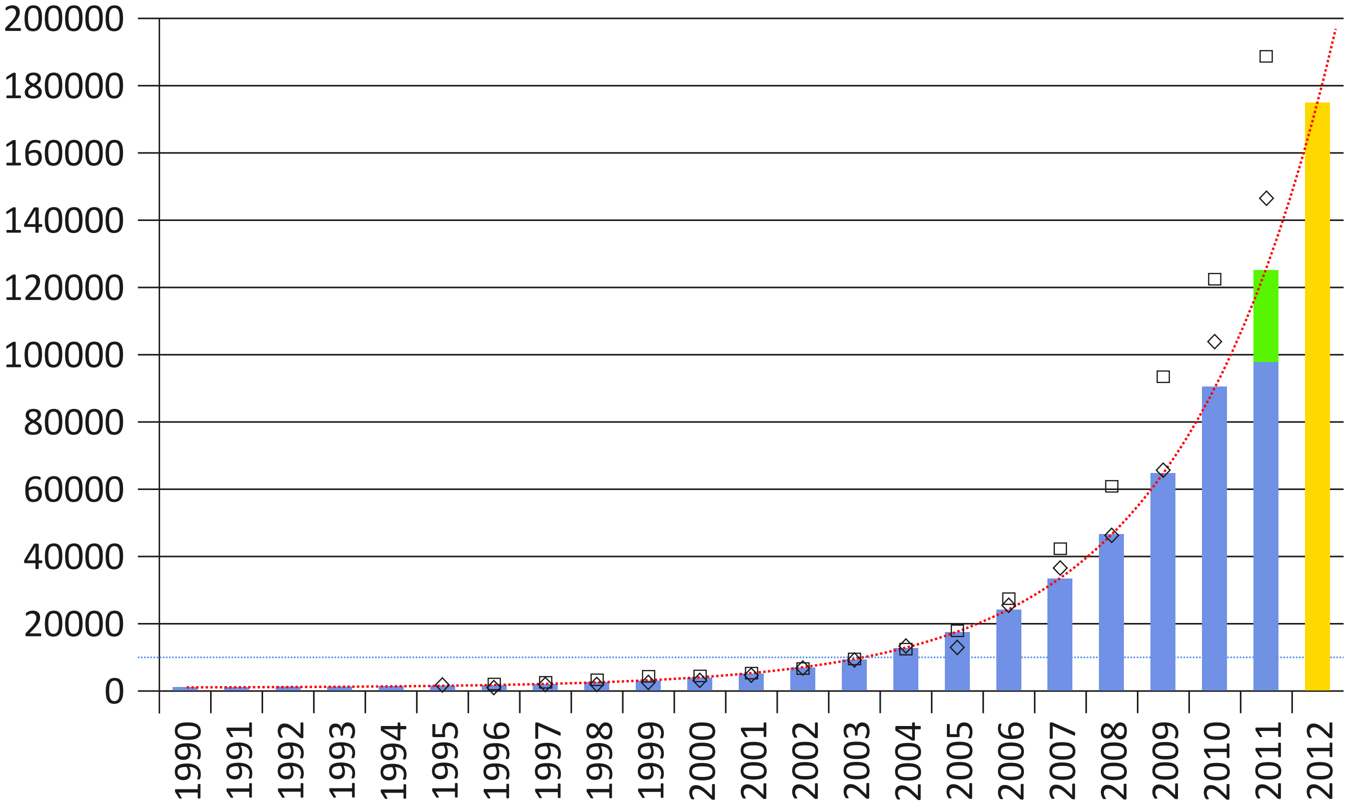
NSC News

2011:3

National Supercomputer Centre at Linköpings universitet

Fighters of the Tide



Fighters of the Tide

Bioinformatics experience a super-exponential increase in data volumes, far superior to the growth curves for both processing power per dollar and bytes storage per dollar, says Joel Hedlund.

