

PhylPro

A Graphical Method for Detecting Genetic Recombinations in
Homologous Sequences

31 July 1998

Beta Version 0.8

Georg Weiller

An Introduction to the Program (draft)

written by An Van den Borre

weiller@rsbs.anu.edu.au

borre@rsbs.anu.edu.au

Bioinformatics Laboratory




Research School of Biological Sciences

Australian National University

GPO Box 475

Canberra A.C.T. 2601

Table of Contents

Introduction	4
System requirements	4
Installing the program	4
Methodological background.....	4
Standard for displaying commands and hyperlinks in this manual.....	5
Summary of steps to generate a Phylogenetic Profile.....	5
1. The Project (VOP).....	6
1.1 Creating a new Project	6
1.2 Importing sequences.....	6
1.3  The Project View or VOP View	6
1.4 Saving the Project.....	8
2. The Dataset.....	8
2.1 What is a Dataset?.....	8
2.2 Creating a Dataset	8
2.3 Composing the sequence set	8
3.  The Sequence View window	9
3.1 Sequence Residue field	9
3.2 Sequence Name field.....	9
3.3 Column field.....	10
3.4 Feature Field.....	10
3.5 Sequence View Preferences	10
Window	10
Sequence.....	10
Columns	11
4. Selecting, Including and Excluding Columns	12
4.1 What is a Column?	12
4.2 Selecting Columns.....	12
Manual selection	12
Using the Menu	12
Using Features.....	13
Using the Profile.....	13
4.3 Including and Excluding Columns from the Dataset	13
5. Defining Features	13
5.1 What is a Feature?	13
5.2 Creating Features.....	13
5.3  The Feature View	14
6. Generating a Phylogenetic Profile.....	14
6.1 What is a Phylogenetic Profile?	14
6.2 Profile Parameters	14
Distance Scores	15
Sliding window limits	15
Position restrictions	15
Scoring matrix.....	16
Correlation measure	16
6.3 Profile View window	16

Zoom Options.....	16
6.4 Interpreting the Phylogenetic Profile	17
Visualising the Informative Columns	17
Relationships	17
6.4 Navigation in the Profile View.....	18
Selecting a Sequence	18
Selecting a region	18
6.5 Profile View Preferences.....	18
Profile View	18
Show location	19
Snap	19
Summary line	19
Defaults	19
6.6 Maximising the Phylogenetic Profile signal	20
6.7 Printing a Profile View.....	20
Profile Printing Preferences	20
Colouring Sequences.....	21
Pasting a Profile to the Clipboard	21
7. Properties of objects in the Project.....	21
7.1 Sequence Properties	21
General	21
Links.....	21
Note	22
Sequence.....	22
7.2 Profile Properties.....	22
7.3 Feature Properties.....	22
8. Creating other Datasets within a Project.....	22
9. Navigation between Views.....	22
9.1 Navigation between Views of the same Datasets.....	23
Locating the current position.....	23
Locating the Analysis Region	23
Locating a range	23
9.2 Navigation between Views of different Datasets.....	23
9.3 Using the Window and View menus.....	24
Window menu	24
View menu	24
10. Saving the Dataset and Project (VOP)	24
11. Exporting Data	25
11.1 Exporting a Phylogenetic Profile Graph	25
11.2 Exporting Sequences	25
Positions	25
Character Mapping.....	25
Wordlist.....	25

Introduction

PhylPro is a computer program that implements the Phylogenetic Profile algorithm as discussed in Weiller 1998. Phylogenetic Profiles constitute a novel way of graphically displaying the coherence of the sequence relationships over the entire length of a set of aligned homologous sequences. Using a sliding-window technique, this method determines the pairwise distances of all sequences in the windows and evaluates, for each sequence, the degree to which the patterns of distances in these regions agree. This method is suited for exploring data consistently as well as detecting recombinant sequences.

The PhylPro program is built using the VOP, an acronym for Virtual Object Pool, a new object oriented database (Weiller in prep.). PhylPro uses the VOP for all its storage requirements i.e. user data as well as interior program data for a given Project. This has the advantage that the entire project is stored in a style file. The terms VOP and Project are interchangeable in PhylPro.

System requirements

Phylpro will run on the following systems: Windows NT, Windows 95 and Windows 3.1 (with win32s.exe installed). Note: Mac systems will need a Windows emulator (e.g. Virtual PC, Softwindows).

Installing the program

At the moment the program does not have a help menu, but a draft of the Introduction manual is included with the program. Windows NT

This Beta Version is shipped without installation procedure and needs to be installed manually. Create a subdirectory PhylPro for instance within C:\Program Files directory (for NT and Win95) or within C:\Programs directory (for Win 3.1) and copy PhylPro.zip to it. Unzip PhylPro.zip in the same directory using Winzip or Pkzip (this file). The following files will be present:

1. PhylPro.exe - The PhylPro program.
2. Phylpro.rtf - The PhylPro introduction documentation draft. This introduction can be read or printed using any editor that can read rtf formatted files.

You can start the program by double clicking its icon in the Explorer window (WinNT or Win95) or its file name in the File manager (Win 3.1). For more convenience we recommend that you put a shortcut icon to the Program on either your Desktop or in the Windows 95/NT start menu.

On the first execution of the program PhylPro creates an initialisation file (PhylPro.ini) in your windows directory which will keep track of the PhylPro customisations. Deleting this file will restore the original settings.

Methodological background

See Weiller, G.F. (1998) Phylogenetic Profiles: A graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.* 15(3):326-335.

Standard for displaying commands and hyperlinks in this manual

Examples:

File:New (i.e. Main menu option:Submenu option): Stands for “Open the main menu *File* and choose the *New* submenu option”.

Save: Any text from buttons or fields in the dialog windows or names of view windows

Sequence Residue field: Indicate areas in the windows that have not been explicitly named within the program.


1. THE PROJECT (VOP): Embedded hyperlinks. These guide you automatically guide to the various Chapter headings (Word 97 only). At any time when using these hyperlinks (in Microsoft Word) you can return to your original position in the document by using the ⇐ (back arrow button) in the Word toolbar or by pressing the <Alt> + Num 4 keys.

