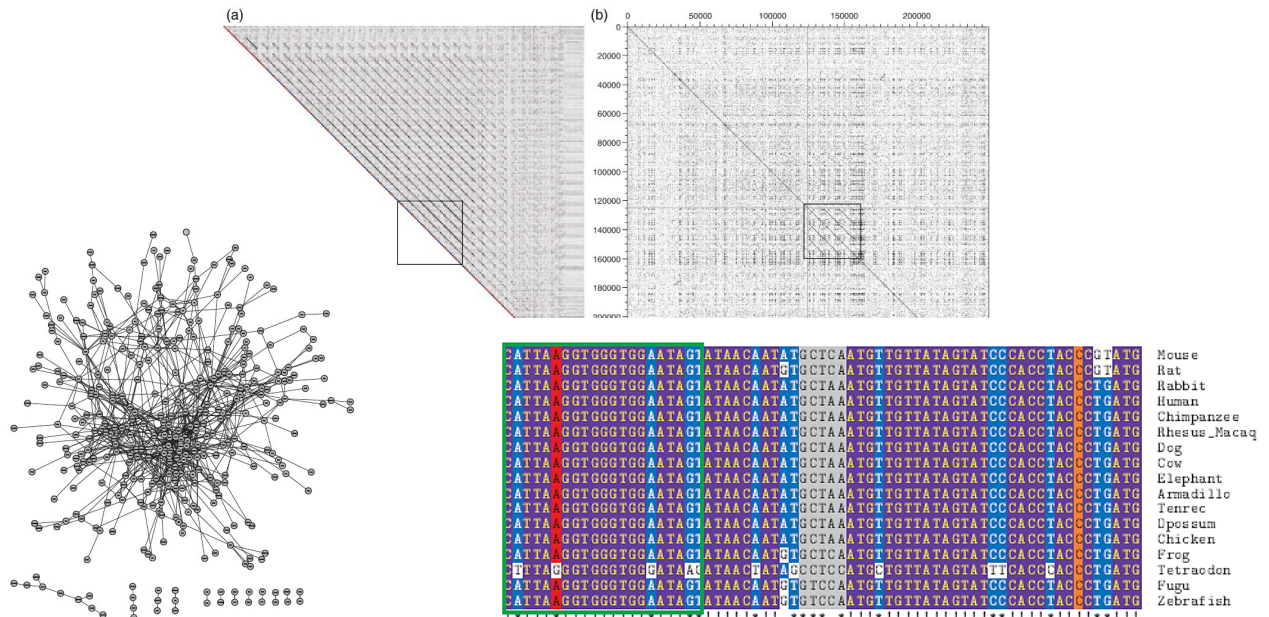


# Stockholm Bioinformatics Centre Annual Report 2010



## Director's summary

In 2010 a very important decision was made – the relocation in 2011 of four SBC groups (Arvestad, Lagergren, Käll, and Sonnhammer) to the Science for Life Laboratories on Karolinska Institutet campus. SciLifeLab is a collaboration between Stockholm and Uppsala which has received “strategic research area” support of 140 MSEK/year in molecular biosciences. The SBC move is considered of great strategic value to both SciLifeLab and SBC, and will no doubt catalyse future collaborations between computational and experimental biology research in the Stockholm-Uppsala area. It will make SciLifeLab better equipped to meet the challenges of modern high-throughput biology, and will give SBC first-hand access to unique genomics, proteomics, and molecular profiling data.

In the other “strategic research area” that SBC is involved in, the Swedish e-Science Research Centre (SeRC), the main activity in the bioinformatics community in 2010 was to recruit senior lecturers at Stockholm and Linköping University, and at Karolinska Institutet. The only finalised position (at SU) was awarded to SBC group leader Lars Arvestad, congratulations!

The only change in senior staff at SBC during 2010 was that Erik Lindahl took a professorship in biological physics at KTH school of physics, and moved his group from Frescati to the Albanova campus. There is a risk that other bioinformatics PIs may also leave DBB, in which case DBB ought to activate the bioinformatics assistant professorship promised in 2005.

Scientifically, 2010 was yet another productive year for SBC, witnessed by the solid publication record and the large number of conference presentations listed in this report. 4 PhD students were promoted: Anna Henricson, Åsa Björklund, Kristoffer Illergård, and Per Larsson. The pictures above are taken from Björklund et al., Östlund et al., and Ensterö et al. (see publication list below).

