INTER
FACE

DEVELOPMENTS IN BIOINFORMATICS

→ Coverstory
Reducing the complexity of biological systems

→ Interview
Lars Juhl Jensen
Content

EDITORIAL  Research as support strategy  3

COVER STORY  Reducing the complexity of biological systems  4

COURSE  In silico drug hunting  7

INTERVIEW  Lars Juhl Jensen  Data mining specialist  Novo Nordisk, Denmark  8

BUSINESS  KeyGene: jumpstart for phenotyping with bioinformatics  11

PROGRESS  Designing metabolic models and estimating their fluxes  by Thomas Binsl  12

PROGRESS  Unravelling the genetic basis of complex diseases  by Sandra Waaijenborg  14

HANDS ON  Open source software package for subtype discovery  16

THESIS  Finding relationships in biological data  18

PORTRAIT  Six questions for Christin Christin  20

IN THE PICTURE  NBIC news  21

COLUMN  Han Brunner: Tools that are never used  24

COVER PHOTO: finding the meaningful building blocks in the immense pile of complex stones that construct life.

→ I’m always on the lookout for exciting data

→ Identifying distinct subgroups within a large patient group suffering from the same disease.
Research as support strategy

Just before the summer break, NBIC granted funding to 22 internationally reviewed projects of the BioRange-II research programme. BioRange-II comprises the sub-programmes: Bioinformatics for sequencing; Genotype-Phenotype Modelling; and Systems Bioinformatics. A fourth sub-programme – Bioinformatics for Proteomics – was implemented at an earlier stage in conjunction with the Netherlands Proteomics Centre.

The chosen scope of BioRange-II safeguards good alignment with the research programmes of several NGI genomics centres and ensures that BioRange-II projects directly support their genomics research. BioRange-II also includes a few fundamental research projects, the results of which may not find application today but which are expected to do so in the future. We are confident that BioRange-II will result in new insights and methodology to advance Dutch life sciences as a whole.

Researchers want to research and BioRange-II provides them an opportunity to do so. Yet, we all realise that NBIC’s most important task is to provide support to the Dutch life sciences community. Moreover, one could argue that the success of NBIC depends on its success in supporting life sciences. But what is support? Support includes consultancy, providing tools and databases, human capacity to perform bioinformatics services (e.g. statistical analyses), research collaboration (BioRange), e-bioscience approaches to the implementation of support platforms (BioAssist), education and training to increase expertise and capacity (BioWise), and a mixture of the above. Each of these support strategies has its advantages and disadvantages with respect to availability, scalability, throughput, responsiveness, pay-off for involved parties, motivation and enthusiasm, user satisfaction or costs, e.g. for providing a service or establishing a specific support strategy. It would be interesting to analyse different strategies, but it is clear that projects in which bioinformaticians and biologist jointly address a question are the most successful and satisfactory.

This view of support implies that bioinformatics research is part of support. Indeed, support occasionally requires research to develop new methodologies. So a research programme is essential for NBIC to complete its most important mission. But it must be defined and implemented with ‘support’ in mind. BioRange, therefore, complements research programmes such as those implemented by NWO and others. Finally, BioRange needs careful alignment with the BioAssist support and BioWise educational programmes.

To strengthen the link to BioAssist, BioRange also promotes the re-use, sharing and joint development of software by requiring that said software become available as open-source with a proper OS license. NBIC will support open-source development through facilitating an open-source repository and development environment (gforce), as well as by providing advice with respect to the choice of a proper license. Similarly, NBIC advocates open access to data.

Research, support and education have their own objectives, but all contribute to enhancing and advancing life sciences, as has been nicely illustrated in this issue of Interface. I hope you enjoy reading the lively contributions, among others an inspiring interview with data mining specialist Lars Juhl Jensen, the clear presentation of the R-SDisc package, a tool for subtype discovery, and the enthusiastic report on the computational drug discovery course.
Reducing the complexity of biological systems

BY ESTHER THOLE

As research takes more and more of a systems approach, computation is becoming the key bottleneck. Collaborations between bioinformaticians and biologists face great challenges.

Systems biology focuses on the complex interactions in biological systems. New emergent properties that may arise from this systemic view are used in order to better understand the entirety of processes that take place in a biological system. Bioinformatics tools are indispensable for interpreting the wealth of information from seemingly disparate datasets and for integrating it in the system. Therefore, NBIC operates a dedicated bioassist platform on systems bioinformatics.
More and more, biologists are expanding their vision to tackle the functioning of complete systems. In what ways can bioinformatics contribute to systems biology? Amongst other ways, tackling ‘hairy monsters’ appears to be on the wish list.

When asked for his definition of systems biology, Bas Teusink, professor of Systems Bioinformatics at VU University Amsterdam, became slightly irritated. “I really thought we had finished that discussion by now. It is about understanding the behaviour of a system based on the behaviour of the individual components and their mutual interactions. It’s ecology of molecules.” Perry Moerland (Academic Medical Centre, Amsterdam) and Lodewyk Wessels (Netherlands Cancer Institute, Amsterdam), programme leaders of NBIC’s BioRange sub-programme Systems Bioinformatics, reacted with mild scepticism. “Oh, does he have a clear definition? We would love to hear it.” But luckily for the overall atmosphere, Teusink’s interpretation is accepted. Moerland: “I see his point; it is not a bad description.”

TOO COMPLEX Generally, systems biology is defined by two approaches. In the top-down approach, you start with the data to build models that work towards understanding the system. This approach is also what many view as ‘bioinformatics’. In the bottom-up approach, you start with detailed descriptions of individual components or a very small system and gradually add new layers of complexity to build a larger system. Bas Teusink works in both directions, but primarily bottom-up, on metabolism of microorganisms. “We want to understand how metabolism is regulated, how choices between metabolic pathways are made and, more interestingly, why these choices are made. What are the benefits of a certain pathway and under which circumstances do certain choices emerge?”

To understand such complex networks of interactions, he also performs what he calls straightforward systems biology: building models for metabolic pathways. “It will prove to be too complex to really fit a complete living cell into a model. But by detailed modelling of the sub-systems, we can reduce those components to the essential inputs and outputs. Subsequently coupling and modelling that information on a higher level allows us to understand the functioning of a complete cell.”

Ritsert Jansen, professor of Bioinformatics at the University of Groningen, takes the top-down approach. “With systems genetics, we add yet another component to the complexity of systems. Even if we succeed in modelling a complete individual, what does it say about the 6 billion individuals worldwide? My personal motivation in this field is to understand the phenomenal variation in phenotypical traits.” This type of work requires genetic information from large populations as well as phenotypical descriptions and molecular data. “The key is to trace the genetic networks that are linked to the phenotypical trait you’re studying. We generate hypotheses on potential networks, sugges-

tions for new pieces of the big puzzle of life. Basically, we generate input for systems biologists to work with.”

STIMULATING CALCULATIONS “In the BioRange programme Systems Bioinformatics, we try to bring these two approaches together”, says Lodewyk Wessels. As to the main objective of the programme, he says: “It’s a cliché, but the leading question is still how to get meaningful information out of the huge datasets that are being generated in all kinds of fields.” Besides the BioRange research programme, NBIC also operates a BioAssist platform on Systems Bioinformatics. “How do you describe a model in such a way that a computer can process it and that it can be used by researchers who all work with different programmes and different systems? That is the foremost question in this platform”, explains platform leader Hans van Beek (VU University Medical Centre, Amsterdam). “These activities represent the ‘supporting’ role of bioinformatics.” Van Beek admits he prefers the other side of bioinformatics: calculating. “Using quantitative measurements from a network to really calculate with your model to come to new estimates of missing parameters. Once that starts going and the model works, that to me is the most stimulating part of bioinformatics.”

INCLUDE CONTEXT How can bioinformatics contribute to progress in systems biology? What is the field expected to deliver? Bas Teusink mentions the need to reduce complexity by moving towards an ‘engineering’ type of approach using modules. “The way we currently build networks results above all in the creation of so-called hairy monsters and nobody knows how to deal with them. We should therefore focus on trying to reduce parts of the monster to a definition in terms of essential inputs and outputs. What exactly is the function of such a part? That requires identification and validation of modules using different datasets based on more than just statistical correlations. Other biological information, for example binding sites of transcription factors, needs to be included as well. I really see an important contribution from bioinformatics here.”

