Coverstory
Struggling with terabytes from the next generation sequencers

Interview
Ron Appel
## Content

<table>
<thead>
<tr>
<th>Editorial</th>
<th>Developing an algorithm</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverstory</td>
<td>Struggling with terabytes from the next generation sequencers</td>
<td>4</td>
</tr>
<tr>
<td>Course</td>
<td>Browsing genes and genomes with Ensembl</td>
<td>7</td>
</tr>
<tr>
<td>Interview</td>
<td>Ron Appel, director of the Swiss Institute of Bioinformatics (SIB)</td>
<td>8</td>
</tr>
<tr>
<td>Spin Off</td>
<td>Bio-prodict builds research databases for protein families</td>
<td>11</td>
</tr>
<tr>
<td>Progress</td>
<td>Computational prediction of protein flexibility in structure-based drug design by Sander Nabuurs</td>
<td>12</td>
</tr>
<tr>
<td>Progress</td>
<td>Searching life science literature and discovering new knowledge with Anni 2.0 by Martijn Schuemie</td>
<td>14</td>
</tr>
<tr>
<td>Hands On</td>
<td>Wikipathways: community-based data curation</td>
<td>16</td>
</tr>
<tr>
<td>Thesis</td>
<td>Genome-wide approaches towards identification of susceptibility genes in complex diseases</td>
<td>18</td>
</tr>
<tr>
<td>Portrait</td>
<td>Seven questions for Rudi van Bavel</td>
<td>20</td>
</tr>
<tr>
<td>In the Picture</td>
<td>NBIC news</td>
<td>21</td>
</tr>
<tr>
<td>Column</td>
<td>Bas Teusink: to model or not to model, that is not the question</td>
<td>24</td>
</tr>
</tbody>
</table>
Developing an algorithm

Imagine a puzzle of many seemingly incompatible pieces. A great challenge indeed. That is how it is when making up a well-balanced NBIC programme. The recently approved NBIC-II business plan provides a unique opportunity to continue part of the current programmes and to initiate several new activities. I am very grateful to everyone who has made a contribution and acknowledge all (junior) researchers of NBIC-I who contributed to this recent success through their research projects. Thanks to all of them who did not meddle with the business plan and thereby helped in not complicating the bioinformatics puzzle further.

Let’s consider a few ingredients of the business plan. The needs of the biologists were clearly driven by biological questions. However, these were heterogeneous and could potentially lead to a programme with isolated projects. In addition, discussions made clear, again, the great need for bioinformatics. In fact, the extent and broadness of bioinformatics needs were overwhelming.

Discussions about the role of NBIC in the Dutch life sciences field were triggered and the definition of bioinformatics was at stake since other disciplines increasingly affect its nature. Bioinformatics, imaging, medical informatics, computer science and systems biology are slowly but surely converging. Yet, there is no consensus on how to approach these developments. Expertise currently embodied by Dutch bioinformatics groups provided another constraint, and research, support, education and exploitation needed to be balanced.

Then there is the issue of ‘doing what the biologist thinks is best’ versus ‘doing what the bioinformatician knows is best’. The bioinformatician is afraid of ending up doing routine tasks resulting in early (academic) retirement, while the biologist is afraid of not getting an answer within the next five years. There are, indeed, examples of this. Therefore biologists want to have bioinformaticians in their labs but I’m sure that there are bioinformatics groups out there that have been thinking of getting a wet-lab next to their computer to really make progress in bioinformatics. Nevertheless, past performance showed that NBIC is very focused on solving biological questions. A member of a past review committee once remarked that too many ‘biological papers’ resulted from our programme. We promised to do ‘better’ in future.

We had to solve this puzzle. A coherent and strong programme was to be established and funding had to be prevented from vaporizing. The resulting business plan presents our algorithm that, I believe, will be successful in further strengthening the Dutch bioinformatics community, in addressing the most urgent needs of the genomics community and in contributing solutions for the ever increasing bioinformatics demand. Unfortunately, we could not satisfy everyone this time but I’m confident new opportunities will arise.

NBIC-II will start in 2009. Naturally, Interface will keep you informed. The same way we do in the present issue which is by informing you of current developments within NBIC-I. The various articles report on progress in bioinformatics programmes including research, support, education and exploitation. The cover article focuses on the next generation of DNA sequencers and the accompanying need for new bioinformatics tools. We also pay attention to the Swiss Institute of Bioinformatics who have just celebrated their 10th anniversary. We are proud to present an interview with its scientific director, Ron Appel. I’m sure he can teach us more about how to develop algorithms.
Struggling with terabytes from the next generation sequencers

Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by more than two orders of magnitude. These new technologies are rapidly evolving, and near-term challenges include the development of robust protocols for generating sequencing libraries and building effective new approaches to data analysis.
Genomics researchers expect that hundreds of second-generation sequencers will soon be in use in the Netherlands. This revolutionary development in DNA sequencing technology will challenge bioinformaticians. Strong computing power and advanced software tools will be needed to analyse the overwhelming dataflow. NBIC’s BioAssist platform ‘high throughput sequencing’ is planning a (central) facility to support local platforms and to stimulate collaboration.

The Plant Sciences Group in Wageningen intends to sequence all the DNA-base pairs from the tomato and the potato. We are talking about nearly one billion base pairs for the genomes of the two crops. For this enormous undertaking – which has not yet been done anywhere in the world – the plant researchers will make use of different second-generation sequencers this autumn. The problems will probably start shortly after they have put the DNA material in these (table model) machines. By then the computer will have already stored 80 million DNA fragments for the tomato alone. “All these 80 million pieces have to be combined in the right order,” explains Roeland van Ham, who is responsible for this bioinformatics. “We are facing a huge challenge.”

NEW MACHINES ARE POPULAR The Wageningen plant researchers won’t be the only genomics scientists in the Netherlands struggling with data from second-generation sequencers. At the moment there are already twelve second-generation sequencers in the Netherlands, and many institutes and companies intend to buy one or more, because the new machines are so popular with their researchers. “Ultimately, we expect there to be hundreds of second-generation sequencers in the Netherlands,” says Johan den Dunnen from the Leiden University Medical Center, whose group has recently sequenced the whole genome (3 billion base pairs) of his colleague Margriet Kriek using a Solexa machine. “Every hospital will have them: first to sequence the DNA of patients with a hereditary disease, and later to sequence the DNA of healthy people as well. By trying to analyze the data of an entire genome now, we hope to be prepared by the time sequencing has become a common technology in hospitals.”

Second-generation sequencers have been on the market since 2005. They have far more possibilities than first-generation sequencers, which were based on the ‘Sanger’ method developed 30 years ago. The first sequencing of the human genome using the Sanger method was completed in 2001. It took 13 years and cost 3 billion dollars. The sequencing of Marjolein Kriek’s genome cost around 50,000 euros and took just a few months. This is partly because with the previous methods the DNA had to be cloned, i.e. built into *Escherichia coli* before the base pairs could be read. This is no longer necessary, which saves a lot of time and money. Another difference is that instead of tens of fragments, hundred thousands to millions of DNA fragments can be read at the same time.