The Master in Bioinformatics programme at DBB changed leadership during 2010 and managed to increase its enrolment from 2 to 7 students. However, starting fall 2011, a tuition fee of 140

KSEK/year will be levied on non-european students, and this is likely to reduce the enrollment again (currently 4 of the 7 are non-european). Another drawback of the tuition fee is that it will increase the bureaucracy of teaching, as all students will need to register via studera.nu for every single course in order to verify that their tuition is paid. On the other hand, there is some hope that the fee will help to attract better and more motivated students.

### **Personnel during 2010:**

Assoc. Prof. Lars Arvestad

Raja Hashim Ali	PhD student
Mehmood Khan	PhD student
Kristoffer Sahlin	PhD student

Prof. Arne Elofsson\*\*

* Åsa Björklund	PhD student
* Kristoffer Illergård	PhD student
* Per Larsson	PhD student
* Wiktor Jurkowski	Postdoc

Marcin Skwark	PhD student
Rauan Sagit	PhD student
Christoph Peters	PhD student
* Karin Julenius	Postdoc
Nanjiang Shu	Postdoc
Sikander Hayat	Postdoc

Prof. Jens Lagergren

Hossein Farahani	PhD student
Joel Sjöstrand	PhD student
Owais Mahmudi	PhD student
Ikram Ullah	PhD student

Ass. Prof. Lukas Käll\*\*

Luminita Moruz	PhD student
Viktor Granholm	PhD student
* Jesper Lind	Post-doc

Prof. Erik Lindahl

Aron Hennerdal	PhD student
Pär Bjelkmar	PhD student
Sander Pronk	Research Associate
Berk Hess	Research Associate
Szilard Pall	PhD student
Arjun Ray	PhD student
Christine Schwaiger	PhD student
Rossen Apostolov	Post-doc
Samuel Murail	Post-doc
Björn Wallner	Research Associate
Teemu Murtola	Post-doc

Prof. Gunnar von Heijne\*\*

Patrik Björkholm      PhD student

Prof. Erik Sonnhammer (Director of the SBC)

Fabian Schreiber      Postdoc  
Kristoffer Forslund      PhD student  
\* Anna Henricson      PhD student  
Gabriel Östlund      PhD student  
David Messina      PhD student  
Oliver Frings      PhD student  
\* Sanjit Roopra      PhD student  
Thomas Schmitt      PhD student

Erik Granseth\*\*      Postdoc (independent)  
Bengt Sennblad      Senior researcher  
Erik Sjölund      System administrator  
Roman Valls      Assistant system administrator

\*) Left during 2010

\*\*\*) Group located at Arrhenius Laboratory, Frescati

### **Collaboration partners**

SU Molecular Biology & Functional Genomics (Prof. Marie Öhman)  
KTH Biotechnology (Prof. Vincent Bulone, Prof. Peter Savolainen, Prof Joakim Lundeberg, and Prof. Mathias Uhlén)  
KI Aging Research Center (Dr. Lars Bäckman)  
KI MMK (Prof. Anna Wedell)  
KI CMB (Prof. Björn Andersson)  
KI CCK (Prof. Arne Östman)  
KI KBC (Janne Lehtiö)  
KI Biosciences and Nutrition (Virpi Tökönen)  
Uppsala University (Dr. van der Spoel)  
Uppsala University (Prof. Hans Ellegren)  
Linköping University (Prof. Fredrik Elinder)  
Göteborg University (Prof. Bengt Oxelman)  
LU Clinical Genetics (Mattias Höglund)  
AstraZeneca (Prof. Hugh Salter)  
DTU, Lyngby, Denmark (Søren Brunak)  
Bioinformatics Laboratory, BioInfoBank Institute, Poznan (Dr. Leszek Rychlewski)  
Institut Pasteur, Paris (Dr. Marc Delarue)  
Stanford University (Prof. Michael Levitt, Prof. Vijay S. Pande, Prof. James Trudell)  
University of Wyoming (Dr. David Liberles)  
McGill Centre for Bioinformatics (Dr. Mike Hallett)  
University of British Columbia (Dr. Wyeth Wasserman).  
Yale University, New Haven, CT. (Dr. Mark Gerstein)  
University of Buffalo (Dr. Daniel Fischer)  
Cornell University, Ithaca, NY. (Dr. Klaas van Wijk)  
The Sanger Institute, Hinxton, UK. (Drs. Richard Durbin & Alex Bateman)  
Janelia farm, VA, USA. (Dr. Sean Eddy)  
University of Valencia, Spain (Dr. Gustavo Camps-Valls)  
University of Rochester Medical Center (Dr. Fred Hagen)

