Specific Objectives of this vlab

This vlab has already started to bring people involved in the development of the "R virtual lab" part of the Lifewatch ESFRI project, together with leading experts in R during an Rvlab Workshop (*R for Lifewatchers*) organized and held in HCMR. Members of the R core group are currently working on the implementation of an optimized (with respect to computational speed-up) R online environment, which will be hosted at the Lifewatch website and will be readily available for all users to perform their analyses.
Hands-on experience has been attained in optimizing some commonly used routines of the VEGAN package, using parallel processing techniques like MPI and the parallel inbuilt pdbR package of R, and we are currently working on refining certain other aspects, e.g., improving the way R constructs are being used, using database structures (like PostgreSQL) to store large matrices.

Some of the routines we are interested in are written in R, others are in C and Fortran. And there are interdependencies among them, for instance metaMDS is calling postMDS which calls vegdist; while the former two are written in R, the latter is written in C.
As such, we need to work on all levels.
We have already parallelized certain functions (taxa2dist, taxondive, outer product, etc.), using MPI and the pbd R package, obtaining some promising preliminary numbers. Yet, as we dwell deeper into how R operates, it would really help us obtain more user case scenarios from real biological problems, a task which we are currently. To be more specific, we are currently investigating:
1. General practices that improve the performance of R codes (this is something we already study from online and other resources, but we are also investigating real-world biological user scenarios).
2. Suggestions on tools and methodologies that are needed to improve the C/Fortran codes and to couple them smoothly with the R environment.

This will help us to address the three major issues of this vlab namely:

1. How to handle big dataset big matrices (bigger than available soft memory).
2. How to Speed-up and break down the problem into sub-programs
3. Integration of vlab in the Lifewatch infrastructure

Why R for Biology

The idea behind the use of R as a statistical software package is to draw biologists away from the traditional Excel worksheet and allow them greater flexibility in analyzing their data. This is particularly important in fields like Systems Ecology and Biodiversity studies due to the large amount of data generated, that; often is impossible to be manipulated by conventional spreadsheet methods.

R is a powerful open-source statistical package that is used by many statisticians and is becoming increasingly popular in Biological related fields. One possible drawback of R is that there is a steep learning curve in order to master the language. However, once this is overcome the potential is definitely worth the time spent to become acquainted with this software environment

Advantages of R:

Free and well supported by its community

R is completely free, and can be downloaded from **www.r-project.org**. Moreover, R supports all major operating systems (Windows, Mac, Linux, Unix, etc.). In addition, there is a large R user community and core developers that further enhances availability of new packages, online free manuals and documentations as well as forums and wikis.

Excellent graphics

R allows for the flexibility in user specific construction of publication-quality graphical output supporting numerous formats (PDF, PS, EPS, JPEG, PNG, TIFF, etc.).

Can import files from other statistical programs

R allows for integration of data files from other statistical software such as Minitab, S, SAS, SPSS, or Stata, which can be imported into R. R supports all commonly used file types such as tab-delimited text, CSV, and even Excel can easily be imported.

New version every six months

A new version is released every six months, which means that any bugs are quickly fixed.

GUIs available

R has a command line interface, which some new users may find a bit intimidating. However, more intuitive point-and-click graphical user interfaces (GUI) are also available (RCommander, RStudio). More advanced users, or those planning on writing lengthy functions can also run R from Emacs.

Learn to program

In addition to being a statistical package, R is also a programming environment, and learning R gives basics of computer programming skills to the user. These skills can allow for a user to write programs or scripts in order to automate repetitive tasks, reduce errors and free up time.

Speak the language of your bioinformatics/statistics colleagues

Modern biomedical research is increasingly becoming a multi-disciplinary field. The knowledge of a high caliber language like R which, is commonly used by the bioinformatics, computational biology, and statistics communities, allows for the better communication with between biologists and a more informatics oriented collaborators.

Why VEGAN?

We selected the vegan package because it provides a variety of tools for descriptive community ecology analysis. It has most basic functions of diversity analysis, community ordination and dissimilarity analysis. Most of its multivariate tools can be used for other data types as well.

The functions in the vegan package contain tools for diversity analysis (as shown in the vignette vegandocs("diversity")),ordination and analysis of dissimilarities. Together with the labdsv package, the vegan package provides most standard tools of descriptive community analysis. Package ade4 provides an alternative comprehensive package, and several other packages complement vegan and provide tools for deeper analysis in specific fields. Package BiodiversityR provides a GUI for a large subset of vegan functionality.

Moreover we selected two basic functions of the VEGAN package (taxa2dist, taxondive) due to their high usability in the area of Ecology and Biodiversity data analysis. The taxon2dists function is commonly used to obtain a distance matrix from species abundance data derived from a specific environments. This is then commonly used as input to the taxondive function in order to obtain indices of taxonomic diversity and distinctness, from the averaged taxonomic distances among species or individuals in the community.

