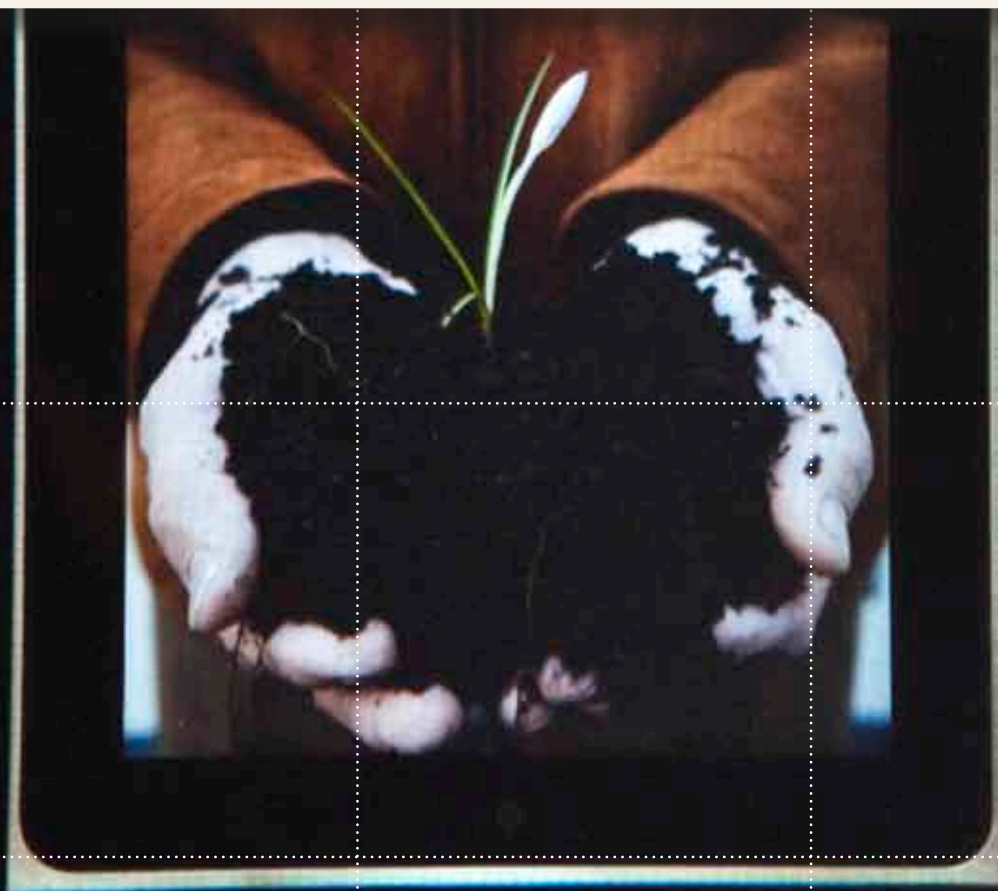


INTER DEVELOPMENTS IN BIOINFORMATICS FACE

→ **Cover story**
**Metagenomics:
unravelling microbial
communities**

→ **Interview**
Jeroen Raes

Issue 7 | April 2011
Netherlands Bioinformatics Centre



Content

COVER PHOTO: The microbial ecosystem in the soil and plants plays a vital role in the health and productivity of crops. Metagenomics enables to investigate such microbial communities.

EDITORIAL	Taking NBIC to Broadway	3	
HIGHLIGHTS	News and Retrospective	4	
COURSE	Computational approaches to biomolecular structure and function	7	→ <i>It was a great introduction for my research work</i>
COVER STORY	Metagenomics: tackling genomes on a post-individual level	8	
BUSINESS	R-Consultancy: consulting services in bioinformatics	11	
INTERVIEW	Jeroen Raes (University Brussels) Professor of bioinformatics and (eco) systems biology	12	→ <i>We are not purely human: we live in symbiosis</i>
PORTRAIT	Seven questions for Jan Bot	15	
HANDS ON	Pindel: fast algorithm detects variants in small fragments	16	→ <i>Detecting simple insertions and deletions as well as complex structural variants</i>
THESIS	Back tracing evolution of the Ras signal transduction pathway	18	
PROGRESS	How genomes as information structures impact human identity By Jan van Baren	20	
PROGRESS	Playing games with our future By Sven Warris	22	
COLUMN	Hans Roubos: Converging technologies	24	

Taking NBIC to Broadway

Barend Mons

NBIC SCIENTIFIC CO-DIRECTOR AND BIOASSIST CHAIR



Following up on the editorial by Jaap Heringa in the previous issue of Interface, 'Bioinformatics comes of age', the question arises of whether the NBIC community is ready for transition to maturity. Actually, more is expected of us than organic maturation; we need to make a leap or two in

our evolution. A quote from a colleague concerning the life sciences field: "it isn't difficult to detect a mounting sense of panic". If predictions are half-true, the data deluge will drive bioinformatics to the centre of all future data-intensive biological research. Are we ready?

Not really, but collectively we can do it, provided we avoid parochialism and we implement respectful collaboration of specialists. We know the disasters when cultures clash without mutual trust and respect: biologists thinking they are autodidact software engineers and informaticians who think they understand biology after reading three textbooks. NBIC is there to support genuinely collaborative efforts between informatics and biology, rather than 'bioinformatics as an afterthought'. Let's use the metaphor of a musical. It is obvious that the script and music are the basis of success. Singers and dancers are equally important. Singers sometimes have to make a few moves, which with some goodwill can be taken for dancing, while the dancers sometimes fill in some minor vocals. Mutual respect is needed. However, even if the performers are top class, in the absence of a conductor the performance will go nowhere. Creative people do not like to be conducted, but real singers and dancers have learned their lesson:

orchestration and direction is an integral part of success! Similarly, NBIC is not just a foundation in Nijmegen, but rather a growing community of excellence with one common goal: to face complex biological questions and massive datasets. The NBIC office works behind the scenes on crucial issues that may seem uninteresting to scientists. Solid partnership agreements were reached with all relevant universities and research institutions, as well as a growing number of companies. These address arrangements on Intellectual Property, confidentiality, societal implementation (valorisation) and more of such 'mundane' issues. Building on the scientific excellence of BioRange, the BioAssist programme has been expanded and reshaped to answer to the evolving needs of the scientific community in the form of an Open Innovation Environment. The last BioAssist meeting I attended was vibrant with energy and had about 40 participants from all over the country. 'Power users' such as BBMRI-NL and CTMM are increasingly discovering NBIC's Open Innovation Environment. At the same time, the BioWise educational activities are developing to broaden support for the field. The first group of 40 scientists has accepted NBIC Faculty status and they can now propose more faculty. This growing group will be engaged in shaping NBIC's future.

All these efforts are aimed at consolidating the growing bioinformatics community and preparing it for the next wave of challenges, for which we will need each other badly. Obviously, this rapid climate shift is not taking place just on the few square kilometres of The Netherlands. If a small country like ours wants to play any role in the international scene, we need to take our musical from Nijmegen to Broadway. So, sing, dance, write papers, but also cherish and contribute to the added value of our NBIC community.

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 9101
6500 HB Nijmegen
t: +31 (0)24 36 19500 (office)
f: +31 (0)24 89 01798
e: office@nbic.nl
w: <http://www.nbic.nl>

THE FOLLOWING PERSONS CONTRIBUTED TO THIS PUBLICATION:

Jan van Baren, Jan Bot, John van Dam,
Philip de Groot, Sacha Hijum,

Daniel Hoffmann, Michiel Kleerebezem,
Keiran Raine, Hans Roubos, Eddy Smid,
Gert Vriend, Sven Warris, Christoph Wilms,
Kai Ye, Barbara Zarzycka

EDITORIAL BOARD NBIC

Ruben Kok
Marc van Driel
Karin van Haren

COORDINATING AND EDITING

Marian van Opstal
Bèta Communicaties, The Hague

DESIGN

CLEVER[®]FRANKE, Utrecht

LAY-OUT

t4design, Delft
PHOTOGRAPHY

Thijs Rooimans
Stockphotos from iStockphoto

PRINTING

Bestenzet, 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

News & Highlights

Formal launch of the NBIC Consortium with 21 partner organisations

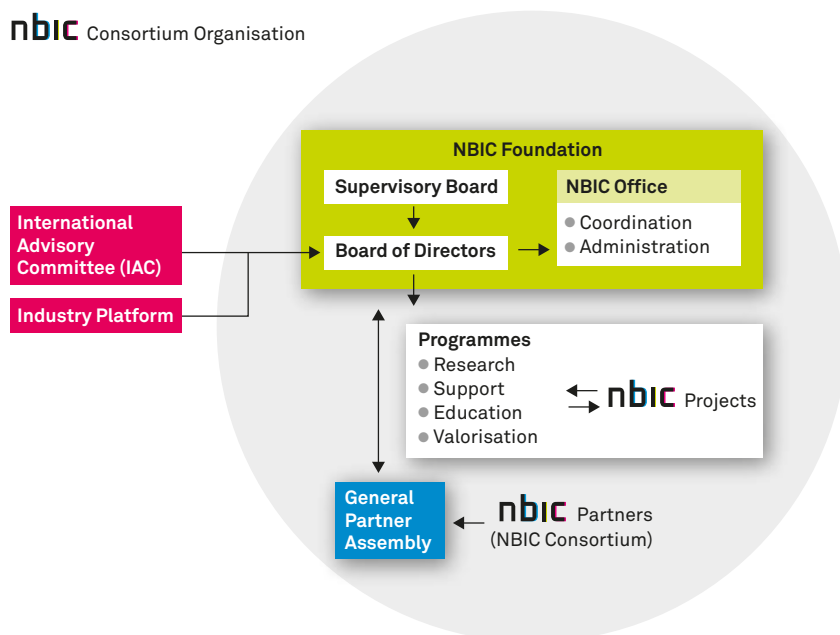
A crucial step in the sustainable development of NBIC as an overarching public-private initiative is the formalisation of the partnership between the connected organisations, with transparent rules of play. Building on what was available in the field, it still took two years of preparation to finally launch the NBIC Consortium in late 2010. The process has involved representatives of the Dutch leagues of universities and

university medical centres, as well as industry. The focus was on creating a flexible consortium model to support NBIC's development towards an open innovation environment.