Bioinformatics can also make a difference when it comes to the semantic web, according to Teusink. “Data should convey more than just the presence or absence of genes. The conditions under which measurements were performed are essential to the outcome and should be taken into account.” Perry Moerland also foresees a role for bioinformatics in this respect. “Take for example the KEGG pathway database. The pathways are roughly defined without information on the context, although we know that in different cell types under different conditions the same pathway can greatly differ. Information on the context is crucial to the usability of the data.”

Finally, the role bioinformatics can play in experimental design is brought up several times. Supporting biologists in effectively exploiting the experimental possibilities of bioinformatics is very important to prevent overkill,
According to Lodewyk Wessels, “sometimes you do whole genome studies and biologists will just focus on their favourite gene and elaborate on those results; that is not an optimal approach.”

“The key is to devise clever measurements”, says Ritsert Jansen. “In the conventional molecular biology approach, you tweak a single factor and measure what happens to the overall system. That is like trying to locate Mont Blanc by walking from Eindhoven to Groningen to find the highest point and then continue looking in the orthogonal direction. But by including genetics, you can give the system a multifactorial ‘kick’. This way, you can ask a lot of different questions in just one single experiment.”

**ultimately biology** Real collaboration with biologists is crucial to any project, as both Moerland and Wessels emphasize. Biologists may need bioinformaticians to make sense of the data streams, but the dependency goes the other way as well. Wessels: “Testing and validating your modelling results takes a lot of time and you need biologists to do the experiments. But do they have the time and the interest? Again, it comes down to collaborating.”

As collaborating with biologists has become the standard, the question of what bioinformaticians and systems biologists see as their ultimate goal arises. Is it answering the biological question or developing the method? “For me, the ultimate goal is to contribute to better cancer therapies, not to develop a generally applicable method”, Wessels explains. Moerland feels that this goes for everybody in systems biology. “Answering the biological question is the main objective. That is what everybody will relate to when you develop a method. What is its use in biological research?” Wessels agrees: “No matter how refined your algorithm is people will always ask about the biological relevance. How can we use it to increase our biological knowledge?”

The focus on the biology side of the questions harbours the risk of seeing biological explanations everywhere. Discerning between ‘real’ biological results and mere statistical correlations is essential and an important task of bioinformaticians. “Look for causal connections”, says Hans van Beek. “Even though I strongly believe that combining existing data can lead to new insights, just blindly measuring and comparing will get you nowhere. In this field, you need targeted questions, and every now and then a wild hypothesis to create a breakthrough.”

**work done?** Ritsert Jansen emphasizes the need for a critical attitude towards the technology. Make sure that you really see what you think you see. “It is very important that we are comfortable with the results. They should really say something about the biology and not be based on artefacts or flaws in the analysis.” That does happen. His group recently published a paper in PLoS genetics in which they re-analysed the findings of another group. “We thought their results were too good to be true. and we were right; it turned out there was a mistake in their analysis protocol.” It must be frustrating to have your results knocked down like that. “Oh well, let’s not forget that in the end we’re all human”, says Jansen. But still, it must be hard to look back at your earlier findings every now and then using better methods and to discover that you need to adjust your results.

At what point can you consider the work done? According to all five interviewed here, the answer is simple: never. The nature of science is that there is always something new to explore and you are always dealing with uncertainties. Except in one respect, according to Bas Teusink. “In modelling, you know one thing for sure. and that is that after version one, there will be a version two and three and so on.”

**acknowledgment**

We thank the interviewees for their collaboration in the discussion about the role of bioinformatics in systems biology.

- Hans van Beek, platform leader of BioAssist support programme Systems Bioinformatics, (VU Medical Centre, Amsterdam)
- Ritsert Jansen, professor of Bioinformatics (University of Groningen)
- Perry Moerland, programme leader BioRange sub-programme Systems Bioinformatics, (Academic Medical Centre, Amsterdam)
- Bas Teusink: professor of Systems Bioinformatics (VU Amsterdam)
- Lodewijk Wessels, programme leader BioRange sub-programme Systems Bioinformatics (Netherlands Cancer Institute, Amsterdam)

The focus on the biology side of the questions harbours the risk of seeing biological explanations everywhere. Discerning between ‘real’ biological results and mere statistical correlations is essential and an important task of bioinformaticians. “Look for causal connections”,
In silico drug hunting

These days many computer programs help in discovering and designing the right drug. However finding one’s way within this ‘in silico drug hunting’ world is not easy. First, knowledge about different disciplines is required to translate the large amounts of structural and sequence information. Besides that you need to know which tools are available and when and how to use them. The International Computational Drug Discovery Course shows the ropes.

“It was quite an intensive course; I learned a great deal from it”, says Markus Kossner, PhD student from the University of Braunschweig, Germany. This summer Kossner followed the 2-week computational drug discovery course in Nijmegen. Kossner works on anti-viral and antibiotic agents in the department of Pharmaceutical Chemistry. He decided to follow the course because he wanted to get a good overview of the complete drug development process, from target discovery by microarray analysis to, finally, the clinical stage. “My background is mainly pharmaceutical sciences and in the beginning I hardly knew anything about programming and useful programs. To fill this gap, I followed some workshops and I took a summer-school course on chemoinformatics in Strasbourg, but I felt I still lacked some knowledge.”

Kossner especially acknowledged the lectures of experienced drug discovery researchers from industry. “Most courses I visited were given by academics, but people from industry have a different view and experience in drug discovery. As an academic researcher, one usually does not take part in all the different steps from target discovery to the clinical stage. Especially the later stages are expensive and therefore are chiefly done in collaboration with or exclusively by industry.” The different views from the lecturers, with biological, pharmaceutical or chemical backgrounds, also taught him a lot.

In the morning there were lectures and after lunch the practical part of the course started. He explains: “We worked on concise problems; for example we had to design leads for nuclear receptors, using diverse computer programs available for the different stages of drug discovery. The lecturers walked around to answer questions. That way you got to know the programs and you could find out which ones were useful for your own research.”

The programs originated from different sources and developers. Some of the programs were developed by the group of Professors Jacob de Vlieg (Schering-Plough), one of the lecturers. Other programs are free of charge and available on the internet. However the majority have to be purchased from companies. Kossner was quite captivated by the Spotfire program: “It is a very easy and useful statistical program. All kinds of fancy statistical analyses are possible and you don’t need to use statistical programming to get quick and useful results.”

On the last day of the course, the different groups had to give a presentation on their drug discovery problem. “This was very useful because we discussed how the different problems were tackled. The lecture given by Jose Duca (Schering-Plough) on unexpected molecular behaviour in particular conveyed a lot of hands-on experience.”

The International Computational Drug Discovery Course covers recent advances in drug discovery informatics, with a focus on the application to real life problems. Scientists from both industry (Schering-Plough) and academia (Radboud University Nijmegen Medical Centre) introduce and discuss the scientific concepts and tools that are part of the modern genomics-based drug discovery pipeline from target discovery and validation to lead discovery and optimisation. The practical component of the course will provide participants the opportunity to work with state-of-the-art in silico tools and databases. The course is targeted at advanced master students, PhD students and postdocs who are interested in the field of computational target and drug discovery.

For more information: http://www2.cmbi.ru.nl/groups/computational-drug-discovery/icdd2009/
“I’m always on the lookout for exciting data”
Professor Lars Juhl Jensen is 34 years young. He is a scientist and entrepreneur; he blogs and tweets, and tries to convince every biologist that bioinformatics is not just a service, but a true science. This year he moved his office from the EMBL in Heidelberg, Germany, back to what may be called his hometown: Copenhagen. Here he is building up his own research group, which hopes to unravel all the secrets of protein phosphorylation regulation in the living cell. "A perfect opportunity", according to the data mining expert.

