SWEGENE Bioinformatics Survey

Report on Current Status and Suggestions for Future Directions

Jari Häkkinen
Complex Systems Division
Department of Theoretical Physics, Lund University
Sölvegatan 14a, SE-223 62 Lund, Sweden
jari@thep.lu.se, http://www.thep.lu.se

August 14, 2002

Executive Summary

An account of meetings with people and research groups associated with SWEGENE is given. Conclusions are drawn from these meetings and used to write a tentative plan for how to proceed with a SWEGENE bioinformatics platform.

The main purpose of a bioinformatics platform within SWEGENE should be to provide computational facilities and logistical support [databases & infrastructure] for scientists whose primary research in biology and medicine benefit from computational approaches.

A secondary goal that should be sought by a bioinformatics facility is to make algorithms and analysis methods accessible for non-computational biologists.

Contents

1 Background

The increasing pace of research in the 21st century has expanded the need for, and the duties of, bioinformaticists. Scientists within life sciences are generating data faster than they can deal with it. At a laboratory level, this can be addressed with LIMS (Laboratory Information Management Systems), which allow for collection and processing of large data sets. Although there are many off-the-shelf LIMS solutions, customisation or implementation of these packages are usually required. At research level, another need must be met. Easy access to collected data from custom built software, flexibility of the data management system, (ideally) non-proprietary solutions to make data exchange between researcher and research group seamless, access to algorithms, ...

There are, in conjunction with the genomics platforms built with SWEGENE support, in general systems for the immediate management of the large amounts of experimental data generated. Most of the platforms are expected to be connected to resources for handling various types of user interactions, conceivably with a back-end database management system and a front-end web based technique to handle user interactions. Various software resources are also required to enable users to conveniently connect their experimental results with publicly, and application specific, bioinformatics databases and tools. In this way, every platform has common and specific needs.

The purpose of a bioinformatics platform within SWEGENE is to provide computational facilities and logistical support [databases, algorithms & infrastructure] for problem oriented scientists in an optimal way.

In the SWEGENE application to KAW and annual reports, readily available through the SWEGENE web site, a handful references are made that might lay out the general setting of this survey. The hopes in these documents are that a bioinformatics effort should cover a large part of the computational biology area, with islands of specific problem domains and very little discussion on how to coordinate a computational biology effort. The coordination might well prove to be impossible or not even desirable.

The first 18 months into the SWEGENE project, the fourth main effort within SWEGENE, bioinformatics, has not produced a proposal on how to meet the needs of the current biological problem domain, or how they should advance into the near, and long term, future research areas within computational biology. Some measures have been done to meet the most immediate needs; sponsoring of algorithm knowledge, database building for specific applications, some computer related acquisitions such as software and databases.

I was contacted to make a survey of the current status of the bioinformatics and computational biology efforts within the groups and researchers with connections to SWEGENE facilities. The idea was also to propose a framework where the SWEGENE platform users would get computerised support for their research projects.

This report is divided into four sections, beginning with this introduction concluded by quotes from SWEGENE documents below. This section is followed by discussion and conclusions from my meetings with different researchers and groups leading to the section on suggestions for action to be taken within a SWEGENE bioinformatics programme. Interested readers have a chance to read more detailed notes from my meetings in Appendix A

Excerpts from SWEGENE Documents

I include this section of statements found in the SWEGENE application to KAW and the first two annual reports in order to provide a feeling of the computational biology visions within the SWEGENE project.

Apropos attracting people and create interaction

"... to create a truly attractive high quality research environment in bioinformatics within the SWEGENE initiative."

"Modern information technology will be used to reinforce interaction within the SWEGENE network."

"Chalmers University of Technology will be allocating 200 million SEK in the field of bioscience over a five-year period, starting in 2000, part of which is Chalmers contribution to the 220 million SEK jointly contributed by the three collaborating universities (see above). The research programs included in the final evaluation process at Chalmers are tentatively titled "Bioengineering", "Bioinformatics", "Biological Physics", "From Gene to Product", "Structure and Function" and "Tissue Engineering". Large parts of or the complete content of these programs are directly related to and synergistic with the SWEGENE program."[1]

[SWEGENE Affymetrix, Lund] "Ann-Sofie Albrekt, employed in October the same year, is responsible for bioinformatics support as well as for helping customers with experimental design and data analysis."

"Bioinformatics is an essential component in all research programs supported by SWEGENE. One of the key problems facing groups involved in these programs has been the scarcity of qualified scientists in this field. Largely as a consequence the allocations to bioinformatics within SWEGENE have been of a different nature from those in the other areas of activity."

Apropos database building

"... the development of new automated on line database searching algorithms which allow stringent identification of post-translational modifications. It is important to automate all steps and integrate these into one platform to allow very high throughput sample analysis."

"There is also a rapidly growing yeast expression database. Building a core competence in this area is important not only in itself but also to support expression data modelling. We aim at fusing these activities with already existing bioinformatics databases and WWW-servers in the South Western region."

"Several biological databases are based in the South Western region[2]. One of the most prominent is RATMAP, ..."

"In addition SWEGENE has supported a full time data base curator at the rat genome database, RATMAP."

"A microarray database, which complies with EMBL standards, has furthermore been developed."

[1] The final plan differs substantially from this statement.

[2] I have only found RATMAP on the Web.

Apropos software/analysis tools

"A collaborative network with bioinformatics research groups should be created to support linkage analysis. Given the problems with defining the level of genome wide significance, there is a need to simulate the data before analysis. This requires access to super computer capacity. The tools of genetic bioinformatics are still rather immature. It is therefore extremely important to have outside collaborators in this field. An exchange programme allowing statisticians to go to centres abroad is essential."

"Much of our research efforts will be in the second wave bioinformatics, where emphasis is on modelling cellular processes in detail and analysing different large scale expression experiments in order to understand basic biological phenomena and/or human disease pathogenesis."

"This particular field is large and many different modelling levels and techniques are likely to be important e.g. statistical inference techniques, differential equations, stochastic differential equations, artificial neural networks and Bayesian networks."

Apropos life after this

"The universities will also take responsibility for the continuation of the programme after the initial five-year period."[3]

[Proteomics, Lund] "An essential component, the Laboratory Information Management System (LIMS), will be delivered in the spring of 2002."

2 Survey Conclusions and Discussion

A more detailed account of my meetings with different groups and people is available in Appendix A. Here I discuss recurring issues.