Partner organisations are connected to the NBIC Consortium by signing a generic Partner Agreement. This agreement governs issues at the level of governance and generic project IP aspects. To stimulate

a flexible collaboration between the NBIC partner organisations, a dedicated Project Agreement formalises the project-level details among the partners involved. The project partners have conditional access to project results and to services developed within the NBIC Consortium. All partner organisations are represented in the yearly General Partner Assembly, which advises with respect to organisational matters.

nbic Consortium Organisation



Since its launch in November 2010, the NBIC Consortium has attracted 21 partners, among which are 8 universities, 7 university medical centres, 4 private research institutes. The Consortium also includes Philips and DSM as the first industrial partners, while new partners in industry have already shown serious interest in joining. Partnership is open to organisations with an active interest in bioinformatics, including academia, private research institutes and companies.

For details, please see the NBIC website (www.nbic.nl/about-nbic/organisation/nbic-partners/) or contact the NBIC office (office@nbic.nl).

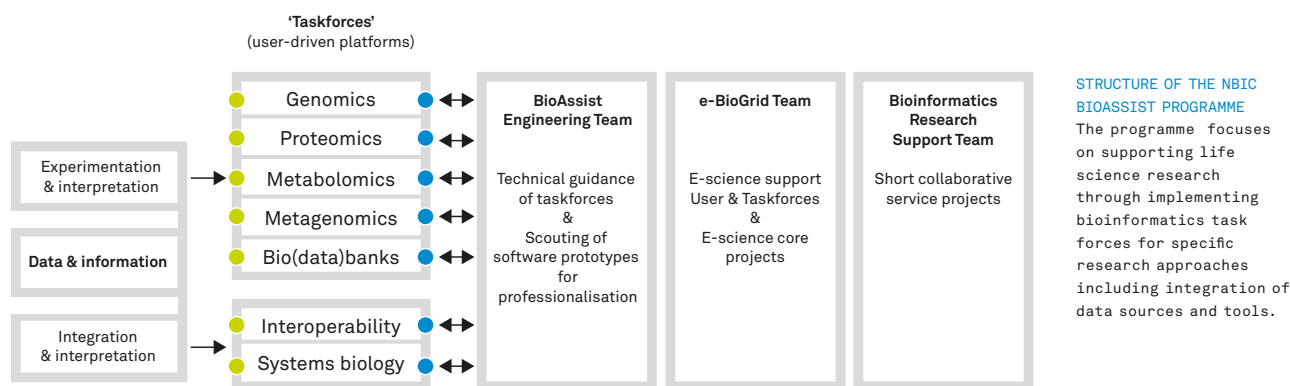
INTERNATIONAL REVIEW COMMITTEE APPLAUDS DUTCH BIOINFORMATICS

On March 24/25 2011, NBIC received an international committee to review its programmes in bioinformatics research, support and education that have been set up with the NBIC partners. The review was part of an external review organised by the Netherlands Genomics Initiative (NGI). The independent committee consisted of prof. Jan van Bemmelen (NL),

prof. Burkhard Rost (USA, Germany), prof. Michal Linial (Israel), dr. Ionannis Xenarios (Switzerland), prof. Graham Richards (UK) and dr. Niklas Goldberg (Sweden).

In a two-day meeting in Amsterdam the committee received a detailed insight in NBIC as the Dutch platform for collaboration in bioinformatics. Unanimously, the committee members were highly enthusiastic about what has been set up in recent years. Across the board, the

NBIC partnership was valued very high in terms of its programmes and organisation. The committee applauded the critical mass reached through the convergence of the Dutch bioinformatics field in NBIC and strongly advises continuation of NBIC as a whole. NBIC was urged to strengthen its outreach activities and to play a more active role in the international field.



BIOASSIST

IMPULSE TO e-SCIENCE FOR BIOINFORMATICS: e-BIOGRID

In order to improve accessibility of life science researchers to the high-end computing infrastructure and e-science expertise available in the Netherlands, a new team has recently been formed in the NBIC support programme BioAssist. Led by Timo Breit (University of Amsterdam) and

Joost Kok (Leiden University), the so-called ‘e-BioGrid’-team was set up in the framework of Big Grid (www.biggrid.nl). NBIC Partner SARA is also closely involved, with Machiel Jansen as the technical project leader of e-BioGrid and member of the BioAssist engineering team. The Big Grid infrastructure contains computer clusters and data storage, combined with specific middleware and

software to enable research that needs more than just raw computing power or data storage. The e-BioGrid team will support life science projects to use grid, cluster and cloud computing, as well as GPUs. First focal areas will be next generation sequencing, metabolomics, metagenomics, biobanking and nanoscopy.

For further information, please see: www.nbic.nl/en/support/e-science/

NBIC GROUPS HIGHLY ACTIVE IN EUROPEAN INNOVATIVE MEDICINE INITIATIVE (IMI)

The Open Pharmacological Space (OPS) project and the Drug Disease Modelling Resources (DDMoRe) consortium have recently started. Both are part of the European Innovative Medicine Initiative (IMI, www.imi.europa.eu/), a partnership between the European Commission and the European Federation of Pharmaceutical Industries and Associations (EFPIA). NBIC researchers play a central role in these two large-scale projects, which focus on modelling (DDMoRE) and interoperability (OPS).

Joost Kok of Leiden University is closely involved in the DDMoRe mathematical modelling project, which involves 9 European academic organisations and 14 companies, including major Pharma and SMEs. DDMoRe aims to develop the Drug Disease Model Library Framework as a gold standard in drug/disease modelling.

NBIC scientific director Barend Mons (Leiden University Medical Centre) is the linking person in the OPS project. OPS aims for improved in silico reasoning for knowledge discovery in Pharma research using the semantic web and focuses on the

interoperability of research data and information.

The recent introduction of the revolutionary ‘nanopublication’ concept and its role in (biomedical) science communication, a key approach in OPS, has lately been published in *Nature Genetics* (April 2011, Vol 43, nr. 4). OPS involves 22 core partners, 8 of which are major Pharma companies. The Dutch groups involved in these projects make effective use of programmes and infrastructures established within NBIC. A dedicated NBIC BioAssist taskforce on data interoperability has already been set up, in a close association with the OPS project.

BIORANGE

RECENT THESES

- Martijn van Iersel: *Data Integration with Biological Pathways*, Maastricht University, November 10, 2010
- Jeroen de Bruin: *Service-Oriented Discovery of Knowledge Foundations, Implementations and Applications*, Leiden University, November 18, 2010

- Thomas Binsl: *Fluxes of Life – Bioinformatics for Metabolic Flux Quantification in Isotopic Non-Steady-State*, VU University Amsterdam, March 11, 2011
- John van Dam: *Evolution of Ras-like GTPase signaling pathways*, University Medical Centre Utrecht, March 30, 2011
- Jeroen de Ridder: *Computational*

- approaches for dissecting cancer pathways from insertional mutagenesis data*, Delft University of Technology, January 31, 2011.
- Miaomiao Zhou: *Genome scale prediction of protein subcellular location in bacteria, with focus on extracellular and surface-associated proteins*, Radboud University Medical Centre Nijmegen, January 4, 2011

FORTRAN STATEMENT

RETROSPECTIVE

On the rapid rise of computers and the amazing prospects for this in the (life) sciences. Senior bioinformaticians retrieve memories from the bygone days so recently passed.

Gert Vriend

Professor of Bioinformatics of Macromolecular Structures Radboud University Nijmegen

“ At the beginning of the nineties we launched a website on G-protein coupled receptors (www.gpcr.org). I think it was one of the first websites in the field of bioinformatics. In those days the website was visited only about four times a day, whereas today this has increased to about four times every second. I was able to track the visits to the website via a small screen on my computer. The Internet was still in a developmental stage and it was not easy to find our website, since search engines did not yet exist. Bob, a colleague in Denmark, wanted to visit and use our website and asked my advice on how to get there. I told him to install a browser and then type in the internet address. At first he didn't know what a browser was, so I explained where he could get the software to browse the Internet. After that he was ready to surf the net. Bob also had already brought a webcam, an IndyCam from Silicon Graphics, which he had connected to his computer. However, he was not yet completely familiar with the possibilities of the IndyCam. The following Monday I noticed through my computer that Bob was visiting our website. I immediately sent him an email. Bob was clearly perplexed by my email. He asked me how on earth could I look through his IndyCam? I emailed him that I could see an awful lot more. I could even see that his secretary was not sitting on his lap. ”

BIOWISE**NBIC PhD SCHOOL BROADENING**

The NBIC PhD School, launched in 2009, is gradually being filled with advanced courses for PhD students in bioinformatics. Courses are repeated every 2 years, allowing PhD students to plan relevant training throughout their PhD period. In January 2011, the second edition of the Pattern Recognition course was held in Delft by the group of Dick de Ridder. Upcoming courses in 2011 are 'Comparative Genomics: from evolution to function', developed by Berend Snel and colleagues (June 27 - July 1, Utrecht) and 'Structural Bioinformatics' (Fall 2011, Nijmegen, Gert Vriend and co-workers). NBIC and the Netherlands Consortium for Systems Biology (NCSB) have joined forces to implement courses relevant for systems biology, including the necessary bioinformatics topics. E-learning modules will be developed as part of this collaboration.

YOUNG BIOINFORMATICIANS VISIT AGENDIA

On November 23 2011 the Regional Group Netherlands (RSG) organised its first company visit to Agendia, an innovative cancer diagnostics company in Amsterdam. RSG Netherlands – co-founded (2008) and supported by NBIC – brings together PhD students in

bioinformatics and computation biology in the Netherlands and is part of the global network of the ISCB. Agendia is the developer of the MammaPrint, a prognostic test for breast cancer patients based on gene expression profiling. The visit, in which 19 students participated, was a success. RSG Netherlands plans to organise more company visits in the near future.

TRAINING HIGH SCHOOL TEACHERS IN BIOINFORMATICS

As part of the bioinformatics@school project, NBIC organised its first training session for high school teachers on November 26 2010, together with the Freudenthal Institute for science and mathematics education. During the session the teachers were challenged to use bioinformatics in their classes. In order to prepare appropriate lessons, they were supported by NAVIGENE, an instructional guide developed by NBIC to navigate through bioinformatics tools. Besides the teachers, a number of high school students also participated in the training session. The participants, teachers as well as students, were all very enthusiastic about the session, motivating the organisers to plan a follow-up meeting in the near future.