When did you become interested in bioinformatics? At the Technical University of Denmark, where I studied chemical engineering, there was no strict five year plan. During my studies, I gradually got interested in biology and I picked courses like organic chemistry, biochemistry and molecular biology, slowly drifting in the direction of biology. At the time I needed to choose a bachelor project, I saw an advert from the group of Søren Brunak for a project on bioinformatics. That word intrigued me immediately as it seemed to combine two of my favourite topics: biology and maths, and since then I have known that I wanted to be a bioinformatician.

Why do you like maths? I guess I just have a logical mind set. As a kid I was already programming, just for fun. I tried, for example, to make an old fashioned, round clock on the computer, which is rather challenging if you haven’t learned trigonometry yet. But my father, who is a programmer, taught me about sine and cosine, and I managed get the clock working. I have no formal training in computer science, but I have been programming since forever.

“We need more data”

And what fascinates you in biology? It is probably the complexity of life. For decades people have been trying to predict how a protein will fold. Yet, it is still a very hard job. The same holds true for phosphorylation, a topic that I work on myself. Our predictions of which kinase will phosphorylate which protein sites are still often wrong; it is clear that there is still a lot to learn!

Are we making progress? Yes, a lot. But we need more data. I would never have worked on phosphorylation networks if it hadn’t been for huge advances in mass spectrometry that provides the data to do so. I’m always on the lookout for new exciting data.

You are starting up your own research group at the University of Copenhagen within the new Novo Nordisk Foundation Centre for Protein Research. Back to your roots?

It’s such an amazingly perfect match. After being a staff scientist at EMBL for almost three years, I was just starting to look for a position as a group leader somewhere in Europe. Then the Novo Nordisk Foundation made this unique, large donation for a research centre in my hometown that conducts the kind of science I do, and I was approached to come and work there.

You made career quite rapidly. What makes you a good scientist? I’m pretty good at spotting the core of a complex problem rather than becoming confused by the size of it. That’s probably my main strength. I can see where the problem really lies, which is a very first step in solving it. It’s not so much a question of intuition, but of being able to keep the overview. But it’s not a quality that I cultivate or think about a lot, it’s just something that I’ve always been good at.

“Online networking is simply a worthwhile investment of my time”

Which of your publications are you most proud of? In 2005, we published a paper in Science on the dynamics of protein complexes during the yeast cell cycle. By mapping temporal data on protein interactions with gene expression, we discovered that most protein complexes consist of periodically and constitutively expressed subunits. We proposed the idea of ‘just-in-time-assembly’, which implies that complexes are regulated by temporal expression of some of their subunits. By the time the Science paper was published, we were working hard on a follow-up, as we figured out a way to test our own hypothesis. If our just-in-time-assembly idea was correct, the identity of the regulated subunits should have changed during evolution, but not the principle that the regulated subunits in each complex are expressed shortly before the complex is needed. We studied four different organisms and found exactly what we predicted and that publication made it into Nature. A wonderful pure bioinformatics project; no new data or wet experiments were necessary to make these discoveries.

Why does a bioinformatician like you tweet? The most important scientific news I hear comes through electronic sources like RSS feeds and shared items on Google Reader, Twitter, FriendFeed or Facebook. Many people in the bioinformatics community automatically channel information between these resources. If I share something in Google Reader, it will be posted to all of them. But actually, it’s not about technology, it’s about people. There is a really good bioinformatics community using these Web 2.0 tools, and also very progressive people from electronic libraries and cloud computing too. Online networking is simply a worthwhile investment of my time.
Can you give an example of what you learned through online networking?

Last spring, for example, Sean Eddy from the Howard Hughes Medical Institute released a new version of the HMMER package. HMMER is a very heavily used program for database searches. It can model families of sequences and search for new members of those families. HMMER is very powerful, but the problem was that it was a lot slower than BLAST. The new version, however, is faster. I don’t care about a ten percent speed up, but a factor of a hundred makes for a whole new ball game. That news was first published on Sean Eddy’s blog.

“It’s good to see that the young biologists find it worthwhile to learn programming”

You are a blogger yourself. Why?

Actually, the internet is a good place to publish. When I put up a blog post, I typically get a couple of hundreds hits. Of course, I don’t know how many people read more than just the title, yet it isn’t a bad score. Blogging is also a nice outlet for the information that you don’t want to publish in a peer-reviewed paper. For example, you don’t want to spend peer-reviewers’ time on studies that don’t have an interesting outcome. At the other hand, if you put the information on the internet, others can find it with Google and may avoid wasting time on doing the same experiment.

You have more than thirty years ahead of you before retirement. What would you like to accomplish in that time?

I hope I can contribute to a change in the biologists’ attitude towards bioinformatics. At the EMBL, I once worked on a project with a young wet lab biologist doing microarrays. After a while, he got tired of calling me every time he needed to analyse some data and decided to learn Python himself so he could at least manage the simple things on his own. It’s good to see that the younger generation of wet lab biologists think it is worthwhile to learn some programming themselves. But many others still don’t recognize that you can actually make scientific discoveries just by reanalysing available data. Furthermore, I hope to prove the usefulness of bioinformatics in a commercial setting as well. For this reason, I have recently cofounded the company INTOIMICS, which offers to perform advanced data mining services as contracting research. Industry often has problems getting access to the latest data mining technologies because of confidentiality. Universities don’t want to sign research contracts that restrict their ability to publish or patent the results or that make them financially liable if confidential data are leaked. A company like INTOIMICS can offer industry the services without these problems.

Which data is on your wish list?

Currently, there are good methods for monitoring all the proteins in a cell globally, but with little resolution. To do a global mass spectrometry study you need millions of cells. There are also techniques, like live cell microscopy, where a couple of different, labelled proteins can be followed in real time. You can see how these proteins move inside an individual cell and in what amounts. It would be fantastic if these two ends would meet; if we had data of thousands of proteins within a cell in real time. I don’t know how we will get there, but I’m optimistic that it will happen.

Is there anything you want to say to the readers of Interface?

As a researcher in data mining, I always take the opportunity to call upon biologists to share their data instead of locking them up on a hard disk after they are analysed. With the kind of technologies we have today, someone might be able to combine the data with other sources and discover something brand new. So please, upload your data on the web or, if applicable, put them in the right repositories!
Plant biotech company KeyGene supports the vegetable and field crop seed companies with innovative technology to improve their breeding strategy. With DNA sequencing technology, KeyGene maps important traits and supports the development of new commercial varieties. The new added focus is phenotyping.

“We are not breeding the crops ourselves. We develop tools to gain knowledge of DNA sequences, biomarkers, SNPs or specific properties”, explains Harold Verstegen, Vice President BioInformatics at KeyGene. With this information breeders can develop strategies for improving their plants. In all these activities KeyGene relies heavily on bioinformatics and information technology.

Verstegen continues: “We not only provide tools to generate content but tools to explore and make predictions about the breeding strategy as well. For example we developed a gene-browser: a dedicated tool for breeders that helps them scout the genome data of their crop. In this way we support the breeding companies in their decision making and development of cross-fertilizations schemes. That is added value. A breeder remains a breeder; his main business is to improve and sell seeds.”

PHENOTYPING The next big thing in plant breeding is tool development for phenotyping. Plant phenotyping is a rapidly evolving concept that links genomics with ecophysiology and agronomy. The basis of this concept is that the functional plant body (phenotype) originates during plant growth. Phenotyping is now solely a skill of breeders, but it is open to interhuman variation in judgment of the plant’s qualities. For instance colour observation depends on the colour of the surroundings. A fully automated measuring system for high throughput phenotyping should do the trick. Robots can do it cheaper, faster and, especially, more robustly and objectively. “We believe that the best approach is to move the plants around on a conveyor belt through the imaging installation, where their snapshots are taken in standard surroundings. It is an elegant and non-destructive method. The images are stored and interpreted with dedicated software. With all the experience we have gained by analyzing genome sequences and bioinformatics, we can make a jump start in connecting the phenotyping of plant properties with genetic information. This asset gives our company a unique market position.”