Overall Impression

I met about 20–30 groups/people during my visits to different departments in Göteborg, Lund, and Malmö, and the lasting impression was that there is an underestimation of the power of computers for data management and analysis among many non-computational researchers. There are exceptions from this, but the majority of the groups need to become self reliant vis-à-vis biological computer knowledge. People seem to feel "bioinformatics stress", and do not really know what to do.

Another impression was that the bioinformatics need for many biologists are very basic, far beyond what I expected starting with the survey. This could partly be explained by the fact that not all people I met really need advanced bioinformatics tools. Along this line I also got the feeling that the junior members of the research projects were responsible for computer related issues, and that senior researchers seemed not to take part in the new

[3]Comment: This might be simpler to do with programmes within one university, whereas inter-university group programmes such as a bioinformatics facility might be harder to maintain due to the need of joint funding (i.e. politics) after an initial SWEGENE funding.

tools and thus become dependent on younger staff. This need not be negative, but usually the younger staff is moving on, and in consequence, remove intangible knowhow (visions, implementation details).

Computational Biology and Programming

Some groups have started joint projects with research groups with algorithmic, analytical, and computer skills, while others are too late into the game. Too late meaning that they will have a hard time to find skilled people that have the time to take part in another project. This was especially noticeable during my visits in Lund and Malmö, where much hope is put on the Complex Systems Division at Department of Theoretical Physics in Lund. Complex Systems has of course limited resources, and it is questionable if everybody should look to external help for analysis of data. They should rather learn themselves.

The scarceness of analysis knowledge can only be solved through education of computational biologists, or helping problem oriented researchers to reach reasonable knowledge level of essential computational biology. These need not all be very advanced algorithmically but should possess a knowledge of what can be done with computers especially since much of the currently needed computer skills are relatively basic. Until this new breed of biologist emerges en masse, there are needs for analytical services.

So, this boils down to supplying analysis services to biologists, and this creates a problem. The algorithmic, and computer literate, researcher does not want to (and, really, cannot career wise) supply this on a service basis, but rather as full members in research projects. However, currently there is not enough computational people to sustain this need.

This problem might be solved on a short, and mid, time range basis by external contractors for pure computer related problems, and in this way minimise the efforts of the analysts. The services provided by the contractors should be payed for by the users. The problem for the contractors is then that they might only get short term assignments and might prefer another commission. This could be solved by SWEGENE guaranteeing a minimum number of assignments to the contractors.

In the long term, the biology departments must realise that every serious bioinformatics interested department must accept the consequences and have at least one person that is responsible for the development and maintenance of software packages of general (departmental) interest. This person should also keep the documentation up to date and introduce new members of the staff to the available bioinformatics facilities. This is especially important for short term employees and master thesis students.

There is another synergy effect of using external contractors. In the currently functioning projects where biologists and analysts cooperate there are new methods developed that need to be made accessible for non-computational users, i.e. researchers that cannot create computer programs implementing new algorithms and ideas themselves. This is sometimes done by project members, but is often neglected. Here professional programmers could be used to perform implementation of algorithms usable for a wider audience.

The long term goal of a SWEGENE bioinformatics venture should be to create an infrastructure to support future use and development of the outcome of the initial funding. The

primary goal should be to create a platform to be used by bioinformatics, and problem oriented, researchers, and the secondary goal should be to create easy access to the infrastructure. These two goals are, fortunately, not mutually exclusive and can both be met with proper funding.

Current Bioinformatics

The number of bioinformatics tools and services available for a wider audience is small. Appendix A.3 describes a few tools or services I recognised that provide a service for a wider audience (than their own projects).

The polygenic centre in Malmö [cf. A.1.5] has plans of giving their users extensive computer infrastructure support and the polygenic centre [cf. A.1.4] in Göteborg is also pending on the proposition from Malmö.

Other groups have solved the computer infrastructure needed to give service to their users, but most of them have no extensive plans to support (computer wise) their users when they leave the facility. Their plans are to poorly developed and vague to be considered here.

Of the different tools and services available today, RATMAP and BASE seem to be the more developed ones.

Facility Building Advice

Facilities planning to give more extensive computer support to their users must understand that this will take resources into account, and take these costs into account when planning their support policy. Convincing users to invest time and resources in a badly funded system will result in bad-will for the facility (and SWEGENE), cf. A.1.7 the GCG/SEQWEB package is slowly dying, partly because of funding issues.

Obviously, facilities should also try to streamline data management, and begin to do so in an early phase of their development. Service facilities should also forsee what the users wants to do with their data when they return to their respective department, and build in support for this into their computer service.

Communication

Several groups felt that they need a work group communications tool, that could help them to assemble knowledge and information about the project and create a "project memory" available for all participants. There surely are commercial packages that addresses this need, and I know at least Wiki Wiki Web[4] that is a free package that allows people to communicate information through a collective maintenance of shared web pages.

There have been a few suggestions to create a help desk facility that could be used to communicate bioinformatics knowledge. I personally have problems to see how this can work, how to find skilled people that is willing to work within a help desk function. I am

[4]http://c2.com/cgi/wiki?StartingPoints

somewhat biased here, but my personal experience with software vendors help desks are not too positive. Communication of knowledge can only occur on a peer-to-peer basis, or attending classes.

3 Recommendation

There is a need to solve many short and long term issues within bioinformatics and computational biology as discussed above. Assuming limited resources, we cannot expect to solve all of them, some of the problems can only be solved by programmes such as SWEGENE, WCN, or Chalmers BioScience[5] while others need to be addressed by government and university policies. In this section I discuss a number of concrete measures that should be taken within a SWEGENE bioinformatics programme.

The main purpose of a bioinformatics platform within SWEGENE should be to provide computational facilities and logistical support [databases & infrastructure] for scientists whose primary research in biology will benefit from computational approaches. Another goal is to create a facility that makes tools and algorithms accessible for non-computational biologists.

Here an outline is given of a scenario that assumes a group of 4–5 people to maintain and implement such a platform and facility in a reasonable time. More resources will allow one to address more issues, while less resources will, at best, prolong the implementation time.

3.1 The Vision

The first stage of a bioinformatics facility implementation is a three fold project. First, and fairly straightforward, it should create a in-house database repository of common and specialised databases. This is readily handled by Sequence Retrieval System, SRS[6] (or a similar tool). The argument to create this locally, as opposed to use available facilities such as EMBL, is to build bioinformatics knowledge and to get access to the databases through application programming interfaces (APIs).

