EMBnet.news

Volume 13 Nr. 1
March 2007

- Bioclipse
- A Grid Approach
- The MitoRes database
- Electronic publishing in e-Learning platform

and more ...

Editorial

The present issue shows several aspects of our activity. An article on a course delivered in Kenya where a new node has been formed last year, several contributions on topics related to the creation of new tools and the use of technological innovations, etc. We will now place some efforts at obtaining texts that also depict our outreach in terms of geographical distribution and outreach. Enjoy reading and, by all means, let us know if you wish to contribute with your experience in Bioinformatics to EMBnet News.

The editorial board: Erik Bongcam-Rudloff, Domenica D'Elia, Pedro Fernandes, Kimmo Mattila and Lubos Klucar.

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 issue 80. Please let us know if you like this inclusion.

Cover picture: *Lyriocephalus scutatus* (Linnaeus, 1758).
Sri Lanka, February 2007 [© Erik Bongcam-Rudloff]

Contents

Editorial Board:

Course Report:

Nairobi, Kenya, August 2006

The course organizers, Erik, Alvaro and Etienne, in front of Kilimanjaro

Erik Bongcam-Rudloff and Etienne de Villiers

Dept. of Animal Breeding and Genetics, SLU and The Linnaeus Centre for Bioinformatics, SLU/UU. Sweden

International Livestock Research Institute, Nairobi, Kenya

The ILRI\BECA EMBNet node located in Nairobi Kenya held a three-day introductory course in Bioinformatics for scientists in East and central Africa in collaboration with the Swedish University of Agricultural Sciences and the Linnaeus Centre for Bioinformatics, (SLU-LCB) and the Swedish EMBnet node. The course was funded by the Swedish Formas - Sida/SAREC project: "Sustainable development bioinformatics project between SLU and ILRI, Nairobi, Kenya ". The objective of the course was to introduce young scientists to and encourage application of bioinformatics/computational biology in their research and to present some of the biological resources available on the ILRI\ BECA bioinformatics platform.

High demand

In the beginning the number of places on the course was limited to 25 but the number of registrations was overwhelming and 35 students were selected among 65 applicants. The selected students were the scientists showing the strongest evidence and potential to apply the knowledge and expertise gained from the course to research and also to train others in an effort to increase capacity in bioinformatics at their home institution. The course consisted of lectures and theoretical sessions and hands-on practical training sessions in a computer laboratory. The participants received training materials and documentation on main topics covered by the course, and will have ongoing access to the support services provided by ILRI\BECA. All participants received a copy of the "Live wEMBOSS CD" created by Martin Sarachu and Diego Bellante at the Argentinean EMBnet node.

Course schedule

Day 1

- Lecture: Pair wise sequence comparisons and Sequence Alignments (Blast, FASTA)
- Lecture: Patterns and Profiles (HMMs and PSI-Blast)
- Lecture: ILRI/BECA Bioinformatics Platform services
- Tutorials: Homology search (BLAST, HMM)

Lecturers: Etienne de Villiers, Erik Bongcam-Rudloff

Day 2

- Lecture: Biological databases I
- Lecture: EMBOSS
- Tutorial: wEMBOSS

Lecturer: Erik Bongcam-Rudloff

Day 3

- Lecture: Biological databases II
- Lecture: ENSEMBL and web based resources
- Tutorial: ENSEMBL

Lecturers: Erik Bongcam-Rudloff, Alvaro-Martinez Barrio

- Examination
- Course evaluation

The course participants

The computer room

Limitations

One limiting factor during the course was the very low Internet bandwidth at ILRI. This problem was solved by using the bioinformatics resources available on the ILRI\BECA bioinformatics platform and a new local installation of the ENSEMBL system.

The problem with low bandwidth is shared by many research institutes and Universities worldwide. For that reason the organizers of this course are working on a project creating a portable bioinformatics teaching kit (PBT-kit) installed on a Mac-mini. But more about that project in the next issue of EMBnet.News.

Course evaluation

The resulting course evaluation gave the course 4.7 points out of 5. The comments given by the participants could be summarised as follows:

"Most participants were very happy with the course, but think that three days is not nearly enough given that the course is so comprehensive and detailed. Too much information in too short time. Suggested time for the course reaches from one week up to two months. Most participants want to learn more in the field of bioinformatics and request a second, more advanced course following this first. They also wish to be updated on present research. Many participants would appreciate if bioinformatics courses were given regularly."

This short training course was very successful and the positive criticism received encouraged us to organize a longer course (8 days) in March, 2007.

Etienne de Villiers and Erik Bongcam-Rudloff

Acknowledgment

The organizers of the course want to thank Laurent Falquet for educational materials and Martin Sarachu and Diego Bellante for the wEMBOSS LiveCD. The course was funded by a grant from Formas-Sida/SAREC nr. 2005-303.

Comment from one participant:

"Make this training available also for more people, especially in the education sector and the medical and research fraternity in the country so that everyone can be able to know more about it and be able to compete with other scientists in the world because of the information acquired. Refresher courses and in-servicing should be encouraged. Thank you, I am more informed and I can be able to relay this information to others who did not have a chance to attend this training. I enjoyed it myself so much."

Links:

```
www.becabioinfo.org
www.slu.se
www.hgen.slu.se
www.embnet.se
www.lcb.uu.se
```

The course material

Using Bioclipse to integrate bioinformatics functionality

Ola Spjuth

Department of
Pharmaceutical
Biosciences, Uppsala
University, Box 591, 751 24
Uppsala, Sweden

ola.spjuth@farmbio.uu.se

Introduction

Bioclipse [1] is an open source project that aims to integrate chemoinformatic and bioinformatic functionality into a single user-friendly workbench. The project started in October 2005 and during the first 12 months over 12 developers contributed to the application. Bioclipse version 1.0.1 was released February 1st, 2007, and contains features to edit, visualize, and analyze molecules, proteins, spectra, scripts, and sequences. This article gives an introduction to Bioclipse and covers the basics how to integrate new functionality into the framework.

Architecture

Bioclipse is built on Eclipse [2], a universal tool platform for constructing software applications. Eclipse is mainly known as an IDE (Integrated Development Environment) but the framework is general and the advanced plugin architecture makes it possible to extend it in any direction. The building blocks in Eclipse are called plugins, and the minimal set of plugins required to build an application is referred to as the Rich Client Platform (RCP). Most other integration frameworks simply provide features for connecting new algorithms and separately installed applications. In Bioclipse it is straightforward to extend almost every part in the workbench such as views, editors, wizards, menus, algorithms, visualizations, or even the object model. For more information about Bioclipse and the Eclipse plugin architecture, see [1].

Features

The main features provided by Bioclipse are:

Molecular management

Molecular management is provided by the Chemistry Development Kit (CDK) [3]. It provides features such as I/O in various file formats (mol, xyz, cml, pdb, sdf), calculation of chemical properties, structure diagram generation, SMILES parsing and generation, atom typing, and substructure searches.

Sequence management

The software library BioJava [4] is used to provide sequence I/O and analysis. It supports many sequence file formats and parsing of output files, e.g. from sequence alignment software.

2D editing of chemical structures

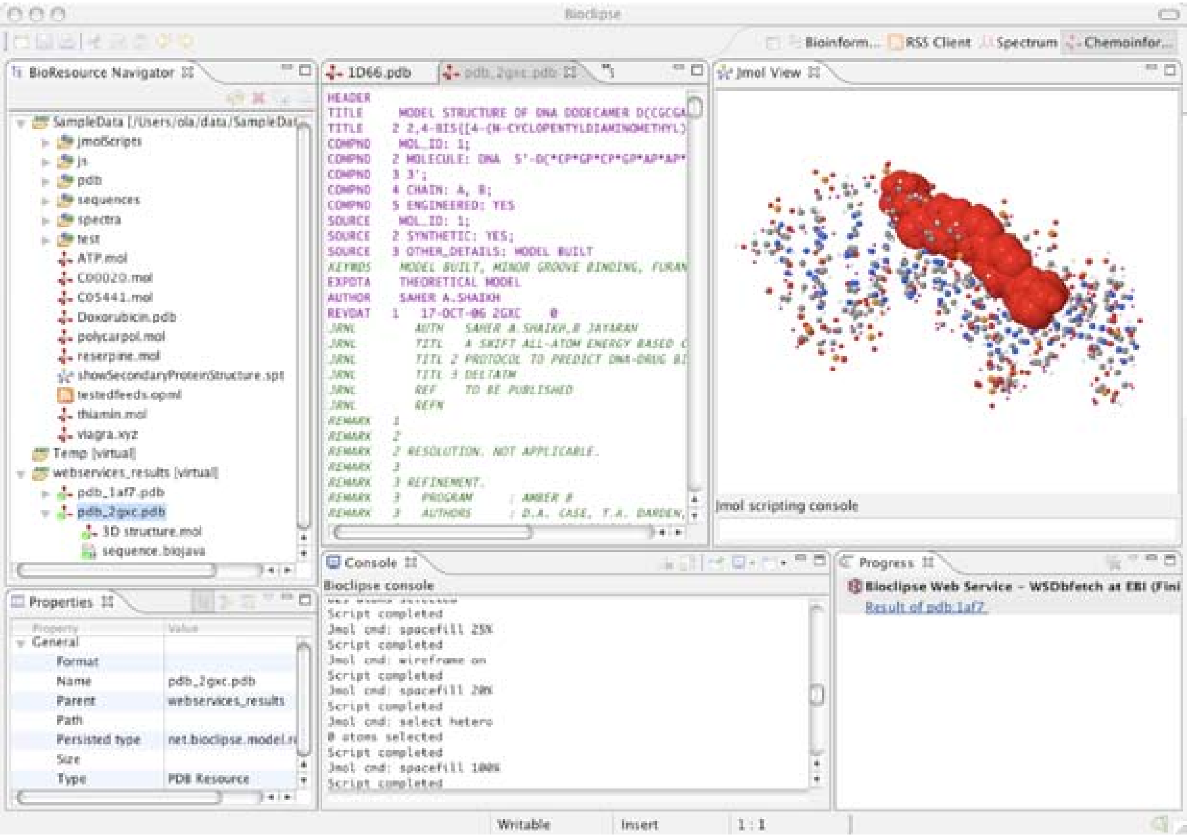
The application JchemPaint [5] is integrated into Bioclipse as a multi page editor. It has graphical 2D editing of chemical structures on one tab and the source in text format on a second; both completely synchronized with each other.

3D visualization of chemical structures

For 3D visualization, Jmol [6] is integrated into Bioclipse. Jmol is a fully-featured java-based 3D visualization toolkit, similar to RasMol and Chime, with advanced graphics and scripting capabilities.

Spectrum analysis

Bioclipse supports editing and visualization of several types of spectra, such as NMR, MS, and IR. A plugin for assignment of peaks to molecular structures, as well as a plugin for computer-aided structure elucidation, is in development.



Properties

Property	Value
General	
Format	
Name	pdb_2gxc.pdb
Parent	webservices_results
Path	
Persisted type	net.bioclipse.model.r
Size	
Type	PDB Resource

```
HEADER
TITLE     MODEL STRUCTURE OF DNA DODECAMER D(CGCGA
TITLE    2 2,4-BIS[[4-(N-CYCLOPENTYLDIAMINOMETHYL)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: DNA  5'-D(C*CP*GP*CP*GP*AP*AP*
COMPND   3 3';
COMPND   4 CHAIN: A, B;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 SYNTHETIC: YES;
SOURCE   3 OTHER_DETAILS: MODEL BUILT
KEYWDS    MODEL BUILT, MINOR GROOVE BINDING, FURAN
EXPDTA    THEORETICAL MODEL
AUTHOR    SAHER A.SHAIKH
REVDAT   1   17-OCT-06 2GXC    0
JRNL        AUTH   SAHER A.SHAIKH,B JAYARAM
JRNL        TITL   A SWIFT ALL-ATOM ENERGY BASED C
JRNL        TITL 2 PROTOCOL TO PREDICT DNA-DRUG BI
JRNL        TITL 3 DELTATW
JRNL        REF    TO BE PUBLISHED
JRNL        REFN
REMARK   1
REMARK   2
REMARK   2 RESOLUTION. NOT APPLICABLE.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : AMBER 8
REMARK   3   AUTHORS     : D.A. CASE, T.A. DARDEN,
```

```
Script completed
Jmol cmd: spacefill 25%
Script completed
Jmol cmd: wireframe on
Script completed
Jmol cmd: spacefill 20%
Script completed
Jmol cmd: select hetero
0 atoms selected
Script completed
Jmol cmd: spacefill 100%
Script completed
```

Figure 1: Screenshot of Bioclipse showing the 3D structure of the PDB file 2GXC retrieved via the WSDbfetch Web service, with the bound ligand highlighted.

Web service support

Bioclipse has a well developed infrastructure for Web services. The WSDbfetch Web service at EBI [7] is one example that has been integrated with a dedicated wizard.

Scripting

The Mozilla Rhino [8] engine has been integrated into Bioclipse and provides scripting functionality using the JavaScript language, which greatly simplifies pipelining of tasks.