University of Paris René Descartes (Prof. Jean-Laurent Casanova)  
Max Planck Inst. für Genetik, Berlin. (Alexander Schliep)  
Duke University (Dr. Joel Meyer)  
EU bioinformatics network Biosapiens  
EU bioinformatics network Embrace  
EU bioinformatics network Genefun

### **Scientific publications**

2010 was again a very fruitful year for SBC in terms of publications, with 27 bioinformatics papers in total. Many of them are collaborations with experimental groups, showing SBC bioinformatics innovations and expertise are being applied to solve biological problems in related fields.

Based on <http://www.sbc.su.se/publications>:

Bertaccini, E.J., Wallner, B., Trudell, J.R. and Lindahl, E. (2010) Modeling Anesthetic Binding Sites within the Glycine Alpha One Receptor Based on Prokaryotic Ion Channel Templates: The Problem with TM4. *J Chem Inf Model* 50 (12) : 2248-2255.

Lima, M.F., Eloy, N.B., Pegoraro, C., Sagit, R., Rojas, C., Bretz, T., Vargas, L., Elofsson, A., Oliveira, A.C., Hemerly, A.S. and Ferreira, P.C. (2010) Genomic evolution and complexity of the Anaphase-promoting Complex (APC) in land plants. *BMC Plant Biol* 10 (1) : 254.

Ray, A., Lindahl, E. and Wallner, B. (2010) Model quality assessment for membrane proteins. *Bioinformatics* 26 (24) : 3067-3074.

Hennerdal, A., Falk, J., Lindahl, E. and Elofsson, A. (2010) Internal duplications in alpha-helical membrane protein topologies are common but the nonduplicated forms are rare. *Protein Sci* 19 (12) : 2305-2318.

Kall, L. (2010) Prediction of transmembrane topology and signal peptide given a protein's amino acid sequence. *Methods Mol Biol* 673: 53-62.

Runesson, J., Sollenberg, U.E., Jurkowski, W., Yazdi, S., Eriksson, E.E., Elofsson, A. and Langel, U. (2010) Determining receptor-ligand interaction of human galanin receptor type 3. *Neurochem Int* 57 (7) : 804-811.

Larsson, P. and Lindahl, E. (2010) A high-performance parallel-generalized born implementation enabled by tabulated interaction rescaling. *J Comput Chem* 31 (14) : 2593-2600.

Lind, J., Lindahl, E., Peralvarez-Marin, A., Holmlund, A., Jornvall, H. and Maler, L. (2010) Structural features of proinsulin C-peptide oligomeric and amyloid states. *FEBS J* 277 (18) : 3759-3768.

Moruz, L., Tomazela, D. and Kall, L. (2010) Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *J Proteome Res* 9 (10) : 5209-5216.

Bjorklund, A.K., Light, S., Sagit, R. and Elofsson, A. (2010) Nebulin: A Study of Protein Repeat Evolution. *J Mol Biol* 402 (1) : 38-51.

Henricson, A., Forslund, K. and Sonnhammer, E.L. (2010) Orthology confers intron position conservation. *BMC Genomics* 11 (1) : 412.

Kasson, P.M., Lindahl, E. and Pande, V.S. (2010) Atomic-resolution simulations predict a transition state for vesicle fusion defined by contact of a few lipid tails. *PLoS Comput Biol* 6 (6) : e1000829.

Illergard, K., Callegari, S. and Elofsson, A. (2010) MPRAP: An accessibility predictor for alpha-helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinformatics* 11 (1) : 333.

Frygeliuss, J., Arvestad, L., Wedell, A. and Tohonen, V. (2010) Evolution and human tissue expression of the Cres/Testatin subgroup genes, a reproductive tissue specific subgroup of the type 2 cystatins. *Evol Dev* 12 (3) : 329-342.

Stranneheim, H., Kaller, M., Allander, T., Andersson, B., Arvestad, L. and Lundeberg, J. (2010) Classification of DNA sequences using Bloom filters. *Bioinformatics* 26 (13) : 1595-1600.

Niemela, P.S., Miettinen, M.S., Monticelli, L., Hammaren, H., Bjelkmar, P., Murtola, T., Lindahl, E. and Vattulainen, I. (2010) Membrane proteins diffuse as dynamic complexes with lipids. *J Am Chem Soc* 132 (22) : 7574-7575.

