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.

			RStudio		- 8 <mark>-</mark> ×
File	Edit Code View Plots Session Build Debug Tools Help				
• •	🕂 🚰 📲 🔚 🛛 🧀 Go to file/function				B Project: (None) •
ж	M83_at_Himton_tutorial_mod.R × SarahTaxa2Dist.r × Tax2DistAndTaxonDix.R* ×	MetaMDS.R × ENV × OTU_table × OD	SEQ.R × small × small ×	»	Environment History
	10 D		1683	observations of 6 variables	😭 🕞 🖙 import Dataset= 🎻 Clear 💿 🔤 List=
	Species	Genus	Family	Ordo ^	Global Environment • Q
	1 Aaptolasma_americana_(Pilsbry,_1916)	Amptolesme_[Arthropode]	Bathylasmatidae	Sessilia	Data
	2 Aaptolasma_brintoni_Newman_&_Ross,_1971	Aaptolasma_[Arthropoda]	Bathylasmatidae	Sessilia	acq 1687 obs. of 6 variables
1	3 Aaptolasma_callistoderma_(Pilsbry,_1911)	Asptolesms_[Arthropode]	Bathylasmatidae	Sessilie	Species: Factor w/ 1687 levels "Aaptolasma_americana_(Pilsbry,_1916)",: 1 2 3
	4 Aaptolasma_fosteri_(Neuman_&_Ross,_1971)	Aaptolasma_[Arthropoda]	Bathylasmatidae	Sessilia	Genus : Factor w/ 222 levels "Aaptolasma_[Arthropoda]",: 1 1 1 1 1 1 1 2 2 2
1	5 Aaptolasma_leptoderma_Newman_&_Ross,_1971	Aaptolasma_[Arthropoda]	Bathylasmatidae	Sessilia	Family : Factor w/ 181 levels "Aberrantidae",: 38 38 38 38 38 38 38 168 168 16.
	6 Aaptolasma_nolearia_Foster,_1978	Aaptolasma_[Arthropoda]	Bathylasmatidae	Sessilia	Ordo : Factor w/ 83 levels "[unassigned] Caenogastropoda",: 69 69 69 69 69 69
1	7 Aaptolasma_triderma_Newman_&_Ross,_1971	Aaptolasma_[Arthropoda]	mathylasmatidae	sessilia	Class : Factor w/ 35 levels "Actinopterygii",: 22 22 22 22 22 22 22 11 11 11
1	8 Aaptos_aaptos_(schwidt,_1864)	Aaptos_[Porifera]	Suberitidae	Hadromerida	Phylum : Factor w/ 13 levels "Acanthocephala",: 3 3 3 3 3 3 3 3 13 13 1.
1	9 Aaptos_alphiensis_Samaai_&_Gibbons,_2005	Aaptos_[Porifera]	Suberitidae	Hadromerida	values
1	0 Aaptos_bergmanni_de_Laubenfels,_1950	Asptos_[Porifera]	Suberitidae	Hadromer1da	Gtaxdis Large dist (1422141 elements, 10.9 Mb)
1	<pre>Aaptos_ciliata_(Wilson,_1925)</pre>	Aaptos_[Porifera]	Suberitidae	Hadromerida	
1	2 Amptos_confertus_Kelly-Borges_B_Bergquist,_1994	Asptos_[Porifers]	Suberitidae	Hadromerida	
1	3 Aaptos_duchassaingi_(Topsent,_1889)	Asptos_[Porifers]	Suberitidae	Hadromerida	
1	4 Aeptos_durissime_(Certer,_1882)	Aaptos_[Porifera]	Suberitidae	Hadromerida	
1	5 Aaptos_globosum_Kelly-Borges_&_Bergquist,_1994	Aaptos_[Porifera]	Suberitidae	Hadromerida	Eller Bioly Parkager Main Viewer
1	6 Aaptos_horrida_(Carter,_1896)	Aaptos_[Porifera]	Suberitidae	Hadromerida	
1	7 Aaptos_kanuux_Lehnert,_Hocevar_&_Stone,_2008	Aaptos_[Porifera]	Suberitidae	Hadromerida	🗘 🗘 🖉 zoom 🦛 seport - 🥺 🔮 Clear All
1	<pre># Aaptos_laxosuberites_(sollas,_1902)</pre>	Aaptos_[Porifera]	suberitidae	Hadromerida	
1	9 Aaptos_niger_Hoshino,_1981	Aaptos_[Porifera]	Suberitidae	Hadromerida	Cluster Dendrogram
2	<pre>0 Aaptos_nuda_(kirkpatrick,_1983)</pre>	Aaptos_[Porifera]	Suberitidae	Hadromeniida	
2	<pre>Aaptos_papillata_(Keller,_1800)</pre>	Aaptos_[Porifera]	Suberitidae	Hadromerida	0
2	2 Aaptos_pernucleata_(Carter,_1870)	Aaptos_[Porifera]	Suberitidae	Hadromerida	
2	3 Aaptos_robustus_Plotkin_6_Janussen,_2008	Asptos_[Porifers]	Suberitidae	Hadromerida 🧹	
<				>	
Disp	skyed 1000 rows of 1667 (667 omitted)				
Co Typ The Typ Co Co Co Co Co Co Co Co Co Co Co Co Co	<pre>convertised and the thread here the transmission of transmission of the transmission of transmission of the transmission of transmission</pre>	R R R R R R R R R R R R R R R R R R R			
	i 🌐 🐨 🌍 🕥 🖬 🚛 🔽 💽 🚺 👔	2 🚯 💶			123 PM