Summary of steps to generate a Phylogenetic Profile

1. Generate an optimal alignment for a homologous set of sequences. File formats accepted by PhylPro are NBRF/PIR, FastA, GDE and Genbank.
2. Run PhylPro and create a new Project (see 1. THE PROJECT (VOP))
3. Import your optimally aligned set of homologous sequences from an alignment file into the Project. Save the Project (see 1.2 IMPORTING SEQUENCES and 1.4 SAVING THE PROJECT).
4. Create a new Dataset and choose which sequences to include (see 2. THE DATASET).
5. Use a Sequence View window to select which regions of the alignment you wish to include in your analyses (see 2. THE DATASET and 4. SELECTING, INCLUDING AND EXCLUDING COLUMNS).
6. Optional: Specify Feature regions for identifiable areas (e.g. intron, exon, etc.) in the alignment (see 5. DEFINING FEATURES).
7. Generate a Phylogenetic Profile using the default parameters and examine it in the Profile View window (see 6. GENERATING A PHYLOGENETIC PROFILE).
8. Optimise the Profile by changing Profile Parameters, included Sequences, included Columns, etc. (see 6. GENERATING A PHYLOGENETIC PROFILE and 6.6 MAXIMISING THE PHYLOGENETIC PROFILE).
9. Save the Dataset with a unique name within the Project (see 10. SAVING THE DATASET AND PROJECT (VOP)).
10. Repeat from 4-9 to generate other Datasets if desired. Duplicate a Dataset to make slight changes to the Profile parameters and compare both (see 8. Creating other Datasets within a Project).
11. Use the results for publication illustrations (see 6.7 PRINTING A PROFILE VIEW and 11.1 EXPORTING A PHYLOGENETIC PROFILE Graph)

1. The Project (VOP)

1.1 Creating a new Project

To create a new Project (or VOP) choose *New* from the *File* menu (*File:New*), or click the  button on the toolbar, and select *Project* from the appearing popup menu. An empty *Project View* (*VOP View*) will be displayed. You must now import previously aligned sequences into the project.

1.2 Importing sequences

Once a new project is created, or an existing project is opened, the *Import Sequence* option on the *File* menu becomes available (*File:Import Sequence*).

Aligned nucleotide or amino acid sequences can be imported in the following formats: NBRF-PIR (*.pir), FastA (*.fa), GDE (*.gde), Genbank (*.gb)

When the option *File:Import Sequences* is selected a standard file selection dialog window appears in which you can browse your computer to find the desired alignment file. The *Files of Types* field allows you to choose between different file formats (*.fa, *.pir, *.gde, *.gb). Choose a file and click the *Open* button. A new window opens and specifies the number of sequences in the chosen file as well as the type of data (nucleotides or amino acids). It also allows you to specify several *Options*. You can click on the *Options* button to modify *Gap characters*, *Missing characters* and the *Genetic Code* translation table you wish to use (in the case of nucleotide data). Default parameters:

- *Gap characters*: -.~
- *Missing characters*: BbDdHhKkMmNnRrSsVvWwXxYy?
- *Genetic Code* (translation table): Universal

All sequences are selected by default. Click the *Select Sequences* tab to import only a subset of sequences from this file.

Click *Import now* when all parameters are specified. A summary table will let you know how many sequences have been imported, the number of *Residues*, the number of *Gap* characters, the number of *Unknown* characters and the *Total* number of characters in the specified set of sequences. Press *OK* to accept and view the raw data in the *Project View* (*VOP View*).

1.3 The Project View or VOP View

This window provides a summary of all the objects associated with the Project in tabular format and specifies various characteristics of these objects.

- *Acronym*: The short name of the object. For sequences this is the first 12 characters of the sequence name as specified in the imported alignment file. For Datasets, this is the name under which the Dataset was saved.
- *Type*: The Type of the object. PhylPro currently uses the following types of objects:

Nuc(6): Nucleotide Sequences. There are 6 types of characters specified for nucleotide sequences. These are four nucleotide symbols A (a), T (t and also U, u), G (g), C (c), - (gap), ? (unknown characters; i.e. all other characters).

Pep(23): Amino Acid Sequences. There are 23 characters specified for Amino acid sequences. These are ACDEGFHIKLMNPQRSTVWY, - (gap), ? (unknown) and * (stop)

SeqSet: A set of either nucleotide or amino acid sequences in a Dataset.

SeqFea: A list of sequence Features in a Dataset.

CorDat: Phylogenetic Correlation Data, i.e. a Profile in a Dataset.

DataSet: All data required to produce a Phylogenetic Profile. This includes a *SeqSet*, *SeqFea*, *CorDat* and *Included Positions*. This means that *SeqSet*, *SeqFea* and *CorDat* objects are embedded in a *DataSet* object and therefore normally not visible in the VOP View window.



- *Ref*: The Reference Count of each object. Various Datasets can include the same sequences. These sequences are present only once in the VOP. Sequences that are used (Referenced) by a Dataset cannot be modified or deleted from the VOP, unless the Dataset using the sequences is deleted first. The Reference Count shows how many other objects depend on a given object, whereby the first number specifies memory references (open Datasets) and the second number specifies VOP references (saved Datasets).
- *Modified*: Gives the last modified date and time that an object was modified. If the modified date or time differs from the Created then an * is placed at the end of the entry.
- *ID*: The unique ID of the object in the VOP.
- *Char(acters)*: Total number of characters (applies to Sequences only).
- *Mis(sing characters)*: Total number of missing characters (applies to Sequences only).
- *Created*: Gives the date and time that an object was created.
- *Gap*: Number of Gap characters present (applies to Sequences only).
- *Stop*: Number of Stop characters present (applies to amino acid Sequences only).

Most of the above characteristics are listed in the VOP View by default. Other characteristics can be added using the *View:Preferences* option, many of which are included for debugging reasons. They are not of importance to the user and not further described in the documentation.

The listing of the objects in the *Project View* can be sorted according to each characteristic by clicking on the respective characteristics label on the top bar.

Any additional information about the sequences that was available in the comment line of the original alignment file is still available for a selected sequence by double clicking a given sequence in the *VOP View* and selecting the *Notes* window (See 7. Properties of).

1.4 Saving the Project

At this stage you have created the Project and specified a set of sequences in a Dataset to analyse for recombination signals. It is recommended that you now save the Project and give it an appropriate name. Use the *File:Save VOP as* option (or use the  button) to save the Project in a desired directory. The extension *vop* will automatically be added to the given filename. Saving the VOP will also result in renaming of the present VOP window into the specified name. A VOP window is characterised by the icon  in the top left corner of the *VOP View* window.

2. The Dataset


2.1 What is a Dataset?

A Dataset combines all the data required to produce and examine a phylogenetic profile. Each Dataset specifies the set of Sequences and includes the Columns of the alignment, the Parameters to generate a Profile, the Phylogenetic Correlation data and the defined Features. A number of different Datasets can be created and compared within a Project.

There are three Views associated with each Dataset: a *Sequence View*, a *Profile View* and a *Feature View*. The name of the associated Dataset is given on the Titlebar of each View window.

The Views represent different ways of displaying aspects of the same Dataset and are linked (e.g. when you double click a certain part of a sequence in any View window the cursor in the other windows will move to that position (this is the default, this linking system can be turned off if desired, see 3.5 SEQUENCE VIEW PREFERENCES)).

2.2 Creating a Dataset

Use the *Dataset:New* option (Or use the  button) to create a Dataset.

2.3 Composing the sequence set

The first step after creating a Dataset is to select the set of Sequences included.

The *Compose Set of Nucleotide Sequences* dialog window allows you to choose a set of sequences. This can be done in two ways,

1. By using the mouse to select individual sequences from the *VOP Sequence Pool* (right listbox) and press the *To set* button to transfer them to the *Dataset Sequences* (left listbox). Holding the <Ctrl> key allows you to select more than one sequence at the same time; holding the <Shift> key will select all sequences between the first and the next selected sequence.
2. By pressing the *Select* button to use any of the available options to select the desired Sequences. The *Select* button options are:
 - *All sequences*: Allows you to select the whole set of available sequences.
 - *Find an Acronym*: Allows you to do a search for an acronym in the sequence acronym list

and select the sequences accordingly.