Alexeyenko, A., Wassenberg, D.M., Lobenhofer, E.K., Yen, J., Linney, E., Sonnhammer, E.L. and Meyer, J.N. (2010) Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity. *PLoS One* 5 (5) : e10465.

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlen, M. and Berglund, L. (2010) Prediction of the human membrane proteome. *Proteomics* 10 (6) : 1141-1149.

Sehat, B., Tofigh, A., Lin, Y., Trocme, E., Liljedahl, U., Lagergren, J. and Larsson, O. (2010) SUMOylation Mediates the Nuclear Translocation and Signaling of the IGF-1 Receptor. *Sci Signal* 3 (108) : ra10.

Kauko, A., Hedin, L.E., Thebaud, E., Cristobal, S., Elofsson, A. and von Heijne, G. (2010) Repositioning of Transmembrane alpha-Helices during Membrane Protein Folding. *J Mol Biol* 397 (1) : 190-201.

Enstero, M., Akerborg, O., Lundin, D., Wang, B., Furey, T.S., Ohman, M. and Lagergren, J. (2010) A computational screen for site selective A-to-I editing detects novel sites in neuron specific Hu proteins. *BMC Bioinformatics* 11 (1) : 6.

Jornvall, H., Lindahl, E., Astorga-Wells, J., Lind, J., Holmlund, A., Melles, E., Alvelius, G., Nerelius, C., Maler, L. and Johansson, J. (2010) Oligomerization and insulin interactions of proinsulin C-peptide: Threefold relationships to properties of insulin. *Biochem Biophys Res Commun* 391 (3) : 1561-1566.

Ostlund, G., Lindskog, M. and Sonnhammer, E.L. (2010) Network-based Identification of Novel Cancer Genes. *Mol Cell Proteomics* 9 (4) : 648-655.

Ekman, D. and Elofsson, A. (2010) Identifying and Quantifying Orphan Protein Sequences in Fungi. *J Mol Biol* 396 (2) : 396-405.

Hedin, L.E., Ojemalm, K., Bernsel, A., Hennerdal, A., Illergard, K., Enquist, K., Kauko, A., Cristobal, S., von Heijne, G., Lerch-Bader, M., Nilsson, I. and Elofsson, A. (2010) Membrane Insertion of Marginally Hydrophobic Transmembrane Helices Depends on Sequence Context. *J Mol Biol*

Biol 396 (1) : 221-229.

Finn, R.D., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Res* 38 (Database issue) : D211-22.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O. and Sonnhammer, E.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38 (Database issue) : D196-203.

### **Courses**

Applied Bioinformatics 7.5 hp DD2397 (KTH) / DA2397 (SU) by Lars Arvestad

Algorithmic Bioinformatics 6 hp DD2450 (KTH) by Jens Lagergren

Omic Data and Systems Biology 7.5 hp DD2399 (KTH) by Jens Lagergren and Lars Arvestad

Advanced Bioinformatics 15 hp KB8014 (SU) by Erik Sonnhammer

Advanced Bioinformatics 30 hp KB8015 (SU) by Erik Sonnhammer

Comparative Genomics 7.5 hp KB8007 (SU) by Erik Sonnhammer

Bioinformatics 7.5 hp KB7004 (SU) by Arne Elofsson

Structure prediction of globular and membrane proteins 7.5 hp KB8008 (SU) by Arne Elofsson

Molecular modelling 7.5 hp KB8005 (SU) by Berk Hess (Erik Lindahl)

Protein physics 7.5 hp KB8011 (KTH/SU) by Erik Lindahl

Analysis of Data from High-throughput Molecular Biology Experiments BB2490 (KTH) / KB8018 (SU) by Lukas Käll

### **Invited lectures and seminars**

CIAM workshop on analysis in biomedicin, KTH, Stockholm, 2010-03-24, Lars Arvestad

"membrane protein Bioinformatics", "Systems biology on microorganisms", Mar 2010, Paris, Arne Elofsson

"Protein structure refinement" "3d-sig, ISMB", Boston, July 2010, Arne Elofsson

"membrane protein Bioinformatics", Tokyo University, Oct 2010, Arne Elofsson

"membrane protein Bioinformatics", Nagoya University, Oct 2010, Arne Elofsson

"membrane protein Bioinformatics", CBRC, Oct 2010, Arne Elofsson

"membrane protein Bioinformatics", Saarbrucken University, Germany, Nov 2010, Arne Elofsson