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.

8			RStudi	0	- 6	×
File Edit Code View Plots Session	Build Debug Jools Help					
Q.• @•					3 Projec	t (None) -
					And the second second	-
envo_gen_neapmap.x = e_ Mo3_at	inkton_tutonal_mod.R × V_Sarahiaka2Dist/ >	 Marchistry and another K. X. M. Wei 	BWD2K = ODECLK = OUNE =	Dune.taxon × mat × 2> —	Charlen instaly	
00 2				20 observations of 1607 variables	😅 🔚 🖙 Import Dataset - 🖉 Clear 💿	Let-
Aaptolasma_americanaPilsbry1916	Aaptolasma brintoni Newman . Ross. 1971	Aaptolasma_callistoderma_Pilsbry1911.	Aaptolasma fosteri Newman . Ross. 1971.	Aaptolasma_leptoderma_NewmanRoss.	Global Environment - Q	
1 0	0	•	0	•	Data	^
2 0	0	•	0	•	Qagg 1687 obs. of 6 variables	
3 1	1	1	1	1	Species: Factor w/ 1687 levels "Aaptolasma_americana_(Pilsbry,_1916)",: 1	23.
4 0	•	•	0	•	Genus : Factor w/ 222 levels "Aaptolasma_[Arthropoda]",: 1 1 1 1 1 1 2 2	2
5 0	8	a	a	8	Family : Factor W/ 181 levels Aberrantidae: 36 38 38 38 38 38 38 108 10	5 10.
6.0		•	•	•	Class : Eactor w/ 35 levels [unassigned] Caerogastropoda , 09 09 09 09 09	11
7 0	0	•	•	•	Phylum : Factor w/ 13 levels "Acanthocephala": 3 3 3 3 3 3 3 3 13 13 13	
8 0		•	8	•	dune 20 obs. of 30 variables	
3 0	•	•	•	•	dune.taxon 30 obs. of 5 variables	
10 0	•	*	8	•	Omat 20 obs. of 1687 variables	-
11 0		•	•		values	
12 .					Ornod List of 10	
13 0					Species : num [1:20] 23 24 695 2 7 0 0 171 215 0	
54 0					species i new fatter, es es este este este este este este es	
15 0	8	8	8		Files Plots Packages Help Viewer	-0
16 0	8	8	8	•	🕼 💿 🤌 Zoom 🔍 Export+ 🔍 🗹 Clear All	8
17 0						
18 0			8			
10 0						
					0	1
					° °	
					8 -	
1					- 0 0	
•					· •	
Console C:/Users/Anastasis/Desktop/R fo	lifewatchers/Theodore Codes/codes for experime	ntal evaluation/datasets/ 🕫			0	
24.0000 58.3/4/	58.3/4/ 422./990 58.3/4/ 1400.9	9		^		
2.0000 58.3393	58.3393 0.0000 58.3393 116.6	8			• 7 °	
7.0000 48.3568	18.3568 341.6593 48.3568 338.5	0				
0.0000					S	
171.0000 47.3685	17.3685 139.5746 47.3685 8100.0	2			0	
215.0000 56.3205	56.3205 317.9504 56.3205 12108.9	2			6 -	
0.0000						
25.0000 62.0541	52.0541 448.7880 62.0541 1551.3	5				
5.0000 26.2913	26.2913 5.9012 26.2913 131.4	6				
0.0000					0	
215.0000 63.1470	53.1470 425.9681 63.1470 13576.6	1			6 -	
0.0000					0	
89,0000 53,2977	53,2977 628,2232 53,2977 4743,4	9				
144.0000 58.5764	58.5764 312.1711 58.5764 8435.0	0			0 100 200 300 400 500 600 70	a
Expected Inf	91.5376 91.5376				0 100 200 300 400 300 000 70	·
> view(mat)					Number of Species	
>				×		
🚍 🚞 ave 📂 🙈 🛛		🛃 😰 🚯 🕅				
						Sep-14

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.