Read further on page 3

Large-scale data handling

Data storage is an increasingly important part of SNIC's tasks. At the recent workshop in Swestore (cf. separate article in this issue) the upcoming challenges were presented together with the new storage solutions. These will provide cost-effective and sustainable storage for Swedish scientists. International collaborations will also be important – both at the Nordic and at the pan-European level. SNIC is part of NDGF on the Nordic level and in EUDAT on the EU level. Furthermore, security is crucial when providing long-term storage. For accurate access control, Swestore utilises certificates, which are issued by the universities. Thus, if the researchers should get access to the national storage facilities, all universities need to provide these certificates. This is not the case today, but will hopefully be that in the near future! It is also important that the SNIC centres keep close collaborations with data-rich research infrastructures. Within the life science domain, the infrastructure BILS has already ongoing collaborations with NSC, and additional collaborations are currently initiated with UPPMAX.

The second implementation phase of the EU project PRACE has now started. All SNIC centres will contribute, and the three centres that have participated in PRACE since long will be the major providers. NSC will contribute within our expertise areas of code optimisation, large-scale storage, and large-scale computer rooms. Furthermore, we will arrange a PRACE workshop together with HPC2N in autumn 2012.

NSC is continuing expanding and recruiting more personnel. Currently, we have 5 positions announced – two system administrators, two system developers and one computational scientist. Please check our web site http://www.nsc.liu.se for more details.

For large-scale data handling, NSC has recently acquired two nodes in the Kappa cluster, equipped with 1 TiB RAM each. These nodes are aimed at special applications requiring very large internal memory, e.g. large genome assemblies and large quantum chemistry calculations. Even though the nodes are financed by Linköping University, they will be made available also for users at other universities.

BENGT PERSSON, NSC DIRECTOR

BILS

BILS (Bioinformatics Infrastructure for Life Sciences) is a distributed national research infrastructure supported by the Swedish Research Council. The aim is to provide necessary infrastructure in the form of databases, data storage and analysis tools. Furthermore, BILS provides bioinformatics support to life science researchers in Sweden. BILS is also the Swedish contact point to the European infrastructure for biological information ELIXIR. BILS was established in 2010 and has currently a staff of 15 people and will during 2012 expand to about 25 persons. Present BILS staff has expertise in protein bioinformatics, mass spectrometry (MS), next generation sequencing (NGS), large-scale data handling, metagenomics, systems biology, and RNAseq. Additional expertises will be added during 2012.

BILS is predominantly offering bioinformatics support in various projects, depending on the user needs. In the projects, the BILS persons are working close to the research group, and they spend part of their time to teach the users in order to propagate the bioinformatics knowledge. Furthermore, BILS provides infrastructure and tools for bioinformatics analyses in order to facilitate these analyses for the users.

BILS is together with SNIC developing systems and strategies for long-term large-scale storage of bioinformatics data (MS proteomics data, NGS sequence data). BILS is also working in close collaboration with Science for Life Laboratory in Stockholm and Uppsala.

To get in contact with BILS, please find contact details on http://www.bils.se

BENGT PERSSON

NSC contributes to PRACE

PRACE (Partnership for Advanced Computing in Europe) is a Research Infrastructure of top level HPC ecosystem in Europe. Currently PRACE-RI includes three Tier-0, petaflop systems namely JUGENE (Blue Gene/P), CURIE (x86-infinband) and HERMIT(CRAY XE6) and several Tier-1 systems. In the near future two more upcoming Tier-0 systems namely SuperMUC and FERMI will be added to PRACE-RI.

PRACE-RI encourages researchers from across Europe to apply for access to its world class resources for research via a central peer review process. PRACE Tier-0 resources are available through three forms of access: Programme Access, Project Access and Preparatory Access. Program Access is available to major European projects that can benefit from PRACE resources. Project Access is intended for individual researchers and research groups including multi-national research groups. Preparatory Access is intended for efficient resource use required to prepare proposals for Project Access. PRACE Tier-1 resources are available through DECI calls.