- *Find a Note*: Allows you to search for sequences that have a specific text in their *Note* fields. A *Note* is a text field of maximum 255 characters associated with a sequence. Sequences (and other objects) can thus be described in a more detailed way than the short *Acronym field* permits (see 7. Properties of).
- *Members of a Set*: This option is only useful when other Datasets have already been created and saved in the Project. It allows you to select a set of sequences that have been used in another saved Dataset. When selecting this option you are presented with a window of all existing Datasets to select from.

When pressing the *To Set* button any selected Sequences in the *VOP Sequence Pool* (right listbox) are included in the *Dataset Sequences* (left listbox). Once all sequences you wish to include in the Dataset have been transferred to the *Dataset Sequences* listbox, press *OK*. You are then presented with a *Sequence View* window.

3. **The Sequence View window**

The *Sequence View* window (*SeqView*) has four *fields* that each display a different aspect of the Sequences.

3.1 Sequence Residue field

The *Sequence Residue field* shows the sequences that are present in the Dataset and can be horizontally and vertically scrolled. On the right and left edge of the bottom bar of the window the numbers of the first and last visible Columns are displayed. When the *Sequence Residue field* is active (i.e. when you click in it with the mouse) the current cursor position is indicated by a flashing rectangle. When the field becomes inactive (e.g. click outside the *Sequence Residue field* with the mouse) the current position is indicated by a yellow rectangle.

The *Sequence Residue field* can be subdivided into 2 horizontal subfields by pulling the top margin line down with the mouse in a click and drag movement. It can be subdivided into two vertical subfields by pulling the left margin line to the right. All 4 subfields can be present at the same time and the cursor position is marked in all of them so it can be used as a reference point for navigation.

Once a profile has been generated the *Sequence Residue field* will respond to double clicking any position in the Profile graph by greying the sequence areas that were used to calculate the Phylogenetic Correlation for the clicked position. This response is a default setting that can be changed in the *Sequence View* Preferences settings (see 3.5 Sequence View Preferences and 6.4 Interpreting the Phylogenetic Profile).

3.2 Sequence Name field

The *Sequence Name field* is located to the left of the *Sequence Residue field*. It displays the acronyms of the sequences in the Dataset. Its user interface varies slightly from standard list boxes.

A sequence can be selected by double clicking its acronym. Holding down the <Ctrl> key will allow you to toggle the selection status of one sequence and holding down the <Shift> key will extend the selection (**Note:** the current sequence is not necessarily selected). The current sequence is marked by a thin frame around the sequence acronym. On the bottom bar for the *Sequence Name field* the number of highlighted sequences and the total number of sequences in the Dataset are given.

From this field sequences can be moved in position using the *Sequence:Move* option (or, alternatively, using the shortcut menu by clicking the right mouse button in the *Sequence Name* or *Residue fields*). With this menu the Sequences can also be given a colour (*Sequence:Set Colour* option) or be hidden/unhidden (*Sequence:Hide / Sequence:Unhide* options).

3.3 Column field

Immediately above the *Sequence Residue field* is the *Column field* with default + signs for each Column of aligned data units (nucleotides or amino acids). The *Column field* indicates which Columns are included (+) in or excluded (−) from the Dataset (see 4. SELECTING, INCLUDING AND EXCLUDING COLUMNS).

3.4 Feature Field

Above the *Column field* is the *Feature field*, which is empty when you create a new Dataset. Any Features that will be defined for the Dataset will become visible in this field (see 5. DEFINING FEATURES).

3.5 Sequence View Preferences

Use the *View:Preferences* option (or use the shortcut menu that appears when clicking the right mouse button in the *Sequence Name* or *Sequence Residue fields*) to set various viewing parameters to modify the *Sequence View*.

Window

- *React to others:* Click this box to allow the simultaneous navigation between different Views (see 9. NAVIGATION BETWEEN VIEWS and 6.4 NAVIGATION IN THE PROFILE VIEW).
- *Show Features:* Click this box to display the *Feature field* in the *Sequence View*.
- *Show column usage:* This box is checked by default. It determines whether the *Column field* will be present on the *Sequence View* or not.

Sequence

- *DNA:* *Sequence Residue field* displays nucleotide data with equal spacing between the nucleotides (option not available for protein sequences).
- *Coding:* *Sequence Residue field* displays nucleotide data showing the coding triplets (option not available for protein sequences).

- *Protein*: Nucleotide data are translated using the genetic code associated with each particular sequence and the *Sequence Residue field* displays amino acid data. (to change the translation coding table select the Sequence and use *Sequence:Properties* option, see 7. PROPERTIES OF).

Note: Once a Profile has been generated some of the Sequence parameters become unavailable. If you wish to make any changes to these parameters you have to delete the Profile (use *Dataset:Profile Remove* option), make the changes to the *Sequence View* Preferences and then regenerate the Profile.

Columns


- *Raw*: *Sequence Residue field* displays all included (+) and excluded (-) Columns of the original imported alignment.
- *Included*: *Sequence Residue field* displays only the Columns that are included (+) in the Dataset (excluded Columns are hidden).
- *Informative*: Option only available when a Profile has been generated. When checked the *Sequence Residue field* displays only Columns that are Informative i.e. that contribute to the calculation of the Profile.

Note: The type of Informative Columns is specified in the *Position restrictions* Parameter of the Profile parameter window (*Dataset:Profile Parameters* option) by selecting either *Variable* or *Parsimonious* Columns (see 6.2 PROFILE PARAMETERS).

- *Font*: Allows you to set the style and size of the nucleotide/amino acid font (current font is displayed to the right of this button).
- *Nuc-colours*: Nucleotide colour of the *Sequence View* can be set by nucleotide (default), by purine vs. pyrimidine, in monochrome or by sequence.
- *Pep-colours*: Can be set by chemical properties (default), hydrophobicity, monochrome or by sequence.
- *Display in blocks of 10*
- *Grey Analysis region*: This option is checked by default. It is only applicable when a Profile has been generated. When checked it will grey the positions of the Sequences that were used to calculate the Phylogenetic Correlation for a specific sequence and position in the *Sequence Residue field* or in the *Profile View* (see also 6.5 PROFILE VIEW PREFERENCES, SHOW LOCATION).
- *Store Defaults*: Allows you to store the present *Sequence View* preference settings as a default.
- *Restore Defaults*: Allows you to convert the present *Sequence View* window to the stored default settings.

4. Selecting, Including and Excluding Columns

4.1 What is a Column?

A Column is a position in a multiple sequence alignment. The Columns included in the analysis (Datasets) can be specified individually. Each Column is represented by a + or – sign in the *Column field* of the *Sequence View* window (see 3.  THE SEQUENCE VIEW window).