RAPIDLY GROWING APPLICATIONS There are many problems to solve, according to Den Dunnen: “For the sequencing of one human genome we already generated somewhere between 4 and 6 terabytes of data. That was too much for our hard disks, which can only store hundreds of gigabytes. So we had to go to the shop several times to buy 1-terabyte hard disks.” One terabyte is 1000 gigabytes, which is enough to store all the information in a big university library. The Solexa machine in Leiden has generated over 500 million pieces of DNA of 35 letters. These pieces have to be placed in the right order using the human reference genome. The Leiden researchers have already extracted the mitochondrial DNA from the disks, but this only represents a first, tiny step in simplifying the analysis. Three leading companies sell the new equipment: 454 Life Sciences (Roche Diagnostics), Applied Biosystems and Solexa (Illumina). These companies deliver software, but only for relatively easy applications, such as ordering a manageable amount of sequences (‘reads’) when there is already a ‘reference’ genome. These are the kind of analyses that many researchers are already used to doing with the Sanger method. That software may be useful, for instance when medical researchers want to compare a specific gene in different cancer patients, or when plant researchers want to compare specific genes of different ‘ecotypes’ of *Arabidopsis thaliana*.

However, with the advent of second-generation machines, sequencing has become so cheap that many more applications are possible. A very popular one is sequencing all messenger RNAs (mRNAs) in a cell, to find out which genes are expressed and in what amount, for instance after the administration of a medicine or a toxic compound. Another one is sequencing all intestinal bacteria in mice or humans to find out what role they play in intestinal diseases. A third new application is sequencing all microRNAs in a cell. Cells can store hundreds of thousands of these regulatory molecules. It is known that they can influence the translation of genes in proteins, and thus the development of a cell, but it is not known exactly which translation steps each type of microRNA influence. The Hubrecht Institute in Utrecht has bought a second-generation sequencer for this fundamental research. “We see it as an enabling technology,” explains Edwin Cuppen. “With these sequencers in house, research questions arise easily.”

NEED FOR SUITABLE SOFTWARE There is no suitable software yet for most of these new applications. Johan den Dunnen for instance wants to know which DNA variations Marjolein Kriek’s genome has because these variations can tell us something about possible genetic predisposition for specific traits and diseases. The researchers in Leiden expect more than a million variants. First they have to find out which differences in bases are true or due to sequence errors. Den Dunnen: “We see some DNA reads 40 times and thus have high confidence, but others we only see 3 times. When do we have enough certainty to decide that a difference is not a sequence error but a variation? We need software to solve this problem.”
The Leiden researchers also need to know which base differences are already described in the literature, and what consequences they have, and which differences are new. What makes this puzzle even more complex is that for the 600,000 to 700,000 variations already known, the findings are collected and described in dozens of different databases. Another task is therefore to develop suitable searching tools that not only can quickly find a specific variation in a genome, but also can link this variation to relevant literature concerning the possible health consequences.

The International Tomato Sequencing Project has already read 66 million base pairs using the Sanger method (in 5 years). But because not all cloned tomato DNA is accepted by the bacteria, the Sanger method alone cannot provide complete insight. “So it is time to do it in a new way,” explains Van Ham. The Wageningen researchers will start sequencing the tomato and potato again, using a Solexa SOLID and their own Life Sciences 454 machine, which they have already been using since January. This combined strategy will lead to more reliable outcomes, but comparing analyses from two machines also requires specific software.

Every single read has to be compared with every other read: pieces that have both their ends sequenced (“paired ends”) will help to build up scaffolds of chromosomes and ultimately the entire genome. Assembling tools have been developed for ‘de novo’ sequencing using the Sanger method, but the Sanger machines can only generate one or two million bases per run, while the second generation machines can produce up to 10 billion bases in a single run. Scientific programmers funded by the NBIC BioAssist program will benchmark and implement the software. The software and results from this project can be used for ‘sequence-based breeding’ in the future: seed companies will have their tomato or potato (partly) sequenced, and will be able to breed using the plants with the best DNA.

NBIC BIOASSIST PLATFORM Finding solutions will be easier and cheaper if the Dutch research centers collaborate, according to the genomics researchers. The Hubrecht Institute for instance has now combined the computing power resulting in a cluster of 150 central processing unit (CPUs). The UMC Utrecht, which has bought a second-generation sequencer this year, also uses these 150 CPUs of the Hubrecht Institute. “But this ‘in-house’ computing power won’t be enough,” says Edwin Cuppen. “The primary processing after one sequencing experiment already takes a whole day, and the number of experiments is growing fast.”

A (central) facility for storage and complex analysis could help local platforms such as the one in Utrecht. This is the idea behind the NBIC BioAssist support platform ‘high-throughput sequencing’ that Roeland van Ham is managing. “The platform is collaborating now with the computing and networking center Sara in Amsterdam to set up just such a facility,” Van Ham says. Once this is realized, researchers in Groningen for instance will be able to send their millions of base pairs to the facility for e.g. analysis. They will also be able to choose specific pipelines, including pipelines (software) for gene expression data, de novo sequencing or re-sequencing. Having got this analyzed data back, they can then use their own software to address more detailed questions. The platform can also play a role in choosing and developing the software for these pipelines. “At the moment we are working to make these services accessible,” concludes Van Ham.
Browsing genes and genome databases is an essential skill for many life sciences researchers. Following a course in which you are guided through all the tools of the Ensembl genome database in two or three days saves a lot of time. The basic Ensembl workshop from the postgraduate school of Erasmus MC is therefore very popular. Recently, a one-day advanced course was also organised for the first time. What can you learn at these courses?

Sylvia Brugman, postdoc researcher at the Erasmus University followed both the basic and advanced workshop on using the Ensembl Genome Browser developed by the European Bioinformatics Institute (EBI) and the Sanger Institute. She explains why using a gene database is essential for her research: “I study inflammation processes in the gut of zebrafish. The animal serves as a model for diseases like Crohn’s disease and Colitis Ulcerosa. In the inflammation processes cytokines play an important role and therefore I want to study the expression of the cytokine genes under different circumstances. However, thus far not all zebrafish cytokine genes are known/sequenced, in contrast to mice and other species. So I need to rely on cytokine gene information from other species, for example to generate probes and primers for my experiments.” Brugman used the Ensembl database to find information about the cytokine genes, but she could not find all the details she needed. She also felt that she was not using all the possibilities the database offered. “I was able to track down information on exons within the genes, but I did not know how to find the location of exon/intron transitions, for example. Furthermore, there are also other databases like the ones offered by NCBI (National Center for Biotechnology Information in the USA) in which similar information can be found. How do you know which database is the best for finding your information or which is the most up-to-date?”

She decided to follow the Ensembl basic and advanced course, which she approved of, especially the basic course: “I learned how the different databases are maintained, how their data is collected and what the connections are between them.” The Ensembl database integrates other genome database like NCBI and is therefore the most complete. During the course the teachers demonstrated all the tools, one by one. “You could follow all the steps on a large screen. There were also exercises in between which were very useful. For example, they asked you to find a certain gene and then go to specific elements within the gene. The course book contained all kinds of exercises as well as the answers,” says Brugman.

The advanced course was given for the first time and participants could ask specific questions they come up with when using the database. They were asked to send in questions in advance by e-mail. The questions were discussed during the one-day course. Brugman: “The participants had very diverse questions, which probably were not relevant to everyone but were interesting to explore. I may need some of the tools discussed (finding SNPs) in the future.” Teacher Bert Overduin and his EBI colleague Bronwen Aken emphasized that the participants should contact the Ensembl Helpdesk whenever they have questions. The helpdesk people are glad to receive questions. This way they also know what needs to be improved or added to the tools and website.

Besides Ensembl, MolMed organizes other bioinformatics courses such as the International Phylogeny Workshop, a UCSC gene browser course given by Jim Kent, and basic courses on Applied Bioinformatics.

For more information go to: www.molmed.nl
“Major challenges in bioinformatics for clever computer scientists” →
In July 1993, the word proteomics had not yet been invented and the World Wide Web had only 150 web sites. But the 34-year old computer scientist Ron Appel immediately saw its potential. He made his gel electrophoresis protein database available for other scientists through ExPASy, the Expert Protein Analysis System. Fifteen years later, Ron Appel is the director of the Swiss Institute of Bioinformatics (SIB), which employs nearly 300 bioinformaticians. “The World Wide Web just seemed to us a much easier way of exchanging information with other scientists.”

