A primer in Phylogenetic Biogeography
using the Spatial Analysis of Vicariance

J. Salvador Arias
CONICET, INSUE, Instituto Miguel Lillo y Facultad de Ciencias Naturales, Universidad Nacional de Tucumán, Tucumán, Argentina
jsalarias@csnat.unt.edu.ar

May 2011

Preface

In recent times, an increased amount of phylogenetic papers include a biogeographic chapter, even in some cases the objective is the biogeographic analysis rather than the phylogeny itself. This reflects the important place of biogeography in current research. Unfortunately, a closer look of most of this papers, reveals that an explicit methodology of the biogeographic analysis is lacking. This contrast with other parts of the methodology in which every step is fairly detailed (for example, the molecular methods, or the phylogenetic analysis).

My principal aim in this manual is to present in a compact (but I hope, complete) way the logic and methodology of the Spatial Analysis of Vicariance [AjE11]. But, incidentally, I want to stress out that biogeography must require a quantified approach. So even if the reader does not want to use the Spatial Analysis of Vicariance, I hope he/she acknowledge at the end the importance of providing an explicit methodogy on how the biogeographic analysis was done.

The main audience of this book will be systematists that want to use their phylogenetic results in a biogeographic framework. So it is expected that the reader is well familiar with the phylogenetic systematics (a good start is [SrB09]). An knowledge of some historical biogeographic topics will be also useful (for example [CjE03][Mj09]), although that is not crucial, I recommend the reader to check at least one of that references. First, there are a lot of subjects related that field, that will not be covered here. Second, an informed reader can contrast their previous knowledge, with the principles outlined here, and judge if the Spatial Analysis of Vicariance is the adequate way to deal with his/her own data.

As I want to give a close attachment between theory and practice, the most important part of this book will cover how to do this kind of analysis thorough a computer program, VIP [Aj10]. VIP is a free and multiplataform program that implements the Spatial Analysis of Vicariance, and provide a simple way to dealt with phylogenies coupled with geographic data. As far as I knew, there are few applications that do it in an explicit way (e.g. [KdL08]), or are only available for a single problem (i.e. phylogeography [LaL08][LpE10]). VIP is the only one that provide analytical tools to use these data beyond the visualization or phylogeography. With the included exercises, I hope that the reader will be able to use and understand all VIP features.

Also, I most chapters I include a “FAQ” section, based on some popular ideas (as I see in paper) or direct questions about the method from colleagues that I receive personally, or by mail. I will be happy to answer any question about the method in my e-mail, so eventually, I hope the FAQ section will be more complete.

There are several people that help me in several ways, with ideas to improve the method, discussion and encouragement. P. Goloboff and C. Szumik, are helping me with this project since its beginning (In fact, they give me the idea!), P. Hovenkamp, is the real mind behind the logic of the method, and he is always open to discuss about it. D. Casagranda, S. Catalano, and M. Mirande, provide always first hand help about anything. T. Crowe gives me the encouragement to start the writing of this manual, and the first draft was written during my visit to South Africa. Of course, it is almost certain that any good point here was made by some of them, nevertheless, they do not necessarily endorse any or all of the positions presented in this manual, and as well as all the errors, are my solely responsibility. Maps in screen captures are taken from NASA blue marble (http://earthobservatory.nasa.gov/Newsroom/BlueMarble/). I was funded by several institutions in several times, including CONICET (a doctoral fellowship), FONCyT (PICT 1314 to P. Goloboff), and The Willi Hennig Society (Mary Stoppes travel award).

1. Introduction

1.1 What is "phylogenetic biogeography"?

Biogeography, is the branch of biology that studies the geographic distribution of the organisms. This is a very diversified field (see for example [BjL98]). Among one of the subjects of biogeography is the so-called "historical biogeography" (e.g. [CrE03], [Mo09]) which try to cope the aim of biogeography from a systematic/phylogenetic point of view.

Hovenkamp [Hp97] observes that there are two main approaches to historical biogeography. The first one, he call "Earth history", and the second "Taxon history." Whereas in Earth history the objective is to understand the biotic history of Earth (through the simultaneous analysis of several and unrelated taxa); in taxon history the aim is to understand the history of a particular taxon.

Here, I equate the taxon history approach with the "phylogenetic biogeography." Phylogenetic biogeography is the name used to describe the intuitive approach developed by Hennig [Hw66] and Brundin [Bl66][Bl72] (see [CjE03]), which tries to understand the geographical history of a particular clade. Their main tools were the drawing of a cladogram over a map. The same idea was re-invented by Avise [AjE87] under the label of "phylogeography" and for populations as terminals (instead of species), that use exactly the same tools of Hennig and Brundin. Then *Phylogenetic biogeography* is the study of the biogeographic story of a particular clade under the light of its phylogeny and geographic distribution.

The Spatial Analysis of Vicariance, the approach described in this manual, is a quantitative approach for phylogenetic biogeography. As well as Hennig and Brundin it uses a cladogram and explicit maps of distributions of the terminals.

1.2 Barriers and disjunctions

Most methods in historical biogeography are focused in relationships among predefined areas. Methods for phylogenetic biogeography (like DIVA [Rf97] and DEC [RrE05]) also use predefined areas, although its objective is to infer an ancestral area rather than any particular relationships among areas.

Spatial Analysis of Vicariance in the other hand, focused on detection of disjunct distributional patterns. In biogeography, a disjunct distribution among two related taxa is know as "vicariant distribution" [NgP81]. Among sister species this is usually labeled as an "allopatric distribution." When sister groups have disjunct/vicariant/allopatric distributions, the disjunctions is associated with a barrier. Then looking for disjunctions, is a way to look for barriers [Hp97][Hp01]. The barrier is the causal factor that keeps both distributions disjunct.

Then, a barrier provide a causal explanation for a phylogenetic (cladogenesis) and biogeographic (allopatry) phenomena. It is important to remark that saying that the barrier is the causal factor/explanation is not the same than saying that *the formation* of the barriers is the cause of the disjunction. When a barrier pre-dates the cladogenetic event, it means that the barrier was crossed (a dispersal event). Nevertheless, is the difficulty to cross the barrier that maintains both sister groups disjunct.

1.3 The aim of Spatial Analysis of Vicariance

The main objective of the Spatial Analysis of Vicariance, is inherited from the objective of Phylogenetic biogeography: to infer the particular biogeographic history of a particular clade. But instead of looking for the ancestral areas of each node, it search for the disjunctions/barriers among sister groups.

At first, this objective seems strikingly different from traditional phylogenetic biogeography, which is to found the ancestral area. But any biogeographic explanation on phylogenetic biogeography is deeply attached to a barrier: either vicariance or dispersal require a barrier, either its formation, or a crossing opportunity. In both quantitative methods proposed until date [Ro97][RrE05], the important result is the change in the ancestral areas by dispersal or vicariance, that is, the barriers rather than the areas themselves.

But in both DIVA and DEC, the barriers are defined beforehand (i.e. the limits of the predefined areas). In the other hand, in Spatial Analysis of Vicariance, the barriers are the result of the analysis. Spatial Analysis of Vicariance is about discovering barriers.

1.4 FAQ

1.4.1 Spatial analysis of vicariance looks for disjunct distributions, that it means that it require that allopatric speciation is the only acceptable mean of speciation?

No, it doesn't. It assume that allopatric speciation is the only one that can be explained directly from the phylogeny and the distribution. By definition, sympatric speciation does not leave any particular geographical mark in a phylogeny. So, if sympatric speciation is a common factor in the speciation, then it means that it is not expected to produce a geographical explanation of the cladogenetic event.

What can be happen, if that in presence of large amounts of sympatry, it would be more difficult to found disjunct distributions, and possibly, that disjunctions will have low supports. In section 5 (specially 5.9) it is detailed how reconstructions are evaluated and how that is related with the objectives of the method.

1.4.2 Can spatial analysis of vicariance distinguish between "vicariance events" and "dispersal events"? Other methods do it.

Actually, the method itself can not differentiate among both events, it just detect the barrier. Note that methods that infer ancestral areas are able to do it, because the user, beforehand define which barriers are important (the limits of predefined areas) and which are not (the inside part of the areas).

So before going to a method that is able to infer those events, try to test how robust are these inferences to different barriers (i.e. different definition of areas), and how they are affected by the scale.

There are some particular cases in which a dispersal can be distinguished from a vicariant event. The obvious example, is when the taxon is undoubtedly younger than the barrier. Another instance of dispersal is when the barrier fall outside the zone of other barriers for the group, as in the case of a group in a continent (barriers inside the continent) with a descendant in an oceanic island (the barrier above the ocean).

See section (5.8) for a discussion on the use of barriers as results.

1.4.3 Can I infer ancestral areas? Other methods do it.

No, just the barriers. But there is no reason to despair! When a barrier is detected, it is possible to infer that in someway, the taxon is associated to this barrier (either it is in one side of the barrier, of the the ancestral distribution was split by the barrier). If several successive barriers are close together, then it will be certain than the taxon is around that barriers. Of course, such inference can not be exact.