When composing the data for your Dataset (which will be used in the Profile) you may want to exclude certain Columns from the sequence alignment (e.g. introns, highly variable areas, areas with many missing characters, etc.).

The set of Columns that is included to calculate the Phylogenetic Correlation can be modified by first selecting the Columns of interest and then excluding or including the selection.

4.2 Selecting Columns

Manual selection

Click and drag movement with the mouse in the *Column field* area of the *Sequence View* window allows you to select a range of Columns (+ or – signs). Alternatively you can select one Column and use the <Shift> key to select all Columns between this and a subsequently clicked Column. Use the <Ctrl> key to toggle the selection status of Columns.

Using the Menu

Use the main *Columns* menu to display several other options e.g. *Columns:Select all*, *Columns:Select Special*, *Columns:Invert Selection*. The *Columns:Select Special* option offers the following choices:

- *Used in other Dataset*: Allows you to select the included Columns from another saved Dataset in the Project.
- *Used in Profile*: Allows you to select the Informative Columns that have contributed to the current Profile. This option is available even if you have not yet generated a Profile; in that case it will use the default settings for the Profile Parameters to determine which are the Informative Columns. Since either the Variable Columns or the Parsimonious Columns were used for generating the Profile (see 6.2 PROFILE PARAMETERS) it will be either of those two sets of Columns that will be selected. Alternatively, either of these two sets are also available in the following options below (i.e. irrespective of which setting was used in the Profile).
- *Variable Columns*: Select all Columns that are variable in nucleotide/amino acid composition.
- *Parsimonious Columns*: Select those *Variable Columns* that have a minimum of two occurrences of each nucleotide/amino acid (i.e. the Variable Columns with autapomorphies are excluded).
- *Missing Characters*: Select all Columns that have x number of missing characters, where x is defined by the user and can be a number between 1 and the number of sequences.

Using Features

Selection of Columns can also be obtained using the Features (see 5. DEFINING FEATURES). By double clicking a Feature in the *Feature View* window the Columns to which this Feature applies will become selected in the *Column field* of the *Sequence View* window.

Using the Profile

The same is applicable when selecting regions in the *Profile View*. A rectangular region can be selected in the Profile by using a click and drag movement with the mouse (see 6.4 NAVIGATION IN THE PROFILE VIEW). This region is then also highlighted in the *Column field* of the *Sequence View* window.

4.3 Including and Excluding Columns from the Dataset

Once a desired set of Columns is selected, exclusion of Columns can be executed by choosing the *Columns:Exclude Selection* option. This will convert the selected Column symbols in the *Column field* into – symbols. Alternatively the *Columns:Include Selection* option will convert them into + symbols.

These menu options are also available using the right mouse button when clicking in the *Column field* of the *Sequence View* window.

5. Defining Features

5.1 What is a Feature?

A **Feature** is a coloured, labeled rectangular field that is defined by the user and associated with a region in the sequence alignment.

5.2 Creating Features

Features can be created using the *Edit:Add Feature* option. A window opens in which you can specify the beginning and end numbers (in Row Position numbers) of the Columns to which the Feature applies. The maximum available number of Columns (equalling the total length of the alignment i.e. the Row positions) is specified at the top of the window. In the *Text* box you can specify an appropriate acronym for the Feature which will be overlaid on the Feature in all *Feature fields* of the Dataset. The *Comments* box allows for a more elaborate explanation but is only visible from the *Feature View* window. There are four options of *Feature Styles* that can be chosen from the options window and a range of *Colours* for each, as well as the possibility to customise your own colours.

This menu is also available as a shortcut menu when clicking the right mouse button in the *Feature field* of the *Sequence* or *Profile View* windows, and anywhere in the *Feature View* window.

A more visual way to specify a Feature is to select a region within the *Profile View* or select a range of Columns in the *Column field* of the *Sequence View* window (see 6.4 NAVIGATION IN THE PROFILE VIEW, SELECTING A REGION and 4. SELECTING, INCLUDING AND EXCLUDING COLUMNS). When subsequently clicking the right mouse button in the *Feature field* of the respective View the shortcut menu will now present the same Feature options window as discussed above but the selected area is automatically written into the *From... To...* fields.

5.3 The Feature View

The Features can be modified using the shortcut menus (clicking the right mouse button in the *Feature field*) or by using the *View:Feature* option. The *Feature View* summarises all used Features within a Dataset. It specifies their respective Column coordinates, acronym, *Type* and *Notes*. Each of these Features can be deleted within the Dataset or selected, cut, copied and pasted from one Dataset to another using the various *Edit* menu options when the *Feature View* window is active (or, alternatively, by clicking the right mouse button in the *Feature View*).

6. Generating a Phylogenetic Profile

6.1 What is a Phylogenetic Profile?

A Phylogenetic Profile is a graph of Phylogenetic Correlation measures.

For each position of a given sequence, distance data are calculated (by pairwise comparison) from an upstream window of aligned Columns and from a downstream window of aligned Columns. The correlation between these two sets of distance data is called the Phylogenetic Correlation. If the Phylogenetic Correlation for a particular position of a given sequence is very low this position is a likely recombination site in that sequence.

Only the Informative Columns will contribute to the calculation of the two sets of distance data since Columns that show no variation of data (uninformative Columns) will not contribute to the distance calculation (i.e. distances for such Columns will be 0 in pairwise comparisons).

Depending upon the data that need analysing you may wish to analyse the entire sequence only a part of the sequence (e.g. you may only want to analyse the exons). The process of how to select the Columns you wish to use has been explained in 4. SELECTING, INCLUDING AND EXCLUDING COLUMNS. **Note:** You do not need to exclude non informative columns explicitly.

Once you have specified the areas that you wish to use for the calculation of a Phylogenetic Profile you are ready to specify the Parameters to calculate a Phylogenetic Profile.

6.2 Profile Parameters

The calculation of the Phylogenetic Profile is influenced by a variety of parameters that define: the type of Informative Columns that are used, the type of data used, the size of the sliding window and the type of algorithm used to calculate the correlation of the distance data.

Phylogenetic Profile Parameters can be set by using the *Dataset:Profile Parameter* option. The following discussion goes through all the Parameter settings sequentially while a more pragmatic approach is presented in 6.6 MAXIMISING THE PHYLOGENETIC PROFILE SIGNAL.

Distance Scores

Specifies the type of data that are used for the pairwise distance calculations.

- *Score Nucleic Acids*: Default for nucleotide sequences, option not available for protein sequences.
- *Score Amino Acids*: If nucleotide sequences are used this option will translate the Included Columns of the Dataset to amino acids and calculate distances from these data.

Note: Columns that have been excluded from the Dataset before a Phylogenetic Profile was generated do not participate in this translation, e.g. excluded columns will be removed prior to translation. This flexibility allows you to exclude e.g. introns from the analyses and still maintain a correct reading frame.

- *Treat gaps as differences*: By default gaps do not contribute to the distance calculations. Ticking this box will result in a difference being scored each time a gap is present in one of the pairwise compared sequences.

Sliding window limits

Specifies the size of the upstream and downstream windows of aligned Columns by defining the number of *Differences* or *Comparisons* that are used to determine the window width.