You’ve been involved in proteomics right from the beginning of the field of research. Looking back, what were the largest leaps or milestones in the field?

Definitely, the transition to high-throughput analysis of the life sciences. That started with the use of 2D-gel electrophoresis for the analysis of proteins, and later got a boost with the introduction of mass spectrometry. The huge amount of data produced had to be accompanied by the development of bioinformatics. From a more personal viewpoint, the most important leap was the advent of the World Wide Web. It created the possibility to access and exchange results and data by all members of the life sciences community.

And what—in your opinion—is a current major obstacle in the field of bioinformatics?

The fact that not all biologists have yet recognized that in current life sciences research projects you need as much dry lab— that is bioinformatics—as wet lab resources.

How did you—a computer scientist—become involved in the life sciences?

Ever since I was a kid, I loved playing with numbers. At the end of my masters in computer science in 1983, I wanted to continue ‘playing’ by pursuing a Ph.D. However, I also wanted to work on a project that was useful to society. I anticipated that a project in medicine would be valuable and searched for a physician who was interested in computer science. I found one in the person of Denis Hochstrasser of the University Hospital in Geneva. He was working on 2D-gel electrophoresis of blood proteins as a diagnostic tool. I started to develop software to identify the thousands of spots. Denis also reconnected me with a former fellow student: Amos Bairoch. Amos was working on the protein sequence database, which would become SWISS-PROT. Together we decided to build the web server ExPASy to make our tools available for other scientists.

ExPASy was actually the very first database on the World Wide Web. How did you come up with the idea?

It just seemed to us a much easier way of exchanging information with other scientists, and that turned out to be true. However, at that time we had quite a job convincing people at the university hospital that allowing scientists from outside to use our server was a good idea. The head engineer only saw the extra work he would have to do and refused at first. My boss had to pull some strings to get things going.

Are you still a big supporter of open source tools?

Yes, open source provides access to many, sometimes outstanding programs, and it creates world-wide development communities. However, it also raises a number of questions, such as the guarantee for continuity as well as the funding of such developments.

You are the director of the Swiss Institute of Bioinformatics (SIB), which has just celebrated its 10th anniversary. What has the institute achieved in these ten years?

A lot. We brought together the most prominent bioinformaticians in Switzerland and became the backbone for IT in the life sciences. But there were also many individual achievements. To give an example, we just launched the complete annotation of human proteins, a huge project lead by Amos Bairoch. This new protein encyclopedia, which required SIB’s Swiss-Prot collaborators to read 45 thousand articles, is the result of the work of thousands of health researchers around the world. The data covers a total of 20,325 human proteins.

“We need stable, long term funding to safeguard the future of the bioinformatics’ infrastructure and data”

And that wouldn’t have been possible without SIB?

It’s always difficult to say what would have happened if……, but I think indeed that it probably wouldn’t have been possible. SIB was actually born from a severe crisis in bioinformatics funding in Switzerland. ExPASy was used a lot, but there was no long-term funding. In May 1996, no one received a salary. The twenty bioinformaticians from the Universities of Geneva and Lausanne then joined forces in SIB to acquire money to continue their work. Without this action, ExPASy and Swiss-Prot may not even have survived.

SIB unites 25 different Swiss bioinformatics groups and about 300 scientists. Wouldn’t it be better to create one central location?

The structure of SIB follows the long federal tradition of Switzerland with its 26 cantons. I consider this an ideal structure. Not only for a bioinformatics institute, but also for education, politics and public administration in general. The federal concept combines the best of two worlds: the decentralized and centralized world. For SIB, it guarantees the development and maintenance of top quality.
facilities, an activity funded by the central Swiss government. The universities are financed through the cantons and independently try to attract the top researchers in bioinformatics who join SIB.

“You can hardly be a biologist today without using a computer to understand the data you collect.”

Yet, SIB is quite unique in Europe. Spain has a similar institute, as do you in the Netherlands with NBIC. We seek cooperation with these institutes as we are natural partners. I, for example, am a member of the scientific board of NBIC. Recently, NBIC and SIB also agreed to open up their summer courses for each others students. There is certainly a need for more coordination in bioinformatics throughout Europe. ELIXIR, the European Life Science Infrastructure for Biological Information, is an important initiative in this perspective. We need stable, long-term funding to safeguard the future of the bioinformatics’ infrastructure and data, and a European-wide collaborative agreement on who does what.

How do you create a sense of community within a network of 25 different groups?
As I mentioned before, SIB started out as a small group of bioinformaticians who struggled to survive. We fought side by side for the ‘life or death of bioinformatics’ in Switzerland, which created a very strong sense of community. Over the past ten years, SIB has grown to 300 people today. To celebrate our 10th anniversary, we organized events at all locations and many people volunteered to help, and more than 300 people participated in the celebrations in Bern. So, I think the sense of community within SIB is still very much alive. But indeed, it is one of our challenges to keep it up. A key issue is to keep up our high quality standards. SIB must be an institute with which every self-respecting bioinformatician likes to affiliate him- or herself.

ExPASy offers many databases and software tools in bioinformatics. Which one is the dearest to you? And why?
That must be the ‘Contact us page’: www.expasy.org/contact.html. We provide bioinformatics services that we enjoy seeing used. Through this page, fellow scientists can tell us about the problems they encounter when using our services and make suggestions for improvement and put up proposals for collaborations. And, they may express their satisfaction.
The first itch to start a company was felt by bioinformatician Dr. Henk-Jan Joosten six years ago. At that time he was working with two other students on a master thesis for Professor Gert Vriend at the Radboud University Nijmegen (CMBI). “We worked for Organon (present Schering-Plough) on the generation of a system that contained all the information that was available on the internet for a receptor protein. We thought this was a good product to build a company around but it was a bit too early and we all went our own way.” Now, Joosten is director of Bio-Prodict, a company in the field of bioinfotech.

After finishing his master study, Joosten started a PhD research project in the laboratory of Microbiology at the Wageningen University on a complete other protein. His work in Nijmegen concerned a receptor protein and had shown the enormous amount of information that is available for one class of proteins. Joosten already knew that the trick is to store it in a clever way, so that it becomes easy to answer complex research questions regarding protein stability, specificity, activity, and effects of mutations. “To solve my research questions in Wageningen I needed the same kind of system, but I did not want to spend one and a half years building it. So I hired an informatics student to write a program which does the generation of such systems automatically. In that way I would be able to do the same trick over and over again for other proteins or applications. What first took one and a half years can now be done in three to four weeks.”

The competition for business ideas organized by the Wageningen Business Generator (WBG) started the ball rolling. Joosten won first prize, which meant that WBG helped with financial and legal advice and to develop a business plan. “They also provide financial support and take a 51% interest. They demand that you invest in the company too; a kind of risk sharing and a big stick to make it work.”

Bioinfotech Company  Now two years later Joosten is the proud director of the brand-new bioinfotech company Bio-Prodict which constructs databases of protein families for its customers. Dr. Peter Schaap, his co-promoter, and Professor Gert Vriend are both advisors at Bio-Prodict. Housed in his former laboratory, Joosten stays close to the scientific scene. Half of his time is spent programming and managing a research fellow within an NBIC project. Together they work on the further development of the system, called 3DM. The other half he is out meeting customers and planning the future for Bio-Prodict. For a starter company the list of customers is already impressive. DSM, Danisco, Schering-Plough and even a university in South Africa are included. “Just building and selling databases as an activity is too little for a healthy business,” says Joosten. He sees opportunities to expand the applications of 3DM for DNA-diagnostics for instance for breast cancer. “A mutation in specific proteins related to a certain disease might be the reason the disease appears. The databases can be used to predict whether the mutation has an effect on the protein and could therefore be the reason that the disease has developed within the patient. This is a large research project we are now discussing. It would take a few years development to get the system operational, and a few more hands.” His ambition is to have in about three years four men working within the company making it possible to sell 3D systems all over the world.