PRACE-RI provides HPC expertise to prospective users for effective use of PRACE resources. PRACE-RI has an extensive education and training effort through seasonal schools, workshops, scientific and industrial seminars etc throughout Europe to aid users and potential users of PRACE systems. The 1st implementation phase (PRACE-1IP) started 1st July 2010 and will continue until 30 June 2012. The 2nd implementation phase (PRACE-2IP) started 1st September 2011 and will continue until 31st August 2013. Twenty one European countries are members of the PRACE research infrastructure. In Sweden SNIC is partner in PRACE-1IP/2IP with the participating centres NSC, PDC, HPC2N, UPPMAX, C3SE and LUNARC.

The contributions from NSC in PRACE-2IP will be in providing expertise to users of PRACE systems for porting and optimizing their codes to effectively utilize the Tier-0 and Tier-1 systems, helping prospective users in submitting proposals for PRACE resources, benchmarking PRACE systems. NSC will also contribute to investigation in future technologies, especially with respect to energy efficiency. This includes studies of highly energy efficient HPC components and systems, as well as power and cooling technologies. Furthermore, NSC will work on novel programming techniques, models and tools in order to achieve good efficiency on petascale systems.

More details about PRACE-RI is available at http://www.prace-ri.eu.

CHANDAN BASU

Bioinformatics – Fighters of the Tide

Biological research is necessary. Not only to get a better understanding of the myriad of intricate and interwoven processes that go on in our very own bodies in order to ensure our continued existence, but also for combating new diseases and understanding our role in the environment. The benefits of biological research are evident in our current simple cures to old plagues and crippling ailments, but the need for further progress is equally apparent, for example in our lack for cures for genocides like malaria, and the threat of emergent pathogens like H1N1.

Unfortunately, biological research is also very costly. This is mostly because biology is life; it's horribly complex, and we don't understand it! Even a simple experiment, like for example culturing bacteria in a test tube and measuring their reaction to certain stimuli, is influenced by innumerable variables that need to be precisely controlled in order to ensure consistent results. What's worse, many of these variables are difficult or impossible to measure, and an unknown number of these variables are simply unknown and therefore impossible to even assess. There are also often numerous confounding factors. For example, the bacterium in question may have several mechanisms in place to react to that specific type of stimulus, only some of which produce the response that is being measured. Furthermore, biological experiments are nearly always very time consuming. The measurements in our simple example would probably only take hours, but would likely be preluded by days of rigorous preparation, growing the bacteria under exact and reproducible conditions, and painstakingly ensuring that no contamination occurs along the way.

There is also of course the ever present ethical imperative. In our society it is thankfully unthinkable to take the most direct route to that new biological knowledge that is most relevant to us humans, so instead of trying out new drugs on humans directly, we tend to take the long way around, starting with test tubes and yeast cells and slowly and laboriously moving up to animals and eventually people, progressing only at the slow pace set by the rigours of acceptable safety.

We obviously can't do all the experiments we want to do. Money, time and ethics set universal limits, so we have to prioritise. Preferably, we should start with those experiments that will teach us the most, and to make those experiments count, we should pry the maximal amount of knowledge out of every speck of data that we collect.

Bioinformatics is the science of handling information on biology. One of its aspects is to ensure that experimental results are stored in an accessible and orderly fashion, so that scientists worldwide can best benefit from them. Another aspect is to use the collected data; to process it in various ways in order to synthesise new theories, find new genes, and explain infection mechanisms for new viruses.

Our success in the first aspect is a great aid and a great challenge for the second. Laboratory methods develop and constantly move toward higher throughput and larger data volumes, so as accessibility and interoperability increases, the possibilities for new discoveries of course increase exponentially, but at the same time we are drowning in that same data we strove so hard to collect (cf. Fig. 1). Smarter and faster algorithms and better ways of using bigger and faster computers are perpetually in high demand, and automation becomes a necessity in order to stay afloat.

So to summarise, bioinformatics entails using computers to analyse huge amounts of very complicated data, taken from a field that is only partially understood, to see the hidden trends and connections, and draw useful conclusions.