But, take a look on the other methods that calculate ancestral areas. As in the preceding question,

test how good is the inference to changes in the area definitions, and the scale of the analysis. As here is the ancestral areas the important question, then it is also worth to examine how areas themselves can be defined, and how exact is such definition. A cursorial view of the uses of these methods in the literature will show you that in most of the cases, the ancestral area is a very crude and extensive area (like “Africa” or “Asia”), so the quality of the inference is not as precise as it sounds at first.

1.4.4 Why the method has the word “vicariance” in its title? After all, the method just seek for barriers, not “vicariance events”

The use of the name is for historical reasons. First, it was developed under the ideas of the “vicariance analysis” of Hovenkamp [Hp97][Hp01], then the name indicates the close relationship with his method. Second, in its original meaning “vicariance” just indicates disjunction [NgP81], it is a word about a pattern, rather than a process.

But actual usage of the word was changed to be almost in most circles to be equating vicariant distributions, with a “vicariant event” that creates the disjunction. I think is pointless to fight for a return to the original meaning, so in this manual, I will use “disjunct” or “allopatric” to refer to this distributions, without assuming any particular process on how the disjunction was formed.

1.4.5 How similar are modern phylogeography with phylogenetic biogeography?

In the section (1.1) I equate “phylogenetic biogeography” with the intuitive approach developed by Avise. Nevertheless, there is a lot of changes in the actual phylogeography as the initial approach of Avise was labeled as descriptive. This “new” phylogeography (e.g. [Kl09]) is strongly orientated to population genetics, so in most cases the explicit link between the geography and the phylogeny, is lost (geography is reduced to a label, or to distance matrices, or phylogeny is ignored). This approaches are completely distinct from the geography and phylogeny oriented approach defended here.

2. VIP

2.1 What is VIP?

VIP (from “Vicariance Inference Program”, for the reason of the name, see 1.4.4) is a computer program that implements the Spatial Analysis of Vicariance. It is a multiplataform, free and open program, that means that you can use it, share it, mix it either as an executable, or as source code. The only condition is that you must recognize and cite its original version [Ar10].

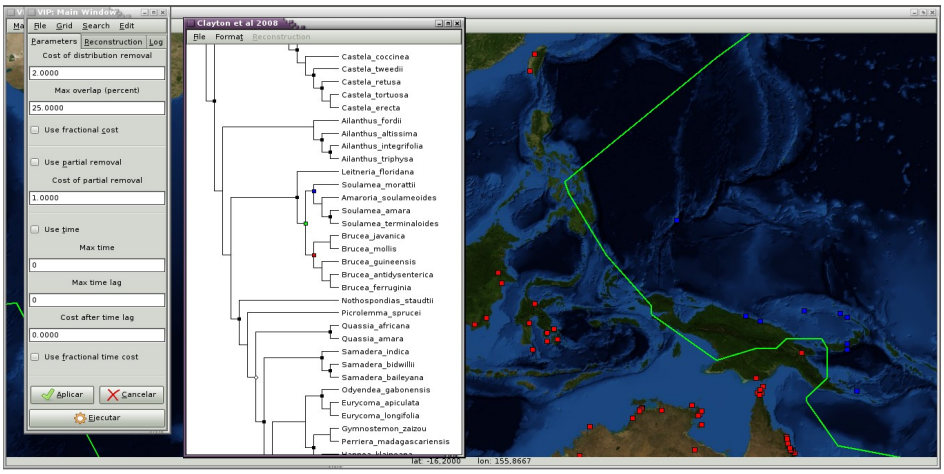

Fig 2.1 VIP interface. Map from NASA blue marble.

The program was developed under a Graphic User Interface paradigm. That means that it uses a windowing system to interact with the user (Fig 2.1). It is written in C, with the GTK+ library. So in order to run it (and compile it) you must have GTK+ installed on your computer. If you are a linux user, is almost certain that your distribution has already installed GTK+, windows users can download a recent version of the GTK+ runtime here: http://sourceforge.net/projects/gtk-win/, and install it before running the program.

After making sure you have GTK+ installed, you can run VIP. You can download the latest version of VIP (As well as the documentation and source code) on its web page: http://www.zmuc.dk/public/phylogeny/vip. It is useful to take at least initially a cursory view of the user's reference, so this will give you a familiarity with the main options of the program.

2.2 VIP's interface

Now, go to a computer and start up VIP. Also, from the VIP's web page, download a sample data file.

The interface of VIP is divided in three windows (Fig 2.1). The main window, the tree-view window, and the map window. Each one has its own menus, and its own suite of keyboard shortcuts.

The **main window** is the window to change the parameters of the program, and to see the basic results (i.e. statistics, a log) of an analysis. In the main window menu, use File to open the sample file that you just download.

The main window, has three flaps. The first one (**Parameters**) allows the user to change the basic parameters of the reconstruction. The second flap (**Reconstruction**) reports the actual cost of the viewed reconstruction, as well as the tree and node selected. The third (**Log**) stores the series of instructions made by the user, as well as some search reports, this can be useful to repeat/remember a previous analysis.

Always, take a look of the log, and save it to future reference, maybe to repeat some particular search, or just to have a list of what you do fits the materials and methods section of your manuscript. To save the log, go to menu Edit>Save log. If the log is to large to be useful (for example, you are making several and different experiments, so there is a lot of changes in parameter values and searches), you can clean it with the menu Edit>Clean log.

The **tree-view window** is the window in which the tree is displayed. When data is loaded, you can select any node, just by clicking on its tip (on terminals) or in the “coalesce” (on internal nodes). Using the arrow keys or the dragging the mouse, you can move the tree, and with the arroys and control key, you can resize the tree.

The **map window**, shows the geographic position of the records in the data set. The points that pertains to the selected node are shown in green. If you click with the mouse on any of this points, the program will show you a dialog box with the particular information for that record. With the mouse button hold, you can move the map. Also, there is an status bar at the bottom, that shows you the current position of the mouse pointer.

By default, VIP does not load any map image. You can open an map image from the menu Map and choose open. The most popular graphic formats can be read by VIP.

Now take your time to move between all menus and windows, pick options and look what happens, and check the description of the option in the user's reference. Check for keyboard shortcuts, so you can learn the ones that you like/use more. As more familiar will you with the program, the better will be your user experience.

2.3 FAQ

2.3.1 I see a Linux and a Windows version, Why not a Mac version?

I don't have direct access to a Mac machine, so I never develop on it. But if you have some experience on programing, or know any one that has it, then you can made your own mac version! Be sure to have any of the basic requirements to compile the program (see below), and check out for the particular caveats for a mac project under GTK+.

When you get your Mac binary, let me know, and post it on the web, so I will redirect any mac user to it!

2.3.2 Has VIP a collection of maps? Where can I found maps?

No, VIP comes without any map. But you can download it from many places on the internet. Make sure that the map that you download has an academic user license that allows you to use it for publication.

I find several high quality maps that you can download from Blue Marble NASA website (http://earthobservatory.nasa.gov/Newsroom/BlueMarble/). They are free and can be used for non-commercial and academic purposes.

2.3.3 What I need to compile VIP?

VIP is written in C, so the first thing that you need is a C compiler. I use the GNU gcc, but I expect it will run well with any compiler as long as you can found the gtk libraries for that compiler. I use a IDE that made most of the things for me, in particular I use Code::blocks, which is available for Linux and Windows as well (http://www.codeblocks.org/). In most distributions of Linux, gcc is already installed, otherwise, you can go to your preferred application manager and download it.

You need GTK+ developer library (which is different from the runtime library!), for Windows it can be download here: http://www.gtk.org/download.html. Linux users will surely need to download it also, it can be done easily with any application manager.

For the linking process in windows, apart of the basic gtk libraries, gdk-win32.lib, gdk_pixbux.lib and pango.lib must be included.

3. Cladograms and Maps: The data for Phylogenetic biogeography

3.1 “Life and earth evolve together”

This catchy phrase of Croizat is the motto of historical biogeography. My own interpretation of this lemma, is that the better way to understand the biogeographic phenomena, is to take a look on both the living organisms diversity, and the geomorphology where they live as well.

The way to understand the diversity of the living world is through a phylogeny, the evolutionary history of life. Thanks to the development of quantitative methods and its computer implementations, the project of looking for a tree of life seems now to be a reachable goal (e.g. [GoE09]). New technologies allows also the search for new sources of characters (ultrastructure, molecular data), and the phylogenetic trees are also an extremely simple form to describe the whole complexity of this character data [Fj79]. All of this advances made phylogenetic analysis to be one of the most rigorous fields of comparative biology, with a continuous examination of every detail, on the quality of methods, data sources and results.

Geomorphology of earth, has its own revolution thanks to the development of GIS and GPS technologies, as well as satellite and aerial data. The level of detail of this measurements allows the production of high quality maps. As part of the earth itself (“Life is the last geological layer” another Croizat quotation), biodiversity data are now highly accurate. Thanks to international efforts, the GBIF portal (http://data.gbif.org) has made available now several millions of biological records, and the number continue to grow.