Bio-Prodict B.V. is a spin-off of the Wageningen University and founded by Henk-Jan Joosten with help of the Wageningen Business Generator, the CMBI and NBIC, and is built around an information system called 3DM. The company is specialized in generating protein-family databases for research intensive companies and institutes.

More Information: Bio-Prodict B.V.
www.bio-prodict.nl (under construction)
3DM System: http://3dmcsis.systemsbiology.nl
In drug design it is of great value to understand how (potential) drugs bind to their protein target. Unfortunately, this is still very difficult to predict, foremost due to flexibility of the protein receptor. Fleksy is a prediction method which can accurately consider protein flexibility.

Many efforts in bioinformatics are ultimately aimed at identifying the genes causing, or involved in, disease. However, once these target genes have been identified, there is still a long way to go before a drug is developed and the disease can be treated. One of the first steps in developing a new drug is identifying potential drug candidates that can inhibit, or in some cases activate, the protein target of interest. In the subsequent lengthy process of optimizing a drug candidate into a drug, understanding just how this drug candidate interacts with its protein receptor is very valuable. This, together with the increased availability of crystallographically determined receptor-ligand complexes, has resulted in a tremendous increase in the application and impact of structure-based drug design (SBDD) in drug discovery over the past years. SBDD comprises a tight integration of biomolecular structure determination with computational methods to predict and optimize molecular complexes. Unfortunately, at the early stages of drug development crystallographic information on the receptor-ligand complex of interest is often not or only limitedly available. It is in these situations that computational predictions can come to the rescue.

**CHALLENGE** To predict the three-dimensional structure of a ligand bound to a protein receptor many so-called ‘docking’ programs have been developed throughout the years. One feature most docking programs share is that they traditionally aim at positioning a flexible ligand into a rigid binding site. Over the years however, it has become increasingly clear that protein structural flexibility plays a crucial role in receptor-ligand complex formation and really should be considered during the drug design process [1,2]. Most of the existing methods do not, unfortunately, consider protein flexibility, which is the main reason why these methods often only produce moderately accurate results. This shift of focus from the traditional ‘key-and-lock’ concept to the ‘induced fit’ model has prompted the need for computational tools which are able to consider protein plasticity.

**FLEKSY APPROACH** We have developed a multi-stage protocol which is capable of considering protein flexibility in the prediction of receptor-ligand interactions. This protocol, which we named Fleksy, is graphically depicted in the Figure. The first stage of the Fleksy procedure is aimed at identifying and sampling (potentially) flexible amino acids in the target protein. For this we make use of any information we can get our hands on: this can range from experimental crystallographic information to sequence conservation. With this information a subset of flexible binding site amino acids is selected. Subsequently, alternate conformations accessible to the selected residues are sampled. To this end a knowledge based rotamer library is applied, which predicts alternate orientations available to the different selected flexible side chains (stage A in Figure). In the second stage the orientations accessible to amino acids that can readily adopt alternate conformations such as asparagine, glutamine and histidine are also sampled (stage B in Figure). The structure models generated in stage A and stage B are combined into one large family, or ensemble, of structures, which encodes the different degrees of flexibility that can be sampled by the receptor. The next challenge is then to correctly position the ligand of interest into this protein ensemble (stage C in Figure). To this end, we make use of the FlexX-Ensemble docking engine [3], which is capable of considering multiple protein structures simultaneously during the docking procedure.
To allow for additional flexibility to occur, the protein structure is consid-
ered as ‘soft’, allowing the generated ligand conformations to slightly pro-
trude the receptor. In this way small steric clashes, which could otherwise
prevent the ligand from entering the binding pocket, are easily overcome.
Typically, the twenty most promis-
ing ligand orientations obtained after docking are retained.
A drawback of the described soft-
docking procedure is that the com-
plexes that are generated are physi-
cally not very realistic. The selected
structures are therefore subjected to
a minimization protocol, in which the
goometry of the complexes is opti-
mized (stage D in Figure). The result-
ating optimized complexes are sub-
sequently scored using a so-called
consensus scoring function (stage E in
Figure). In this consensus score, three
different energy scores are combined
to generate one overall ranking of the
generated complexes. The most high-
ly ranked complex is selected as the
final prediction.

VALIDATION To assess the perform-
ance of the Fleksy approach, we eval-
uated it on a large number of phar-
maceutically relevant structures, of
which a single example is shown in
Figure 1. In addition to this example,
Fleksy was validated using over 300
docking experiments utilizing some 35
different receptor-ligand complexes
taken from the Protein Data Bank.
Averaged over the different datasets
Fleksy accurately reproduced the
observed binding mode for 78% of the
complexes. This compares favourably
to the rigid receptor FlexX program
which reaches an average success
rate of 44% for these datasets.

The obtained results clearly show
the ability of our docking pipeline to
consider receptor plasticity during
small molecule docking. As such, the
method is suitable to be readily imple-
mented within a computational drug
discovery environment.

REFERENCES
flexibility for drug discovery. Nature Reviews
Drug Discovery, 2, 527-541.
protein flexibility into docking and struc-
ture-based drug design. ExpertOpinionDrugDis-
covery, 1, 335-349.
3. Claussen, H., Buning, C., Rarey, M., Len-
docking considering protein structure var-

CONTACT
Sander B. Nabuurs, PhD
Center for Molecular and Biomolecular
Informatics (CMBI)
Radboud University Nijmegen Medical Centre
(RUNMC)
PO Box 9181
6500 HB Nijmegen
Web: http://sander.nabuurs.org
E-mail: sander@nabuurs.org

KEY CONCEPTS
Fleksy is a novel approach to consider both ligand
and receptor flexibility in small molecule docking.
Averaged over several validation sets Fleksy repro-
duces the observed binding mode for 78% of the
complexes. This compares favourably to the rigid
receptor FlexX program which on average has a suc-
cess rate of 44% for these datasets.
Fleksy is developed jointly with scientists at the
pharmaceutical company Schering-Plough in Oss,
The Netherlands (formerly N.V. Organon), as part of
a BioRange project coordinated by NBIC. The result-
ning software is actively used by molecular modellers
at various Schering-Plough sites, in The Netherlands,
Scotland and the United States.
The author’s plan is to make the Fleksy protocols
available for download to the academic community
by the end of 2008.
A paper on Fleksy was published in the Journal of
Medicinal Chemistry: Nabuurs, S.B., Wagener, M.
fit docking. Journal of Medicinal Chemistry 50, (26),
6507-6518.
Sander Nabuurs was recently awarded a VENI grant
by the Netherlands Organization for Scientific
Research (NWO) based on a proposal which builds on
the work described here. With this grant he aims to
develop novel methodology which can consider even
larger degrees of protein flexibility, such as loop and
domain movements.

LEGEND
A visual outline of the Fleksy approach, illustrated using the
induced fit docking of progesterone into the apo structure of the pro-
gesterone antibody DB3. The Fleksy procedure consists of five differ-
ent stages (A-E), which are discussed throughout the text. In this
example the agreement between the modelled and co-crystallized lig-
and, shown in blue and orange respectively, is excellent.
One of the major challenges in efficient biological research today is to make full use of information available from the exponentially growing amount of scientific literature. To use this information one must take the ambiguities of natural language into account, and sometimes combine information across different disciplines. We developed Anni as a tool to find hidden information in literature.