Read more about the bioinformatics@school projects: www.nbic.nl/education/high-school-programmes/bioinformaticsschool/teacher-training/

Training high school teachers



AUTUMN SCHOOL 2010

BY LILIAN VERMEER

Computational approaches to biomolecular structure and function

At the end of 2010, an autumn school on computational approaches to biomolecular structure and function took place at the University of Duisburg-Essen, Germany and the University of Nijmegen. The two week course was the result of a joint effort by Gert Vriend (CMBI, Nijmegen) and Daniel Hoffmann (University Duisburg-Essen). The course was very intensive but highly appreciated by the fourteen participating students.

"It was a great introduction for my research work," says Christoph Wilms who had just become a PhD student in Hoffmann's group at the Department of Bioinformatics when the autumn school started. "I will focus on the prediction of the effect of mutations on the stability of proteins. To investigate this, I will use existing software or probably write my own, if the existing software does not suffice. Since this work will involve a lot of computational work, the first week in Essen was very useful for me. It was more computationally oriented, whereas during the week in the Netherlands we learned how to work with the data and how to interpret them."

GREAT JOB "Hoffmann and Vriend did a great job," says Wilms. "It was a very balanced programme. All the lectures were connected to tasks. These tasks were carried out in groups of two to

three students who presented their results at the end of the day. The autumn school was also the first activity in Barbara Zarzycka's PhD. She started in October 2010 in the 3D-structure function group of Gerry Nicolaes at the Cardiovascular Research Institute at Maastricht University. "My PhD research is focused on proteins which play a role in haemostasis, the mechanisms of the body to prevent blood loss. When something goes wrong in haemostasis it can result in bleeding problems or thrombotic disorders. I hope to find new inhibitors of proteins involved in thrombotic disorders by studying the structure of proteins purely computationally." For her the week in Nijmegen was the most interesting

since it focused on various aspects of working with molecule structures to answer biological questions. "Gert Vriend's lecture on structure validation made me aware of the pitfalls I can encounter in this process."

USEFUL Zarzycka found the first week in Essen useful as well, but she thinks that she will use that knowledge more at a later stage in her PhD. "Both weeks were very intensive, so it was good to have a weekend in between to reflect on the things learned." Zarzycka also found it a great honour that Nobel laureate, Robert Huber gave a lecture and listened to the students talk about their results.

THE ORGANISERS


Gert Vriend is professor of Bioinformatics at the Radboud University Nijmegen and Daniel Hoffmann is professor of Bioinformatics at the University Duisburg-Essen. Both are involved in the field of structural bioinformatics. Hoffmann says: "Our expertise overlaps, but is also complementary in some aspects. We focus deeply on the physical modelling of biomolecular structure, dynamics, and function, whereas Vriend's research group focuses more on the application of the tools: which tools are suitable for which biological questions and how can one extract the desired results." Both bioinformaticians realised that by combining their expertise they could provide students with a complete

overview of structural bioinformatics. A grant awarded by DAAD (German Academic Exchange Service) allowed them to develop this course. Vriend: "We decided to address excellent master students and starting PhD students. To this end we created a questionnaire for interested participants in order to select for the most excellent students." Vriend and Hoffmann are considering organising the school once every two years. They admit that it costs a great deal of effort and time to get the course running and to attract good lecturers, but both are happy with the student's achievements. Vriend: "They made a big step forward in their scientific endeavours." Hoffmann: "I feel that it was an exciting experience for the young researchers to see how good science unites people from different countries."

BY ESTHER THOLE

METAGENOMICS TACKLES GENOMES ON A
POST-INDIVIDUAL LEVEL

Unravelling microbial communities



MOVING BEYOND INDIVIDUAL GENOMES, METAGENOMICS ALLOWS THE MAPPING OF ALL GENETIC INFORMATION PRESENT IN A COMMUNITY OF MICROORGANISMS. IN THIS WAY METAGENOMICS HAS CREATED A BREAKTHROUGH IN THE STUDY OF THE SO-CALLED UNCULTURABLES – MICROORGANISMS THAT CANNOT READILY BE CULTURED AND THUS DEFY CONVENTIONAL ANALYSIS TECHNIQUES. AS SUCH, METAGENOMICS IS PERFECTLY SUITED TO TAKE ON THE LARGE POPULATIONS OF UNKNOWN MICROBES THAT INHABIT EVERY CORNER OF THE WORLD. AND THAT INCLUDES OUR OWN BODIES. THE COMPLEXITY OF SUCH STUDIES IS DAUNTING AND PUSHES BIOINFORMATICIANS TO THE LIMIT. THE NEWLY ESTABLISHED NBIC METAGENOMICS TASKFORCE HELPS THEM TO MEET THE CHALLENGES AHEAD.

When it comes to complexity, metagenomics projects will be hard to beat. Trying to unravel the composition of highly diverse microbial communities, map all the present genetic information and mould that into new biological insights is not for the faint-hearted. But it may shed light on why the world-famous Gouda cheese tastes so good.

Wandering through the gastronomical landscape, it will be a real challenge to find a lunch menu that is as easy to prepare as the quintessential Dutch dish 'broodje kaas'. A roll of bread, some butter and a slice of Gouda cheese: that's it. Although it baffles many non-Dutch that this is actually called 'lunch', the dish itself harbours no mysteries. Or so it seems. Because who would have thought that an everyday product like cheese could be the topic of advanced research? As it turns out, the process of making cheese has been known to mankind for a long time, but exactly what happens during the fermentation process remains unknown. "Many artisanal cheeses, including Gouda, are prepared using so-called undefined starter cultures, mixtures of lactic acid bacteria of which the composition is not known in detail," explains Eddy Smid (Wageningen University / Top Institute Food and Nutrition). In other words, it is unclear which bacterial strains are present in the starter culture, why they are present and what each strain contributes to the reactions by which milk is turned into cheese. This gap in knowledge is not due to a lack of research interest, but of technical limitations, says Smid. "Until the 1980s, a lot of work was done on unravelling the composition of these bacterial mixtures, but by then the technical possibilities had been stretched as far as they could." One of the major problems in the study of microbial consortia, whether it concerns relatively simple ones like those in cheese or the extremely complex consortia found in for example soil, sea water or the human gut, is that the traditional cultivation-based methods are unable to catch the entire span of biological diversity present in such samples. Smid: "With the ascent of genomics and especially metagenomics, the field has gained momentum again. Metagenomics allows us to map the complexity of the consortium and to explain the need for and origins of this complexity."

MULTICELLULAR ORGANISM Metagenomics refers to the study of genetic information on the level of communities instead of focusing on the genome of an individual organism. In short: you take a sample of whatever microbial habitat you want to study (starter cultures, soil, waste water, faeces – to name a few) and sequence all DNA present. "In metagenomics, we view the consortium as a multicellular compartmentalised organism," says Sacha van Hijum (NIZO food research / Radboud University Nijmegen Medical Centre), who heads the new NBIC Metagenomics Taskforce. "It is about mapping all the genes in a certain environment. The metagenome encodes a network of functionalities. Until now, a lot of the ongoing research focused on model

systems, but the time has come to include the microbial diversity that characterises real biological systems." And that diversity seems unlimited. Van Hijum: "The 2,000 to 4,000 genes that make up the genome of an individual bacterial strain are in fact only a small selection out of the genetic repertoire of 10,000 to – in some cases – 20,000 genes present in all strains of a given species. Studying one strain provides you with a genetic snapshot of the total genetic diversity of a species. Furthermore, next to the various strains, a bacterial consortium also contains phages and plasmids, which further contribute to the genetic diversity of the metagenome. We need to move up from the strain to the level of all species present in a consortium in order to get the complete picture and be able to make predictions on the behaviour, functionalities and engineering possibilities of bacterial consortia."

With the latter, Van Hijum touches upon on the application of the knowledge unearthed through metagenomics since apart from satisfying the scientist's innate curiosity, why is learning about the functioning of microbial consortia relevant? "First of all, in virtually all ecosystems, microorganisms do not operate on their own, but fulfil all kinds of functions as part of a consortium. For industry, consortia are relevant because they are more robust and more versatile than individual strains." For the food industry, fermentation is an essential production process. "This type of research can help to make the artisanal methods of producing cheese applicable on an industrial scale," says Smid. "Being able to predict how a starter culture will operate should enable the design of mixed starters that produce high quality cheeses with, for example, reduced salt or fat levels. These are important consumer demands for the food industry to address."

LEAN VERSUS OBESE Understanding the workings of a microbial consortium is also relevant in order to gain more insight into different kinds of fermentation processes, i.e. the ones that take place in our own body. Michiel Kleerebezem (NIZO food research / Wageningen University) works on gut metagenomics – mapping the genetic network of the substantial and highly diverse microbial communities that inhabit our intestines. "The underlying question in metagenomics research is what the causal relations are between the microbial communities in the gut and a number of diseases such as obesity." In 2010, Qin et al. published a 'catalogue' of the microbial genes in the gut in *Nature*. Their work showed that humans share a major part of bacterial genes in the gut, but that the gut metagenome also encompasses sets of genes that are not commonly found in all individuals. "Studies performed in the USA suggest that there is a causal relation between gut bacteria and obesity and that the ratio between the two major phyla *Bacteroidetes* and *Firmicutes* is a key factor," says Kleerebezem. In this field, the primary focus is on the colon, which is one of the reasons why Kleerebezem and colleagues decided to focus on the small intestine. "The small intestine is where the readily available nutrients are absorbed by the host mucosa. It is therefore likely that

changes in the diet will have a large impact on the bacterial communities residing here,” he explains.

They are working on a metagenome catalogue of the small intestine, and so far it has become clear that this part of the intestine harbours a community of lower diversity compared to the colon. On the other hand, the small intestine microbiota is more dynamic and is dominated by other microbes. “We have a fair view of the species and the genes. Right now, we are figuring out the functionalities using metatranscriptomics. The DNA level is ‘only’ the functional blueprint of an ecosystem; in the end you want to know the functional behaviour of a consortium of species. Through such approaches one may be able to identify the differences in functionalities of the microbiota in lean and obese individuals, which may fuel the rational design of interventions.” That this type of data-rich research cannot take place without advanced bioinformatics tools seems obvious. Kleerebezem: “Our metatranscriptomics studies have yielded 50 million sequence tags of genes that are expressed. What we need are ways to extract the relevant biological information in a condensed format. For example: how many glucose importers do we see? We need the collaboration with bioinformaticians to get the biology on the table.” Eddy Smid agrees: “As biologists, we want to connect genetic information to phenotypic diversity. That is the interesting part for us.”