ONGOING SHIFT Bioinformatics is gaining importance. About ten years ago KeyGene started their bioinformatics group with just a few people. Now Verstegen’s group consists of 30 people with backgrounds ranging from biology to physics and informatics. In the next five years Verstegen expects that most new employees will be bioinformaticians. The reason is the ongoing shift from traditional lab work towards more in silico desk work. By that time about 70% of the research work will be done by bioinformaticians. The high demand for data storage capacities and advanced bioinformatic software tools requires a good IT-infrastructure that follows the speed of innovation. “We have invested enormously in data storage and data management systems”, says Verstegen. “That is also one of the reasons why KeyGene has participated as a clusternode in BIG Grid of NBIC, which enables public and private collaboration initiatives. With these initiatives NBIC has made the rather fragmented field on research and education of bioinformatics more coherent and transparent. The organization operates like glue that binds the Dutch Informatics field all together. The same is needed for the business side.”
Knowledge about metabolic fluxes is of great value in drug development and bioreactor improvement. However, quantitative flux determination is difficult in animal tissues. LIPSSS and FluxEs sort out these problems and may be uniquely capable of quantitating metabolic fluxes in biopsies at much higher resolution than possible with noninvasive imaging.

The functioning of cells is reflected in metabolite levels and metabolic fluxes. Dynamic adaptation to changing environmental conditions is accomplished by changes in both of them. Hence, knowledge of metabolic fluxes is desirable for designing new efficient bioreactors and for drug development. The determination of metabolic fluxes in vivo is difficult, especially at high spatial resolution. This problem is usually tackled by measuring time series of metabolite levels or tracer experiments using stable isotopes. The isotope incorporation in key metabolites is measured by Nuclear magnetic resonance Spectroscopy (NMRS) or Mass Spectroscopy (MS) and subsequently analysed by appropriate software.

**ISOTOPE LABELLING** The determination of relative metabolic fluxes with stable isotopes is possible from data observed in isotopic equilibrium. Absolute metabolic fluxes can be quantified from time series data on metabolite levels or isotope incorporation. However, the turnover rates of many components in animal tissues are low and metabolic steady states are brief. Therefore, it is often not possible to label these biological systems with stable isotopes until an isotopic steady state is reached. Furthermore, time series data is difficult to obtain at high spatial and metabolic resolution and cannot be obtained from single tissue biopsies. To overcome these problems, we developed a novel approach we call Labelling with Isotope for a Pre-Steady-State Snapshot (LIPSSS) [1]. LIPSSS enables researchers to quantify absolute metabolic fluxes from single time point measurements in isotopic non-steady state. Like common tracer experiments, the LIPSSS approach starts with the infusion of substrates enriched with stable isotopes (e.g. $^{13}$C or $^{18}$O). However, in our case the metabolism in target samples is arrested at a predefined time point before reaching an isotopic equilibrium. Subsequently, the isotope incorporation in key metabolites up to this time point is measured using NMRS or MS. Using this data the metabolic fluxes are then quantified by computer analysis. Computational analysis is crucial to successfully perform the metabolic flux measurements. To this end we developed a computational framework called FluxEs in which metabolic pathway models can be flexibly assembled, simulated, analysed and metabolic fluxes estimated. FluxEs is based on our previously developed program FluxSimulator [2] and is run in the R computational environment. Hence, it was easy to port FluxEs from a desktop to the Dutch Life Science Grid where all computations were performed. FluxEs can be used without computer programming expertise since it offers an easy way to specify metabolic models in a plain text file. The model specification is automatically translated into the mathematical representation consisting of ordinary differential equations. Finally, flux parameters are determined by a non-linear minimisation of the difference between the simulated and measured stable isotope distributions.

**VALIDATION** The LIPSSS approach was experimentally validated in vivo on the cardiac energy metabolism of anaesthetised pigs under two different conditions: basal state and cardiac stress induced by dobutamine. The experiments were performed by David J.C. Alders under the supervision of A.B. Johan Groeneveld and Hans van Beek at the VU University Medical Centre Amsterdam. The LIPSSS protocol started with the infusion of $^{13}$C enriched substrates for the tricarboxylic acid (TCA) cycle. The TCA cycle, which is also called the citric acid or the Krebs cycle, is a critical part of the cell metabolism generating energy coupled to oxygen consumption. After 5.5 minutes heart tissue biopsies were taken and isotope incorporation into TCA cycle metabolites measured. Additionally, the oxygen consumption of the whole heart was calculated from ‘gold standard’ blood-gas based measurements. Using LIPSSS we were able to show the differences in metabolic conditions (basal state and cardiac stress) reflected in the metabolic fluxes and oxygen consumption. We found that...
the TCA cycle runs with an increased speed of about 50% under cardiac stress induced by dobutamine compared to basal state. The inflow of carbon from acetate into the TCA cycle also increases. Exchange fluxes between TCA cycle intermediates and amino acids are estimated to be lower during dobutamine stimulation. Due to the appropriate quantification of the TCA cycle flux and related parameters, it was possible to calculate the oxygen consumption for the entire heart. Despite the fact that the ‘gold standard’ physiological measurements were done for a much larger area than covered by the tissue samples of the LIPSSS experiments, both oxygen consumptions agreed with an overall correlation coefficient of 0.88.

FLEXIBLE TOOL The LIPSSS approach offers scientists a convenient way to quantify metabolic fluxes in situations where time series measurements of metabolite levels are not possible or the lack of an isotopic steady state makes it impossible to use equilibrium methods. Other advantages of LIPSSS are: (1) the possibility to measure in a single tissue sample at a single time point; (2) a more sensitive indication of changes in the biochemical pathway than provided by measurement of metabolite levels; and (3) the opportunity to take very local measurements using measurement techniques with high chemical resolution. The latter is not possible with in vivo imaging methods. LIPSSS experiments can comfortably be analysed by our computation framework FluxEs. Due to its simple input format, users are able to flexibly design metabolic models and estimate their fluxes even without computer programming skills.

Possible future application areas for the methodology we developed are: (1) the study of drug intervention effects to improve cardiac metabolism under ischemic conditions in vivo in small tissue regions; or (2) measuring the turnover and synthesis of neurotransmitters in the central nervous system.

REFERENCES

CONTACT
Thomas W. Binsl, MSc
Centre for Integrative Bioinformatics VU
VU University Amsterdam
De Boelelaan 1081A
1081 HV Amsterdam
Web: http://www.few.vu.nl/~tbinsl
E-mail: tbinsl@few.vu.nl

SMALL TOY MODEL OF A METABOLIC PATHWAY
The model consists of four metabolites (A, B, C and D) represented by their carbon atoms (c1, c2 and c3). Carbon transitions are given by arrows. The 13C isotope incorporation via flux v into the pathway is highlighted blue. NMR measurement reflects the amount of label incorporation into key metabolites. Each set of peaks corresponds to certain combinations of 13C (upper panel). Using the bioinformatics algorithms in FluxEs, the metabolic fluxes (here v) can be determined (lower panel).

KEY CONCEPTS
- Quantification of metabolic fluxes with time series data on metabolite levels is difficult in animal tissue.
- Labelling with isotope for a Pre-Steady-State Snapshot (LIPSSS) enables one to quantify absolute metabolic fluxes from a single time point measurement in a small biopsy.
- LIPSSS is an isotope tracer experiment with metabolism arrested in target samples at a predefined time point before reaching an isotopic equilibrium. The isotope incorporation up to this time point is measured in key metabolites by NMRS or MS.
- FluxEs is a computational framework to quantitate metabolic fluxes from LIPSSS data by non-linear minimisation of the difference between computed and measured stable isotope distributions.
- FluxEs provides a way to specify metabolic models in a plain text file and works on all operating systems on which the R computational environment is implemented.
- LIPSSS and FluxEs were successfully validated on the energy metabolism of porcine heart biopsies under two metabolic conditions: basal state and cardiac stress.
- The research was done in collaboration with the Department of Intensive Care, Institute for Cardiovascular Research (VU Medical Center). Its results are currently submitted to a bioinformatics journal.
- The program is available on request and will soon be added to NBICs GForge repository.
Unravelling the genetic basis of complex diseases

Complex diseases like cardiovascular disorders and cancers have a major impact on the Western world. They are the most important causes of death in modern society. By increasing our understanding of such diseases and unravelling their genetic basis, we can add a major contribution to the development of early detection methods, therapies and cures.