Despite efforts such as the Gene Ontology and UniProt, a substantial portion of all available information about genes and proteins can still only be found in the literature, where it resides in the form of natural language. In order to find this information, biologists traditionally have to read through many papers, and while doing so must sometimes combine information from several articles from different fields in order to answer their question.

**TEXT-MINING** A wide range of initiatives has been developed to mine the literature. One of the emerging approaches is text-mining, which infers associations between biomedical entities by combining information from multiple papers. Anni is an online available text-mining tool (http://biosemantics.org/anni). The goal of Anni is to aid biologists by linking the literature to concepts described in ontology in order to find relationships between various biological entities such as genes, proteins, biological processes, diseases and drugs. Even the language used in scientific literature is riddle with ambiguities. This is especially true when it comes to names of genes and proteins. For instance, the Prostate Specific Antigen can also be referred to by its abbreviation PSA, or its systematic name KLK3. The latter, in turn, can also be spelled as KLK-3 or KLK-III. In order to deal with this variety, we have created an extensive ontology, where synonyms and spelling variations are mapped to a single concept [1]. Furthermore, gene names are also often homonyms: for example, PSa can also stand for PSoriatic Arthritis, or even the Poultry Science Association! We have developed several simple rules for distinguishing between different meanings of a word, achieving remarkably good results. For example, we achieved a precision of 75% in the BioCreative II gene name recognition task [1]. In contrast, if one were to simply link all occurrences of gene names to the respective genes, the precision would be about 9%, with similar recall [2].

**CONCEPT PROFILES** In order to find relationships between concepts, Anni makes use of the notion of co-occurrence: if two concepts appear together more often than would be expected by chance, then we assume that they must be related. For each concept, we list all the related concepts in its concept profile, weighted by the strength of the association. For instance, in the concept profile of a particular disease, we will find the genes that are often mentioned in the same articles as that disease. By comparing concept profiles, we can even find indirect relationships between concepts, including new relationships that have never been mentioned explicitly in literature. For example, if we compare the concept profile of a drug to a disease, we might find that both profiles have many genes in common. Based on this, we could formulate the hypothesis that the drug might be used to treat the disease, even though they have never been mentioned together before. This process is called knowledge discovery, and was first proposed by Swanson. He showed that he could predict that fish oil can be used to treat Raynaud's disease, based on an overlap of concepts such as platelet aggregation and blood viscosity. His findings were subsequently verified in clinical trials. We have been able to reproduce Swanson's study with Anni 2.0.

**MICROARRAY ANALYSIS** The original application of Anni was the interpretation of microarray experiments. These experiments typically result in a large list of significantly up and down regulated genes. Anni is organized in concept sets, and in order to analyze such a list, it must first be converted into a concept set in Anni. An important step in this process is the mapping of genes in the list to concepts in the ontology. For this, Anni can use the gene names, but as stated earlier, these are ambiguous.
Anni therefore also supports a wide range of unique gene identifiers, such as Entrez-Gene, UniProt, and Affymetrix identifiers. Once the gene list is imported into Anni, we can perform a cluster analysis. Genes with similar concept profiles will cluster together. It is then possible to investigate clusters, and determine which concepts are common in the concept profiles of the genes in a cluster.

The concepts that link the genes together can help to understand the biology behind the gene list. These concepts could be biological processes as defined in for instance the Gene Ontology, diseases or other important biological concepts. An important property of Anni is that it is always possible to go back to the original literature on which the associations are based. Researchers can check whether Anni’s findings are correct. Currently, the ontology behind Anni consists of the Unified Medical Language System, containing a wide variety of biomedical concepts, and our gene thesaurus with genes of human, mouse and rat. Other organisms are currently not supported.

**PROMISING TOOL** Because of the wide range of concepts, Anni can be used for many different purposes. In collaboration with our partners at the urology department of the Erasmus MC and the Center for Human and Clinical Genetics at the LUMC, we have already shown Anni to be useful for several applications such as micro-array analysis and drug discovery [3]. We envision many other applications of Anni to be possible.

### KEY CONCEPTS
- The large amount of biomedical literature seriously challenges the biomedical researcher to keep abreast of the current developments.
- The Biosemantics Group of the Erasmus MC developed a tool, Anni 2.0, designed to aid the researcher with a broad range of information needs. This work is part of the BioRange programme of NBIC.
- Anni 2.0 provides an ontology-based interface to Medline and retrieves documents and associations for several classes of biomedical concepts.
- The text-mining tool can be used for simple queries, such as: give me all the genes that are associated with ‘prostatic neoplasms’. Another typical application is to explore the associations between a set of concepts, such as a list of genes that were found to be differentially expressed in a DNA micro-array experiment. Anni can also be applied for literature-based knowledge discovery.
- Anni 2.0 uses a statistical approach to deal with the huge amount of literature data, and currently includes all 11.8 million articles published in Medline since 1980.
- Anni 1.0 could only be used for micro-array data interpretation. Anni 2.0 is designed for many other types of knowledge discovery as well, and has an improved user interface.
- Anni 2.0 is freely available to all at http://biosemantics.org/anni

### REFERENCES

### NOTE
The BioCreative (Critical Assessment of Information Extraction systems in Biology) consists of a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. (http://biocreative.sourceforge.net/)

### CONTACT
Martijn J. Schuemie
Biosemantics Group, Medical Informatics department, Erasmus University Medical Center of Rotterdam
PO Box 2040
3000 CA Rotterdam
Web: http://biosemantics.org
E-mail: m.schuemie@erasmusmc.nl

---

By the editors
Through the years Wikipedia has become a very well-known and much used free online reference resource for encyclopaedic information. Since it allows users worldwide to create and edit articles at any time the information is kept very much up-to-date. Researchers have discovered that Wiki-based platforms can also be very useful for sharing scientific information. WikiPathways has recently been created for biological pathways.

Biological pathways are critical in understanding the functions of individual genes and proteins in terms of the systems and processes that contribute to normal physiology and to disease. Pathways provide intuitive views of the myriad of interactions underlying biological processes. A typical signalling pathway, for example, could represent receptor-binding events, protein complexes, phosphorylation reactions, translocations and transcriptional regulation, with only a minimal set of symbols, lines and arrows.

Making these tools available for computational methods of analysis will allow researchers to connect pathways to databases of biological annotations and experimental data. Pathways data, however, presents a special case since the information is not directly coupled to data collection. One does not sequence or measure a pathway. Pathway data must be collected from a mass of biological information distributed across multiple publications and databases. Until recently there was no central platform where all the known data about pathways was collected and kept up-to-date. Researchers from the Department of Bioinformatics—BIG Cat at the Maastricht University and from the Conklin Lab at the Gladstone Institutes at the University of California, San Francisco have now developed such a platform, called WikiPathways. The initiators hope that the growing challenge presented by the influx of biological data is met with this platform.

WikiPathways provides an open, public platform dedicated to the curation of biological pathways. It facilitates the contribution and maintenance of pathway information by the biology community. Researchers all over the world are now able to contribute to keeping biological pathways data up-to-date. Each pathway has a dedicated wiki page displaying the current diagram, description, references, download options, version history and component gene and protein lists. The pathway diagram is edited with an embedded applet version of PathVisio. The Description and Bibliography sections can be edited in-page as well through applets that facilitate entry.
er the two groups maintain WikiPathways and develop software behind the platform for pathway analysis and drawing. Entering and changing data through WikiPathways is much more dynamic than the traditional way of submitting data in databases," says Kelder. "With a few mouse clicks one can change data immediately whereas in the traditional way you would have to submit and wait for the people who maintain the databases to review and enter your results." The developers are now working hard on improving PathVisio, the pathway editor. "The more user friendly the software, the more people will use it and this will eventually improve the data." Sometimes people may not want to enter their data since they haven’t published it yet. Kelder explains: “For that specific purpose we are creating a feature called ‘personal pathways’ which allows people to shield their results temporarily from other users. They may want to share these pathways with specified users. After the data is published, the pathways will be released to the public.” Furthermore portals for specific pathways have been and are being developed. Since the research field of pathways is very wide the initiators of WikiPathways enable researchers to create their own sub-communities.