"Protein domain evolution", Saarbrucken University, Germany, Nov 2010, Arne Elofsson

"Semi-supervised machine learning for the peptide identification problem in shotgun proteomics", keynote at the 5th Conference of the Hellenic Society for Computational Biology and Bioinformatics,

Alexandroupolis Greece, 17- 19 October 2010, Lukas Käll

"Semi-supervised machine learning for the peptide identification problem in shotgun proteomics", Instituto Gulbenkian de Ciência, Oeiras, Portugal - 15 September 2010, Lukas Käll

Weizmann Institute, Rehovot, Israel, 2010-02-15, Erik Lindahl

Exascale 2010, Barcelona, Spain, 2010-12-14, Erik Lindahl

Bioscience 2010, Juelich, Germany, 2010-11-16, Erik Lindahl

Korea Advanced Institute of Science & Technology, South Korea, 2010-12-06, Erik Lindahl

Samsung Advanced Institute, South Korea, 2010-12-07, Erik Lindahl

ISC'10, Hamburg, Germany, 2010-05-31, Erik Lindahl

Bridging Nanotechnology Conference, Stanford University, USA, 2010-02-23, Erik Lindahl

Center for Scientific Computing, Helsinki, Finland, 2010-10-28, Erik Lindahl

Molecular Physiology of Channels, Sigtuna, Sweden, 2010-09-17, Erik Lindahl

ISBRA2010, Paris, France, 2010-09-13, Erik Lindahl

Biocontrol Workshop, Stockholm, Sweden, 2010-04-27, Erik Lindahl

Institut de Biologie Structurale, Grenoble, France, 2010-03-19, Erik Lindahl

“FunCoup: global networks of functional coupling and medical applications”, Computational Techniques in Medicine, Center for Molecular Medicine, Karolinska Institute, Stockholm, 19 October 2010, Erik Sonnhammer

“FunCoup: global networks of functional coupling”, COMPAS workshop “Cancer Systems Biology for benefit of patients: achievements, challenges and prospective”, Stockholm, 13 April 2010, Erik Sonnhammer

Leibniz Graduate School of Molecular Biophysics Annual Meeting (invited keynote lecture). Freiburg, FRG. March 2010. Gunnar von Heijne

Annual Meeting and Conference of the Canadian Society of Biochemistry, Molecular and Cellular Biology (invited keynote lecture). Banff, Canada. April 2010. Gunnar von Heijne

Distinguished Seminar Series of the Max Planck Institute for Developmental Biology, Tübingen (invited lecture). May, 2010. Gunnar von Heijne

35th FEBS Congress (invited lecture). Gotheburg, Sweden. June 2010. Gunnar von Heijne

Symposium “Evolution of Transport Systems” (invited lecture). Frankfurt, FRG. July 2010. Gunnar von Heijne

18th Meeting on Methods in Protein Structure Analysis (invited lecture). Uppsala, Sweden. August 2010. Gunnar von Heijne

11th conference of the International Society for Endocytobiology (EMBO plenary lecture). Tromsø, Norway. August 2010. Gunnar von Heijne

International Symposium on Protein Community (invited lecture). Nara, Japan. September 2010. Gunnar von Heijne

Frontiers in Membrane and Membrane Protein Biophysics Symposium: Experiments and Theory (invited lecture). Irvine, USA. September 2010. Gunnar von Heijne

TranSys Conference Membrane Protein Structure, Biogenesis and Bioinformatics (invited lecture). Acquafredda di Maratea, Italy. October 2010. Gunnar von Heijne

3rd Annual NIH Roadmap Structural Biology Membrane Protein Technology Meeting (invited keynote lecture). La Jolla, USA. November 2010. Gunnar von Heijne

### **Computer infrastructure**

The SBC employs a very standardized computer system in which each workplace has an identically set up desktop computer. All user disk storage is done at PDC and is accessed via the AFS file system (in 5-10 Gb volumes). Heavy computation is carried out on the 5440-core compute cluster Ferlin, also maintained by PDC. A summary of the infrastructure is listed below.

Desktop computers:

~40 desktops running Ubuntu Linux

Ferlin compute cluster:

In total 5440 cores, 2.66GHz CPUs on 672 8-CPU compute nodes with 8 Gb RAM (shared with SNIC).

