GeneDirectory-A new tool for genetic fingerprint analysis

Introduction

Phylogenetic studies often generate vast quantities of data, a sizeable proportion of which is genetic fingerprint information. This information is used for, forensic science, paternity studies and indeed any application where a genetic profile has to be compared with one, or several, others. gel band pattern comparisons were revolutionized by the arrival of gel analysis packages such as GeneTools. GeneTools allows a user to store the information associated with a gel in a GLP manner, perform bandmatching across a gel to a known standard and export this information. However, gel analysis packages do not allow a user to compare gel band patterns between gels. In the past if such analysis were required, it would be necessary for either the data to be exported to another package or to do the calculations by hand.

GeneDirectory is a gel data storage and analysis package that works in conjunction with GeneTools enabling a user to perform complex genetic fingerprint studies across an unlimited number of gels.

Storage and Representation of Gel Data

To store large amounts of gel data efficiently, it is necessary to place it into an ordered structure. GeneDirectory utilizes a relational database and SQL queries to store and retrieve the gel data. Having a formal database structure means that it is possible to lock records whilst a user is accessing them. Therefore it is possible to have multiple database users and allow networking of the database. The formal structure also allows for the addition of further modules to GeneDirectory (e.g. sequence information could be associated with a track on a gel).

The way in which the gel information is displayed to the user is very important in a package storing large amounts of data. If the data were simply displayed as a list of tracks it would soon become impossible to navigate. In GeneDirectory the database is represented by a hierarchical tree (explorer-type) view (*Figure* 1).
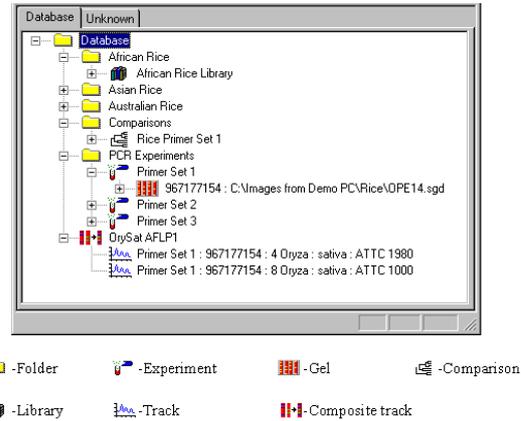

Figure 1- The hierarchical tree representation of the GeneDirectory database

Features of GeneDirectory

There are a number of features of GeneDirectory that allow the user to fully investigate the similarity of tracks stored in the database. Whichever type of analysis is to be pursued, the first step is to open a data set of tracks. This data set could be a gel, an experiment (a collection of gels), a library (a collection of classified tracks) or a comparison (a collection of unrelated tracks). There are currently three independent views available when a data set is open; cluster analysis/similarity matrix, genotyping and VNTR. Initially when the data set is opened, the track data and the track images are displayed (*Figure* 2).
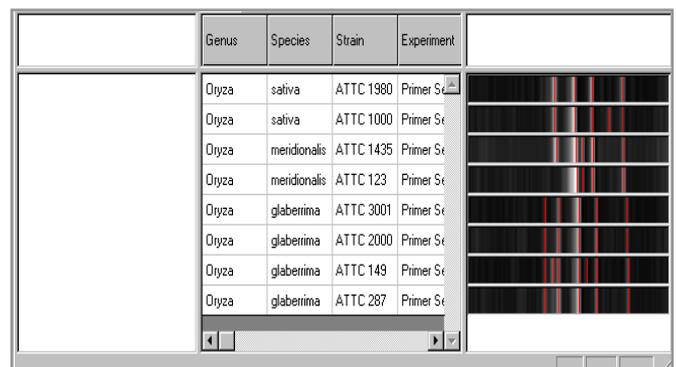

Figure 2- A collection of tracks (comparison, Rice Primer Set 1) following opening from the main tree view.

Cluster analysis

Cluster analysis is a powerful method to establish similarities between items. Cluster analysis process involves the identification and grouping together of similar tracks into clusters. In the case of gel tracks, this variable feature is the migration-distance values of bands on a track. A similarity measure is calculated between an individual track and every other track eventually generating a similarity matrix (*Figure* 3). In GeneDirectory, the similarity measure runs from zero (no similarity) to one (identical). The similarity parameters band tolerance (percent variation of migration-distance acceptable), migration type (MW, Rf or position) and similarity coefficient (Jaccard or Dice) can be altered by the user.
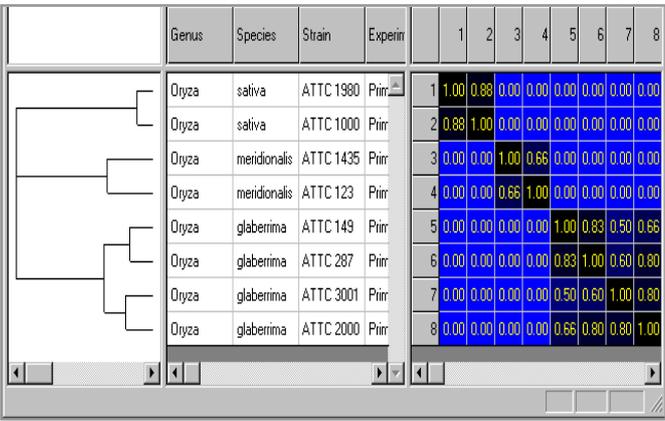

Figure 3- A dendrogram and similarity matrix generated from the track collection in figure 2.