Biomedical techniques have advanced greatly over the past decades. And with an ever-growing amount of information on genetic risk factors obtained by microarray technology (proteomics, gene expression, single nucleotide polymorphisms, copy number variations, etc.), the identification of the genetic risk factors of complex diseases seems closer than ever. However, good statistical methods that can deal with the complexity of such high-throughput data are rare and thus many challenges remain.

Most existing studies have divided their study population into controls and cases, not realising that such classification is likely to cause heterogeneity within the two groups. This heterogeneity is an intrinsic result of the complexity of gene regulation, as well as many extracellular and intracellular factors. The same disease can be caused by (a combination of) different pathogenetic pathways. And due to this heterogeneity, the genetic markers responsible for, or involved in, the onset and progression of a complex disease are difficult to identify.

In order to overcome these problems, rather than dividing the study population into cases and controls, it is preferable to identify the phenotype of a complex disease by a set of intermediate risk factors. As a result, the question shifts from ‘What is the genetic difference between healthy and sick people?’ to ‘Which genetic factors are related to phenotypic changes that are involved in the progression (and onset) of a disease?’ The latter question takes this vast complexity into account and is therefore more likely to provide the desired answer.

HIGH-DIMENSIONAL DATA When associating two sets of variables – where one contains the phenotypic information and the other the genetic information of a group of subjects – common underlying structures can be revealed, thereby unravelling the genetic basis of the disease under study. This association can be achieved by a multivariate statistical method known as canonical correlation analysis (CCA). CCA can find underlying structures in datasets by maximising the correlation between a weighted linear sum of all the variables in one set and a weighted linear sum of all the variables in the other set. It’s these weights that tell us something about the importance of the corresponding variables: the higher the absolute weight, the more it contributes to the underlying structure.

However, like many statistical methods, CCA fails in the presence of high-dimensional, complex data, which is exactly the kind of data that we are interested in. As a consequence, a number of problems arise when analysing high-dimensional microarray data: (i) the end-result can be difficult to interpret due to the high number of variables; (ii) multicollinearity, caused by e.g. genes acting in a complex network or single nucleotide polymorphisms (SNPs) that are located closely to one another, which results in unstable weights; and (iii) due to the high number of variables compared to the number of subjects in the study, association studies will give many different solutions that are equally good (called over-fitting). So, in the presence of high-dimensional data, no conclusions can be made about the importance of the variables based upon the corresponding weights.

NEW APPROACH Many researchers in the field recognise these bottlenecks and different methods have been developed to solve these problems in the regression context. Among such methods is penalization – like ridge regression, the lasso [1] and the elastic net [2] – where a penalty is posed upon the size of the weights. Consequently, the weights become more stable in the presence of multicollinearity (ridge regression); the number of variables is reduced by means of continuous variable selection (lasso); or – by combining the previous two penalization methods – highly correlated variables get similar weights and are in or out of the model together (elastic net). To present a much needed adaptation of these methods into multivariate models lags behind.

By adapting these penalization methods to CCA, we were able to obtain stable, sparse, interpretable results. We applied this method to...
different datasets containing two sets of variables, one with phenotypic markers and one with genetic markers. From within these sets, variables were selected which resulted in a high correlation between the phenotypic and genetic sets and a high correlation within these two sets. Clustering within the sets resulted in the preservation of haplotypes and genetic networks. Based on this initial success, we extended this so-called penalized canonical correlation analysis [3] in such a way that the method was able to handle even more complex data-structures as encountered in genetics and proteomics.

PROMISING RESULTS The first hurdle we wanted to tackle was to make our datasets capable of containing categorical data (caused by e.g. the presence of SNPs). To cope with this we implemented optimal scaling to penalized CCA. We then made even further extensions to associate a large set of genetic markers with repeatedly measured phenotypes – like cholesterol – that may vary over time and are therefore often measured at different time points. This way we were able to relate the (stable) genetic basis to (dynamic) phenotypical changes over time. Finally, because of the complexity of factors involved in the onset and development of complex diseases, we extended our method to allow for the association of more than two sets of variables if needed. The results of the different studies are promising: many known, but also several unknown association were revealed, the latter of which could be of interest for future biomedical studies. We thus provide researchers with much needed, newly developed, statistical tools to explore and analyse the vast amount of data obtained. It is clear that many challenges still remain on the path to unravel the genetic basis of complex diseases. But these tools can bring us one step closer to understanding their genetic complexity.

REFERENCES

CONTACT
Sandra Waaijenborg, MSc
Department of Clinical Epidemiology, Biostatistics and Bioinformatics
Academic Medical Centre, Universiteit van Amsterdam
PO Box 22798, 1100 DE Amsterdam
Web: www.amc.nl
E-mail: s.waaijenborg@amc.uva.nl

KEY CONCEPTS
- The risk of many common diseases, like cardiovascular disorders, metabolic defects or cancer, is thought to be influenced by multiple genes as well as by environmental factors.
- Microarray technology aimed at elucidating pathogenesis of complex diseases generates a growing amount of data.
- Appropriate statistical techniques are needed for quantifying the association between the multi-dimensional (phenotypic and genotypic) data.
- Canonical correlation analysis (CCA) provides a suitable multivariate statistical technique, but fails in the presence of high-dimensional, complex data as generated by microarray experiments.
- The department of Clinical Epidemiology, Biostatistics and Bioinformatics at AMC, UvA developed multivariate statistical methods that can deal with the complexity of high throughput data. The work is part of NBIC’s BioRange programme.
- Adapting penalization methods to CCA has resulted in a method to handle even more complex data-structures.
- The developed tool can bring biomedical researchers one step closer to understanding the genetic complexity of common diseases.
- R package is still under construction, but R codes are available upon request.
- A part of the work as described in the enclosed article will be published in Bioinformatics. The author is now finishing her PhD thesis based on this research.

By the editors

---

PENALIZED CANONICAL CORRELATION ANALYSIS
To reveal an underlying structure shared by two sets of variables, penalized CCA finds a weighted linear combination of a selection of variables in one set which correlates maximally with a weighted linear combination of a selection of variables in the other set. These weighted linear combinations are summarised in the underlying structures \( \omega \) and \( \xi \) for the linear combination of the phenotypic markers and genetic markers, respectively. Some markers contribute little to nothing to the underlying structure and can subsequently be deleted from the model (dotted lines).
The identification of homogenous patient subgroups for diseases presenting clinical heterogeneity is of major interest. Osteoarthritis, Parkinson’s disease or major depressive and anxiety disorders represent illustrative examples. Finding subtypes may be helpful in unveiling pathogen mechanisms and subsequently in developing tailored prevention strategies and therapies.

Fabrice Colas from the Leiden Institute of Advanced Computer Science (LIACS) developed a methodology to facilitate the search for homogeneous subtypes within heterogeneous diseases. As no automated subtyping methodology existed, he and his colleagues developed the so-called ‘RSDisc’ package for subtype discovery. It identifies and evaluates subgroups within large patient groups defined by different variables. The scenario was implemented as an open source software package in the R platform for statistical computing and graphics. Roughly, the package consists of a data preparation part, a clustering algorithm part to find the subpopulations, and specific tools adapted to the evaluation and representation of the discovered subtypes. Data can be all kinds of variables, for example the severity of certain symptoms, from very severe to no symptoms at all. Data have to be transformed into scale-invariant quantities or quantities that can be used in the algorithm for cluster analysis. This is what the beginning of the program does: data transformation. After this, different subtypes or clusters are calculated based on the variables. Many different algorithms are available for cluster analysis. Colas and colleagues chose the one from the Statistics Department from the University of Washington because, “It is well understood and it leads to
Tineke van Veen is assistant professor at the Psychiatry Department of the LUMC. Her research focuses on patients with depression and anxiety. The disease is a heterogeneous disorder. “We assume that different types of depression exist, which have distinct pathogeneses, different disease courses, and specific treatment requirements. Our hypothesis is that each of the subtypes has a different set of genetic and environmental factors. To identify these factors that contribute to depression, we first need to find subgroups.”

