E

BMT/13/31
ORIGINAL: English
DATE: November 21, 2011

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES, AND DNA-PROFILING IN PARTICULAR

Thirteenth Session
Brasilia, November 22 to 24, 2011

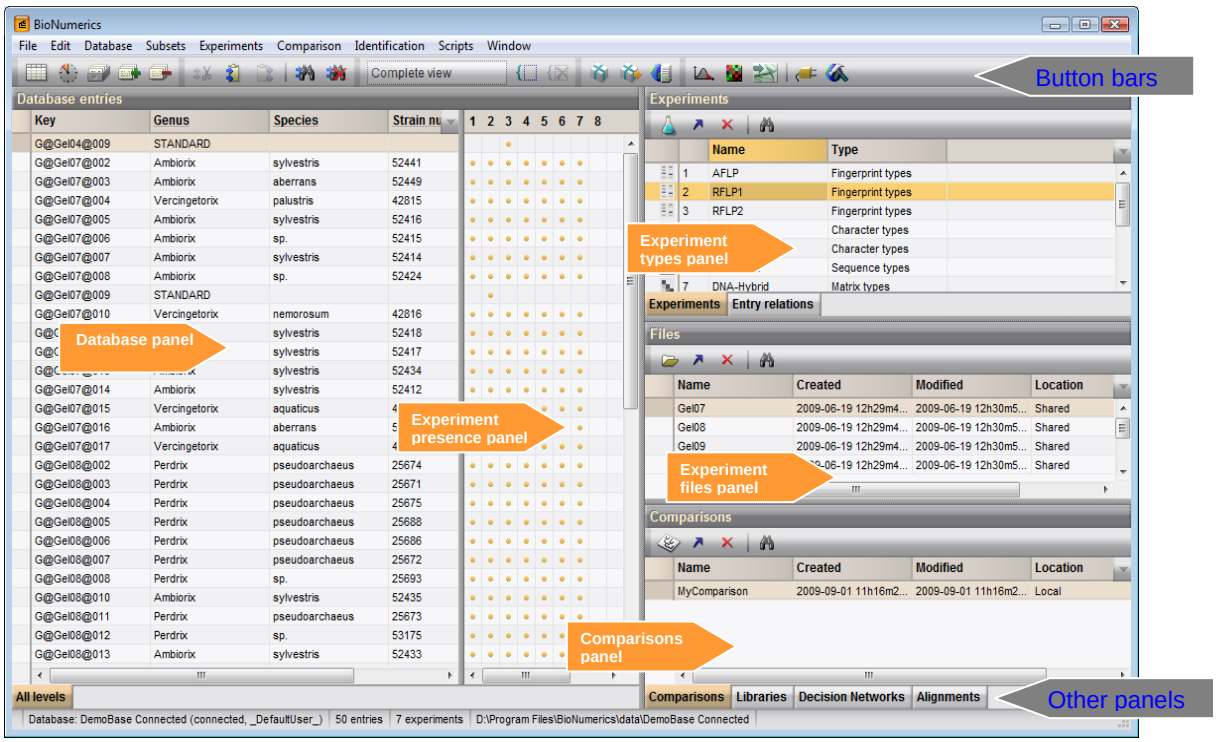
BIONUMERICS: A UNIVERSAL PLATFORM FOR DATABASING AND ANALYSIS OF BIOLOGICAL DATA

Document prepared by an expert from the Netherlands

INTRODUCTION

1. Applied Maths, a Belgium company, develops cutting-edge software for the biosciences. Since its establishment in 1992, the company has gained worldwide recognition with its software GelCompar for the analysis of 1D gel patterns, and later with BioNumerics, a software suite for integrated databasing and analysis of biodata. BioNumerics has grown into a universal platform combining databasing integrated networking and a wide range of analysis and decision-making tools, including data mining, querying, clustering, identification, statistics and analysis technologies for genomics, metabolomics and proteomics. The strength of the software lies in the non-obvious combination of bio(techno)logical and mathematical know-how combining the broad expertise in all areas of bioinformatics and the competence of designing powerful and innovative algorithms. BioNumerics is a universal solution to store and analyze all your biological data in one system.

