E

BMT/11/10 Rev.
ORIGINAL: English
DATE: August 28, 2008

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Eleventh Session
Madrid, September 16 to 18, 2008

CONSTRUCTION OF AN INTEGRATED MICROSATELLITE AND KEY MORPHOLOGICAL CHARACTERISTIC DATABASE OF POTATO VARIETIES ON THE EU COMMON CATALOGUE
PART 2: THE DATABASE (REVISED)

Document prepared by experts from the United Kingdom

CONSTRUCTION OF AN INTEGRATED MICROSATELLITE AND KEY MORPHOLOGICAL CHARACTERISTIC DATABASE OF POTATO VARIETIES ON THE EU COMMON CATALOGUE
PART 2: THE DATABASE (REVISED)

Dr Alex Reid[1] & Dr Ir Lysbeth Hof[2]

[1]SASA, Edinburgh, UK
[2]Naktuinbouw, Wageningen/Roelofarendsveen, NL

INTRODUCTION

1. During the course of the 2 year project to construct an integrated microsatellite and key morphological characteristic database of potato varieties on the European Union (EU) Common Catalogue large amounts of data were generated. These data included the results of the 9 microsatellite markers or simple sequence repeats (SSRs) which were scored as binary data, the morphological data scored in a multi-state format, textual information about the samples them selves and photographs of light sprouts. The requirements for a software platform were therefore straightforward. It had to be able to handle different types of data, be able to store a large amount of data and analyze these data. The most important aspect of these requirements is the ability to analyze these data and a particularly desirable feature would be to reliably identify unknown samples. The solution was to use the BioNumerics software package (Applied Maths), an integrated package that enables the user to link the results from numerous molecular techniques (known as 'Experiment Types' in BioNumerics) to a single entry in the database. These experiment types can vary from electrophoresis gel images (including 2D protein gels), DNA sequences and character data both binary and multi-state. A variety of analytic techniques can then be applied either on single or on composites of several different experiment types.

2. Data can either be entered directly in BioNumerics or more effectively stored in an external database such as MS Access which is linked to BioNumerics via Open DataBase Connectivity (ODBC). Using MS Access as the linked database has the advantages that data can easily be exchanged between laboratories even if they do not have access to a copy of BioNumerics themselves. Updates made to the MS Access database are updated in BioNumerics with a simple command.

3. For use in this project a further useful feature of BioNumerics is the facility to create a library of reference varieties against which unknown isolates can be screened resulting in accurate identification of said isolates.

Database Structure

4. A database was constructed in MS Access containing 11 tables and was designed so that it can serve BioNumerics. In the Access database there are tables for each of the nine markers, a table for the morphological data and a table containing general information for each sample in the database (Table 1). As BioNumerics requires a unique identifier for every entry and as some varieties had multiple samples, the DNA sample code has been used for the SSR data (we tried to analyze every variety as duplicate samples from separate tubers, and many samples have been analyzed in the United Kingdom as well the Netherlands to test robustness of the data). Entries for the morphological data have only one entry for each country and here the unique identifier is the variety name and the country code where the description was carried out. For example 3 samples of *cv.* Ditta were submitted for SSR analysis from the

Netherlands, Poland and the United Kingdom and the unique identifier for these samples are the DNA extraction code (NL-089, PL-021 and UK-0598) respectively. All 4 countries submitted a morphological description for Ditta and the unique identifiers are Ditta_DE, Ditta_NL, Ditta_PL and Ditta_UK. Note that for reasons of confidentiality all other variety names have been removed from this document.

Table 1. The data recorded for each sample in the database and explanations for each field. Note that the samples used for DNA analysis are not necessarily the same as those used for light sprout and/or morphological characterization. Also some varieties have full morphological descriptions, others only light sprout descriptions and that light sprout photographs are not necessarily linked to full descriptions.

Field	**Explanation**
Key	The unique identifier for use in BioNumerics
Variety Denomination	The name of the variety
Origin of sample e.g. Breeder (B), office (O) or other (T)	Where the sample was obtained from either the breeder or from one of the partners collections or from another source
Submitting office (S SASA, B BSA, C COBORU, N Naktuinbouw)	The office which submitted the sample for DNA analysis or submitted the morphological description
Harvest year	The year the sample analyzed was harvested
DNA extraction laboratory (S SASA, N Naktuinbouw)	The laboratory the DNA sample was extracted (either United Kingdom or Netherlands)
Extraction year	The year the DNA extraction was made
Place of storage of DNA sample (S SASA, N Naktuinbouw)	The place where the DNA sample is kept in long term storage
SSR analysis performed at (S SASA, N Naktuinbouw)	The laboratory where the SSR analysis was performed
SSR analysis year	The year the SSR analysis was performed
Technical protocol used for morphological description	The technical protocol used for the morphological description (if known)
Description year (either before 1995 or actual year)	The year that the latest description was carried out, N.B. not necessarily from official description
Place description carried out	The office the description was carried out at
Photograph availability (and link to photograph)	The file name for the photograph (if available) and hyperlink
Place photograph taken (office)	The office the photograph was taken at
Photograph year (if known)	The year the photograph was taken
National Reference	The national reference number of the sample (if there is one)
Comments	Any other comments (e.g. is the variety a mutant of another variety)

5. There is a table for the morphological data and 9 tables for the each SSR markers. All of these tables follow the same basic format (Table 2). Each entry has its own unique identifier (Key), variety name and then the data itself. It was decided to store the SSR data as individual markers as this meant that it would be possible to analyze the data on a marker-by-marker basis as well as a combined data set.