Integrating new functionality

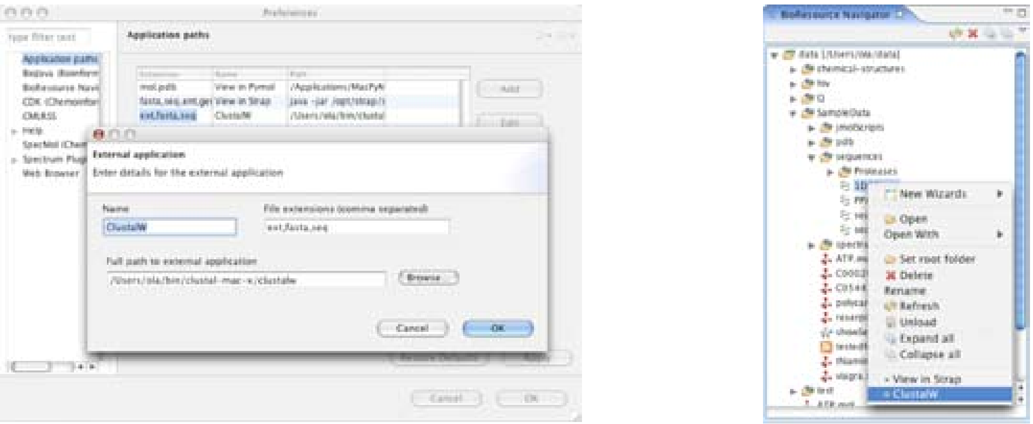
Integrating new functionality into Bioclipse can be done in several ways. This section will demonstrate some of these ways by showing how to integrate the third party application ClustalW [9], which is used for multiple alignments of protein sequences.

Command line invocation

Separately installed applications can easily be integrated into Bioclipse at runtime by specifying their command line syntax in the preferences (Figure 2). To create an action that runs ClustalW, simply open the preference page from the main menu, enter the file extensions for which the application should be available (e.g. fasta, seq), and specify the complete path to the ClustalW application. If you now, in the BioResource Navigator, right-click a file with multiple sequences in Fasta format there is a new option ClustalW (Figure 2) that invokes the local ClustalW application.

Writing a new plugin

The most flexible way to integrate new functionality into Bioclipse is by wrapping an existing library, application, or Web service in a new plugin. By creating custom actions and menu options it is straightforward to make new features available for users with advanced GUI components that are able to interact with other plugins. This will be demonstrated here by outlining the main steps



Extension	Name	Path
mol,pdb	View in Pymol	/Applications/MacPyM
fasta,seq,aml,gei	View in Strap	java -jar /opt/strap/
ext,fasta,seq	ClustalW	/Users/ola/bin/clustal

Figure 2: Bioclipse preferences for integrating third party applications using command line invocation (left) and the generated context menu option in the BioResource Navigator (right).Figure 2: Bioclipse preferences for integrating third party applications using command line invocation (left) and the generated context menu option in the BioResource Navigator (right).

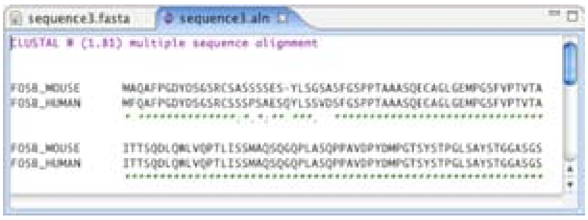
for how to create an editor with basic highlighting for displaying and editing ClustalW result files in Bioclipse. The complete source code is available from the Bioclipse website [10] in the downloads section, as well as from the Bioclipse Subversion server [11]. Start by setting up a new plugin project in Eclipse named bc _ clustalw and create the following classes:

```
class ClustalWEditor extends net.
bioclipse.editors.keyword.KeywordEditor

class ClustalWKeywords extends net.
bioclipse.editors.keyword.Keywords

class ClustalWRuleScanner extends
org.eclipse.jface.text.rules.
RuleBasedScanner
```

ClustalWKeywords defines the keywords for the editor that should be highlighted. The ClustalWRuleScanner defines how the editor should display these keywords. The ClustalWEditor is the actual editor and is linked to the Bioclipse workbench to operate on file extensions .aln and .dnd via the plugin manifest (the file plugin.xml) in the following way:

```
<extension point="org.eclipse.
ui.editors">
  <editor
    id="net.bioclipse.editors.clustalw.
ClustalWEditor"
    name="ClustalW Editor"
    icon="icons/sample.gif"
    extensions="aln,dnd"
    contributorClass="org.
eclipse.ui.texteditor.
BasicTextEditorActionContributor"
    class="net.bioclipse.editors.
clustalw.ClustalWEditor">
  </editor>
</extension>
```

The resulting editor is shown in Figure 3.



```
CLUSTAL W (1.81) multiple sequence alignment


FOSB_MOUSE      MAQAFPGDYDSGSRCSASSSSES-YLSGSASFGSPPTAAASQECAGLGEMPGSFVPTVTA
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
                * *************.*.*.** ***.   ******************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
                ************************************************************
```

Figure 3: The implemented ClustalWEditor with basic highlighting

Adding a Web service

It is easy to add new clients for Web services to Bioclipse. The user-friendliest method is to create a Dialog or a Wizard that is integrated with the job scheduling in Bioclipse. That way, when the Web service is executed, it is run in a separate thread, leaving the workbench available for other tasks. When the job is completed it signals to the workbench and the user can view the results of the operation. This section will describe how to extend bc _ clustalw with a Dialog to invoke the ClustalW Web service available from EBI, called WSClustalW [7].

Download the WSDL file from [7]. Generate a java client (e.g. using axis) and copy the files into a new package, `uk.ac.ebi.www.WSClustalW` in the bc _ clustalw plugin. In Eclipse this is a trivial step using the Web Tools Platform (WTP).

Creating the Action

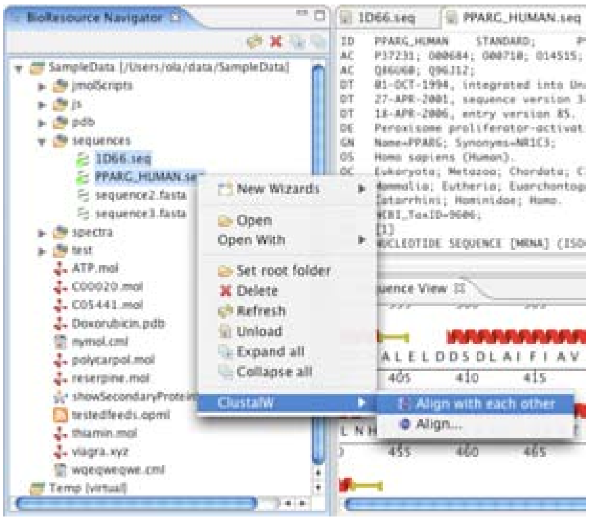
We want to be able to access the Web service by right-clicking on two or more sequences in the BioResource Navigator in Bioclipse. For this we create the following Action:

```
class AlignWithClustalWActionDelegate
implements IViewActionDelegate
```

Next we implement the method Run(…) to open the ClustalWDialog (see below) and then call WSClustalW. We hook the action into the Bioclipse framework via the plugin manifest:

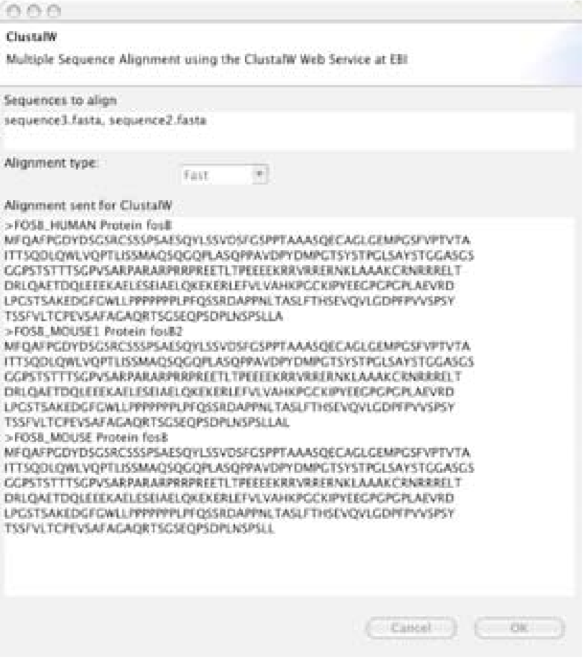
```
<extension point="org.eclipse.
ui.popupMenus">
  <viewerContribution
    id="net.bioclipse.contribution.
popup.clustalw.bioresourceview"
    targetID="net.bioclipse.views.
BioResourceView">
  <visibility>
    <objectClass name="net.bioclipse.
model.BiojavaResource"/>
  </visibility>
  <menu
    id="net.bioclipse.views.
BioResourceView.ClustalWMenu"
    label="ClustalW"
    path="additions">
    <separator name="group1"/>
  </menu>
  <action
    label="Align with each other"
    icon="icons/sample.gif"
    menubarPath="net.bioclipse.views.
BioResourceView.ClustalWMenu/group1"
    class="net.bioclipse.
plugins.bc _ clustalw.actions.
AlignWithCLustalWActionDelegate"
    id="net.bioclipse.
plugins.bc _ clustalw.actions.
AlignWithCLustalWActionDelegate"
    enablesFor="2+">
  </action>
  </viewerContribution>
</extension>
```

The *viewerContribution* node defines that we want this action to appear as a popupMenu in the BioResource Navigator View. The *visibility* node makes this Action visible only if a user has right-clicked on a BiojavaResource, which is used to manage sequences in Bioclipse. The property *enablesFor*="2+" enables this action only if two or more objects are selected in the BioResource Navigator (see Figure 4).



```
ID   PPARG_HUMAN     STANDARD;      P
AC   P37231; O00684; O00718; O14515;
AC   Q86U60; Q96J12;
DT   01-OCT-1994, integrated into Un
DT   27-APR-2001, sequence version 3
DT   18-APR-2006, entry version 85.
DE   Peroxisome proliferator-activat
GN   Name=PPARG; Synonyms=NR1C3;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; C
     Mammalia; Eutheria; Euarchontog
     Catarrhini; Hominidae; Homo.
     NCBI_TaxID=9606;
     [1]
     NUCLEOTIDE SEQUENCE [MRNA] (ISO
```

Figure 4: The created action is only available and enabled in the context menu when two or more sequence files are selected in the BioResource Navigator.



ClustalW

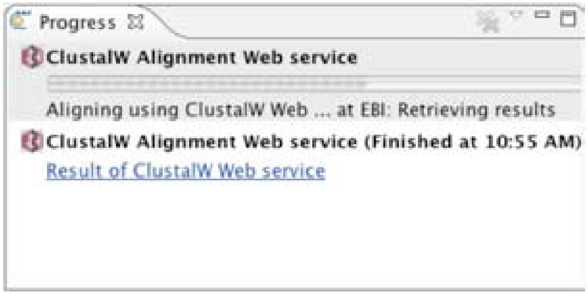
Multiple Sequence Alignment using the ClustalW Web Service at EBI

Sequences to align: sequence3.fasta, sequence2.fasta

Alignment type: Fast

Alignment sent for ClustalW

```
>FOSB_HUMAN Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLA
>FOSB_MOUSE1 Protein fosB2
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLL
```



ClustalW Alignment Web service

Aligning using ClustalW Web ... at EBI: Retrieving results

ClustalW Alignment Web service (Finished at 10:55 AM)

Result of ClustalW Web service

Figure 5: The ClustalWDialog provides alignment settings and concatenates the selected resources into an input file for ClustalW (left). The Bioclipse Progress View displays jobs running in the background and signals when results are available (right).

Creation the Dialog

To visualize the ClustalW input and set alignment settings we create the following class:

```
class ClustalWDialog extends
TitleAreaDialog
```

TitleAreaDialog is a standard dialog with a title area, and is suitable as a base class for building our dialog. We override the *createDialogArea*(…) and create our graphic components (Figure 5, left). As this dialog is only opened from our action, we don't need to define it in the manifest. When the user clicks OK, a job is set up and added to the job scheduling of Bioclipse. If desired, the job can be run in the background and is then visible in the Progress View (Figure 5, right). Upon completion the Progress View displays that results are available, and when a user clicks on the job the results are displayed.

Writing a script

It is straightforward to create new scripts in Bioclipse using the integrated Javascript editor. You can take advantage of functionality from all installed plugins in Bioclipse and create scripts to execute complex tasks, such as pipelining steps in a workflow. Figure 6 displays a script that re-

```
//Use WSDbfetch Web Service to get
protein sequences from Uniprot
wsdbfetch = new
DbfetchServiceServiceLocator();
sequence1 = wsdbfetch.getUrnDbfetch().
fetchData(“uniprot:Q3UMM4”, “fasta”,
“raw”);
sequence2 = wsdbfetch.getUrnDbfetch().
fetchData(“uniprot:Q4KM47”, “fasta”,
“raw”);


...

//Concatenate strings for ClustalW
alignmentString=sequence1 + “\n” +
sequence2;

//Create ClustalW data object to operate
on, choose full alignment type
clustalwData=new ClustalWData(alignment
String,”full”);

//Create action and run Web service
action = new AlignWithCLustalWActionDel
egate(clustalwData);
action.doAlignment();
```

Figure 6: Excerpt from a Bioclipse script to retrieve two sequences from Uniprot and perform a sequence alignment with ClustalW using Web services at EBI. The complete script is available from the Bioclipse Wiki [12].

trieves two sequences from Uniprot and performs a sequence alignment with ClustalW by invoking the WSDbfetch and WSClustalW Web services at EBI [7].

License

Bioclipse is licensed under Eclipse Public License (EPL) [13], an open source license that allows plugins to be licensed individually. Hence, any third party application, framework, or library can be integrated - even commercial ones. Open source promotes fast bug fixing and close collaboration between users and developers and more and more organizations discover these advantages. We are convinced that an open integration framework is the best way of creating and delivering software in life science.