YES OR NO? To a certain extent, metagenomics projects require similar bioinformatics solutions, but there are clear differences as well, emphasises Sacha van Hijum. “Different projects require different tools. The cheese project requires an analysis pipeline that allows very detailed mapping of the differences between the strains down to the level of SNPs and indels [single nucleotide



NBIC METAGENOMICS TASKFORCE

“With the Metagenomics Taskforce we aim to bring together bioinformaticians who are involved in various metagenomics projects so that we can combine scattered expertise,” says principal investigator Sacha van Hijum. “Next, we will analyse what the major bioinformatics-related questions are, identify the generic aspects in those questions and then get to work on tackling the most pressing bottlenecks.” The taskforce is not primarily focused on solving urgent, practical problems. “There is a clear research component in our activities. We aim for long-term solutions and will also establish the links to grid and cloud computing activities within NBIC.” Several research groups have already joined the effort and the taskforce’s kick-off meeting is planned for mid 2011. Researchers interested in participating can contact Sacha van Hijum at: sacha.vanhijum@nizo.nl

insertions/deletions, ed.], but such a pipeline is not suitable to analyse the much larger and much more diverse gut metagenome. However, there will be generic aspects and those can be addressed by the Metagenomics Taskforce.” One of van Hijum’s personal interests leans towards using Boolean decision models to distil functional interactions from the huge metagenomics datasets. “It boils down to finding the simplest way to describe a system. Trying to bring choices down to ‘yes’ or ‘no’. You aim for reasoning like ‘if we see A and we see B, but not C, then this sample belongs to group X’. This approach helps you to make these complex data manageable and to create a simplified yet comprehensive outline. It also identifies the questions that you can answer right away as well as pinpoints major gaps in biological knowledge.” Another problem he will sink his teeth into concerns classification techniques, which provide focus in the analysis process and also are used to validate assumptions used for mining the data. “I never just assume that the data, sample descriptions and hypotheses used for data mining are correct. You need to be sure that what you interpret is not due to experimental error, but really says something about the biology you are studying.” So besides the sheer size of the datasets, the various levels of complexity that need to be unravelled and the different approaches that are needed to translate everything into biological information, there is the challenge of ensuring you can actually draw conclusions from the data you generated. Instead of being overwhelmed, van Hijum appears to be inspired by it all. “The complexity of metagenomics pushes you to give it all you got. You really need to pull out all the stops and that’s what I like about this field.”

ACKNOWLEDGEMENTS

- Sacha van Hijum, Senior Scientist NIZO food research; Group leader Bacterial Genomics (CMBI, NCMLS, Radboud University Nijmegen Medical Centre); Bioinformatics coordinator, TI Food and Nutrition; Principal Investigator Metagenomics Taskforce, NBIC.
- Michiel Kleerebezem, Principal Scientist NIZO food research; Professor of Bacterial Metagenomics, Wageningen University.
- Eddy Smid, Associate Professor of Food Fermentation, Wageningen University, Group leader, TI Food and Nutrition.

R-CONSULTANCY

BY LILIAN VERMEER

Consulting services in bioinformatics

To analyse microarray and high speed sequencing data, researchers increasingly use the so-called Bioconductor package. However, using Bioconductor and R, the statistical programming language for the package, is not an easy task. Philip de Groot, an expert in R and Bioconductor, noticed that many researchers require help with these and other bioinformatics analyses. He started his own business to fill this need.

“When I started working for the Amsterdam Medical Centre, and later on Wageningen University after finishing my PhD, I had not yet come up with the idea of starting a business,” admits Philip de Groot, CEO of R-Consultancy. “But when I won the NCI Venture Challenge 2007 along with two other contestants, the idea of setting up a business began to take shape.” De Groot won the Venture Challenge for aspiring life sciences entrepreneurs with the idea for MADMAX, a management and analysis database for multiplatform microarray experiments. He used the award money to conduct a market analysis to investigate exploiting MADMAX, including providing R and Bioconductor expertise.

VIABLE COMPANY R is a free software environment for statistical computing and graphics. The software is popular in the life sciences, e.g. for microarray analysis and high speed sequencing analysis utilising the Bioconductor package. Bioconductor provides tools for the analysis and comprehension of

high-throughput genomic data. In science, the software is used to gain a better understanding regarding the biology that underlies diseases like cancer and diabetes, as well as to increase the understanding of the interaction between nutrition and health. “It appeared that serving only academic customers would not create a viable company.” Nevertheless, in January 2009 De Groot started his company, R-Consultancy, in addition to continuing his four-day per week job as a bioinformatician at Wageningen University. De Groot has noticed that experts with a thorough knowledge of R and Bioconductor are still scarce and that the market for his business is picking up. “Although most of my customers still are academics, commercial businesses are increasingly calling for my services. My first client was a large Japanese bioinformatics company. They designed Simbiot – a system for management and analysis of biological data. Within this system I incorporated the SNP-microarray analysis application.” For another client De Groot changed the existing script for a specific bioinformatics programme in such a way that the running time was much shorter.


“Since I have already worked for ten years with R and many of the applications in which it is used, I am usually quite fast in detecting the bottleneck in a programme.”

EXPERTS POOL In some assignments De Groot works together with other experts in (bio)informatics. “Some problems are too large and complicated to be solved by one person because of the different expertise required. For example, to arrange restricted access to selected users in a system with many different bioinformatics applications, I asked for the assistance of an ICT specialist in this area.” De Groot is also part of the Experts Pool of the Delft based engineering/software developing company S&T Corp (www.stcorp.nl/step/bioinformatics), which set up this pool to fill the increasing need for unique and/or rare technical expertise. “NBIC is also thinking about setting up an experts pool,” says De Groot. “Increasingly, new applications in the life sciences are being developed and scientists usually start using these applications very quickly. The data analysis is frequently not ready or is still user-unfriendly. That’s where I can help and I hope that people find their way to my business ever more often.”

R-Consultancy Bioinformatic Services aims to help researchers, pharmaceutical companies, and life-sciences companies in solving R and Bioconductor related problems. This can range from answering a simple R-related question to performing a full next-generation sequencing analysis, performing all sorts of (large scale) R-calculations, to writing (and maintaining) e.g. a dedicated R-library for maximising research efforts.

More information: www.r-consultancy.com
Philip de Groot, CEO
info@r-consultancy.com





INTERVIEW WITH JEROEN RAES
BY MARGA VAN ZUNDERT

BIOINFORMATICIAN JEROEN RAES DISCOVERS
THE SECRETS OF HUMAN FLORA

“We are not purely
human: we live in
symbiosis”



Metagenomics is in its discovery phase. The composition and interactions of complex microbial ecosystems such as our intestinal flora are revealed today by bioinformaticians, thanks to high-throughput sequencing data. “Cool science”, according to Professor Jeroen Raes at the University of Brussels. He enjoys fishing in the unknown. His group is identifying many new bacteria. Ninety percent of all bacteria cannot be cultured and therefore went undetected and unnamed. The work provides better insight into ourselves because man is actually a walking bacterial colony. Our human cells, genes and nuclear bases are outnumbered by those of the microorganisms we carry along. Are we human?

You became a professor rather young.

How did you manage that?

In 2005 when I started working in metagenomics, the field was very new and hot, a true niche. Since then, the field has developed very quickly. I swiftly acquired unique, much sought-after expertise. I consider myself lucky because there are many good, hard-working postdocs around but only a few openings as a group leader or professor.

You have two master degrees: do you feel you are more a biochemist or a bioinformatician?

Actually, I'm a chemist. I debated with myself for a long time whether to study chemistry or biology. After receiving my bachelors in chemistry, I started a masters in biochemistry. Soon after I started a PhD in bioinformatics; the university started offering an MSc in bioinformatics, which I followed too. In my field you need a lot of disciplines: I need to be a biologist, a biochemist, ecologist, a computer scientist, mathematician, statistician, immunologist, ... All those disciplines are important.

“I don't think we would survive without our bacterial friends”

What inspired you to specialise in metagenomics?

Just before I had a job interview with Peer Bork for a postdoc position at EMBL, I attended a lecture by Greg Venter. He talked about his expedition over the world's oceans to collect samples for finding unknown species by metagenomic sequencing. I thought that was real cool science, true discovery. When Peer asked me if I would be interested in working on metagenomics, I immediately said yes. Metagenomics started in oceanography and in these first years we focused on oceanic ecosystems because that data was available. But soon other scientific disciplines got interested. Today much of my work concerns the human intestinal

flora. In bioinformatics it is often difficult to explain your work to the grocery man or your neighbour, because it is often too technical. But intestinal flora is 'ordinary'. Everyone knows he has intestinal flora and that it is important that it functions well. When I talk about how we are figuring out what microorganisms our flora contains, how they live together and influence our health, people are keenly interested. I enjoy that very much.

Which microbiomes do you study?

Besides the human microbial flora of the gut, we have projects on skin flora and microbiomes in the throat and upper airways. We are also a partner in the Human Microbiome project in which samples are taken from 120 people from various body sites: mouth, throat, airways, small and large intestines, skin and, for women, the vagina as well. And we are also interested in the flora of model organisms such as mice.

“We need piles of 10 to 20 hard disks to exchange data-sets with other groups”

How great is the influence of microorganisms on a human?

We don't know, and that is the biggest challenge in this field. Our immune system and bacterial flora have a very intimate relationship. Our immune system doesn't fight the bacterial flora off as it does with pathogenic bacteria. Why? I think our microbiome keeps out pathogens by ecological competition. Moreover, intestinal bacteria provide essential nutrients and vitamins. I don't think we would survive without our bacterial friends. We live in symbiosis: we are not purely human. It is a cliché in our field, but for every human cell there are ten bacterial cells in your body. We are completely outnumbered. We are a walking bacterial colony. That is a fact. Are we human? I don't know. It might be science fiction, but I can imagine that intestinal bacteria also provide signals to man, their hosts, for example to stimulate eating.

What news does your work give us about our microbiome?