Disk servers: 2 servers, ~8 Tb in total

Internal servers: mail, cups, life, mickey

Web servers:

<http://www.sbc.su.se>:

Intel Core2 Quad 2.40GHz, 4 Gb RAM, 250 Gb RAID2 disk, Centos Linux.  
accessed from 9000-15000 unique IP numbers per month.

Hosted services:

- \* PRIMETV: Visualize tree reconciliations
- \* PrIME-GSR: A Bayesian integrated model for genes, sequences, and rates
- \* MapDP: factorizing branchlengths into divergence times and rates
- \* PrIME-GEM: Probabilistic orthology analysis (binaries downloadable)
- \* Pmembr A threading method for membrane proteins.
- \* HMMER High capacity site for use of HMMER to search SCOP or Pfam
- \* ProQ A protein model quality predictor.
- \* PeroxiP Predict peroxisomal proteins and Pfam domains
- \* PRODIV-TMHMM Topology and reentrant predictions.
- \* TMHMMfix TMHMM with optional fixing and reliability score calculation.
- \* DAS Prediction of Transmembrane Regions.
- \* NucPred Nuclear localization prediction.
- \* DRIP-PRED Disorder/order prediction for proteins.
- \* GPCPRED Contact map prediction for proteins.
- \* SVMHC Prediction of MHC class I binding peptides.
- \* PhylProM Phylogenetic profiles

- \* OVOP automatic view generation for protein structures (source code available)
- \* modhmm A modular HMM program used in PRO(DIV)-TMHMM and other studies..
- \* LGscore A program to measure the similarity between proteins.
- \* Palign Our alignment/threading programs.
- \* ssHMM Secondary structure HMMs based on HMMER
- \* LEPRO Protein modelling C++ /library.
- \* TAED The Adaptive Evolution Database.
- \* www.genefun.org GeneFun EU collaboration
- \* www.perlgp.org PerlGP, The Open Source Perl Genetic Programming System
- \* www.socbin.org Society for Bioinformatics in Northern Europe
- \* prime.sbc.su.se Probabilistic Integrated Models of Evolution.
- \* SeqXML.org Standardised format for sequence- and metadata.
- \* OrthoXML.org Standardised format for orthology information.

#### neon.sbc.su.se:

Intel Quad CPU, 4 \* 2.66GHz, 8 Gb RAM, 1 Tb RAID2 disks, Centos Linux

Hosted services:

- InParanoid.sbc.su.se A comprehensive database of orthologs and inparalogs in eukaryotes
- Pfam.sbc.su.se A comprehensive database of protein domain families.
- FunCoup.sbc.su.se Comprehensive protein networks of functional coupling.
- Phobius.sbc.su.se A combined transmembrane topology and signal peptide predictor.
- Avdist: A tool for analyzing haplotype differences.
- Repeatalign: Binary Repeat Align server.
- RefSense: An alternative to Pubmed.
- Excap.sbc.su.se
- facs.sbc.su.se
- modelestimator.sbc.su.se
- octopus.cbr.su.se
- scampi.cbr.su.se
- topcons.cbr.su.se

#### helium.sbc.su.se:

2\*2.80GHz Pentium 4, 2 Gb RAM, 750 Gb RAID2 disks, Centos Linux

Hosted services:

- jSquid.sbc.su.se A java tool to visualize networks and edge scores in FunCoup.
- Sfinx.sbc.su.se Prediction of functional and structural features in proteins.
- Sfixem.sbc.su.se A java viewer for Sfinx.
- GPCRHMM.sbc.su.se A hidden Markov model for GPCR detection.
- Humanoid.sbc.su.se Human ortholog groups and functional shift analysis of subfamilies.
- MSA.sbc.su.se Multiple alignments and assessment of alignment accuracy.
- DASHer.sbc.su.se A Java DAS client for displaying protein sequence annotations.
- MultiParanoid.sbc.su.se
- DAS services for Phobius, signalP, HMMTOP, PhD, Toppred, etc.  
(<http://das.sbc.su.se:9000/das/>\*)
- Funshift.sbc.su.se Functional shift analysis between the subfamilies of a protein domain family.

#### argon.sbc.su.se:

4\* 2.8 GHz Pentium 4, 4 Gb RAM; 1.25 Tb disks, Ubuntu Linux

#### uranium.sbc.su.se:

8\* 2.8 GHz Intel i7, 8 Gb RAM; 2 Tb disks, Ubuntu Linux