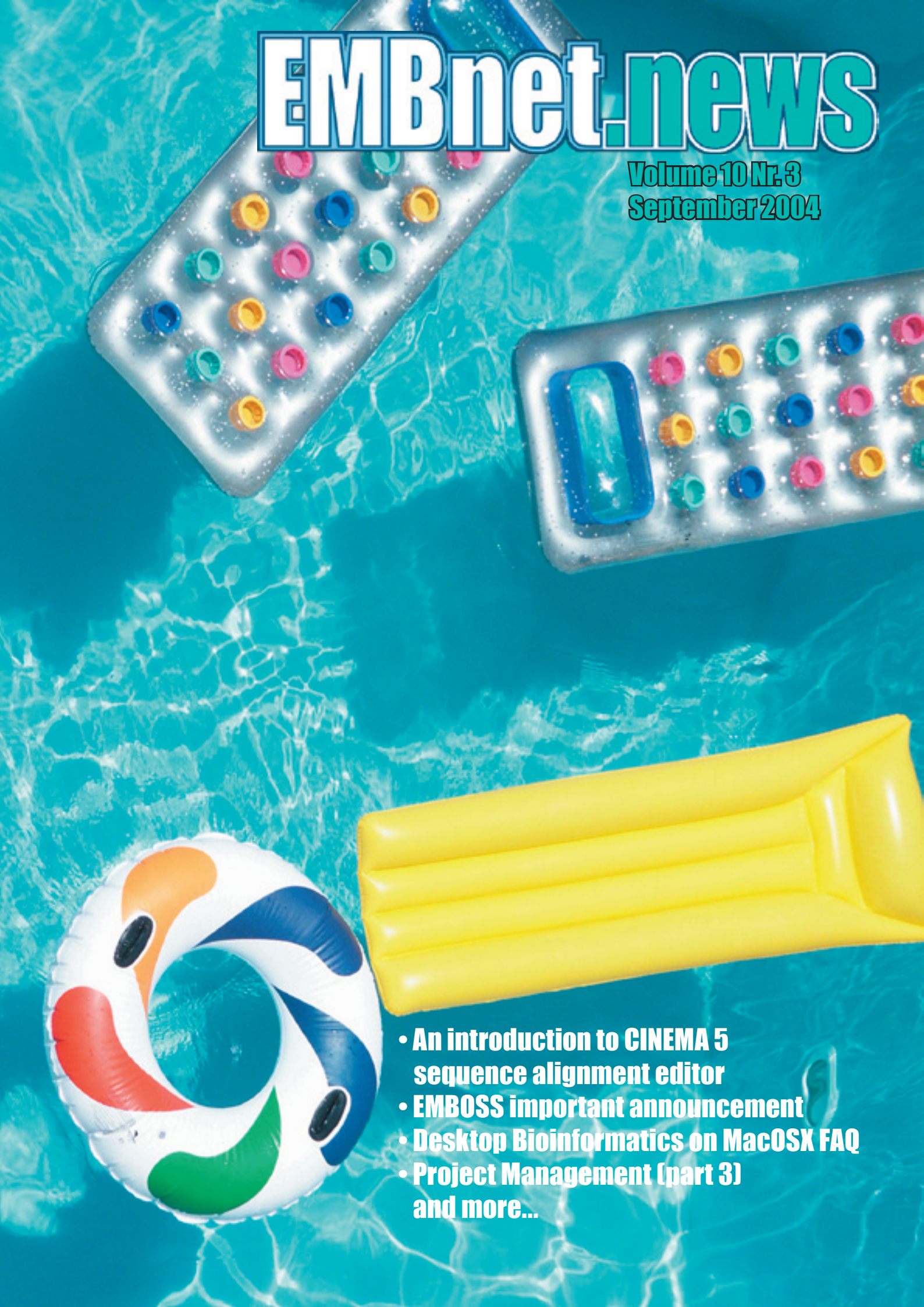


# EMBnet.news

Volume 10 Nr. 3  
September 2004

- 
- An aerial photograph of a swimming pool with clear blue water. In the upper left, a rectangular inflatable mat with a grid of colorful circular holes (pink, yellow, blue, green) is partially submerged. To its right, another similar mat is floating, featuring a blue rectangular inflatable ring in one of its holes. In the lower left, a colorful inflatable ring with orange, blue, red, and green sections is visible. In the lower right, a long, yellow inflatable lounge chair is floating. The water's surface is rippled with light reflections.
- **An introduction to CINEMA 5 sequence alignment editor**
  - **EMBOSS important announcement**
  - **Desktop Bioinformatics on MacOSX FAQ**
  - **Project Management (part 3) and more...**

## Editorial

The NHGRI (National Human Genome Research Institute) announced in early August that they have started to sequence no fewer than 18 new organisms!

Their strategic mix includes 9 mammals among which are orangutan, elephant, hedgehog, guinea pig, cat, and rabbit, as well as 9 non-mammals: a slime mold (*Physarum polycephalum*), a ciliate (*Oxytricha trifallax*), a choanoflagellate (*Monosiga ovata*), a placozoan (*Trichoplax adhaerens*), a cnidarian (*Hydra magnipapillata*), a snail (*Biomphalaria glabrata*), two roundworms (*Pristionchus pacificus*, *Trichinella spiralis*) and the lamprey (*Petromyzon marinus*).

This enormous amount of sequence data will help researchers to gain new insight into the organisation of genomes. But as a consequence this will increase the need for better bioinformatics tools and databases to analyse and store the data. Can EMBnet help the biomedical researchers to benefit from these data? Of course!!

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



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

We provide the EMBnet community with a printed version of issues 49 and 50. Please let us know if you like this inclusion.

Cover picture: Pool, summer 2004 [© Erik Bongcam-Rudloff]

## Contents

Editorial	2
An Introduction to the CINEMA5 sequence alignment editor	3
EMBOSS: announcement	9
Desktop Bioinformatics on MacOSX FAQ	10
Project Management (Part 3)	14
Protein spotlight 49&50	22
Nodes information	26

### Editorial Board:

Erik Bongcam-Rudloff, The Linnaeus Centre for Bioinformatics, SLU/UU. SE  
 Email: [erik.bongcam@bmc.uu.se](mailto:erik.bongcam@bmc.uu.se)  
 Tel: +46-18-4716696  
 Fax: +46-18-4714525

Laurent Falquet, SIB, Lausanne. CH  
 Email: [laurent.falquet@isb-sib.ch](mailto:laurent.falquet@isb-sib.ch)  
 Tel: +41-216925954  
 Fax: +41-216925945

Pedro Fernandes, Instituto Gulbenkian. PT  
 Email: [pfern@igc.gulbenkian.pt](mailto:pfern@igc.gulbenkian.pt)  
 Tel: +315-214407912  
 Fax: +315-214407970

Oscar Grau, IBBM. AR  
 Email: [grau@biol.unlp.edu.ar](mailto:grau@biol.unlp.edu.ar)  
 Tel: +54-221-4259223  
 Fax: +54-221-4259223

Gonçalo Guimaraes Pereira, UNICAMP. BR  
 Email: [goncalo@unicamp.br](mailto:goncalo@unicamp.br)  
 Tel: +55-19-37886237/6238  
 Fax: +55-19-37886235

### Publisher:

EMBNET Administration Office  
 c/o Jack Leunissen  
 Dept. of Genome Informatics  
 Wageningen UR, Dreijenlaan 3  
 6703 HA Wageningen, The Netherlands  
 Email: [Jack.Leunissen@wur.nl](mailto:Jack.Leunissen@wur.nl)  
 Tel: +31-317-484074



## An Introduction to the CINEMA 5 sequence alignment editor



James Sinnott<sup>1</sup>, Steve Pettifer<sup>1</sup> & Terri Attwood<sup>1,2</sup>

<sup>1</sup>Department of Computer Science & <sup>2</sup>School of Biological Sciences  
University of Manchester  
Oxford Road  
Manchester M13 9PL, UK

### Introduction: a UTOPIAn sequence editor

UTOPIA (User-friendly Toolkits for OPERating Informatics Applications) is a project that aims to develop an easy-to-use bioinformatics analysis environment, based on a workbench of re-usable widgets that can be combined to form tools that address a particular scientist's needs. A key philosophy of UTOPIA is the construction of applications that are genuinely user-friendly and the project has drawn together specialists from the field of bioinformatics and computer science to concentrate on the issues of human computer interaction in today's bioinformatics tools and data sources. A bioinformatician should be able to launch a UTOPIA application and immediately begin working in a way that is intuitive, familiar and predictable.

The first application built with the UTOPIA widgets is CINEMA 5, an interactive tool for the visualisation, construction and manipulation of multiple protein and DNA sequence alignments. CINEMA 5 is the most recent development of the CINEMA (Colour INteractive Editor for Multiple Alignments) alignment editor, which has existed in several different forms over the past few years, including its original incarnation (Parry-Smith, et al, 1998) and CINEMA-MX (Lord et al, 2002). CINEMA 5 aims to provide a highly intuitive, flexible and full-featured interface, along with several features that are unique to multiple sequence alignment editors. Consistent with the UTOPIA design philosophy

of effortless interaction, CINEMA 5 supports the familiar 'point and click/drag and drop' interaction metaphor of contemporary office and design applications, and is optimised for the efficient rendering of extremely large sequences.

This article provides an overview of the most important aspects of CINEMA 5, including novel features such as breakpoints and search-guided alignment; some practical pointers are also supplied where appropriate. Extensive online help documentation is provided with CINEMA 5 and should be consulted for full details of how to use the application. CINEMA 5 is freely available for Windows, Linux and OS X systems from <http://aig.cs.man.ac.uk/utopia>.

### Opening, saving and importing sequences.

CINEMA 5 recognises a number of different sequence file formats, including FastA, NBRF-PIR & PHYLIP (reading/saving), and GCG-MSF, CLUSTAL & NEXUS (reading only). The native CINEMA 5 sequence alignment format can also be used, which stores CINEMA-specific information (such as motifs and breakpoints) along with the sequence data, and is recommended for saving alignments between sessions. A standard folder, My Sequences, is provided as a suggested repository for locally stored sequences and alignments. This folder is located in My Documents on Windows systems, and in the top-level directory of your user account on Linux and Apple OS X systems. Sequences are easily imported into an open alignment, and can either be appended to the end of the alignment or inserted into it at a specific point.