THE DEVELOPER. “Most of the data about pathways is published in articles or admitted to databases or presented in PowerPoint slides,” says Thomas Kelder, bioinformatician at the Maastricht University and involved in the development of Wiki-pathways. Since this type of pathway information is often static and not amenable to computation or data exchange, not many colleagues can profit from it. Kelder explains: “We wanted to develop a program in which you can save your biological pathway data in a graphical form that can also be edited easily by other researchers. Also everyone should have access to it. That’s how we came to think of developing a Wiki-based platform for biological pathways.” At the same time Californian colleagues came up with a similar idea and so the groups decided to work together on WikiPathways at the beginning of 2007. Together they set up the Micronutrient portal which is being maintained by Suzan Wopereis from TNO at Zeist. Wopereis is very satisfied with the new possibilities created by WikiPathways. “It’s great that we can visualize all our (recent) knowledge on micronutrients in pathways pictures.” Wopereis studies the effect of nutrients on health. “We specifically want to know which processes take place at the onset of diseases which are connected with lifestyle, like diabetes type II, atherosclerosis and cardiovascular diseases.” TNO also performs dietary intervention studies with a focus on pre-diabetic symptoms. “By studying the metabolites, genes and proteins with functional genomics techniques before and during the dietary intervention, we hope to find biomarkers which may be indicative of certain diseases. Micronutrients play an important role in oxidative stress, inflammation and hormonal regulation, which are important processes in the development of lifestyle associated diseases” “With PathVisio you can create your own pathways under conditions you specify yourself,” says Wopereis. Commercially available software is expensive and presumes cellular measuring conditions, whereas metabolomics techniques, for example, are mostly applied to extracellular biofluids such as blood and urine. PathVisio is user friendly and it is freeware. It allows you to draw and visualize how for example metabolite X, metabolite Y and gene Z are connected. The pathway data created with PathVisio is easy to export to other programmes because it can be downloaded in different formats. Visitors can log in at WikiPathways after (free) registration and extend the data according to their specialized knowledge.” The Micronutrient platform started recently and therefore the group of users is still limited, but this will probably soon change according to Wopereis. “We are in different consortia, such as the European NutriGenomics Organisation (NuGO), Eurreca and the Netherlands Metabolomics Center (NMC), where we try to stimulate the participants to work with the portal. We have already noticed that the group of users is growing.”

THE USER. One of the recently created portals of WikiPathways is the Micronutrient portal which is being maintained by Suzan Wopereis from TNO at Zeist. Wopereis is very satisfied with the new possibilities created by WikiPathways. “It’s great that we can visualize all our (recent) knowledge on micronutrients in pathways pictures.” Wopereis studies the effect of nutrients on health. “We specifically want to know which processes take place at the onset of diseases which are connected with lifestyle, like diabetes type II, atherosclerosis and cardiovascular diseases.” TNO also performs dietary intervention studies with a focus on pre-diabetic symptoms. “By studying the metabolites, genes and proteins with functional genomics techniques before and during the dietary intervention, we hope to find biomarkers which may be indicative of certain diseases. Micronutrients play an important role in oxidative stress, inflammation and hormonal regulation, which are important processes in the development of lifestyle associated diseases” “With PathVisio you can create your own pathways under conditions you specify yourself,” says Wopereis. Commercially available software is expensive and presumes cellular measuring conditions, whereas metabolomics techniques, for example, are mostly applied to extracellular biofluids such as blood and urine. PathVisio is user friendly and it is freeware. It allows you to draw and visualize how for example metabolite X, metabolite Y and gene Z are connected. The pathway data created with PathVisio is easy to export to other programmes because it can be downloaded in different formats. Visitors can log in at WikiPathways after (free) registration and extend the data according to their specialized knowledge.”

The work in Maastricht is part of the BioRange programme, coordinated by NBIC.

REFERENCE

INTERNET
www.wikipathways.org

Wikipathways is committed to open access and open source. All content is available under a Creative Commons License.

SOURCE CODE
All source code for WikiPathways and the PathVisio applet is available under the Apache Licence, Version 2.0. You can download the code from: http://svn.bigcat.unimaas.nl/wikipathways or http://svn.bigcat.unimaas.nl/pathvisio

Hands on section

interface

issue 2 | 2008

17
FOR AN INCREASING AMOUNT OF DISORDERS IT IS BECOMING CLEAR THAT MANY GENETIC VARIANTS ARE INVOLVED. FINDING THESE VARIANTS AND THE AFFECTED DISEASE GENES IS LIKE LOOKING FOR A NEEDLE IN A HAYSTACK. BIOINFORMATICIAN LUDE FRANKE HAS DEVISED NEW STATISTICAL METHODS TO IDENTIFY SUCH VARIANTS. RECENTLY HIS WORK WAS CROWNED BY RECEIVING A PHD DEGREE, CUM LAUDE.

Although Lude Franke claims that the various types of research he conducts differ widely, the title of his thesis, and especially the word 'genome-wide' in it, covers much of the content. Most of his studies have followed genome-wide approaches. “Only a few years ago, genome-wide DNA oligonucleotide arrays became available. With these chips, we are able to assess single nucleotide polymorphisms (SNPs), deletions and duplications for hundreds of thousands of loci at once. Before, we could only assess a subset of these variants using rather time consuming PCRs and assays,” explains the young doctor.

During his PhD research at the Complex Genetics department of the UMC Utrecht, Franke has been working predominantly on data from these chips. Clinical researchers at various hospitals provided him with the data from thousands of patients suffering from coeliac disease and amyotrophic lateral sclerosis (ALS). Franke used DNA chips that compared patients with controls for no less than 300,000 SNPs. “Such studies have proven to be very valuable to identify associated SNPs, but also additional information can be extracted from these chips,” says Franke. "Deletions and duplications have turned out to be much more common than we had initially thought. We developed a method to genotype small but common deletions with considerable accuracy. I applied the method to
my own DNA, resulting in the identification of many deletions. The results of this analysis are printed on the sides of the pages of my thesis,” he explains.

**UNUSUAL APPROACH** The approach Franke took to identify these deletions is different from usual approaches. When no deletion is present three genotypes (AA, AC and CC) are possible for a SNP, but in the presence of a common deletion at the place of the SNP three additional deletion genotypes emerge. However, when intensity measurements are plotted, these genotypes form six overlapping clusters as these DNA chips had not been specifically designed to assess deletions. To overcome this, Franke reasoned that “nearby SNPs” provide information about the likely deletion genotype of the SNP under investigation. He used this “linkage disequilibrium” to improve genotyping.

“Through resequencing we corroborated that our method indeed can accurately assign deletion genotypes to these SNPs.”

Not only sick people carry deletions, as testified by Franke’s own genome. He thinks that paralogs are one of the reasons that those deletions do not necessarily cause disease. “A paralog is an evolutionary duplicate of a gene and often has a function that is comparable to the original. If the original gene is not present because of a homozygous deletion, the paralog can sometimes take over its function. We also found, by assessing the number of biological interactions these genes have, that genes that are often deleted generally have a less important biological function.”

