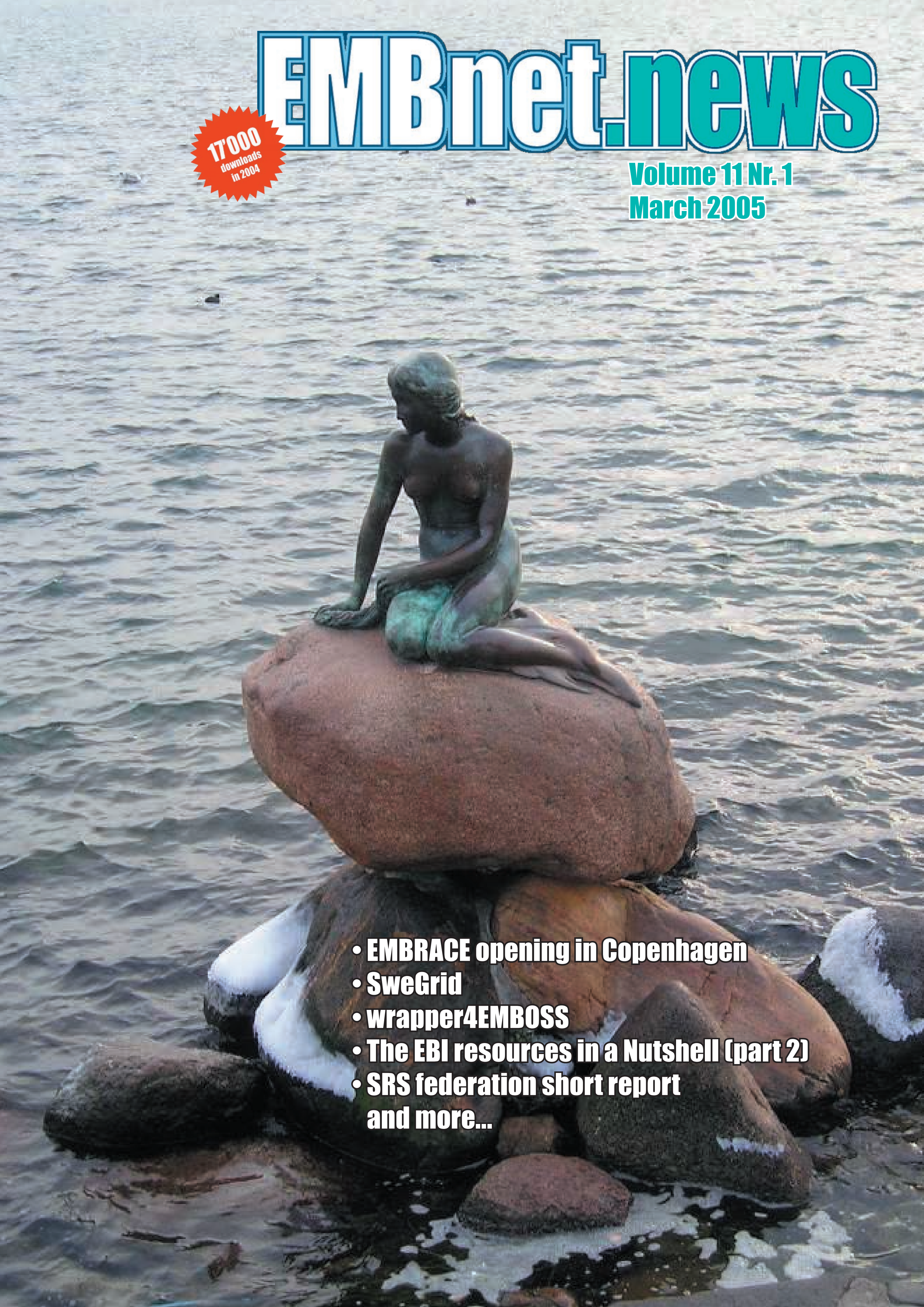
EMBnet.news

17'000 downloads in 2004

Volume 11 Nr. 1
March 2005



- EMBRACE opening in Copenhagen
- SweGrid
- wrapper4EMBOSS
- The EBI resources in a Nutshell (part 2)
- SRS federation short report

and more...

Editorial

As 1993 was the year the World-Wide-Web was born, this year marks the beginning of a new era of computing. Exactly as for WWW, the CERN is at the origin of the new GRID concept. However the GRID means many things for different persons. From the high throughput distributed computing power, to synchronized spreading of data, accessing intelligent user-friendly integrated tools, distributed databases of medical images, and the list is far from exhaustive!
The road is still long to reach the ease of use and acceptance of the web, but let's face it one day we will be able to execute tasks without knowing where the databases are located and in which format they are available, we won't know on which computer in the world the real calculations are done. We will only care about how fast we will get the results...
So with the growing number of projects in Europe and around the world, one can legitimally predict that 2005 is going to be the **Year of the GRID!**

The editorial board: Erik Bongcam-Rudloff, Laurent Falquet, Pedro Fernandes, Oscar Grau, and Gonçalo Guimaraes Pereira.

proteinspotlight

> ONE MONTH, ONE PROTEIN <

Protein Spotlight (ISSN 1424-4721) is a periodical electronic review from the SWISS-PROT group of the Swiss Institute of Bioinformatics (SIB). It is published on a monthly basis and consists of articles focused on particular proteins of interest. Each issue is available, free of charge, in HTML or PDF format at
`http://www.expasy.org/spotlight`

We provide the EMBnet community with a printed version of issues 54&55. Please let us know if you like this inclusion.

Cover picture: The Little Mermaid, Copenhagen, January 2004 [® Nick Daniels, www.nickdaniels.com]

Contents

Editorial Board:

Publisher:

EMBnet Executive Board
c/o Erik Bongcam-Rudloff
Uppsala Biomedical Centre
The Linnaeus Centre for Bioinformatics, SLU/UU
Box 570 S-751 23 Uppsala, Sweden
Email: erik.bongcam@bmc.uu.se
Tel: +46-18-4716696

SweGrid

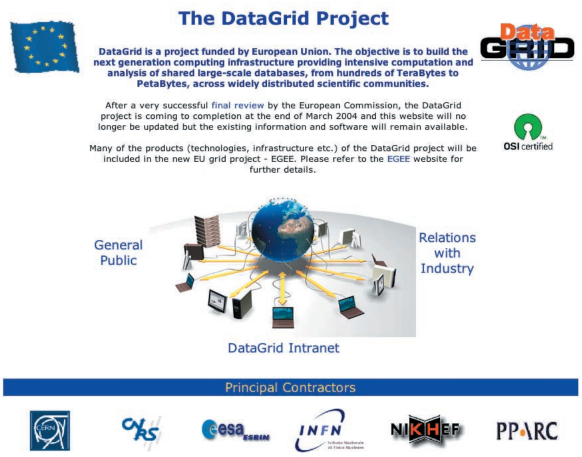
Nils-Einar Eriksson
Swedish EMBnet node, The Linnaeus Centre for Bioinformatics and Uppsala Biomedical Centre, Computing Dept.

Introduction

A few years back it was obvious that the data processing needs of some large-scale scientific projects would far exceed the capacity offered by conventional solutions. For example, huge amounts of data were anticipated from the Large Hadron Collider (LHC), the planned particle accelerator at the CERN, that will start in 2007.

At the same time, the processing power of ordinary desktop computers was reaching interesting levels. The idea of joining lots of computers into a grid of tightly interconnected computers had been around for at least 10 years. Excellent data communications also enabled the distribution of processing tasks and data to widely spread groups of computer resources.

An initiative from the CERN, the **DataGrid** project, got funded by the European Union. Its purpose was to join local and remote computing resources into an infrastructure for powerful computation and analysis of large-scale databases. The DataGrid (now part of **EGEE** EU project) was to be organized according to a multi-tier model with a tier-0 node at CERN and tier-1 nodes in a number of European countries. Tens of thousands of inexpensive commodity computers were to be connected.

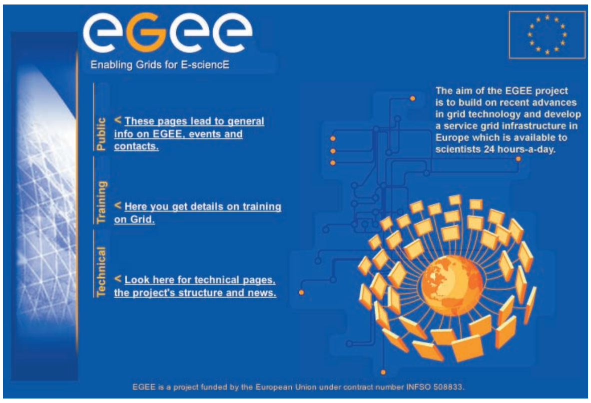


The DataGrid Project

DataGrid is a project funded by European Union. The objective is to build the next generation computing infrastructure providing intensive computation and analysis of shared large-scale databases, from hundreds of TeraBytes to PetaBytes, across widely distributed scientific communities.

After a very successful final review by the European Commission, the DataGrid project is coming to completion at the end of March 2004 and this website will no longer be updated but the existing information and software will remain available.

Many of the products (technologies, infrastructure etc.) of the DataGrid project will be included in the new EU grid project - EGEE. Please refer to the EGEE website for further details.



< These pages lead to general info on EGEE, events and contacts.

< Here you get details on training on Grid.

< Look here for technical pages, the project's structure and news.

The aim of the EGEE project is to build on recent advances in grid technology and develop a service grid infrastructure in Europe which is available to scientists 24 hours-a-day.

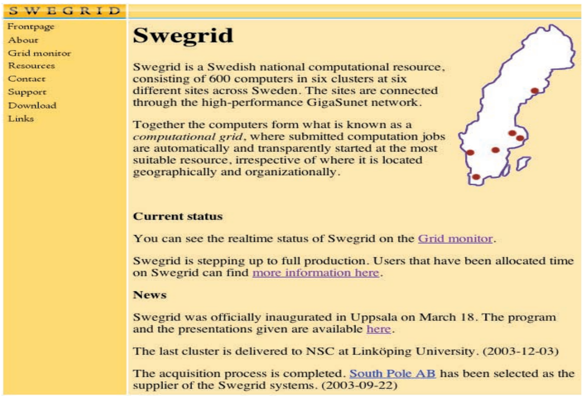
EGEE is a project funded by the European Union under contract number INFSO 508833.