### Visualisation and navigation of sequence alignments

Flexibility of visualisation is a key aspect of CINEMA 5. Alignments can be displayed at

a range of different resolutions, from close-up for detailed editing, to a single-pixel-per-residue overview; the appropriate level of detail is easily set via an intuitive mouse-controlled zoom tool. Multiple views can be opened onto a single alignment, allowing different regions to be visualised and edited simultaneously, at different resolutions if required. Each view has a colour scheme associated with it, which dictates how the amino acids are coloured. Default schemes are provided for both protein and DNA sequences, and a colour scheme editor allows customisation of the scheme for a particular view. When there is more than one view of the same alignment, all views except that currently active are dimmed slightly, and only regions of the alignment coinciding with the active view are shown at full intensity. Figure 1 illustrates a number of these visualisation features in CINEMA 5. In addition to standard scrollbars, an alignment can be navigated using the cursor keys and

the home/end/page up/page down keys in a manner that will be familiar to users of word processing software. Support for the mouse wheel is also provided and, if present, this control can be used to scroll through an alignment or, with the shift key held down, to zoom the view in or out.

## Basic editing of sequence alignments

CINEMA 5 has three main interaction modes for working with a sequence alignment: insert/remove gaps, slide sequences & select regions/edit motifs. The active interaction mode is chosen by selecting the appropriate button from the main toolbar. Basic alignment editing operations provided include mouse-controlled insertion or removal of gaps and the sliding of entire sequences left or right, relative to the rest of the alignment. Simply

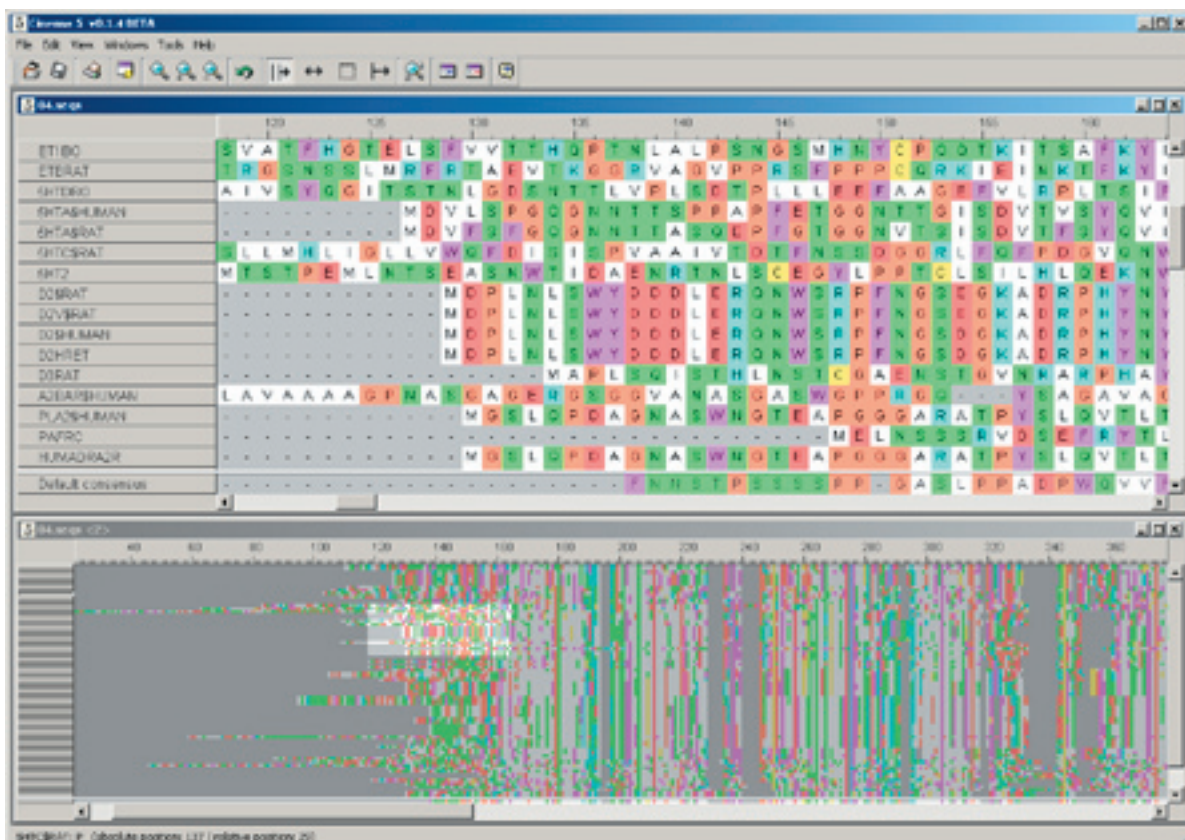


Figure 1. Visualisation in CINEMA 5: two views of the same alignment; the bottom view provides an overview while the top view concentrates on a particular region. The highlighted area in the bottom view illustrates the location of the region displayed in the top.

select the appropriate interaction mode from the toolbar and click-and-drag a sequence with the left mouse button to begin editing. Most other operations that can be performed on a sequence alignment are accessed via the context menu, which appears when the right hand mouse button is clicked in an alignment window.

## Sequence Groups

Sequence groups provide a convenient mechanism for simplifying the task of editing an alignment. By combining two or more sequences together as a group, editing operations can be applied to all the sequences simultaneously. A temporary group of sequences is created by holding down the shift and/or control keys, and then selecting the desired group of sequences with the left mouse button. Once selected, an insert/remove gaps or sliding operation can be performed on the entire group at once by interacting with one of the grouped sequences. If a non-grouped sequence is selected at any point then the group is disbanded. Persistent sequence groups can be created to simplify repeated group-based operations. After selecting the desired sequences a persistent group is formed by

right-clicking the mouse over one of them and choosing **Group selected sequences** from the menu. A coloured block in the label of each sequence contained in the group illustrates the new grouping and enables easy identification of its members. Right-clicking on a grouped sequence provides menu options for either ungrouping that sequence or disbanding the entire group. A sequence grouping can be temporarily overridden during editing by using the middle mouse button, rather than the left. Existing sequence groups can also be combined with other groups and/or individual sequences, to form hierarchical sequence groups. The *Sequence Group Manager* tool provides greater control over the creation and organisation of hierarchical sequence groups; see the online help for full details.

## Breakpoints

A *breakpoint* is a vertical divider that is inserted along a 'column' of residues in the alignment, logically splitting it into sub-alignments at that point. During subsequent editing, the sub-alignments either side of the breakpoint are treated as separate entities. Breakpoints can thus be used to preserve specific regions of an alignment whilst

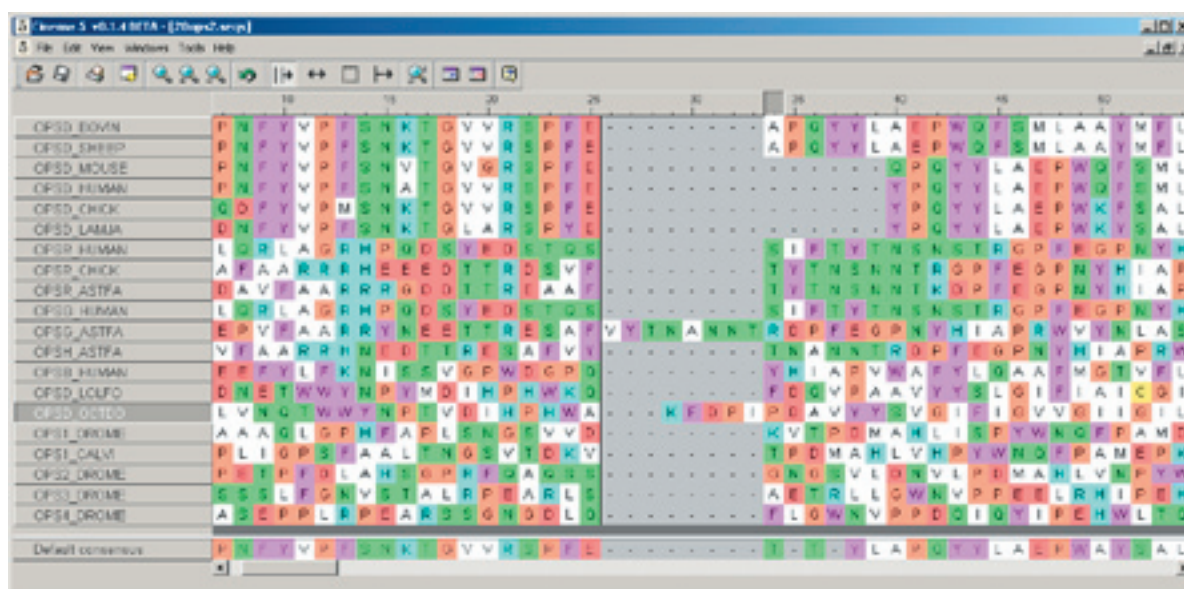


Figure 2. A breakpoint (the vertical black bar) has been inserted into this alignment, preserving the region to the left whilst editing continues on the right hand side.



editing others (see figure 2). For example, if a region of an alignment has been successfully aligned, a breakpoint may be inserted at one, or both sides of it; subsequent changes to other areas of the alignment will not affect the region contained within the breakpoints, which has effectively been 'fenced off'. To insert a breakpoint into an alignment, position the mouse pointer over the column at which the breakpoint is to be added, right-click and choose **Insert breakpoint** from the context menu. Existing breakpoints may be removed at any time, with the two sub-alignments either side of the breakpoint reunited. To remove a breakpoint, right-click it and select **Remove breakpoint** from the context menu.

## Moving and Deleting sequences

A sequence can be repositioned within an alignment by selecting its label with the left button and dragging the mouse up or down. A sequence may be deleted from an alignment by right-clicking on the sequence

and choosing **Remove selected sequence** from the menu. Multiple sequences can be deleted by first creating a temporary group of the sequences to be removed. It is also possible to remove one or more entire "columns" of residues from an alignment, and this is achieved by right-clicking a column in the numerical key at the top of the alignment and selecting **Delete selected column** from the menu. Multiple columns can be removed by first group-selecting them in the same manner as for sequences.