Last year we participated in a study that resulted in the publication of a catalogue of the human intestinal flora in *Nature*. The article provides an overview of the amount and types of species present. We also identified a minimum parts list for the intestinal flora in the form of 6000 genes. Of course this pool contains bacterial housekeeping genes, but also all genes that are necessary for the interaction within the ecosystem and for the interaction with the host. We are currently developing the tools to unravel how the interaction in the ecosystem works. We study which microorganisms work together and which are in competition. We are building the first interaction

networks, not on a protein level, but at the individual level. In a forthcoming paper we will report that the intestinal microbiomes we find group around in three clusters, which we call enterotypes. They differ widely in type and mix of species. We haven't yet figured out why. It's fascinating that there appears to be no link to nationality, age, sex or health. Are they representatives of various states during the day or seasons? Or do the enterotypes link with the genetic makeup of the host? People are primarily interested in the characteristics of disturbed flora, but even the healthy state has a number of riddles.

How's your own flora?

I don't know. But I'm becoming very curious about what enterotype I am myself, thus I will probably investigate that soon.

“All bioinformatics software needs to be rethought to keep up with high-throughput sequencing”

What technological breakthrough are you eagerly awaiting?

At this moment, we have great difficulties in keeping pace with the continuing breakthroughs in high-throughput sequencing. With every new generation of sequencers, the amount of data increases by two or three orders of magnitude, while the costs are dropping fast. We are happy to be working together with the Beijing Genomics Institute (BGI) on the European metagenomics project called megaHIT. At one point, BGI alone had as much sequencing capacity as all American institutes together; China is investing enormously in science. The Chinese sequencing power has provided Europe a worldwide jump in metagenomics. Sequencing technology, however, develops much faster than chip technology, which leads to data storage and analysis problems. We now need to post piles of 10 to 20 hard disks to exchange datasets with other groups. Also bioinformatics itself is becoming a bottle neck. BLAST, for example, is starting to become too slow. All bioinformatics software needs to be rethought or reinvented to keep up with high-throughput sequencing. But a technology breakthrough that would be spectacular is high-throughput single cell sequencing. The illumina technology we now use cuts the genomes into pieces that need to be puzzled together again after sequencing. A nice job for bioinformaticians, but multiple single cell sequencing would provide the complete metagenome at once.

You are going on expedition yourself, I've heard.

Yes, I hope to board the Tara sometime next year. Tara is a 36-meter sailing ship making a three year

expedition over the world's oceans to study the marine microbial ecosystems. It's a multidisciplinary initiative involving marine scientists, climate scientists, microbiologist, microscopists, bioinformaticians, etc. Researchers aboard take samples and determine all kinds of parameters at the sampling sites such as temperature, salinity, nutrient density and oxygen levels. The samples are stored and shipped to EMBL. In Brussels we analyse some of the metagenomic data that result from sequencing the samples. It will be pleasant to do some actual lab work for a change; we bioinformaticians usually sit behind our computers. Just as with our intestinal flora, we have no idea about plankton's biodiversity and ecology. The oceans are the most important lung of our planet and we know nothing about it!

JEROEN RAES (1976)



- 2010 Publication of the human gut microbial gene catalogue in Nature
- 2009 Professor of Bioinformatics and (eco)systems biology University of Brussels (VUB) and Group leader of Flanders life science research institute VIB
- 2005 Postdoc, later scientist, at EMBL Heidelberg
- 2004 Postdoc at CropDesign
- 2003 PhD in bioinformatics, University of Ghent, Belgium
- 2001 MSc Bioinformatics
- 1998 MSc Biochemistry



Seven questions for Jan Bot

BY BASTIENNE WENTZEL

Name: Jan Bot
Date of Birth: 11th of March, 1983
Place of Birth: Rotterdam, The Netherlands
Nationality: Dutch
Study: Media and Knowledge Engineering, Delft University of Technology
Career: Scientific programmer, Delft University of Technology
Hobbies: Squash, cooking for and with friends

Who is Jan Bot?

I was born and raised in Rotterdam. After high school (VWO) I studied Media & Knowledge Engineering (MKE) at Delft University of Technology. After obtaining my Master's degree I was offered a position as scientific programmer by Marcel Reinders, who at that time was setting up a bioinformatics group. I was and still am attracted to bioinformatics because I feel I am creating software with a purpose. The challenge for me is operating between the areas of biology and informatics.

What is your job as a scientific programmer at Delft University?

As a scientific programmer I have the freedom to follow my own path to solve a given problem. Creating a design and carrying out the implementation is done in collaboration with the colleagues who will eventually be using the tools. PiCaS is an example of such a tool. I developed it for Jeroen de Ridder, a colleague in my group, who wanted to run his algorithm on the Life Science Grid (LSG) and needed a framework to manage his output. The LSG is a high performance computing network setup by SARA and Big Grid, which consists of about 4,000 computer cores.

What do you enjoy about working with the Grid?

To me, the Grid is a big playground with a wide variety of hardware to play with. It is very rewarding to get something to work. For example, Jeroen wanted to investigate relationships between insertions in mouse tumours and gene expression. To do this, he developed an algorithm that required 15 million small computing jobs. The Grid is not suited for that many small jobs. To solve this, we built our own framework which is now up and running and which reduced the computing time of no less than a century to just three weeks.

What else do you work on?

I also work on Cytoscape, a programme to visualise biological networks. We added the module CytoscapeRPC with which you can write scripts in any programming language to adapt networks in Cytoscape, instead of doing this by hand for every new dataset. That makes

the job efficient and reproducible. The fun part of this is that it has led to interesting collaborations, and it is inspiring to see that my work is used around the world.

"It is very rewarding to get something to work"

How does the NBIC BioAssist programme work for you?

BioAssist is a natural way to work with many different people from proteomics to systems biology. For the exchange of ideas and expertise this actually works fine, but it is challenging to have these people work together because everyone has their own speciality and is solving their own set of problems. For me the greatest benefit lies in exchanging knowledge about tools and getting a monthly update on what people are working on.

How do people find you for cooperations?

Within BioAssist people know what I do and what my specialities are. I make an effort to advertise CytoscapeRPC and the grid work by writing tutorials and presenting this work at conferences. Also word-of-mouth marketing is very effective. Additionally, I publish application notes such as for CytoscapeRPC, which has just been submitted to Bioinformatics.

Who do you work with?

For the grid work I collaborate heavily with the people at SARA, without whom this project would never have succeeded. I also try to get people to use the software by visiting them and showing them how it works. I have done such a project with Victor de Jager who works at the CMBI. CytoscapeRPC has led to a number of interesting contacts, the most prominent one being Paul Shannon. He works for the ISB and has implemented RCytoscape, a Bioconductor module which relies on CytoscapeRPC to communicate with Cytoscape. His bug reports and feature requests have led to a more useful plug-in.

PINDEL

BY BASTIENNE WENTZEL

Fast algorithm detects variants in small fragments



Next generation sequencing has delivered very high numbers of reads of small fragments of about 100 bp (base pairs), resulting in high coverage and high accuracy. “But short fragment length is a problem for identifying large sequence variations”, says Kai Ye. He developed efficient and versatile software to detect simple insertions and deletions as well as complex structural variants. Keiran Raine used the application to successfully identify a genuine tumour suppressor gene in renal cancer.

It is his mission to detect variations other than SNPs in next generation sequencing data, says Kai Ye from the Leiden University Medical Centre. One of the problems is that these variations, such as insertions or deletions, may be large. The break point of a fragment after sequencing may span the deletion. Mapping these deletions to the reference genome can then be done with only one of the ends of the fragment, while usually both ends (paired-ends) are mapped. Existing software for the analysis of next generation sequencing data suffers from poor resolution. Ye therefore developed Pindel[1],

a data analysis programme which can detect insertions and deletions (indels) and other complex structural variants at single base resolution from next generation sequencing data. The speciality of Pindel is the re-alignment of a sequence based on an anchor position to find larger or more complex events that standard mapping tools are unable to handle, according to user Keiran Raine from the Sanger Institute. “The applied algorithm in Pindel is in general slower than those used for initial mapping of whole genome datasets. The trade-off is a gain in accuracy. By pre-selecting reads that we would expect to map,

we give Pindel a hint in the form of an anchor position to start searching for an appropriate mapping, which allows it to still have a relatively swift throughput.” This saves time and CPU. Furthermore, Pindel allows the use of multiple samples and tags the source of the data. This makes it simple to filter out likely germ line mutations (inherited variations present in all tissue) and sequencing artefacts as they appear in both samples, and allows the identification of somatic variations only. Looking at somatic changes in the genome (that have been acquired during life) in cancer patients can elucidate how a cancer arises.

DEVELOPER Kai Ye developed the programme Pindel during his postdoctoral study at the European Bioinformatics Institute EBI near Cambridge in the UK. Pindel uses a so-called pattern growth algorithm to identify breakpoints of large deletions and medium sized insertions from paired-end short reads. A small portion of the reads may have only one end mapped to the reference genome, but these unmapped reads contain information about the precise break points of the deletion event. “If we can find a proper position to split the read into two fragments, which can be mapped back to the reference separately, we will be able to compute the exact positions of the break points and thus the fragment deleted compared to the reference. If we collect multiple reads that support the same incidence, we will be more confident about the deletion

event in the test sample,” explains Ye. “Using this new idea we can now detect variations of any size.”

“Using Pindel we can detect any type of structural variants of any size”

Detecting deletions is relatively easy. Insertions, on the other hand are complicated. Longer insertions may be spread out over various fragments, making it difficult to determine the fragments. Therefore the precise break points must be computed and the fragment inserted in the medium sized range of 20 bases for 36 bp reads. The developers showed that 80

percent of deletions and insertions (10 bp to 100 kb or 1-16 bases respectively) were correctly detected from 36 bp paired end short reads. Ye is the first to develop software that can achieve this, although others are now developing similar programmes he says. Additionally, he achieved a significant improvement in speed of about 100-fold. “CPU time is valuable. Also, when we approach the 1000 dollar genome in about five years time, analysing genome variants may become a standard test requiring the data of thousands of patients to be analysed. We are in the process of improving the performance of Pindel to achieve the required efficiency.” Ye is now developing methods to identify other variants such as inversions, duplications, large insertions, inter-chromosome variants and to analyse RNA sequence data.