Conclusions

Bioclipse has been met with great approval by the research community. The idea of having a user-friendly workbench with a powerful plugin architecture that is released as open source is an appealing thought for both users and developers. For end users it means that they can easily get access to the latest functionality, and for developers it makes their implementations readily available for users. The innovative use of Eclipse in life science resulted in Bioclipse being awarded the third price and the audience award at JAX Innovation Award 2006 [14].

Bioclipse has great potential for the future. Many features and new plugins are in development, e.g. for molecular dynamics, database persistence, systems biology, QSAR, and chemical reaction modeling. We encourage researchers and developers to contribute to Bioclipse by adding new features, integrating existing software, develop editors for new file formats; only your imagination sets the boundaries.

Bioclipse is already used by scientists and teachers all over the world in areas such chemistry, biology, biotechnology, and pharmacology. Some organizations have even declared Bioclipse as their standard platform, and have stated that they will release all their new applications based on the framework.

Bioclipse is already a feature-rich workbench and integration platform, and with the right support it could evolve into a standard platform for delivering bioinformatics functionality to scientists and teachers all over the world.

Remarks

Bioclipse is an international collaboration with the main contributors from the Department of Pharmaceutical Biosciences at Uppsala University, Cologne University Bioinformatics Centre (CUBIC), and the Department of Chemistry at University of Cambridge. For the latest information regarding the project and more information about how to get involved, see the Bioclipse website [10].

Acknowledgements

The author would like to thank the Bioclipse developers, and all others supporting the project.

References

[1] Ola Spjuth, Tobias Helmus, Egon L Willighagen, Stefan Kuhn, Martin Eklund, Johannes Wagener, Peter Murray-Rust, Christoph Steinbeck, Jarl E.S. Wikberg: Bioclipse: An open source workbench for chemo- and bioinformatics. BMC Bioinformatics 2007, 8:59 (22 February 2007)
[2] Eclipse universal tool platform [http://www.eclipse.org].
[3] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 2003, 43(2):493–500.
[4] BioJava [http://bio java.org/].
[5] Krause S, Willighagen E, Steinbeck C: JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. Molecules 2000, 5:93–98
[6] Jmol [http://jmol.sourceforge.net/]
[7] Labarga, A. et al: Web services at EBI. EMBnet.news, 11(4) 18-23 (2005)
[8] Mozilla Rhino [http://www.mozilla.org/rhino/].
[9] Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and

weight matrix choice. Nucleic Acids Res. 1994 Nov 11;22(22):4673-80

[10] Bioclipse homepage [http://www.bioclipse.net]

[11] Bioclipse Subversion server [https://bioclipse.svn.sourceforge.net/svnroot/bioclipse]

[12] Bioclipse wiki [http://wiki.bioclipse.net]

[13] Eclipse Public License [http://www.eclipse.org/org/documents/epl-v10.php].

[14] JAX innovation award 2006 [http://jax-award.de/jax_award06/gewinner_en.php]

Bioclipse homepage: www.bioclipse.net

Announcement

Embrace Workshop on Bioclipse 2007 (EWB '07)

May 23 -25 2007
Uppsala Biomedical Centre (BMC), Uppsala, Sweden

Contents:
Bioclipse is an open source workbench for chemo- and bioinformatics with rich functionality for molecules, sequences, proteins, spectra, and scripts. Bioclipse has advanced plugin architecture which facilitates integration of new functionality, such as algorithms, editors, visualization, Web services, and third party applications. The Embrace Workshop on Bioclipse 2007 (EWB '07) will consist of lectures and hands-on labs to demonstrate the features of Bioclipse, the power of the plugin architecture, and how to integrate new features into the framework. Participation in the workshop is free of charge.

For registration and more information, please see the workshop homepage: http://teacher.bmc.uu.se/BioclipseWS07 and http://www.bioclipse.net.

Proof of concept:

Embarrassingly Parallel Bioinformatics Code on the Grid

Heinz Stockinger, Marco Pagni, Lorenzo Cerutti, Laurent Falquet

Swiss Institute of
Bioinformatics, Quartier
Sorge - Bâtiment Génopode,
CH-1015 Lausanne,
Switzerland

Introduction

Several popular bioinformatics packages for sequence analysis (i.e., the HMMER [10], SAM [22] or the PFTOOLS package [20]) use the so-called profile-HMM approach to search and align sequences. They are very CPU intensive (in fact, HMMER and PFTOOLS run at more than 99 percent of CPU on a single core processor) and typically have the characteristic of being embarrassingly parallel. In other words, the input dataset (i.e. the collections of sequences and profile-HMMs) can be partitioned, and several calculations can run in parallel. This feature is utilized by several parallel implementations of the algorithms such as in [4, 17, 24]. However, existing implementations are mainly tailored to clusters or supercomputers on a local-area network with a homogeneous hardware infrastructure.

Grid computing, and in particular Data Grids, provide a powerful infrastructure for running CPU intensive, high performance applications [14]. The bioinformatics community has also started to use Grid technology to "gridify" and speed up their most important applications [15, 11]. A successful example can be found in [3]. Furthermore, the EMBRACE Grid [9] as well as the Swiss Bio Grid [23] project aim at improving the collaboration of bioinformaticians as well as the daily work of biologists. We propose a Grid approach, which can be used by biologists who do not have access to conventional clusters.

As a proof of concept we have chosen the profile-HMM to be executed on the Grid for improving the overall execution time of the application. We have created a service that automatically partitions the input data set, creates a parallel workload and then submits it to a Grid infrastructure for execution. Note that our approach leverages the existence of Grid resource brokering and scheduling algorithms but optimizes it by adding fault-tolerance and higher level, application specific features to existing Grid infrastructures.

We have decided to leave the bioinformatics package unchanged and provide a wrapper tool that takes care of all the necessary preprocessing steps (splitting the input data into chunks) and the submission to a Data Grid infrastructure. This allows us to use *a generic approach*, which can be applied to a broad class of similar problems and not only to the one considered here. Moreover, we do not change the implementation of an existing algorithm, which shields us from version change etc.

In more detail, we have designed and developed a Grid tool called ***wg*** (Workload Generator) that can execute several different bioinformatics tools such as *hmmsearch* and *pfsearch* in a Grid environment. The tool is built in a modular way to allow for easy integration of different Grid platforms where a Grid job should be executed. In our current implementation we support EGEE [8], ARC [7] as well as cluster schedulers such as LSF [16]. The tool provides two different algorithms, **a-priori algorithm** and **run time sensitive algorithm** (details in next section), to execute an embarrassingly parallel program on many nodes of a Grid or a cluster. In this paper we show that our system (using a run time sensitive algorithm) outperforms small or medium-size clusters if their scalability limits are reached.

Algorithmic Approach: featuring "Large-Scale" on the Grid

Before we run an application on the Grid, one first needs to find out if it is reasonable to do it. We emphasize "large scale", which means that the number of sequences and profiles needs to be reasonably large in order to be regarded as a "Grid problem". There is no general statement about the required size to qualify as a "Grid problem", but empirical studies are required to determine this. However, let us first discuss the *characteristics of a Grid environment* in more detail before we elaborate more on this topic.

- A Grid typically has an **intrinsic latency**: A Grid is typically a world-wide distributed computing infrastructure with lots of interacting services such as the system to discover services, resource brokers to find suitable computing elements etc. In addition, cluster batch systems are used at remote sites. Consequently, for each Grid job we need to consider a certain latency in terms of response time that is spent mainly on resource brokering, scheduling, job monitoring and the retrieval of the output.

- **Heterogeneity** and **"unpredictable" performance**: Since each site in a Grid is an independent administrative domain, each site is free to deploy different batch systems, storage systems, computer hardware etc. with different latency features. It is rather difficult to predict the exact execution time of a standard job. Having this latency in mind we can deduce empirically which jobs would run faster on a single CPU. In such a case, it is suggested to run the application on a local machine. The system that we designed takes care of this problem by looking at the number of profiles/ sequences to be processed by the specific application.

Mapping the Problem to a Distributed Environment

We use the following terminology: a **job** is the smallest piece of work submitted to a Grid. Since *wg* is a wrapper around existing tools and needs to execute on remote worker nodes in the Grid, in our model one job corresponds to one remote execution of *wg*. Since each job requires data chunks to be worked on, each job can run 1 or many **tasks** where a task includes 1 or many sequences as well as 1 or many profile-HMMs. Given that, a task itself can run 1 or several sequential instances of *hmmsearch*, assuming that one instance can only run with 1 profile-HMM but

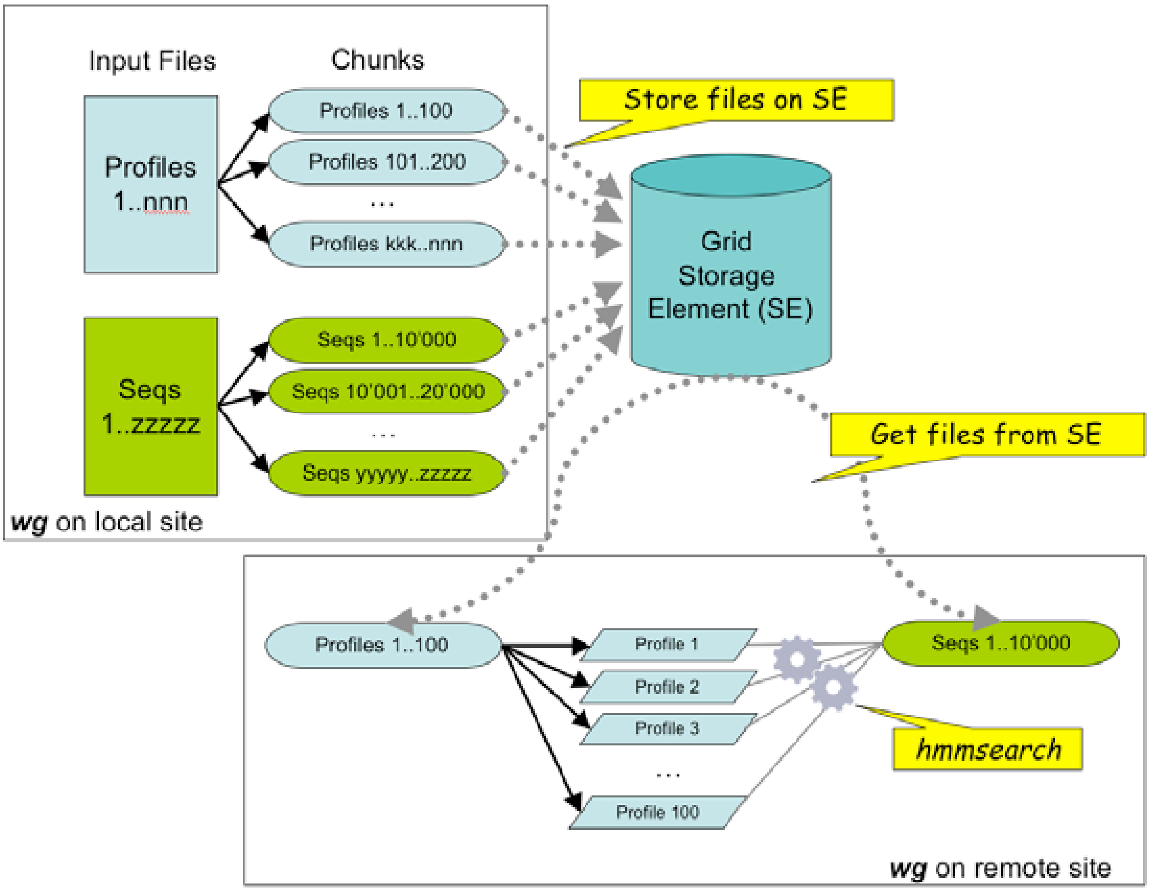


Figure 1. Creation of data chunks and the execution of wg on the local as well as remote machines.

with many sequences. Consequently, Grid jobs run in *parallel* whereas each job executes its tasks in *sequential order*. In order to achieve the overall goal, the following steps are needed (and executed by *wg* as shown in Figure 1):

1. **Data preprocessing**: Determination of the size of the problem. If it is large enough, data chunks are created and sent to specific Storage Elements in the Grid where they can then be retrieved by the nodes processing the data (cf. Figure 1). The creation of these data chunks is actually a bi-dimensional problem where the number of chunks of profile-HMMs and of chunks of sequences has to be determined to be executed in parallel.

2. **Creation of Grid job descriptors**: In order to submit a *wg* job to the Grid, one first needs to write a job descriptor that has information on the executable, its input parameters and its output parameters. This step takes care of creating a job description file for each job that can run in parallel with others.

3. **Job submission**: Jobs can then be submitted to the Grid for parallel execution. Our system supports several different Grid types (implementations), and the user can select which one to use.

4. **Remote execution**: All jobs that are submitted to Grid computing services will then execute on the worker nodes of the remote systems. In detail, an instance of *wg* is executed remotely, fetches the necessary data from a Grid storage element, creates chunks and then executes *hmmsearch* (cf. Figure 1). The jobs write their output into the local directories of the computing nodes and prepare them for transfer back to the user.

5. **Merging of results**: Since each distributed job creates its local output, all outputs need to

be merged and made available to the end user. The Grid middleware system takes care of that.

Scheduling/Execution Algorithms

Data Grid infrastructures such as EGEE or ARC have built-in brokering algorithms that are used for job scheduling. Such an algorithm tries to optimize for *high throughput* of many concurrent jobs. In principle, we can make use of this but we are more interested in *high performance* of our sequence search and alignment process, i.e. we want to minimize the wall clock time of the overall execution. Our system has two built-in scheduling-execution algorithms, which optimize the steps 3 and 4 mentioned in the previous subsection:

- **A-priori algorithm (AP-algorithm)**: This simple algorithm is based on the fact that all jobs get roughly the same amount of work to be computed. For a given number of tasks we decide at the beginning of the job submission phase to run a certain number of jobs in parallel, each with the same number of tasks to be executed. It is assumed that all jobs finish within a certain time. In scheduling theory this is commonly referred to as a **push model** where tasks are pushed to an execution environment without checking all the details of the run-time environment. Note that this is a very valid approach used by several schedulers such as the one in gLite (EGEE's middleware system). We consider the AP-algorithm to be non-optimal but list it here for the sake of comparison.