## Printing alignments & Generating images

An alignment can be printed out to any printer attached to your system. Alignments may be printed in colour, and on any paper format your printer supports; you can specify the output size of the alignment in terms of pixels or millimetres. If the alignment is too large to fit the chosen paper size then it can either be wrapped vertically down the page,

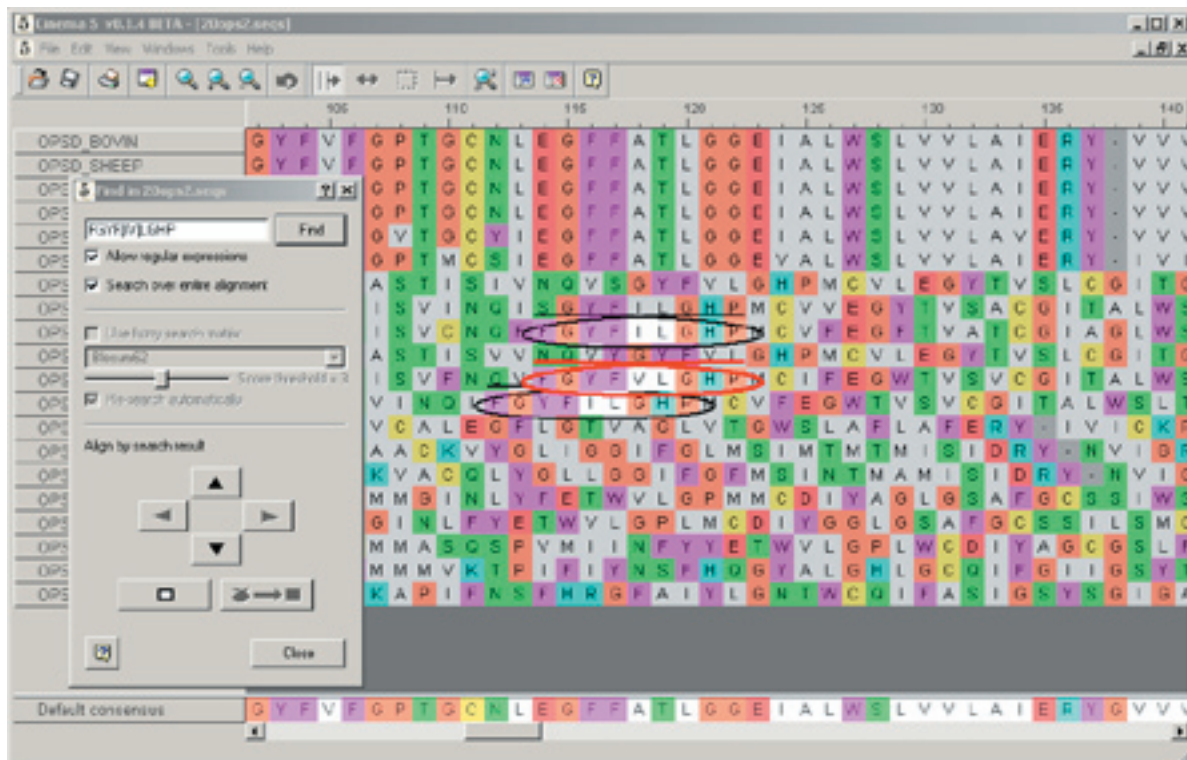


Figure 3. Searching for sub-sequences within an alignment. In this example, a simple regular expression has been used to broaden the search.

or split up into a set of page-sized tiles, which can be joined together to recreate the entire alignment. An alignment can also be saved as a bitmap image file and a number of standard image file formats are supported: bmp, xpm, xbm, png; the size of the image to be saved can be specified in terms of pixels or millimetres.

## Searching for sub-sequences

CINEMA 5 includes a powerful and flexible search facility, enabling the user to quickly locate instances of particular sub-sequences within a sequence or alignment. Pressing **Ctrl+F** calls up the search dialog, from which a variety of different searches may be performed (see figure 3). After entering an appropriate search string, clicking on the Find button will locate all exact matches within the alignment. It is also possible, after selecting the appropriate check box, to broaden the search using regular expressions. More general 'fuzzy' searches may be performed with substitution matrices, and a number of standard *Blosum* and *Pam* matrices are provided. The results of a fuzzy search can be refined by using a slider to alter the threshold used to accept/reject potential matches. Once a search has been performed the

results are navigated using the four arrow buttons, in the lower half of the dialog, which allow for fast location of a particular sub-sequence. The current search result is highlighted with a red outline. By default searches are made over all sequences in the alignment, although they can be restricted to specific sequences by first selecting those sequences in the alignment and then unchecking the **Search over entire alignment** box.

## Search result guided alignment

We present a novel facility for using the results of a sub-sequence search to help guide the sequence alignment process: pairs of sequences can be progressively aligned according to the relative positions of specific search results. To begin, one search result is *anchored* by navigating to the result and then pressing the **Anchor result** button at the bottom left of the dialog; its sequence will remain in a fixed position throughout the process. A search result from another sequence can now be aligned to the anchor by navigating to the result and clicking the **Align result to anchor** button at the bottom right of the dialog; this sequence will slide left

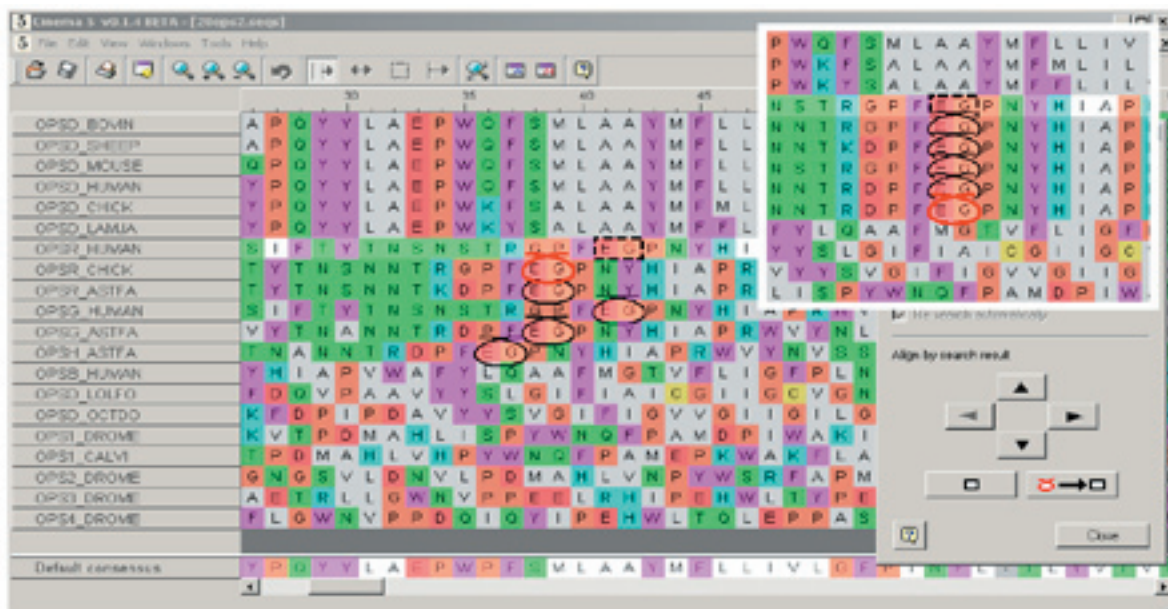


Figure 4. Using search results to help guide the alignment process. The main image shows a set of search results, with the first result anchored (shown by a dashed outline). The inset image shows the state of the alignment after the other search results have been automatically aligned to the anchor.

or right until the result and anchor are aligned. Search results found on further sequences can also be aligned to the anchor in the same way. It is possible to change which result is anchored at any point by pressing the **Anchor result** button when the appropriate result is active; changing the anchor will only affect subsequent actions. Figure 4 illustrates the search result guided alignment process.

## Motifs & Fingerprints

Extensive tools are supplied for defining, compiling and managing motifs within a multiple sequence alignment. Motifs are grouped together into *fingerprints*, representing either the main family of the alignment or any appropriate sub-families. A motif is defined by selecting a rectangular region of the alignment and instructing CINEMA 5 to convert this region into a motif. To select a region, switch to select regions/edit motifs mode and then, with the left mouse button held down, drag a box around the region of interest. The box has a dashed outline to indicate that it represents a temporarily selected region. Once selected, the region can be converted into a motif and added to a fingerprint. To add a motif to the main family fingerprint, right-click within the selected region and choose **Convert to motif and add to main family fingerprint** from the context menu. If the selected region covers just a sub-set of the sequences in the alignment then it may, as an alternative, be added to a *sub-family* fingerprint for those sequences; this option will also appear on the context menu if applicable. Once converted, the outline of the region will become solid, indicating that it is now a motif. To simplify the definition of motifs belonging to the main family fingerprint, holding down the **shift** key whilst selecting a region will force it to span the entire 'height' of the alignment.

## Working with motifs & fingerprints

A motif can be saved, either as a standard motif file, a sequence alignment file or a bitmap image, or it can be printed out. It

is also possible to perform operations on a fingerprint of motifs collectively, e.g. saving all motifs within the fingerprint at once. These operations are all accessible by right-clicking within a motif region and choosing either the **Motif...** or **Fingerprint...** submenus from the context menu. When a motif is first defined it is automatically given a name, derived from the name of the fingerprint it belongs to and its position in the alignment, from left to right, relative to the other motifs in the fingerprint. Fingerprints are automatically named after either the alignment they belong to (in the case of the main family fingerprint) or the set of sequences they cover (for sub family fingerprints). A fingerprint can be renamed by selecting **Rename Fingerprint...** from the **Fingerprint...** submenu. All of the motifs defined within the fingerprint will automatically be renamed to reflect this change.

## The Motif Manager

The Motif Manager is a tool for managing and working with motifs and fingerprints. Each fingerprint defined for an alignment has a folder in the Motif Manager and each motif has a corresponding entry in the appropriate fingerprint folder. This provides a convenient virtual repository for the motifs defined within a particular alignment. One of the most useful features of the Motif Manager is the ability to quickly locate a specific motif within the alignment. By right-clicking the motif's entry in the Motif Manager list and choosing **Show motif in view** from the menu, the currently active view of the alignment will be scrolled so that the selected motif is visible. You can also use the Motif Manager to save, print, rename and delete motifs and fingerprints.

## Future developments

CINEMA 5 is the subject of constant development, while at the same time the UTOPIA toolkit is expanding to provide further applications, most notably Ambrosia, a 3D molecular visualisation tool. Future versions of CINEMA 5 will provide seamless interaction



with Ambrosia so that conserved regions of the alignment marked as motifs can be automatically highlighted in the 3D protein structure. Conversely, it will be possible to highlight structurally important residues using Ambrosia and have them pinpointed within the protein sequence shown in CINEMA 5. Work on seamless access to remote services such as databases and 'grid enabled' resources is also under way.

UTOPIA is funded by EMBNet, e-Science North West, the UK Department of Trade and Industry and the EPSRC.

## References

Parry-Smith, DJ, Payne, AWR, Michie, AD and Attwood, TK (1998). CINEMA – A novel Colour INteractive Editor for Multiple Alignments. *Gene*, 221, GC57-63.

Lord, PW, Selley, JN and Attwood, TK (2002). CINEMA-MX: a modular multiple alignment editor. *Bioinformatics*, 18(10), 1402-1403.

## Important announcement: EMBOSS and Jemboss in danger!



## Closure of EMBOSS development site

The UK Medical Research Council is to close the Research and Bioinformatics Divisions of the RFCGR (the current home of EMBOSS) in the middle of 2005. The MRC Press Office has stated:

«All MRC can say at this stage is that Council have made a decision to close the Research and Bioinformatics Divisions. However, the Director has been asked to draw up a closing down plan for consideration by Council in July.»