The birth of SweGrid

A Swedish high-energy physicist, Tord Ekelöf, was among those that suggested that the Nordic countries (Denmark, Norway, Finland, Iceland and Sweden) should take an active part in the European grid work. Partly inspired by a successful history of early grid efforts in the **NorduGrid** project, Ekelöf invited representatives from computationally and data-processing intensive Swedish research areas to participate in a proposal to the Knut and Alice Wallenberg foundation. The intention was to build a grid that could serve as a test-bed for a future Nordic tier-1 node.

The proposal resulted in a grant to establish **SweGrid** - a network equipped with 600 computers and generous data storage facilities – in total 120 Terabyte disc storage and 120 Terabyte tape storage, located at six Swedish universities.

Among the SweGrid funding participants



Swegrid

Swegrid is a Swedish national computational resource, consisting of 600 computers in six clusters at six different sites across Sweden. The sites are connected through the high-performance GigaSunet network.

Together the computers form what is known as a *computational grid*, where submitted computation jobs are automatically and transparently started at the most suitable resource, irrespective of where it is located geographically and organizationally.

Current status

You can see the realtime status of Swegrid on the Grid monitor.

Swegrid is stepping up to full production. Users that have been allocated time on Swegrid can find more information here.

News

Swegrid was officially inaugurated in Uppsala on March 18. The program and the presentations given are available here.

The last cluster is delivered to NSC at Linköping University. (2003-12-03)

The acquisition process is completed. South Pole AB has been selected as the supplier of the Swegrid systems. (2003-09-22)

were representatives for bioinformatics. One of the main classes of data processing in bioinformatics involves the analysis of large amounts of data, normally stored in databases. For SweGrid to be useful in this context, database information must be made available to the grid end-nodes.

At one of the first meetings of what would become the EMBRACE project (see next article in this issue) and at EMBnet's AGM 2003 the idea of re-using xNDT, a well tested database distribution mechanism, was suggested (AGM 2003 Minutes, section 7.1, Eriksson).

xNDT is well known to EMBnet managers from the early 90´s. It was created by the former Swedish EMBnet node manager Peter Gad to redistribute databases from the Swedish EMBnet node to a number (6-10) of national EMBnet subnodes. It soon got widespread use among EMBnet nodes all over Europe as shown at one of the conferences 'Genes Proteins and Computers' in Chester, by Peter Rice, EMBnet chairman at the time. xNDT thus got thoroughly tested in a heavy-duty production environment during some years.

The need to keep databases on grid fileservers updated from a central source is very similar to one of the main tasks of most EMBnet nodes. The new and challenging part is the development of flexible mechanisms to provide each end-node of the grid with a selection of data optimised for the analysis to be performed. Efforts of this kind are now being made by the Swedish EMBnet node in collaboration with Peter Gad and a first version of xNDT on SweGrid is now being tested.

A much-wanted feature for SweGrid and similar grids is a dual-queue system. Being restricted to a single multi-hour 'batch-queue' limits its use to well-prepared 'heavy-duty' applications.

More information on the GRID? Visit the GridCafé!

http://gridcafe.web.cern.ch

GridCafé
The place for everybody to learn about the Grid
English | Français
What is the Grid?
How does it work?
What can it do?
A brief history
The Grid and you
Grid @ CERN
Grid projects worldwide
I want to run a job on the GRID!
Can the GRID help my business?
What do experts think about the GRID?
Will it really work?
Check out our GRID in a FLASH animations!
CERN
About I Search I Site Map I Contacts I Credits

- GridCafé is now available in Chinese!
- Other news

Other web addresses

CERN
http://www.cern.ch

Gridhistory
http://gridcafe.web.cern.ch/gridcafe/Gridhistory/ancestors.html

DataGrid project
http://eu-datagrid.web.cern.ch

EGEE project
http://www.eu-egee.org/

Nordugrid
http://www.nordugrid.org

SweGrid proposal
http://www.grid.tsl.uu.se/swegrid/SweGridApp.html

SweGrid
http://www.swegrid.se

EMBnet
http://www.embnet.org

The Linnaeus Centre for Bioinformatics
http://www.lcb.uu.se

Uppsala Biomedical Centre
http://www.bmc.uu.se

The EMBRACE project opening meeting in Copenhagen

Press Release

Hinxton, February 1, 2005 – The Commission of the European Union has awarded €8.3 million to a pan-European task force which will improve access to biological information for scientists throughout and beyond Europe. The EMBRACE Network of Excellence, which encompasses computational biologists from 17 institutes in 11 countries and is coordinated by the European Bioinformatics Institute's Associate Director Graham Cameron, will use these funds to simplify and standardize the way in which biological information is served to the researchers who use it.

Scientists now depend on databases to access the avalanche of information that they produce. For example, geneticists are trawling through the human genome for genes that are involved in diseases. Data providers put a huge amount of effort into providing data resources that are comprehensive, user-friendly and cross-linked to other databases; but different data providers use different methods. This means that a researcher might have to search ten or more different databases to find all the information pertaining to a particular set of candidate genes. If they're doing these kinds of searches on a regular basis, they'll want their own local copies of the databases. Maintaining up-to-date and fully functioning versions of all those databases and the tools to search them is a huge and complex task.

Vincent Breton (CNRS, Clermont-Ferrand, France), a member of EMBRACE's Executive Board, describes the problem as analogous to the use of electrical items before the electrical grid. "You didn't know whether your gadget's plug would fit the socket," he says.

EMBRACE will turn the relationship between user and provider on its head by enabling data providers to provide well-defined interfaces to their databases that will conform to the same standards, essentially creating a 'data grid' – the EMBRACEgrid – that will allow users to make the most of dispersed data resources.

To ensure that EMBRACE's efforts are immediately useful to biologists, Europe's most heavily used biomolecular databases and tools will be integrated into the EMBRACEgrid.

A 'technology watch' will ensure that the EMBRACEgrid doesn't become locked into technology that is quickly superseded. The grid will also receive regular workouts using test problems, such as identifying candidate genes for a disease or linking viral mutations to their ability to cause disease. Disseminating information about the EMBRACEgrid will be vital to ensure that scientists throughout Europe not only use the new technology, but also help to expand the capabilities of the EMBRACEgrid by 'grid enabling' their own data resources.

"Many elegant and powerful computational biology tools are under-utilized," says EMBRACE Executive Board member Erik Bongcam-Rudloff (LCB-SLU, Sweden). "EMBRACE will allow us to unlock their potential by standardizing access to them."

Participants

1 European Molecular Biology Laboratory – European Bioinformatics Institute, EMBL-EBI, UK, Germany
2 Istituto di Tecnologie Biomediche – Sezione di Bari, CNR, ITB-BA, Italy
3 University of Manchester, UMAN, UK
4 Swiss Institute of Bioinformatics, SIB, Switzerland
5 Medical Research Council, MRC, UK
6 Swedish University of Agricultural Sciences Linnaeus Centre for Bioinformatics, SLU-LCB, Sweden
7 Centre National de la Recherche Scientifique, CNRS, France
8 Technical University of Denmark, CBS-DTU, Denmark
9 Consejo Superior de Investigaciones Cientificas, CSIC, Spain
10 Stockholms Universitet, SU, Sweden
11 InstitutNationaldelaRechercheAgronomique, INRA, France
12 Max-Planck-Society, Max-Planck- Institute for Molecular Genetics, MPI-MG, Germany
13 CSC - Scientific Computing Ltd, CSC, Finland
14 University College London,UCL, UK
15 Weizmann Institute of Science, WIS, Israel
16 Stichting Katholieke Universiteit, KUN-CMBI, The Netherlands
17 Instituto Nacional de Tecnica Aerospacial (Centre de Astrobiologia), INTA-CAB, Spain

wrappers4EMBOSS

a fast-and-easy way to integrate BLAST and other 3rd party software under EMBOSS

Guy Bottu
Belgian EMBnet node (BEN)
ULB Campus de la Plaine,
blv. du Triomphe,
1050 Brussels, Belgium

The EMBOSS software package contains a lot of programs, but of course it is not possible to (re)code all the algorithms and functions the users might need. Therefore people have turned to integrating 3rd party software. One way to do this is to take the original source code and modify ("embossify") it ; usually there are the input and output modules that are changed. This is what has been done for most of the "Embassadir" packages (HMMER, PHYLIP, MEME,...). A second way is to write an EMBOSS "wrapper" program; which launches the actual program. This is what emma and eprimer3 do for the CLUSTAL and Primer3 software, respectively. At the Belgian EMBnet Node we had already a long tradition of "wrapperizing" software under GCG. In the course of 2002, I hurried to integrate most of these programs under EMBOSS, so that the users of BEN could continue to use their familiar tools.

A BEN wrapper program operates roughly the following way :

- As always, the program begins with the embInit subroutine (which parses the ACD file and the command line and makes sure the input data and parameters are all correct).

- The input sequences are retrieved by the USA mechanism and are stored in an appropriate format (usually fastA but sometimes MSF) in a "temporary" file.

- The program generates a command line for the "naked program below" and launches it.

- If useful, the output file and/or the screen output is processed by a Perl script, in order to remove unneeded or confusing items. Typically, if a line says that the input sequence was /tmp/seq-something, this is replaced by the original USA.

- If the output file contains a set of sequences, these are stored in a "temporary" file and then finally exported in the format chosen by the user, using the EMBOSS "SeqWrite" subroutines.

- If useful, the program will, if it detects an X-Window environment (variable DISPLAY set), open the output file with an appropriate graphical viewer.

During EMBnet meetings collaborators of other nodes showed interest for my work. Especially the possibility of having BLAST under EMBOSS seemed to trigger a great demand. Unfortunately, in their original form the "wrappers" were not readily portable. To mention just one thing, the names of the available "public" databanks in BLAST format are written as such in the ACD file. The ideal would be a subroutine that detects these databanks on-the-fly. Writing such a routine is not a trivial task. The EMBOSS development team showed interest in doing it, but we could of course not wait, and there is the supplementary problem of making this work seamlessly under the GUI's. It is then that Martin Sarachu from the Argentinian EMBnet node, who was already collaborating with us on the development of wEMBOSS (see article in the previous issue of EMBnet.news), decided to devote himself to the problem of making the "wrappers" portable without making the code much more complicated and without tinkering with the installation of EMBOSS itself. He finally came up with the following solution : the local manager must first install EMBOSS as well as the programs to be "wrapped" and make sure that, at least at installation time, they are in the "path".