- **Run time sensitive algorithm (RTS-algorithm)**: The AP algorithm is simple but has a few problems since it does not utilize all available Grid resources in an efficient way due to its lack of monitoring the execution times of jobs. The RTS algorithm provides an improvement in several ways. The algorithm starts in the same way as the AP algorithm: parallel jobs are submitted to Grid worker nodes. However, the RTS-algorithm does not decide a-priori which data will be processed. Once a job has started its execution on a worker node, and the first few instructions are executed correctly, there is a much lower chance that a job will fail (most of the job failures in a Grid environment are due to node missconfiguration or problems with getting the job started in the local batch system). This is particularly true for parallel jobs that always run the same application program such as *hmmsearch*. Consequently, a task is only assigned to a job once this initial launch has been passed, and a higher probability for correct job execution is yielded. This computing model can be compared to the BOINC [2] approach where tasks are **pulled** from a remote site to an execution site. In more detail, once a job has been correctly started at a worker node and has passed the critical phase of job execution, it contacts a **Task Server** for a task to be computed. Once the task is finished, the job can ask for another task until all tasks have been computed (Figure 2). Finally, the Task Server has control over how long jobs are running and can take care of fault-tolerance and recovery in case jobs are failing or not finishing within a certain time limit, i.e. if tasks are not finished within a certain time, they can be marked for re-processing by other processors. This has the advantage that fast processors get more tasks than slower processors. It is important that the tasks are small enough to smooth out run time differences of different processors.

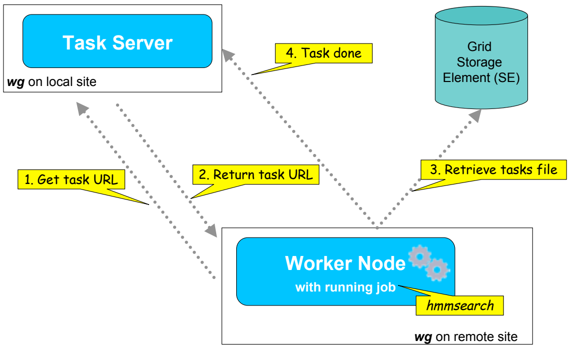


Figure 2. The Task Server (operated as a Web Service) returns locations of the individual task files on the Storage Element (URLs) on demand. The running job (i.e. an instance of wg) retrieves the task file from the Storage Element and then processes it locally.

Implementation and Test bed Environment

In order to achieve the five necessary steps pointed out above and to submit large-scale

Grid jobs, we have written a C++ (gcc 4.0.2 on GNU/Linux kernel 2.6.14) tool called ***wg*** (Workload Generator) that can do the necessary processing on the local desktop as well as on remote machines. The client tool is statically *linked* in order to avoid execution failures due to potential differences of shared libraries at remote Grid sites. The size of the executable is about 2.2 MB, which might cause a submission problem in case the input sandbox of the Grid software is limited. The Task Server is implemented as a Web Service using SOAP for communication (gSOAP 2.7 for C++) published by WSDL. The basic interface of the Task Server is shown in Figure 2. The server also provides an administrative interface for registering task files that are available on a Storage Element. The Task Server keeps track of all the states of jobs and their tasks. In detail, a MySQL database back-end is used to store meta-data such as URLs, start and end dates for tasks. In this way the system can easily find out how long jobs have been running, mark them for resubmission etc. The software has been written in a modular way to be able to cope with several Grid (EGEE, ARC) and cluster platforms (LSF). For instance, if a new type of Grid system needs to be supported, an additional implementation of the Grid interface (mainly for job submission and data transfer) needs to be implemented. The same is true for the interface to the database back-end. Most of the run-time tests and in particular the ones in the next section have been done on EGEE's worldwide distributed Grid infrastructure, using software versions LCG2.7 and gLite 3.0. All sites run GNU/Linux (typically Scientific Linux CERN) whereas we mainly developed on Fedora Core 4, using the native default compiler. For both EGEE as well as ARC, Globus Toolkit 2 with GSI needs to be available on the local machine where the *wg* client is started.

For the experimental results we used the following setup:

- EGEE Grid infrastructure with a certain VO (virtual organization) to have access to more than 80 distributed sites. Each of the sites deploys LSF, PBS or similar on a local cluster with 2 to 128 processors, i.e. in total we have potential access to more than 1000 CPUs. The EGEE infrastructure is deployed in production use, which means that many of the clusters are heavily loaded and only allows us to use a subset of processors at any given time

- Local cluster at SIB in Lausanne with 64 nodes, LSF and HP's Lustre shared file system. The nodes of the cluster run Red Hat Advanced Server 4.1 and Linux kernel 2.6.

- Our bioinformatics tool we mainly used is *hmmsearch* version 2.3.2. In order to create synthetic profiles for the experiments in the next section we used the tools *hmmemit* and *hmmbuild*, both from the same HMMER package 2.3.2.

- We use a biological benchmark dataset which consists of a profile-HMM database with **7,868 profile-HMMs** (PFAM version 17 March 2005) and a synthetic sequence database with **10,923 sequences** (Swiss-Pro splice variants, Swiss-Prot 46 from February 2005). In addition we use synthetic databases defined to have a guaranteed match of sequences in the order of 1% in the biological alignment process.

Experimental Results and Discussion

We are interested in the overall performance (wall clock time) needed to analyse data. In particular, we have designed and run the following three experiments:

1. **Benchmarking an heterogeneous Grid environment**. We show different execution times and job failure rates of identical jobs.

2. **Comparisons of algorithms**. Here, we are interested in the performance comparison of the AP vs. the RTS algorithm.

3. **Single CPU vs. cluster vs. Grid**. A parameter study is conducted to show when it is more efficient to run the job either on a single CPU, a cluster or on a Grid.

Heterogeneous Grid environment

Given the heterogeneity of the hardware used in the EGEE infrastructure, it is rather difficult to give precise performance predictions of existing jobs. In order to show the different job execution times, we submitted 100 identical jobs (20 profile-HMMs against the identical benchmark sequence database) to EGEE. Most of the jobs were running for 40 to 60 minutes but there were a few ones, which were running much longer (more than 6 hours) or were aborted. This is a general phenomenon observed on a Grid environment.

Comparisons of Algorithms

In order to measure the performance of the search and alignment process applying our algorithms we used the benchmark dataset mentioned in previous section. We ran the hmmsearch program over all profile-HMMs and all sequences using different algorithms and platforms as depicted in Figure 3. The overall execution time for on a single CPU is more than 144 hours, which is our performance reference. In order to benefit from the parallel/distributed infrastructure we produced chunks of the profile-HMM database and ran them against the full sequence database of our benchmark. For the RTS algorithm we created 400 tasks, each of them containing about 20 profile-HMMs to cover all 7,868 profile-HMMs of the benchmark dataset. In the performance comparison between single CPU, AP and RTS algorithm we measure the wall clock time for the execution of parallel jobs running on a subset of data. This time includes all the steps mentioned above, i.e. includes preprocessing, job submission, data transfers and the retrieval of output.

We varied the number of jobs for both the AP (64, 128 parallel jobs) and the RTS (64,128, 256, 512) as well as the cluster (64, 128 and 256 jobs) running LSF. For the RTS algorithm we use a Task Server that monitors the progress of running jobs and asks for resubmission of tasks in case they are not executed within a certain time frame (in our tests 1 hour). Therefore, even if jobs fail during their executions, these tasks can be reassigned to other processors in order to avoid errors in the overall job execution.

In all cases, the RTS algorithm outperforms the AP algorithm due to its better way to execute more tasks on faster machines than on slower ones. It is also fault tolerant since the system makes sure that all tasks have finished. The RTS algorithm works very well for large-scale jobs and can outperform medium size clusters.

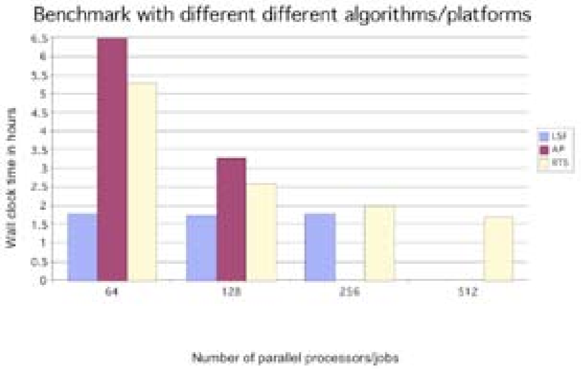


Figure 3. Wall clock time of parallel jobs running on 64, 128, 256 and 512 processors, each running one benchmark job. We compare the two Grid algorithms (AP and RTS) with the execution on a LSF cluster with 64 CPUs.

Single CPU vs. Cluster vs. Grid

In the previous subsection we have seen that a Grid can outperform clusters only if the problem size is sufficiently large, and the Grid offers a significantly bigger number of processors than a cluster. In order not to degrade the performance of the job execution it is therefore of major importance to decide if a given problem should either run on a single CPU, a cluster (assumed we know the *maximum* number of nodes to be used) and a Grid where we need to know at least the *minimum* number of sites and processors available. We therefore conducted another experiment (again sequence search and alignment using *wg* and *hmmsearch* using a very large sequence dataset (up to 5,000,000 sequences) and 256 profile-HMMs. This represents a rather realistic biological use case. In order to find out when a single CPU, a cluster or a Grid performs better, we changed the number of sequences from 10 to 5,000,000 while we kept the number of profile-HMM constant at 256. For the Grid case, we only used the superior RTS algorithm with 256 parallel jobs. Details of the performance comparison can be found in Figure 4. Running only the smallest benchmark (256 profile-HMMs against 10 sequences) is executed faster on a single CPU whereas in all other cases the cluster is faster. In contrast, the RTS take 10 minutes for executing small jobs (up to 5000 sequences on 256 parallel nodes. However, this time includes the job submission time (jobs where submitted with 3s difference from each other in order not to overload the resource broker), data transfer times etc. which results in the more or less minimal ex-

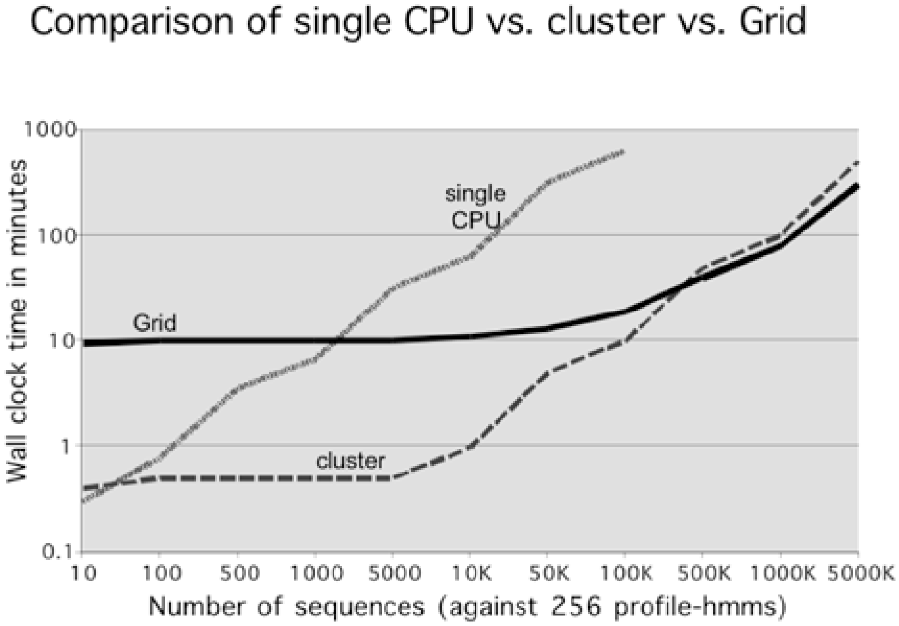


Figure 4. For small to medium size problems the cluster performs better than the Grid. However, for large scale problems (more than 500,000 sequences vs. 256 profile-HMMs) the Grid gives a better performance.

ecution time of 10 minutes for the benchmark jobs. In general, a simple HelloWorld program can be executed in about 5 minutes on the EGEE infrastructure.

Once we increase the problem size, the additional overhead of the Grid becomes insignificant due to the longer execution times of the jobs. We can see that with about 500,000 sequences against 256 profile-HMMs the Grid performance is better than the performance of the 64-node LS cluster. Once this point is reached, problems with larger size and similar characteristics can run faster on the Grid than on the cluster.

Related Work

The project that is most related to our work in term of the Task Server is BOINC [2]. BOINC is used to implement projects such as SETI@home and provides a very good architecture and implementation for *volunteer computing* which is partly in contrast with a model for Grid computing that is often based on a dedicated infrastructure. BOINC's Task Server is highly scalable and is also based on a MySQL database back-end. Another interesting approach is called *persistent redundant computing* where several clients calculate the same task (redundantly) to allow for correct results. The main difference to our approach is that *wg* actively uses a Grid scheduler and also a Grid information system to find possible sites to execute tasks whereas in the BOINC approach potential clients contact a BOINC server, which then schedules the tasks. One of the most known bioinformatics tools is BLAST, which can be easily parallelized and executed on clusters or Grids. For instance, BLAST is often run on clusters using Condor [12]. However, our tool focuses on supporting both, clusters and Grids. Nimrod/G [1] is another related project that can utilize free CPU cycles in both cluster and Grid environments. Its commercially available version has been successfully applied to bioinformatics problems for protein modelling. Other types of bioinformatics problems that were ported to Grid environments are both computation and *communication* intensive. A representative example is the CHARMM application for protein folding that was successfully run on Legion [19]. Such applications typically use MPI, which is not the focus of our work.