- *Differences* (default 10): Makes the specified window size dependent upon the number of differences found for each pairwise comparison (i.e. the number of differences used (as specified) will be the same for all sequences but for different pairwise comparisons the positions that contribute to the comparison will differ).

Note: This option is new and not described in Weiller 1998.

- *Comparisons* (default 40): Makes the specified window size dependent upon the (specified) number of aligned Columns that are compared (i.e. the alignment Columns used in the comparison will be the same for all sequences but the number of differences per pairwise comparison may differ).
- *Unlimited*: The window will be as large as the remaining upstream and downstream ends of a given sequence and all Columns are used.

Position restrictions

Of the Columns that have been included in the Dataset only the Informative Columns will actually be used to calculate the pairwise distances. This parameter allows you to specify the type of Informative Columns.

- *Variable*: All Columns that have more than one type of nucleotide/amino acid
- *Parsimonious*: Only those Variable Columns that have a minimum of two occurrences of each nucleotide/amino acid (i.e. Columns with autapomorphies are excluded)

Scoring matrix

Specifies the scoring matrix used to calculate the pairwise distances.

Correlation measure

Specifies the Correlation measure used to express the Phylogenetic Correlation between the upstream and downstream distance data.

- *Correlation* (default): Documentation to be completed.
- *Manhattan distances*: Documentation to be completed.
- *Bay-Curtis distances*: Documentation to be completed.
- *Rank-correlation*: Documentation to be completed.

Save Graph between sessions: If checked the graphics data (Phylogenetic Correlation) will be saved together with the Dataset. Otherwise only the Profile Parameters are saved. This option ensures that the program does not have to regenerate the Profile before displaying it. This can save time when the Dataset is large or you are working on a slow speed computer (see also 10. SAVING THE DATASET AND PROJECT (VOP)) however, it may result in big file sizes.

6.3 Profile View window

Save the Profile Parameters by clicking *OK*. The Phylogenetic Profile can now be calculated by using the *Dataset:Profile Generate* option. Display the Phylogenetic Profile by using the *View:Profile* option.

The *Profile View* window has two fields: The *Feature field* (a mirror of the *Feature field* in the *Sequence View* window) and the *Profile field* which shows the Phylogenetic Profile.

The X axis of the Phylogenetic Profile indicates the Column position in the alignment. When looking at the unzoomed Profile this starts at 0 and ends with the last position of the longest sequence in the alignment. The Y axis gives the Phylogenetic Correlation on a default axis of -1 to 1. The green line indicates the mean of Phylogenetic Correlation of all sequences. Its presence on the graph can be switched off using the Profile Preferences (see 6.5 PROFILE VIEW PREFERENCES, SUMMARY LINE)

Zoom Options

Using the *View:Zoom* option you can zoom into any region of the Profile. When no area is selected on the Profile the area you wish to enlarge can be *Specified* in XY axis coordinates. Alternatively, an area can be selected in the Profile using a click and drag motion with the mouse. This selected area is then automatically enlarged using the *View:Zoom:Selection* option in the same menu. To return to a previous zoomed area use the *View:Zoom:Previous* option and to return to the full Profile use *View:Zoom:All*.

The *Zoom* menu is also available using the right mouse button anywhere in the *Profile field*.

6.4 Interpreting the Phylogenetic Profile

See also: Weiller, G.F. (1998) Phylogenetic Profiles: A graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.* 15(3):326-335.

Recombination signals appear on the graph as areas of low Phylogenetic Correlation, visualised by single sharp-pointed downward peaks in the graph.

The accuracy of the Phylogenetic Correlation slightly decreases towards the edges of the graph. This is due to the reduction in size of one of the compared (upstream or downstream) windows that are correlated when the point of comparison moves closer to the each edge.

Visualising the Informative Columns

To interpret a recombination signal it can be useful to look at the area of the sequence alignment that was used to calculate the Phylogenetic Correlation for the peak position in the graph (i.e. those pairwise comparisons that contributed to the distance calculations in the upstream and downstream window). This area can be viewed as follows:

- *Profile View*: Use the *View:Preferences* option and select the *Show Location, Analysis Region* option. Press *OK*. Then double click any position in the Phylogenetic Profile graph: the selected Sequence will be shown in red and the other sequences will be partly blue. The blue areas indicate the Analysis Region of each sequence that contributed to the Phylogenetic Correlation calculation for the clicked position of the selected (red) sequence.
- *Sequence View* (providing the *Grey analysis region* box is checked in the *Sequence View Preferences*, see 3.5 SEQUENCE VIEW PREFERENCES): Execute the steps discussed above and then look at the *Sequence View* window. The blue areas shown in the *Profile View* correspond to the grey areas in the respective sequences of the *Sequence View*. This enables you to find those Informative Columns that contributed to the pairwise distance calculations for the clicked position in the Profile graph. The clicked position itself will be visible as a yellow rectangle in the *Sequence Residue field*.

Relationships

The relationship between the two adjacent sequence segments (the total sequence areas upstream and downstream from the clicked position) is expressed as % differences. These numerical data can substantiate the evidence for recombination at a certain position in a certain Sequence. To show these data for the previously clicked position (e.g. the lowest point of a recombination signal peak) click the **Relationships** option of the **View** menu. A list of relationships between the current sequence upstream/downstream segments and all other Sequences that are present in the Dataset is then displayed.

This list is subdivided into the two adjacent columns (upstream and downstream) listing the % difference of each segment of the clicked Sequence with each of the other sequence segments. This list could clarify which of the other Sequence segments are closest to each of the recombined segments in the clicked Sequence.

6.4 Navigation in the Profile View

Selecting a Sequence

A sequence in the *Profile View* window can be highlighted by double clicking a line of the graph at any position. This will result in two changes in the *Profile View*.

Firstly the Sequence that was clicked will become highlighted (red) and a Crosshair indicates the selected position.

Secondly, information about the highlighted Sequence will appear on the bottom bar of the program window. This information includes (from left to right): the number of selected positions (one in this case) and the total number of available positions, the number of the clicked sequence in the Dataset together with its acronym, and the Column number of the clicked position. The Column number depends on the settings of the Profile Parameters and whether the sequences have been translated: *VI* indicates Variable Informative Columns, *VI(t)* indicates Variable Informative Columns (translated), *PI* stands for Parsimonious Informative Columns, *PI(t)* stands for Parsimonious Informative Columns (translated) and the *Raw* number gives the reference number of the (nucleotide) Column to the original imported alignment in the Project alignment. Which of the *VI* or *PI* numbers will be displayed depends upon the *Profile* Parameter settings for *Position Restrictions* (Variable or Parsimonious).

Simultaneously, the *Sequence View* window will display the clicked position (yellow rectangle) in the left top corner of the *Sequence Residue field* and highlight the appropriate sequence Acronym in the *Sequence Name field*.