Then he must run an installation script, which does the following things :

- ask interactively a number of general things : which (groups of) "wrappers" to install, if the "temporary" area to use is /tmp of something else, etc.

- detect the location of the "naked" programs

- for each (group of) "wrappers" ask for a number of specific things, e.g. which databanks in BLAST format are available and which one should be selected by default

- modify the C code files, ACD files and Perl scripts as needed by the local configuration

- modify the EMBOSS makefiles and compile the programs

This of course entails that each time something is changed to EMBOSS, to the "naked" program or to the collection of available databanks it will be necessary to re-run the install script again.

The present distribution of wrappers4EMBOSS contains the following groups (the local manager can select which one(s) he wants to install) :

Interface to the NCBI BLAST suite

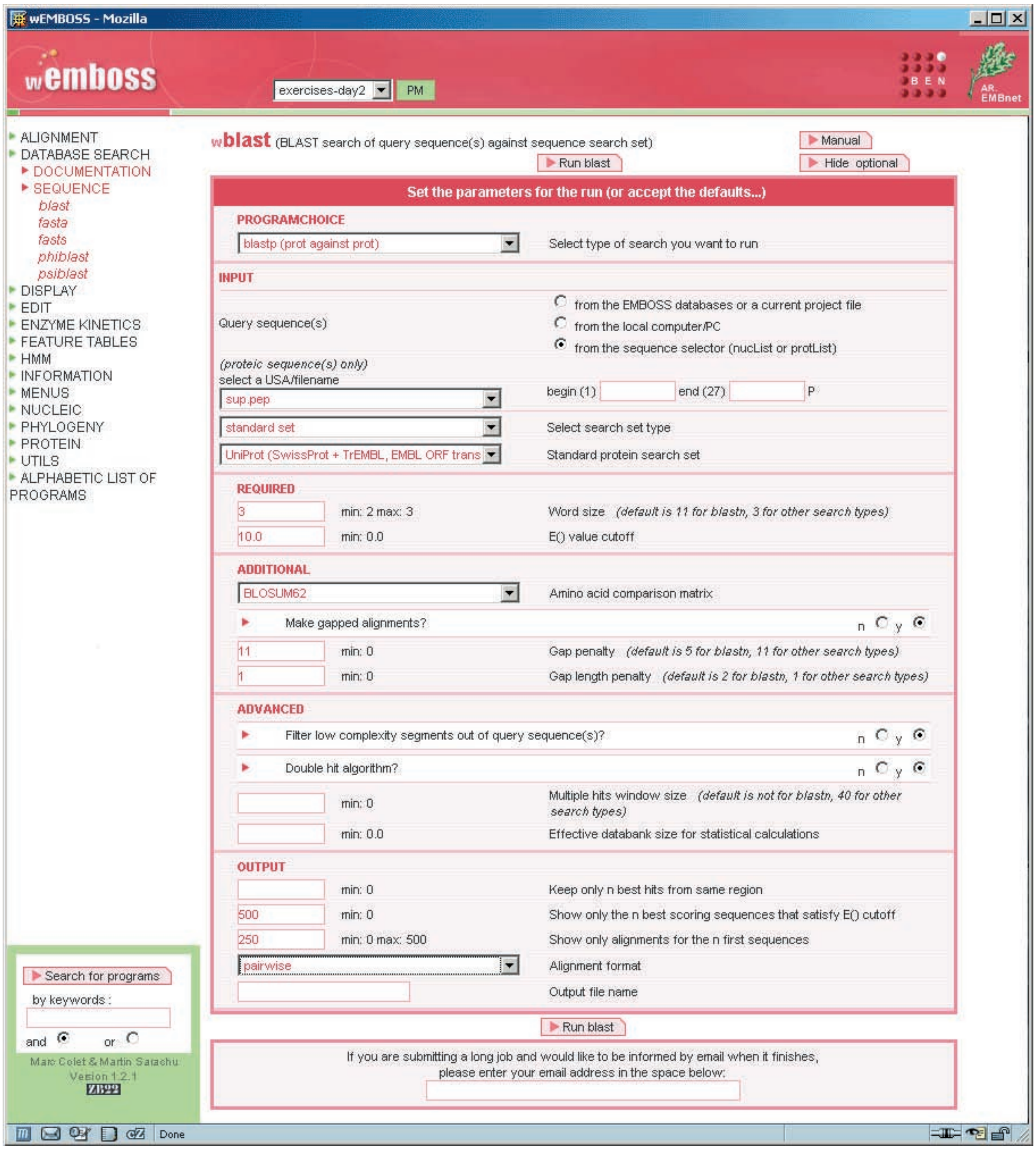
blast : "wrapper" for blastall, options blastn, blastp, blastx, tblastn and tblastx. It performs a similarity search of a sequence against a sequence databank in any combination nucleic acid/protein. It is possible to submit several query sequences at the same time, the "wrapper" will submit them one after the other.

psiblast : "wrapper" for blastpgp operating in PSI-BLAST mode and for blastall, option psitblastn. It searches a protein sequence against a protein databank and generates a protein profile from the best "hits" ; the search is performed iteratively until it converges or otherwise up to a maximum number of rounds (by default 10). The profile can be exported and searched against another protein databank or against a nucleic acid databank.

phiblast : "wrapper" for blastpgp operating in PHI-BLAST (or combined PHI-BLAST + PSI-BLAST) mode. It searches a protein sequence together with a PROSITE style protein pattern against a protein databank.

blast2seq : "wrapper" for bl2seq, makes local alignments between 2 sequences

makeblastdb : "wrapper" for formatdb, makes a BLAST format databank from sequences



wblast (BLAST search of query sequence(s) against sequence search set)

Set the parameters for the run (or accept the defaults...)

PROGRAMCHOICE

blastp (prot against prot) — Select type of search you want to run

INPUT

Query sequence(s)

○ from the EMBOSS databases or a current project file
○ from the local computer/PC
● from the sequence selector (nucList or protList)

(proteic sequence(s) only)
select a USA/filename: sup.pep — begin (1) end (27) P

standard set — Select search set type

UniProt (SwissProt + TrEMBL, EMBL ORF trans — Standard protein search set

REQUIRED

3 — min: 2 max: 3 — Word size (default is 11 for blastn, 3 for other search types)

10.0 — min: 0.0 — E() value cutoff

ADDITIONAL

BLOSUM62 — Amino acid comparison matrix

Make gapped alignments? n ○ y ●

11 — min: 0 — Gap penalty (default is 5 for blastn, 11 for other search types)

1 — min: 0 — Gap length penalty (default is 2 for blastn, 1 for other search types)

ADVANCED

Filter low complexity segments out of query sequence(s)? n ○ y ●

Double hit algorithm? n ○ y ●

min: 0 — Multiple hits window size (default is not for blastn, 40 for other search types)

min: 0.0 — Effective databank size for statistical calculations

OUTPUT

min: 0 — Keep only n best hits from same region

500 — min: 0 — Show only the n best scoring sequences that satisfy E() cutoff

250 — min: 0 max: 500 — Show only alignments for the n first sequences

pairwise — Alignment format

Output file name

If you are submitting a long job and would like to be informed by email when it finishes, please enter your email address in the space below:

Figure 1 : An example : Blast under wEMBOSS. The user intends to search a 27 amino acids long peptide against the UniProt databank.

provided by the user, either as a file in fastA format or with an EMBOSS USA.

For the databank searches, the user can select one of the "public" databanks installed by the manager, but he can also search a "private" databank or provide a set of sequences with an EMBOSS USA (which will be transformed on-the-fly into a databank).

Interface to W. Pearson's FastA suite.

fastasearch : "wrapper" for fasta, fastx, fasty, tfasta, tfastx, tfasty and ssearch.

fastapid : "wrapper" for fasts, fastf, tfasts and tfasf. "pid" stands for "protein identification"; it searches a set of peptides from a protein

```
#  indexsearch output (EMBOSS List File)
#  Mon Jan 17 16:57:19 2005
#   query  :  [dbs={uniprot_swissprot  uniprot_trembl}-des:protease*  |
proteinase*] & [dbs-org:carica* & papaya* | carica papaya*]
#  10 entries found
UNIPROT_SWISSPROT:LSPI_CARPA
# Latex serine proteinase inhibitor.
UNIPROT_SWISSPROT:PAP2_CARPA
# Chymopapain precursor (EC 3.4.22.6) (Papaya proteinase II) (PPII).
UNIPROT_SWISSPROT:PAP3_CARPA
# Caricain precursor (EC 3.4.22.30) (Papaya proteinase omega) (Papaya
# proteinase III) (PPIII) (Papaya peptidase A).
UNIPROT_SWISSPROT:PAP4_CARPA
# Papaya proteinase IV precursor (EC 3.4.22.25) (PPIV) (Papaya peptidase
# B) (Glycyl endopeptidase).
UNIPROT_SWISSPROT:PAP5_CARPA
# Cysteine proteinase (EC 3.4.22.-) (Clone PLBPC13) (Fragment).
UNIPROT_SWISSPROT:PAPA_CARPA
# Papain precursor (EC 3.4.22.2) (Papaya proteinase I) (PPI).
UNIPROT_TREMBL:Q670P1
# Proteinase omega (EC 3.4.22.6) (Fragment).
UNIPROT_TREMBL:Q7M1Q8
# Proteinase omega (Fragments).
UNIPROT_TREMBL:Q42673
# Papaya proteinase omega (EC 3.4.22.6).
UNIPROT_TREMBL:Q39561
# Cysteine proteinase inhibitor (Cystatin).
```

Figure 2 : A (short) example of an indexsearch output.

digest against a sequence databank.

fasta2seq : "wrapper" for lfasta and plfasta from the older fastA suite version 2, makes local alignments between 2 sequences. By default the "wrapper" produces an output in text format, but the user can request instead a graphic in PostScript format, which under X-Window is opened with GhostView.

Interface to the ClustalW program

clustal : takes input sequences and makes a multiple sequence alignment

clustalnj : takes as input already aligned sequences and makes a phylogenetic tree by the Neighbour-Joining method. It has options for Kimura distance correction and for "bootstrapping". Under X-Window the tree is opened with NJplot.

CLUSTAL had been implemented at the BEN site as an alternative for emma, which at the time looked very unsatisfactory. With release 2.9.0 of EMBOSS emma has been much improved, but the clustalnj wrapper remains very useful.