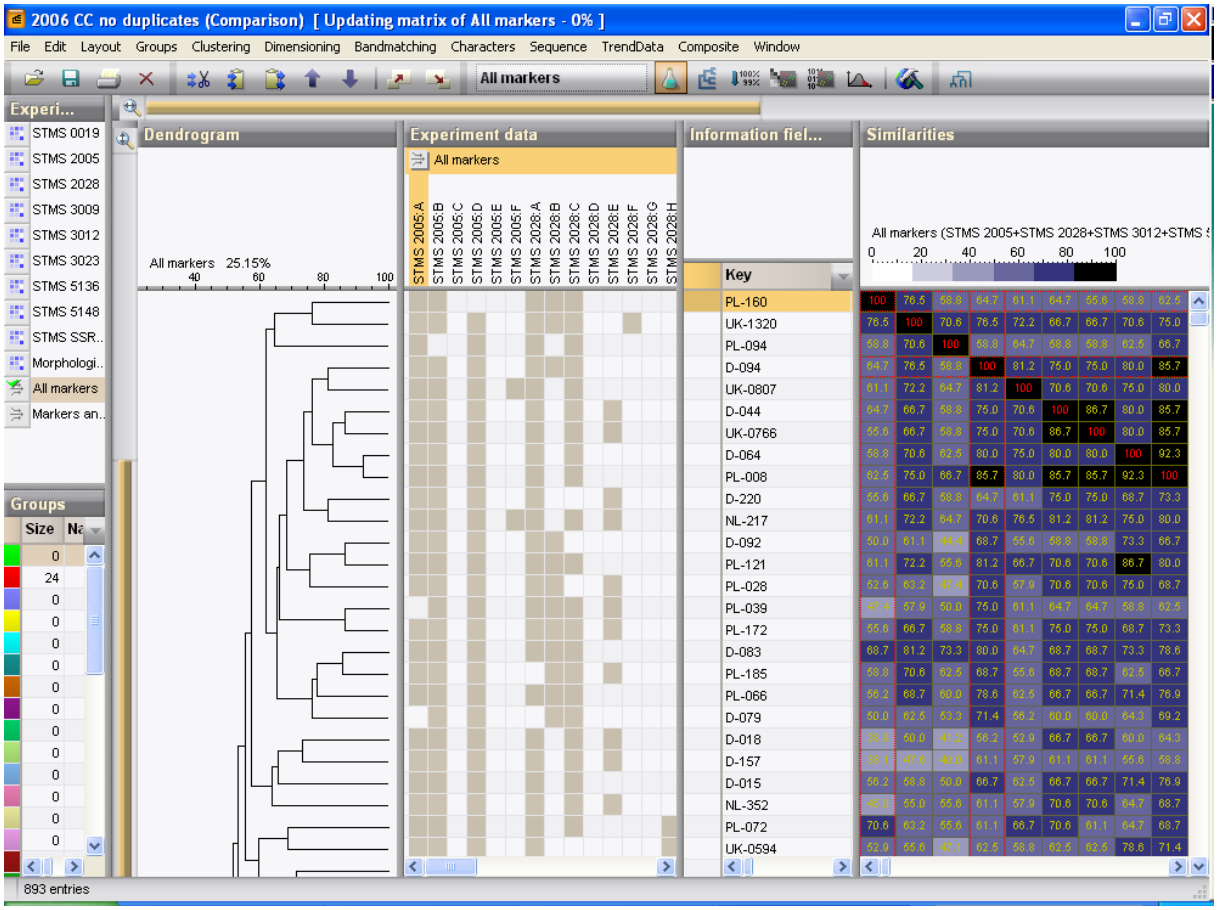
Table 2. An example of SSR data (from marker MS SSR1)

Key	Variety denomination	A	B	C	D	E	F	G	H	I	J	K	L	M	N
NL-001	A	1	0	0	1	0	0	0	0	1	1	0	0	0	0
NL-002	B	0	1	0	1	0	0	0	0	1	0	0	0	0	0
NL-003	C	1	0	0	1	0	0	0	0	1	0	0	0	0	0
NL-004	D	0	0	0	1	0	1	0	0	1	0	0	0	0	0
NL-005	E	1	0	0	1	0	0	0	0	0	0	0	0	0	0
NL-006	F	1	0	0	1	0	1	0	0	1	0	0	0	0	0
NL-008	G	0	0	0	1	0	1	0	0	1	0	0	0	0	0
NL-009	H	1	0	0	0	0	1	0	0	1	0	0	0	0	0
NL-010	I	1	0	0	1	0	1	0	0	1	0	0	0	0	0

BioNumerics

6. Transferring data from MS Access to BioNumerics is easily achieved by the use of drop down menus and once the data has been imported comparisons can be made using a variety of techniques. These include various types of cluster analysis (Figure 1), maximum parsimony and principle components analysis. Analyses can be performed on the entire database or a selection of entries. After addition or deletion of specific entries, results can easily be re-calculated with a simple command.