The second leg is to create, maintain and curate, (one or more) databases of general interest. This could attract research people with interest in creating databases for their needs. One application that seems to be very attractive to build is a pathway (biological network) database. With a computer readable pathway database at hand, this would have utility within microarray and proteomics and beyond. A pathway database should also form the basis for future research within, as well as outside, the facility and become the facility profile for the biological science community.

The third part is a pure service function and should facilitate computer infrastructure and support for projects within the SWEGENE umbrella that is to be published as web services or need database connections. Research groups should also have a possibility to suggest bioinformatics (computational biology) projects and contribute to the bioinformatics facility. An example of an interesting project is BASE [cf. A.3.1]. BASE is an application built to

[5] http://www.bioscience.chalmers.se
[6] http://www.lionbioscience.com/solutions

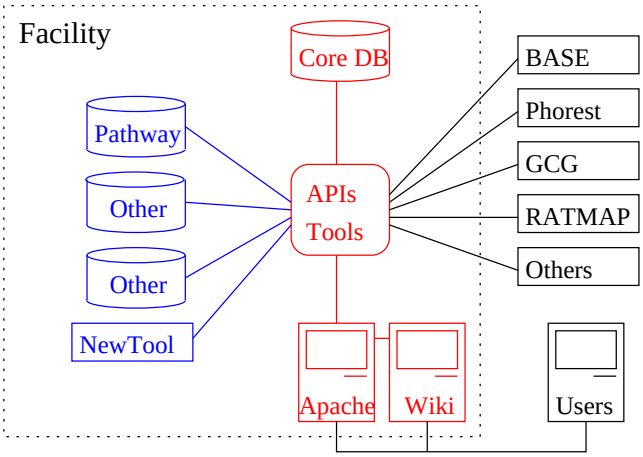

Figure 1: Organisational plan is indicated with colours and the bioinformatics facility is indicated with the dotted box. The red parts indicate the bioinformatics core responsibilities, the blue corresponds to in-house R&D, and the black to external users. The bioinformatics core is responsible for maintenance of support databases, APIs, and tools, but also a web server front-end and a notice board facility. All users connects to the core through the APIs or the Web.

meet a research group's internal data management needs but has a more general audience than the group itself. The project is now growing into a full-fledged software development project, and further maintenance should be supported by others than the original creators. In this way the researchers can go on to other projects, and the application can be further developed for the scientific community (of course assuming that the application is interesting enough). This sort of exit for researchers could create a more positive attitude toward technology transfer of research results.

A properly maintained and concentrated effort would create a web site whereto people could return to use, and get information about, available tools. However, if a scenario as described above is to work, non-research staff is needed to maintain the computer infrastructure whereas scientific staff should coordinate the efforts.

The vision outlined here is pictured in Figure 1.

3.2 Organisation

To reach Shangri-la skilled man power is needed, but there is no real organisation at place today and a new organisation must consequently be put together. Required knowledge is bioinformatics, software development, biology/medicine, and algorithms.

Software developers are needed to maintain, develop, and implement databases, create analysis tools and web services. The developers should also be a part of the approved projects suggested by external and facility internal parties. Software developers should be compared to laboratory technicians hired to assist researchers in the experimental situation. Software developers *are* the laboratory technicians of bioinformatics, and as such they are to support people with their computer (programming) related problems.

Recruitment of software developers is expected to be successful since these positions offer challenging and interesting problems to solve. Working in an environment with algorithm, database and web ingredients will give developers good prospects for future careers.

The facility should be coordinated by a scientific steering group. This group should make decision and priorities for the software development and the research team, and provide the facility with scientific guidelines, international contacts and collaborations. The facility should be run by a scientific management that is to execute the visions and requirements of the steering group. The steering group and facility management should also scout for interesting projects for the facility to take part in. The scientific management positions are expected to be attractive for people who wants to be a part in shaping, and contribute to, the bioinformatics community as a whole and the future research projects within the facility.

3.3 The Route

The following is a short sketch of how to proceed to create a facility outlined in Figure 1. The organisation should meet the chosen plan, and speed, for the creation of the facility.

Short Time Range Actions [Expected time span is 6 months]

Construction of bioinformatics core structures. The aim of these are to support external and internal users in development of more complex tools and databases.

Create documents, these should include howtos, training, tools, why, information.

Thorough examination of a basic curated pathway database and further investigation of available pathway databases.

Setup a Wiki Wiki[4] web site for the interested research groups. Wiki is a system that enables users to change content of a web site, and in this way information could be shared between researchers. The content should be maintained by the users themselves, whereas the computer infrastructure is supplied by the facility.

Setup information on how biologists can suggest projects of interest for them. These should be projects with expected general interest such as BASE, PHOREST, and RATMAP. Furthermore, these projects should be driven by the ones suggesting them with support from the bioinformatics platform.

Mid Time Range Actions [6–12 months]

The core system is up and running.

Create and add APIs for advanced users. Setup necessary computer infrastructure, and begin adding tools and documentation.

Collect interesting local projects into an umbrella organisation where active and well maintained projects are welcome.

Expand pathway database. Connections to tools such as BASE and other tools that can utilise pathway data.

Workshops and seminars arrangements to market the facility.

Long Time Range Actions [>12 months]

The facility is fully up and running. Research can now fully be performed within the facility and in cooperation with external partners. Bioinformatics core continues to develop the system, adding features requested by users.

Pathway database accepted by science community. Bioinformatics research utilising the database. More tools added and new projects started.

Continued marketing through workshops and seminars.

3.4 Investment and Time Estimates

Even though this is not a proposal for a bioinformatics facility I provide some rough estimates of funding needed to pursue the plan outlined above.

The largest expense in a bioinformatics project is related to personnel, but funding is also needed for computer hardware, databases and computational machines. I have calculated costs for three different tentative bioinformatics facilities. The estimates are compiled into Table 1 covering personnel, travelling budget (75kSEK per facility manager and year), and computer infrastructure (300kSEK averaged per year).

Minimum cost scenario, MCS. This translates to the longest implementation time of the above plan. However, it is questionable if this kind of funding will create a sustainable, and a competitive, bioinformatics facility. No real support of external request can be offered.

Time optimal scenario, TOS. The aim of this scenario is to reach the long time range goals well within the SWEGENE funding time. Support of external project requests will be available after an initial start-up phase.

Maximum service scenario, MSS. This is basically the same as TOS, but with an added service level, and a divided facility between Göteborg and Lund. In this scenario as much as possible of the conclusions [cf. Section 2] are fulfilled.