2. Overview:



3. Outline of a BioNumerics database:

1. Database panel with information fields. Every entry has a unique key that links all data from the information fields and the experimental data for this entry.
2. Experiment presence panel. The dots show that there is data present for the particular entry for the particular experiment.
3. Experiment type panel: List of different experiments and type of experiments.
4. Experimental files panel: Original files with raw data that are imported into BioNumerics.
5. Comparisons panel: List of comparisons in which a subset of entries can be compared for data from a combined set of experiments.

4. What you can store:

(a) Information fields

- Up to 500 fields (each up to 80 characters)
- Link with external databases

(b) Attachments

- Bitmap images
- HTML and hyperlinks
- Word documents
- Excel spreadsheets
- PDF files
- Text documents

(c) Fingerprints

- 1-D electrophoresis gels scanned as bitmaps (RFLP, PFGE, Ribotyping, RAPD, DGGE & TGGE, etc.)
- Sequencer chromatogram files (AFLP, VNTR, HDA, etc.)
- Spectrophotometric files
- MALDI & SELDI profiles
- All other kinds of densitometric profiles

(d) Character data

- Phenotypic test panels
- Antibiotic resistance profiles
- Fatty acid and quinolone profiles
- Hybridization blots
- Biochemical & morphological features
- Microarray & Genechip data
- Etc.

(e) Sequence data

- Sequence trace (chromatogram) files
- Formatted sequences from public databases (EMBL, GenBank)
- Aligned sequences
- Amino acid sequences

(f) 2-D gels

(g) Trend curves and kinetic reading data

5. <u>What you can do:</u>

(a) Querying

- Using fields
- Using experiments
- Using ranges of values
- Using any combinations

(b) Cluster analysis

- UPGMA, Nearest Neighbor, Furthest Neighbor, Ward...
- Minimum Spanning Trees
- Consensus trees
- Calculation of degeneracy, error estimation on trees

(c) Identification

- Library construction
- Statistical confidence
- Neural Networks
- Decision Networks

(d) Dimensioning, ordination

- Principal Components
- Multi-Dimensional Scaling
- Self-Organizing Maps

(e) Phylogeny

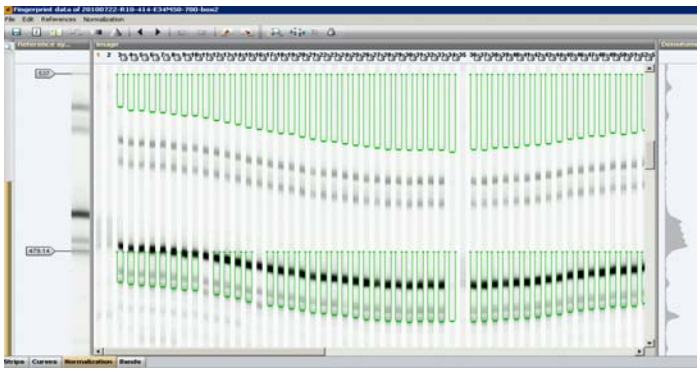
- Pairwise & multiple sequence alignment
- Neighbor Joining
- Parsimony
- Maximum likelihood

(f) Statistics

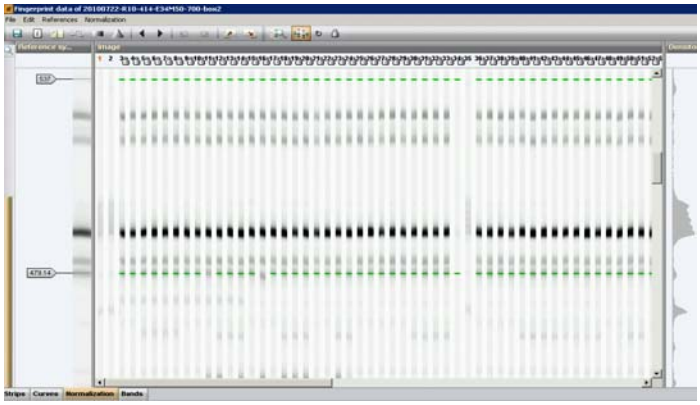
- MANOVA
- Discriminant analysis
- K-means partitioning, Jackknife,...
- Numerous statistical tests and charts

