in the field of early data-handling and analysis. Both teams will closely collaborate with the

VLe project and investigate which VLe aspects can be used beneficially for their projects. Also dissemination will occur in collaboration, so that proteomics research groups that still have to start today, can have access to these results. The first results of this project include the set-up of GRID infrastructure, the storage of proteomic data in GRID infrastructure, while XML models (mzXML, pepXML, mzDATA and analysisXML) for proteomics data have been evaluated.

Bioinformatics for Metabolomics and Fluxomics

Dr. J.H.G.M. van Beek, hans.van.beek@falw.vu.nl

Now that several genomes have been sequenced, functional genomics is getting more attention. Cell function and pathology are directly reflected in metabolite levels and fluxes in metabolic pathways. Large scale measurements of many metabolites and their intracellular levels in parallel are captured under the term 'metabolomics' and large scale quantitation of metabolic fluxes is termed 'fluxomics'. This project will develop and test bioinformatical methods and tools for these new scientific fields. These tools will be tested and demonstrated in collaborative projects. Databases for storing raw and processed information, obtained by mass spectroscopy (MS) and nuclear magnetic resonance spectroscopy (NMRS), are needed.

42 High-throughput metabolome data must be processed, stored, and analyzed, and algorithms for compound identification are required. New tools for qualitative characterization of metabolites and pathway topology will also incorporate subroutines to analyze stable isotope enrichment. Dynamic isotope enrichment, found with MS and NMRS, can be interpreted when spectral peaks and isotope distributions can be identified and analyzed. Then, metabolic fluxes can be quantified in vivo, and dynamic responses of metabolic networks after treatments can be analyzed. Results in the first period were the design of a computational pipeline for automated compound identification, the implementation of a database of small molecules (MetBase), and the development of models for prediction of physicochemical properties. For flow of isotopes in metabolic pathways a FluxSimulator package was developed.

Proteomics based biomarkers for clinical research

Dr. ir. H.C.J. Hoefsloot, huubh@science.uva.nl

Proteomics and metabolomics research is partly focused on biomarker discovery. Finding biomarkers requires clean proteomics and metabolomics data. However: 1) very many compounds are measured, 2) each measured compound can be distributed over different instrumental channels, 3) different experimental designs are used, 4) the data can be time-resolved, 5) the noise levels can be considerable, 6) inherent biological variation, 7) usually few samples are available, 8) large fluctuations in dynamic range

may occur. This calls for bioinformatics/biostatistics methods that are robust, simple and reliable. Then, validation must ensure the quality of the biomarker selection based on these methods. The work in this project will be consequently divided in two parts. Part 1 will focus on new validation tools for existing methods to discover biomarkers, while part 2 will deal with making new methods for biomarker discovery with built-in validation facilities. A new statistical biomarker discovery tool, based on rank products, has been developed, and a paper on the effects of bootstrap aggregating of metabolomics data has been accepted.

Protein interactions at multiple scales

Prof. dr. J. Heringa, heringa@cs.vu.nl

Genotype defines phenotype by a network of molecular interactions: key to all cellular processes. This project develops and tests bioinformatics tools that analyse interactions based on a variety of (genomics) data. The scale at which the interactions take place, vary, as do the interacting substrates. In spite of these variations, underlying mechanisms are quite similar. The complementary expertise of the various groups provides an excellent basis for collaboration. This project will study protein complexes to illuminate the association/dissociation dynamics of protein complexes in solution, and even the ferocious problem of complex and cluster formation in different cellular environments. Pair-wise protein interactions will be studied by machine learning techniques. Various types of -omics data will be used, but also sequence and structure-information of the interacting partners. Protein-DNA interactions will be predicted for the transcription regulatory networks in bacteria by 'phylogenetic foot printing'. This will enable detection of regulatory elements in whole genomes. The Golgi apparatus functions as a central delivery system in the cell. Are there signals present in protein sequences that interact with this apparatus. Finally, two extremely complex signal-transduction pathways, associated with oncogenesis, will be analyzed. This will in part be driven by previous results, and will direct new wet-lab experimentation. This project has already finished an analysis of ancient, parallel duplication in pathway evolution, as well as the fate of duplicated genes for function prediction. The STRING database (prediction of protein function) has been updated, and potential new proteins were predicted. A new tool for finding functional determinants in protein multiple sequence alignment was developed. A fast and reliable method for protein sequence and structure comparison was devised, and plant glycan specific monoclonal antibodies were screened. Localization of several Golgi enzymes was studied, and transfection experiments with the labeled Golgi enzymes (DNA plasmids) were carried out.