Scenario	Management	Developers	Total annual cost (kSEK)
MCS	0.75	1	1800
TOS	0.75	4	3400
MSS	1.5	8	6600

Table 1: Running cost estimates for tentative facilities, see text for a description of the scenarios. The facility management is not fully supported by the facility to allow them to pursue research projects. This will enable seamless interaction between developers within the facility and researchers.

A Survey Notes

The survey is divided into several parts; i) an investigation of the current status of bioinformatics at the different SWEGENE facilities, ii) a survey of the expectation from users of the SWEGENE facilities, iii) a description of tools and services developed within groups I met.

The survey covers people and groups who voluntarily agreed to meet me during April to June 2002. No real effort was made to force meetings, and this is somewhat reflected in the fact that I meet more groups in Lund/Malmö area than in Göteborg and, thus, have a geographically uneven sampling of meetings. Judging by the meetings, I would say that Göteborg has, in general, conceptually embraced bioinformatics more than Lund, This can maybe be explained by the larger number of meetings in Lund giving a fuller spectrum of bioinformatics development in Lund.

The quality of the meetings varied a lot, from fruitful discussions about current and future needs to a situation where people tried to argue for that their needs should solved by a SWEGENE bioinformatics project.

A.1 SWEGENE Facilities

A.1.1 Automatic Sequencing of Non-Model Species Facility, Lund

This is a facility hosted at Microbial Ecology, Lund University, and is a late entry into the SWEGENE project, but is expected to be up and running in early autumn 2002. The facility is to provide EST sequencing for non-model organisms. There is software, PHOREST, to aid the researcher with storage and annotations of the ESTs. The group has used their annotation software over a longer time, and has thus created a curated annotation of EST sequences involved in *mycorrhizal symbiosis*.

The users of the platform will not run their sequencing themselves. Samples are to be handed in, and results are returned in some format to the user (a file containing FastA sequences, maybe clustering). PHOREST support will probably not be included but this is an issue concerning the SWEGENE platform to resolve. Should the platform service be extended (from sequencing only) to give database support for its users? This will create a cost for maintenance and should be budgeted for. Furthermore, there is no real hard-core computer software/hardware knowledge currently within the group after Dag Ahréns (a former Ph.D. student) departure from the group, i.e. currently there is no bioinformatics knowledge at the department.

Ahréns departure also affects a recent research track within the facility. The group has started to run microarray experiments using these ESTs, and found a need to connect the BASE [cf A.3.1] microarray software with PHOREST. The group faces problems with this since the developers (Ahrén and Carl Troein) of PHOREST has left the group. The developers are still physically in Lund, so a knowledge transfer is possible but should be solved as soon as possible.

The feel within the group is that the hardware, experimental setup, is not the real problem,

but rather how to deal with data, especially from non-human and model organisms. Data analysis and visualisation is expected to be a bottle-neck.

Suggestions

One direct suggestion on what should be done is to create some kind of forum where ideas and needs could be exchanged such as discussing creation of new applications such as PHOREST.

Another idea is to setup a pool of bioinformatics service people. These should be available, at cost, for needing groups. SWEGENE should guarantee the salary for the pool. A pool of this sort is suggested to be located at Complex Systems Division at Department of Theoretical Physics in Lund. It is expected to be hard to assemble a pool like this and one way to do it might be to allow these people to do part-time research and paying competitive salaries. The research part is important since competing with salary, for the needed competence, is probably out of the question.

SWEGENE should fund more service people.

In general there should be examples of what can be done experimentally with the different SWEGENE platforms, what can be analysed, how to do analysis connected to the platform. Dag thinks one important thing with a bioinformatics platform would be to inform people of what the different SWEGENE platforms do and why, the analysis software they use.

Tunlid believes in customised databases for the experiments and research you want to do. This of course requires the users to design their experiments and databases before actually running the sequencing. There will be a need for database expertise, should this be provided by SWEGENE? An expert should probably help with the experiment designs at least initially since proper input is needed before experiment design.

A.1.2 cDNA Microarray Facility, Lund

The SWEGENE service part is expected to be up and running sometime after summer. The users are going to be introduced to the facilities and are expected to perform their experiments themselves. The Oncology Department has BASE [cf. A.3.1] as their bioinformatics resource, and consequently experiments are going to be fed into BASE, and the users can continue using their BASE accounts from their home sites. The users will also have an option to get their data on a CD. However, the feeling is that a CD with data is not going to be sufficient, but rather more computer support is needed and this is hopefully accomplished with BASE.

Suggestions

There was a positive attitude toward having (university) local copies of databases (NCBI and others), and create links to these databases in BASE. This to reduce the number of failing points. Another idea discussed was to create some uniform proxy solution to access databases in a transparent way (for the application programmer). Someone still needs to do the proxy implementation.

A.1.3 Microarray Resource Centre, Lund

The SWEGENE Microarray Resource Centre[7] is dedicated to Affymetrix[8] microarray technology. The centre was started in 1999, but was not accessible for SWEGENE users until early 2001. The staff is 3–4 people.

Several software packages are currently in use at the microarray facility; three different packages from Affymetrix, GeneSpring from Silicon Genetics[9], Spotfire[10], and Clonedike [cf. A.3.2]. External database used are LocusLink, Medline, and more through the software packages in use at the centre; www.netaffx.com, Spotfire, GeneSpring.

The facility itself acknowledge a problem of managing their data. They are in a desperate need of a database application to manage the output from experiments. The centre feels that this is their most important problem to solve currently, and are dependent on outside help since Immune Technology has no programming skills in house. Software vendors has very high license fees for LIMS software and the centre would prefer to use free (costless?) software solutions. The centre was not aware of the BASE application developed within SWEGENE.

Users of the facilities get a CD with their data and a copy of Affymetrix Microarray Suite (one (basic) part of the three packages in use at the centre). If the users need more software they must acquire it themselves. The people I met at the centre said that the users seem to be satisfied with the provided support. However, during the discussion it become clear that the users are dependent on extensive support, i.e. analysis and research support. The users do not seem to be able to treat their data. The centre bioinformatics support actually performs the analysis and provides the users of list of genes. The users then make an effort to biologically analyse the lists (gene by gene), but usually comes back for more support. People at the centre interpret the users data, and the user makes some kind of biological interpretation effort. This situation cannot be acceptable. The users have little prior knowledge so the facility should guide them initially so that they can avoid the mistakes made due to inexperience. Education of the users is needed, work shops, lectures and meetings are provided. The users learn but there are new users dropping in all the time. There is a high degree of bioinformatics stress shown in the discussions with the centre. So far everything has been done manually but the users expect a more professional attitude toward data management.