Selecting a region

A region of the Phylogenetic Profile can be selected by using the mouse with a click and drag movement to highlight a rectangular region in the graph. Holding down the <Control> or <Shift> key will ensure that the entire height or width of the Profile is selected. The bottom right area of the program window then displays the total of positions that are selected. Simultaneously, the selected sequences will be highlighted in the *Sequence View*. The last Column of the selected rectangle will be visible in the *Column field* as well as within the *Sequence Residue field*.

This selection process is convenient when you wish to exclude regions from the Phylogenetic Profile calculation (see 6.6 MAXIMISING THE PHYLOGENETIC PROFILE SIGNAL).

6.5 Profile View Preferences

Some of the options of the *Profile View* Preferences have already been discussed. Below is an overview of all the available options.

Make the *Profile View* window active (click in it) and select the *View:Preferences* option. It will display the following viewing parameters:

Profile View

- *React to others*: This box is checked by default. It allows the simultaneous navigation in the

Sequence or *Feature Views* (see 9. NAVIGATION BETWEEN VIEWS). For example when a position of a sequence is clicked in the *Sequence View* that sequence will become highlighted (red) in the *Profile View*. When this option is not selected the *Profile View* will not react to mouse movements in other Views

- *Show Features*: Check this box to display the *Feature field* in the *Profile View*.
- *Graph: Show Informative Columns Only*. When this box is checked the Profile view will be scaled to visualise only the Informative Columns that were used to generate the Profile (i.e. Variable or Parsimonious).
- *Smooth*: Smooths the Profile graph by averaging several adjacent Phylogenetic correlation values.

Show location

- *Crosshair*: The Phylogenetic Correlation of a clicked position in the Profile graph will be highlighted by a Crosshair pointing to the exact position on the X and Y axes of the graph.
- *Analysis Region*: The upstream and downstream windows that were used to calculate the Phylogenetic Correlation for a clicked position will be highlighted on the Profile graph (blue).

Note: The size of the blue area on the graph is dependent upon the settings of the *Dataset:Profile Parameters*. If a number of *Comparisons* was specified in the *Limit Sliding Window* by option e.g. *Comparisons* 40 (default) each sequence will have an equal length highlighted in blue. If, however, a number of *Differences* was specified then the length of the blue area will vary for each sequence, dependent on its similarity with the selected (red) sequence (see also 6.2 PROFILE PARAMETERS).

- *Off*: This switches the *Crosshair* and *Analysis Region* options off. The result is that only the clicked sequence is highlighted in red.

Snap

- *To minimum*: Clicking on a certain position in the Phylogenetic Profile graph will move the cursor to highlight the position that has the minimum value in that section of the graph (this facilitates finding the exact tip of the downward peaks).
- *To maximum*: Clicking on a certain position in the Phylogenetic Profile graph will move the cursor to highlight the position that has the maximum value in that section of the graph.
- *Off*: Switches the snap *To minimum* / *To maximum* facility off.

Summary line

Show: The summary line (green by default) shows the mean of all Profile Correlations. Its presence on the Profile graph, colour and line width can be set here.

Defaults

- *Store Default*: Allows you to store the present Profile Preference settings as a default.
- *Restore Default*: Allows you to convert Preference settings of the present *Profile View* window to the stored default *Profile View* Preference settings.

Printing options are also available from this menu but are discussed in 6.7 PRINTING A PROFILE VIEW.

6.6 Maximising the Phylogenetic Profile signal

A good way to explore your data is to use the default parameters to generate a Phylogenetic Profile and then explore the various Profile Parameters to maximise the signal. There is no universal way of maximising the signal(s) and the results will depend highly upon the data that you are analysing. The following steps may be useful:

- Use the *Score amino acids* option (*Dataset:Profile Parameter* option) when dealing with coding sequences.
- When highly conserved regions are present in the sequence the Phylogenetic Correlation is best calculated by setting *Limit Sliding Window:Differences* (in *Dataset:Profile Parameters*). This will increase the number of Informative Columns that are used and make sure all the pairwise comparisons use the same number of Differences (see 6.2 PROFILE PARAMETERS).
- If the Phylogenetic Profile has many downward peaks (i.e. areas of low Phylogenetic Correlation indicative of possible recombination sites) in various sequences the individual signals can be singled out by removing all sequences that show these peaks (use the sequence selection method discussed in 6.4 NAVIGATION IN THE PROFILE VIEW and then exclude the sequences using right mouse button shortcut menu in the **Sequence Name field** of the **Sequence View**) and then reintroducing each Sequence separately (i.e. creating several new Datasets which each have one of the recombination sequences reintroduced, see 8. Creating other Datasets within a Project)

6.7 Printing a Profile View

Profile Printing Preferences

Click the *View:Preferences* option while the *Profile View* is active and select the *Printing* tab. This dialog window allows you to set some printing settings for the *Profile View*.

- *Profile Parameter*: A summary of the Profile Parameter settings used to calculate the Phylogenetic Correlations will be printed at the top of the graph when this option is selected.
- *Sequence Features*: The Features will be present on the Profile View print if this option is selected.
- *Add title line*: When this box is checked the program will query the user, when printing the Profile, for a title comment (limited to 255 characters). This title will be aligned to the right margin of the Profile on the print.
- *Rulers*: The presence of *Horizontal* or *Vertical* rulers can be switched of or on
- *Font size* of the rulers can be set in 1/10 pt.
- *Store Defaults*: Allows you to store the present *Printing* settings as a default.

- *Restore Defaults*: Allows you to convert the present *Printing* settings to the stored default settings.

Use the *File:Print* option to print the Profile.

Colouring Sequences

Each sequence can be given a user defined colour on the Profile print. Select one or more sequences in the *Sequence View* or in the *Profile View* (by selecting a small area). Then select the *Sequences:Set Colour* option. This will allow you to specify a colour for the sequence(s) from a palette.

Note: To remove the green summary line from the Profile graph use the *View:Preferences* option and deselect the *Show Summary Line* option.

Pasting a Profile to the Clipboard

The Profile can be copied to the Clipboard and then pasted into your favourite Graphics Editor Program to edit the print. Each Profile line is copied as an object.

Use the *Edit:Copy Profile* (or <Ctrl> + C) option to copy the Profile to the Clipboard.

Note: Phylogenetic Profiles can consume a large amount of your computer's memory when copied to the clipboard. It is recommended to release this memory as soon as possible by copying another small object to the clipboard.

7. Properties of objects in the Project


All objects in the VOP have Properties. The Properties dialog window has several pages (tabs). The type of page available for an object will depend upon the type of object.

Documentation incomplete.

7.1 Sequence Properties

At any time the Properties of a sequence can be viewed and some settings can be manipulated. Select a Sequence in the *Sequence Name field* by double clicking its Acronym. Then use the *Sequences:Properties of ...* option to display a window with the following information:

General

Lists the Acronym of the selected Sequence; the date and time the Sequence was *Created* and *Modified*; *Reference Count*; a unique *ID* for the sequence; and the *Object Type*. For definitions see 1.3  The Project View or VOP View.

Links

Can be used to find the Datasets that refer (link) a sequence. These datasets need to be deleted before a sequence can be modified or deleted.