Interface to the pftools package

For searching a protein against the PROSITE databank, uses the Perl script ps_scan.pl from the ExPASy web site. The manager also needs to install P. Bucher's PFTOOLS suite.

ps_scan : searches a protein sequence against the complete PROSITE databank (patterns + profiles + rules). The user can provide a motif databank of his own (with patterns in PROSITE syntax and/or profiles in Bucher format) and/or restrict the search to selected motifs. Note that the EMBOSS program `patmatmotifs` searches only the PROSITE patterns.

pf_make : "wrapper" for PFTOOLS pfmake, takes as input a sequence alignment and makes a profile

Blocks

For searching a protein or nucleic acid sequence against the FCHRC Blocks protein motif databank, uses the FCHRC BLIMPS suite. This group only contains program **bscan**.

The special story of indexsearch

Recently a 6th group was added to wrappers4EMBOSS, **indexsearch**, a "wrapper" for the SRS internal search engine **getz**, which reproduces the functionality of the GCG program **lookup**. This program has a whole history of its own. With GCG version 8.1 appeared lookup, based on the code of SRS version 4. lookup has less functionality than SRS, but has one advantage : it stores the result of a search as a sequence "List File" in the UNIX working directory of the user, where it can directly be used by other GCG programs. The problem was that people might not like to waste disk space and CPU by having both a native SRS and lookup make indices in parallel. Reinhard Doelz, the ineffable manager of the Swiss EMBnet Node in the days that it was in Basel rather than in Lausanne, found a way to "hack" lookup so that it could use the indices made by SRS. With the arrival of SRS 5 the format of the "section files" (binary files with information about the databanks) changed. I found a way to make a functional system with the lookup of GCG 8 or 9, SRS 5 with indices, SRS4 without indices, and a lot of "hacking". A proposal to publish a "how to" in this magazine was rejected because of "too much proprietary stuff". With the arrival of SRS 6 the format of the indices changed. I then developed a Perl script to "wrap" getz and called it indexsearch (this gets rid of the proprietary name lookup and furthermore makes a nice contrast with GCG stringsearch and EMBOSS textsearch, both of which search in the text rather than in indices). Finally, when BEN and other EMBnet Nodes made the transition of GCG to EMBOSS, I adapted it to EMBOSS.

indexsearch.pl can be run as such in a simple text terminal. It presents successive screens for databank choice, query composition and output management. A small EMBOSS program indexsearch (with accompanying ACD file and system call to indexsearch.pl) allows for integration behind a graphical user interface. The user can choose between 3 different output formats :

- a simple List File with just the names of the sequences
- a List File with (outcommented) description lines added. This reproduces the output format of GCG lookup and of course helps the user in the further handling of the result. Note that indexsearch uses a call to SRS with custom view EMBOSS _ List _ File. To make it work the local manager must add to the SRSSITE/views.i and SRSSITE/loader.i a few extra lines.
- a multiple sequence file in fastA format

About wEMBOSS and other GUI's

wrappers4EMBOSS has been optimized to operate under wEMBOSS and both packages are distributed together (you can download them from the Web site `http://www.wemboss.org`). They are however not tied and the local manager can perfectly decide to install the one without the other. The "wrappers" run well at the command line, except for a cosmetic problem with makeblastdb (the default basename for a databank derived from a sequence set is not automatically set to myblastdb[1]). As for the many other GUI's that have been proposed for EMBOSS, it depends very much on how good they are in handling the computed attributes in the ACD syntax. In our experience it did not work well with (the early version of) Luke McCarthy's EMBOSS-GUI[2], but it should work with Jemboss and I managed to make it work with Staden (at the price of some "hacking" to get around data type name changes in the later versions of EMBOSS[3]).

References

- Clamp M., Cuff J., Searle S. M. and Barton G. J. (2004).
  «The Jalview Java Alignment Editor»
  Bioinformatics, 12, 426-7.
- Sarachu M. and Colet M. (2004).
  «wEMBOSS: a web interface for EMBOSS»
  Bioinformatics. 2005 Feb 15;21(4):540-1.
- Zmasek C.M. and Eddy S.R. (2001).
  «ATV: display and manipulation of annotated phylogenetic trees»
  Bioinformatics, 17, 383-384.

[1] The reason is that the expression "@(@($(seqtype) == 1) ? myblastdb : $(userfastadb))" produces an error message when userfastadb has a NULL value. wEMBOSS overrides the evaluation of this expression by setting –outfilebasename=myblastdb on the command line.

[2] http://bioinfo.pbi.nrc.ca/~lukem/EMBOSS/

[3] The Tcl/Tk files for Staden spin must be generated from altered ACD files in which "standard" and "additional" have been replaced by "required" and "toggle" by "boolean".

SRS Federation

a progress report

Last September many EMBnet delegates attended the SRS Federation workshop[1] we organised in Elewijt (Belgium), in conjunction with the EMBnet AGM2004. During these two days, with the input from Thure Etzold and his colleagues from Lion Bioscience, we designed the backbone of a new kind of infrastructure in which the EMBnet nodes would collectively build an architecture to access permanently updated databanks. With the support of LION, three nodes (Belgium, Slovakia and Sweden) began in October the exercise consisting in building a network structure where each member would take care of indexing only part of the databanks, the complement being made available by the other members. Around the turn of the year, Brazil, Colombia and Poland joined the effort. Today, for OMIM and UNIPROT, members of the SRS federation actively build "tar-balls" consisting of daily datasets together with the corresponding SRS index files and these datasets are shared among the group. Further efforts are still needed today in order to generalise this data sharing schema to all six datasets that were chosen for the experiment (EMBL, UNIPROT, PDB, PROSITE, KEGG and OMIM). Once this stage has been reached, the SRS Federation will be ready to offer its products to the other members of EMBnet and allow them to build a local functional SRS server while not having to sustain the burden of building all indices locally. The next phases of the SRS Federation will consist in establishing the appropriate level of regional redundancy, QoS, security and other developments presently in the pipeline of SRS development at LION.

Keep an eye on the SRS Federation website (http://www.srsfed.org) for information about the further developments of the project. We are confident that there is a future for this infrastructure, both inside EMBnet and beyond.

Robert Herzog
Belgian EMBnet Node

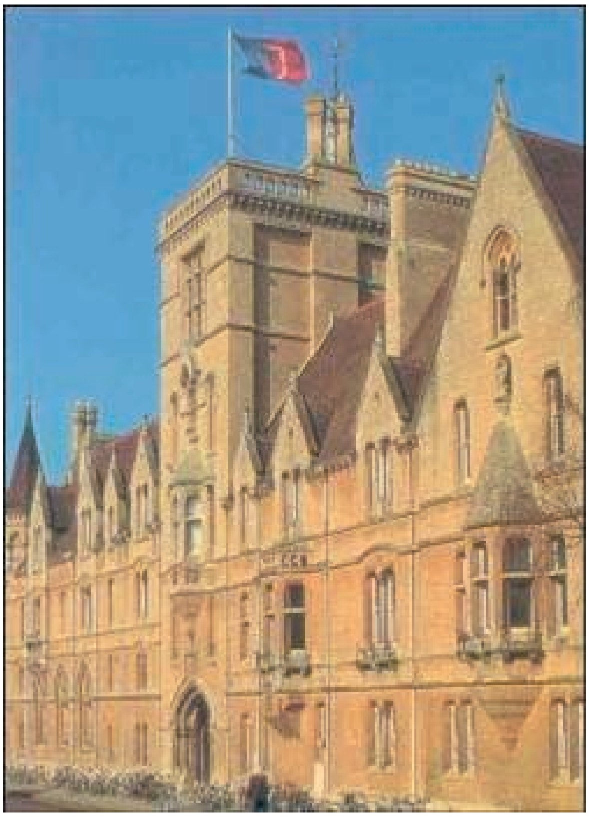
[1] See EMBnet.news 10.4, Dec. 2004.

Announcement

HealthGRID 2005
7th - 9th April
Oxford, UK

The registration for the HealthGrid 2005 conference is open.

Healthgrid 2005 will provide a forum for GRID projects in the medical, biological and biomedical domains, as well as for grid projects that seek to integrate these. With an emphasis on results, it is hoped that functionality being developed within such projects will be demonstrated. The overall objective is to reinforce and promote the awareness of the deployment of grid technology in health. Participation is encouraged across the spectrum of bio/medical informatics from technology developers and researchers through research network representatives and regional health authorities to healthcare clinicians and administrators. For more information please see the conference web site. (http://oxford2005.healthgrid.org/)



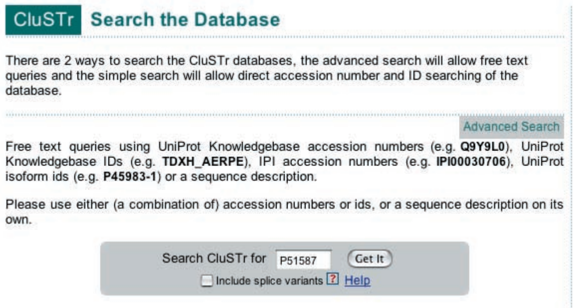
The EBI resources in a Nutshell (part 2)

Lisa Mullan
EMBL Outstation - Hinxton,
European Bioinformatics
Institute, Wellcome
Trust Genome Campus,
Hinxton,
Cambridge, CB10 1SD,
United Kingdom

This practical was designed for the EBI's Small to Medium Sized Industry Programme and aims to offer a practical overview of some of what the EBI can offer in terms of bioinformatics resources in the form of web services (see Part 1 in previous EMBnet.news issue 10.4).

The CLuSTr database

Access the CluSTr database at `http://www.ebi.ac.uk/clustr/`. Follow the Search link on the left hand side of the page and enter the SwissProt accession number for the protein (try with P51587 or BRCA2 _ HUMAN) into the appropriate search field and hit the "Get It" button. It may take some time to retrieve the results of the query as the search is done on the fly. The Z scores will increase towards the right as the clusters get smaller.



CluSTr Search the Database

There are 2 ways to search the CluSTr databases, the advanced search will allow free text queries and the simple search will allow direct accession number and ID searching of the database.

Advanced Search

Free text queries using UniProt Knowledgebase accession numbers (e.g. **Q9Y9L0**), UniProt Knowledgebase IDs (e.g. **TDXH_AERPE**), IPI accession numbers (e.g. **IPI00030706**), UniProt isoform ids (e.g. **P45983-1**) or a sequence description.