Conclusion

We have designed and implemented an efficient and flexible method to accelerate embarrassingly parallel problem such as *hmmsearch* and *pfsearch* on a Grid. Given that large-scale Grid infrastructures have rather unpredictable performance figures, our system takes care of these performance fluctuations by applying the Run Time Sensitive algorithm that dynamically adapts to given and changing performance parameters in the Grid. Therefore, we gain "maximum" of overall performance for large-scale computations and provide a high-performance solution to scientists that do not have access to a large computing cluster. We have also shown a benchmark that allows making a decision of where it is better to execute the job (cluster vs. Grid) and conclude that sufficiently large scale, embarrassingly parallel problems can achieve better performance on Grids rather than on small or medium scale (up to 64 nodes clusters).

To sum up, our proposed algorithm, together with our tool can significantly help bioinformaticians to run their CPU intensive problems using the Grid.

Acknowledgements

A longer version of this article has been published in the 2nd *IEEE International Conference on e-Science and Grid Computing (e-Science 2006), Amsterdam, Dec. 2006.*

This work was funded in part by the EU project EMBRACE Grid which is funded by the European Commission within it FP6 Programme, under the thematic area "Life sciences, genomic and biotechnology for health", contract number LUNG-CT-2004512092. We also acknowledge the EGEE project for the usage of the Grid infrastructure.

References

[1] D. Abramson, et al. High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid? Int. Parallel and Distr. Processing Symp., Mexico, May 2000.

[2] D. Anderson, E. Korpela, R. Walton. High-Performance Task Distribution for Volunteer Computing. IEEE Int. Conf. on eScience and Grid Technologies, Australia, Dec. 5-8, 2005.

[3] C. Blanchet, C. Combet, G. Deleage. Integrating Bioinformatics Resources on the EGEE Grid Platform, Int. Workshop on Biomedical Comp. on the Grid, Singapore, May 16-19, 2006.

[4] F. Brugnara, R. De Mori, D. Giuliani, M. Omologo. A Family of Parallel Hidden Markov Models, IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, California, March 1992.

[7] M. Ellert et al. The NorduGrid Project: Using Globus Toolkit for Building Grid Infrastructure. Nucl. Instr. and Methods A502:407-410, 2003.

[8] Enabling Grids for E-sciencE (EGEE) project: http://www.euegee.org

[9] EMBRACE Grid project: http://www.embracegrid.info

[10] HMMER: Profile HMMs for Protein Sequence Analysis. http://hmmer.wustl.edu

[11] D. Hull, et al. Taverna: A Tool for Building and Running Workflows of Services, Nucleic Acids Research, 2006.

[12] M. Karo, et al. Applying Grid Technologies to Bioinformatics. IEEE Symp. on High Perf. Distr. Computing, San Francisco, California, Aug. 7-9, 2001.

[14] E. Laure, H. Stockinger, K. Stockinger. Performance Engineering in Data Grids, Concurrency and Computation: Practice and Experience, Wiley Press, 17(2-4):171-191, 2005.

[15] P. Lord, et al. Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt. International Semantic Web Conference, Hiroshima, Japan, Nov. 7-11 2004.

[16] LSF, Platform Computing: http://www.platform.com/Products/

[17] B. J. Moreno. Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments. Digital Cambridge Research Laboratory Technical Report, CRL 97/7, Sep. 1997.

[19] A. Natrajan, et al. Studying Protein Folding on the Grid: Experiences Using CHARM on NPACI Resources under Legion, IEEE Symp. on High Perf. Distr. Computing, CA, Aug. 2001.

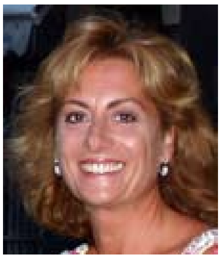
[20] PFTools, http://www.isrec.isb-sib.ch/ftp-server/pftools/

[22] SAM, http://www.cse.ucsc.edu/research/compbio/sam.html

[23] Swiss Bio Grid: http://www.swissbiogrid.ch

[24] W. Zhu, Y. Niu, J. Lu, G. R. Gao. Implementing Parallel Hmm-pfam on the EARTH Multithreaded Architecture. Computational Systems Bioinformatics, Stanford, CA, Aug. 2003

The MitoRes database



Domenica D'Elia

Institute for Biomedical Technologies, CNR, Via Amendola 122/D, 70126 Bari, Italy

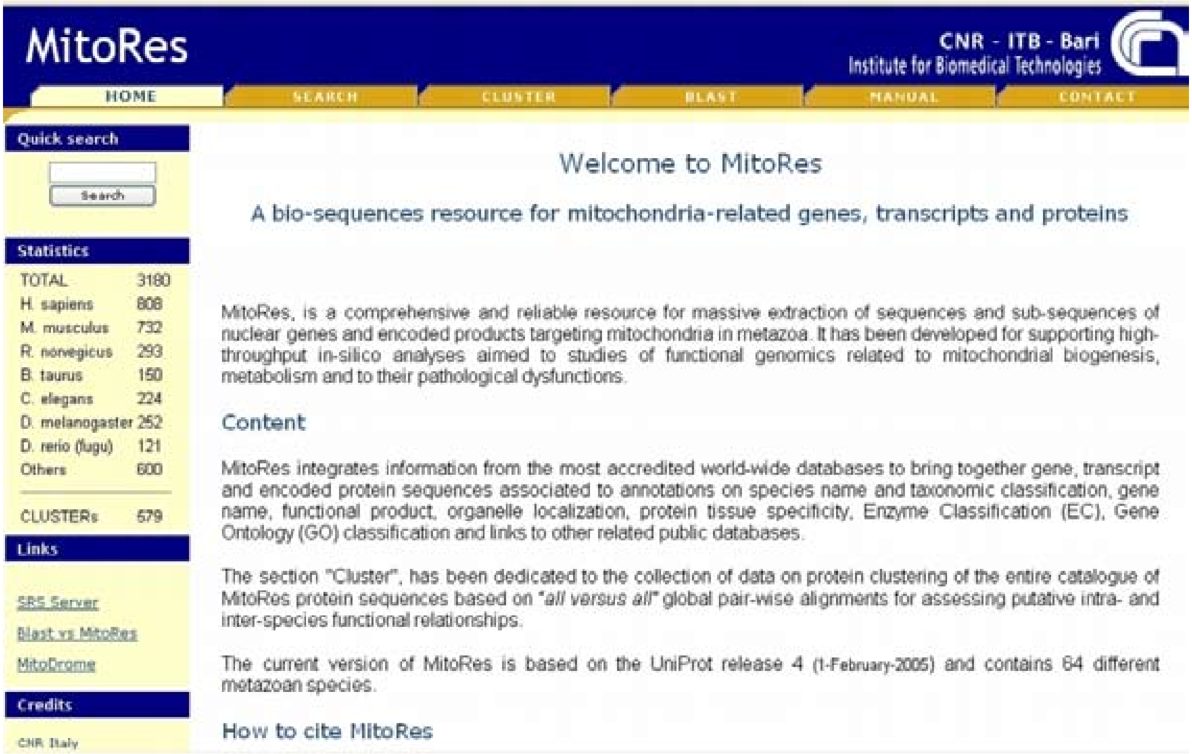
The MitoRes database is an integrated resource of nuclear-encoded mitochondria sequences. It collects and integrates protein, transcript and gene sequences from every metazoan species available to date and implements tools for the retrieval and export of different categories of sequences. The aim is to support researchers interested in the functional characterization and comparative analysis of mitochondria targeting sequences, in overcoming problems they face when searching for different subsets of related sequences from different resources and for different organisms. As a further support, MitoRes also makes a comparison and carries out clustering of protein sequences based on their similarity using an "all versus all" pair-wise global alignment and integrating data obtained from protein comparison with data on their gene structure. Clusters are collected in the MitoRes section named "CLUSTER". This article is aimed at giving an overview of the tools provided by MitoRes and some suggestions on how they can be used to exploit the database in the best way.

Structure and content

MitoRes is a MySQL relational database which consolidates information from different biological databases e.g., UniProt, ENSEMBL, RefSeq, UTRdb et al. Currently, the database provides annotations from 64 different metazoan species with a total of 3180 records and 519 protein clusters and is public available at the following address: `http://www2.ba.itb.cnr.it/MitoRes/`. A more detailed description of MitoRes appears in BMC Bioinformatics 2006, 7:36.

HOW TO USE MITORES

The MitoRes home page (Figure 1) provides the direct access to different resources:



Welcome to MitoRes

A bio-sequences resource for mitochondria-related genes, transcripts and proteins

Statistics

TOTAL	3180
H. sapiens	808
M. musculus	732
R. norvegicus	293
B. taurus	150
C. elegans	224
D. melanogaster	252
D. rerio (fugu)	121
Others	600
CLUSTERs	579

Links

SRS Server

Blast vs MitoRes

MitoDrome

Credits

CNR Italy

MitoRes, is a comprehensive and reliable resource for massive extraction of sequences and sub-sequences of nuclear genes and encoded products targeting mitochondria in metazoa. It has been developed for supporting high-throughput in-silico analyses aimed to studies of functional genomics related to mitochondrial biogenesis, metabolism and to their pathological dysfunctions.

Content

MitoRes integrates information from the most accredited world-wide databases to bring together gene, transcript and encoded protein sequences associated to annotations on species name and taxonomic classification, gene name, functional product, organelle localization, protein tissue specificity, Enzyme Classification (EC), Gene Ontology (GO) classification and links to other related public databases.

The section "Cluster", has been dedicated to the collection of data on protein clustering of the entire catalogue of MitoRes protein sequences based on "*all versus all*" global pair-wise alignments for assessing putative intra- and inter-species functional relationships.

The current version of MitoRes is based on the UniProt release 4 (1-February-2005) and contains 64 different metazoan species.

How to cite MitoRes

Figure 1. MitoRes database home page. The database resources are accessible from the tool bar on top.

- the SEARCH page, to perform queries and export both the query report and sequences;

- the CLUSTER section, where proteins are grouped in clusters;

- the BLAST tool, through which the user can perform database searches using any type of sequence against the whole MitoRes sequence collection;

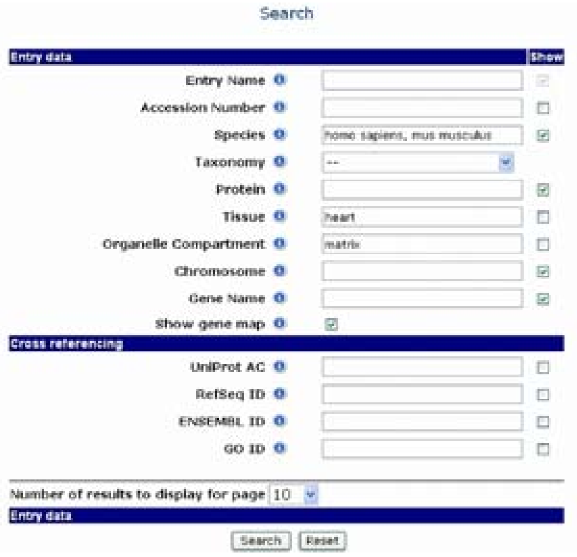
- the MANUAL, which describes the structure and content of the database and supports users in how to use it;

- the CONTACT page, providing details on how to contact the MitoRes authors.

How to search the database

To search the database, you can use a quick search field available on the database home page or the Search page. The quick search field supports lists of MitoRes ID, UniProt accession number (AC) and gene name while the MitoRes search page supports queries based on a wider range of search criteria. An example is shown in Figure2.

The "show" check box placed on the right side of each search field allows changing the format of the query report depending on your own needs. By default, the report of matching records includes MitoRes identifier (ID), species and gene name, chromosome number and protein function. Using the "show" check boxes, any of this information can be excluded and other categories of major interest for your specific aim can be added. Therefore, thanks to this option MitoRes can also be used to build up lists of gene and/or of sequence identifiers. For example, if you need the list of UniProt AC, RefSeq and/or ENSEMBL ID related to the proteins or genes you are searching for, it is enough for you to select the relevant check boxes in the search page. The database identifiers of each record matching the query criteria used, will be listed in the DB X-ref column in the query report. The query report can be saved using the "Save search" option located on the top of the table and exported in a format readable by Excel as well as by any text editor program.