This action will more than halve the current core development team and will therefore adversely affect the development and support of EMBOSS. We hope that alternative sources of funding can be found. In the mean time all the EMBnet nodes urge you to support the EMBOSS development team by either filling in the survey or by sending a letter of support or both.

## Survey

**It would help the EMBOSS project tremendously if you complete this survey:**

[http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/emboss\\_survey.html](http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/emboss_survey.html)

This will help us to both gain support for EMBOSS and to steer the future development of the project down lines suggested by you.

## Letters of support

**If you feel that EMBOSS is valuable, please send your letters of support to:**

EMBOSS User Survey  
MRC Rosalind Franklin Centre for Genomics Research  
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SB, UK

or by E-mail to: [survey@rfcgr.mrc.ac.uk](mailto:survey@rfcgr.mrc.ac.uk)

We will keep you informed of further developments.

## Desktop Bioinformatics on MacOSX: FAQ



**Erik Bongcam-Rudloff**  
Assistant Professor  
Swedish University of  
Agricultural Sciences  
The Linnaeus Centre for  
Bioinformatics, Uppsala

In our previous issue of EMBnet.news we described how to install and connect embossRUNNER to our installation of EMBOSS. After writing our previous article we announced our package on Apple's MacOSX download pages ([http://www.apple.com/downloads/macosx/math\\_science/](http://www.apple.com/downloads/macosx/math_science/)). This became a big success and the package was the third most popular download during the month of July. It also resulted in an enormous boost in the download of EMBnet.news issues.

This increased number of users meant also an increase of E-mails asking for support. In this article I compiled the most common questions and the answers we gave.

### Questions and answers

#### 1- Do you have a list of what ebiotools installs?

Yes, here is a list:

clustalw1.83.pkg  
hmmmer\_altivec2.2g.pkg  
shlibs.0.5.pkg libraries needed by some programs e.g.. libpng  
libjpeg etc  
kaptain-EMBOSS.0.95.pkg  
staden-1.4.pkg  
kaptain.0.71.pkg  
t\_coffee.1.37.pkg  
emboss-db.2.0.pkg (= prosite, cutg, transfac, aaindex, prints, rebase. etc.)  
ncbi-tools.pkg  
wwwblast.pkg  
emboss-2.8.0.pkg  
nedit-5.3.pkg  
primer3.0.9.pkg  
glimmer2.pkg

You can see a list and customize your installation clicking on the "Customize" button. See figures 1 and 2:



Figure 1. To see the list of installed packages click on Customize.

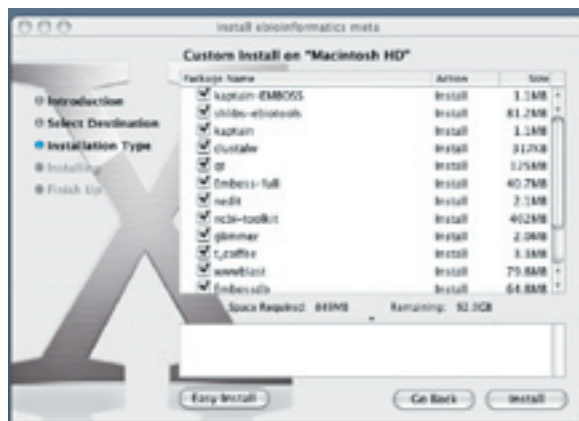


Figure 2. Shows the list of packages that eBiTools installs. We recommend you install most of the packages unless you know what you are doing. eBiTools uses 790 MB of hard disk space.

#### 2- I installed eBiTools but I do not see the applications.

The eBiTools package installs in /usr/ebiTools. Most of the programs installs in /usr/ebiTools/bin. Staden in /usr/ebiTools/staden/macosx-bin/, Blast for command line in /usr/ebiTools/blast/bin, hmmmer\_altivec in /usr/ebiTools/hmmmer\_altivec/bin/ and Web-blast in /Library/WebServer/Documents/blast/

#### 3- I downloaded eBiTools, but can't see how to get it up and running.

eBiTools is not a program, but a package containing more than 120 programs.

**4- I wanted to run the Staden package. I managed to open the "terminal" application, and typed "pregap4", but I obtained several error messages.**

You must have X11 installed to run Spin, pregap4 etc. See "Practical tips" for details.

**5- I downloaded and installed eBiotools, as well as the latest issues of EMBnet.news, but I am confused with all the manual editing mentioned on the first issues. You also have different packages for download.**

Please download only the latest release of eBiotools (one for the G4 and one for the G5). The first releases are on the download site only as reference. Read this article and the article in EMBnet.news 10.2 for embossRUNNER. For the advanced user only: if you want to activate and use the Web-based blast installed in /Library/WebServer/Documents/blast, please read EMBnet.news 9.2.

**6- I would like to use blast using our own sequences. I am not a Unix person therefore I am very interested in running blast through a web-interface**

In MacOSX you have a web-server built in (Apache). Using Apache and the instructions in Embnet.news 9.2 you can also connect your own blast-databases and use the blast web-pages installed by eBiotools. Please try starting the web-server from

System Preferences -->Sharing -->Web

Then open a web-browser and write  
<http://localhost/blast>

You should see a blast-server installed on your machine. Now you can activate it and connect your sequence database(s) to it (follow the step-wise instructions in EMBnet.news 9.2).

**7- I have been using these tools on OSX 10.2 for some time now... I have a new installation of 10.3.3 on a G4 powerbook and I'm a little confused by the migration to the bash shell. I edited the .cshrc file to reflect the move to /usr/ebiotools/ but this file does not seem to be recognized by bash.**

The eBiotools installation also automatically

fixes all manual editing to get the correct Paths and Environment values. But sometimes this does not work on some systems.

If you have a new installation of 10.3.x bash is the default shell. If you upgraded from 10.2.x to 10.3.x you still have tcsh as the default shell. This makes all new "Terminal" users very confused. To avoid any problems related to shells please see "Practical tips". If you want to change your default shell you will find instructions here:

<http://www.macosxhints.com/article.php?story=20020213084259198>

**8- A user wrote to us: it appears that the path to the man pages on OSX gets broken by sourcing the file staden.profile. I found it could be fixed by adding /usr/share/man to the definition of MANPATH in staden.profile.**

Thanks Jeff!

**9- For stand alone blast set up, I am following your article in EMBnet.news 9.2. I am already having difficulties in the first step, the pico .ncbirc is not saved. The "save" comes with an error message "Cannot open file for writing".**

Please forget the previous packages and use the latest one. The new packages fixes many things discovered by the users of eBiotools. But the problem that affected this user is that he tried to save files in a directory where only root has write privileges. To be able to save changes on files owned by root you must belong to the administrator group and then start the command line by sudo. E.g.

```
%sudo pico .ncbirc
%password:
```

Then it will be possible to save the new edited file.

### Practical tips:

Many readers have concerns about using the command line tools. It is difficult to remember names paths etc. eBiotools install several programs with Graphical User Interfaces (GUI's). But they must be started from the command line. One nice way of avoiding this is to make a list of programs on the X11 Applications list. You do this once and you can forget about command line paths.



How to start command line programs from Apple's X11 Applications list?(see Figures 3-6 below).

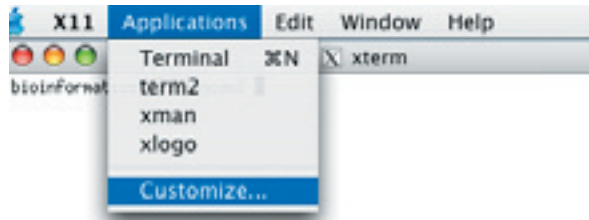


Figure 3. Start Apple's X11. Click on Applications, choose Customize.

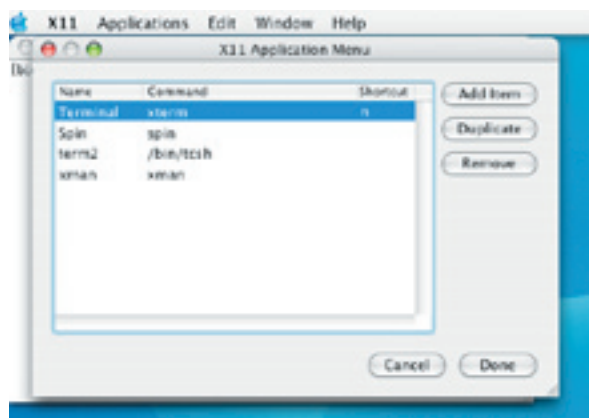


Figure 4. This opens a configuration window

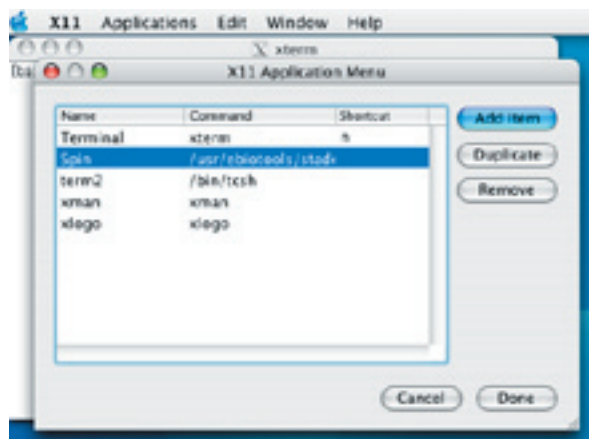


Figure 5. Click on the Add Item button. Click twice on the empty field to make it split in two fields one for the name you choose (a good and easy to remember name) and one for the actual path to the program you want to run.

Now you can add all the eBioinformatics programs you want. E.g `EMBOSS /usr/ebiotools/bin/embosslauncher.kaptn` that launches the GUI `embosslauncher.kaptn`.

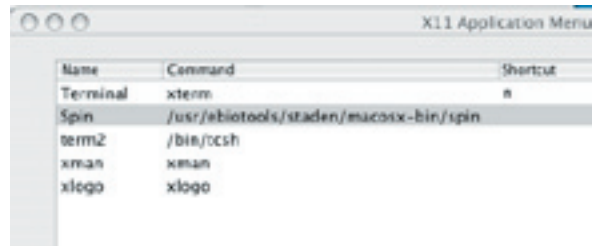


Figure 6. Here you can see in detail the path to the Staden program Spin.