Both pair of developments, in phylogenetics as well in geography and georeferencing has producing an enormous interest on biogeography, with the focus turn of “geophylogenies” [Kd10]. Unfortunately, most of this approaches were left just as a “data visualization” line of research. Analytical methods on the other hand, although use up-to-date methods and results of phylogenetic methods, dispose every of the developments of GIS and GPS geography.

To my knowledge, the only approach that explicitly uses a geographic location and a phylogeny in an analytical fashion is the random-walk approach used in phylogeography [LaL08][LpE10]. But that method only can take a single location per terminal. This is natural for their phylogeographic approach, but it isn't for a more general phylogenetic biogeography approach in which each terminal might represent several museum specimens.

Spatial Analysis of Vicariance, and its implementation (VIP), provides an analytical framework, that tries to fill the gap, between powerful phylogenetic applications, with detailed geographic information.

3.2 Phylogenies

VIP is not a phylogenetic analysis tool, so it took already defined trees from the user. This trees provide the phylogenetic framework for the analysis. There are many people that suggest that the use of a tree in biogeography assume that the phylogeny is “known without error” (e.g. [NjE08]). That is not true. The assumption is that the tree represent the better explanation of the data at hand. Of course, part of the support of the reconstruction is dependent on the support of the original phylogeny.

VIP uses an XML format to read and store trees. If you are a TNT [GpE08] user, you can transform a tree to XML using the macro “toxml.run” that is available in VIP website. Users of other programs might use Archaeopteryx (http://www.phyloxml.org/) to transform the data into phyloXML, and then open it with VIP.

Now open the file, using the menu File>Open in the main window. The tree will be appear on the

tree-view window. If the tree has a name, then the window will be renamed accordingly. The menu File (on the tree-view window) allows the user to change the name, or to save the tree in several formats, including SVG format, a highly flexible graphic format. The menu Format, allows to change the outlook of the tree. Play with it to choose your preferred tree display!

Remember that you can use the arrow keys holding the control key to change the size of the tree at your own desire. You can move on the tree with arrow keys, by dragging the mouse with the left button press, or with the mouse wheel. Also, position key Begin and End, and PageUp and PageDown can be used.

To select nodes (terminal or internal) click the left button of the mouse when the pointer is above the tip of the node, that is, at the end of horizontal lines. When selected a node is marked in green. If you hit the right button, a dialog box that allows the edition of the node name, and age (when available) is shown.

3.3 Maps and distributions

In conjunction with the phylogeny, the distributions are a key component of a phylogenetic biogeography analysis. VIP allows a simple edition and display of the distributional data. It is not a sophisticated GIS/database program. But has enough tools to provide a clean management of the data, and the use of detailed modern maps.

In the map window, you can open a map in raster format (i.e. an image), using the menu Map>Open. The requirement is that the map is on a isometric projection, that is each pixel has a constant size in terms of degrees. By default, the program assumes that it is a map of the whole world. You can change the limits of the map in the menu Map>Map limits.

To move on the map, use drag the mouse with the left button pushed. Another option, is to center the map on an specific localization, to do this use the menu Map>Center on.

Now it is time to add some records to the data! There are two ways to do it, first, I will describe the interactive way, which is the more simpler. It is useful when you have to add/edit few records on an already mature data set, or if your source of data are shaded area maps.

First, go to main window, and select the menu Edit/Hot mouse. Now if you open the menu Edit, the option house mouse must be checked. That means that the pointer of the mouse is “hot”, you can add or delete data with it. Depending on the desired accuracy, viewing a grid can be of help, in the menu Grid>Grid settings of the main window you can change the size of each cell grid (the unit measurement are degrees). To show, or hide the grid, go to map window and select the menu Drawing>Grid alternatively.

Select a terminal node using the mouse in the tree view window, or the node browser in the reconstruction flap of the main window. Only terminals can be edited.

To add a point, just click on the map with the left button holding the shift key. You can add as any points as you want. To edit/delete a point, click the right button of the mouse over the point while holding the ctrl key, the record browser dialog will be open, now with the option to save or delete activated, you can change manually the localization, or some metadata stuff (like a collection, a catalog number, a reference) and save it, or just click on delete and the point will be gone. There is no undo option implemented so, be careful!

Sometimes, you want to delete a bunch of records around a particular position (for example, a group of records with flipped coordinates, or records from an old definition of a species). That can be done by click on the left button and both control and shift key pressed. A dialog box will ask you the limits to do the deletion (the maximum distance on digress), so all the records within that limit will be deleted. As always, be careful with this option to avoid data loss.

When having records on the map, they might be so small, and difficult to see or click. You can change that with the menu Drawing>Record radius. Records that do not pertains to an active node are shown as empty boxes, sometimes, when there is a lot of records, this inactive records are hide the map, or you just want to see just the records of the actually selected node (this can make the drawing more faster), to do that go to Drawing>All records, at any time, you can change this option to see again all the records.

To finish edition, just go to main window, and deactive the hot mouse in the edit menu. You also might be interested in save the data. In the main window you can do it with File>save data.

The displayed map can be saved in any time, with the menu map>save map. You can save the whole map (the whole “world”), or just the portion of the displayed in the screen. Both in JPG format. You can also save the current selected node as KML to be open in any earth browser (like google earth).

3.4 Adding records from an external source

The preferable way to acquire data in VIP is to read some data from a some form of data base. First you must take the records and export them to a text file separated by tabs. The first row of the file must contain the headers of each column. The required fields are name or scientific name, latitude and longitude. Other names are also possible (for record metadata), and VIP can understand them. See the example data set.

There are a lot a records available from on-line collections and catalogs. Most of them can be downloaded as tab delimited tables, DarwinCore (GBIF standard), KML files (Google Earth, geographic standard), NDM xydata files [ScE03][Gp05]. All of them can be read by VIP. Also, you can read records from a previous VIP file. The requirement to be read is that the names on the records will be identical (case sensitive) to the name of the terminals (or included in a list of synonyms).

The menu option to read the records, is File>Feed records, in the main window. If any record is read it will be reported in the log flap. Unread records are stored in a pair of files called unmatch.xml and unfeed.xml. Unmatch.xml includes a list in alphabetic order of all names that can not be identified, whereas unfeed.xml keeps the unread records. The unmatch.xml file is designed so you can easily transform it into a synonym list.

For example, suppose that we have a terminal named Homo sapiens, after feeding the records, there are some unread records. The unmatch.xml file contains, among other things, this names:

```
<Folder>
      <name>Homo sapiens sapiens</name>
      <synonym> Homo sapiens sapiens</synonym>
</Folder>

<Folder>
      <name>Homo Sapiens sapiens</name>
      <synonym> Homo Sapiens sapiens</synonym>
</Folder>
```

Then you just only have to change the <name> field to the correct name, Homo sapiens

```
<Folder>
      <name>Homo sapiens</name>
      <synonym> Homo sapiens sapiens</synonym>
</Folder>

<Folder>
      <name>Homo sapiens</name>
      <synonym> Homo Sapiens sapiens</synonym>
</Folder>
```

Or, in a more compact way

```
<Folder>
      <name>Homo sapiens sapiens</name>
      <synonym> Homo sapiens sapiens</synonym>
      <synonym> Homo Sapiens sapiens</synonym>
</Folder>
```

After replacing the unread names with the correct names in the data set, you can save the file (preferable with another name, like “synonyms”) and then feed it again with VIP. This file has no records, but it contains a list of synonyms, so unfeed.xml (or the original data) can be read, this time, VIP will be able to identify the previously unidentified records, thanks to the synonyms.

Please note, it is preferable to read always the unfeed.xml. Otherwise, if some records of the original data set were read, they will be read again. This produce no harm in the results, just redundant data, but in some cases, these redundant data can be a very large set or records.

As always, after a successful data read, it will be advisable to save the data.

Try to read the XML appendix in the VIP user's reference, an take a look on a web browser, or a simple text editor (like notepad) of the XML files produced by VIP, try to understand what and how it is stored. This files are not intended to be edited by a human, but it is important that you will be able to understand, at least superficially, the data stored in it. This will give you an advantage at the moment of preparing data for your analysis.

Always check your data after a feeding season. It is possible (it happens in several public databases) that several records were on a wrong position (flipped coordinates for example) or well outside the “range” (based on unreliable source, or old taxonomic definitions, for example). Check section 3.3 to a brief explanation on how to edit data with the mouse.

3.5 On the importance of record data

Any reader familiar with taxonomy and phylogeny, known the importance of museum specimens to this fields. A thoughtful taxonomic revision or a phylogenetic analysis always include a section on the examined material. This is important, because it reveals the original data in which the analysis is based. It is possible to return to that specimens to challenge the initial interpretation of the data.

Historical biogeography is like the little brother of taxonomic revisions. Distributions and assignation to an specific taxon are key ingredients in both fields. Then, a biogeographic analysis must stands explicitly which is used to assign particular distributions to the data. In this way, going back to the source data, it will be possible to test this knowledge claims.