Professor Joost Kok, Scientific Director of LIACS, advised Van Veen to contact Fabrice Colas to assist her in finding these subtypes with the R SDisc package. She explains: “Although I am not yet a very skilled user of the package, I think it is a very good tool to assist in finding subtypes. The calculation is done automatically and very fast. In a first analysis, we found subgroups of patients who experienced different frequencies of adverse events, like divorce or losing a relative.” Van Veen remarks that close collaboration with a statistician is required in order to identify subtypes. “As Fabrice mentioned it is important to discuss how to select and code variables and how to interpret results.”

To validate the subtypes from the first study, a second analysis was performed on independent data. Possibly because of different processing of the data in the different studies, the results are not as similar as expected.

“I think it is a very good tool to assist in finding subtypes”

“We need to identify the differences in data processing and find out if we can rectify this.” Soon Van Veen will be doing the analyses herself.

THE DEVELOPER  Fabrice Colas specialised in data mining after his studies in electrical engineering in France. As a PhD student at the LIACS, Colas started working on a question from Ingrid Meulenbelt researcher at the department Molecular Epidemiology, LUMC. Colas explains: “Ingrid wanted to find homogeneous subgroups within an osteoarthritis patient group under study. The goal was to assess whether the spread of the disease was random or whether it followed a particular pattern.” To address Meulenbelt’s problem, Colas started his work on the subtype discovery data mining scenario. It involved many discussions with the researchers. “As a data mining specialist, I do not understand the meaning of the patient data well: how they should be interpreted, how the results should be evaluated and represented, et cetera. So we need to exchange to select the right data transformation, the most appropriate cluster algorithm, or the right statistical tests that will ‘justify’ the results. Furthermore, the researchers ask for reliable and reproducible subtype analysis.

“The most awkward element of each analysis remains the data preparation”

This means that if one wants to repeat the numerical analysis, all the information for this should be logically archived within the package output.” Ingrid Meulenbelt used the prototype of the data mining scenario developed by Colas to identify subtypes of osteoarthritis. Based on these subtypes, she now performs genetic analyses. Yet, because of the generality of the research question it addresses, the scenario was soon applied to other complex pathologies like Parkinson’s disease and diagnosis of aggressive brain tumours. Most recently it was also applied to research on psychiatric disorders. Hence, the scenario was gradually turned into an R package. Now it is available on the internet with a tutorial that leads you through all the steps. Still, according to Colas, the most awkward element of each analysis remains the data preparation. “You need to decide which data you want to enter and in which way. The decision can influence the nature of the results greatly.”

THE USER  Tineke van Veen is assistant professor at the Psychiatry Department of the LUMC. Her research focuses on patients with depression and anxiety. The disease is a heterogeneous disorder. “We assume that different types of depression exist, which have distinct pathogeneses, different disease courses, and specific treatment requirements. Our hypothesis is that each of the subtypes has a different set of genetic and environmental factors. To identify these factors that contribute to depression, we first need to find subgroups.”

Professor Joost Kok, Scientific Director of LIACS, advised Van Veen to contact Fabrice Colas to assist her in finding these subtypes with the R SDisc package. She explains: “Although I am not yet a very skilled user of the package, I think it is a very good tool to assist in finding subtypes. The calculation is done automatically and very fast. In a first analysis, we found subgroups of patients who experienced different frequencies of adverse events, like divorce or losing a relative.” Van Veen remarks that close collaboration with a statistician is required in order to identify subtypes. “As Fabrice mentioned it is important to discuss how to select and code variables and how to interpret results.”

To validate the subtypes from the first study, a second analysis was performed on independent data. Possibly because of different processing of the data in the different studies, the results are not as similar as expected.

“I think it is a very good tool to assist in finding subtypes”

“We need to identify the differences in data processing and find out if we can rectify this.” Soon Van Veen will be doing the analyses herself.
Finding relationships in biological data

Bayesian networks, the topic of Anand Gavai’s recent thesis, is a lot closer to everyday life than one might expect. In addition to analysing complex biological data, the technique can be used for analysing sports results such as football, predicting stock markets, wireless sensor networks, money laundering, and crisis management.

“The very first applications and a classical example of Bayesian networks is the Microsoft Office Assistant”, Gavai says. “But you have to have some background knowledge to make sense and money out of it.”

There are two aspects of Gavai’s thesis. First, he wrote a user friendly software program for storing and analysing biological data and then he used Bayesian networks techniques for interpreting these data. Classical techniques for the analysis of microarray or LC-MS data are focused on multivariate statistics. These techniques are very good at describing the data, but cannot take into account indirect relationships between genes of interest, for example. In
this way a correlation but not a causal relationship can be found between data, and no future prediction can be made. Gavai explains: “You find a relationship between two genes, but you do not find the nature of this relationship. For example, if you see a young man and an old man who look very much alike talking to each other, you may think that they are father and son. But they may be strangers and accidentally look alike. Classical statistics merely notices that the men look alike.”

A new technique called Bayesian networks (officially termed a ‘probabilistic graphical model’ and having nothing to do with Bayesian statistics) can find causal relationships. The technique uses statistical concepts, but cannot be considered a statistical technique, says Gavai. “Using Bayesian networks has two big advantages. We can analyse massive amounts of data because the technique is developed from a computer science background. At the same time, it represents the data in a graphical, intuitive way, which makes it possible to find complex relationships when many variables are involved.”

**PRIZE-WINNING SOFTWARE** The project started in 2005 as a collaboration between Prof. Michael Muller’s Human Nutrition group and Prof. Jack Leunissen’s Bioinformatics group, both at Wageningen University. “High throughput technologies such as microarray technologies were gaining momentum at that time. A system was needed to manage and store this data. The next step was to glean knowledge from this data.”

Anand Gavai started his research by designing a microarray management and analysis database called MADMAX. It won the NBIC venture challenge in 2007 for the best bioinformatics program. “The unique selling point of my program is that it is able to manage and analyse data from many different types of microarrays”, explains Gavai. “The devices come in various shapes and sizes. The data which is generated from them is also in different formats. We showed that we have a system that can take care of data coming from any chip. The other advantage is that no programming is required for the biologist using the software. We created a visual interface around the core program. This also reduced the analysis time from several months to days or hours.” The software started out as a tool to be used within the combined project of BioRange and the Netherlands Nutrigenomics Consortium (a part of the Top Institute Food and Nutrition) and now has about 80 users around Europe and the US. “It is actively used in the Nutrigenomics Consortium, and the co-developers Ke Lin and Philip de Groot are now the dedicated scientific programmer and statistician.”

**SELECTING GENES** Next, Gavai set out to find out the metabolic state of an organism using the MADMAX program to store and analyse data. In human nutrition, people are interested in finding the effect of food on a certain biological condition, he explains. “From a list of mice genes we selected a subset of 1000 genes that we think determine whether an individual has been fasting or being fed. We focused on five metabolic pathways, and using Bayesian networks, we were able to narrow the list down to three or four genes that are responsible for determining an individual’s metabolic state.” To keep things relatively simple, homogeneous data coming from one microarray experiment was used, taking no clinical parameters into account. The next step in the research was using heterogeneous data, this time blood samples from twenty smokers. The goal was to find the key genes that cause an elevated cotinine level in smokers’ blood. The clinical data was collected at Maastricht hospital. “We combined the microarray data with the clinical data from the blood samples”, Gavai explains. “We found genes having an effect on small cell lung cancer. This was the first time that different types of data were used in epidemiological research. We collaborated with biologists to verify that the relationships we had found were making sense.” Finally, Gavai wanted to show how to use all these techniques for metabolomics data, which is completely different from microarray data. For this, three different types of tomatoes were studied. Tomato breeders are interested in biomarkers (metabolites) for ripening of tomatoes. The oxilipin pathway that indicates this ripening was reconstructed using Bayesian networks. “A certain combination of nodes in a Bayesian network can indicate a biomarker for a specific trait”, says Gavai, showing that his methods can indeed be applied to metabolomics data. This is promising, he says, since microarrays may become obscure in the future. But the methods he developed will not.