Search

Entry data		Show
Entry Name		□
Accession Number		□
Species	homo sapiens, mus musculus	☑
Taxonomy	--	
Protein		☑
Tissue	heart	□
Organelle Compartment	matrix	□
Chromosome		☑
Gene Name		☑
Show gene map	☑	
Cross referencing		
UniProt AC		□
RefSeq ID		□
ENSEMBL ID		□
GO ID		□

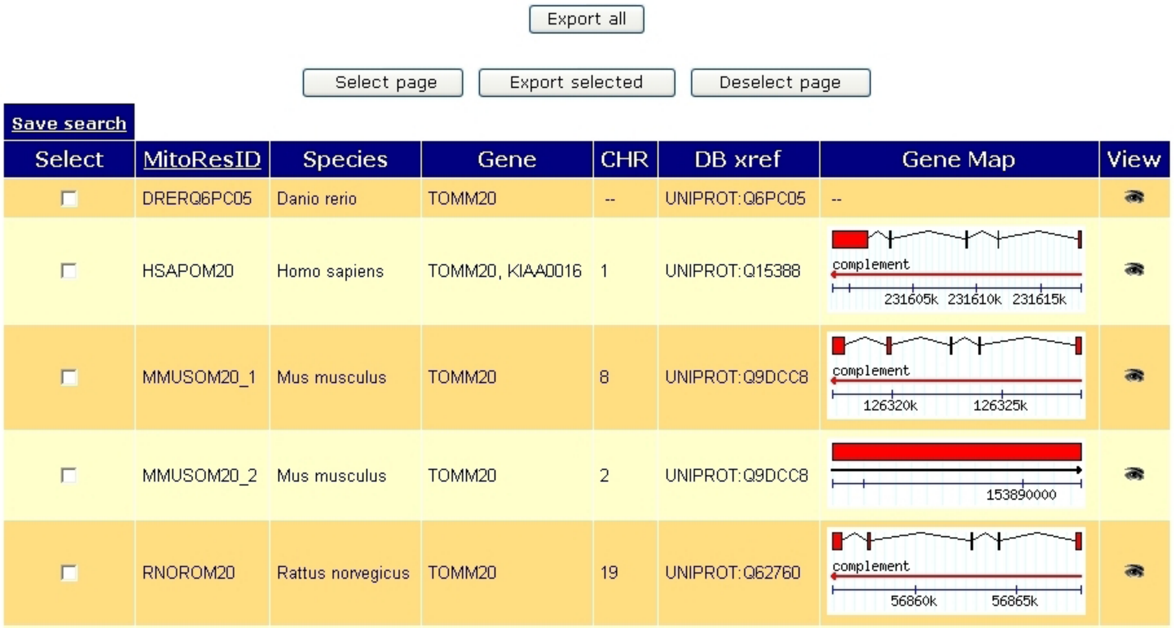
Number of results to display for page 10

Entry data

Figure 2. The database query page. In this figure is shown a search for matrix mitochondrial proteins in *H. sapiens* and *M. musculus* whose expression has been tested in heart. The cross-referencing search fields can be used with lists of sequence identifiers from UniProt, RefSeq, ENSEMBL and GO respectively, and the "show" check boxes to select information the user wants to be displayed in the query report.

Another useful option provided by MitoRes is the inclusion, in the query report, of the gene map. If you have carried out a query for the same protein in different organisms the inclusion of the gene map in the query report allows you to have immediately the picture of the availability of genomic annotation about the protein you are interested in, how many copies of the gene are annotated in each species and the conservation of the gene structure in the species investigated (exon/intron organization). Figure 3 shows the result obtained searching for the Tom 20, a central component of the TOM mitochondrial receptor complex. In this case, it has been chosen to include into the query report, in addition to prefixed information, also the UNIPROT accession number and the gene map.

As you can see, from a view of the query report, a lot of information can be easily deduced. Annotations about the protein are available in four different species, while annotations about the gene are available for only three of them (*H. sapiens*, *R. norvegicus* and *M. musculus*). The structure of the gene seems well conserved in all the species examined, but *H. sapiens* and *R. norvegicus* have only one copy of the gene, while the *M. musculus* has two copies. The two gene



Select	MitoResID	Species	Gene	CHR	DB xref	Gene Map	View
□	DRERQ6PC05	Danio rerio	TOMM20	--	UNIPROT:Q6PC05	--	👁
□	HSAPOM20	Homo sapiens	TOMM20, KIAA0016	1	UNIPROT:Q15388	complement 231605k 231610k 231615k	👁
□	MMUSOM20_1	Mus musculus	TOMM20	8	UNIPROT:Q9DCC8	complement 126320k 126325k	👁
□	MMUSOM20_2	Mus musculus	TOMM20	2	UNIPROT:Q9DCC8	153890000	👁
□	RNOROM20	Rattus norvegicus	TOMM20	19	UNIPROT:Q62760	complement 56860k 56865k	👁

Figure 3. An example of query report. The format chosen for the query report integrates information coming from UniProt about the protein (organisms in which the protein has been characterized, along with the UniProt AC) and information about the related gene from ENSEMBL (structure and nuclear localization). The figure of the gene map is descriptive of gene structure, sequence orientation and genomic localization.

copies in *M. musculus* are located on different chromosomes and one of the two gene copies, the one located on chromosome 2, has not a typical exon/intron organization: hence, there are good probabilities that it is a copy of the ortholog coming from a species-specific retro-transposition event. The information deduced on the basis of the general picture given by the query report can be further investigated having a view of the database entries and/or by extracting and analyzing the sequences of interest.

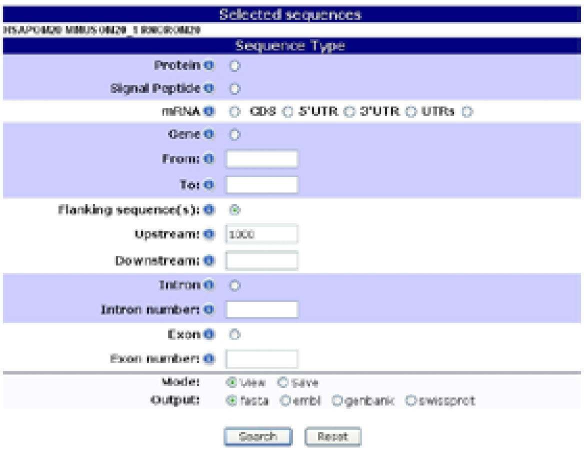
Visualization of entry content and export of sequences

The database entries listed in the query report can be viewed clicking on the "eye" icon in the last column of the query report. The entry shows the annotations about the protein and gene and transcript/s structure, provides links to related databases (ENSEMBL, UniProt and RefSeq) and allows the direct access to the interface for the extraction and export of sequences as well as to the cluster of homologues, if any, in other species. If the gene codes for more than one isoform, for each one of them, information is reported into the entry, including also the links to the related RefSeq and UTRef database entries.

The extraction of sequences can be performed from the query report interface or from within each entry using the "Export sequence" button on top. By Using the "Export" option from the query report, it is possible to perform the extraction of sequences from all the retrieved entries or only from those selected checking the box in the "Select" column of the report. The database interface for the extraction of sequences allows to extract and export, in different file formats, any type of sequence you want such as the protein or only the signal peptide; the complete mRNA sequence, the CDS or the UTRs regions; the complete gene sequence or part of it and the flanking gene regions up to 5000 bp; all the intron and exon sequences or only those specifically chosen by you. A picture of the "Export sequence" interface is shown in Figure 4.

CLUSTER section

The Cluster section of the database collects proteins on the basis of their sequence similarity. Proteins are clustered in the same group, if they have a sequence similarity up to 60%. Clusters can be accessed both from the home page and from the database entry; in the last case, if a



Selected sequences

HSAPOM20 MMUSOM20_1 RNOROM20

Sequence Type

Protein ○

Signal Peptide ○

mRNA ○ CDS ○ 5'UTR ○ 3'UTR ○ UTRs ○

Gene ○

From:

To:

Flanking sequence(s): ●

Upstream: 1000

Downstream:

Intron ○

Intron number:

Exon ○

Exon number:

Mode: ● View ○ Save

Output: ● fasta ○ embl ○ genbank ○ swissprot

Figure 4. MitoRes interface for the extraction and export of sequences. In this picture the extraction and export of the first 1000bp upstream the gene is reported as an example of the potential application. The example is referred to the sequences export from the entries of TOMM20 orthologues identified by the query report shown in Figure 3.

protein belongs to a Cluster, in the entry appears a clickable button on top named "Associated cluster". Clicking on this button the system shows the cluster entry which reports the list of putative homologues in other species and also the gene name and map for each organism. You could investigate on each entry associated to the cluster by clicking on the MitoRes ID listed in the Cluster entry and/or extract any type of sequences from all the entries belonging to the same Cluster using the "Export all" button placed on top of the entry. Beside being a useful instrument "per se", the CLUSTER section could also be used as a tool to overcome difficulties commonly encountered in database searching and due to the well-known lack of standardization in biological annotations. The example used to describe the functionality of the MitoRes query and sequence export tools illustrates the case of a search for the same gene in different species. In many cases, the same gene has a different name in different species or can be edited in a different manner (i.e., COX4, COXIV, COX4I1); hence, a database search carried out using as search criterion the gene name could be unfruitful. A big help in this case could come from the use of the Cluster section. Indeed, although only one entry is retrieved by the query system, if this entry belongs to a cluster, by accessing the Cluster entry, you will be able to have a view of similar proteins present into the database, proteins which are very likely the homologues in other species.

BLAST tool

The BLAST tool provides you with the possibility to search the database also using a sequence of any type against the whole collection of MitoRes sequences. In this case, before to execute the Blast search you have to perform your registration. This procedure is required only for the management of the work flow on our server and is absolutely free of charge. The results of the Blast search will be send to your e _ mail address as soon as they are ready.

Remarks

MitoRes is periodically updated through an automated procedure making use of a suite of BioPerl and C programs that retrieve, collect and integrate data from centralized databases and populate the database records automatically. The Contact page, accessible from the database home page, provides name and address of the MitoRes authors; you can contact us to ask for further support or send an e _ mail for comments or suggestions.

Announcement

EMBO Workshop: Molecular Biodiversity and DNA Barcode

Accademia dei Lincei, Rome (Italy)
17 - 19 May 2007

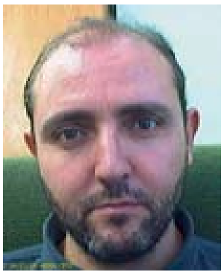
The goal of the workshop:

- Molecular Biodiversity and Evolutionary Genomics
- DNA Barcode
- Molecular Biodiversity in different lineages
- Methodology and Computational Biology
- Applications

If you are interested in attending please register online at `http://cwp.embo.org/w07-28/`

Abstract submission deadline: 2 April 2007
Registration deadline: 30 April 2007

Some key issues of electronic publishing in e-Learning platform



José R. Valverde

EMBnet/CNB, Centro
Nacional de Biotecnología,
CSIC Campus Univ.
Autónoma Cantoblanco,
Madrid 28049, Spain

Doubtless, the most difficult and important task that all humans have to face in their life is that of learning and, by extension, teaching. There has been no single parent (and we have constance since the first written records) that has not complained of the difficulty of the task nor wished she/he could benefit from the experience of others.

In modern society the problem is compound by the need to teach students to work and solve problems cooperatively, in sharp contrast with traditional, individualistic academic approaches. Indeed, it is even ironic that those charged with this task have so little experience with sharing our teaching experiences, materials and methods with other colleagues, often excused as shyness about quality or secrecy protection of superior crafts.

In these articles we want to introduce you to new approaches to share your training experience, materials and methods with your colleagues, showing why and how this is good for everybody and, mostly, for you and how both you and your students can benefit of the cooperative culture empowered by the Internet.

These first two articles are relevant to two main issues of electronic publishing: the upload of authors' material (courses) on a common platform (Advice to authors) and the hard task to choose a licence under which authors have to release their works (Choosing a license).

Advice to authors

First the legal advice:

Please, try to state always the licensing terms explicitly for any materials you add or use. If you use materials from other sources in your course, try to make sure that:

- they are legally usable under Fair Use, academic or educational terms

or else

- they are prominently noted as external contributions (including a link to the original site whenever possible)
- they are available or have been released under a license equally or less restrictive than the one you use
- you attributed properly authorship of materials
- they include the original license so readers will know which terms apply to which materials

Include a history page on the header of all your courses. This page should detail in a historic way (much like the historic annotations of source code) the original author, date and license, as well as all changes made to the original materials, stating for each change who did it, a description of the changes made, why they were done and the date when they were applied.

I always try to make it prominent and evident, usually stating it forefront in the course headers and/or the introduction.

Choosing a name

When one starts something it is tempting to let the horses run loose (like a pioneer in the prairies of Far West) and start producing materials as quick as you can. I know, I have done it myself and now regret it (well, no, sort of) for my lack of provision for the future.

Your first step when you create a course will be to *give your course a name*. This is fine, it is your course and you get to name it, that's fair. However you must keep in mind that other people are using the system as well, and they may as well want to use the same name, or even have chosen it before you came up with it. That is one reason to be cautious when choosing a name, but not the only one.

A more compelling reason has less to do with the adventurous world of explorers and more with the clerical routine of the real world. You are creating a course for a reason, and that is to have students use it. Think of it: these students will probably want a graduation certificate. Good, you think, it is easy to create one and mail it over. But that is not the end of the story: now you have a liability to prove the grades you assigned should any questions arise, and hence you must conserve this specific instance of the course for some time. So, if you want to reuse the materials you will need to create a new course, import everything (except user data) from the old one, maybe add some changes and go ahead. Now you have a name conflict: you stepped in your toes.

How long do you need to keep these materials? Well, it depends on how long you will need them. If you can save all relevant data to a safe place for later referral, you may be able to do without it and delete the course after several months, a year or whatever you are required the grades to remain unchallenged. The alternative probably is that *a course must be maintained for life.*

Maintained for life! How comes? you ask. Well, people forget or lose things that they will need later. When the time comes, maybe ten years after the course took place, you are guaranteed to get a request from an old student asking for a new certificate to prove his grade somewhere. And when this time comes, where are you going to look into to find out what happened? Was actually this person a student of yours? In which course? Which year? What happened at the time?

Hence my advice for you is simple: plan that any course you create may stay around for a long time. You don't want conflicts with other teachers, not even with yourself later on, and you want to be able to identify an odd course that took place long ago easily.