Suggestions

There should be a resource for design of experiments – users have done all the possible mistakes before they come to the facility and they don't understand the use of doing more than two chips (this is probably a budget issue since one chip costs approximately 1000–6000SEK).

[7] http://www.immun.lth.se
[8] http://www.affymetrix.com
[9] http://www.silicongenetics.com
[10] http://www.spotfire.se

A.1.4 Profiling Polygenic Diseases Centre, Göteborg

Tommy Martinsson is probably becoming the manager for the Göteborg profiling polygenic diseases platform with Staffan Nilsson as part time statistician. My feeling after meeting them is that Staffan Nilsson has the computer visions, and is involved more than as statistician only.

The hope was to get the facility up and running early in the autumn, but the deadline cannot be met due to delays. The new deadline is set to the end of 2002. The people involved in the project seem to have a clear vision of what they want to accomplish, but some computer/data management issues are unresolved.

The centre is aiming for a fully automatic (high throughput) system for genotyping of SNPs, and is going to let researchers run their samples themselves.

The facility is to cooperate with its sister organisation in Malmö, but there are some question marks regarding the computer facility cooperation. The Malmö centre is pushing two solutions, one where Göteborg should use the computer facilities in Malmö, and one where Malmö helps Göteborg to acquire a license from BioComputing OY in Finland for their database and data analysis needs. This makes a confusing impression, and the discussions progress slowly. Göteborg is going for one of them, and this will probably be the StatGene package provided by BioComputing OY (if the licensing cost is acceptable).

Discussing the polygenic diseases projects in Göteborg and Malmö is cumbersome since Malmö shows a dual face to me and the Göteborg group. My interpretation is that Malmö has not really defined what their service level would be. Depending on the service level offered, Göteborg must decide what to do for their data management.

Before they decide on whether to use the computer facilities offered by Malmö they would like to see a proof-of-concept. However, if the progress is too slow, another solution will be implemented.

A LIMS is needed to keep track of samples and the data flow in the laboratory. This is unresolved so far since the machinery is not decided for yet (due to delay in funding). There is a LIMS product at Thermo LabSystems[11] (Nautilus) that has a reasonable cost around 100kSEK per user.

Computer Skills

The group has only one programmer and uses mainly statistics programming languages such as R. People work in Excel sheets and converts these into tab-delimited text files (and a fair amount of time is used to support people with Excel).

The users of the facility will have problems analysing their data since they have poor theoretical knowledge. A couple of standard analyses will be provided but this is not going to be enough.

[11] http://www.thermolabsystems.com

A.1.5 Profiling Polygenic Diseases Centre, Malmö

An ongoing project to assemble phenotypes (proteomes, transcriptomes, physiology) and genotypes (SNP, micro satellites, DNA sequences). The phenotypes and genotypes are linked together with data driven algorithms. The group is approximately 30 people whereof three are computer specialists.

There are two different databases in use at the centre that solve the same biological problem.

The centre thinks that their users have a clear picture of what they want to accomplish, but this thinking is somewhat contradicted by the fact that they are worried about how to meet more advanced needs from their users. An initial basic support need is expected, but the expectation is that this need will decrease over time.

Somehow there is a discrepancy between the users needs and the service the centre wants to offer. Actually the service offered by the facility was not entirely clear and there was two modes. (The service offered has been described in a document that hopefully is a compromise of the two modes.) Holger Luthman seems to want users to suggest project and then to cooperate with the interesting ones. The vision is to create cooperations/joint projects with their users, and participate with algorithmic input, otherwise the users must manage themselves. Holger has a clear vision of what he thinks should be implemented, although Leif Groop understands that users might want to use the facility without always joining a project with people from the centre, and is afraid that if everyone must participate in joint projects, the users might feel they loose too much control. Leif also acknowledges the need to educate people and give them more service than just the hardware platform (and the planned database service).

The implemented database solution will be available (late 2002 or early 2003) for other groups as a part of the platform services. One idea is to get users to use the database facilities in Malmö, the benefits are obvious; users do not need to maintain the system and security, the centre will have a strong case for future funding to sustain the system, the users will interact with the system through web browsers. Another solutions is that the users can install the systems for themselves, but must then maintain their systems themselves with very little support. This will, on the other hand, give these users better control.

There is a number of programs that users want to run. It is expected that most of them will be installed within the centre and available for their users.

There is now a document defining the bioinformatics support to be expected from the facility. The support offered is quite ambitious. Highlights from the document:

- The SWEGENE Resource Centre for Profiling of Polygenic Diseases in Malmö will be possible to access from a home page with information about services provided by the centre and how to contact the centre.

- The databases and analysis tools (programs) will be accessible with a conventional browser interface via the home page.

- The centre will provide an introduction to the use of the databases in the format of one-day courses and, in addition, provide help-desk support to solve everyday problems.

- All projects have to appoint a contact person, who is responsible for the dissemination of this information within the projects and for the introduction of new project collaborators in their use. Via the same project accounts users will get access to the genetic analysis tools (programs like Genehunter, Mapmaker, Linkage, etc.).

- The use of each analytical tool is introduced on the home page to guide users during their analytical work-up of the data.

- The aim is to allow direct interfacing with analytical tools, to be able to analyse genetic data without the use of the databases provided by the centre.

The facility also realises that an extensive support as described above will need personnel resources.

There must be a well defined basic service level for the users. This should include pre- and post-experiment support in addition to the more natural laborative work. An example of post support would be to offer a set of more or less standardised analyses, bioinformatics support. A typical pre support task is to allow users to ask how to design experiments, if their questions can be answered with the services offered, pointing out how to *not* do experiments, creation of tutorials.

Suggestions

SWEGENE should setup/support a contract with an external party that can provide programming knowledge and skills to projects. SWEGENE should also educate and support people in programming and in bioinformatics.

Expected needs

There is a need to facilitate for (micro)array data, but the current databases cannot do this. Could BASE be utilised for this, with links between databases? There are similar problems for proteome/ics data.

Luthman feels a need for some kind of communication facility where groups can store data and information that might be relevant for others in the group. Basically a work group tool is needed.

A.1.6 Proteomics Research and Development Centre, Lund

The centre aims at providing "access for the large medical, natural science and technology research communities to resource centres for medium-throughput protein analysis and identification. This encompasses the whole range of techniques from sample preparation to computer analysis of the results." (Quote from SWEGENE web site.)

The customers will not be running the mass specs themselves, but might be allowed to do the sample preparation. There are a few workstations available for analysis of data on site.