Biomathematics in mass spectrometry based proteomics for identification and modeling of protein networks

Prof. dr. ir. J.H. van Schuppen, J.H.van.Schuppen@cwi.nl

Cellular functions are controlled by regulatory pathways and networks of transcription factors. These pathways and networks are mapped in studies on the regulation of the genome. In international collaborations, the regulation of genome function is addressed at a higher level: genetic regulatory nodes are mapped, as are networks that control the activity of embryonic stem cells, and the process of differentiation to specific cell types. Such studies rely heavily on informatics and mathematics. This project will focus on the mathematics to analyze and model the genetic circuitry that controls the formation of the blood system.

BioRange SP 3: Integrative bioinformatics

MCSIS heterogeneous data handling

Prof. dr. G. Vriend, vriend@cmbi.ru.nl

44

The project is executed in close collaboration with Organon. The MCSIS (Molecular Class Specific Information System) technology deals with heterogeneous data that are available for G protein-coupled receptors, as well as for nuclear hormone receptors. It extracts data from many large, normally monolithic, databases, and combines these in one information system. In this project the MCSIS software will be automated so that bioscientists can make their own MCSIS for the molecule of their own interest. Recently developed software products will be integrated, and we intend to GRID-enable the MCSIS technology. As a first task, this project has updated the database GPCRDB.

Structure validation for modeling purposes

Prof. dr. G. Vriend, vriend@cmbi.ru.nl

3D-modeling is crucial for the study of proteins, but very complicated. Homology-modeling is the preferred approach, but very imperfect. Many errors in the details of protein structure can only be detected when structure validation tools are used in combination with experimental data. In this project we will concentrate on the validation of structures solved either by X-ray, or NMR. The validation software will provide automatic feedback to protein structure refinement software so that re-refinement of (old) structures with modern software will be possible in cases where the experimental data is available.

This project has already included the supercomputer at SARA in the process to enhance hundreds of structures overnight, instead of just a dozen. Also, articles were published and our results were applied in life science research.

Exploiting structural genomics information to incorporate protein flexibility in drug design

Prof. dr. J. de Vlieg, jacob.devlieg@organon.com

Structural bioinformatics allows, in principle, the design of small molecules, such as potential pharmaceutical drugs, that bind specifically and tightly to a protein target. However, a protein may change its shape in such an interaction, and such changes are extremely difficult to predict. We need to know more about protein flexibility in order to understand how drugs exert their biological effects. In this project we want to predict the conformational changes, or alternatively predict the selection of protein conformations, that occur in a protein upon binding with a ligand, in particular at the binding site. The compilation of structures of proteins within the same family is thought to be instrumental in understanding protein flexibility. As a first result, a protocol to deal with flexible residues in combination with combination with docking was developed. This project also saw its first publication.

45

Prediction of gene function and regulation

Prof. dr. C.J.F. ter Braak, cajo.terbraak@wur.nl

To predict the function and regulation of genes, evidence from various data sources have to be pieced together. The challenge is to do this on a genome-wide scale and with significant sensitivity and specificity. The project distinguishes 3 work packages. WP 1 will build on a probabilistic framework for gene function prediction, based on *Saccharomyces cerevisiae*. This framework will be expanded, and extended to more complex plant genomes. WP 2 will study the role of alternative splicing in gene functionalization in plants, as well as evaluate its impact on computational methods of gene function prediction. WP 3 will focus on animal species. Genomes will be explored for functional elements and genetic variations, concentrating on non-protein coding genomic elements (e.g. promoter elements, but also microRNA-coding sequences). This project has already resulted in a computational pipeline for the identification of microRNAs from whole genome comparisons. This has led to the identification of many novel candidate microRNA genes. Two manuscripts in which this pipeline is implemented have been submitted.