As a courseware system is usually a shared facility, you should find out what the convention is in your server. *A sensible example of a naming policy* might be:

- Choose a name for your course
    - Anything you like, better if it is descriptive and related to the topics, and even better yet if it sounds attractive
    - Add to it something to identify yourself (it may be the two-letter country code, the acronym of your institution, etc...)
    - Add to it a date, may be the year, or a year and month,... depending on the periodicity you expect for the course
- Choose a short name for the course
    - Following the same advice but making it shorter

An example might be "Molecular Dynamics, EMBnet/CNB, Feb 2006", with a short name like "MD-ES-2006".

If you are to reuse the materials of a course later on, all you need to do is create a new course with the updated name, and then use the import utility of the system to carry over all the materials from the old course, or may be use a backup (lacking user data) to restore all the materials to the newly created course.

Always include a backup of your course

Have you ever used a reference book in your discipline? Ever wondered why it was so? Ever wished to write one?

It is very easy: the continued success of most reference books depends on two factors, one is the quality of the contents, of course, the other, less evident, relies on the continued quality maintained by contributors. Most reference books are no longer written by their original author (possibly dead several decades ago), and still they keep their original name alive for ever. This is because the original author did not mind others adding or enhancing to his work, and because those others kept his name to acknowledge his seminal work.

You can do the same. Write your course and pay close attention to quality. Then make sure others can use and add to it easily. Let them help you enhance it.

The only problem is that if you just publish a work it will be difficult for others to use or add to it unless they all use your own server to work and lecture... something that is very difficult to do and coordinate on a world-wide basis. The only practical way is to give them the contents for their own use. However, copying all the contents of a feature-rich environment such as Moodle one by one is a daunting task (I know, I've done it myself many times), and almost nobody will take the pains.

That is why you should *always include on your courses*:

- a full backup in moodle format
- a history page detailing all modifications
- a license notice

This way it will be easy for everybody to know that they can make a copy and use your materials elsewhere, it will be trivial to do so, and acknowledging your work will be a snap.

It would be wise if you also *request **cordially** that significant changes be submitted back* to you for merging in the original course and further sharing, or at least, that they notify you if they use your materials.

Even if you don't get contributions, you will be able to state in your curriculum that your works have reached a level of quality high enough to guarantee their use at others institutions -which is an acknowledgement by them of the quality of your work-.

Think of it, what does look better in your curriculum: stating you are a great professor with terribly good course notes that you only offer as a private added-value for your students or stating that your course notes may not be that terribly good as you yourself would describe them, nor exclusive to your students, but they are good enough for Harvard, Yale or Cambridge to use them on their own courses?

Personally, I know what would make me more proud of my work. And certainly praise by others ranks higher in my list than contemptuous praise by myself.

Make it easy for others to use your notes and they will. Make it difficult and they will write their own.

To provide a course backup in Moodle, first create a backup of your course, then move it to the files area (as the backup area is not accessible to other people) and finally make a link to this accessible backup from the header of the course.

Always include a history page

Whether you used materials from others or not, whether you allow for modifications or not, it is very difficult that any work will remain unchanged for a long time. May be only you make changes to it, but it is still important to clearly annotate all changes done, why they where done, and include dates and authorship details.

This is akin to common practice in software development and is important for the same reasons:

- this is the only way to properly acknowledge all authors involved in writing a work
- this is the only way to know who made what, when and why
- the history serves as a track of changes
- a properly written history allows you to prove when some important piece of work was actually done (and defend your authorship in case of legal disputes)

Including a history page allows you to acknowledge the original authors, as well as any further contributing authors, know what each person has done and understand why, so you can further enhance knowingly if needed, and of course allows you as well to state which changes you have contributed, explaining your reasons.

This allows everybody to know the relative importance of each contribution as well as to understand why they were done and if it makes sense to add, remove or revert changes to suit local needs. *It converts the course in a living work*, not just a text cast in stone that will become an obsolete historic curiosity sooner than expected.

Allow provisions for recovering from major changes

Accordingly, you should keep in mind an important derivation of this principle: you want a history because you admit that a course is a living work subject to continual changes, driven by the authors' perceived needs at a given time. It is possible then that some changes may be too specific or time-dependent, or even that some decisions may prove to be plainly wrong (we all make mistakes and I'm no exception). Hence, *you should allow provisions for recovering from major changes*.

The simplest way to do so is to *make backups of the course* and annotate them accordingly (e.g. by renaming them with significant names). This is simple, but has some drawbacks, the main one being that backups are not carried over with the course when copying it, and hence previous materials may be lost when the course is imported to other institutions. Still, you should make and keep backups of your courses after every major change (e.g. adding new materials or activities), and download a copy of them to a safe place. Failure to do so, will result in much sorrow sooner or later.

Ideally we would like to use something like RCS or CVS to keep version control of a course material. We may just uncompress the backup (which is actually an XML file) and store it on *a version system* but this entails that all course contents are stored together, making it difficult to recover partial changes (like e.g. deleting a single lesson or a number of assignments). If at any point it becomes easier to make partial backups of course resources or sections, this might become an option of choice

An alternative solution is to *split courses* in meta-course sections and maintain each separately. This means that a complex course is built of smaller sub-courses that can be combined freely and re-ordered to make for the whole. For complex courses this is probably the best and most versatile solution, allowing as well others to use, substitute and combine those portions of your work they need. It simplifies also management of sections and their backups.

A compromise solution when changes do not apply to whole sections is to *hide course materials instead of deleting them*. In this scenario, instead of removing what you do not need or substituting it for your own version, you just add the new materials and hide the non-needed ones from view. This way, if you later discover you want to recover them, all you need to do is show them again and you are done. In due time, after various releases, if it becomes obvious you will never use them again (maybe because you have better new materials now) you can finally delete them once you feel strongly safe they are no longer needed.

Still, make backups, backups, backups, download them, annotate them and keep them around for emergencies and -most of all- history record. It may all start as a pet play project, but there is no guarantee others won't buy into it and make it into a wonderful ultimate work of art. You have a duty to History and your descendants to keep a record of changes so others can later study and learn how this wonder came out to be and into existence. Ever wondered how the pyramids were built? It is still puzzling us, and that's because we have not been able to find any record of the process. Mind you, your great-great-great- ... -grandson three thousand years from now may be a new Indiana Jones and wonder how your work could succeed. Be nice: leave them some hints.

Don't write, reuse

This is closely related to our next topic, quality. The best way to ensure quality and save work is to rely on proven works of excellent quality already produced by others.

In a world of ever increasing complexity no man(woman) is any longer an island: it is ever more difficult to become the humanist wise knowing everything. Do not misconstrue me: I'm not saying it is impossible, I pride myself in trying to, but -and I can tell from experience- it is ever more difficult. As a result, we rely every day more on work group and collaboration to create things.

Even if you are the final humanist genius, you can build nice works like the paintings of Picasso or Dalí, but you cannot go any further: no actor, film director, architect, music composer, etc.. can work alone. Even Beethoven in all his genius needed an orchestra to play his works, Gaudí needed legions of masons to construct his buildings, Spielberg needed executives, actors, cameramen, and a legion of people to film his movies, heck! even Caesar needed legions to build the Roman Empire!

You may have in mind the ultimate course, but building it will surely need the help of others. Announce your intentions, share your ideas with others, collaborate and start working towards your goal.

It should be obvious by now after all we have said: *sharing our work saves us efforts and allows us to reach greater quality.* Just do not forget all the previous advice given:

- look around for good materials to use (more on this later)
- make sure you are allowed to use those materials (either by fair use or because their licenses allow you to)
- always, always, always attribute works to their authors (you don't want your works steal, do you? then don't steal those of others)
- use the history page to keep track of all contributors and changes
- make backups, backups, backups, save a copy, annotate them and keep them around for emergencies and most important of all, for the record of History

So, how do you find good materials?

The first approach is to just look into the Internet: you surely have lots of bookmarks to nice materials you have found out there, and it is very easy to just make a cursory search on your preferred search engine (Google, Yahoo, Altavista, Lycos) to find the most popular ones.

This usually works. Just try to use content neutral search engines. Some engines are too tightly tied to some private company whose main interest is not the search market itself, and who will shamelessly give higher precedence without warning to hits that point to themselves, give them some advantage or in which they have a vested interest. These engines will make your life a lot more difficult. Run away from them like the pest. You know who they are.

As we have said, this usually works. It has one major drawback: *most materials you will find this way just happen to have only a copyright notice pointing to their author or even none at all* (in which case default copyright terms apply, i.e. you can not reuse those materials). In these cases you must *contact the author and ask for permission* to use their work and redistribute it. Whenever you do, please consider asking them to visit the Creative Commons web site or their local legal advisory team to select a license and state it clearly in the work so others in the future will know and save the hassle of contacting them over and over.

But there is a much better approach: Creative Commons and open licenses are now flourishing everywhere. It's nothing new: almost everybody who published their works on the Internet originally intended them to be openly and publicly available, only they didn't care to state it explicitly (it was assumed by default). In the mid-1990s many commercial-minded people started to join the Internet and to include restrictive licenses in their works. This confused things a lot and gave birth to all these new open documentation movements: there was now a need imposed by these money-avid newcomers to clarify usage terms.

If you use the first approach you will actually find that many materials were indeed intended to be openly used by their authors, only they didn't say so explicitly and later on never took the trouble to change every single page they wrote to affix a license. Still, getting in touch is a hassle, often compounded with the fact that the original address in the document has changed and the author needs to be tracked to his new location.

So, back to option two. We said there is a better way, thanks to the uprise of the new explicit open licensing movements, so which way may this be?

Easy: the major search engines include an advanced search option where you can state clearly the licensing terms you want associated with the pages returned. So, just go to Google, Yahoo, Altavista or any open document portal (like the Creative Commons, Wikipedia, MIT courseware or other web sites) *select the advanced search option and restrict your search to openly licensed available documents*. Now just navigate those hits and verify the license is indeed what you expect. You'll have saved a lot of work.

One additional word of advice: *do not look only at English pages* (or for the sake of it, pages in your target language). Many English works are linked to from non-English pages and reachable only from them (e.g. because the author wrote the material in English but linked to it from a local language page). Often you'll find excellent materials available directly only for students (i.e. you won't be able to add them to the course but may want to link to them). Do not be afraid of following links in Chinese, Russian, Arab, Egyptian, Greek, Spanish, whatever. You'll find lots of nice surprises. And *a trick navigating foreign pages*: often pages are named in English, so although contents may be unreadable, putting the cursor over links will show the english name of the page pointed to giving you an ad hoc translation of the link.

Finally, although this is not needed, think that you will probably want to give a certificate to students to prove they made your course. It is not a bad idea to *make the blueprint of the certificate an integral part of your course materials, and even may be each personalized certificate once issued*. You may store certificates hidden from user eyes so they can't access them, and you will find it handy when the time comes to issue the certificate. Not only after the course ends, but also if some years later a student comes asking for a copy. Furthermore, you will have saved the work for future instances of the course (a significant saving if you like to or must produce elaborate certificates with your institution logo, etc.).

Check your course quality

Last we come to the most relevant point. If you followed our advice until now you will already have lots of excellent materials to start from. Most probably you will not need to do anything at all besides glueing them together and adding a few minor odds and ends (forums, workshops, exercises). Your course will have popped out of nothing just by the magic of the Internet.

Now, if you add or start from scratch, you must pay close attention to quality. This is easier said than done today.

Today most materials are or have been written by people who learned in the last (20th) Century, and used to ancient learning methods. We are reaching to a Brave New World, and need to change our way to think and do things in order to adapt to this ever faster changing World.

Some advice is always safe: *keep scientific properness to a maximum*. Be precise, use language properly, be clear and concise. *Add a grain of salt* (humour). This has always and will always be good.

Beyond this, learn *to be attractive to students*: first, stay close to them (use but not abuse colloquial language, you are trying to make scientists, not -shrug- lay news(wo)men). Second, try to find the equilibrium point between a teacher's distance and a group member (you want them to join the group of elite scientists, your group), you should rather *behave like a welcoming group leader* integrating a newcomer than like the classic professor in a crystal tower. Third, work in group (you already do when using other people's materials) both with other teachers and with your students, and teach them to work in group (that is what they will be required to do in the Real World later on) and most important of all

get rid of your outdated teaching/learning prejudices

A good starting point is the moodle documentation, to which you can find a link in the EMBnet training web site. It is a lot less important to know the tool than to understand the education principles involved. Any good teacher knows that the tools he uses are irrelevant, it is the attitude that matters. Any good teacher knows that knowledge itself is irrelevant, it is the ability to evoke and use it that matters. But most important of all, any good worker and manager knows that it is not personal abilities that matter, but the efficiency to do quality work.

Personally, my experience has shown me that Life itself is a complex thing. It doesn't matter whether you know a lot or not, but whether you can get the job done. It doesn't matter whether you do the work yourself or not, but whether you can get the job done. Natural selection acts on results, not on how these are accomplished: to live in water an animal must adapt to it, it doesn't matter if it is a fish, a mammal or a turtle, each of them will find different solutions, but they all will live in water.

When I do evaluate students I emphasize two things: that they can and should cooperate to pass the exam, and that I don't care how they do it as long as they pass. This entangles two things: first, obviously, I need to design different exams for each student, but second and most important, I evaluate their ability to pass, not how they do. In real life they will be asked to solve problems, but nobody will ask them to know the academic way to solve problems: they may have a huge memory and remember all possible solutions, or be quick thinkers and find it on the fly, or good readers and look it up, or good at personal relations and get others' help, or even charming leaders and get others to do their work (what's a good manager after all?). Actually they will need to be conversant in all of these abilities. By forcing a single approach (the classical exam) I would be castrating them of all other abilities. By giving them freedom I expose them to all of these strategies, they will know they also work, will learn to use them by themselves, and will learn to impose limits on abuse, and -in one word- will be really prepared to live in the Real World.