The throughput is expected to be 50 gels per week. Every gel is 4x20MB. You choose approximately 100 spots from each gel, resulting in 200MB of MALDI mass spectra data per gel. Of these 100, approximately 20 are expected not to be identified and must be further analysed on MSMS which will yield another 15–20MB of data. This adds up to 300MB per

gel, and a total of 15GB of data will be produced per week. The customer however will only get tiff formated images of the gels and the identification results on a CD (approximately 10MB per gel).

Expected needs or wish list

The customers need further support when they leave the facility since they have a CD with a few hundred proteins they need to work with.

Two packages would be useful for the facilities customers; a) GenoMax from InforMax[12]. This package is expected to help customers automate common task through GenoMax protocols. There is also a vague promise (from the software vendor) that customers can add their own modules to GenoMax. There are, however, some question marks about the economic stability and company control for InforMax. This will of course have impact on the development of the software and it is rumoured that InforMax should consolidate its current products. b) A text based data ripping algorithm is needed such as Virtual Adapt from Virtual Genetics[13].

Suggestions

The centre thinks that GCG is a proper tool to use in the analysis of the results.

Peter James thinks we should study Lion BioScience, maybe get a deal with them, where they help us (for free?) and they can use us a showcase. SRS alone is too basic.

Two kind of support programmes are needed; a) A research level help-desk where people can get help with analyses. This should of course be in project form. I have some problem seeing this, people should probably find their partners themselves, but maybe a list of competences could be useful. b) Tutoring is needed for the different software packages and databases (InforMax, Virtual Genetics, home brewed, ...) otherwise it is a waste of money and time.

A.1.7 Structural Biology, Lund

Met Guoguang Lu from Department of Molecular Biophysics. Lu is not supported by SWEGENE but is a part of Anders Liljas group.

Lu was in the process of sending a document to SWEGENE describing, and requesting more funding for, his bioinformatics server. This is a short recap of the document. There is a server for nucleotide and protein sequence analysis tools for local research and teaching activities. This server runs a GCG/SEQWEB package with 20 frequent users from local researchers. The number of users was previously larger but is declining. However the number of users is expected to increase if the server is properly maintained, training is provided, and the service endorsed. The document also lists a number of tasks that will be performed if SWEGENE allocates proper funding. All tasks, except one, concerns the GCG package. The exception concerns a new alignment technique developed at Molecular Biophysics.

The reason for the decline in the number of GCG users (from approximately 100 to 20) is that there is a lack of resources, and no resources for training. The resource problem cause the system to become obsolete (no software or database upgrades), and in consequence people leave the GCG service. The software is currently upgraded with Lu's own funds. Lack of

[12]http://www.informaxinc.com
[13]http://www.virtualgenetics.se

training is a natural consequence of lack of training facilities.

Lu has web service where protein structure alignment is offered. The algorithms used are quite old, but in near future a new service will be opened. This utilises sequence and structure information for alignments, and will in the future lead to an algorithm for protein folding prediction.

Expected needs

New hardware for the web services is needed.

Suggestions

Lu feels that there is a need for a general bioinformatics service. He actually gets requests but cannot respond due to lack of resources.

There should be specific SWEGENE funding for training. As it is now, buying equipment is possible, maintenance somewhat funded but no real training budget. Also students should get full access to facilities. This will lead to a larger user base and better economy after SWEGENE funding is ended.

A.1.8 Structural Biology Platform on Membrane Proteins, Göteborg

The facility is working with membrane proteins, and are currently setting up the facility. The facility is aiming in unique experimental questions and has no real references to use in their build up. Hardware has been bought and the platform is hiring lab researchers. The facility is inviting different research group to take part in the creation of the site, and several groups has responded positively on the invitation.

The facility specialises on a few membrane proteins, and is aiming for in depth knowledge about the ones chosen. This is in contrast with a WCN effort that tries to make high throughput work on membrane proteins.

There is a feel that the platform has sufficient bioinformatics knowledge associated to the facility; Jonathan Mullins is modelling membrane proteins from sequence (and motif) information, Department of Mathematical Statistics, Chalmers Göteborg, is involved in the experimental setup, and there is a Ph. D. student focused on bioinformatics questions. Also the bioinformatics issues are to be solved within the different projects.

A.2 Current and Expected Users of SWEGENE Facilities

A.2.1 cDNA Microarray Facility, Lund [cf. A.1.2]

Researchers from Automatic Sequencing of Non-Model Species Facility [Sec. A.1.1] are using microarray experiments in addition to their own activities, and they had some questions when they started with microarray experiments. How to work with microarray data? How does Clustering/promoters/... work? Sampling times? Design of experiment? Spotting has been done with assistance from the facility. Early impressions are that the facility should have a more streamlined data management.

They were aware of BASE and has their own annotation tool, PHOREST.

Needs

There is a fungi genome sequencing project at Whitehead Institute in US. How to connect BASE, PHOREST, and the future fungi genome data?

Comments from an experienced microarray user (Ph.D student).

The analysis sophistication levels are different between users. Some users will only use microarrays to complement their research whereas others will develop new algorithms.

As a biologist, one needs at least an intuitive feel for algorithms used (and needed), but the feeling is that seniors does not participate in the new bioinformatics trend. The knowledge transfer within the department is within the Ph.D students, and (fortunately) since the students overlap with each other the knowledge stays within the department. Maybe postdocs could help to fill the knowledge gap, but currently there is not any (external) postdoc tradition.

The outcome of an experiment depends heavily on prior, personal or departmental, knowledge. Tools to decrease the importance of this prior knowledge is needed, i.e. there should be tools for finding gene relations, gene-pathway connections, and more.

BASE is a quantum leap. At the beginning of microarrays everything was done manually with Filemaker Pro and Excel. BASE organises, visualises, ..., but no analysis. The implemented analysis tools are not sufficient (clustering and visualisation, MDS), so cooperation is needed between groups that can contribute to BASE.

Needs

BASE solves a part of the data management of a microarray experiment but there is still a lot of manual work done. The output from the microarray experiments, and BASE, are gene lists. These lists are examined manually by using LocusLink, Medline, Clonedike and Transpath. This introduces randomness, i.e. one needs some luck to get interesting hits.

Analysis tools to find systematic errors are needed, as it is now some researchers do it, others don't.

Comments from newer microarray users.

Suggestions

Presently there is a need for basic computer and data infrastructure, easy accessibility. The efforts in these basics should be coordinated since it is a waste of resources if every department has to solve the same problems. For our personal benefit a molecular biology server would be nice. GCG should be there, up to date and maintained, accessible for every one at Lund University. There is a GCG server already, but it is poorly maintained. To avoid maintenance problems the server should be sponsored by SWEGENE.