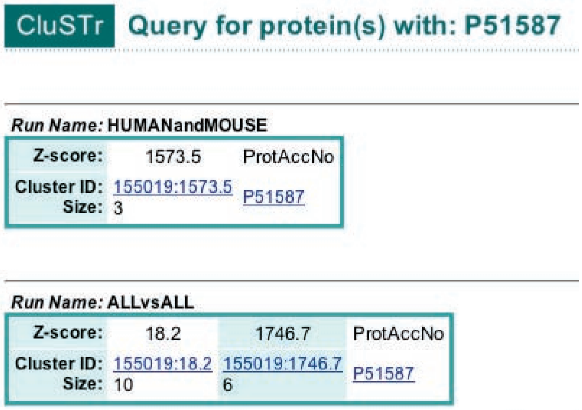
Please use either (a combination of) accession numbers or ids, or a sequence description on its own.

Search CluSTr for P51587 Get It
□ Include splice variants ? Help

The upper box is a search of the protein within Human and Mouse species only. The lower table details the results of an all-against-all search. The Z scores are created for each cluster and remain the same provided no new proteins are added. Proteins are re-clustered when the database is updated.

The identification number of a cluster is dynamic and created each time the database is searched. It is composed of a protein accession number together with the Z score of the cluster. The accession number used is the same one that is used to search the database. If for some reason the information in the database has been accessed by other means, then the protein ID will be selected from one of the proteins in the cluster. Thus there could be more than one ID for the same cluster.

Look at the results of the ALL against ALL search. There are two clusters visible.



CluSTr Query for protein(s) with: P51587

Run Name: HUMANandMOUSE

Z-score:	1573.5	ProtAccNo
Cluster ID: Size:	155019:1573.5 3	P51587

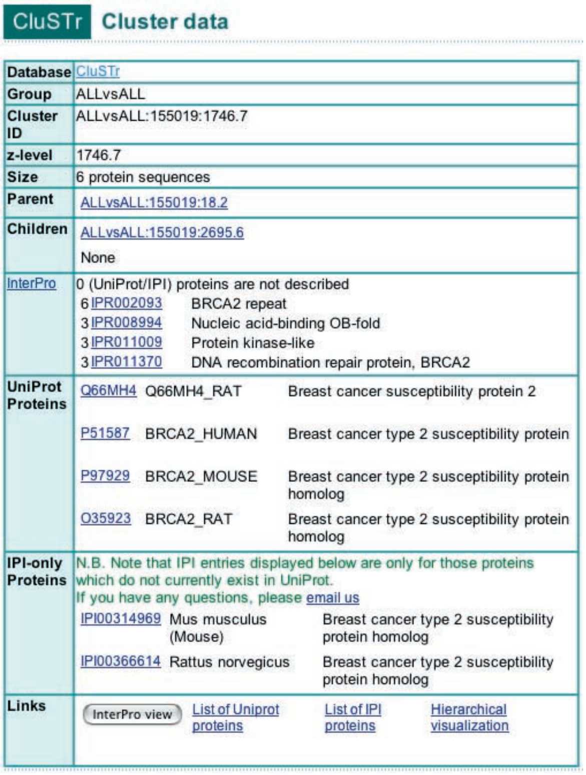
Run Name: ALLvsALL

Z-score:	18.2	1746.7	ProtAccNo
Cluster ID: Size:	155019:18.2 10	155019:1746.7 6	P51587

The smallest cluster (the one with the highest Z score) contains six proteins. However, the CluSTr database uses all the information available in Uniparc1, but only reports those proteins that are annotated by SPTrEMBL or IPI – thus the eventual protein count may be lower than initially suggested.

Follow the link to the cluster represented by the highest Z score in the ALL against ALL search (155019-1746.7). There are 6 proteins in this cluster, with only 5 represented – three from Uniprot and two from the IPI database.

The Clustr database is a repository for storing clustered sets of proteins. Each protein is aligned in a pair-wise fashion against all others in the database. Those scoring above a specified threshold undergo Monte Carlo simulation to determine homology. Homologous proteins are then clustered into groups, moving up the hierarchy from singletons to large families. Each cluster receives a Z score resulting from the simulations. Pair-wise alignments that do not achieve a Z score of at least ten are not included in any subsequent clustering.



CluSTr Cluster data

Database	CluSTr
Group	ALLvsALL
Cluster ID	ALLvsALL:155019:1746.7
z-level	1746.7
Size	6 protein sequences
Parent	ALLvsALL:155019:18.2
Children	ALLvsALL:155019:2695.6 None
InterPro	0 (UniProt/IPI) proteins are not described 6 IPR002093 BRCA2 repeat 3 IPR008994 Nucleic acid-binding OB-fold 3 IPR011009 Protein kinase-like 3 IPR011370 DNA recombination repair protein, BRCA2
UniProt Proteins	Q66MH4 Q66MH4_RAT Breast cancer susceptibility protein 2 P51587 BRCA2_HUMAN Breast cancer type 2 susceptibility protein P97929 BRCA2_MOUSE Breast cancer type 2 susceptibility protein homolog O35923 BRCA2_RAT Breast cancer type 2 susceptibility protein homolog
IPI-only Proteins	N.B. Note that IPI entries displayed below are only for those proteins which do not currently exist in UniProt. If you have any questions, please email us IPI00314969 Mus musculus (Mouse) Breast cancer type 2 susceptibility protein homolog IPI00366614 Rattus norvegicus Breast cancer type 2 susceptibility protein homolog
Links	InterPro view List of Uniprot proteins List of IPI proteins Hierarchical visualization

The individual hierarchies can be investigated using the parent and child links to move up and down the clusters.

The upper "Children" link points to the cluster which has accepted our search protein as a homologue. The "none" specification on the lower "Children" display is a likely indication that the protein added to achieve the current cluster was a singleton. Follow the upper link (ALLvsALL:155019:2695.6) and the five annotated proteins that formed the previous cluster – i.e. the one that accepted our protein as a homologue - will be visible.

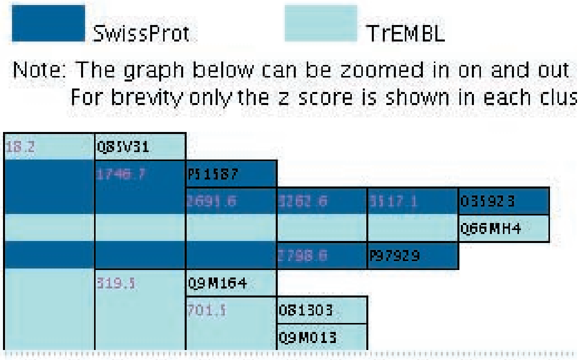
Continue to follow the lower link to the children and a pattern based on the clustering – and therefore the similarity of proteins to each other will emerge - offering a potential insight into their evolution. Notice the Interpro information on specific motifs and domains. All proteins taken from SpTr in the cluster have BRCA2 repeats and two further domains in common. The IPI proteins share only the BRCA2 repeats.

As the OB fold domain is highly contributory to the DNA repair family, it is expected that the IPI proteins would not show this latter family domain.

Return to the original 6 protein cluster (although it will display only 5) and follow the "Hierarchical visualisation" link. This will show all those proteins obtained from the Uniprot, SwissProt and TrEMBL databases and the relative clustering. Click on a score to zoom in further information.

Cluster hierarchies are only displayed for those clusters that were constructed from similarities based on a maximum proteomic group of 10,000 peptides. Clusters above this threshold are not displayed as the graphic would be too large to render effectively.

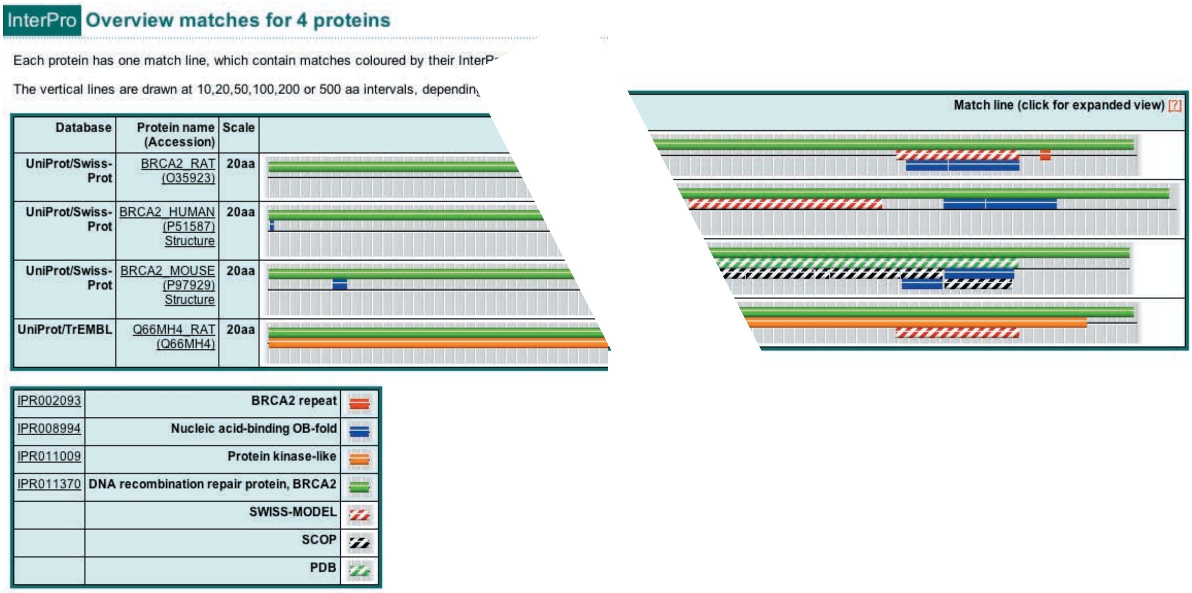
Follow the link from the left most cluster (with the same Z score as the P51587 cluster).



Note: The graph below can be zoomed in on and out
For brevity only the z score is shown in each clus

The display shows how two cluster sets have been merged together to form the second (and lower scoring) ALL against ALL cluster originally noted after the initial search. A search for the proteins in the lowest cluster, Q9M013 and O81303, in Uniprot reveals that these are Arabidopsis proteins also containing the BRCA2 repeats and a nuclear-binding OB fold. The protein Q9M164 clustered in the next hierarchy does not contain either the repeat region or the OB fold. A separate alignment of this sequence with O81303 displays a region of similarity over 324 residues, none of which are involved in the repeat or fold regions.