USER “When I first started working with Pindel it was a relatively memory hungry piece of software,” recalls Keiran Raine. The software was originally designed for 36 bp sequencing reads and required exact matching on either side of a non-reference event. “Now Pindel has a much improved memory footprint, allows mismatches around non-reference events and has had several new types of event detection added along with parallel processing capabilities,” Raine says. The senior computer biologist works for the Cancer Genome Project at the Wellcome Trust Sanger Institute. He recently published a Letter in Nature^[2] which describes mutations in a newly found renal cancer gene. Using Pindel, the researchers specifically looked for somatically acquired mutations. Pindel was essential for finding the inactivating frameshift mutations in

the cancer gene PBRM1. Raine explains: “Without this information, it would have been virtually impossible to identify this gene as a genuine tumour suppressor gene. At the time we were building our pipeline, there were no tools we were aware of that allowed data from multiple samples to be analysed at the same time while retaining source information.

“Without Pindel, it would have been impossible to identify PBRM1 as a genuine tumour suppressor gene”

The only other tool available at the time for identifying indels was SAMtools, which was specifically

designed for analysing germline genomes, and therefore presented some difficulties for modifying for detecting somatic indels. We also felt that Pindel was more sensitive at that stage.” The tool is easy to use for anyone with some experience of Unix/Linux and next generation sequencing data. “Generating the input format from BAM data with the appropriate records was probably the most complex part of using Pindel,” says the user. The programme is reliable and relatively fast to run with a good output format. Raine has helped development not only in reporting bugs but also with code. “I created an early pre-processing perl script and module to generate relevant inputs from BWA aligned BAM files. Also some of the testing was done using known cancer variants to identify problem areas where we provided Kai with small datasets to test changes,” Raine says.

REFERENCES

1. Ye, K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21), 2865–2871. (<http://www.ncbi.nlm.nih.gov/pubmed/19561018>)

2. Varela I. *et al.* (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469, 539–542.

NOTES

¹ Cooperation with NBIC
Kai Ye and NBIC cooperate on the software

engineering part of Pindel. Focus is on re-factoring and testing the code, extending the documentation and enhancing support for work flow software.

² Open source software
<http://sourceforge.net/projects/pindel/>

BY ASTRID VAN DE GRAAF

Back tracing evolution of the Ras signal transduction pathway

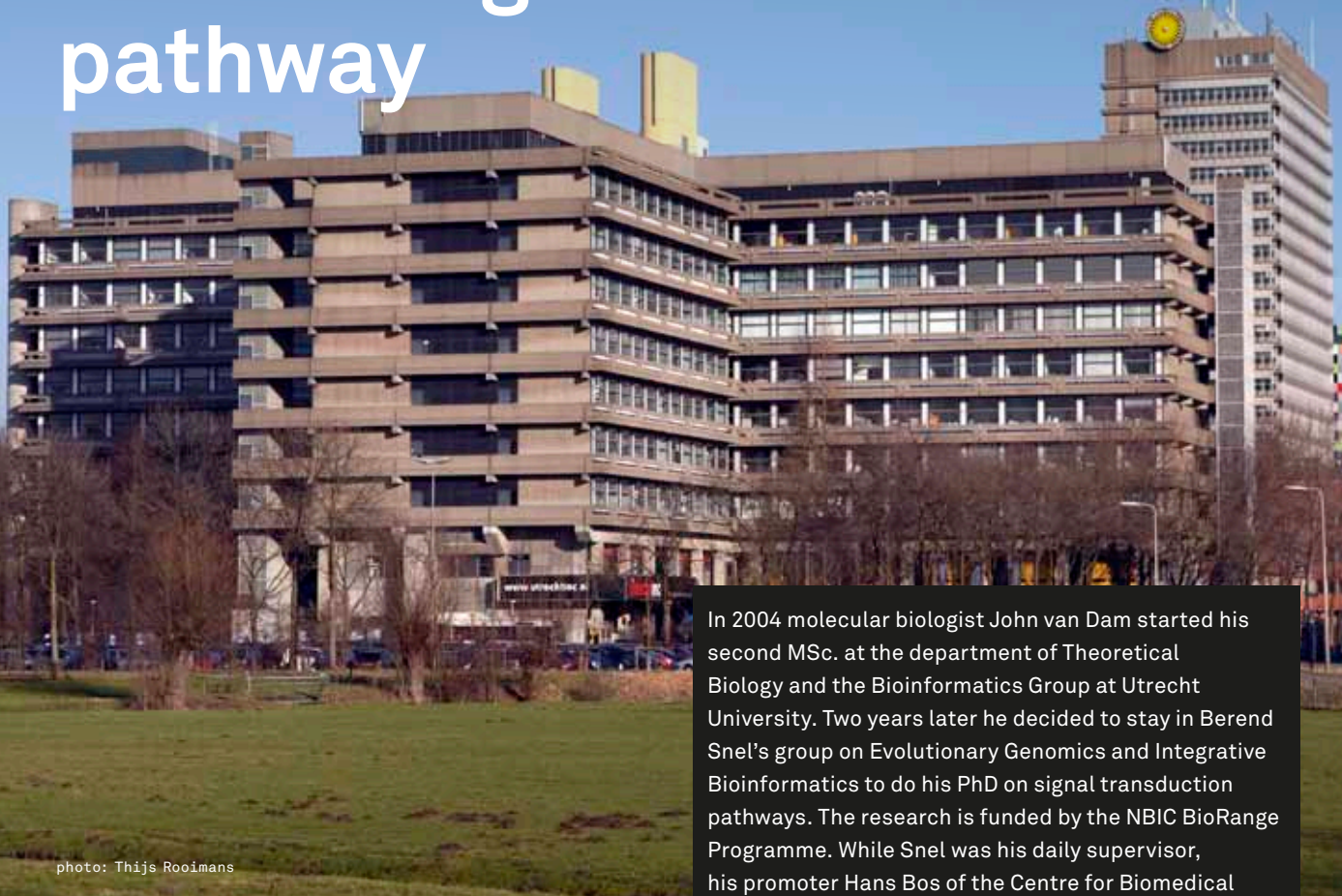


photo: Thijs Rooimans

THE RAS SIGNALLING PATHWAYS REGULATE A WIDE VARIETY OF CELLULAR PROCESSES SUCH AS CELL PROLIFERATION, CELL DIVISION AND CELL DIFFERENTIATION. JOHN VAN DAM ATTEMPTS TO UNRAVEL THE EVOLUTION OF THESE PATHWAYS USING COMPARATIVE GENOMICS. BESIDES NEW EVOLUTIONARY INSIGHTS, HIS RESEARCH MAY LEAD TO A BETTER UNDERSTANDING OF HOW THESE PATHWAYS OPERATE.

Signal transduction pathways are crucial for cells. Complex networks of interacting proteins regulate all cellular processes and depend on internal and external stimuli. Since the pathways also play a role in many diseases, thorough molecular research has been done to uncover how they operate. “We chiefly know *how* a pathway is connected. If we also knew *why* it is connected as it is, then we would be able to fully understand the mechanisms and use it for therapy development,” explains John van Dam. “So by investigating how it has changed over time

In 2004 molecular biologist John van Dam started his second MSc. at the department of Theoretical Biology and the Bioinformatics Group at Utrecht University. Two years later he decided to stay in Berend Snel’s group on Evolutionary Genomics and Integrative Bioinformatics to do his PhD on signal transduction pathways. The research is funded by the NBIC BioRange Programme. While Snel was his daily supervisor, his promoter Hans Bos of the Centre for Biomedical Genetics was at the University Medical Centre (UMC) in Utrecht. There, signal transduction is one of the research themes of the group Molecular Biology of Cancer. The interaction with the UMC was very valuable for him in terms of being able to check the relevancy of this work as a bioinformatician. During his stay the Theoretical Biology group expanded in size. The group tackled many different evolutionary subjects with bioinformatics such as population dynamics, HIV, host-pathogen interactions and plant development studies. <http://bioinformatics.bio.uu.nl>

and how it was connected before, we try to answer that question.”

Van Dam has focused on the evolution of the Ras signalling pathway in eukaryotes and the role of its key proteins, Ras GTPases. GTPases behave like switches which can be on or off depending on whether they bind guanosine triphosphate (GTP) or guanosine diphosphate (GDP). Ras GTPases also have oncogenic properties. In over 15 percent of all kinds of cancer Ras GTPases are mutated. “What we wanted to know was how these proteins had

evolved, why there are so many Ras-like GTPases and how these proteins relate to each other.” His research project was part of a collaboration between the Physiological Chemistry Department at the University Medical Centre, where his promoter Hans Bos is situated, and Utrecht University’s Bioinformatics group. For his daily research, Van Dam worked at the department of Theoretical Biology and Bioinformatics under the supervision of Berend Snel.

COMPARING DATA Van Dam used a comparative genomics approach to investigate evolution of a signalling pathway. “The way to do that is to look at specific protein families within the signalling network, how well they are conserved and how they evolved. At the end you can make an evolutionary reconstruction of the entire signalling pathway,” says Van Dam.

One of the protein families Van Dam studied was the RasGEF domain. Guanine nucleotide exchange factors (GEFs) replace the GDP with GTP and thereby activate Ras. “A GEF is regulated by many processes and is therefore an interesting drug target. Furthermore, GEFs are ideal for investigating evolutionary dynamics and the role of protein domain acquisitions, since GEF proteins have a common ancestor but many different protein domains with specific functionalities. We compared human proteins with many sequences of different species such as protists, fungi and animals. We discovered that animal and fungal Ras signalling are quite divergent but that there are also more similarities than we expected.” Van Dam continues: “The combination of bioinformatics and evolution is mainly about comparing sequences. To do that we use the best tools available, but these tools are not always directly suited for our purpose. So you first have to adapt some programmes you work with. We are expert users, Snel always tells us.” Not that Van Dam is developing software himself, but he is adding different applications together to research his data. “We use the programmes to couple data and develop our own databases. But it takes a while to learn how to use them optimally,” Van Dam relates from experience. In using sequence data from the public domain, he encounters other difficulties. “It is very important to know which genome sequences and which identifiers were used in the analysis. If people don’t mention which version they used, it is difficult to compare results or use the data.”

TWO WORLDS One of the most important aspects while doing his research was having a sparring partner in the biological research field. That’s why Van Dam attended the staff meetings at the UMC every week. “You can read a lot of literature, but to know what is really going on in that field, you need an expert. I had a lot of support from Hans Bos and inspiring discussions about the results. When I found patterns of relations, I wanted to know if that meant something. The interaction with UMC was very valuable for that. It was like living in two worlds.”