Unfortunately, this is not the picture on most modern biogeographic accounts. In most cases, distribution data was just cited as taken from “the literature” without an explicit citation for each

terminal. This is a consequence of the “area” approach to biogeography, in which a taxon is scored to a huge area, without taken into account the particularities of its distribution. Then, the data became opaque to the readers: a direct contrast of the data is not possible, as it is unknown.

The use of direct record data in VIP follows completely the opposite direction. As the method use record data, this means, that eventually, researches must made explicit statements on the data used in the analysis. The own infrastructure of VIP, that allows the storing of collection metadata encourages this behavior.

As in taxonomic revisions, historical biogeography research must include a direct reference to the material (museum specimens) in which the study is based.

3.6 FAQ

3.6.1 Can I read #Nexus trees on VIP?

No in the current version. But you can transform your files into phyloXML using Archaeopteryx (http://www.phyloxml.org/), phyloXML files can be open with VIP.

3.6.2 Can I edit my tree in VIP?

No, you can not change the topology of trees with VIP. There is a lot of programs that allows you to edit your trees, so I think this will not be problematic. In any case, I think that a future version of VIP might be able to edit trees.

3.6.3 Why VIP use everything in XML?

XML is a flexible way to store data for sharing on an internet framework. Although the files produced are usually big, its redundancy allow an efficient compression. The XML language is called a “human readable” format. This means that looking a file in a web browser (preferable) or in a text editor, it is possible to understand, in some way the content of the file. Also, the format is standardized so it can be read with any XML parser (it is also “machine readable”).

Also, XML can be extensible, that means that additional metadata, not initially included in the design can be introduced without destroying its initial meaning.

These possibilities are good for VIP. It allows researchers to share their data (e.g. as supplementary information) in a easy way, but also in a quite complete descriptive file. It also, allows the user to keep track of its own data, because metadata allow a precise identification of each record.

As far as I know, VIP is the first program that tries to store distributional data along with a phylogeny, for such a complex task, XML provide an intuitive way to store the data.

3.6.4 Can I export the data to other formats?

Yes, there are several formats to export the data, so VIP data can be read/used by another programs. You might be able to import files in KML format, so you can open it on Earth browsers (like Google earth) and most modern GIS programs. If you are interested on some form of endemism analysis [ScE02][ScG05], you can also save the xydata to be read in NDM [Gp05].

You can also will be able to save the record data as tables, or the phylogenies in phyloXML.

3.6.5 What map projections can be used on VIP?

Any isometric projection, that is, any projection in which the size of each pixel is constant in terms

of degrees (a cylindrical model of the earth). Note that the height and width might not be equal (although most maps use a 1:1 format).

3.6.6 I don't have a map. Can I see and edit my data?

Yes, but I suspect, it will be more challenging. By default VIP assumes that a 720x360 map of the whole world is load. So just change the limits of this “map” to match your preferred scale. Nevertheless, see (2.3.2) for an small and simple list of some available maps.

3.6.7 I have shaded maps instead of records. Can I use these data?

Yes, of course. I usually prefer museum specimens (see 3.5). But shaded maps produced by authoritative sources are also a valid source of data (this can be specially important for taxa in which collection of individuals is difficult or it is limited by ethical reasons, as for example, several mammal taxa). The important thing about this shaded maps, is that allows a contrast of the data: it is possible to look for specimens or to the field to check the distribution range. Also, giving the source of the map will be of valuable help, as data capture can be contrasted with more detailed maps or different sources.

3.6.8 Why you say that predefined areas are not testable? Are not they just like shaded areas?

Of course, you can use a similar argument used for shaded areas (3.6.7) to justify the use of predefined areas. But there is a difference between a shaded area and a predefined area. Shaded areas might be quite different for each terminal, the argument can be used just when a predefined area exactly match the shaded area. Otherwise, there are several parts of the predefined area that are not occupied by the terminal. Then there is not a direct relation between the “predefined area” and the distribution of the taxon: actually the distribution of the taxon is not completely congruent with the area, but this is hidden by the methodology itself, as you will be unable to identify which terminals are more congruent with the area that others (at least from the same data set). Moreover, in several studies, “small” overlap over neighbor areas was usually ignored.

This has a practical consequence, whereas in a shaded area, the area occupied by a taxon is more precise, but in general terms it will remain similar to the previous area, changing the scale of predefined areas does not guarantees that.

4. Biogeographic reconstructions

4.1 Disjunct distribution among sister groups

Until now, no analysis was made, we just have the data for the analysis. But before starting with the analysis is important to known, how the method work in detail.

Disjunct distributions were detected by eye since long time ago. Among a pair of species, a disjunction can be noted just by seeing its distributions in a map. The same inference is made for supra-specific taxa, like genera, tribes or families. In such cases the distribution of the supraspecific entity is the sum of the distribution of all of its descendants. It is not estrange to found a taxonomic revision which such kind of inferences. Even without any attempt of a biogeographic inference, distribution of supra-specific taxa is given in taxonomic revisions in that way.

Spatial analysis of vicariance is built on such tradition. The union of the distributions of the descendants of a node is taken as the distribution of a node [Hp97][Hp01]. This does not mean that this distribution is an ancestral distribution. Like in taxonomic revisions, this is just a pattern inference.

Once you have a distribution of a group, it can be compared with the distribution of its sister group [Hp97][Hp01]. If both distributions are disjunct, then you can hypothesize a barrier to explain the disjunction and the cladogenetic event. Otherwise, the distribution of the sister groups are overlapped. Sometimes the overlap is so small, compared to the distributions, that you might want to ignore that minimal overlap, and propose that the distribution of that sister groups is a disjunct one.

This is more or less the mechanics of a part of the method proposed by Hovenkamp [Hp97][Hp01].

4.2 Grids

The union of shaded maps of each node can be done by hand (e.g. [Hp01]) although it will be a long and error prone task in even modest data sets. Hopefully this can be implemented into a computer program, that can speed up the calculation, and eliminated the errors associated with repetitive works.

But shaded maps can not be worked automatically in a computer framework. This areas are continuous surfaces, and computer programs (in general) require discrete entities. Fortunately this problem was solved by greeks thousands of years ago: You can represent a complex area as a set of small squares, as more smaller the square, the better the representation. VIP takes this approach: it uses a grid to store the distribution of each node.

A grid cell is considered as a “presence” for a terminal, if there is a record inside the cell, otherwise, the cell is considered an “absence” for the terminal. In a internal node, the set of present cells is the union of the present cell in its descendants.

With this system of grids, overlap can be measured by just checking the intersection of both presence-sets in a sister group pair. If the intersection is void, then the pair have disjunct distributions.

4.3 Grid settings

Although the grid is a convenient and natural way to store geographic data, it is not without problems. It is important to decide the size and origin of the grid.

Ideally, a more fine grid will be always the better option. But there are practical limitations to this.

First, in an ideal data (i.e. one with “infinite sampling” like a shaded map), an small grid size will require a lot of memory. In the other hand, real data will be affected with sampling problems, as each record might will end with its own cell, so even in cases of overlap, might be accepted as vicariant because of the gaps in the distribution.

So, the first important thing is to made a balance between a fine scale (to give more detail), and a broader scale (to compensate incomplete sampling).

The second problem, is the grid origin. Take for example a cell grid of 5x5 degress, starting at point (0, 0). There are two record in the position one on (2, 6) and other on (2, 9). Then the first cell of the second row is scored as present. But if the starting point is (0, 2), two cells (the first cell of the first and second column) will be scored as presences. Note that the problem of the origin becomes less notorious with smaller cell sizes.

In VIP the grid always start at 90º N, 180º W. This does not mean than the problem of the origin is solved, it just simple to give a constant reference point. A solution, that can be also useful for the problem of the grid size, is to use a “filling algorithm” [Gp05][AlE08].

The filling algorithms used by VIP are fairly simple. The first (the default) is based on Manhattan's distance (city block distance), and it is called a Von Neumann's neighborhood. The second one, is based on Chevichev's distance (chess king distance), and it is called a Moore's neighborhood. Whatever is the case, the program uses a bound given by the user and if the distance of a cell with an observed record (using a defined distance measure) is less than the bound, the cell is scored as a presence.

Why this help to solve both scale and origin problem? Because it allow the user to set an smaller cell size, and then reducing the effect of the origin, but also, some cells around the observed cell are filled, and then reducing the effect of the sampling problems. For example, instead of using a cell of 5x5 degrees, it might be better to use cells of 1x1 degree, and filling around up to 2 cells, with a Moore's neighborhood. This will produce a set in which we have “cells” of 5x5 more or less centered on each record. Note that, if we assume that for a set of 5x5 cells, the central point is the place of the record, then the new cell set using 1x1 with a fill of 2, will have exactly the same area assigned to the 5x5 cell grid (although it will be stored in a larger memory chunk).

The grid settings can be changed/viewed with the menu Grid>Grid settings.

If you use a shaded map, it is preferable to set the scale beforehand starting to set the grids. Just one point per grid will be enough. Also, it is preferable to stick to that scale, and set the fill distance to 0 (as there is no reason to fill on shaded maps).