**CURRENT JOB** At present Anand Gavai works at the VU department of Sciences as a postdoc in a BioRange project in collaboration with NCSB (Netherlands Consortium for Systems Biology). He is a liaison officer in addition to being a researcher. His research includes the prediction of metabolites from an existing metabolic reconstruction database. With this knowledge, biological pathways such as fermentation processes may be optimised. “In my current job I do not use my program anymore, but in my spare time I sometimes analyse data with it, just as a hobby. In the future I would like to take on a project like this.”

**ADDITIONAL INFORMATION**
Who is Christin Christin?
I come from Tangerang, Indonesia. I obtained my bachelor's degree in computer science from the University of Indonesia. I already realized then that I wanted to do something with biological data. Bioinformatics was unknown there, but my supervisor knew about aligning DNA sequences and advised me to write my thesis about it. He also advised me to go to Europe to continue in bioinformatics. I obtained my master's degree at the Chalmers University in Gothenburg, Sweden, with a thesis about scatter plot partitioning of NMR data for the prediction of protein structure from the NMR spectra using Bayesian networks. In 2006, on the advice of my supervisor, I applied for the PhD position with Professor Rainer Bischoff’s Group at the University of Groningen in the Netherlands. I ended up at their project ‘Biostatistics for proteomics-based clinical biomarker discovery’ headed by Peter Horvatovich. I am very pleased to be working in this Groningen research group with its extensive facilities and collaborations.

How do you fit in the research group as a bioinformatician?
When I started, I was the only bioinformatician in the group. Most problems I had to solve myself by reading the references on the internet. Now there are three bioinformatics PhD’s and one scientific programmer in the group. Each of us has his own subject: data pre-processing and LC-MS analysis; time alignment like me; and statistical validation. I get my data from the biochemists working in this group. It is often difficult for them to imagine what I am doing on my computer so I have to explain what I do in a very general way. But we are eager to learn from each other.

With whom do you cooperate, inside and outside of your institute?
Besides Groningen, I work with my second promotor Professor Age Smilde and his research group Biosystems Data Analysis at the University of Amsterdam. They are experienced in statistics and bioinformatics. In addition, we also cooperate with the VU Medical Centre. We also have a collaboration with IBM research in New York, which has a research centre for computational biology.

What advice do you have for bioinformatics students?
To be a bioinformatician, you have to be very flexible and adaptable because you have to deal with different kinds of data and with people who have no idea what you are doing. Furthermore, you have to have strong analytical and computational abilities. However, when you are ambitious and motivated, it is easy to become a bioinformatician. Almost all the information such as computer programming and data structure is available on the internet. You need to understand programming logic and concepts. I got this knowledge from one of the best teachers in Indonesia.
NBIC IN STOCKHOLM

The annual meeting of the International Society for Computational Biology (ISCB) is held in Europe once every two years, where it merges with the European Conference on Computational Biology (ECCB). This year the 17th annual conference on Intelligent Systems for Molecular Biology (ISMB) and the 8th European Conference on Computational Biology (ECCB) were organised in Stockholm, Sweden.

For a full week, Stockholm was transformed into a lively meeting place for bioinformatics researchers from all over the world. Traditionally, the Student Council Symposium is organised the day before the official start of the conference. This year a representative from the Dutch network of PhD students (RSG Netherlands), Jeroen de Ridder (TU Delft) was co-chair of the organising committee. The Symposium was a tremendous success, with a record attendance of 125 young scientists. Together with many Dutch scientists, NBIC was well represented at the ISMB/ECCB 2009 as gold sponsor of the conference and the Student Symposium. Many young, enthusiastic researchers visited the NBIC booth to find out more about NBIC programmes, and in particular about the new BioRange projects with vacancies for PhDs and PostDocs. Barend Mons, leader of NBIC’s support programme BioAssist, presented the CWA initiative (see also short news: Concept Web Alliance) on several occasions, which received much attention and strong support from international scientists.

DUTCH PHD STUDENTS ELECTED CHAIR AND SECRETARY OF ISCB STUDENT COUNCIL

The elections for new members of the Student Council of the International Society for Computational Biology (ISCB) turned out in favour of the Dutch candidates, both of them PhD students. Jeroen de Ridder (TU Delft) has been elected Chair and Jayne Hehir-Kwa (UMC St Radboud Nijmegen) will be taking up the position of Secretary. Jeroen already has some issues he would like to address through the Student Council. He explains: “Personally, I would like to expand the Regional Student Groups to a larger number of countries with special focus on developing nations. Through the Student Council, we can bring those students into contact with the bioinformatics community in the US and Europe.” Both Jayne and Jeroen are currently board members of the Regional Student Group in the Netherlands, co-founded by NBIC. Their one-year term with the ISCB Student Council will start March 2010.

More information on the Student Council: www.iscb.org/student-council

Dutch PhD Students Elected Chair and Secretary of ISCB Student Council

The elections for new members of the Student Council of the International Society for Computational Biology (ISCB) turned out in favour of the Dutch candidates, both of them PhD students. Jeroen de Ridder (TU Delft) has been elected Chair and Jayne Hehir-Kwa (UMC St Radboud Nijmegen) will be taking up the position of Secretary. Jeroen already has some issues he would like to address through the Student Council. He explains: “Personally, I would like to expand the Regional Student Groups to a larger number of countries with special focus on developing nations. Through the Student Council, we can bring those students into contact with the bioinformatics community in the US and Europe.” Both Jayne and Jeroen are currently board members of the Regional Student Group in the Netherlands, co-founded by NBIC. Their one-year term with the ISCB Student Council will start March 2010.

More information on the Student Council: www.iscb.org/student-council
CONCEPT WEB ALLIANCE
Scientific progress is hampered by the very large and ever increasing amounts of scientific information, and with data that are often inaccessible or not interoperable. NBIC has proposed setting up an international alliance of scientific organisations and industry to join forces and collaborate on finding solutions to these problems. For that purpose, a meeting was organised in New York (8 May, 2009) with over 50 representatives from academia as well as from the private sector, worldwide, which resulted in the establishment of the Concept Web Alliance (CWA) initiative. The CWA is foreseen as a not-for-profit international organisation, aimed at leading by example, and so hopefully becoming a trusted party when it comes to data models for semantic triples, setting de facto standards of interoperability, best practices, et cetera. http://www.nbic.nl/about-nbic/affiliated-organizations/cwA/

NBIC AFFILIATE OF INTERNATIONAL SOCIETY
In June NBIC became an affiliate of the International Society for Computational Biology (ISCB). ISCB has approved NBIC’s application in the category of Affiliated National Centres and Networks. The ISCB Affiliates program is designed to forge links between ISCB and regional non-profit memberships. These members are groups, centres, institutes and networks including researchers from various institutions and/or organisations within a defined geographic region involved in the advancement of bioinformatics.

NBIC EDUCATION: BIOWISE HIGHLIGHTS
SIB – NBIC COLLABORATION ON EDUCATION
NBIC and the Swiss Institute of Bioinformatics (SIB) are joining forces in educating their PhD students. Agreements have been made to open up their mutual PhD courses for both Dutch and Swiss PhD students. Recently, the first exchanges took place: a Swiss PhD student participated in the Course ‘Managing Life Science Information’ at the NBIC PhD School, while a Dutch PhD student from Delft participated in the SIB Summer School in Lugano, Switzerland. SIB and NBIC are actively promoting these kinds of exchanges. NBIC has travel funds available for PhD students who want to go to a Swiss PhD Course.