Protein families

So how can we avoid getting washed away in the flood? It is obvious at this point that humans can never hope to keep up with the escalating data generation rate, at least not on the level of individual entries. However, many entries share many common properties, and by clustering entries we could conceivably do better in our race against the machines. One useful way of clustering proteins and genes is to define protein families; groups of closely related proteins likely to have the same physicochemical properties, expression patterns, interactions with other cell components, tissue specificity, subcellular localisation and so on, grouping together the evolutionary counterparts from different species ranging from man through mouse, octopus, banana and various bacteria and even on to viruses. The alcohol dehydrogenases working busily in our livers from time to time is one example (cf. Fig. 2). As a side note, apart from alleviating the annotation effort, protein families are what allows us to do experiments on lab rats rather than people.

Borrowing from speech recognition, profile hidden Markov models (HMMs, as used in HMMER [3]) have enabled us to teach computers to recognize new members of known protein families, by feeding them good examples of already known members. A computer with a library of HMMs could then automatically detect all new members of known families in a newly sequenced genome, on the fly, and at virtually no cost [4].

The problem with this excellent first line of flood barrier automation is that someone has to teach the computers about the protein families in the first place, which is a labour intensive work even for seasoned field experts, and there are plenty protein families that are little known or even need discovering (cf. Fig. 3). The second line of flood barrier automation is therefore currently under construction, where computers automatically generate models that can automatically detect new members of new families [5]. Welcome to the new millennium!

JOEL HEDLUND

National Data Storage

On 28 October, the "Swedish National Data Storage Initiative" Swestore arranged a workshop to inform STAC about status and user needs. The Swestore project was initiated 2008 and should be fully integrated in SNIC with a tight collaboration with large scale initiatives/projects. The core services in the project are:

* Harmonized centre storage solutions at all six SNIC centers.

* Cross-site backup

* National distributed transparent-access storage.

This work is lead by the Swestore working group (lagringsgruppen) with members from all SNIC centres.

Swestore is based on three pillar – Centre storage, Tape infrastructure and National storage.

Centre storage is a center wide cluster storage that should be harmonised from a user perspective. The SNIC centres collaborate on procurement and at the technical level.

Tape infrastructure is now up and running. All centres have backup and cross site backup between three tape libraries at HPC2N, PDC and NSC. There are also storage projects that use this tape infrastructure.

National storage. The national storage are now up and running, currently with about 20 different storage projects. Storage pools are distributed between all six SNIC centres. Currently, there is about 1200 TiB storage available, and further expansion is planned. All SNIC clusters have grid access to the storage. For easy access to this storage, the users need a certificate, which can be issued by the university. However, there are some initial difficulties with these certificates that need to be solved. These includes:

* User client tools are hard to use.

* User documentation is poor

* User certificates are difficult for users to manage.

* User management.

The conclusion from the workshop is that the Swestore project is up and running. Centre storage and tape infrastructure will be maintained. Furthermore, even though the national storage is already in use, there are needs for additional developments to create user environments suitable for each large user group.

The next Swestore workshop is planned for spring 2012.