Whenever you have record data, try to explore the effects of changing the grid size/filling distance settings.

4.4 More on filling

Using the basic grid settings a the fill distance is set to a maximum. But in some cases, you might have a very restricted species, or group of species, so you might want to have an small or no filling at all.

In VIP you can define a filling distance for a particular clade. The condition is that the filling distance will be equal or smaller than the filling distance defined in the grid settings. When you change the filling distance of a node, the filling distance of all of its descendants will be changed to the new value.

To change the filling distance value of a node, you must click the left button of the mouse while holding the shift key, on the tip of the node to be modified. To restore the fill distance to default

settings go to Grid menu in main window and select set fill in all nodes.

4.5 Creating the grid

Now the grid settings are established, a grid can be created. This can be done with the menu Grid>Create grid in the main window. After the grid is created Grid parameters can not be changed, except for the a change filling distance on nodes (that require a re-calculation of the grid (Grid>Redo grid, on main window menu). Also, data can not be edited.

Results saved or read are always interpreted under the actual grid, so be careful to maintain the same settings between seasons. Of course, it will be an interesting experiment if a reconstruction under a different grid setting remains stable.

The grid can be destroyed with the menu grid>destroy grid on the main window. When the grid is destroyed, the results were removed from memory. So before destroying the grid, be sure that you already saved the actual results.

Once the grid is created, it is displayed in the map, you can hide/show the grid with drawing>grid menu on map window. Also, as the grid is created you might choose to display the cells instead of the records. This will be might be useful in a large data set, to reduce the time of repainting.

4.6 The default (OR) reconstruction

After creating the grid, a default reconstruction, using ORing is calculated automatically. The ORing process is the same process described on (4.2). This is a computational formalization of the ideas brings by Hovenkamp [Hp97][Hp01] and the taxonomic practice (4.1).

On the main window, go to reconstruction flap. It will show you the basic statistics of the reconstruction. For the moment, the only important one is the number of pairs of disjunct sister groups.

In the tree view window, you can identify the disjunctions because nodes in which their descendants are disjunct are displayed with a black square. This reconstruction can be saved as a SVG file (to be open it almost any vector image editor) under the menu File>Save tree graphic. If you select a node associated with a disjunction, then one descendant will be marked on blue, and the other in red. Look at the map, and records for each descendant will be showed in their respective color, so you can see the disjunction in distributional terms.

You might be interested in a barrier proposal. VIP propose some barrier based on Voronoi's tessellation. That means that the “barrier” is a line in which the distance to most closed points to each descendant is equal. The barrier can be activated with the menu Drawing>Barrier in the map window. If points are close, the barrier will be well defined, and with more probably coincides with the real barrier. As points get farther, the barrier will become just a line in the middle. In such case, maybe it is better to draw the “space” between the points as the barrier. This can be done selecting Barrier>Delaunay triangles. To return to Voronoi lines select the menu Barrier>Voronoi lines.

For some nodes, you might be interested in see the barriers on ancestors (for example, if you suspect that the group is product of a continuous dispersal). You can see them selecting in the menu Barrier>Barriers on ancestors in the map window. The barriers in ancestors will be shown as a thin broken line.

It is important to remark that the barriers drawn by VIP must be seen as an heuristic to detect the barrier, rather than the barrier itself. So the user must be not afraid to drawn his own barrier (as long as it takes into account the distributional disjunction).

As explained in section (3.3) the maps of each reconstructions, can be saved as images. The map

will be saved following the drawing status of grids, cells vs. records, or the barrier, that is, following the options activated on the actual display.

4.7 Measuring and taking into account the overlap

When using default parameters no overlap is allowed. So if you are seeing the records, in a disjunction every record will be red or blue. But if you see cells, maybe some cells appears as green (i.e. with overlap) how this can happen?

Although it use filled cells, VIP counts overlap just on the observed cells. To measure overlap, VIP compares the distribution of observed cells in one node against the active cells (observed and filled) on the other, the number of cells in the intersection divided by the number of observed cells of the first node is the overlap percentage. As the comparison on a pair of nodes is not reciprocate, the overlap is measured taken each descendant node as reference, and then selecting as the overlap associated to the pair the greater of both values.

As depicted in (4.1), it is possible to think that taking into account a partial overlap might be desirable. You can do it on VIP in the parameter flap of the main window. Put your preferred value on the max. overlap edit box.

The value acceptable to overlap varies. 0% means no overlap. In some cases, this might be to strict, because even if every node is perfectly allopatric, some records can fall in the same cell (this can be pronounced if there are also some filling. Then if you use filling, you also must take overlap into account.

On the other hand, higher values would produce that every pair of sisters will be counted as disjunct, even when the evidence for the disjunction is lacking (i.e. the high overlap). Values between 10-25% might be good in most cases. As always, if you are not convinced with this parameter, experiment what happens if it is modified.

4.8 Other things that can be done with a grid

You can use the grid created by VIP to other kind of analysis different from the spatial analysis of vicariance. You can save the grid as a table in the menu Grid>Save grid table of the main window.

There are different options to save the grid table, you can save the richness of each cell in a tab delimited table or in a KML to see it in Google earth or other earth browser, in that file, the values where scaled with respect of the highest value, and given a color based on an scale that goes from deep blue (for low values) through green (middle values) to deep red (higher values).

A phylogenetic diversity based on PD [Fd92], that assume an ultrametric tree, in which every node is always close to the terminals, and the value assigned to a cell with a single terminal, is the maximum number of node-levels in the examined trees. The values can be stored into a tab delimited table, or in a KML file, that use the same scale as in the richness file.

Also, you can save a matrix format for NDM, although I recommend to use the xydata format (which can be done in the File>Save data as, in the main window) instead of a table.

4.9 FAQ

4.9.1 If is not an "ancestral distribution", what is the distribution of a node?

In the Spatial analysis of vicariance, distributions associated with nodes are not ancestral distributions. Rather they are a prohibitive area for its sister. That is, more than detection of the ancestral area, is the detection of the area in which the taxon is absent.

The key of the Spatial analysis of vicariance, is to found the barriers among taxa, so distributions are more valuable in the sense that they provide the actual limits of a terminal, and then the sum of various distributions can be seen as the limit of a whole clade.

In this regard, even species with broad distributions would be useful to the analysis [Hp97], without the requirement of so called “biogeographic assumptions” [NgP81].

4.9.2 Is not the grid the use of “predefined areas” just at a finer scale?

Yes and no. Yes, because the use of a grid is imposed to the continuously distributed data. But as explained in (4.2), a grid is really a form to express an area surface. So it is not a predefined area in the sense used commonly in biogeography, as it is not expected that an species pertains to an area, rather, the area is defined by the taxon itself.

It is worth to notice that other methods that handle predefined areas can not handle grids, because they definition of areas is deeply attached to the method [AjE11].

4.9.3 Why I can not edit my data after the creation of the grid?

Although grids provide a natural way to deal handle biogeographic data, they also introduce a enormous waste of memory space, as most nodes (terminal and internal) are not so widely distributed, then a huge number of cells will be never used. This is not just a waste of memory, but also, they made the calculations slower. Then VIP compress the grid to minimize the number of cells in the grid, without losing any spatial information (you can see how many active cells has the data set in the log, just after creating the grid). Unfortunately, this compression depends on the state of the data, and then made some modifications very difficult to handle.

So, before you start your analytical part, be sure that you complete all the data edition. Of course, if you found that you need to change something, you can always destroy the grid, and go to edition (hot mouse) mode.

4.9.4 I have a polytomy and every terminal is allopatric among then, why it is not showed/counted as a disjunct distribution?

In VIP polytomies are always counted as a overlapped node. It will be desirable to detect a disjunction associated with a polytomic node. But the main problem is that in such cases the barriers will became ambiguous, as the barriers are defined just for a pair of taxa. As each possible resolution of the polytomy can produce their own suite of barriers, then the barrier associated with the node will became ambiguous.

Although I will prefer a formal solution for the problem, the actual treatment is consistent with the actual meaning of a polytomy: ambiguity of the resolution.

4.9.5 Why the barriers extends well beyond earth?

As explained in section (4.6), the barriers are based using a Voronoi tessellation algorithm. Also they are based on a cylindrical model of the earth. Nevertheless, the earth forms a continuous so the pacific basins are already connected. Then, the program take into account that in the calculation. As a consequence, a disjunct distribution that proposes a polar barrier (from pole to pole) require also another side of the earth. So in much cases, you see the “other” side barrier.

Although the real barriers are pretty restricted, it will be computed as if they split the whole world. Then, barriers sometimes goes beyond their logical place. Finally, the algorithm requires some imaginary points well beyond the study region, so lines in the borders can be completed.

Just remember, displayed barriers are an heuristic to detect the real barriers, rather than an accurate depiction of the actual barries.

5. Biogeographic reconstructions under an optimality criterion

5.1 Comparing reconstructions

When you got the OR reconstruction you can go directly to the disjunctions, and discuss their consequences. Now suppose than another researcher use your data and proposes that if we do not take one of your terminals into account, the remaining sister groups of the clade (say, 6 nodes) are now disjunct.