Someone should compile information about the SWEGENE resources. The current web page is a scandal and must be fixed. A resource centre for the platforms and bioinformatics should be created. This centre should inform people of what can be done with the platforms, why it is interesting and how to do it. Potential users might not be not aware of several of the available resources, and a proper presentation of facilities could attract more users.

The bioinformatics platform should be made available for users at modest cost. These users should include students and people involved in education (this opinion stems from the

understanding that SWEGENE funds cannot be used in education).

Needs

The cDNA microarray facility is expected to make a few workstations connected to the experimental setups available for computer analysis and processing of experiments. This will be awkward and cumbersome for the users who might prefer to do the processing at their own department. The access to image analysis (software) might become a bottleneck. This could be solved by trying to get a site license for the analysis software. A single user license cost about USD4000.

Another issue is the question what the users are supposed to do with their data when they leave the SWEGENE resource centres.

We have a DNA sequencer ourselves for use at cost price. However, there is a need for software.

A.2.2 Proteomics Research and Development Centre, Lund [cf. A.1.6]

Comments from a user of the Odense protein identification facility.

At Odense people are allowed to run their experiments themselves, but there are several hired people to (exclusively) maintain the machines, and educate users. Thus, as a user of the Odense facility, it is a pity that users are not allowed to run their experiments themselves at the SWEGENE Proteomics Research facility. Experience from Odense seems to be that a pure service organisation was less efficient. At Odense only public databases are used (in which form was somewhat unclear), and a software package, GPMAW[14], is used.

The prices of the analyses might become an obstacle in attracting users. The quoted prices are 4000SEK/2D-gel to take the user to the MALDI identifications. There is approximately 20 rejects (on a 96 plate) that needs an another qTOF run at a price of 1000SEK/spot. If the user compare two gels, and pick one plate of spots from each, this analysis will cost approximately 50000SEK.

A.2.3 Imaginary (non-existent) Bioinformatics Facility

I have met several computational biologists and algorithm people that would make up a pool of advanced users of a future bioinformatics/computational biology facility. They have many ideas for bioinformatics facility, and some of these ideas are reproduced here.

Software

If one would like to promote software sharing one should maybe setup a SWEGENE sourceforge[15] where people could exchange software. This has been done with BASE and the mailing list is fairly busy.

Create validation tools, eg. tools that could sort out a group of genes on the basis of function. This sort of tools are useful when all the clusterings, normalisations, ... are done, and you must start to think and use biology knowledge to get further.

[14] http://welcome.to/gpmaw
[15] http://sourceforge.net

Communication

A communication platform that people within SWEGENE could use. One might use Wiki Wiki Web[4]. Wiki allows any user of a web page to add and change information on it. If properly used this can be utilised for communication of results, experiences and more.

A forum for discussing algorithms (clustering, normalisation, theories, how to extract information).

Practical education about the platforms are needed. This will enable the users to find proper information.

Databases

There is a need to make peoples lives easier, more atomised. A lot of tasks are done ineffectively, and can be atomised with proper database and programming support. A centralised database repository might be useful, since one usually must retrieve data from various servers on the web. This might clog servers especially if large arrays are used. Biologists wants simple interfaces to retrieve information. For analysis they will form collaborations.

Create a database infrastructure that gives easy access for users. This will automatically lead to further development of the system. A major task is to create a protein database with a single front-end to all databases.

In order to create something new, as compared to just mirroring, a database facility should give users non-web access to the databases through APIs. This might need communication encryption to complicate eavesdropping. And, instead of mirroring databases, maybe one should create a software interface that makes the accesses to the proper servers. This will be dependent on external servers and the software must be monitored. The database mirroring monitoring is changed to server monitoring.

It would be very nice to access pathways in a computer readable form. Actually everything should be in computer readable form, cf. KEGG[16] has GIF graphics to present information. Establish systems that enable the connections from mRNA, sequences, proteins, and more to biology.

Services

NIH has a service organisation for biologists where they can put requests. These requests are treated as projects, and billed to the department doing the requests. The result of the projects are usually a new interface through a web page. This service organisation is supported by a bioinformatics core that maintains and develops the bioinformatics facility at NIH.

In general people with programming knowledge and theoretical skills are needed. This could be solved, in part, by hiring Ph.D. students with "proper" knowledge and make them to assist (as part of their duties) the rest of the group(s).

[16] http://www.kegg.org

A.3 Software and Services

This section describes different software packages and services I encountered during visits at the different groups. There are different amount of information presented for the projects, with more or less information about their weaknesses and strength. This should not be interpreted as a grading of the different projects but is a mere reflection of my understanding of what is offered by the project, and the amount of information presented to me.

A package that is not created within SWEGENE but made available to researchers is the license The Wallenberg foundation acquired from Celera. This is a nationwide 3 year license to the Celera databases and software tools. Nationwide means that there is 101 principal investigator licenses (each with 3 or 4 designated licensees) which the licensee must pay approximately USD2000 for (depending on demand). Several groups have access already.

A.3.1 BioArray Software Environment – BASE

BASE[17] is a SWEGENE funded project, and is developed at Complex Systems Division, Theoretical Physics, Lund University in cooperation with Department of Oncology, Lund University by three developers; Lao Saal, Carl Troein, and Johan Vallon-Christersson (all Ph.D. students). Others have also made contributions. Lao and Johan are research biologists and have been involved in the application definition and as cDNA microarray experts. Carl has been *the only* major software developer in this project.

- First release version 1.0 was shipped May 17, 2002 under GNU General Public License (GPL)[18]. The developers fix problems reported through the mailing list. The project has a project page at sourceforge[19], and has attracted two external software developers. There seems to be enough interest around the project that it should be properly funded and extended. There is already external contributions to the software, someone contributed a small package to connect BASE to the GeneSpring analysis package.

- Currently the project is poorly documented. Only installation documentation is available[20].

- There is support for plug-ins, and plug-in development is encouraged on the BASE web site. There are two summer workers employed to implement algorithm plug-ins to BASE. However, plug-in creation documentation is poor and needs to be improved.

- Currently only one file format is supported, base-format. MAGE-ml[21] support is under construction.

- A Java 3D-viewer is under development.