Van Dam discovered that there was another world at the end of his MSc. in Molecular Biology in Leiden. “I met Paulien Hogeweg during a day at the Utrecht

University’s Theoretical Biology group. We clicked right away and everything fell into place.” Fully convinced that bioinformatics is necessary to research complex and dynamic cellular processes, Van Dam did his second MSc. on Theoretical Biology & Bioinformatics in Utrecht. “At that time, I realised that all the analysis with microarrays and sequencing are generating enormous amounts of data. But what to do with all these data? Molecular biologists have to deal with that.” He has never regretted his choice, although he did have a hard time finishing his thesis, as last October he started his postdoc quite early. “I was not near the final date of my PhD, but when I saw the job description hanging on the bulletin board, I thought: this is me. This is what I want to do.” So Van Dam sent in his application and ended up the favourite.

COMPLETELY NEW Currently he is working as a postdoc at the Centre for Molecular and Biomolecular Informatics at UMC St. Radboud in Nijmegen on a completely new subject: the evolution of cilia and flagella. “We know these structures chiefly from the hairy whiptails on cellular organism or sperm, but we have them everywhere, for instance, in our lungs to transport mucus upwards or the non-motile cilia as photo sensitive regions on retina cells to transfer light signals. A simple gene mutation can affect the function. We are going to look at the evolutionary part of it with comparative genomics. This time I will stay in touch with the human genetics group in Nijmegen to get the same interaction as I had in Utrecht.”



NAME: John van Dam
UNIVERSITY: Utrecht University
PROMOTOR: Prof. J.L. Bos (UMC Utrecht)
and B. Snel PhD (Utrecht University)
THESIS TITLE: Evolution of Ras-like signalling pathways

Promotion : 30 May 2011

MORE INFORMATION

1. Van Dam, T.J.P. *et al.* (2011) Evolution of the Ras-like small GTPases and their regulators. *SmallGTPases* 2(1).
2. Van Dam, T.J.P. *et al.* (2009) Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi. *Cellular Signalling* 21 (11) 1579-1585.
3. Van Dam, T.J.P. and Snel, B. (2008) Protein Complex Evolution Does Not Involve Extensive Network Rewiring. *PLoS Comput Biol* 4(7): e1000132. DOI:10.1371/journal.pcbi.1000132

How genomes as information structures impact human identity

The widely held view that genomic processes are information processes is a historically grown view. Critics of this view in science studies argue that seeing genomes as information structures maintains genetic determinist views of cellular organisation and inheritance, while ignoring other aspects like cell metabolism and the organism's environment. In my research I look both *if* and *how* this idea still plays a role in genomics.

The Centre for Society and Genomics (CSG) is a national research organisation much like NBIC. As the name already suggests, our research is aimed at how genomics influences and interacts with society. While implementations of genomics research results currently only affect a limited number of people, developments and ideas within genomics have greatly impacted society, not least in aspects of human identity. By human identity I mean collective as well as individual identity, such as ideas of what makes humans different from other species, what it means to be ill or healthy, why we behave the way we do and how we relate to other humans. By facilitating storage and analysis of large amounts of data, bioinformatics has a huge influence on how knowledge is gathered and shaped in genomics and thus on the changes in how we view ourselves as humans. But this influence is always indirect,

always through other genomics fields that bioinformatics cooperates with. The project 'Bioinformation and human identity' that I am working on targets analysing how bioinformatics, in its processing of genomics information, contributes to shaping identity. The project is a collaboration between CSG and NBIC.

GENETIC THINKING Genes and genomes are widely seen as containing information. This has not always been the case. Before the 1950s genes were mainly seen as having (biochemical) specificity, that is, a specific molecular arrangement, which triggers a specific reaction. But by the 1990s the idea of genomes as information had taken hold completely. Philosopher Donna Haraway observed that, "[F]rom the point of view of the 1990s, the genome is an information structure that can exist in various physical media." As media she mentioned the DNA sequences in natural chromosomes in the organism, yeast artificial chromosomes, bacterial plasmids, and the "computer programs that manage the structure, error checking, storage, retrieval, and distribution of genetic information." This observation of genetic thinking in the 1990s also demonstrates the reason that genomes being seen as information matters to human identity. Information is more controllable than matter. If that information can be transferred to a

computer, manipulation of organisms, including humans, becomes very easy. If all that humans are, body as well as behaviour, is controlled by one source of information, it could mean that genes contain our destiny, a destiny that is unaffected by individual choices. If scientists then were to have the key to that information, it should not be surprising that this idea frightened people in the 1990s.

COMPLEX INTERACTIONS What the effect of that could be is very well illustrated in the film *GATACA*, also from the end of the 1990s. In this film, the dream of a boy to become an astronaut is made impossible; not by his abilities, but by society's acceptance of the idea of genetic destiny. His genes are unsuited for being an astronaut, so no one will invest in him. If genetic scientists were then able to change this genetic destiny, it could at once assume the role of saviour and have more control over people's destiny than they themselves have. The images I describe here are genetic images from the 1990s, critiquing the genetic determinism of that time: the idea that genes are the single source of heredity and development. In the meantime, genomics has moved on. The idea that single gene single trait connections are the norm is something belonging to the past: genetic processes are now widely acknowledged to be complex interactions of genes and

their environment, inside as well as outside the organism. But the idea of genomes as information structures is still criticised for its genetic determinism, its over-emphasis on genomes in the constitution of organisms. So what is left of the idea of genes as information in the more complex context of the genomics era?

SYNTHETIC GENOME In order to get an answer to this question, I looked into a recent science practice that fits the idea of information as contained in different media: the J. Craig Venter Institute's synthetic genome. What Gibson *et al.* claim to have done fits the idea exactly: they claim to have proven the principle that a bacterial cell can be digitally redesigned in the computer, through its digitalised genome. I took this synthetic genome as a starting point. The reactions of eighteen scientists with a genomics background, which were published in the *Journal of Cosmology*, give an idea of how information plays a role in the current scientific discourses. In the majority of these reactions

it was argued that the synthetic genome was not reprogrammed and achieving reprogramming remains far off. But most interesting is that in the reactions the idea of genomes as information structure is maintained. Not only that, the idea was used in more genetic determinist as well as more complex views of cellular organisation. Even with the possibility of other elements of this organisation also being part of an information process, the emphasis came back to what was supposedly the source of this information: the genes.

NEXT STEP So while views of cellular organisation have become more complex since the 1990s, the idea of genomes as information is still very much alive. That it also features in more complex views of cellular organisation does not automatically mean that its connection to genetic determinism is broken. It seems that genes as main carriers of information still have a central role when the information concept is applied to a broader cellular process.

The relevance of this for human identity is that how functioning of genes in organism is portrayed will have its influence in all areas of genomics. How this works out in fields more related to human genetics and what the role of bioinformatics is in this context will be the next step in my research.

REFERENCES

1. Benner, S.A., *et al.* (2010) Commentaries: Artificial life. Scientific revolution? Or the end of life as we know it? *Journal of Cosmology* 8.
2. Gibson, D.G., *et al.* (2010) Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome, *Science*, 329, 52-56.
3. Haraway, D.J. (1997) *Modest_Witness@Second_Millennium.FemaleMan@Meets_OncoMouse™: feminism and technoscience*. Routledge, New York.

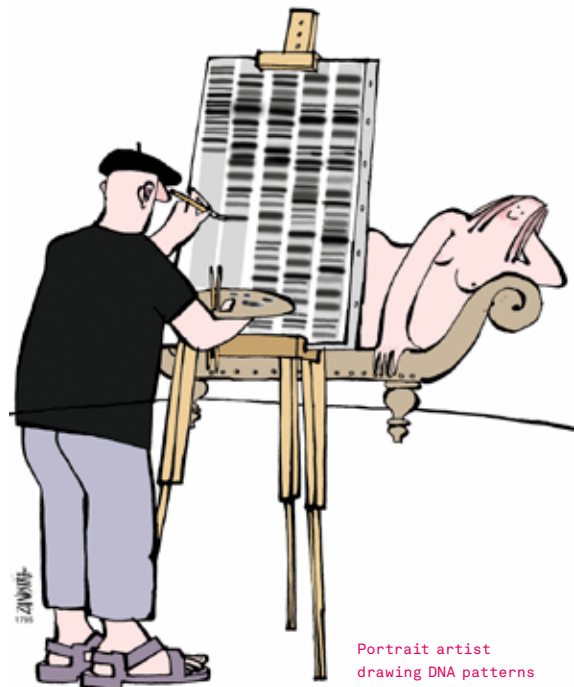


CONTACT:

Jan van Baren, MA
Radboud University
Centre for Society and
Genomics
PO Box 9010
6500 GL Nijmegen
j.vanbaren@science.ru.nl
www.janvanbaren.nl

→ KEY CONCEPTS

- **Information** in biology does not have one unambiguous meaning. It has however been a driving force in important developments in molecular biology: it inspired efforts to crack the 'genetic code'. It is central to Crick's dogma and it was an important driver for the Human Genome Project. Currently it drives the efforts of systems biology. Over time, its meaning has shifted. While in Crick's central dogma information is one directional, in systems biology it is a complex element, part of a network of cellular processes and their environment.
- **Human identity** is a complex concept with cultural and biological elements. It consists of *ideas* of what constitutes human beings but is also rooted in material elements like genes. At the same time it directly influences societal processes such as politics, law and health. Science has an important role in the translation between these aspects.
- **Situated knowledges**, a concept from Donna Haraway, refers to the view in Science and Technology Studies that (scientific) knowledge always depends on the position and worldview of scientists and as a result is valuable in *specific* situations. Rather than disqualifying science as knowledge producing practices, this is what puts science *in* this position.



Portrait artist
drawing DNA patterns

- The research described in this article is based on the project 'Bioinformation and human identity', a collaboration between CSG and NBIC.

By the editors

Playing games with our future

Big screens with high end graphics cards fill up our classrooms and research facilities. The computers are fitted with top notch processors and plenty of memory and disk space. Students are organising LAN parties, and even the Accreditation Organisation for The Netherlands and Flanders rates the bioinformatics facilities at the Hanze University Groningen as 'Excellent'.