These gene interactions also play a role in a different part of Franke’s thesis. He has predicted functional gene interactions into a network called GeneNetwork.nl. Interacting genes may encode for different parts of the same protein complex. Or one gene may be coding for an enzyme which metabolizes a protein from another gene. The third mode of interaction is a gene coding for a protein that influences the expression of another gene. Franke used the network to prioritize potential disease genes in loci that had been identified in linkage studies. A program called Prioritizer analyses these loci and determines whether some of the genes are potentially biologically related, as they are likely to be controlled by another upstream gene. Therefore devised a method to remove this genotypic effect. “We then observed a considerable increase in co-expressing genes, enabling us to identify more biological relationships, highlighting the relevance of genetical genomics. I am currently following this up as a post-doc at the Genetics department of the UMC Groningen and at Queen Mary, University of London.”

**META-ANALYSIS OF A PHD PERIOD** Apart from being a researcher, Franke is also graphic designer. His thesis, obviously designed by himself and his colleagues at CleverFranke, is packed with graphical gadgets. The dust cover in fact is a poster which not only cites a rather disturbing statement from a personalized genetic testing company (implying their results should be regarded worthless), but also depicts a word-by-word statistical analysis of the entire thesis. Inside, Franke’s period as a PhD student is graphically depicted through a network analysis, an analysis of how scientific papers develop and an analysis of his e-mail correspondence. “I like generating images from data,” is his simple explanation. “As there are no photos in my thesis, I wrote some software in the hope of creating a few interesting and visually appealing illustrations.”

**ADDITIONAL INFORMATION**

1. **Who is Rudi van Bavel?**

I was born in the south of the Netherlands, in the province of Zeeland, and I grew up in the far north, namely Leeuwarden. After the mavo I obtained my havo/mbo certificate and I worked for a few months before starting to study mbo systems engineering. I then worked for two years as a systems engineer, after which I completed the HBO (bachelor) course in bioinformatics. Currently I work at Keygene, a (bio)tech company with an ‘academic’ atmosphere.

2. **Why did you choose the bachelor study of Bioinformatics?**

It was fun but not really my calling. I did an internship and worked for two years as a systems engineer at a biotech company and found the biological part very interesting. In 2002 I heard that a new bachelor study in Bioinformatics had started at the Hanzehogeschool Groningen and I thought, ‘Why wasn’t this study there before?’ I decided to start and I became one of the first groups ever in this study.

3. **What did you learn?**

Everything was new, of course. Only the first two years were fixed and the curriculum in the third and fourth years was made up as we went along. Each year consisted of half biology subjects, including lab experiments such as preparing gels, and half IT subjects such as programming, simulations of bacteria growth or data mining. My final project was really a bioinformatics project. We created a website that projected the expression patterns of microarray data on images of biological pathways made by KEGG (Kyoto Encyclopaedia of Genes and Genomes). I still use the experiences gained during that period today.

4. **What does your job at Keygene entail?**

I started chiefly as a web application developer. I developed a sequence analysis suite for the quality control and visualization of Sanger sequences for example. This year I moved on to the lead discovery team where I needed to apply my biology knowledge as well as my IT knowledge. I develop analysis programs, for example with data from a 454 Life Sciences sequencer. This machine can sequence a lot of DNA in a short time but produces fairly short read-outs. My program is designed to extract as much information as possible from this data while integrating data from other sources and analyzing patterns.

5. **How does your work fit with your study?**

What I do now fits perfectly with what I have learned. I am a Perl specialist for designing web tools. Furthermore I learned a lot about Linux during my study. Sometimes I help the ICT group when they have a Linux problem. It is very nice that Keygene, as part of its innovation strategy, grants time for ‘hobby projects’. I study programming on video cards partly during office hours. This is potentially much faster than on regular processors because data processing is done massively parallel which is an advantage for the enormous amounts of data generated during sequencing.

6. **Who are your colleagues?**

The department is large and still growing. We work for all the departments at Keygene and also do commercial software development for third parties. I work with statisticians, biologists who have more biology experience and less IT experience and with IT specialists with little biology experience. People with an academic background are in the majority. They are more concerned with research and details while I am more practical. I also visit the lab so I know what is going on there. It makes it easier to build useful applications.

7. **Any tips for (future) bachelor students in bioinformatics?**

In my experience many biotech companies do not realize that we exist. That’s a pity. There is a lot of work in this field and it is growing fast. This is the ideal combination if you are interested in computers and also in the study of living things and biological processes. A little more publicity about the bachelor study Bioinformatics would help. The educational institutes should go to companies and even to high schools to introduce themselves. I would advise more advertising. Bioinformatics is the best of two worlds. It’s just so much fun!
SIB VISITS NBIC

The Swiss Institute of Bioinformatics (SIB) is one of the leading bioinformatics institutes worldwide. Founded in 1998 by pioneers such as Ron Appel (ExPaSy) and Amos Bairoch (SwissProt, nowadays: UniProtKB), SIB celebrates its 10th anniversary this year. The Institute now includes 250 bioinformaticians who work at selected universities and institutes all over Switzerland. A delegation of SIB scientists came to visit NBIC in Amsterdam on July 6 and 7 of this year. The aim of this meeting was to establish contact between the scientists of both organizations, exchange recent developments within both institutes and to identify topics for future collaboration.

Key scientific topics discussed were bioinformatics for proteomics and high-throughput sequence analysis, genotype-phenotype modelling, systems bioinformatics and e-bioscience/grid computing. In addition, the delegations exchanged experiences and ideas on bioinformatics education and at the organizational level. The meeting took place in a very positive and open atmosphere and it was immediately clear to all participants that there are major parallels and complementarities between the scientific programmes of SIB and NBIC. This gave rise to ideas for collaboration in projects and between the organisations.

Initial collaborations have started in the fields of proteomics, while project groups are being formed around other subjects. SIB service unit Vital-IT has serious interest in the NBIC BioAssist support programme. NBIC and SIB management have also identified a great deal of similarity between the institutes, in terms of their national focus, level of organisation and approach. A memorandum of understanding has been drawn up between SIB and NBIC to stimulate further interaction of scientists on both sides. Both institutes will stimulate the exchange of students and intend to link their education programmes. NBIC will facilitate these exchanges by making some budget available for students to attend SIB courses. There will also be some budget for researchers to visit their colleagues in Switzerland, e.g. for a short period of collaborative research. SIB, on its part, will make similar arrangements for its students and researchers to visit Dutch bioinformatics labs. SIB and NBIC will set up a series of seasonal schools, the first of which will be held in Switzerland in summer 2009.

www.isb-sib.com

NGI AWARDS

The Netherlands Genomics Initiative (NGI) has recently accepted the 25 M€ business plan which NBIC has formulated with the involvement of a great number of bioinformaticians in the Netherlands. NGI will contribute a significant 13.7 M€ to the NBIC programme for the period 2009 - 2013. With the available budget we can continue and strengthen the bioinformatics collaborations with many of the NGI genomics and technology centres. Projects are also planned to establish bridges to programmes outside the NGI.

www.isb-sib.com

NBIC 2009-2013

NBIC-cake at the kick-off meeting on October 31 to celebrate the granting of the NBIC-II programme.
network, such as Parel snoer, CTMM, Ti-Pharma, Ti-Food and Nutrition and Ti-Green Genetics. We are continuing essential collaborations with e-science centres VL-e and BiG-Grid through the further joint development of BioAssist support platforms. All together, scientists of 15 academic institutes and five industrial parties will participate in the new programme, which will yield new positions for about 20 PhD students and post-docs and 20 programmers.