Figure 1. Cluster analysis of varieties using the Jaccard coefficient and UPGMA. The panel on the left allows the experiment type to be selected for analysis (in this case all SSR markers), the next panel shows the dendogram calculated from these data. The middle panel shows the data itself either as a graphical representation or as a numerical value. The last two panels show the information for the samples and the similarity matrix for all pair-wise comparisons.

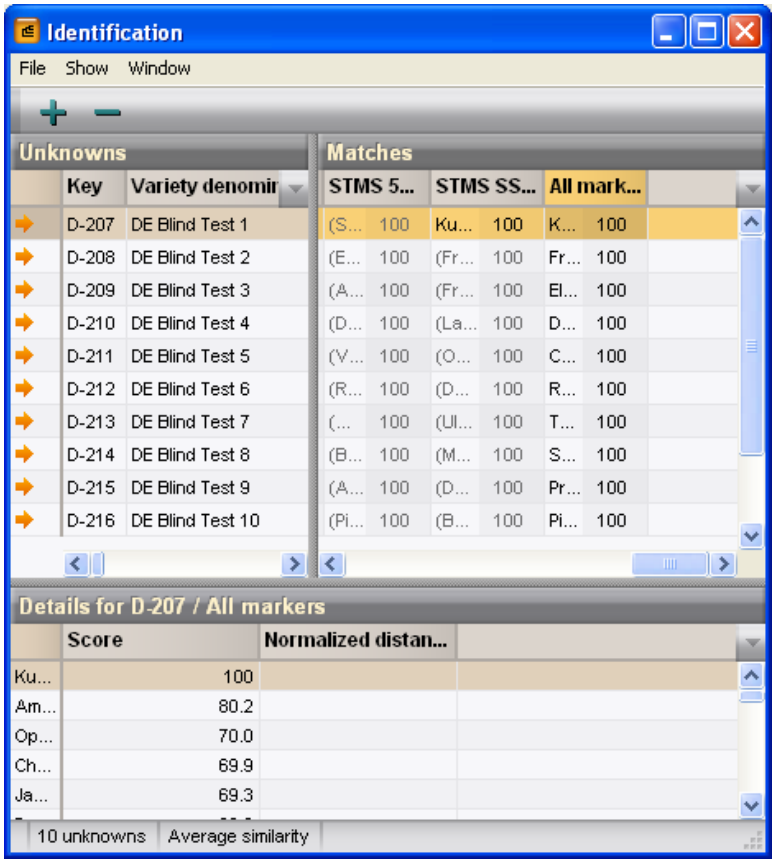


7. Identifications are performed by screening against a library of previously entered taxa (Figure 2). The results obtained for the blind test samples analyzed during the project were all successfully identified with the exception of two varieties previously highlighted as pair with identical profiles that could not be explained.

8. Two libraries were set up in the database: one contains the SSR data; and the other the morphological data. As the majority of the molecular data is consistent from one sample to another each library taxa is linked to a single representative entry in the database. The few exceptions are one variety where a polymorphism appears to be present and a few varieties where different samples yield significantly different profiles (presumably these are actually different varieties with one sample being mislabeled at some point in time). See document BMT/11/9 "Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue. Part 1: discussion of morphological and molecular data".

9. Morphological data were less consistent and, therefore, the morphological library is slightly different in that each library unit is linked to multiple samples (for those varieties where data has been submitted from more that one source). For this reason it is virtually impossible to achieve accurate identifications on the basis of morphology alone using the database.

Figure 2. Results obtained from screening blind test samples against the SSR library. The top half of the window shows the samples that have been screened on the left hand side. The right hand panel shows the percentage match for each sample against the individual markers and all of the markers combined. The lower half of the window contains the results for individual samples and shows the top hit followed by the next most similar and the percentage matches for each. For reasons of confidentiality the columns with variety names have been narrowed in order not to reveal full names.



Conclusions

10. One of the main problems encountered during a project such as this is the enormous amount of data generated in a short of time period. For example, each variety has a 95 digit binary number (the total number of possible alleles) and there are over 1,100 samples in the database. The Access database linked directly to BioNumerics enabled us to store these data in an industry standard and exchangeable format and also perform complex analyses of these data and perform identifications of unknown samples.

Acknowledgements

11. Funding of this project was provided by the Community Plant Variety Office of the European Community, The Dutch Ministry of Agriculture, Nature and Food Quality (Netherlands), Naktuinbouw (Netherlands), SASA (United Kingdom), the Bundessortenamt (Germany) and COBORU (Poland). The microsatellite profiles for the Netherlands were provided by Plant Research International, Wageningen (Netherlands).

REFERENCES

BMT/11/9 (2008) Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue. Part 1: discussion of morphological and molecular data

[End of document]