Current use of BioNumerics at Naktuinbouw

6. As can be observed from the above paragraphs, the BioNumerics software has a broad spectrum of tools and applications. Probably not many users will apply all tools for their routinely day-to-day work in the laboratory. For Naktuinbouw the software is applied to build up databases. For every crop/species a separate professional relational database (MySQL) is constructed. Depending on the crop and the availability of crop-specific molecular marker systems, fingerprint data based on either AFLP and/or on SSR is imported. The software can handle to import gel images or raw data files generated by capillary sequencers. In our lab the DNA profiles are processed on a Licor (slabgel) system using fluorescent dye's. The resulting TIFF image is directly imported into BioNumerics. Before analysis, the lanes on the gel have to be defined. The lane strips are linked to the sample and all information fields belonging to the sample using a unique key. Every entry in the database has such unique key. To be able to compare different gels together, it is of crucial importance to normalize the gels. Therefore we add an internal reference system to every sample on the gel.

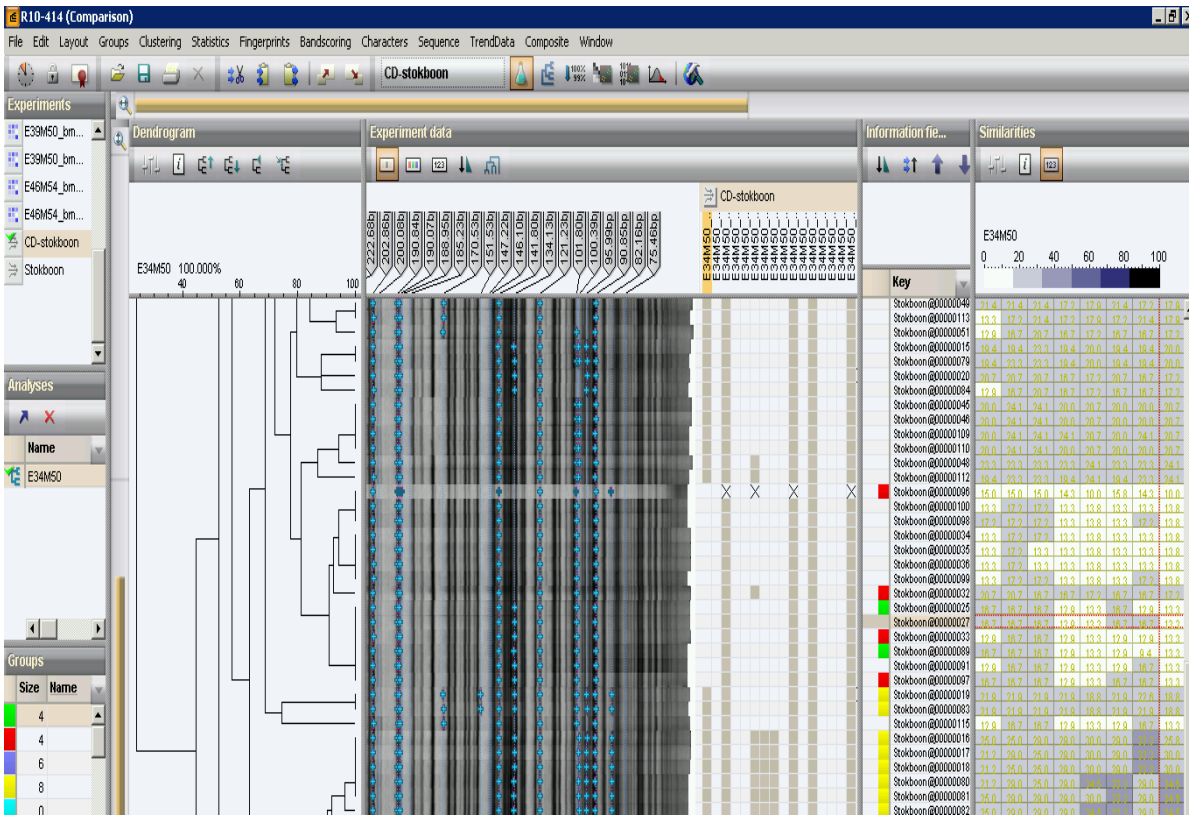


Before normalization



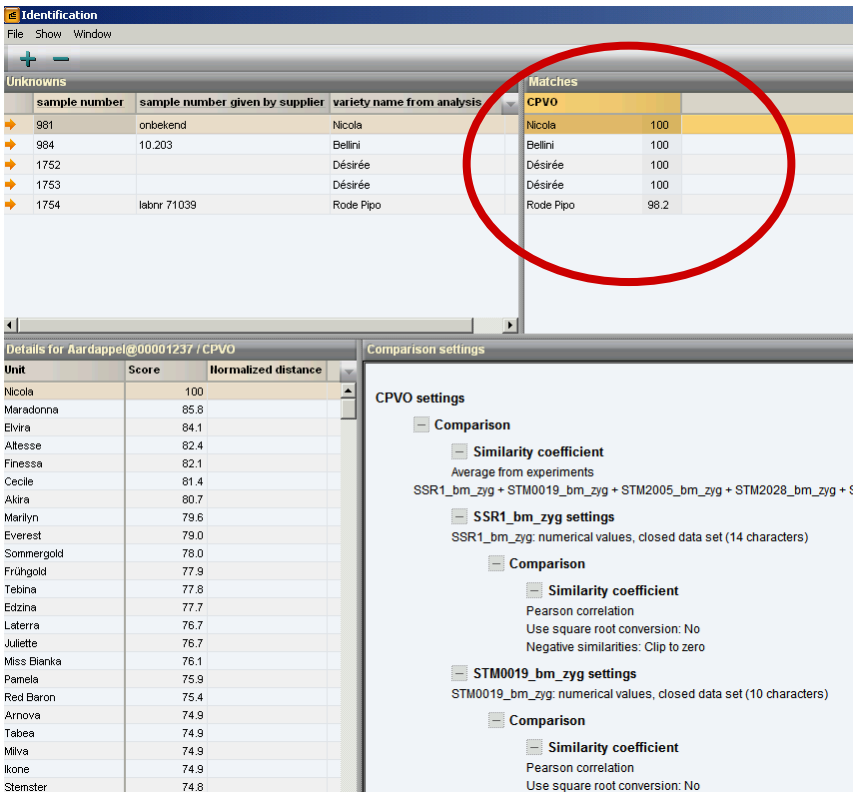
After normalization

7. Genetic analysis is performed on all or on a subset of samples in the database. First of all a comparison window is opened. The selected samples are displayed and also for every experiment the gel strips can be exposed. Depending on the order of the samples (numerical, alphabetical) an artificial gel is composed in which the polymorphic bands can be scores in a dominant manner (absent/present). Markers (bandclasses) are defined and the presence of a band is depicted by a "+". These scores can be saved as character data using the band scoring plug-in. Clustering is performed by choosing one of the available similarity algorithms that are fit for binary data (Jaccard, Dice, Simple Matching) and additionally the preferred cluster analysis method (e.g. UPGMA, neighbor joining, single linkage, complete linkage). The dendrogram, the similarity matrix, the scoring table, the raw data (gel image) and the sample information can be shown all together in one screen. That gives us the opportunity to double check the raw data (gel image and scoring) when unexpected clusters or variation is observed in the dendrogram since the most related gel strips are now close together. This verification system allows us to decrease the (human and technical) error rate drastically which effects the reliability and quality of the analysis in a positive way.



8. One of the major advantages of the BioNumerics software is the possibility to perform a single genetic analysis on a combination of different types of data in composite data sets. All kinds of different types of data (e.g. fingerprints, character data both numerical or categorical, sequences) in different weights reflecting the difference in reliability of the data can be used.

9. When the number of samples in the database increases, there is an alternative way to compare new samples' DNA profiles with the complete database. For these purposes identification libraries can be used. Identification can be performed using several methods (e.g. mean similarity, maximum similarity, K-nearest neighbour).