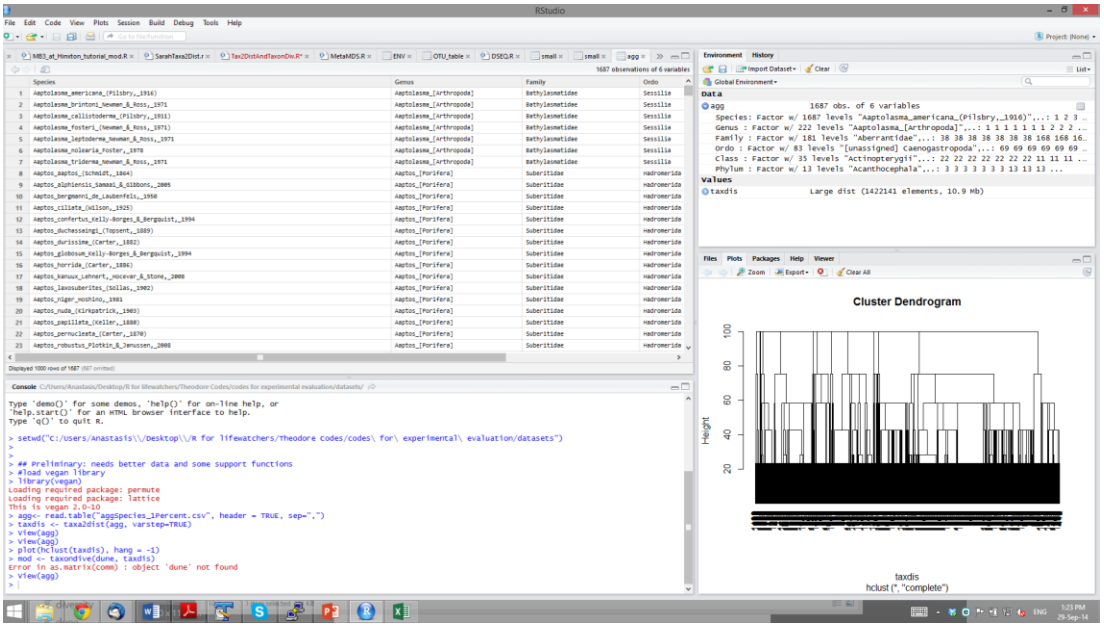

Figure x. Screenshot showing a test case scenario of a classification table with variable step lengths used as input to taxon2dist. Header represents taxon label from phylum to down to the species level.

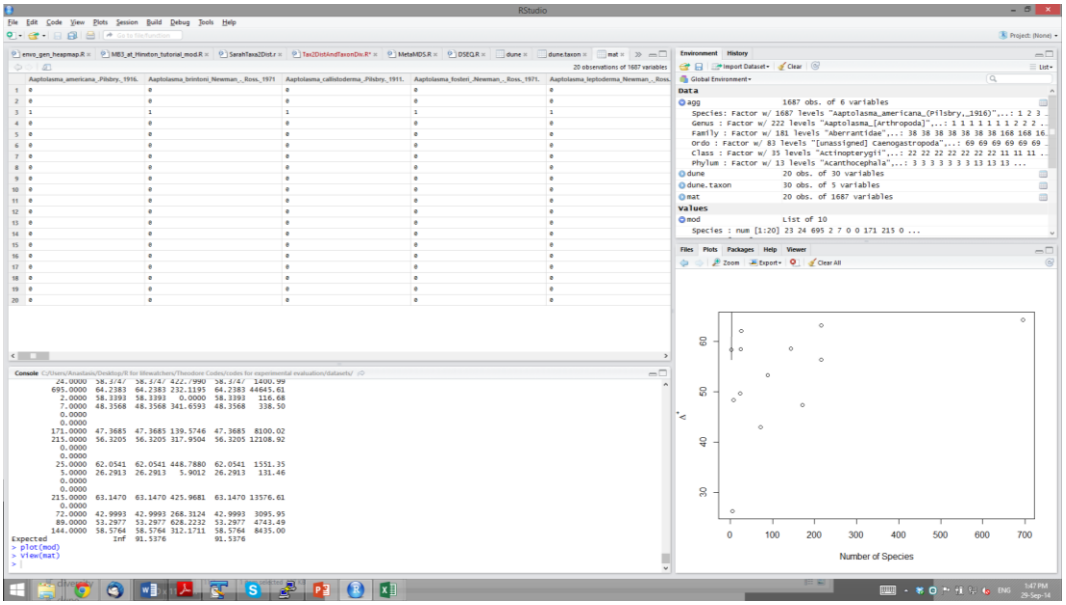

Figure x2. Screenshot showing a test case scenario of an abundance table with variable step lengths used as input to taxondive together with output from taxon2dist. Header represents species name and the rows represents samples. The integers are the abundance of the specific species in any given sample.

Integration of the R vlab to the general Lifewatch architecture/infrastructure.

We are also currently investigating ways to integrate the R vlab to the general Lifewatch infrastructure and taking in consideration possible was to achieve this in the future Specifically (in close collaboration with the Lifewatch data core repository implementing team at ICS) we are looking into:

- Data accessibility - retrieving data directly from Lifewatch and other sources and performing analyses in R vlab.
- Use the postgresQL structure to temporarily store data that end user wants to experiment on.
- Integration with triple store framework
- Implementation of user friendly interface similar to that publically available index function list created by colleagues in MPI Bremen (http://mb3is.megx.net/eatme/).

- Direct upload of user Rscript, parsing and modification according to the available implemented parallel functions of R and execution in an optimum way on the Lifewatch PC cluster architecture.