Protein knowledge building through comparative genomics and data integration

Dr. P.M.A. Groenen, peter.groenen@organon.com

Currently, the function of only about 50-80% of proteins in each genome is known or predicted. This fraction can be increased by comparative genomics and integration. The project aims to construct a repository of (the most) accurate sequence similarity information from all fully sequenced genomes. These sequence similarities will form the basis of a phylogeny-based protein database. Also, enhanced methods of functional annotation based on sequence homology and non-homology methods will be developed. Finally, a data warehouse for enriched protein information, coupled with improved and robust visualization techniques, will be developed. This will allow sophisticated data mining and knowledge building in the areas of biomedicine and biotechnology. Important steps toward the integrated database and its maintainability have been realized. In collaboration with the Rolf Apweiler group at the EBI we currently use the Uniparc-based pairwise alignments as our base. Also, a new tool for detection of functional sites in protein sub families was released, and new ideas for the functionally relevant interactions in poliovirus replication were developed.

46

Phenotyping clustering of multi-factorial diseases

Prof. dr. A.K. Smilde, asmilde@science.uva.nl

Different people react differently to drugs: this simple fact is a major challenge for drug design. Biology is an integrated system of genetic, protein, metabolite, cellular, and pathway events that are in flux and interdependent. These biological elements largely define the phenotype. With advanced bioinformatics tools, interlinked data repositories with species-specific molecular genetic information can be mined, and new phenotype-related cluster strategies can be developed. This will generate model descriptions, based on systems biology, and enable personalized medication and nutrition. Metabolomics data have been measured in a cohort of twins. Also, an intensive effort is started to cluster relatives on basis of the measured metabolomics data. A literature study led to a definition of system structure and boundaries for cholesterol homeostasis and dyslipidemia in the mouse.

Development of generic integration methodology towards a life-sciences problem-solving environment by modeling of data and knowledge

Dr. T.M. Breit, breit@science.uva.nl

This project develops a problem-solving environment for computational life-sciences research. Such an environment allows the e-bioscientist to do computer-assisted research on complex biological problems. It will facilitate the correlation and review of (~omics) data in various ways, which will generate new ideas and hypotheses. The study of well-defined biological problems using this environment will enable the discovery new fundamental phenomena. A problem-solving environment needs a formal and non-ambiguous representation of data for robust and flexible data integration, as well as for the application of knowledge models. The latter includes how semantic models (ontologies, knowledge models) facilitates the generation of hypotheses and the development of 'knowledge-rich' analysis methods that incorporate biological knowledge next to data. This project has already accomplished the integration of 'R' in a Grid environment. A workflow-based micro array PSE for an e-Biolab was started, and key technologies and components as the basis for a semantic framework were identified. An improved TextBLAST was integrated in a tool (provisionally) called 'Anni', and the project embarked upon a Wiki approach to distributed annotation and updating of ontologies.

47

BioRange SP 4: Virtual Lab e-Science (VLe) for bioinformatics

PHASAR/BioMeta - mining metabolite data from literature

Prof. dr. C.H.A. Koster, kees@cs.ru.nl

This project will construct the PHASAR (Phrase-based Accurate Search And Retrieval) system for the automatic extraction of information from large amounts of literature. 'Metabolites' are selected as a test-case, since these are normally only mentioned in passing in articles dealing with other topics, and in such diverse sources that manual extraction is practically impossible. As a result, a detailed thesaurus of metabolite terminology, as well as a database of metabolites and their relations, will be also generated. The resulting system is generic in nature and, given suitable thesauri and ontologies, can be applied to other subject areas. PHASAR uses profiles, generated in interactive analysis of selected documents. A profile consists of linguistic phrase patterns, incorporating information of co-occurring words and ontological information about synonyms, as well as wider and narrower concepts from UMLS. Browsing of whole documents is replaced by sentence browsing and inspection of the index. In experimenting with a sequence of prototypes of the PHASAR system, its functionality and user interface have been optimized, efficient algorithms for server and clients were found. A common API for the PHASAR Server was designed. All