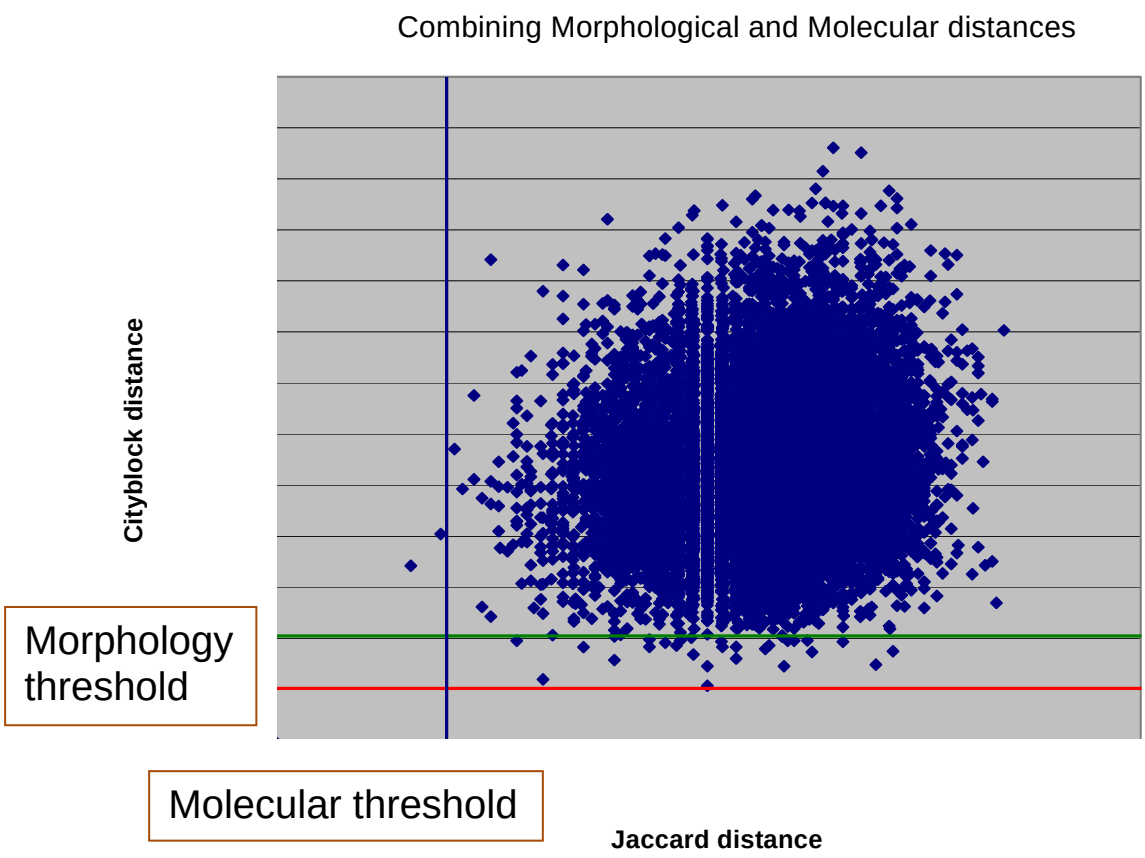


Best hit for unknown sample with samples in library

Future use of BioNumerics at Naktuinbouw

10. The technological developments are improving rapidly. With the introduction of second, and now third generation sequencing technologies and platforms, the costs for sequencing is dropped dramatically. As a result whole genome sequencing is already available for major crops and will become available and affordable for all crops in the (near) future. Therefore, Naktuinbouw investigates the possibilities to perform variety identification studies based on genomic sequences. These sequence data can also be imported in the BioNumerics software as text files in FASTA, GenBank or EMBL format, direct download from online repositories, batch import and assembly of sequencer trace files.

11. Since the composite dataset allows us to combine different type of data, Naktuinbouw would like to use BioNumerics to combine morphological and molecular similarities/distances for the management of reference collections.



Naktuinbouw, PO BOX 40, 2370 AA Roelofarendsveen, The Netherlands:
Dr. Hedwich Teunissen, Daniël Deinum, Menno Hoekstra, (contact: h.teunissen@naktuinbouw.nl)

Applied Maths, Keistraat 120, 9830 Sint-Martens-Latem, Belgium:
Johan Goris (contact: info@applied-maths.com)

[End of document]