TOM LANGBORG
SWESTORE COORDINATOR

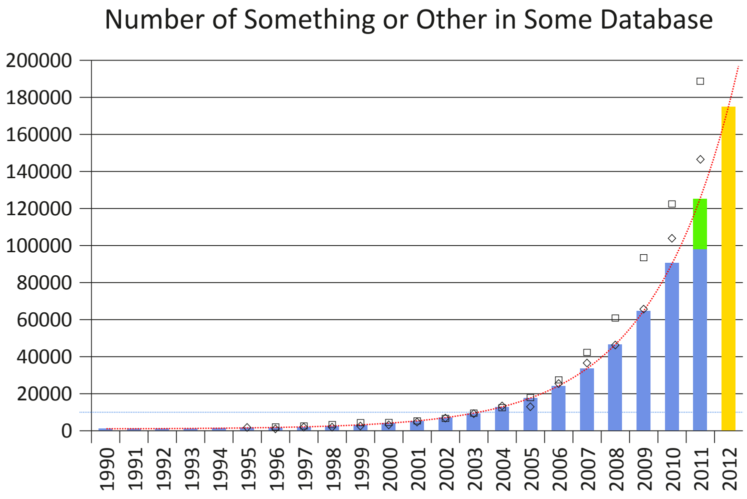


Figure 1: This figure shows the number of something or other, in some database, in arbitrary units. In the field of bioinformatics, this is the picture you will get regardless of how you slice the cake or which cake you slice, be it the number of available protein sequences or completed genome sequencing projects. Therefore, one of these is nearly always included for motivation in any bioinformatic presentation, paper or grant application. The author will draw your attention first to the exponential hockey-stick shape of the trend line, and then to the fact that we have already surpassed last year's number, and that the projected number for next year is even higher. Then the author will point out that human information processing capabilities cap out at about the level of the thin blue line (at around 10% of today's number), before concluding that we humans, as a species, have been out of our depth since 2003, and we're doing progressively worse all the time, with no end in sight, and that's why we need more computers, automation and bioinformatics. In this particular instance of this plot, you can also see the superimposed growth curves for the number of GOLD completed genome sequencing projects (diamonds, currently 2980 [1]) and TrEMBL protein sequences (squares, currently 17651715 [2]). A notable feature of especially the latter curve is that the last half-decade has actually showed a super-exponential increase in data, far superior to the growth curves for both flops processing power per dollar *and* bytes storage per dollar, and this will likely become a major problem in the very near future. So the problems and opportunities in this field are in fact even greater than I have hitherto led you to believe!

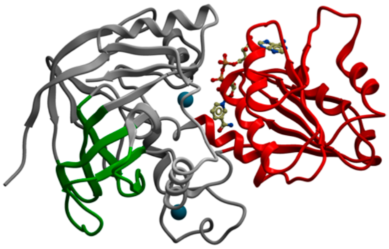


Figure 2: Schematic representation of the three dimensional structure of human class III alcohol dehydrogenase, with its zinc ion and NAD cofactors. The chain of peptide bonds connecting the sequence of amino acids is shown as a ribbon, coloured red in the cofactor binding domain, and green in the folding core of the catalytic domain.

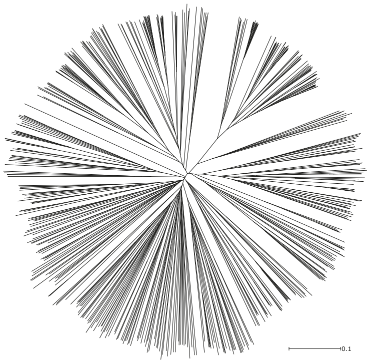


Figure 3: Tree diagram of relations between proteins in the MDR superfamily [5], excluding non-eukaryotic sequences as well as 80 % of the remainder for clarity. In total, 518 out of 16667 known members are shown, and the scale bar represents 10 % sequence differences. The alcohol dehydrogenases (found for example in human liver) can be found as one of the little broom shapes at approximately 1 o'clock. This figure serves as an illuminating example of the intricacies of the field of bioinformatics as a whole, where, after removing layer after layer of size and complexity, there is always ample size and complexity still left to go around.

References

[1] K. Liolios, I.-M. A. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, N. C. Kyrpides, The genomes on line database (gold) in 2009: status of genomic and metagenomic projects and their associated metadata., Nucleic Acids Res 38 (Database issue) (2010) D346D354. doi:10.1093/nar/gkp848.
URL http://dx.doi.org/10.1093/nar/gkp848

[2] T. U. Consortium, The universal protein resource (uniprot) in 2010., Nucleic Acids Res 38 (Database issue) (2010) D142D148. doi:10.1093/nar/gkp846.
URL http://dx.doi.org/10.1093/nar/gkp846

[3] R. D. Finn, J. Clements, S. R. Eddy, Hmmer web server: interactive sequence similarity searching., Nucleic Acids Res 39 (Web Server issue) (2011) W29 W37. doi:10.1093/nar/gkr367.
URL http://dx.doi.org/10.1093/nar/gkr367

[4] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, The pfam protein families database., Nucleic Acids Res 38 (Database issue) (2010) D211D222. doi:10.1093/nar/gkp985.
URL http://dx.doi.org/10.1093/nar/gkp985

[5] J. Hedlund, H. Jörnvall, B. Persson, Subdivision of the mdr superfamily of medium-chain dehydrogenases/reductases through iterative hidden markov model refinement., BMC Bioinformatics 11 (2010) 534. doi:10.1186/1471-2105-11-534.
URL http://dx.doi.org/10.1186/1471-2105-11-534

Joel Hedlund is an application expert in bioinformatics at NSC. He has worked mainly with large scale protein sequence analysis and classification, with a special interest in grid computing.