Note

This window shows the name of the file (and where it was located on your computer) from which the Sequence was imported and the Notes that were provided within this file with the Sequence. This window can be edited.

Sequence

- *Genetic Code*: Allows you to set the coding table to be used for translation of the Sequence to: *Universal* (default), *Yeast Mitochondria*, *Mamalia Mitochondria*, *Drosophila Mitochondria*, and *User Codes 1 to 4* (not yet implemented).
- A summary of the Character Types in the Sequence shows: *Unambiguous Characters*, *Ambiguous Characters*, *Gap Characters*, *Stop Characters*, and *Total Characters*.

7.2 Profile Properties


Documentation incomplete.

7.3 Feature Properties

Documentation incomplete.

8. Creating other Datasets within a Project

Sequence subsets of existing Datasets are often analysed in order to compare them with each other. In such cases it is often quicker to duplicate an existing Dataset and then exclude certain unwanted Sequences (or change certain Profile Parameters) than to start a New Dataset from scratch. This ensures that defined Features do not have to be recreated and it also facilitates the introduction of slight changes in the Profile Parameters without losing the previously saved Dataset and Profile.

Use the *Dataset:Duplicate* option (or the  button) to create an exact copy of your current Dataset. A new *Sequence View* will appear which is identical to the previous Dataset. Then remove/add sequences from it by:

- selecting the unwanted sequence(s) in the *Sequence Name field* and using the *Sequences:Remove* option.
- using the *Sequences:Add/Compose* option, to go back to the *Compose Set of Sequences* window, where you can specify the Sequences that need to be added or removed.

In other situations you may prefer to start afresh. In that case use the *Dataset: New* option and proceed as explained in 2. The Dataset. If you wish to use the defined Features from a previous Dataset in the new Dataset use the *Feature View* to select those Features and the *Edit:Copy/Paste* option, to import them into the new Dataset.

9. Navigation between Views

A high quality navigation system is available between the various View windows that are open in the program.

9.1 Navigation between Views of the same Datasets

Navigation between the *Profile*, *Sequence* and *Feature Views* within the same Dataset is enabled by default (see 3.5 SEQUENCE VIEW PREFERENCES and 6.5 PROFILE VIEW PREFERENCES). Providing the *React to Others* option (default) is selected for all open Views then the following system is in place:

Locating the current position

Using a double click with the mouse in the *Sequence* or *Profile View*.

- When double clicking a Phylogenetic Profile line at a given position in the Profile graph the *Sequence View* window will also display that position (yellow rectangle) in the top left corner of the *Sequence Residue field* and highlight the appropriate sequence acronym in the *Sequence Name field*.
- When double clicking a nucleotide/amino acid in the *Sequence Residue field* the appropriate sequence (red) and position (Crosshair by default) will be highlighted in the *Profile View*.
- When double clicking a Feature in the *Feature View* the range of Columns corresponding to this area will be highlighted in the *Column field* of the *Sequence View* and the *Sequence Residue field* will display this area entirely or partially (when the area is too large to be shown).

Locating the Analysis Region

- If the *Sequence View* Preferences window (see 3.5 SEQUENCE VIEW PREFERENCES) has the option *Grey Analysis Region* selected then double clicking a position on the *Profile View* will colour the Analysis Region blue and at the same time show this region as a greyed area in the *Sequence Residue field* of the *Sequence View*.

Locating a range

Using the click and drag mouse movement in the *Sequence* or *Profile View*.

- When selecting a region of the Profile graph the acronyms of the sequences that are present in that region become selected in the *Sequence Name field* of the *Sequence View* and the Columns of that region become selected in the *Column field*. The selected area of the sequences will also be displayed in the *Sequence Residue field* (unless it is too large to be displayed and then the last Column of the selected area will be visible in the *Sequence View*).

9.2 Navigation between Views of different Datasets

Profile, *Sequence* and *Feature Views* of **different** Datasets can be open at the same time in the program. The inter-Dataset navigation system between these windows is similar to the intra-Dataset navigation system i.e. clicking a position/area in one View window of Dataset A will locate that position in another View window of Dataset B. This facilitates, for example, the location of a given sequence present in two different Datasets, comparing of a Phylogenetic Profile area of a certain Feature in two different Datasets, etc.




Note: The selection of the Analysis Region is not copied to other Datasets.

9.3 Using the Window and View menus

Window menu

Individual Views of the Datasets can also be opened using the *Window:Add View* option which then presents a choice window where Sequence, Feature list or Phylogenetic Profile can be specified. To organise the various views on your screen the option *Window:Cascade* will cascade all Views on the screen. The option *Window:Tile Horizontally* will display all Views in full width underneath each other with the active window at the top, while the *Window:Tile Vertically* will display the Views in two columns on the screen. *Window:Close All* will close all windows.

View menu

Using the View menu all the available *Views* of a Dataset can be displayed or generated. The *Sequence View* is displayed when using the *View:Alignment* option or pressing the  button on the toolbar. The *Profile View* is displayed by using the *View:Profile* option or pressing the  button on the toolbar. The *Feature View* is displayed by using the *View:Feature* option or pressing the  button on the toolbar. The *VOP View* is displayed using the *View:VOP* option.

10. Saving the Dataset and Project (VOP)

The set of Sequences, the Included Columns and the defined Features are all part of a Dataset and hence will be saved when the Dataset is saved.

When saving the Profile Parameter settings only (default option) the program will recalculate the Profile graph each time a previously saved Phylogenetic Profile is reopened. This saves disk space and recommended when you are working on a relatively small Dataset since the graph is calculated in a few seconds.

- *View:Store* option will allow you to store a Phylogenetic Profile within a Dataset before the Dataset is saved. When saving the Dataset (*Dataset:Save* option) the profile is saved as a set of Profile parameter settings.

Saving the Profile graph as graphics data is recommended when you are working with large Datasets or have a low memory capacity and the calculation of the graph takes several minutes rather than seconds.

- Select the *Dataset:Profile Parameters* option and tick the box *Save graph between sessions* to save the graphics data. Then save the Dataset (*Dataset:Save* option). The Phylogenetic Profile is now saved as graphics data (which means settings of the *Profile View* Preferences, including a zoomed view) and when you open the Dataset the program will not have to recalculate the Profile and is able to display it immediately.

By default the program will always show a query window if you try to close a Dataset or Project (VOP) without saving the changes. Remember that saving a Dataset is not the same as saving the VOP and you will have to save the VOP as well before exiting the program. **FAILING TO SAVE THE VOP WILL RESULT IN THE LOSS OF ALL MODIFICATIONS!**

11. Exporting Data

11.1 Exporting a Phylogenetic Profile Graph

The *View:Profile Preferences* option settings and the settings specified on its *Printing* tab will determine how the Profile is exported or copied.

A Phylogenetic Profile can be exported using the *File:Export* option. Two formats are available: *Enhanced Metafile* (*.emf) and *Windows Metafile* (*.wmf). These enable importation of the graph in programs such as Microsoft Word and Powerpoint.

Alternatively the *Profile View* can be copied to the clipboard and pasted by using the *Edit:Copy Profile* (or <Ctrl> + C) option.