You can argue, reasonably, that your reconstruction is better, even if no node in the clade can be explained geographicly, because you take into account all data (this seems to be the initial position of Hovenkamp [Hp97]). The other researcher argue, also reasonably, that although some of the data is ignored, his/her reconstruction has a better explanatory power.

It is necessary to provide a measure to compare reconstructions. That is, an optimality criterion to rank different reconstructions, and select the one that provides the better fit among the reconstruction and the data.

5.2 The optimality criterion of Spatial analysis of vicariance

To propose an optimality criterion it is important to known what are the qualities of a reconstruction that we prefer. Intuitively, it seems that a reconstruction that has 3 pairs of disjunct sisters would be better than a reconstruction without any disjunction: In the first one, you can explain three cladogenetic events by the means of a barrier. Then the first thing that the optimality criterion must take into account is the number of disjunct distribution proposed.

Usually, it is better to think this in terms of a “cost.” If we assign a particular cost for sister group pairs that do not shown any disjunction (sister group with a high overlap among their distributions), then the reconstruction with the lower cost, is the one that will have more disjunct pairs of sister groups.

Under such optimality criterion, we can check what happens if the elimination of a node distribution will increase the number of disjunct sister pairs (as in [Pr04a][Pr04b] in parasite-host systems). This would be somewhat similar to the “maximum vicariance/coespeciation” approach [Pr04b] [Rf02].

Nevertheless, the basic problem, is that in some cases, you might eliminate a lot of node distributions just to increase an small amount of pairs with disjunct distributions. Also, it is evident that a distribution elimination is an *ad hoc* strategy, as it ignores data. So it seems reasonably to minimize the number of node distribution eliminations.

Then the optimality criterion, is a minimization of the (possible weighted) number of overlapped nodes, plus the (possible weighted) number of distribution eliminations.

Here, it is important to introduce another concept, the *j*-nodes [Rp94b]. A *j*-node is a node in which the distribution of all but one descendant was eliminated. Such nodes can not be counted: they behave like a terminal, as they only have “one” descendant. They, they are never counted as overlapped nodes (they do not overlap with anything!) and do not have a pair of descendant nodes that can be disjunct. A *j*-node is always the consequence of one (or several, in a polytomy) node distribution elimination, so in some way, when counting a node distribution elimination, you are counting the transformation of its parent node into a *j*-node.

5.3 The optimality criterion in VIP

VIP uses the optimality criterion described in the previous section. The cost of overlapped nodes is always set to 1. The user can change the value of node distribution elimination. Because the

elimination of a distribution will always produce a *j*-node (in a binary tree), it is only allowed to set the cost of node distribution eliminations to be equal to 1.0 or more. Otherwise, if you eliminate a node, it will produce a cost that is smaller than an overlapped node. For example, suppose you have two equally sympatric terminals, in this “tree”, the only sister pair is overlapped, so the cost is 1.0. If you allow a cost below 1.0 to distribution eliminations (say 0.5), a reconstruction in which any terminal is eliminated, and then transform the parent node into a *j*-node (and then not counted) will be preferred to the OR one (in this example, 0.5) even if no information is gained (no sister pair will be found!).

In a fully dicotomic tree, the default value of a node distribution elimination (1.00) will produce the same as a “maximum vicariance” approach: the cost of a node distribution elimination, and then, transforming the parent node into a *j*-node will have the same cost as if the node distribution is not eliminated, and the parent node is overlapped, then this is equivalent to minimize the number of sister pairs that can not be explained as disjunctions [Pr94b]. This identity does not hold for polytomic trees, as the association between the cost of a distribution elimination and an overlapped node is lost: several removals on a polytomy would be more costly, even if it produce a new disjunct sister pair (without losing any previous one). So if you want to be sure that your reconstruction produce the “maximum vicariance” you can set the node distribution removal to 0.00, that means that the program will ignore the cost associated with eliminations and just count the number of disjunct sister pairs.

The value of a node distribution elimination is set on the flap “parameters” in the main window, in the edition box with that label. Once you can set this parameter, be sure to apply it clicking on the button apply, in the bottom of the window. If you go to “log” flap, you will find that the change in the value is reported.

You can also take the overlap into account. That is, if you have two reconstruction with the same cost. In one reconstruction a disjunction without any overlap is proposed, that is absent in the second. The second reconstruction in turn has a disjunction absent in the first one, but with an overlap of the 20%. You might prefer the first reconstruction, after all, there is no overlap there. If you check the box use fractional cost in the parameters's flap, then the overlap on disjunct nodes will be take into account.

5.4 Experimenting with reconstructions

Once you have setting different cost to eliminations, you can experiment what happens in a reconstruction. The OR reconstruction, of course, does not change. But you can try your own reconstruction!

First, look again the “Reconstruction” flap in the main window. Now you can understand some of the statistics that resume the reconstruction. The cost box, gives you the cost of the actual reconstruction. In the OR reconstruction as it does not have any node elimination, this number is equal to the number of overlapped pairs of sister groups (In the case that you do not use the amount of overlap in the cost calculation). You only have the OR reconstruction in memory, you can copy it pressing the button copy. The copy will be stored on the reconstruction buffer. Now you will see that there is a report of a single reconstruction is in the buffer. Select the option Reconstruction buffer to move to the reconstruction buffer (with the reconstruction that you just created). If you have more reconstructions, you can use the navigation buttons to move through the reconstructions.

Go to menu Edit>Edit reconstruction in the main window. Now go to the tree view window. If you click the right button of the mouse holding the control key, the state of the node will change: to a node distribution elimination (symbolized with a white circle, if it descendants are overlapped sisters, or as a white square if it is descendants are disjunct), or as an active node. Take a look on the change of the reconstruction in the tree and the statistics of the reconstruction flap in the main window: maybe some new disjunct distributions will be found, the cost of the reconstruction, the

number of disjunct sister pairs and the number of node distribution eliminations will change. Go to parameters, and change the node distribution elimination cost (don't forget to press the apply button) and see what happens, how the cost of the reconstructions change.

5.5 Searching for optimal reconstructions

Manipulating the reconstructions can be fun, and also an useful way to get some particular reconstruction. But certainly, it is not an effective way to search an optimal reconstruction.

Ideally, an exhaustive search of all possible solutions will be desirable, but such approach will be very slow and unpractical for most real data sets: it is a combinatorial problem, so the number of possibilities might be fairly large.

To make a search go to Search>Heusitic search menu in the main window. It display a dialog with a lot of options. Here I will explain the most basic ones.

The searches implemented on VIP were based on a stepwise modification of each node. In the first versions of VIP a node distribution was eliminated, and then the new reconstruction was evaluated. The actual version do that but by default it also restore distributions eliminated, so it “flip” the nodes. This is a better alternative, as it is more exhaustive and more independent of the initial movements (as it can “go back”).

To set the proposing algorithm to just eliminate or flip the nodes, use the check box “flip nodes”. By default, it flip nodes, if the box is unchecked, just eliminations will be try. I recommend that the flip nodes will be used always (I left eliminations just for historical reasons).

To look for the different solutions there are two basic algorithms. First, you can try to flip/remove each node (at random), until a better solution was found, and then start again, if all nodes are checked without an improvement, the algorithm stops. This is the basic “hill climbing algorithm.” The second possibility is to check all possible flips/removals, and when all nodes are checked, the best reconstruction is choose, and start again, if all proposals produce no better solution, the the algorithm stop. This is the Page's algorithm [Pr94b]. Both algorithms has their own advantages and problems. The hill climbing is more faster, and less prone to particular local optima, but might fail to get in global optimal more frequently. On the other hand, Page's algorithm as it is more exhaustive takes longer times, it can be very effective in some data sets, but is highly probable to get trapped in a local optima, if there are some particular nodes that are always choose in the first rounds of the analysis.

To select among Page's and Basic hill climbing algorithm, use the check box “Page's algorithm.” When checked, Page's algorithm will be used, otherwise, the basic hill climbing will be used.

It might be desirable to have an algorithm with the more exhaustive nature of Page's algorithm, but less biased by first choices. It can be done, if instead of looking the whole tree, a group of nodes is analyzed. This is similar to the idea of the sectorial searches [Gp99] used in phylogenetic analysis. There are two modes actually implemented in VIP, in the first, a full sector search was performed, that is the data set is break up in sectors, an each sector was search independently. At the end, a complete round of flip/removal was performed with the whole data set. In the second, just a first group of nodes is used, and then a search with the full data search. Sectors have the advantage of being a little bit faster, and then allowing the use of more exhaustive searches.

The sector policy can be set with a group of select buttons: You can choose No sectors (the default), Start with a sector, or a full sector search. Also, you can set the size (in nodes) of each sector.

Then the search is done by examining a set of nodes in a random way. As it is possible that a particular random combination of nodes will unable to found a global optimum, and to ensure a

more extensive exploration of the solution space, it is important to do several replicates of the search. Each replicate follow the same rules, just start with another node order. This can be set by the edit box, number of replicates.