Once a similarity matrix is generated, it is possible to draw a dendrogram. The dendrogram is a representation of how the items being studied should be grouped, or clustered (*Figure* 3). There are numerous linkage rules that determine how items are clustered, many of which are available in GeneDirectory.

Genotyping

Heredity studies are genetic studies which often involve comparing a number (normally two) of parental genetic fingerprints against a series of related individuals and checking for matching bands. Once a collection of tracks has been opened in GeneDirectory, it is possible to switch to a genotyping view (*Figure* 4). Within this view it is possible to display two types of track collections. In an upper window a composite track is displayed. A composite track is composed of one or more normal tracks that will be band matched to the data set of tracks in the lower window. The user can alter the matching parameters, band tolerance and migration type. It is also possible to include or exclude bands from both the composite track and a member of the track data set.
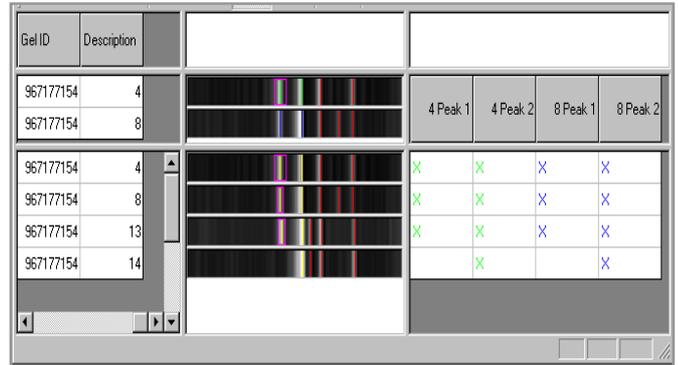

Figure 4- Bandmatching between a composite track and a data set of tracks

The composite track is shown in the upper window and the data sets in the lower window. **N.B.** A number of bands in the composite track have been excluded from the matching process.

Variable number tandem repeats (VNTR's)

If the user is working with a DNA sample that contains short nucleotide repeat sequences such as VNTRs it is possible to calculate the number of repeated elements in each band on a track using GeneDirectory software. After opening a data set of tracks it is possible to view the VNTR data for each track as shown in *Figure* 5.
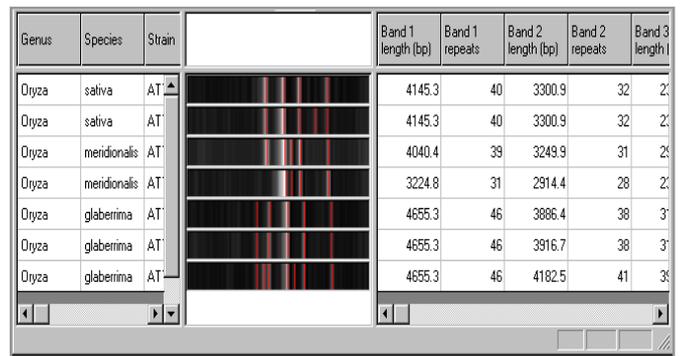

Figure 5- VNTR Data calculations window
Showing the size of each band in a track and the number of times a repeated element is present in that band. In the above example the repeat length and the flanking region were set at 100 bp.

Searching for a Track

A key feature of GeneDirectory is the ability to rapidly identify unknown or unclassified tracks. When establishing a database of tracks, some will be identifiable in terms of the classification data of the organism they originated from. User specified

classification headings can then be associated with that track. Once classified, a track can be added to a Library data set. The gradual build-up of classified tracks in these libraries develops into an increasingly comprehensive data collection. When an unknown track is introduced into the database, a search wizard can be used to identify the nearest matches in terms of band positions and band quantities.

Track basket

GeneDirectory is an extremely flexible package, in that tracks from a gel can be copied to various other structures within the database. As it is possible to copy tracks from many different gels, it is necessary to have a mechanism whereby tracks can be stored prior to being pasted into the new structure. In GeneDirectory this mechanism is the track basket The user can place a number of tracks or even a whole gel into the basket and then copy them into a new structure.

Conclusion

GeneDirectory is an unrivalled product in the track database market. The powerful combination of ease of use, for which Syngene has become renowned, and cutting edge database technology make it an essential tool for a genetic researcher.