[17] http://base.thep.lu.se
[18] http://www.gnu.org/licenses/gpl.html
[19] http://sourceforge.net/projects/basedb
[20] These are under construction.
[21] http://www.mged.org/Workgroups/MAGE/mage.html

- cDNA microarrays were the primary target in the development, but Affymetrix data should work as well. This is not well tried but people who downloaded BASE seem to be interested in getting Affymetrix to work.
- There are no APIs available, this will be a major task and beyond the current developers time horizon. There are discussions on how to implement a database API. This would allow the usage of other database engines than MySQL since only the (future) interface needs to be adjusted to different databases.
- To get the most of BASE (research wise) within a department there should be SQL knowledge, and enough programming knowledge to be able to implement plug-ins.

The fact that Carl Troein has been the only software developer in the project is in principle a weakness, but this seems to be resolved now after the GPL release. The project attracts outside development and support in open source spirit.

Needs

The current support plan from Carls part is that he is working with BASE until mid-July 2002, and must decrease is efforts to a 5% fraction of his time, thus spending most of his time on other projects. Projects ending in a scenario where Ph.D. students get stuck in software development should always be avoided. Projects like BASE (and PHOREST [cf. A.3.3]) where one or two persons become too involved and important will require too much attention. There should be plans to create a self sustainable environment for the project, i.e. make it public and if it is interesting for a wider audience, the project will become independent of the original developers. The project still needs management, and in the case of BASE this should be accomplished without too much involvement by Carl Troein, thus a project manager is needed for BASE.

The links to information about the clones need to be changed to accommodate other chip data. There is also a, near in time, need to facilitate for linkage to other species such as mouse, rat, zebra fish, and more.

A.3.2 Clonedike

Sven Bilke has, together with Thomas Breslin (both at the Complex Systems Division, Department of Theoretical Physics, Lund University), created Clonedike. This is a tool to assess the quality of analysis of microarray data using annotated ontology information. It is a stand-alone software and will soon be released under the GPL and maintained by Sven.

The development of the software is funded by SWEGENE, and has already been installed at the Microarray Resource Centre in Lund [cf. A.1.3].

A.3.3 PHOREST

Developed at, and funded by, Microbial Ecology, Lund University, by two developers; Dag Ahrén and Carl Troein. PHOREST was created to meet a growing need to manage sequences from sequencing programs. The project has escalated since its initiation as people's needs changed throughout the project realising the power of their new tool.

PHOREST deals with cDNA libraries, and searches, comparisons, and clustering (CAP algorithm) are implemented. When a genome is sequenced, the sequence is searched for ESTs, and these are the input to the database in PHOREST. (PHOREST could be used for any kind of sequences if the application makes sense.) The clustering creates unisets of ESTs and gives information about which genes are expressed in large quantities. These uniset could be used to create microarrays for a more precise(?) measurement of the expression levels. An important feature of PHOREST is that the unisets are manually annotated.

- PHOREST is a web based solution with an Apache (SSL or HTTP) server, PHP, and a MySQL database. No API is defined. Disk space is moderate, depends on external database.

- Everything in PHOREST is automatically updated every month, the underlying databases (NCBI non-redundant protein database and more?), homology searches, clusterings, .. and the users are automatically notified of changes.

- PHOREST connects to local copies of Genbank and 4–5 other databases. Mostly used for blast searches. Tunlid would like to link sequences to metabolic pathways. The current annotation is stopped at TCA cycles and KEGG EC number analysis is done. Connections to MIPS[22] database (gene expression in pathways) for further metabolism studies.

- PHOREST has data export possibilities. One could create a microarray experiment (BASE), run it and re-enter the results into PHOREST.

- There is a fungi genome sequencing project at the Whitehead Institute in the US. This is expected to be included somehow into PHOREST.

- Ahrén thinks that algorithm development is needed within PHOREST, e.g. the clustering algorithm is not perfect.

- There are some plans of creating PHOREST–BASE links.

Dag Ahrén has been a part of the creation of PHOREST (together with Carl Troein who made major software contributions), as biologist and partly as a programmer. BioBridge Computing AB[23] has acquired the rights to the software, and Ahrén has a deal with BioBridge, where he is supposed to finish up the PHOREST project (funded by Teknikbrostiftelsen). Anders Tunlids group has full rights to the work, meaning that they can do whatever they want with their PHOREST package, but not redistribute it. It is not clear what this means for other academic groups. I have done some investigations in this matter. I have the understanding that BioBridge has all rights for PHOREST, even future improvements made at Tunlids group, and has no plans to allow free academic licenses. I have tried to persuade BioBridge to reconsider without any success so far.

[22] http://www.mips.biochem.mpg.de
[23] http://www.biobridge.se

A.3.4 Profiling Polygenic Diseases Computer Services

The Malmö facility uses two different applications in their research; a software package supplied from BioComputing OY[24] in Finland, and a database application implemented by Prevas[25] consultants commissioned by Arexis[26] in Göteborg. The plan is to give the facility users access to the applications through the Web.

A.3.5 RATMAP and GAPP, Göteborg

RATMAP[27] is not really developed within a fully SWEGENE supported project but SWEGENE has supported a one year contract (ending in August 2002) for a programmer. The programmer has transfered, and in the process created a new application, the RATMAP database from Filemaker Pro with static web pages to a fully dynamic webified database application. The transfer could not have been possible without external support and knowledge.

The objective of RATMAP is to collect, curate and present rat genome data, maintain the rat gene nomenclature[28] and to define interspecies gene homology, orthology, (rat-mouse and rat-human).

The RATMAP group has also created GAPP (Gene And Position Prediction)[29] where the objective is to present a detailed rat/mouse comparative map and to predict rat gene positions using mouse data (prediction accuracy is 95%). The comparative map is now extended to include rat/human comparisons (currently rat/mouse).

The two applications have 9000 and 5000 logins/month (counted on a daily basis), respectively.

The team consists of eight people, of which two will have funding problems during the autumn 2002. This is said to be temporary, and funding is secured to late 2002 or early 2003.

A.3.6 Structural Biology Web Services

Guoguang Lu at Department of Molecular Biophysics, Lund University, has a few proteins structure related algorithms implemented as web services[30] that are popular and frequently used by external users. Lu is also running the GCG/SEQWEB package acquired by SWEGENE, and the local GCG server has approximately 20 users.

Jari Häkkinen
August 14, 2002

[24] http://www.biocomputing.fi
[25] http://www.prevas.se
[26] http://www.arexis.se
[27] http://ratmap.gen.gu.se
[28] http://rgnc.gen.gu.se
[29] http://gapp.gen.gu.se
[30] http://bioinfo1.mbfys.lu.se