Follow the InterPro link from the original 6 protein cluster. This offers a graphical display of the protein features for the Uniprot entries within this cluster. Mousing over the features



InterPro Overview matches for 4 proteins

Each protein has one match line, which contain matches coloured by their InterP

The vertical lines are drawn at 10,20,50,100,200 or 500 aa intervals, dependin

Database	Protein name (Accession)	Scale	Match line (click for expanded view) [?]
UniProt/Swiss-Prot	BRCA2_RAT (O35923)	20aa	
UniProt/Swiss-Prot	BRCA2_HUMAN (P51587) Structure	20aa	
UniProt/Swiss-Prot	BRCA2_MOUSE (P97929) Structure	20aa	
UniProt/TrEMBL	Q66MH4_RAT (Q66MH4)	20aa	

IPR002093	BRCA2 repeat	🟥
IPR008994	Nucleic acid-binding OB-fold	🟦
IPR011009	Protein kinase-like	🟧
IPR011370	DNA recombination repair protein, BRCA2	🟩
	SWISS-MODEL	▨
	SCOP	▨
	PDB	▨

will indicate their function and location within the protein.

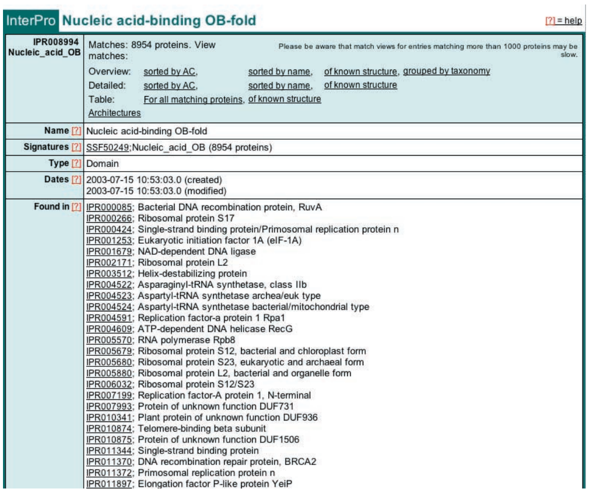
InterPro is an integrated documentation resource for protein families, domains and sites. InterPro combines a number of databases (referred to as member databases) that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful integrated diagnostic tool. InterPro unifies: · PROSITE, home of regular expressions and profiles · Pfam, SMART, TIGRFAMs, PIR SuperFamily and SUPERFAMILY keepers of hidden Markov models (HMMs) · PRINTS, provider of fingerprints (groups of aligned, un-weighted motifs).

Signatures describing the same protein family, domain repeat or site are grouped into unique InterPro entries. Each combined InterPro entry has a unique accession number, an abstract describing the features of proteins associated with the entry and literature references and has links to the relevant member database(s). All UniProt protein sequences that have matches to a particular InterPro entry are listed in the Match Table associated with that entry. There are also links to the InterPro graphical views. The graphical views, which can be sorted by UniProt accession number, structure or taxonomy, show the position of the signatures on the protein, mousing over the signature brings up a pop-box, giving the accession, name and position.
Interpro graphically represents the location of a protein domain and information pertaining to the origin of that domain and the proteins that contain it. Families are also defined and may contain several interpro domains which are often, but not always, in the same order.
InterPro and InterProScan are accessible for interactive use over the EBI web server (www/ebi.ac.uk/interpro), they are distributed as stand-alone copies by anonymous ftp.

The display shows the repeats, OB folds and DNA repair family that have been noted in the CluSTr protein sets.

The "nuclear acid-binding OB folds" on each of the proteins are denoted by a bar 🟦 . There is a small fold region at the start of these proteins and then three larger regions towards the C terminal end of the protein. Conservation across the three available proteins may suggest that this fold is necessary for protein function. Background analysis concludes that the final domain region in the Human protein is in exactly the same place as an alternative splicing mechanism causing an additional exon in the gene.

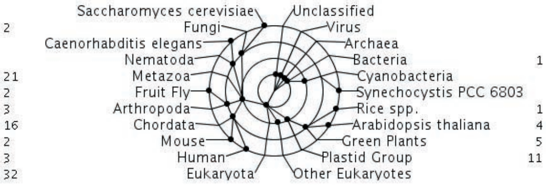
Click on the OB fold bar to see more information on this and other features. The Interpro entry IPR008994 details the OB fold domain as it constitutes a true match to the SUPERFAMILY database. The regions correspond to those found on the human protein. Note the reference to the PDB from this region.

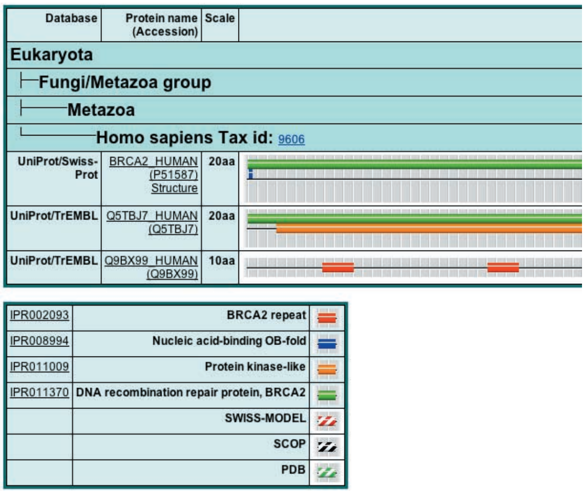


IPR008994 Nucleic_acid_OB	Matches: 8954 proteins. View matches: Please be aware that match views for entries matching more than 1000 proteins may be slow. Overview: sorted by AC, sorted by name, of known structure, grouped by taxonomy Detailed: sorted by AC, sorted by name, of known structure Table: For all matching proteins, of known structure Architectures
Name [?]	Nucleic acid-binding OB-fold
Signatures [?]	SSF50249;Nucleic_acid_OB (8954 proteins)
Type [?]	Domain
Dates [?]	2003-07-15 10:53:03.0 (created) 2003-07-15 10:53:03.0 (modified)
Found in [?]	IPR000085; Bacterial DNA recombination protein, RuvA IPR000266; Ribosomal protein S17 IPR000424; Single-strand binding protein/Primosomal replication protein n IPR001253; Eukaryotic initiation factor 1A (eIF-1A) IPR001679; NAD-dependent DNA ligase IPR002171; Ribosomal protein L2 IPR003512; Helix-destabilizing protein IPR004522; Asparaginyl-tRNA synthetase, class IIb IPR004523; Aspartyl-tRNA synthetase archea/euk type IPR004524; Aspartyl-tRNA synthetase bacterial/mitochondrial type IPR004591; Replication factor-a protein 1 Rpa1 IPR004609; ATP-dependent DNA helicase RecG IPR005570; RNA polymerase Rpb8 IPR005679; Ribosomal protein S12, bacterial and chloroplast form IPR005680; Ribosomal protein S23, eukaryotic and archaeal form IPR005880; Ribosomal protein L2, bacterial and organelle form IPR006032; Ribosomal protein S12/S23 IPR007199; Replication factor-A protein 1, N-terminal IPR007993; Protein of unknown function DUF731 IPR010341; Plant protein of unknown function DUF936 IPR010874; Telomere-binding beta subunit IPR010875; Protein of unknown function DUF1506 IPR011344; Single-strand binding protein IPR011370; DNA recombination repair protein, BRCA2 IPR011372; Primosomal replication protein n IPR011897; Elongation factor P-like protein YeiP

Do the same for the BRCA2 repeat region. There is one region on the Human protein in the same area as one of these regions that is related to a structure in the PDB.

Follow the Interpro link on the accession number for this repeat to access more detailed information on this domain. Scroll down to the taxonomy matches and click on the Human link.





Database	Protein name (Accession)	Scale	
Eukaryota			
├─Fungi/Metazoa group			
├──Metazoa			
└───Homo sapiens Tax id: 9606			
UniProt/Swiss-Prot	BRCA2_HUMAN (P51587) Structure	20aa	
UniProt/TrEMBL	Q5TBJ7_HUMAN (Q5TBJ7)	20aa	
UniProt/TrEMBL	Q9BX99_HUMAN (Q9BX99)	10aa	

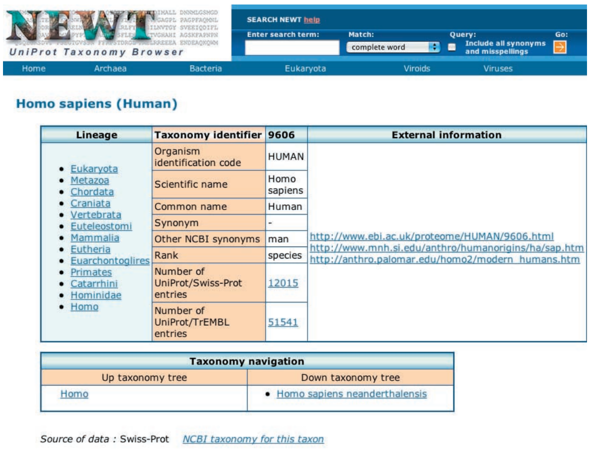
IPR002093	BRCA2 repeat	🟥
IPR008994	Nucleic acid-binding OB-fold	🟦
IPR011009	Protein kinase-like	🟧
IPR011370	DNA recombination repair protein, BRCA2	🟩
	SWISS-MODEL	▨
	SCOP	▨
	PDB	▨

This displays the 3 proteins within the Human proteome which have thus far been shown to contain this particular domain - the original protein and two further human proteins found in the UniProt/TrEMBl database.

The NEWT database

Follow the link at the top of the graphical display to the "Homo sapiens Tax id" and the NEWT database.

The NEWT database is a compilation of the information within the NCBI Taxonomy database together with proteins found in the SwissProt section of UniProt. It is maintained by the SwissProt group in Switzerland. For each species, NEWT displays the following taxonomy data: Swiss-Prot scientific name, Swiss-Prot common name and Swiss-Prot synonym, lineage, number of protein sequence entries in Swiss-Prot and TrEMBL.