HPC Courses Held at NSC

Between October 25 and October 28, two HPC courses open to all SNIC users were held back to back at NSC. The first was a one day course and provided an introduction to Grid computing and accessing grid storage. The second course was a beginners level course in the message passing interface API, MPI. Both course were given in lecture form interspersed with hands-on sessions and exercises.

The Grid and MPI courses were led by the Swegrid coordinator Jonas Lindemann and parallel programming expert Joachim Hein respectively, both from the Lunarc supercomputing centre of Lund University. NSC staff assisted in the practical sections of the MPI course.

The grid course covered the basics from getting a Grid certificate allowing access to grid resources to submitting jobs and accessing storage and also gave an introduction to the large scale Swestore national storage infrastructure.

In the MPI course, the concepts behind message passing and distributed memory computing was introduced and the key MPI calls were explained. The course contained practical, hands-on sessions where point-to-point communications, non-blocking communications and the collective communication calls were used in concrete programming exercises.

Many participants came from outside Linköping including Uppsala University, the Karolinska Institute, Stockholm University and Lund University as well as NSC partners SMHI and Saab.

NSC wishes to thank the lecturers, assistants and all course participants for making this a successful event and we hope to see you again on future events arranged.

JOHAN RABER

Course participant Daniel Filipazzi (left) and course lecturer Jonas Lindemann.

Practical session during the MPI course.

UPCOMING EVENTS

HiPC 2011; 18th IEEE International Conference on High Performance Computing
December 18 – 21, 2011, Bengaluru (Bangalore), India.
http://www.hipc.org

FAST'12; 10th USENIX Conference on File and Storage Technologies
February 14 – 17, 2012, San Jose, CA, USA.
http://www.usenix.org/events/fast12

PPoPP 2012; 17th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming
February 25 – 29, 2012, New Orleans, LA, USA.
http://dynopt.org/ppopp-2012

HPCA-18; The 18th International Symposium on High Performance Computer Architecture, 2012
February 25 – 29, 2012, New Orleans, LA, USA.
http://www.ece.lsu.edu/hpca-18

CCGrid 2012; 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing
May 13 – 16, 2012, Ottawa, Canada.
http://www.cloudbus.org/ccgrid2012

IPDPS 2012; 26th IEEE International Parallel & Distributed Processing Symposium
May 21 – 25, 2012, Shanghai, China.
http://www.ipdps.org

ICCS 2012; International Conference on Computational Science
June 4 – 6, 2012, Omaha, NE, USA.
http://www.iccs-meeting.org

HPCS 2012; The 2012 International Conference on High Performance Computing & Simulation
July 2 – 6, 2012, Madrid, Spain.
http://hpcs2012.cisedu.info

Euro-Par 2012; International European Conference on Parallel and Distributed Computing
August 27 – 31, 2012, Rhodes Island, Greece.
http://europar2012.cti.gr

ICPP2012; The 41st International Conference on Parallel Processing
September 10 – 13, 2012, Pittsburgh, PA, USA.
http://www.icpp2012.org

The 27th NORDUnet Conference
September 18 – 20, 2012, Oslo, Norway.
http://www.uninett.no/NORDUnet2012

IEEE Cluster 2012
September 24 – 28, 2012, Beijing, China.
http://ieeecluster2012.csp.escience.cn

LINKÖPINGS UNIVERSITET ◦ LINKÖPINGS UNIVERSITET

Linköpings universitet

National Supercomputer Centre, Linköpings universitet, 581 83 Linköping, Sweden
Tel: +46 (0) 13-28 26 18, fax: +46 (0) 13-28 25 35, e-mail: info@nsc.liu.se
www.nsc.liu.se

Editor: Peter Münger Use NSC Express, www.nsc.liu.se/express, to subscribe for a paper or electronic version.