At ALIFE, the applied research group at the Institute for Life Science and Technology (Hanze University Groningen), it is all serious business. Within the SIA-RaakPRO funded project 'BioComp' we develop high throughput and high performance infrastructures and applications using commodity hardware. We already have years of experience with advanced grid technologies and are currently running a grid with 178 nodes. This number is still growing, with eleven new computers now being installed. The grid is an integrated part of our education and research.

MORE COMPUTER POWER Our partner institutes – the University of Applied Sciences Arnhem Nijmegen (HAN), the University Medical Centre Groningen (UMCG), the Centre for BioSystems Genomics (CBSG), the University of New South Wales (UNSW), NBIC and KeyGene – recognised that with the current streams of data and processing pipelines, the need has arisen for even more computing

power, without taking a significant portion of research budgets. The general purpose graphics processing unit (GPGPU) offers high speed computing power for less money and is more energy efficient than CPU-based systems. The Chinese Tianhe-1A system for example, the fastest supercomputer in the world, not only consists of 14,336 Xeon processors, but also contains 7,168 NVidia Tesla M2050 GPGPUs. The future for bioinformatics analyses relies on these technologies. Not everyone can afford a supercomputer, however. For just a few hundred euros, you can fit a standard PC with a graphics card to create your own desktop supercomputer. Currently we have 19 GPUs and within a few weeks we'll have 30 GPUs, which will give us about 25 teraflops at peak performance. Having an infrastructure is of course convenient and even necessary for large scale data analyses. Many bioinformatics tools, like BLAST and Interproscan, already run on CPU-based grids without problems. GPGPU technology is, however, a much more recent development, with the first release of CUDA in February 2007 and with the open computing language openCL in October 2010. GPU hardware is completely different from CPU hardware, which makes GPUs very powerful but also requires development of new applications. These applications are released to the public often, and they focus on different aspects of bioinformatics

[1, 2]. GPU-BLAST enables a user to run BLAST on a graphics card; CUDA-MEME is the GPU-equivalent of MEME; and GPU-Hmmer allows you to calculate hidden Markov models.

THE BioCOMP PROJECT The Hanze University Groningen initiated the BioComp project to develop new software for GPUs, to further expand the knowledge and use of grid technologies in Life Science research and education, and to enable new methods of large scale bio-data analyses. Together with our partner consortium, we received a SIA-RaakPRO grant in 2009. The project started in September 2009 and will run for four years.

In the first part of the BioComp project we focused on developing a GPU version of the Smith-Waterman alignment algorithm (SW). SW is an important part of many sequence comparison programmes, but is very slow due to its algorithmic performance of $O(n^2)$. GPU applications do exist and show that the parallel nature of a GPU is very suitable for performing high speed database searches using SW. The Parallel Smith-Waterman Alignment Software (PaSWAS) we developed focuses not on search, but rather on producing the alignment profiles with the location of gaps and mismatches in the hit. PaSWAS enables researchers to align sequences with the benefits of the accuracy of Smith-Waterman and

the speed of the graphics card. For our partner KeyGene in Wageningen we used the software to extract primers and ID tags from 454 sequence data, and for University of New South Wales we classified immunoglobulin sequences and reported the number of mutations in these sequences. In both analyses we retrieved a higher number of sequences and detected more mutations than with BLAST-like software. On a single GPU it took only minutes to perform the over two million alignments. We will soon send the paper on PaSWAS for publication. At the NBIC conference in April we will present a poster with more detailed results. The next phase of the BioComp project consists of several parts. PaSWAS will be extended with a Java interface so it easy to use and it can be integrated with existing pipelines. The UMC Groningen

just started a PhD project to develop a genome wide association algorithm, which the Hanze University Groningen will support with knowledge and hardware. As more and more other applications are released, we will install, test and use these applications where needed. GPU-BLAST appears to be a very nice addition. It is used, for example, together with Hanze Energy Knowledge Centre (EKC) in metagenomics projects aiming at deciphering the key players in bio-gas production.

NEW TECHNOLOGY NEEDED

The numbers in biomedical and bioinformatics research are staggering. Gigabases per run on modern sequencer are commonplace; finding single nucleotide polymorphisms in millions of reads is normal; and performing genome wide association studies with hundreds of

individuals and millions of markers is nothing new. To keep up with all these data and questions, new technologies need to be employed. Who knew that the future of bioinformatics would be in using technology developed for shooting up virtual people in 3D with very realistic graphics. Gaming hardware is now serious business in bioinformatics research.

REFERENCES

1. Vouzis, P.D. and Sahinidis, N.V. (2011) GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics* 27, 182-188.
2. Manavski, S.A. and Valle, G. (2008) CUDA compatible GPU cards as efficient hardware accelerators for Smit-Waterman sequence alignment. *BMC Bioinformatics*, 9 (Suppl 2): S10

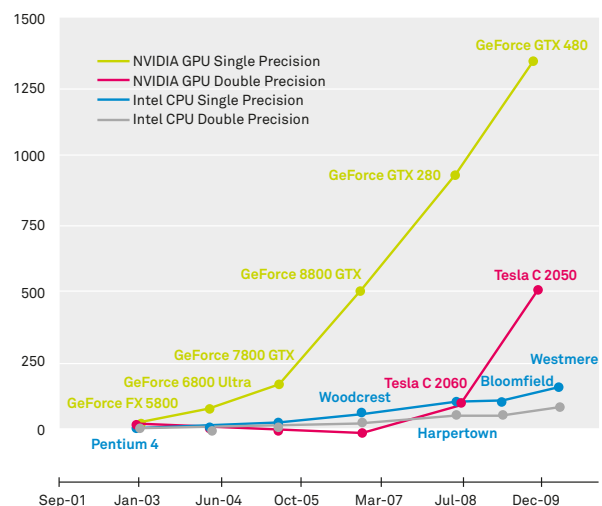


CONTACT:
Sven Warris
Hanze University
Groningen,
ALIFE, Institute for Life
Sciences & Technology
s.warris@pl.hanze.nl
@swarris on Twitter
www.hanze.nl/ilst

→ KEY CONCEPTS

- During the Mastering bioGrid project, students investigated the possibility of using either OpenGL or CUDA for high performance computing in bioinformatics using low-cost graphics cards.
- CUDA is an extension of C developed by NVidia (full C++ support has been available since version 3.0) especially for tasks other than graphics.
- From mid-2010 we have had about 160 Linux nodes in our grid. A 24-core, 96GB ram with 7TB storage will be used for metagenomics research. This computer has two Tesla C1060 GPUs. The grid also contains an additional 17 GT200-based GPUs and at least 14TB of storage.
- Currently we are testing this network with our GPU software and data from our partners.
- Making use of commodity hardware like standard desktop PCs fitted with GPGPU-enabled graphics cards opens the door for low-cost high throughput / high performance bio-computing.
- The Parallel Smith-WatermanAlignment Software was the first GPGPU application developed in the BioComp project and more will follow.
- The Hanze University Groningen initiated the BioComp project to develop new software for GPUs. Along with our partners, we received a SIA-RaakPRO grant in 2009. The project started in September 2009 and will run for four years.

Theoretical GFLOP/s



PERFORMANCE INCREASES OF NVIDIA GPUS COMPARED TO INTEL-BASED CPUS.

Giga floating point operations per second (GFLOP/s) given for both platforms show that the GPU is not only much faster, but also show a much better performance increase. With the release of the Tesla C2050, the same is true for double precision floating point operations per second. For most calculations in bioinformatics tools, however, single precision is accurate enough.

Source: NVidia CUDA Programming Guide for CUDA Toolkit 3.2.

Column

Hans Roubos



Hans Roubos
Senior Scientist
DSM Biotechnology
Centre, Delft

CONVERGING TECHNOLOGIES

Bioinformatics is a crawling robot that is learning to walk. Still in its infant stage, it senses the future with an open mind. Personally, I am highly excited about the pace and movement of what is happening in biotech and bioinformatics. Rapid developments in the field of Nano, Bio-, Info- and Cogno technologies – NBIC in short – are driving the field forward.

DNA sequencing is one such technology. I was introduced to Sanger sequencing in a lab course nearly twenty years ago – it took us half a day to unravel a sequence of only 500 base pairs. The Sanger institute had just been formed at that time (1992), and we experienced the first signals of a World Wide Web with the introduction of Mosaic. The primary goal at Sanger was the elucidation of the human genome in an estimated 15-year public effort. Venter and Smith led the first successful effort to sequence an organism's entire genome: *Haemophilus influenzae* at TIGR (1995). Baker's yeast *S. cerevisiae* and *E. coli*, both important for biotech, followed shortly thereafter. This breakthrough initiated a whole new research area, with many aspects for us just emerging bioinformaticians. Today, the molecular biology database issue of Nucleic Acid Research lists 1330 Dbs. GenBank contains over 100 billion unique base pairs – a number that doubles every 18 months.

Venter believed that shotgun sequencing was the fastest and most effective way to obtain human genome data. Although his idea was controversial – unable to get public funding – he had a mission. Celera Corporation was founded (1998) and already two years later a draft of a human genome – result of both public and private effort – was announced jointly by then US president Clinton and the British Prime Minister Blair. Celera sequenced the human genome at an estimated cost of 10 percent of the public \$3 billion project, helped by automation, advanced bioinformatics and impressive computing power.

A few months ago, Ion Torrent's personal genome machine was launched. The size of a desktop printer and at a cost under 50k\$, it will again revolve the way we work. The basis is a micro-fluidic chip with millions of reactors and direct binary read-out during a sequencing-by-synthesis reaction, delivering 100 Mb of data in a two hour run. Moore's law continues to apply. The belief is that upcoming single-molecule sequencers will be able to provide the entire human genome within 2 hours for less than 1000\$ in 2014.

Now, nearly twenty years after my introduction to Sanger sequencing, I type in 'alcohol dehydrogenase' on Google: 636,000 results (0.07 seconds). I select a protein sequence at Entrez-Pubmed, do my DNA redesign and order the sequence online at one of a dozen available gene synthesis providers. My order will soon be delivered by regular mail. At iGEM, student teams are generating novel functionality in micro-organisms in a summer course. Gibson and Venter – with a new mission – just published the first synthetic organism, the re-programming of a cell now controlled by a synthetic genome. Bioinformatics has become a cornerstone of industrial biotech innovation programs.

2020: a personal genome printer is announced. Imagine... Bright Science... Cloud computing is outdated. Watson stopped playing Jeopardy and grows its artificial brain with biology.