Research projects (BioRange) fall into four themes: (1) Sequence-based bioinformatics (2) genotype – phenotype modelling (3) proteomics and (4) systems bioinformatics. The programmer positions are planned in the bioinformatics support platforms under BioAssist (see picture on page 23). Existing platforms focus on generic pipelines for (1) analysis of gene expression and structural genomics data (‘functional genomics’) (2) proteomics data (3) metabolomics data, and (4) clinical data (biobanking). Two new platforms will be set up to facilitate analysis of data produced with (5) the latest high throughput sequencing equipment, and in (6) systems biology experiments. Intensification and broadening of the current bioinformatics education projects (BioWise) is also foreseen. Here, the first priority will be to set up a national inter-university PhD school in bioinformatics. Many Dutch bioinformatics groups have already welcomed this plan. Up to 10% of the overall new NBIC programme budget (2.4 M€) will be focused towards dissemination and exploitation of the project results.

The new NBIC funding through NGI will give a fair new impulse to the Dutch bioinformatics field. It is already clear that the NBIC umbrella has started to lend a strong international visibility to the bioinformatics activities in our country.

---

**NBIC EDUCATION:**

**BIOWISE HIGHLIGHTS**

**RESEARCH SCHOOL**
The national Research School in Bioinformatics is a collaboration between bioinformatics professors of a number of universities in the Netherlands. An organization framework for teaching bioinformatics PhD students has been initiated based on two main PhD course areas: (i) The Technology Track will include machine learning, pattern recognition, fundamental statistics and information management while (ii) the Applications Track will include.

**NBIC RESEARCH:**

**BIORANGE HIGHLIGHTS**

**PREDICTION METHOD**
In close collaboration with experimental biologists, Aalt-Jan van Dijk and his co-workers developed a method for prediction of sub-golgi localization of transmembrane proteins. The collaboration started with the Theoretical Information group of Frank Neven at Hasselt University in Belgium, with the aim of further developing methods to search for sequence motifs in protein interaction networks. The applications of this method will aid the genome-wide analysis of interaction networks. This work has been published in Bioinformatics.

**VENI / VIDI GRANTS**
Three researchers within the network of NBIC BioRange have been awarded a VENI or VIDI grant from NWO: Aalt-Jan van Dijk (WUR, VENI grant linked to the work described above), Sander Nabuurs (RU, VENI) and Rainer Breitling (RUG, VIDI).

**PARTICLE PROFILER**
The importance of accurately measuring lipid levels in blood is well known. High levels of cholesterol in blood are known to increase the risk of a variety of diseases, such as atherosclerosis and related disorders such as cardiovascular events (stroke/high blood pressure), and diabetes. Daniël van Schalkwijk and co-workers developed a simulation method called Particle Profiler to predict lipoprotein blood profiles. The work has been submitted and will be published soon. An application of the developed software is part of a patent which has been filed.

**MORE WIKI APPROACHES**
In the previous issue of Interface we reported on the launch of WikiProteins by Dr. Barend Mons group (Erasmus MC/Leiden UMC). Community knowledge is not always...
stored in the public databases and WikiProteins provides a method for community annotation on proteins. Another Wiki approach was developed by Dr. Chris Evelo’s group, which allows community annotation of pathways. WikiPathways has been developed in collaboration with the University of California San Francisco. It was published in PLoS Biology (Pico AR et al., ‘WikiPathways: Pathway Editing for the People’). WikiProfessional and WikiPathways enjoyed broad international attention in Science (‘WikiPathways Debuts’) and Nature (‘Big data: Wikomics’ and ‘Molecular biology gets wikified’).

allows distribution of computing and storage resources specifically for the life sciences institutions in the Netherlands. Furthermore, all programmers introduced themselves and briefly explained their projects and expertise. At the second meeting, presentations were given by Machiel Jansen (Sara) on software engineering, Morris Swertz (RUG/UMCG) on Molgenis and by two scientific programmers on the progress in the proteomics platform. Both meetings were received well by the programmers. A lot of (inter)action has been created by this enthusiastic group.

As described in the previous issue of Interface, BioAssist is a collaborative e-bioscience endeavour in which participating biologists and (bio)informaticians contribute to organizing bioinformatics support for selected domains in the Netherlands. Within several bioinformatics groups scientific programmers are creating support platforms by building software based on generic technologies. The scientific programmers are trained monthly in using these generic technologies for storage, information management, GRID usage, creation of web services and workflows (using e.g. Taverna). One of the speakers during the first meeting was Jeroen Engelberts (BioAssist/SARA e-science support). He introduced the BioAssist scientific programmers to the Life Science Grid architecture. The Life Science Grid is the distributed component of the Big Grid e-science infrastructure, which

**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

[Contact information]

**EDITORIAL BOARD**

Antoine van Kampen
Ruben Kok
Marc van Driel
Karin van Haren

**COORDINATING AND EDITING**

Marian van Opstal
Bèta Communications, The Hague

**DESIGN**

Clever°Franke, Utrecht

**LAYOUT**

tedesign, Delft

**PHOTOGRAPHY**

Thijs Roosmans

1. [www.wikiproteins.org](http://www.wikiproteins.org)
2. [www.wikipathways.org](http://www.wikipathways.org)
3. [www.sciencemag.org/cgi/content/full/321/5889/623c](http://www.sciencemag.org/cgi/content/full/321/5889/623c)
“A great risk is that the interpretation of results will be guided by the model development. In a sense the structure of the model will determine the interpretation of the results. So basically the research will ‘prove’ the hypothesis because the model will be designed to prove it. A preferred approach would be to set out to disprove the hypotheses”, runs the referee’s comment.

Despite recent successes of bioinformaticians in the NWO veni and vidi schemes, these sort of referee comments can still be expected in regular biology calls, when biological questions are addressed with a combination of experimental, modelling and theoretical approaches, i.e. with an integrative bioinformatics (or systems biology) approach. The damage to the proposal was done, but the statement is of course ridiculous. Models are the best way to falsify hypotheses: if you can model it, you can understand it (but more of that later).

Hand-waving and anecdotal explanations of complex biological phenomena, for example multifactorial diseases, are great for describing the next knockout mouse model. It is also a way to go on forever – uh well, until roughly 6,000 knockouts in yeast or about 24,000 in mice; many of the double knockouts are now being done in yeast as well – and get it published in high impact journals (see Lazebnik’s hilarious 2002 paper on how a biologist would fix a radio, PMID 12242150). It is not the way to arrive at a real understanding of complex biological systems, and given the interest in bioinformatics, systems biology and now also synthetic biology, it seems that mainstream biology, and also industry, is becoming aware of this. So modelling should be, and is becoming, an integral part of biology.

But what about the referee’s presumption that good scientific practice should disprove hypotheses? There is some irony in this, as Popper based this preferred approach on physics, a science dominated by models. Physicists are actually spending billions near Geneva trying to find a particle that thus far only exists in models! The hypothesis-based paradigm has initially also condemned data-driven approaches in bioinformatics as inferior fishing experiments (with similar referee comments I suspect). I think these approaches have by now demonstrated their value, but also their limitations.

Getting back to the question. Some time ago there was actually an interesting philosophical discussion in Science about whether a model (or hypothesis) that can predict a set of data is better than if that model had been derived from the same data set (prediction versus accommodation of data, see Lipton 2005, PMID 15653494 and comments). Clearly the community is divided, but with a trend towards preferring the prediction. But the issue seems artificial: accommodation you do when you have the data and no hypothesis, prediction you do when you test a (data-driven) hypothesis with new data sets. It is an iterative cycle. The real challenge is to use as many of the old hypotheses (a priori knowledge, or ‘legacy data’) when doing new fishing experiments. That is what databases – and models – are for.