BIOINFORMATICS@SCHOOL
In 2008/2009 another 2837 Dutch high school students became bioinformatician for a day. Beginning in early 2010, pupils from elementary school will also be able to learn about bioinformatics (www.bioinformaticsatschool.eu). Stay informed: subscribe to the bioinformatics@school newsletter by sending an email to office@nbic.nl

NBIC PHD SCHOOL NEWS
Last May, the PhD Course ‘Managing Life Science Information’ attracted 16 participants to Amsterdam. The participants were 11 PhD students and several researchers, as well as 2 scientific programmer(s) from the BioAssist project. The third course of the NBIC PhD School will take place 11-15 January 2010 in Delft. The topic will be ‘Algorithms for Biological Networks’. For more information about the PhD School courses, see www.nbic.nl/education/nbic-phd-school

DISTINGUISHED VISITING SCIENTIST STIPEND
In May, the Netherlands Genomics Initiative (NGI) granted five applications for a Distinguished Visiting Scientist Stipend. NBIC BioRange researcher Prof. Ritsert Jansen received the Stipend to collaborate at Jackson Laboratory in Maine USA. New applications for the NGI Stipend may be submitted by 1 April 2010: http://www.genomics.nl/InternationalActivities/Disting__Stipend.aspx
PHD THERSES
Since the last issue of Interface in May this year, three BioRange PhD theses were successfully defended. We congratulate the young doctors:
- Yunlei Li, 20 September 2009, Delft University of Technology, promoter: Prof. M.J.T. Reinders
  Title: Exploiting noisy and incomplete biological data for prediction and knowledge discovery
- Anand Gavai, 8 June 2009, Wageningen University and Research Centre, promoters: Prof. Jack A.M. Leunissen and Prof. Michael Muller
  Title: Bayesian networks for omics data
- Richard Notebaart, 6 May 2009, CMBl, Radboud University Medical Centre, promoters: Prof. Bas Teusink and Prof. Roland Siezen
  Title: Integrative bioinformatics of metabolic networks

CLS – NEW PROJECTS
The Netherlands Bioinformatics Centre (NBIC), the Netherlands Genomics Initiative (NGI) and the Netherlands Organisation for Scientific Research (NWO Physical Sciences) have launched a second round of the research programme Computational Life Sciences (CLS). Recently, on June 2009, five new projects were awarded:
- The evolution of stochastic heterogeneous networks as hedging adaptations to fluctuating environments. Coordinator: P. Haccou, PhD (Leiden University)
- Reverse physiology of the cortical microcircuit. Coordinator: A.R. Houweling, PhD (Erasmus Medical Centre, Rotterdam)
- Multi-scale modelling of calcification in scleratinin corals. Coordinator: J.A. Kaandorp, PhD (University of Amsterdam)
- The co-evolution of the receptor signalling network of natural killer cells with its ligands. Coordinator: Prof. R.J. de Boer (Utrecht University)
- The regulatory network underlying malaria parasite-host interactions. Coordinator: T.M.H. Dijkstra, PhD (Radboud University Nijmegen)

NBIC SUPPORT:
BIOASSIST HIGHLIGHTS

BIOASSIST ENFORCED BY ENGINEERING TEAM
As of September 2009, a dedicated, centrally managed, but distributed software engineering team supports the BioAssist support platforms to develop professional software and content services.

The NBIC Support programme, BioAssist, is a collaborative endeavour in which biologists and (bio)informaticians organise bioinformatics support in a collaborative fashion. The BioAssist programme picks up ‘code’ and ‘content’ of generic value and assists the original developers in developing these assets for community use. So called Support platforms integrate tools and databases to provide specific bioinformatics functionalities needed for the analysis, integration and interpretation of experimental data from Genomics, Proteomics and Metabolomics approaches.

The tools and databases incorporated in these platforms need not be provided as a central service, but can remain at distributed and connected local computer systems (GRID nodes) of the contributors. Consequently, these tools and databases are continuously updated and maintained by the respective NBIC partners. New and emerging technologies from the e-Science, Web and Grid communities increasingly support this interactive and flexible approach. Moreover, the approach ensures timely adaptation of the platforms to keep pace with the changing needs and increasingly complex questions of the life sciences community. Get involved!

Please contact Barend Mons, programme leader BioAssist, or one of his colleagues: http://www.nbic.nl/support/getting-involved/

COLOPHON

Interface is published by the Netherlands Bioinformatics Centre (NBIC). The magazine aims to be an interface between developers and users of bioinformatics applications.

Netherlands Bioinformatics Centre
260 NBIC
P.O. Box 910
6500 HB Nijmegen
t: +31 (0)24 36 15600 (office)
f: +31 (0)24 36 19395
s: office@nbic.nl
w: http://www.nbic.nl

EDITORIAL BOARD NBIC
Antoine van Kampen
Ruben Kok
Marc van Driel
Karin van Haren

COORDINATING AND EDITING
Marijn van Opstal
Bèta Communications, The Hague

DESIGN
COLEVERFRANKE, Utrecht
LAV-OUT

PHOTOGRAPHY
Thijs Roelmans
Stockphotos from ISTockphoto

PRINTING
Beatenzet, Zoetermeer

DISCLAIMER
Although this publication has been prepared with the greatest possible care, NBIC cannot accept liability for any errors it may contain.

To (un)subscribe to ‘Interface’ please send an e-mail with your full name, organisation and address to office@nbic.nl

COPYRIGHT NBIC 2009

Martijn Huynen, UMC St Radboud Nijmegen
Joost Kok, Leiden Institute of Advanced Computer Science
Marco Rosa, Institute for Informatics
Harold Versteegen, Kyogene N.V.

EDITORIAL ADVISORY BOARD
Maurice Brouwels, Sara Computing and Networking Services
Tineo Brilt, Swammerdam Institute for Life Sciences
Roeland van Ham, Plant Research International, Wageningen UR
TOOLS THAT ARE NEVER USED

Many years ago, I wanted to systematically compare the clinical features associated with different genetic diseases. All I needed was a searchable database that would allow me to analyse phenotypes and then I would see some really interesting biological patterns. I quickly found someone who for over 20 years of professional life had been developing the ultimate database for my purpose. I approached the venerable gentleman who had put the toil and sweat of an entire career into creating this masterpiece. Luckily, he immediately recognised the beauty of my ideas and we agreed to collaborate. All that was needed was for him to get an agreement with a professional publisher who would distribute this wonderful piece of uniquely curated data, for a small and well-deserved fee. So I waited. And waited. And waited some more. But the contract was never finalised, and one year later the creator of this database fell ill and then died taking his database with him to the grave.

This was my first experience of how a perfectly sensible bioinformatics scheme can fail because of the inaccessibility of data. And the problem isn’t always about the data being physically beyond reach. What is much more common is the opposite situation in which a clever bioinformatics tool is made freely available to the community. A paper is written for a major journal, and the superiority of this application for solving the major problems of modern day genomics is proven by P-values and illustrated by specific examples. Previously developed tools are shown to be hopelessly out of date; hence a hopeful vision for the future is painted. But none of this works out and then there is only silence. In spite of everything, the tool is never used at all.

Why is this, and do we need to do something about it? Let’s first determine what is wrong. In the first place, the tool may not solve a problem that is of any real consequence to the biological community. This is relatively frequent and is often related to the fact that to be better than random isn’t actually of any great significance. We biologists have significant prior knowledge about the questions we pose of the data. So we tend to trust our own judgement, and will use a tool only if it is far superior to what we can do ourselves. I freely admit here that the first tool that I proposed and that was designed by my bioinformatics friends in Nijmegen has so far solved exactly one biological question.

The second problem is even simpler. Many bioinformatics tools were designed to be used and applied by bioinformaticians only. Hence they have no appeal to biologists who lack the basic skills to apply them in practice.

Yet, these are not problems that we should try to solve. People must and will pursue their ideas however wacky or inconsequential they may be. Let’s celebrate the natural diversity of tools, and put them in the great museum of inconsequential scientific achievements. Let’s assume that they somehow contribute by throwing just that little bit of light that allows their competitors to do a better job next time. For me personally, I shall continue to plan for building the ultimate contraption to understand the collective genetic diseases or the phenome. And if I fail, I shall at least have had a good time. Let’s enjoy life while it lasts!