Homo sapiens (Human)

Lineage	Taxonomy identifier	9606	External information
• Eukaryota • Metazoa • Chordata • Craniata • Vertebrata • Euteleostomi • Mammalia • Eutheria • Euarchontoglires • Primates • Catarrhini • Hominidae • Homo	Organism identification code	HUMAN	http://www.ebi.ac.uk/proteome/HUMAN/9606.html http://www.mnh.si.edu/anthro/humanorigins/ha/sap.htm http://anthro.palomar.edu/homo2/modern_humans.htm
	Scientific name	Homo sapiens	
	Common name	Human	
	Synonym	-	
	Other NCBI synonyms	man	
	Rank	species	
	Number of UniProt/Swiss-Prot entries	12015	
	Number of UniProt/TrEMBL entries	51541	

Taxonomy navigation	
Up taxonomy tree	Down taxonomy tree
Homo	• Homo sapiens neanderthalensis

Source of data : Swiss-Prot *NCBI taxonomy for this taxon*

As this is a human protein, the human taxonomy ID will be given. Currently there are 12015 SwissProt entries arising from the human genomic sequence. This will change as more proteins are discovered and annotated.

The lineage of Homo sapiens is displayed on the left hand side of the view.

The NEWT data is available from the European Bioinformatics Institute.

http://www.ebi.ac.uk/newt

Hazel? Blue? Greenish-brown? Black?

By Vivienne Baillie Gerritsen

In the early days of the last century, scientists believed that the color of our eyes was a straightforward inherited trait. Mendel's laws of inheritance had become fashionable and eugenists saw in them an elegant and practical way to define our species. However, as the years passed and research in genetics progressed, bestowing the pigmentation of our eyes to the powers of a sole gene soon showed its weaknesses. Pigmentation proved to be a complex biological process. Nevertheless, as the 20th century bows out and the 21st comes upstage, it appears that – though pigmentation as a whole is part of an intricate biochemical network – the color of our eyes does indeed seem to be in the hands of a specific gene which codes for a protein known as the P protein.

Research on the genetics underlying human pigmentation – i.e. the color of our eyes, hair and skin – started in the very beginning of the 20th century by way of two biologists who also happened to be husband and wife: Charles (1866-1944) and Gertrude Davenport. Their idea was to demonstrate the existence of simple Mendelian principles underlying human pigmentation. But things turned out to be a little more complicated than expected. How, for instance, can one define the color of an eye? Would a hazel eye for me be a hazel eye for you? Furthermore, Sewell Wright (1889-1988), the great American evolutionary theorist, demonstrated that eye, hair and skin pigmentation could not be taken as separate entities. All types of pigmentation were the result of a same biochemical process itself inherited.

What exactly is pigmentation? What provides humans with ginger hair, green eyes or black skin? The pigment involved in our natural body colors is melanin. Melanin is synthesized from the amino acid tyrosine and is a light-absorbing biopolymer. It does not have a defined chemical structure, binds quite happily to other chemical entities and is pretty resistant to biochemical degradation. Melanin is produced and stored in cytoplasmic organelles known as melanosomes, which float around melanin-producing cells: the melanocytes. An interesting point is that, within human populations, the number of melanocytes per tissue does not really vary. What does vary however is the number of melanosomes, and hence melanin, per melanocyte. Typically, a dark-haired, dark-skinned, dark-eyed woman would have loads more melanosomes in her melanocytes than her fair-haired, fair-skinned, light-eyed counterpart.

Fig.1 The albino Lucasies family from the Netherlands became one of Phineas Barnum's most popular exhibits in the second half of the 19th century.

So why are our eyes brown? Or hazel? Or greenish-brown? Or black? Or bluish-green? Or chestnut? Or dark brown? Or blue? Or greyish-blue? It all has to do with melanin, naturally, but also light. For one, the lighter our eyes, the less melanin in our melanocytes. And when white light hits our irises, various wavelengths are either absorbed or reflected and give rise to the three common eye colors: brown, greenish-hazel and blue. Although such a classification is merely for practical purposes since eye pigmentation – like skin pigmentation – ranges from dark to light in a continuous manner.

Where does P protein come in? P protein is a medium-sized protein lodged in the melanosome membrane. It has twelve membrane-spanning regions and seems to be involved in as many hypothetical activities. It could transport small molecules such as tyrosine for example. It may act as a stabilizer of the melanosomal protein complex which includes a number of tyrosinases needed to synthesize melanin. P protein has also been proposed to function as a melanosome-specific ATP-driven proton pump thus regulating melanosomal pH. It may also play a part in the processing, sorting and regulation of the levels of tyrosinases, without which no melanin would be synthesized at all. So currently there is not much of a consensus as to its function but what has been discovered is that modifications in its activity, modify directly the color of an iris, so it must be a major orchestrator of the pathway which ultimately leads to the color of our eyes.

Albinism is a direct consequence of pigmentation abnormalities, of which many forms exist. The most current form, known as oculocutaneous albinism II or OCA2, is caused by a mutation in the P protein gene. It is an autosomal recessive disorder characterized by an absence, or reduced amount, of pigmentation in an individual's skin, hair or eyes and is unfortunately associated with relatively severe visual and auditory problems. The occurrence of OCA2 is 1/40000 in most populations worldwide. Historically, in the 19th century, albinos were thought to have all sorts of supernatural powers such as mind-reading or witchcraft and the infamous American businessman, circus man, impresario, politician, journalist and museum owner, Phineus Barnum, invited numerous albinos to be exhibited in his traveling show and museum.

Besides prenatal diagnosis of OCA2 and a better understanding of melanogenesis itself, what is the point of studying human pigmentation in the first place? For one, it helps to demolish prejudices such as 'racism' since biologists are quite incapable of establishing barriers between populations with regards to skin color. Genes involved in pigmentation, such as OCA2, also provide valuable information in the matter of population migrations and evolution, and one particularly interesting development is to be found in the field of forensics. Indeed, if a specific color of eye can be objectively attributed to one gene, then a DNA sample on an unidentified body could ultimately lead forensic investigators to an enhanced physical profile of a victim and perhaps even its identification. The only drawback here though is agreeing on the definition of a color…

Cross-references to Swiss-Prot

P protein, *Homo sapiens* (Human) : Q04671

References

1. Sturm R.A., Frudakis T.N.
   Eye colour: portals into pigmentation genes and ancestry
   Trends Genet. 20:327-332(2004)
   PMID: 15262401

2. Frudakis T.N., Thomas M., Gaskin Z., Venkateswarlu K., Chandra K.S., Ginjupalli S., Gunturi S., Natrajan S., Ponnuswamy V.K., Ponnuswamy K.N.
   Sequences associated with human iris pigmentation
   Genetics 165:2071-2083(2003)
   PMID: 14704187

3. Sturm R.A., Box N.F., Ramsay M.
   Human pigmentation genetics: the difference is only skin deep
   Bioessays 20:712-721(1998)
   PMID: 9819560

The taste experience

Vivienne Baillie Gerritsen

To what end do we need to taste? So that we can enjoy a night out at the restaurant? Fish also taste. Have you ever met one dining at the table next to yours? We taste because we have to be able to distinguish between what is good for us, and what is not. Nowadays, it's easy. All you have to do is march down the road to the nearest supermarket and pick out what you need. Most of us can read and relate the word 'tomato', 'steak' or 'chocolate' to a taste. Thousands of years ago, however, there were no supermarkets (or chocolate) and our ancestors could only rely on their taste buds. They soon learned that sweet-tasting foods were probably edible, whilst bitter ones were probably not. Bitterness usually spelled poison. And how do we distinguish between sweet and bitter? Via taste receptors lodged in the recesses of our taste buds.

Nature has granted us five fundamental sensations – sweet, bitter, sour, salty and umami – the combination of which we call 'the taste of something'. Different tastes demand different types of receptor and scientists are only beginning to untangle the complexities involved in their perception. To date, they have tracked down three receptors which seem to be directly involved in our perception of sweetness as well as the fashionable umami taste[1]: T1R1, T1R2 and T1R3, which combine to form heterodimeric G-protein coupled receptor complexes. This was no surprise really since proteins of this sort are also known to play key roles in the pathway we use to perceive smells for example.

Taste receptors are lodged in taste cell membranes which are found by the hundreds in our taste buds. Relays to our brain are then made via sensory nerve fibers. T1R1, T1R2 and T1R3 are about 800 amino acids long and wind in and out of each cell membrane seven times, where they form twosomes: T1R1 with T1R3, or T1R2 with T1R3. The former heterodimer helps us to perceive sweet tastes whereas the latter triggers off the perception of the umami taste. Without T1R3, the sensation of sweetness would be tasteless, so to speak. Interestingly, T1R3 has an amino terminal glycosylation site which, were it to be used, would add an appendage which in turn would interfere with T1R3's coupling to either T1R1 or T1R2. Consequently, this glycosylation site may well have some kind of regulatory function in the taste experience.

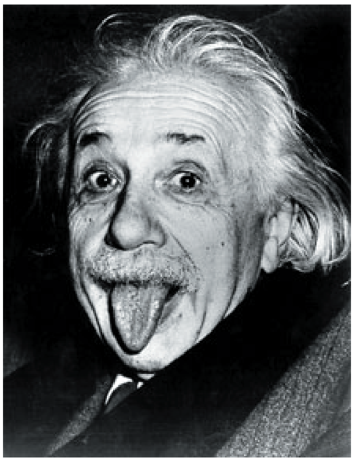


Fig. 1 Einstein's taste buds

How do flavors bind to their receptors? A number of proposals have been put forward. There are 'small flavors' and 'larger flavors', i.e. small chemical entities such as carbohydrates or larger macromolecules like proteins. How do the receptors cope with the difference in size? Small flavors probably slip into pockets formed by the heterodimers, thus triggering off the sensation of taste. Larger flavors are just too big to squeeze into such pockets and some believe that some macromolecules might display 'sweet fingers' which do fit in neatly. This, however, could not account for the great increase in activity sparked

[1] Protein Spotlight issue 17

off by a large sweet molecule which can be up to 100’000 times sweeter than the sensation procured by a smaller one! An altogether different mechanism is no doubt involved. What has been suggested is that larger molecules dock to the taste receptors via large interacting surfaces. This would not only stabilize the whole complex but would also account for the long-lasting and greater sensation of sweetness.