Oh, but this is my personal advice, and as usual, your mileage may vary.

Choosing a license

To decide on a license is usually a hard issue. You should first consider the terms under which you want to release your works. There are many possibilities, from the most restrictive ones (no use or redistribution allowed, all rights reserved) to the most open ones (releasing it to the public domain). A good help on the most relevant options to consider can be found at the Creative Commons web site. In short, some of the **things you may want to consider** are:

- **Basic terms**
  - *Attribution*: do you want others to acknowledge your work? This will usually be the case as it matches standard academic culture and allows you to get recognition for your efforts.
  - *Derivative works*: do you want your work to be modifiable by others? Again, common academic culture favours this, as advancement of our knowledge relies on building upon previous work to reach further heights, but you may be concerned that changes might detract from the perceived quality of your work.
  - *Commercial use*: do you want to allow commercial use of your work? Traditionally, commercial companies had the role of exploiting the knowledge produced by academia. Nowadays this is discussed as academics are required to guarantee a direct return on their works and the public questions whether public funded results should be allowed to be appropriated by commercial companies without any investment or return.
  - *License propagation*: do you require your license to be respected and carried over to derivative works or copies? Again this will usually be the case, but a point can be made if someone adds significantly to your work but would like different licensing restrictions. On the other hand you may require the terms to be propagated and later negotiate additional agreements with interested parties.

- **Special terms**
  - *Jurisdiction*: most normally you will want a generic license with the widest application, but you must always remember that different Countries have different laws and some tuning may be needed.
  - *Public Domain*: this means you leave out any copyright claims and essentially offer your work to the public at large to do with it as they wish.
  - *Geopolitical conditions*: it is also possible to make a distinction on special issues so that different licenses will apply depending on them (e.g. allow broader freedoms to devel-

oping countries while requiring royalties from developed nations).
    - *Partial licensing*: you may want to allow use of parts of the work but not all of it (e.g. sampling of a movie or song, using no more than one chapter form a book...).
    - *Reduced copyright duration*: in today's quickly changing world you may want to retain copyright for a short time acknowledging that it will be obsolete and meaningless afterwards instead of a full legal term.
    - *GNU Copyleft terms*: these allow people to use your under stringent terms to ensure freedom (see the FSF or CC sites for more details).

- **Non-exclusion of rights**: whatever you decide, you must always keep in mind that you, as the author retain all rights to the work, which means you are always free to also release it under different licensing terms if you so wish. This means that you may for instance forbid commercial use by default, but if a company calls your door you may agree to issue them an additional license allowing commercial use under royalties.

Next you must **ensure that you are legally allowed to release your work** under those terms.

- *If it is all your own* (e.g. because you made it at home, on your own time, using your own resources, and without any contractual restriction with your employer), then you can do as you wish.

- *If you did the work under contractual restrictions*, like on paid time (e.g. by your employer, at work, or under a grant from a funding agency) or under contractual restrictions (e.g. under a non-disclosure agreement or under a restrictive employment contract) then you will need to check with your employer, funding entity or contractual stakeholders and verify with them it is OK to release your work under your chosen terms (or negotiate which terms are acceptable).

Finally, you must **draw the legal text** stating unambiguously the terms under which your work is licensed. This is a terribly difficult task (unless you opt for one of two options which we'll talk about later) requiring legal advice.

- If your work is entirely yours, then again you have two options:

    - *look for professional legal advice* (and pay for it out of your own pocket) to write the license for you (this only makes sense if you expect to make up for the costs by selling your work)
    - *use one of the many already existing licenses* that matches your desired terms

- If your work is bound to an organization, it is best to contact their legal department and ask them to either

    - *draft a license stating your terms* (actually leaving all the legalese to professionals) or
    - *sanction or approve an already existing license* you present them

In general, you will find that the best, easiest and more expedite course is to **look for an already existing license** which covers your terms and either use it or ask your legal department to approve of it. Your main responsibility in this case is to make sure that the license chosen is legally correctly written so it is enforceable and will actually defend your interests.

The Creative Commons Web Site

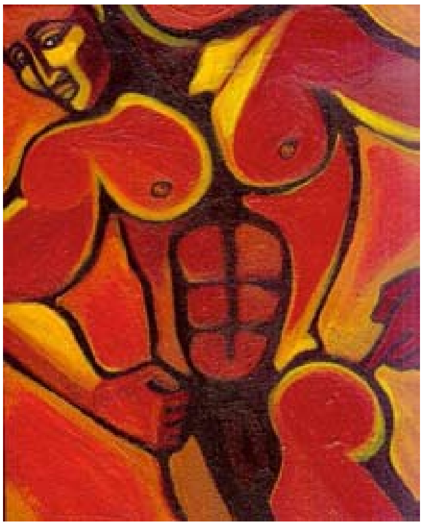
Your best starting point at this stage is the Creative Commons Web site. There you will find help to decide on the most common options and terms related to releasing works in the wild, as well as ready to use licenses written by professionals who have already taken the pain of investigating and drafting the licenses and ensuring they are legally binding and enforceable.

Using the CC site is easy as a breath: just select the terms under which you want to release your work, get the associated license and use it (or produce it to your legal department for approval). In the CC site you will find as well instructions on how to associate the license to your works, short descriptions of the licenses and full legal text of the same, neat logos to allow for easy license identification and links to the descriptions and text.

The tenuous nature of sex

Vivienne Baillie Gerritsen

Everyone knows how to tell the difference between a boy and a girl. The exterior signals are obvious. And yet, despite such a clear statement on Nature's behalf, the molecular pathways underlying our being either male or female are subtle and fragile. It takes very little to make a woman out of a man – at least as far as our chromosome makeup is involved. We were told that boys are XY, and girls XX. But it's not so simple. Some girls are XY, and some boys are XX… So there must be something sophisticated involved. And we are only beginning to discover what. Because of its singular architecture, the male Y chromosome is distinctive under the microscope and it was not long before 19th century scientists caught on that it had a major role in the making of a man. A closer look at it led molecular biologists to a specific region on the Y chromosome and, in the 1990s, scientists announced the discovery of a protein – the Sex-determining region Y protein (Sry) – that had a major role in convincing a foetus to become a baby boy.



Red Nude Male, acrylic

DanceGallery @New Millennia Studio

How an embryo becomes male or female has puzzled many a human for thousands of years. While menstrual blood was believed to be the material with which semen moulded the beginnings of a foetus, it was thought that heat favoured the development of little boys and cold shaped an embryo into a little girl. Semen that came from the right testicle gave boys; the left testicle harboured the wherewithal to make girls. Likewise, embryos that were positioned to the right of the womb became male, and those to the left, female. Once the notion of two germ cells and their meeting was acknowledged, the Ovists believed that a child was held within an egg and just needed sperm to trigger off its development, while the Spermists speculated that it was sperm that sheltered the embryo, and the egg was there to feed it. As time passed and techniques improved, scientists discovered not only chromosomes but also the fact that specific chromosomes defined sex. The X chromosome was the first to be discovered because of its size. Y was discovered shortly after.

The fate of the Y chromosome is bewildering. It all started 300 million years ago when Nature thought up sexual reproduction. To reproduce in this way, you need at least two entities of an opposite sex. It was a great way to promote biological diversity and a basis on which natural selection could work, but it meant compromising on one chromosome, i.e. the Y chromosome as we know it today. 300 million years ago, a large portion of an X chromosome was inverted. As a consequence, it was unable to pair with its sister chromosome. Left to fend for itself, it fast became prey to mutations and over the years it has lost hundreds of genes, gained a lot of rubbish and become very small. So much so that some scientists think that the Y chromosome may just wither away…and hence the male of the species? 'Nonsense!' say fellow researchers, other animals have lost their Y chromosome already and have simply found another way of promoting the male program. There is no reason why *Homo sapiens* wouldn't do the same, and some of the future mutations on Y may even support its survival.

So as long as XY embryos carry Sry, their destiny is a male one. What does Sry do? It triggers off pathways that promote the development of masculine features which depend upon two types of cell: the Sertoli and the Leydig cells. Sertoli cells give rise to the testes and Leydig cells will ensure the production of androgens - the male hormones. How does Sry do this? Sry is a transcriptional factor and binds to the minor grooves of DNA. There – by way of bending the nucleotide structure at angles of 60 to 80° – it either activates or represses the transcription of genes. To date, no one is sure whether Sry activates 'male' genes or whether it represses 'female' genes – though there is growing evidence for the former. Naturally, Sry does not act on its own; without the help of molecules both upstream and downstream of Sry expression, there would be no baby boy at all – or one whose male features are hugely hindered.

Up until the 7th week of pregnancy, a human embryo's phenotype is neither male nor female. If the embryo is XY, Sry is expressed – though only for a short time. In mice, for instance, Sry is active from the 10th day of development until the 12th day. And if it misses the bus, the embryo is heading for serious trouble. In fact, problems related to a person's sex can be the cause of a dysfunctional Sry, either because it has not been expressed, because it is inactive or because it wasn't expressed on time. Besides its role in sex differentiation, Sry is also found in male brains. Could it be involved in what could be described as 'male' behaviour? Possibly. It seems to have a role in the production of dopamine which is involved in motor control. The obvious question then is: are women handicapped in this respect? Of course not. In fact, the female hormone oestrogen also has a role in dopamine production and could replace that of Sry...

It takes a network of molecules to become a boy or a girl. Sex is not the work of only one protein though Sry is certainly sitting on a hinge. Integrated into an XX mouse embryo, it can switch the sex program and a female mouse becomes…male. Mice are not human though. Furthermore, the common belief that girls only become so because Sry is not expressed, is not accurate. Women are not passive in their making. One of Sry's doings is to activate the production of a hormone known as AMH which ensures that any feminine features are impeded.

To cut a long story short, the chromosomal definition of a man is XY, and that of a woman XX. But it takes very little to confuse the nature of one or the other by providing our species with humans that carry both feminine and masculine attributes for example, or by 'turning' what 'should' have been males into females, or females into males. There are XY females where Sry is not expressed and XX males where Sry is expressed thanks to the presence of its gene on an X chromosome. And there are hermaphrodites whose phenotype is neither male nor female, but a bit of both. The making of a sex is not a trivial affair and we should bear in mind that it takes very little to tip the scales and confuse the program. It is a thin line between a man and a woman.

Cross-references to Swiss-Prot

Sex-determining region Y protein, *Homo sapiens* (Human): Q05066
Sex-determining region Y protein, *Mus musculus* (Mouse): Q05738

References

1. Polanco J.C., Koopman P.
   Sry and the hesitant beginnings of male development
   Dev. Biol. 302:13-24(2007)
   PMID: 16996051

2. Mittwoch U.
   Three thousand years of questioning sex determination
   Cytogen. Cell Genet. 91:186-191(2000)
   PMID: 11173854

3. Fox D.
   The descent of man
   New Scientist magazine, Issue 2357, August 2002

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

Chile
Juan A. Asenjo
Centre for Biochemical Engineering and Biotechnology (CIByB). University of Chile
Beauchef 861, Santiago, Chile
Tel: +56 2 6715140
Fax: +56 2 6991084
Email: juasenjo@ing.uchile.cl
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.cn.embnet.org

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://www.co.embnet.org

Costa Rica
Allan Orozco
University of Costa Rica (UCR), School of Medicine,
Department of Pharmacology and ClinicToxicology
San Jose, America Central
Costa Rica
Email: allanorozco@gmail.com
Tel: +506 2074489
http://www.dftc.ucr.ac.cr/

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
Kimmo Mattila
CSC, Espoo
Email: kimmo.mattila@csc.fi
Tel: +358 9 4572708
Fax: +358 9 4572302
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

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
Akash Ranjan
Laboratory of Computational Biology & Bioinformatics facility, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad
Email: akash@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

Pakistan

Raheel Qamar
Department of Biosciences, COMSATS Institute of Information Technology, Park Road, Chak Shahzaad Campus, Chak Shahzaad
Islamabad, Pakistan
Email: Raheel _ qamar@comsats.edu.pk
Tel: +0092-333-5119494
http://www.ciit.edu.pk/Departments _ & _ Faculties/Link=DeptDetail&f=Departments%5F%26%5FFaculties&SMID=10

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: +315 214407970
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 2 5930 7413
http://www.sk.embnet.org

South Africa

Heikki Lehvaslaiho
SANBI, University of the Western Cape, Bellville
Email: heikki@sanbi.ac.za
Tel: +27 (0)21 959 2096
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, Génopode-UNIL, CH-1015
Lausanne Email: Laurent.Falquet@isb-sib.ch
Tel: +4121 692 4078 Fax: +4121 692 4065
http://www.ch.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

IHCP

William Moens
Institute of Health and Consumer Protection
Via E. Fermi 1 - 21020 Ispra (Varese), Italy
Email: william.moens@jrc.it
Phone: +390332786481
http://ihcp.jrc.cec.eu.int/

ILRI/BECA

Etienne deVilliers
International Livestock Research Institute
PO Box 30709, Nairobi 00100, Kenya
Email: e.villiers@cgiar.org
Phone: +254 20 4223000
www.becabioinfo.org

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

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. You can get them from the EMBnet organization Web site:

`http://www.embnet.org/download/embnetnews`

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

Submission deadline for the next issue:

May 20, 2007

EMBnet.news is an official publication of the EMBnet organisation
www.embnet.org