Please also add "nedit", a text editor with very nice features that solves most of the format problems created by the use of MS-Word. `Nedit /usr/ebiotools/bin/nedit` If you want to have a "nedit" with coloured features and many other enhancements as shown in figure 7, then please download the ".nedit" configuration file written by Thomas Siegmund. You can find it at:

<http://userpage.fu-berlin.de/~sgmd/.nedit>

Place this file in your home directory. Please remember that all files starting with "." (dot) are invisible on a Unix system (and MacOSX is one!).

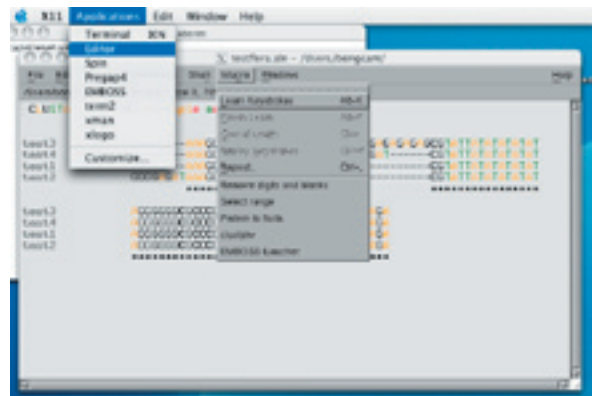


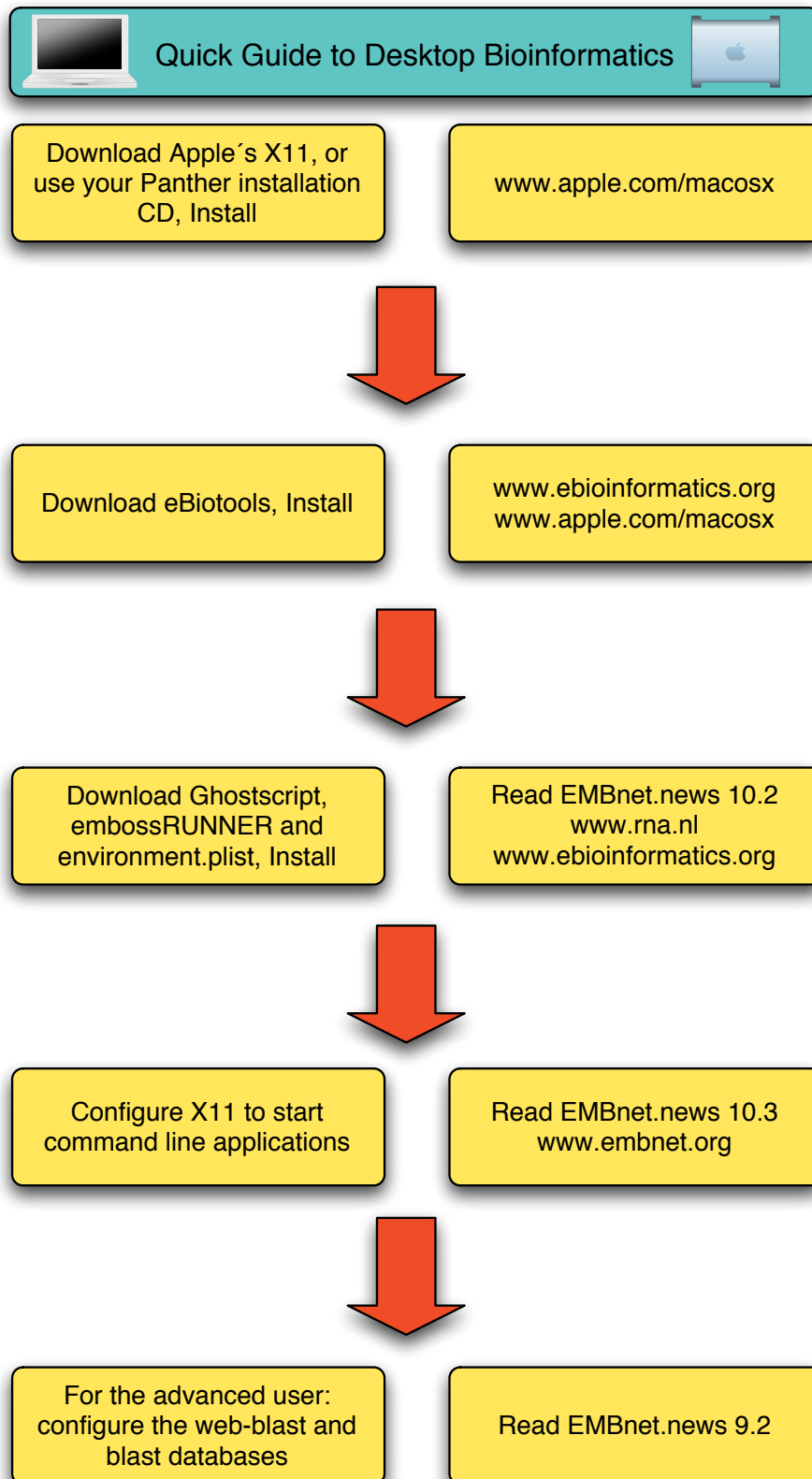
Figure 7. Shows Nedit (I call it Editor to make it easy to remember) with a clustalW alignment. This alignment is fully editable. If you want to produce color pictures with high quality, use also the MacOSX native ClustalX application developed by Anders Nister

([www.ebioinformatics.org](http://www.ebioinformatics.org)).

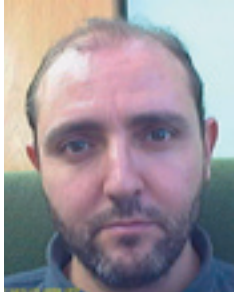


See also next page for a Quick Guide describing the installation process.

We would like to thank all people that write to us as their questions and comments make this package better and better.



## Project Management (part 3): Project Development



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

In this article we will proceed with our review of Project Management techniques delving into the Development phase. Previously we have defined, starting from an initial idea or need what it is that we want to achieve (the goal and scope of our project) and presented our proposal to the appropriate authority for approval, making a business case to defend the interest of the project.

It is now time to move ahead and start planning in detail how are we going to achieve the goals that have been agreed and approved. Probably the single most important element for project success is the planning phase: every aspect should be taken into consideration and carefully analysed in advance if we want to avoid nasty surprises later on.

Once we get an initial plan drafted, we must submit it for approval and start working using it. This being a not-so-ideal world, we may need to make changes to the plan in later phases to adapt to upcoming events, and so we will continue using the techniques and tools depicted here during the execution of the project.

### Getting started

To draw the project plan we will have to *find out exactly what is needed to make the project succeed*, and compare it to our initial expectations and business case.

So, what should be expected from a well-thought out project plan? At a minimum it should consist of

- A detailed breakdown of tasks needed to carry out the project (the **Work Breakdown Structure**)
- A detailed inventory of resources needed to ensure project success and how they are going to be used (the **Cost Analysis**)
- An analysis of factors that may affect the successful completion of our project and possible contingency actions to deal with them (the **Risk Analysis**)
- Details on how the project can be followed up to monitor its evolution and the quality of the results (the **Quality Assurance Plan**)

It should be noted that *all of these should be drafted in parallel, and subject to negotiation with key stakeholders.*

*It is your work as a Project Manager to make sure the plan is accurate, feasible and that it satisfies all key stakeholders; and this will require from you both, good technical and sociological/political abilities.*

### The Project Work Breakdown Structure (WBS)

As we mentioned in our introductory article, a project is defined in the **PMBOK** (the ISO standard for Project Management) as “*a temporary endeavour undertaken to create a unique product or service*”.

This means that it is *a complex process to be undertaken in finite time*. To deal with this complexity we can break down the overall project into smaller parts that may be handled conveniently, and place them in a suitable time frame to fit our time constraints. This is what we intend to achieve with the **WBS**.

To build a WBS we must *break our project*



into smaller **tasks**. While we may try to pin down the project to the very last detail, this is a cumbersome and error-prone task for non-trivial projects. Instead, it is advisable to start by *dividing our project into smaller subprojects*, which can then be further refined separately.

For complex projects, with many subprojects and component tasks (some may go up to thousands of tasks) it is worth *grouping together sets of subprojects sharing a common trait* into coherent bundles known as **work packages**.

*Organising tasks has various benefits*: first it makes the breakdown structure conceptually easier to handle; second, it allows us to divert specification of workpackages and subprojects to the people who will be responsible for them; by doing this, we'll establish a hierarchy or network of responsibility among participants; we'll avoid being too intrusive on the details of how people carry out their work, which is often a serious source of dissatisfaction and problems; and finally this acknowledges them as key stakeholders of our project (as the ones who must make it happen).

Whenever possible, *it is better to rely on people* who will actually do the work to produce accurate estimates of the time and resources required to achieve their sub-goals. This may mean that you will need to call up meetings of developers/implementers to draft the plan, and perhaps include some of them in your meetings to discuss the plan with other key stakeholders.

*Sometimes, however, this will not be possible*, because you do not know in advance who will be responsible for doing what in the end. In this case, we must do it ourselves, drawing from our technical expertise, previous experience or technical advisers, or a combination of all of these.

Almost any PM tool provides facilities for making a hierarchical breakdown of a project into tasks. You will find the more relevant differences between standalone, single-

user tools, where one person must centralize all the breakdown work (possibly collecting estimates from experts), and distributed tools (e.g. Web based) where you may assign subprojects to other project members and defer further breakdown to them, thus sharing the work load (see figure 1).

Name of Project	Begin	End	Category	Status	Priority	Leader
ALION	2003-09-25	2004-09-15	offered	<input type="checkbox"/>	1	J
Analysis phase	2003-09-25	2003-10-15	offered	<input type="checkbox"/>	1	david
First release	2003-10-16	2004-01-15	offered	<input type="checkbox"/>	1	J
Second release	2004-01-16	2004-04-15	offered	<input type="checkbox"/>	1	mad
First polishing	2004-01-16	2004-04-15	offered	<input type="checkbox"/>	2	david
Third release	2004-04-16	2004-07-15	offered	<input type="checkbox"/>	1	david
Second polishing	2004-04-16	2004-07-15	offered	<input type="checkbox"/>	2	mad
Odds & Ends	2004-07-16	2004-09-15	offered	<input type="checkbox"/>	1	mad

Figure 1. A WBS with a hierarchical structure, with specific subprojects assigned to different leaders (snapshot proceeding from PHProjekt). Note that other task properties may be set as well (priority, duration...)

## Estimating task duration

At the same time we break down our project into suitable subprojects and further on down to leaf tasks, we must consider the time duration of each of them. We need to make estimations on how long each of these tasks will take and fit them in a feasible schedule.

*Estimating task duration is not easy as it involves a prediction (and a compromise)*. We can't know the future, but often we may have an intuitive, more or less accurate idea of how long and how much work will be required to carry out each task. If possible, we should ask our experts to produce estimations.