Sweetness comes in all shapes, sizes…and sweetness. How for instance can we distinguish a chocolate bar from a slice of chicken? Our tongue is an orchestra of receptors, there to distinguish between sweet, yes, but also sour, bitter, salty and umami. When you eat a slice of salami, just imagine the amount of molecules which are released onto your taste buds firing off concertos of sensations, the combination of which will make you perceive something definitely on the sweet side though the background noise is filled with sensations of saltiness, perhaps even sourness, a little bitterness and who knows some umaminess.

Over the millennia, the mechanisms underlying the perception of taste have been fine-tuned for the purposes of survival. Sweet tastes, for instance, are attractive because it usually means that we are consuming carbohydrates which are essential to the organism. Nature has also found a crafty way to filter our tastes. There are L-amino acids and their mirror images D-amino acids. L-amino acids are the ones used as building blocks. And guess what? These are the ones recognized by the umami taste receptors. It really does make one wonder whether we have any say at all in the matter.

Besides the academic desire to acquire knowledge on how and why we taste, the prospects of such research are particularly appetizing for those immersed in food flavoring. Taste receptors can be switched off – or on – by designing molecules which mime the receptors’ natural ligands. Funnily enough, Nature has devised her own little fib since the sweet protein miraculin can actually trick the mind into thinking that something bitter is in fact sweet[1]. Food flavoring is not the only issue though, there are serious illnesses related to sweetness, such as diabetes or hypolipemia (an abnormally low concentration of fats in the blood) which could of course also benefit from drug design.

Cross-references to Swiss-Prot

Sweet taste receptor T1R1, *Homo sapiens*, (human) : Q7RTX1
Sweet taste receptor T1R2, *Homo sapiens*, (human) : Q8TE23
Sweet taste receptor T1R3, *Homo sapiens*, (human) : Q7RTX0

References

1. Zhao G.Q., Zhang Y., Hoon M.A., Chandrashekar J., Erlenbach I., Ryba N.J.P., Zuker C.S.
The receptors for mammalian sweet and umami taste
Cell 115:255-266(2003)
PMID: 14636554

2. Temussi P.A.
Why are sweet proteins sweet? Interaction of brazzein, monellin and thaumatin with the T1R2-T1R3 receptor
FEBS Letters 526:1-4(2002)
PMID: 12208493

3. Nelson G., Chandrashekar J., Hoon M.A., Feng L., Zhao G., Ryba N.J.P., Zuker C.S.
An amino-acid taste receptor
Nature 416:199-202(2002)
PMID: 11894099

National Nodes

Argentina

Oscar Grau
IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata
Email: grau@biol.unlp.edu.ar
Tel: +54-221-4259223 Fax: +54-221-4259223
http://www.ar.embnet.org

Australia

Sonia Cattley
RMC Gunn Building B19, University of Sydney,NSW, 2006
Email: scattley@angis.org.au
Tel: +61-2-9531 2948
http://www.au.embnet.org

Austria

Martin Grabner
Vienna Bio Center, University of Vienna
Email: martin.grabner@univie.ac.at
Tel: +43-1-4277/14141
http://www.at.embnet.org

Belgium

Robert Herzog, Marc Colet
BEN ULB Campus Plaine CP 257
Email: rherzog@ulb.ac.be mcolet@ulb.ac.be
Tel: +32 2 6505146 Fax: +32 2 6505124
http://www.be.embnet.org

Brazil

Gonçalo Guimaraes Pereira
Laboratório de Genômica e Expressão - IB
UNICAMP-CP 6109
13083-970 Campinas-SP, BRASIL
Tel: 0055-19-37886237/6238
Fax: 0055-19-37886235
Email: goncalo@unicamp.br
http://www.br.embnet.org

Canada

Canadian Bioinformatics Resource, National Research Council Canada, Institute for Marine Biosciences,
Email: manager@cbr.nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
http://www.ca.embnet.org

Chile

Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
Tel: N/A
http://www.embnet.cl

China

Jingchu Luo
Centre of Bioinformatics
Peking University
Beijing 100871, China
Tel: 86-10-6275-7281
Fax: 86-10-6275-9001
Email: luojc@pku.edu.cn
http://www.cbi.pku.edu.cn

Colombia

Emiliano Barreto Hernández
Instituto de Biotecnología
Universidad Nacional de Colombia
Edificio Manuel Ancizar
Bogota - Colombia
Tel: +571 3165027 Fax: +571 3165415
Email : ebarreto@ibun.unal.edu.co
http://bioinf.ibun.unal.edu.co

Cuba

Ricardo Bringas
Centro de Ingeniería Genética y Biotecnolgía,
La Habana, Cuba
Email: bringas@cigb.edu.cu
Tel: +53 7 218200
http://www.cu.embnet.org

Finland

Eija Korpelainen
CSC, Espoo
Email: eija.korpelainen@csc.fi
Tel: +358 9 457 2030
http://www.fi.embnet.org

France

Jean-Marc Plaza
INFOBIOGEN, Evry
Email: plaza@infobiogen.fr
Tel: +33 1 60 87 37 11 Fax: +33 1 60 87 37 96
http://www.fr.embnet.org

Germany

Sandor Suhai
EMBnet node at the German Cancer Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
http://www.de.embnet.org

Hungary

Endre Barta
Agricultural Biotechnology Center
Szent-Gyorgyi A. ut 4. Godollo,
Email: barta@abc.hu
Tel: +36 30-2101795
http://www.hu.embnet.org

India

H.A.Nagarajaram
Laboratory of Computational Biology & Bioinformatics facility, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
http://www.in.embnet.org

Israel

Leon Esterman
INN (Israeli National Node) Weizmann Institute of Science
Department of Biological Services, Biological Computing Unit, Rehovot
Email: Leon.Esterman@weizmann.ac.il
Tel: +972- 8-934 3456
http://www.il.embnet.org

Italy

Cecilia Saccone
CNR - Institute of Biomedical Technologies
Bioinformatics and Genomic Group
Via Amendola 168/5 - 70126 Bari (Italy)
Email: saccone@area.ba.cnr.it
Tel. +39-80-5482100 - Fax. +39-80-5482607
http://www.it.embnet.org

Mexico

Cesar Bonavides
Nodo Nacional EMBnet, Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos
Email: embnetmx@cifn.unam.mx
Tel: +52 (7) 3 132063
http://embnet.cifn.unam.mx

The Netherlands

Jack A.M. Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen, NL
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074
http://www.nl.embnet.org

Norway

George Magklaras
The Norwegian EMBnet Node
The Biotechnology Centre of Oslo
Email: admin@embnet.uio.no
Tel: +47 22 84 0535
http://www.no.embnet.org

Poland

Piotr Zielenkiwicz
Institute of Biochemistry and Biophysics
Polish Academy of Sciences Warszawa
Email: piotr@pl.embnet.org
Tel: +48-22 86584703
http://www.pl.embnet.org

Portugal

Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +351 2144079070
http://www.pt.embnet.org

Russia

Sergei Spirin
Biocomputing Group, Belozersky Institute Moscow
Email: sas@belozersky.msu.ru
Tel: +7-095-9395414
http://www.genebee.msu.ru

Slovakia

Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
http://www.sk.embnet.org

South Africa

Ruediger Braeuning
SANBI, University of the Western Cape, Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
http://www.za.embnet.org

Spain

José M. Carazo, José R. Valverde
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid
Email: carazo@es.embnet.org,
jrvalverde@es.embnet.org
Tel: +34 915 854 505 Fax: +34 915 854 506
http://www.es.embnet.org

Sweden

Nils-Einar Eriksson, Erik Bongcam-Rudloff
Uppsala Biomedical Centre, Computing Department, Uppsala, Sweden
Email: nils-einar.eriksson@bmc.uu.se
erik.bongcam@bmc.uu.se
Tel: +46-(0)18-4714017, +46-(0)18-4714525
http://www.embnet.se

Switzerland

Laurent Falquet
Swiss Institute of Bioinformatics, CH-1066 Epalinges
Email: Laurent.Falquet@isb-sib.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
http://www.ch.embnet.org

United Kingdom

Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton, Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
http://www.uk.embnet.org

Specialist Nodes

EBI

Rodrigo López
EBI Embl Outstation, Wellcome trust Genome Campus,
Hinxton Hall, Hinxton, Cambridge, United Kingdom
Email: rls@ebi.ac.uk
Phone: +44 (0)1223 494423
http://www.ebi.ac.uk

ETI

P.O. Box 94766
NL-1090 GT Amsterdam, The Netherlands
Email: wouter@eti.uva.nl
Phone: +31-20-5257239
Fax: +31-20-5257238
http://www.eti.uva.nl

ICGEB

Sándor Pongor
International Centre for Genetic Engineering and
Biotechnology
AREA Science Park, Trieste, ITALY
Email: pongor@icgeb.trieste.it
Phone: +39 040 3757300
http://www.icgeb.trieste.it

LION Bioscience

Thure Etzold
LION Bioscience AG, Heidelberg, Germany
Email: Thure.Etzold@uk.lionbioscience.com
Phone: +44 1223 224700
http://www.lionbioscience.com

MIPS

H. Werner Mewes
Email: mewes@mips.embnet.org
Phone: +49-89-8578 2656
Fax: +49-89-8578 2655
http://www.mips.biochem.mpg.de

UMBER

Terri Attwood
School of Biological Sciences, The University of
Manchester, Oxford Road, Manchester M13 9PT, UK
Email: attwood@bioinf.man.ac.uk
Phone: +44 (0)61 275 5766
Fax: +44 (0) 61 275 5082
http://www.bioinf.man.ac.uk/dbbrowser

TECH-MGR

Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes and helps them with maintenance and troubleshooting.
The team consists of experienced system administrators and programmers who ensure the availability of local services for all EMBnet users.

EMBnet.news

ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can print then please let us know. Please send your contributions to one of the editors. You may also submit material by e-mail.

Past issues of embnet.news are available as PostScript or PDF files (ISSN 1023-4144). You can get them by anonymous ftp from:
the EMBnet organisation Web site
`http://www.embnet.org/download/embnetnews`
the Belgian EMBnet node
`ftp://ftp.be.embnet.org/pub/embnet.news`
the UK EMBnet node
`ftp://ftp.uk.embnet.org/pub/embnet.news`
the EBI EMBnet node
`ftp://ftp.ebi.ac.uk/pub/embnet.news`

Submission deadlines for the next issues:

May 31, 2005
August 15, 2005
October 31, 2005