At the moment, VIP just keep a single solution for replicate. Sometimes, for a particular state, the data set has several equally optimal solutions. For the actual solution, no possible improvement can be made, but maybe, from the other equally optimal solutions it is possible, but as only one reconstruction is stored, then no other solutions will be examined. To cope with this problem, VIP can accept randomly an equal solution. To set the probability of accept an equally optimal solution use the edit box probability of accept equals, using a percentage as a probability. This will be useful mostly for Page's algorithm.

When search for reconstructions, the most usual alternative is to keep just the optimal (and different) reconstructions. For some reasons, you might want to keep reconstruction for all the replicates, being optimal or not. To do this check keep all reconstructions, by default only the best reconstructions are stored.

By default, when you make a new search, the program automatically delete all the reconstructions in memory. But you might want to keep previous reconstructions found (for example, you instead of doing 10000 replications do 10 runs of 1000 replications each one), click on retain current reconstructions, so previous reconstruction will be keeped (unless a better reconstruction score will be found).

Once you set your search options, click on accept to start the search.

5.6 The results of the search

After finishing a run, check the log flap in the main window. You will see the configuration of the search algorithms used, and also the number of stored reconstructions. It also indicates the number of hits on the best score, that is, the number of times in which an iteration end in the best score. Usually, if you have the same (or almost the same) number of hits and stored reconstructions, means that the data set is highly ambiguous (if there is a lot of reconstructions) or very hard (if there few reconstructions), and then it is preferable to increase the number of replicates used on the search. Beware when you have several hits to optimal and few reconstructions and using Page's algorithm without any form of sector (it might be a by that the algorithm was trapped in the same solution).

In the reconstruction flap of the main window, you can see the score of the actually selected reconstruction. The reconstructions from a search will be stored in the reconstruction buffer. If there are more than one, then you can use the navigation buttons to look at the different reconstructions. As doing in section (5.4) you can copy and edit different reconstructions.

You might want to save the actual results of your search. You can do it by going to the menu file>save results as in the main window. Remember this reconstruction is dependent on the actual grid and parameter settings, so, if in a later session you don't have the same settings, then the results will have different scores. To open a previous one result go to the menu file>open results in the main window. You only can do this with the active grid!

5.7 Several reconstructions

There are cases in which you found several different reconstructions. Depending on the data set and the length of the search, you might finish with hundreds of different reconstructions, which might be difficult to handle.

In such cases, it is possible to use a “consensus” of the reconstructions found. In VIP, a consensus is calculated with the reconstructions in the reconstruction buffer. If a sister group pair is disjunct in

all reconstructions in the buffer, then the pair is keeped as disjunct in the consensus. If a node distribution is removed in all reconstructions, then it is keeped as removed.

That is, the consensus reconstriction will show the reconstructions that are always common to all solutions. If you have only the best cost reconstructions in the reconstruction buffer, then the consensus will show the disjunctions that are supported by the data.

As this reconstruction is a resume of several reconstructions, and not a reconstruction on their own, no cost is calculated for it.

For the node distributions, the resulting node distribution is the intersection of all distributions assigned to that node in all reconstructions. As some records will be present in some reconstructions and not in others, then they are ambiguously assigned to the node, and showed in white. If you don't want to see them, uncheck the option draw>removed records in the map window menu. If there is a barrier to be calculated, it only will take into account the unambiguously assigned records.

Sometimes, you want to remove a consensus (for example, because you modify the reconstruction buffer). You can do it in the main window menu search>clear consensus. When you do a search, this is done automatically.

Also, you might have to clear the whole reconstruction buffer, with the menu search>clear. Or you might want to filter the reconstructions (for example, because you copy and edit some reconstructions, or change the settings, or keep all reconstructions during search), this can be done in the search>filter menu of the main window.

5.8 Barriers, again

There are some interesting experiments that you might want to check once the barriers were calculated. For example, you might want to see if the barrier on ancestral nodes were somewhat in the same region of the barrier calculated for the actual sister pair. This might be useful if you expect some dispersal (so the barriers will move in the “direction” of the dispersal), a radical change on the barrier geography (also associated with dispersal and subsequent diversification in the new land), or a similar barrier in an ancestral node (that might indicate an equivalent phenomena affects ancestral groups, for example, successive glaciations).

To do this, just activate the barrier in the menu draw>barrier on the map window, and select barrier>ancestral barrier menu on the same window. Ancestral barriers will be shown as dotted lines.

It is also possible to look for a common barrier in other nodes, although as presently implemented it is not well effective. This is, look for a pair of distributions that are coincident in terms of the disjunction and their distribution (as suggested by Hovenkamp [Hp01]). The problem with the actual implementation, is that it is possible that some taxa share the same barrier, but not their distributions, so, that potential coincidences will be ignored. Also, the calculation is very strict, not allowing a partial overlapping between the pairs (although, within pairs they are allowed according to the overlap parameter!). As the idea is that this common disjunctions might be simultaneous, then only independent nodes (i.e. the compared nodes are not in a descendant-ancestor relationship). Also the comparison is always done with respect to the actually selected node, so it is possible that a pair of incompatible nodes will be shown as common. This is just a try on the potential common barriers, rather than a search for them.

This option is available once the consensus is calculated (so if you only found a single reconstruction, you must calculate the “consensus” to do it) and selected. Once the consensus is selected, you can see how many disjunctions are common across the tree (or several trees, if you load more than one), in the log flap of the main window (reported as “supported disjuntions”). In the

tree view window, choose the menu reconstructions to see, all reconstructions, all the supported/common reconstructions or the reconstructions that are shared with the actually selected sister pair group. Note that in some cases, although the selected sister group pair is disjunct, there are no other disjunction that show the same barrier/distribution.

In the map window, if you select the menu barrier>common barriers, it will show you the barrier associated to the intersection of all potentially common barriers for the actually selected node.

5.9 On the biological meaning of the optimality criterion

Now, we have an optimality criterion, and a search procedure. But how biologically significant are the results? There are some special assumptions in the method?

Spatial analysis of vicariance is tightly coupled with an allopatric model of cladogenesis. Only patterns that are disjunct can be explained. This have a reason: this is the only speciation mode in which there is an strong association between the phylogeny and the geographic distributions (Here, peripatric speciation is taken as a form of allopatry, their difference is just in degree).

That means, that we want to search the maximum possible cladogenetic events explainable as disjunctions in the distribution. But this not require that allopatric speciation be the only possible explanation, nor than allopatric speciation will be common.

There are countless ways to explain a disjunct distribution: from plain allopatry, to invoking any process coupled with several extinction and dispersal process. Nevertheless, under such assumption, no method will be able to explain anything. So the method subscript here to a principle of parsimony, similar to the one behind phylogenetic analysis [Hw66][Fj83]: always assume allopatric speciation in the absence of contrary evidence.

This, in a similar fashion as phylogenetic parsimony [Fj83], does not assume that allopatric speciation will be frequent. If sympatric speciation is common, then it is expected that most of the distributions show high levels of overlap (sympatic speciation is a process independent of the geography of the distribution). In such cases, then, it will be difficult to the method to find allopatric distributions, which is adequate because the method only can explain cladogenesis in allopatry.

But it is important to say that failing to find allopatry, does not mean that sympatry will be the preferred explanation. In the spatial analysis of vicariance, any sister group that remains as overlapped, means that it can not be explained under the current reconstruction, no process can be associated with the geography of the distribution. So the method can not be used to measure the different degrees, rates or kinds of speciation processes, that will be an over-interpretation of the results.

In [Pr94b], distribution removal was connected with dispersal (host shift in co-especiation). This can be done because the method is constrained to a host phylogeny, then the removal is also associated with a destination node. Here, this connection can not be done, so it is possible to argue that a removal imply dispersal, but what is not clear, is who/when and the barrier across the dispersal event occurs. In most cases, the removals where ambiguous: they can be assigned to different nodes in different reconstructions. They are “the same” removal, in the sense that this removal is required to find some particular disjunction(s) in more basal nodes. In other cases, they are not ambiguos, but simply uninformative (that is, associated with a vastly widespread node distribution). Again, there is no explanation possible to them. In this sense, removals are purely an *ad hoc* strategy, and that is the reason for trying to reduce them.

5.10 FAQ

5.10.1 What is the better value for node distribution removals?

There is no “better” value for this. It depends, as explained in section (5.1) on how you value the reconstructions. If you want to seek for all disjunctions possible, without caring about the number or distribution removals, then using 1.00, or “0.0” (that is, just maximizing disjunctions, see 5.3) will be the best possibility. In the other hand, if you think that any removal is to be strongly penalized, high values, like 5 or 6 will be fine (although it will probably produce the same answer than the OR reconstruction).

I prefer an intermediate value, of 2. Under such cost regime, removals were accepted if at least a new disjunction will be found.