*Previous experience should always be taken into account*. Even experts may produce inaccurate estimates or task breakdowns for various reasons (ambition, optimism, ignorance...). Thus, at any rate, whenever possible we should try to include previous experience in similar tasks or projects, if possible from within our own Company, using performance records, or in their absence, using standard industry-wide estimations (often available from technical publications).

That is why every project manager should *try to keep a log of work and project*

performance and make post-mortem tests on every project to collect useful statistics that can be used to make future estimations more accurate (we'll see more on this in an upcoming article in this series).

*Time is a relative quantity.* Most often we take it for granted, and tend to forget about its huge variability. *Short projects may specify task duration in hours*, but when planning any project that lasts more than a few days, we should always have a calendar (or more) at hand: when calculating completion dates you should add to the duration estimation holidays, non-working days, staff vacancies and so on. *If you manage an international project you should then take into account local holidays and variations as well* (e.g. working days vary across cultures and countries, summer differs between hemispheres...).

There are many tools that provide support for integration of calendar information from

various countries. Generally these will allow you to tailor the calendar to yearly variations. Still, *you should always find a way to reflect in your schedule the impact of work leaves*, which will not usually be known in advance (See figures 2).

## The dependency graph

Once we have our duration estimates, allowing for work leaves and adjusted to local variations, we need to *place tasks in time*. Of course, we may just schedule one after the other linearly, but most often this will be highly inefficient. Even so, we still need to consider the order in which we are going to do them.

It obviously does not make sense to submit a paper for publication before we have written it. Some tasks exhibit *precedence relationships* among them, and these should

Figure 2 (A) shows a calendar for August 2004. The calendar is organized by week (33, 34, 35, 36) and day of the week (Monday to Sunday). Local holidays are highlighted in different colors: green for Friday, 08/13/2004; pink for Saturday, 08/14/2004; and red for Thursday, 08/26/2004 (Summer Bank Holiday (UK)). Other holidays listed include Maria Himmelfahrt (D), Maria Himmelfahrt (A), and Assumption (F).

Figure 2 (B) shows the user preferences configuration dialog. The settings are as follows:

- Language: automatic (by browser)
- Timezone: Europe/Berlin
- Working Days: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday (all checked)
- Start of Week: Monday
- Holidays: DE, DE\_AT, EN\_UK, ES (checked), FR, US
- Namedays: PL, DE, NL, HU, CS
- Theme: blue
- Layout: classic\_layout
- Changes OK?: Enter
- Reset: Reset

Figure 2. (A top) A view of the calendar tool from a web based PM package (TUTOS) showing the month of August, 2004: note how various local holidays in different countries are noted by default. (B bottom) A section of the user preferences configuration dialog in TUTOS: showing internationalization options for a multinational project (e.g. language, time zone, working days and holidays that apply).

be taken into account when placing them in a time schedule.

It is very important that we *define all task relationships at this phase*: first, this will allow us to realize which tasks are independent among themselves and may take place simultaneously (hence reducing the time to complete the project), second it will allow us to schedule them in an appropriate order.

Significant relationships are "**Finish to Start (FS)**" (i.e. the precedent task must be finished before we start the subsequent task), "**Start to Finish (SF)**" (one task must be started before we can consider the other closed), "**Start to Start**" (one task must be started before the other), "**Finish to Finish**", etc. Of these, the most commonly used is the first one, "**Finish to Start**" (See figure 3).

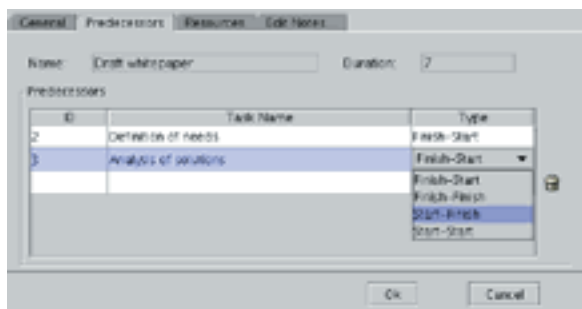


Figure 3. Selection dialog for task relationships in GanttProject.

Establishing dependences is important for another reason: it allows us to identify **sensitive paths** along the road to success in our plan. Sensitive paths define important relationships where delays in one task may have a significant impact on the rest of the project. Among these, the most important is the **Critical Path**, which defines the shortest path through interrelated tasks that leads to completion. This will allow us to set the optimistic **termination date** for our project, also known as **baseline**, the reference time frame against we will measure our performance.

In addition, the critical path is important because it identifies *tasks whose completion is critical to the termination date*: delays or schedule slip-offs along the critical path

will imply delays in the termination of our project.

Relationships *outside the critical path* are also important: they define the impact of each task on their dependences, and will allow us to identify the effect of schedule slippages during the execution phase of the project.

As we shall see later, external factors and prediction inaccuracy may have an impact on task completion. During the planning phase we *should always give at least two estimations*: the optimistic or **earlier completion and finish dates** and the pessimistic or **latest completion and finish dates** for each task. This will allow us (and our stakeholders) to know the exact time frame we are using and the reliability of our estimations, helping get a more accurate picture of our plan.

In order to better convey this information, a number of techniques have been developed that provide graphical representations of our project plan that visualize task relationships and the time schedule.

The most popular are the **network chart** (showing task dependencies), the **GANTT chart**, which places tasks in a time schedule, and the **PERT/CPM chart** which includes relationships and duration estimates to ease identification of critical paths, and a number of extended representations combining features of each of them into a single graph (See figure 4).

## Resource planning

Just as we go along our plan, we need to *identify the resources needed to make it succeed*. This is obviously tightly coupled with the estimation of other factors (e.g. given more resources a task may be finished earlier) and will add to the total cost of the project (and hence will need agreement and approval by our client and stakeholders).

At this stage you need to *consider carefully who will be involved in the project, defining*



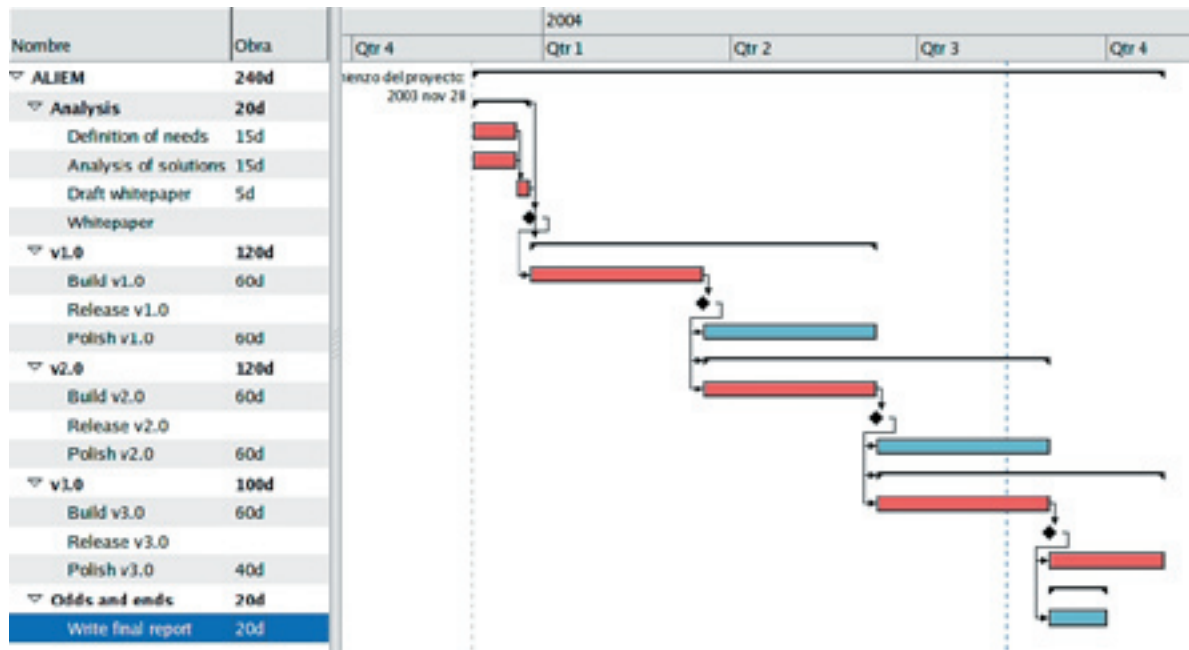


Figure 4. An extended GANTT representation obtained with MrProject. Tasks are identified as bars, with subprojects bundling several tasks depicted as brackets. Relationships are identified by arrows among tasks, and the critical path is highlighted in red. Milestones are identified with rhombic marks.

*clearly their skills, experience and knowledge, etc.* and you should consider to start building the team that will carry out the work.

*Starting to build up a successful team* now has many advantages: it will reduce uncertainty in planning, and increase the number of people and advice you can draw from to make your plan more accurate. Getting them involved helps build up the *esprit de corps* (group identity) and their identification with project goals, and allows everybody to know each other better and ease out differences.

However, *in most cases team build-up must be deferred to the execution phase.* This happens when you depend on plan approval before going ahead to build the team (they may be working in other departments, need to be hired with project funds or subcontracted to third parties, etc.).

Often you will find a *mixed scenario*: you know some of the people that will be involved (e.g. other project managers or research groups) with allocation of additional staff being deferred until the plan is approved and funds secured for the execution phase.

This takes us to a relevant point: *resource allocation in the Civil Service.* As the GEPMBok (**A Government Extension to the PMBoK**) notes, public agencies lack the flexibility of private companies: "A *project manager is not afforded the ability to hire personnel to fill critical skills needs without going through a bureaucratic process. If a hiring is authorized, the process can take several months*"; civil servants often have tenure positions and must attend other duties; they have a duty to the public which may introduce conflicts with project goals. *You should take all of these factors into consideration in any project dealing with civil servants.*

People are probably the most relevant asset in most projects, but not the only one. In most instances *you will also need other resources to complete your work.* This may come in two forms: existing facilities and non-existing resources.

*Existing facilities* are often taken for granted. This is not so, and should not be so: they still have a running cost (e.g. electricity or maintenance) that should be included in the cost statement as overheads. Calculating

**overheads** may become difficult at times and that is why many organisations use as a rule of thumb a given percentage of all project costs as the average overhead to be charged on all projects. *You should find out what your organisation rules are regarding overhead calculation and factor them in your cost statement.*

It is easier to make an *inventory of required, unavailable resources*, but by no means are they that much easier to account: unless you are to buy them all straight away from the start, you will need to make some estimation of their cost when they are needed, and plan in advance for their timely availability. In a fluctuating market, predictions may be highly biased, and that is why many organisations and funding agencies have defined strict limits to the amounts that may be allocated to specific items. *It is your role to find out what these limits are and ensure your project plan complies with them.*

Acquired resources must be further refined in additional categories: *some goods/resources have a durable lifetime*, and funding agents may be reluctant to pay in full for goods that you will keep for your own use after the project is over. You may need to negotiate with them how much of their cost is to be covered or to revert those goods to the client once the project is over. *Other goods have a limited lifetime* and are considered useless afterwards: their cost is usually divided by their lifespan and their cost factored considering the duration of their use during the project. Again, specific factoring terms may have been defined as standards for these situations, and you should be aware of them.