11.2 Exporting Sequences

File:Export option allows you to export your sequences. Nucleotide or protein sequences can be exported in the following formats: *NBRF/PIR*, *FastA*, *Mega*, *Phylip*, *Intelligenetics*, *Nexus*, *GDE*, *Genbank*. Choose a format and press *OK*.

The *Export Sequences* window summarises the specified *File Name*, chosen *File Format*, *Number of Nucleotide* and/or *Amino acid* sequences. By default the sequences will be exported as seen in the *Sequence View*. Press the *Options* button to specify other export options:

Positions

- *Entire Sequences*: Sequences will be exported in full length.
- *Region*: Only the specified region of the Sequences (give Raw Column numbers) will be exported.
- *Included Columns*: The Included Columns specified in the active Dataset will be exported. This facilitates exporting certain areas of the alignment (e.g. combined coding regions).

Character Mapping

Allows the user to specify which symbols will be used (in the exported file) for specifying *gap* characters, *unknown* characters, *stop* characters and *pad* characters.

Tick the *Translate Nucleotide Sequences* box if the sequences should be translated using the sequence specific translation tables.

Wordlist

- **Sequence Name field**: The *Sequence Name field* is located to the left of the *Sequence Residue field*. It displays the acronyms of the sequences in the Dataset.
- **Analysis Region** (of a position in the Phylogenetic Profile graph): Upstream and downstream window of the Sequence alignment that is used to calculate the Phylogenetic Correlation for that position in the Profile graph. See pages 12, 18, 20.

- **Column:** A Column is a position in a multiple sequence alignment. The Columns included in the analysis (Datasets) can be specified individually. Each Column is represented by a + or – sign in the *Column field* of the *Sequence View* window. See pages 10, 11, 12, 13, 14, 15, 16, 19, 25.
- **Column field:** The area above the *Sequence Residue field* in the *Sequence View* window. The *Column field* indicates which Columns are included (+) or excluded (–) from the Dataset. See pages 2, 10, 11, 12, 13, 14, 19, 25.
- **Dataset:** A Dataset combines all the data required to produce and examine a phylogenetic profile. Each Dataset specifies the set of Sequences and included aligned Columns of those sequences. The Parameters to generate a Profile, the Phylogenetic Correlation data and the defined Features will also be unique to a specific Dataset. See pages 5, 8, 9, 10, 12, 13, 14, 16, 17, 18, 19, 24, 25, 26.
- **Excluded Columns** Columns that are excluded from the Dataset and will therefore be excluded for generating a Phylogenetic Profile. They are indicated by a – sign in the *Column field* of the *Sequence View* window. See page 16.
- **Feature:** A coloured labeled rectangular field that is defined by the user and associated with a region in the homologous sequence alignment. Features are displayed in the *Feature field* of the *Sequence* or *Profile View* windows, and in the *Feature View* window. See pages 2, 3, 5, 8, 10, 11, 13, 14, 17, 20, 23, 24, 25, 26.
- **Feature field:** Is the area that illustrates the Features. It is located above the *Column field* in the *Sequence View* window and is mirrored above the Profile in the *Profile View*. See pages 14, 17, 28, 29.
- **Feature View:** Accessible through the *View* menu, *Features* option. This View summarises all used Features within a Dataset. It specifies their respective Column coordinates, acronym, Type and Notes. See pages 13, 14, 25.
- **Included Columns:** Are those Columns of the alignment that have been included in the Dataset and will therefore be used for generating a Phylogenetic Profile. They are indicated by a + sign in the *Column field* of the *Sequence View* window. See pages 16, 26, 27.
- **Informative Column:** A Variable Column which is used to calculate the pairwise distances which will be correlated to calculate the Phylogenetic Profile. The type of Informative Columns can be specified by the user in the Profile Parameters to include all Variable Columns or only the Parsimonious Columns. See pages 3, 11, 13, 15, 16, 18, 19, 20, 21.
- **Parsimonious Column:** A Variable Column that has a minimum of two occurrences of each nucleotide/amino acid (i.e. no autapomorphy or single occurrence). See pages 11, 13.
- **Phylogenetic Profile:** A graph of Phylogenetic Correlation measures. See pages 2, 3, 4, 5, 7, 10, 15, 16, 17, 18, 19, 20, 21, 24, 25, 26, 27.

- **Phylogenetic Correlation:** The datapoints of the Phylogenetic Profile graph. For each position in a given sequence (of an optimal alignment of homologous sequences) distance data are calculated (by pairwise comparison) from an upstream window of aligned Columns and from a downstream window of aligned Columns. The correlation between these two sets of distance data is called the Phylogenetic correlation. If the Phylogenetic Correlation is very low and stands out in the Phylogenetic Profile graph as a downward peak this position is a likely recombination site in that sequence. See pages 15, 16, 17, 18, 20, 21.
- **Profile View:** Accessible through the *View:Profile* option. This View shows a graph of all Phylogenetic Correlations for all positions in all sequences of the Dataset. See pages 2, 3, 5, 12, 19, 20, 21, 22, 25, 26.
- **Project View (VOP View):** Accessible through the *View:VOP* option. The View that lists all the objects present within the Project in a tabular format. See pages 2, 6, 7, 8, 23, 26.
- **Raw Column numbers:** Column numbers as given in the original Sequence alignment imported to the Project (i.e. within the Dataset this is the sequential number given to a Column when all Included and Excluded Columns are counted). See pages 11, 14, 19, 27.
- **Selected Columns:** Are the columns that are highlighted at any particular moment in the *Column field* of the Sequence View window.
- **Sequence View:** Accessible through the *View:Sequence* option. This View of the Dataset is composed of four fields: the *Feature field*, the *Column field*, the *Sequence Residue field* and the *Sequence Name field*. See pages 5, 8, 9, 10, 11, 12, 13, 14, 17, 18, 19, 21, 22, 24, 25.
- **Sequence Name field:** The *Sequence Name field* is located to the left of the *Sequence Residue field*. It displays the acronyms of the sequences in the Dataset. See pages 2, 10, 19, 21, 23, 24, 25.
- **Sequence Residue field:** Is that field of the Sequence View which displays the sequences that are present in the Dataset. See pages 2, 5, 9, 10, 11, 12, 18, 19, 24, 25.
- **Variable Column:** Columns that are variable in their nucleotide/amino acid composition, i.e. contain at least two different characters other than gaps or missing. See pages 13, 16.
- **View:** There are three Views associated with each Dataset, each representing a different way to display its data: Sequence View, Profile View and Feature View.
- **VOP:** Is an acronym for Virtual Object Pool, a new object oriented database (Weiller in prep.). Although any type of object can be stored in a VOP, relatively large objects like sequences, distance matrices, etc. are handled particularly efficient. PhylPro uses the VOP for all its storage requirements i.e. user data as well as interior program data for a given Project. This has the advantage that the entire project is stored in a single file. The terms VOP and Project are interchangeable in PhylPro. See pages 2, 3, 4, 5, 6, 7, 8, 9, 17, 23, 26.