5.10.2 Can I use non-integer values for node distribution eliminations?

Yes, you can, but be careful, as it is explained the actual heuristic procedures made a change in the reconstruction, and check it. But with a non-integer set of costs, solutions will be harder to find, as the change on the reconstruction is step-wise. For example, using a cost of 1.5, is at least initially the same than using a cost of 2, because overlapping sister groups occurs in units! so an “intermediate” solution between 1 and 2 (which is the one expected) will be difficult to find with the current procedure.

I am experimenting with other searching algorithms, like simulated annealing, and I hope, this will be able to solve the problem (under the current implementation, the annealing is just slow).

5.10.3 How many replicates I will need for a search?

This is a question that must be answered to each data set independently. Small data sets, with few terminals, and few overlap, can be solved quickly with the default (100) number of iterations. A more challenging problem, might require 1000 or even 10000 iterations just to be sure that at least a good exploration of the solution space is explored.

Take into account always the number of hits found in your analysis, and the number of stored reconstructions. Few hits with few reconstructions, must of the time indicate that more iterations will be need.

Even, it is certainly possible that solutions for huge data sets, will be never discovered with the actual algorithms, so it might be important to develop new algorithms that might be able to found solutions for this problems in reasonable time!

5.10.4 Which is the better probability of accept equals?

This value is more or less experimental. From my own experience, I set the default at the value that give me the better results in my test data sets, that is about 50%.

Smaller values, are more stricter, so if the data set contains an optimality “plateau” (in which there are several reconstructions of the same cost, but just few of them can lead to a better reconstruction), it will be harder to find the optimal. On the other hand, higher values, will produce that almost all the time a different solution is recover, so instead of finding the escape route to the plateau, it might be just “run in circles.”

5.10.5 Which is the best size of a sector?

As with most questions in this section, this require experimenting with the actually analyzed data set. In general, the idea is that the sectors will not be huge (otherwise, they will not help to scape

form local attractors), but also, that they will not be very small (only few solutions will be selected and then the main part of the search will be the standard one, so is like no using sectors at all). My guess, is to use a number equivalent to a quarter of the number of terminals.

5.10.6 How many time will take the search?

Again, this is highly dependent on the data set size, not just on the number of terminals, but also, in the geographic spread and resolution of the grid: smaller grids will have more cells, so they take more time to evaluate each reconstruction. The increasing of time of the scale is quadratic, because moving from grids of say 1 degree, to 0.5 degrees, will increase the number of cells by 4.

Of course, increasing the number of replications will increase the time of the search. This increase is a bit more than linear, because if the data set include several solutions, then more time will be consumed on comparing the new reconstruction with the stored reconstructions.

If you want to check the time, my solution is to run just 10 or 20 iterations, and then try to estimate the time of a higher number of iterations based on this small run.

At the end, no bother much about the time, surely the analysis will be take less time that the time you need to create/curate/edit the data set! And try to make the runs in the time that you are not working: at the lunch break, in the night, the weekends, etc.

Epilogue

This is the basic knowledge to understand phylogenetic biogeography, and to use VIP. There are some other few things that VIP can do, but most of them are experimental. They are included in the program with the objective to allow the users to try them, but they are actually under investigation, or the algorithms will not able to produce adequate solutions. This options might be discussed in a future version of this primer.

But for the moment, I hope that the theory and practice included here will be enough to allow the users to be familiar with VIP, and to will be able to use it for their own problems. Maybe after reading all the manual, you can read again the most theoretical parts of the manual, now with the advantage of knowing a more completed picture of the approach.

Again, I want to stress out the importance of giving a good description of the procedures used in a biogeographic analysis, from how the data was acquired, the importance to provide a list of the specimens and/or distributions used, and finally how the data was analyzed. In this way, most of the parts of the analysis can be repeated in almost any place of the world, by any researcher.

Also, data transparency will allow the possibility to build up more complete and exhaustive data sets, without losing the whole previous work. Taxonomy has provided transparency of their data from a long time, and it is time that biogeographic analysis follows that route!

References

[Aj10] Arias, J.S. 2010. VIP. Program published by the author. Available at: http://www.zmuc.dk/public/phylogeny/vip/

[AjE87] Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C. 1987. Intraspecific phylogeography. *Annual Review of Ecology and Systematics* 18: 489-522.

[AjE11] Arias, J.S., Szumik, C.A., Goloboff, P.A. In press. Spatial Analysis of Vicariance. *Cladistics*. Doi: 10.1111/j.1096-0031.2011.00353.x

[AlE08] Aageseen, L., Szumik, C.A., Zuloaga, F.O., Morrone, O. 2009. Quantitative biogeography in the South America highlands—recognizing the Altoandina, Puna and Prepuna through the study of Poaceae. *Cladistics* 25: 295-310.

[BjL98] Brown, J.H., Lomolino, M.V. 1998. *Biogeography, (2nd ed.)*. Sinauer.

[Bl66] Brundin, L. 1966. Transantartic relationships. *Kungliga Svenska Vetenskaps Akademien Handling, Series 4* 11: 1-472.

[Bl72] Brundin, L. 1972. Phylogenetics and biogeography. *Systematic Zoology* 21: 69-79.

[CjE03] Crisci, J.V.C., Katinas, L., Posadas, P. 2003. *Historical Biogeography*. Harvard Univ. Press.

[Fd92] Faith, D. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1-10.

[Fj79] Farris, J.S. 1979. The information content of the phylogenetic system. *Systematic Zoology* 28: 483-519.

[Fj83] Farris, J.S. 1983. Logic of phylogenetic analysis. In: *Advances in Cladistics vol. 2*. (N.I. Platnick, V.A. Funk, eds.). Columbia Univ. Press, pp. 7-35.

[Gp99] Goloboff, P.A. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* 15: 415-428.

[Gp05] Goloboff, P.A. 2005. NDM/vNDM. Program published by the author. Available at: http://www.zmuc.dk/public/phylogeny/endemism/

[GpE08] Goloboff, P.A., Farris, J.S., Nixon, K.C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* 24: 774-786. Program available at: http://www.zmuc.dk/public/phylogeny/tnt/

[GpE09] Goloboff, P.A., Catalano, S.A., Mirande, J.M., Szumik, C.A., Arias, J.S., Källesjö, M., Farris, J.S. 2009. Phylogenetic analysis of 73060 taxa corroborates major eukaryotic groups. *Cladistics* 25: 211-230.

[Hw66] Hennig, W. 1966. *Principles for a phylogenetic systematics*. Univ. Illinois Press.

[Hp97] Hovenkamp, P. 1997. Vicariance events, not areas, should be used in biogeographical analysis. *Cladistics* 13: 67-79.

[Hp01] Hovenkamp, P. 2001. A direct method for the analysis of vicariance patterns. *Cladistics* 17: 260-265.

[KdL08] Kidd, D.M., Liu, X. 2008. Geophylobuilder 1.0: an ArcGIS extension for creating 'geophylogenies'. *Molecular Ecology Resources* 8: 88-91.

[Kd10] Kidd, D.M. 2010. Geophylogenies and the map of life. *Systematic Biology* 59: 741-752.

[Kl09] Knowles, L. 2009. Statistical phylogeography. *Annual Review of Ecology and Systematics* 40: 593-612.

[LaL08] Lemmon, A.R., Lemmon, E.M. 2008. A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic Biology* 57: 544-561.

[LpE10] Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A. 2010. Phylogeography takes relaxed random walk in continuos space and time. *Molecular Biology and Evolution* 27: 1877-1885.

[Mj09] Morrone, J.J. 2009. *Evolutionary Biogeography*. Columbia Univ. Press.

[NgP81] Nelson, G., Platnick, N. 1981. *Systematics and Biogeography*. Columbia Univ. Press.

[NjE08] Nylander, J.A., Olsson, U., Alström, P. Sanmartín, I. 2008. Accounting for phylogenetic uncertainty in biogeography: a bayesian approach to dispersal-

vicariance analysis of the Thrushes (Aves: *Turdus*). *Systematic Biology* 57: 257-268.

[Pr04a] Page, R.D.M. 1994a. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 43: 58-77.

[Pr04b] Page, R.D.M. 1994b. Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics* 10: 155-173.

[Rf97] Ronquist, F. 1997. Dispersal-Vicariance Analysis: A new approach to the quantification of historical biogeography. *Systematic Biology* 46: 195-203.

[Rf02] Ronquist, F. 2002. Parsimony analysis of coevolving species associations. In: Page, RDM. *Tangled phylogenies*. Univ. Chicaco Press. Pp. 22-64.

[RrE05] Ree, R.H., Moore, B.R., Webb, C.O., Donoghue, M.J. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* 59: 2299-2311.

[ScE02] Szumik, C.A., Cuezzo, F., Goloboff, P.A., Chalup, A.E. 2002. An optimality criterion to determine areas of endemism. *Systematic Biology* 51: 806-816.

[ScG05] Szumik, C.A., Goloboff, P.A. 2004. Areas of endemism: An improved optimality criterion. *Systematic Biology* 53: 968-977.

[SrB09] Schuh, R.T., Brower, A.V.Z. 2009. *Biological systematics: Principles and applications (2 ed.)*. Comstock Pub. Assoc.