As you can see, *we need to state both the resources needed and the periods during which they will be used.* These will be associated to the duration of the tasks in which they are needed, and hence resource planning needs to be tightly coupled with task planning.

Most tools allow you to define the resources (human and other goods) that

are associated with your project on a by-task basis, specifying the time periods during which they will be needed. This is needed to ensure proper usage of shared resources (e.g. staff, common rooms, etc.) and to facilitate proper accounting of the project costs. Depending on their degree of sophistication, tools may include more or less advanced cost estimation facilities and resource scheduling.

## Risk Analysis

Life being as it is, there is a whole lot of uncertainty about it. No project plan will be complete if it does not take into account these uncertainties. This analysis helps stakeholders take educated decisions and *evaluate the inherent riskiness of a project.*

There are *two kinds of uncertainties* to consider, and both should be evaluated and enumerated in your project plan: **unavoidable and avoidable risks.**

You should try to *identify all the project risks in the planning phase*, together with an estimation of their likeliness. The plan must include a **list of risks** in both categories, **sorted by likelihood and impact** on the project. It may not be possible to identify all risks in advance, *the risk list should be available to everybody involved in the project along its full life cycle*, so that any new risk that may be identified later on can be added to the list and considered.

Besides providing an indicator of the uncertainty associated to your project, you also need to specify how you are going to deal with these risks.

**The ideal way to deal with risks is to altogether avoid them.** This means that you must *identify events that advise of risk proximity and trigger their appearance.* During the execution phase the project team will have to keep an eye on these indicators so as to be able to identify their presence.

Next, you must *specify the corrective actions that you can take when any of these indicators appear in order to prevent the materialization of the risk.*

Finally, you need *to plan and specify which recovery actions you will need to undertake to fix any impact that the risk may have had on your project plan.*

These activities may take the form of *additional tasks in your project plan and WBS that might seem unneeded costs since they only add value if the risk materializes (like performing periodic backups), and hence may need some education of stakeholders to make them see their convenience.*

All these activities have an *associated cost*, which you must assert and factorize with their associated risk likelihood. *These costs should be added to the cost statement and included in the resource planning for your project.* You can view it as your insurance plan (it may just be so), which should be tailored for your project needs.

*There are many tools available for performing risk analysis.* Most of them already include default provisions for the most common risks identified in specific industrial fields, and allow you to add new factors and model their effect on your project.

In general, ***you must assume that a number of risks will materialize and alter your proposed baseline*** project duration estimate. Existing

tools will help you determine how relevant the impact will be and the relative probabilities of project completion at a given date. That is why you should never compromise on a single date, but instead provide an optimistic date (if you get the lucky and nothing happens), a pessimistic date, and a likely date (usually that with 50% probability or the one most likely) (See figure 5).

## Quality Assurance Plan

All through this exposition we have mentioned that the project plan is more than a statement of intentions describing how we intend to carry out the project. Once approved it will constitute a contract binding us to the activities described, and a reference document to carry out the work depicted.

The QA plan attempts to provide a way to measure the advance of the project and the quality of the work produced. It will allow the early detection and correction of problems and ensure the final product lives up to the standards expected and satisfies everybody.

To have a good reference, we need to *include provisions in our plan for measuring project advances and the quality of the work developed.* This will allow us as project managers and our stakeholders and clients to measure the evolution of the project, detect when it gets out of hand or slips out of schedule and take corrective actions in a

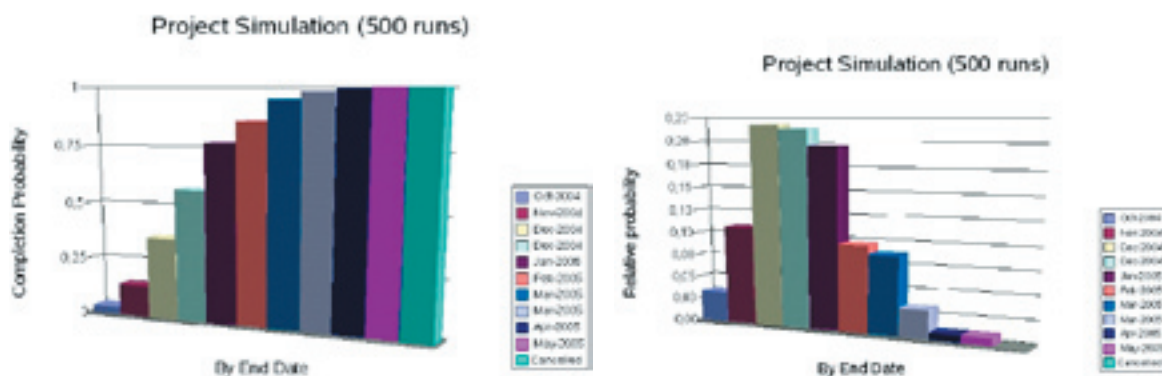


Figure 5. Risk analysis simulation of a project showing the completion probability (A) and the relative probability (B) of completion at various dates, obtained with RISKOLGY.

timely and effective manner. It will allow us as well to carry out a proportional accounting and monitor usage of resources ensuring they are not needlessly wasted.

We need to state what is expected of our project, how can we measure its advances, the quality of the results and what resources can be reasonably devoted to obtain them.

*Quality Assurance in a project can be partitioned following the project WBS.* In this approach we consider each subproject and task as a separate project, defining what is the final product expected, and how can we measure task completion and the final quality of the product. We should as well consider the risk that any intermediate product doesn't live up to our expectations and define what are we going to do to remedy this with minimal impact on the rest of the project.

This means that we must build a list of **deliverables** for our project: not only the final product, but also the intermediate elements that we will build to achieve our goal, indicating their expected completion dates, and how can its development be measured and their final quality be tested (e.g. by defining the standards they must comply with).

While we, as project managers, need to keep an eye on project evolution, it will often be inconvenient for our client or other stakeholders to spend much time closely monitoring the project work. For this reason, it is convenient to *introduce in our project specific dates or milestones at which we expect to have achieved relevant results.* Milestones will punctuate our project serving as indicators of its evolution.

*Useful milestones will usually be associated with major achievements in our project.* They will serve as indicators that a specific phase has terminated with measurable results (deliverables) and that we are ready to go ahead. They may be associated with further gate reviews where the client asserts his/her satisfaction with the development of

the project up to the milestone and decides whether it is worth to go ahead.

## Time for another review

After we are done looking at our project in detail, we may discover that what initially looked like a great idea was, on closer look, not really so.

For this reason, a new **gate review** may need to be conducted after the project plan is finished to *reconsider the project and define if it is worth going ahead once the full details are known.* It is for the same reason that many sites prefer to have a more or less accurate project plan before giving the initial project approval. Actually it varies between companies or funding agencies: in some cases there is no budget allocated for project planning, and promising project ideas must be evaluated before awarding them some money to proceed to planning, while in other situations, project managers coming up with new ideas are expected to document them with a detailed project plan before getting approval.

## Learning more

The PMBoK is an excellent resource and the ISO standard for Project Management. The "Government Extension to the PMBoK" is an addition to the standard for projects dealing with the public administration. There are many other books and resources that are of interest. "The Complete Idiot's Guide to Project Management", by Sunny and Kim Baker (Alpha Books) is a good general reference. "Waltzing with bears: Managing Risk on Software Projects" by T. de Marco and T. Lister (Dorset) is an excellent introduction to risk management. "Slack: getting past burnout, busywork and the myth of total efficiency", by Tom de Marco (Broadway) provides insight on time allocation. There are many references of interest; you can get more details at <http://bioportal.cnb.uam.es>



## Snowy stardom

By Vivienne Baillie Gerritsen

**One higher primate gained recognition for being sent up into space, another for memorizing language signs and yet another for saving the life of a young boy. But none of them reached the heights of fame the albino gorilla ‘Snowflake’ did. Snowflake was the first case of albinism in great apes ever recorded. He suffered from the now well-documented pathology: oculocutaneous albinism type 1 (or OCA1). This type of albinism is the most common form of albinism in humans and is caused by the malfunction of a tyrosinase, an enzyme which has a key role in the synthesis of the pigment melanin.**

Snowflake was a Western lowland albino gorilla (*Gorilla gorilla gorilla*), captured in Equatorial Guinea in October 1966 by a Catalan primatologist Jordi Sabater Pi. He was originally named Nfumi Ngi, meaning ‘white gorilla’ in the native Guinean language, before the National Geographic nicknamed him ‘Snowflake’. Snowflake was shipped to Spain where he became the icon of Barcelona Zoo and was known as ‘Floquet de Neu’, which is Catalan for snowflake. Snowflake fathered no less than 22 offspring, none of which were albinos and even lived to see his grandchildren! In 2001 however, at the mature age of almost 40 – well over the average lifespan of 25 years for wild gorillas – the gorilla was diagnosed with skin cancer, until then a disease quite unknown within the species but sadly quite frequent in human albinos. He was euthanized in 2003 and the news of his death made the headlines worldwide.

The first hints of tyrosinase activity arose in the late 1900s when the French naturalists Bourquelot and Bertrand first observed that the toadstool *Russula nigricans* bore a colorless substance which blackened upon exposure to air. A year following the discovery, Bertrand recognized the substance as tyrosine. In 1904, barely a year later, it was discovered that extracts of a mammalian melanoma could convert tyrosine to melanin. Scientists were beginning to grasp the intricate business of melanogenesis, i.e. the synthesis of the natural pigment melanin. It was the British physiologist H.S.Raper who made the most important studies on melanogenesis from 1920 to 1935 but it was only during the 1950s that some real sense was made out of the melanin biosynthetic pathway. Tyrosinase was the enzyme which acted upon

tyrosine, which would ultimately lead to the formation of the natural pigment melanin.

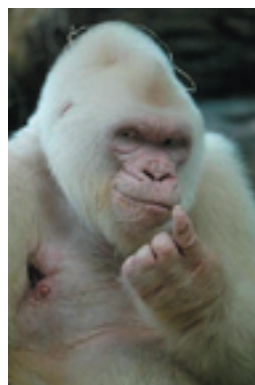


Fig.1 Snowflake, the albino gorilla.