Medline abstracts were parsed and indexed using the LISA system of SARA, Amsterdam. The requirements for thesauri were fixed, and an evaluation plan set up.

User interfaces for scientific collaboration

Prof. dr. ir. A. Nijholt, anijholt@cs.utwente.nl

Collaboration is more than communication. The User Interfaces project complements communication with interfaces that help scientists work with maximum efficiency and effectiveness. The working practices of the intended user group will be analyzed in detail. Secondly, interfaces will be constructed that adapt to their users, and can provide advice during the interaction. The task analysis of the VL-e use case (task flow in e-biolab workflow) was nearly completed. Also, the Taverna workflow tool has been significantly enhanced. A paper on user analysis has been accepted at an international workshop, and project members have contributed to a textbook on human-centered visualization environments.

Advanced information processing in bioinformatics

Prof. dr. J.N. Kok, joost@liacs.nl

Over the last 15 years, data sets have increased dramatically in size and variety. Simultaneously, computational methods for extracting information from large quantities of data have been developed. In collaborations, such as BioRange projects, information from different data streams comes together. This poses a challenge with regard to the integration of heterogeneous data. This project focuses on advanced information processing: algorithms, databases, and tools, that, when integrated in the collaborative information facilities of VL-e, will provide a powerful bioinformatics environment. Semantic models within bioinformatics are needed, and advanced generic tools for data mining, and tools and methods for data interlinking and integration. Also, the design and development of a pattern-based framework to support the knowledge discovery efforts of bioinformatics researchers will be investigated. A framework for inductive querying has been designed, and comparative studies in text mining, and substructure mining on the large NCI database, were done.

BioRange SP X: Theme-bridging research

This sub-programme forms a bridge between the other subprogrammes of the BioRange research programme. The objective is to combine newly developed tools with '~omics' data in a systems biology approach for discovery and hypothesis generation in the area of life sciences research.

SPX: Integrative bioinformatics with data model enabled data analysis: test case industrial microorganisms

Prof. dr. R.J. Siezen, r.siezen@cmbi.ru.nl

Microorganisms are widely used as cell factories. To improve these factories, to generate new products or better performance, these factories must be studied as an integrated system. In this project, we will use '~omics' data for discovery and hypothesis generation in the area of life sciences research based on a microorganism test case. For this, a new strategy is needed to enable integration of heterogeneous models and data, as well as methods for the analysis and visualization of such heterogeneous data. The use of data and knowledge models for data annotation and integration is the basis for a powerful, robust, and scalable integrative bioinformatics methodology.

Notities

The Computational Life Sciences (CLS) research programme is a joint initiative of the boards of the NWO research councils Physical Sciences and Earth and Life Sciences, the Netherlands National Computing Facilities foundation, the Netherlands Organisation for Health Research and Development and the NWO governing board.

The 2007 round is financially supported by the Netherlands Genomics Initiative (NGI) and the Netherlands Bioinformatics Centre (NBIC).

Programme Office
Netherlands Organisation
for Scientific Research
ir. drs. Michiel de Boer
PO Box 93460
2509 AL The Hague, The Netherlands
Telephone: + 31 (0)70 3440 812
cls@nwo.nl
www.cls.nl



Netherlands Organisation for Scientific Research

01 01 01
01200001
00
011001 00
0010110
01001000
11001100
1001 1100
01
0100100
00110011
00101