The discovery of the processes which lurk behind albinism followed close behind the discovery of tyrosinase activity. Already in 1908, Sir Archibald Garrod (1857-1936), a British physician and scientist, suggested that albinism was probably due to the failure of an intracellular enzyme. Though he did mention the fact that tyrosinase did indeed have a role in melanogenesis, he didn't make the link between albinism and loss of tyrosinase activity. Indeed, for decades it remained obscure as to whether albinism was a consequence of the loss of tyrosinase activity or an absence of melanocytes – the cells in which melanin is synthesized. Towards the end of the 1950s, it was shown that albinism – at least the form that was being studied – was not due to the absence of melanocytes but indeed to the inactivity of the tyrosinases.

What exactly is albinism? There are a number of forms of albinism but on the whole albinism is a lack of melanin formation which can bring on a number of complications. The most frequent form is known as oculocutaneous albinism type 1 (OCA1) which results from a malfunction of the tyrosinase enzyme. Individuals with OCA1 suffer from hypopigmentation of the skin and hair as well as visual disorders which are a direct consequence of the non-synthesis of melanin. In humans, this type of albinism occurs in 1 of 40'000 births. Being a very visual phenotype, this disease was in fact one of the earliest genetic disorders to be studied and, interestingly, the analysis of its distribution – once inherited – demonstrated that Mendel's laws of segregation not only applied to plants but also to mammals.

Though it is thought that the tyrosinase enzymes may well function in the form of multienzyme complexes, tyrosinases are essential to melanin formation. They actually trigger off the reaction which will ultimately convert the amino acid tyrosine to the melanin biopolymer. What tyrosinase does in this crucial step is catalyze the hydroxylation of L-tyrosine to 3,4-dihydroxyphenylalanine (DOPA) and then the oxidation of DOPA to DOPAquinone. And DOPAquinone is the key intermediate in the formation of two different types of melanin: the pheomelanins, which are yellow and red, and the eumelanins, which are brown and black.

Tyrosinases perform their function securely lodged within the melanocyte membranes. Their catalytic domain sports two copper atoms ligated to a series of histidine residues. The hydroxylation of tyrosine begins by the binding of a dioxygen to the copper atoms, which lie at a distance of 3.6 Å. When the tyrosinase gene is mutated one way or another (frameshifts, splice site mutations, deletions, substitutions...), what frequently happens is that the enzyme's structure is modified in such a way that the distance between the copper atoms is altered thus leading to the impossibility of tyrosine hydroxylation.

Despite the fact that over 100 different mutations within the tyrosinase gene are now known to cause OCA1, scientists still have not found which one applies to Snowflake's albinism. His condition was certainly due to a malfunction of the tyrosinase enzyme but what was it that caused this? Substitutions in the sequence have not been spotted, however there could be a mutation in an upstream regulatory element. Time will tell.

Besides clarifying the processes involved in Snowflake's albinism, melanogenesis and the function of the tyrosinase enzyme itself, studies within these fields are also helping in the fight against skin cancer. The antigenic properties of tyrosinases, for example, should be helpful as early markers for the detection of metastases whilst the specificity of their promoters could also be used in the event of gene therapy.

## Cross-references to Swiss-Prot

Tyrosinase, *Gorilla gorilla gorilla* (Lowland gorilla) : Q9BDE0

## References

1. Oetting W.S., Fryer J.P., Shriram S., King R.A.  
Oculocutaneous albinism type 1: the last 100 years  
Pigment Cell Res. 16:307-311(2003)  
PMID: 12753405
2. Martinez-Arias R., Comas D., Andrés A., Abello M.-T.  
The Tyrosinase gene in gorillas and the albinism of Snowflake  
Pigment Cell Res. 13:467-470(2000)  
PMID: 11153699
3. Oetting W.  
The tyrosinase gene and oculocutaneous albinism type 1 (OCA1): a model for understanding the molecular biology of melanin formation  
Pigment Cell Res. 13:320-325(2000)  
PMID: 11041207

## A virus for life

By Vivienne Baillie Gerritsen

**Who would have thought that a virus could have anything to do with a tissue as important to life, and its development, as the placenta? A viral protein, now known as syncytin, whose gene was probably integrated into the primate genome over 25 millions years ago, is hugely expressed in placental tissue – especially at the beginning of embryonic development – and is giving signs of bearing an essential role in placental architecture. Could it be then that placental evolution, and indeed mammalian evolution, finds its origins in the doings of a virus?**

The virus in question is in fact a retrovirus known as human endogenous retrovirus : HERV-W. The human genome – as indeed many other eukaryotic genomes – is sprinkled with retroviral genes and sports up to 8% of endogenous retroviral sequences! With the passing of time, many of them become dysfunctional due to mutations but a number of them do seem to survive and their sequence can be translated into healthy proteins. It is not always clear how their expression could benefit the host's biology but the discovery, in the year 2000, of the retroviral protein syncytin may well be a fascinating example of how a retroviral product has been integrated and kept relatively intact throughout time, in the host's genome, and whose function – possibly via natural selection – serves the host's purpose.

Syncytin is a 538 amino-acid viral envelope protein, which exhibits a transmembrane and a cytoplasmic subunit. It is thought that these two subunits are posttranslationally cleaved and meet up at the cell's membrane. In general, a viral envelope protein mediates cell fusion and participates in building up the virus' envelope. What was discovered was that syncytin seems to fulfill the same purpose in humans, i.e. cell fusion, but at the level of the placenta.

The placenta is lined, on the mother's side, by a tissue known as the syncytiotrophoblast – from where the protein's name derives. The syncytiotrophoblast is an essential tissue for the proper development of the embryo, acting as a place of nutritional exchange between the mother and her child, as well as a center for hormone synthesis and what is more, it has a role in the foetus' immunity. The syncytiotrophoblast is a one-celled layer formed

via the fusion of trophoblasts. But what exactly are trophoblasts? Trophoblasts are the cells which accompany the foetus in the very early stages of development giving it all it needs and helping it to anchor itself in the womb. The fusion of trophoblasts creates the syncytiotrophoblast, the tissue which hugs the placenta. And yes, you guessed, syncytin is involved in the fusion of trophoblasts.



**Fig.1** *The Foetus in the Womb*  
Leonardo da Vinci (c.1510-1512)

In 'normal' circumstances, i.e. in the event of viral infection, syncytin would be synthesized thanks to the host's translating machinery and would then be found at the surface of a viral cell. From there it would recognize a specific receptor on a second viral cell, make a beeline for it, and mediate cell fusion. Syncytin has been squatting the human genome for millions of years now; it still does use our machinery to be expressed but instead of initiating viral envelope formation, it is involved in mediating

trophoblast fusion. Indeed, the expression of syncytin during placental formation is high. And what is more, besides some expression in the testis, this seems to be the only place the protein is expressed.

Not only does syncytin have a role in placental morphogenesis but it may well have a role in driving off viral infection too. How? Via receptor interference. When syncytin drives trophoblast fusion to form the syncytiotrophoblast, it probably does so by using cell receptors that are used by other 'healthy' invading viruses. If syncytin grabs the receptor before another viral envelope protein gets the chance, it is in effect interfering with the formation of invading viruses and hence viral infection of the foetus. Which is good news. A little mind-boggling to realize that a virus is at the heart of something so essential as the human placenta and what is more is capable of fighting off ... viral infection of the human foetus. Here is a virus we owe one to!

Hence, from an evolutionary point of view, syncytin could have quite a role in placental development. So far though, it seems that HERV-W is only shared between humans and higher primates which wouldn't make much of a basis for a major role in placental evolution. However, interestingly, there are a number of different HERV-W families which are found

dispersed in various mammals. Could it be then that various viral families are at the heart of the diverse types of placenta observed in mammals today? Certainly, syncytin does appear to have something to do with certain malfunctions of the placenta such as pre-eclampsia (high blood pressure, weight gain and convulsions) or HELLP (hemolysis, elevated liver enzymes, low platelets) syndrome because, in the event of such ailments, it is not expressed at the normal levels.

All this raises hopes in the field of biotechnology and the search for treatments against such diseases during pregnancy – although both diseases are multisystem disorders, and cell fusion is a far more complicated process than the mere recognition of a membrane receptor by syncytin. However, the discovery of specific amino acids within the syncytin sequence, which are paramount in receptor recognition, could provide a basis for the genetic engineering of the syncytin polypeptide. Indeed, it has been shown that the fusogenic protein when expressed in tumor cells can bring about cell fusion and subsequently tumor cell death. The ugly part is that syncytin could even perhaps participate in envelope formation of invading viruses, such as HIV-1 for instance, and potentially contribute to its pathogenesis. Time will tell.

### Cross-references to Swiss-Prot

Syncytin, human endogenous retrovirus-W (HERV-W): Q9UQF0

### References

1. Frendo J.-L., Olivier D., Cheynet V., Blond J.L., Bouton O., Vidaud M., Rabreau M., Evain-Brion D., Mallet F.  
Direct involvement of HERV-W env glycoprotein in human trophoblast cell fusion and differentiation  
*Mol. Cell Biol.* 23:3566-3574(2003)  
PMID: 12724415
2. Knerr I., Beinder E., Rascher W.  
Syncytin, a novel human endogenous retroviral gene in human placenta: evidence for its dysregulation in preeclampsia and HELLP syndrome  
*Am. J. Obstet. Gynecol.* 186:210-213(2002)  
PMID: 11854637
3. Mi S., Lee X., Li X.-P., Veldman G.M., Finnerty H., Racie L., LaVallie E., Tang X.-Y., Edouard P., Howes S., Keith J.C., McCoy J.M.  
Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis  
*Nature* 403:785-789(2000)  
PMID: 10693809



## 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>

### Brasil

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

### Canada

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

### Chile

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

### China

Jingchu Luo  
Room 303, Exchange Centre, Peking University  
Email: luojc@cbi.pku.edu.cn  
Tel: +86-10-6275 9001  
<http://www.cbi.pku.edu.cn>

### Colombia

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

### Cuba

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

### Finland

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

### France

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

### Germany

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

### Hungary

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

### India

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

### Israel

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

### Italy

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

### Mexico

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

### The Netherlands

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

### Norway

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

### Poland

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

### Portugal

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

### Russia

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

### Slovakia

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

### South Africa

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

### Spain

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

### Sweden

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

### Switzerland

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

### United Kingdom

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

## Specialist Nodes

### EBI

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

### ETI

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

### EU

Frederick Marcus  
DG Research - European Commission  
Brussels BELGIUM  
Email: Frederick.Marcus@cec.eu.int

### ICGEB

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

### LION Bioscience

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

### MIPS

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

### UMBER

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

### TECH-MGR

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



## ISSN 1023-4144

Dear reader,

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

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

Submission deadline for next issues:

October 31, 2004  
February 29, 2005  
May 31, 2005