E

BMT/8/14 Add.
ORIGINAL: English
DATE: September 18, 2003

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Eighth Session
Tsukuba, Japan, September 3 to 5, 2003

ADDENDUM TO DOCUMENT BMT/8/14

BAND SCORING, DISTANCES, USE FOR DISTINCTNESS AND UNIFORMITY, DATA STORAGE

Presentation prepared by experts from France and the United Kingdom

The Use of Band Scoring, Distance Estimates and Data Storage in DUS Testing.

John Law[1] and Sylvain Gregoire[2]

[1] **NIAB**, UK
[2] **GEVES**, France

1. Scoring of Molecular Data

Many potentially useful DNA profiling methods currently exist and further variants / major developments are regularly reported in the scientific press.

Scoring 'strategies' based on the genetics of the crop [eg ploidy, allogamous or self-pollinating] and the specific action of the profiling method [dominant marker system]; are required.

1. Scoring of Molecular Data.

Further key issues relate to the information necessary for the specific task in terms of:-

(a) The number of individuals to be assessed
(b) The 'quantity' of molecular information
(c) Scoring Protocols

1. Scoring of Molecular Data.
(a) The number of individuals to be assessed

The key, is the level (known or assumed) of the within variety uniformity. This will generate a 'sample' representative of the specific material under study ~ relatively small in the case of self-pollinating varieties, clonal material or inbred lines; relatively large for allogamous crops.

1. Scoring of Molecular Data.
(a) The number of individuals to be assessed

There is an exact formula for sample size that covers all cases. Experience, coupled with knowledge of the crop and the purpose of the experimentation have a strong role here. Ask a statistician as the precision of the estimates will be strongly influenced by the 'sample size'.

The effect of any reduction in sample size can be estimated *post-hoc* by the method of bootstrapping.

1. Scoring of Molecular Data.
(a) The number of individuals to be assessed

Practical issues, such as the number of 'lanes' on a gel or a piece of expensive equipment, may make for very poor sampling strategies.

1. Scoring of Molecular Data.

(b) The 'quantity' of molecular information

It is popular to screen a relatively small 'variety set' with a large number of 'markers' and to assess the potential power or effectiveness of each marker. While 'PIC' is a power estimate; many other features need to be taken into account when making decisions as to which specific markers to use in scaling up from small 'look-see' to large-scale experimentation.

1. Scoring of Molecular Data.

(b) The 'quantity' of molecular information

Other features include:-

- are markers mapped?
- balanced genome coverage?
- robustness, sensitivity and overall score-ability?
- potential for multiplexing?

… etc

The effect of any reduction in quantity of molecular data can also be estimated, ***post-hoc,*** by the method of bootstrapping.

1. Scoring of Molecular Data.

(c) Scoring Protocols.

To be established before any large-scale experimentation and includes how to deal with

- rare alleles, null alleles or 'faint' bands
- missing data
- scoring of monomorphic bands

An agreed detection precision is advisable (in terms of bp's).

Operating protocols for manual scoring will differ from those based on 'automated' scoring approaches.

2. Distance Measures.

Text books and scientific literature over the past 30 years has led to a confusingly long list of potential distance or similarity estimates.

Choice should be made on the basis of the genetics of the crop, understanding of the marker system in use and not just on the method(s) offered by the software that is available or "the currently favoured method".

Implicit or explicit assumptions needed to be checked thoroughly.

2. Distance Measures.

Binary Data

Zero Weight of Absence (0)

- Nei and Li [x1 presence/presence (1,1)]
- Jaccard's [x2 presence/presence (1,1)]

Full Weight of Absence (0)

- Simple Matching

2. Distance Measures.

Allelic Frequency

- Euclidean
- Roger's
- Nei's standard genetic distance

3. Data-basing of Molecular Data.

Q. Does putting molecular data into an EXCEL spreadsheet constitute "holding the data in a database"?

3. Data-basing of Molecular Data.

Q. Does putting molecular data into an EXCEL spreadsheet constitute "holding the data in a database"?

A. Yes and No!

3. Data-basing of Molecular Data.

Q. Does putting molecular data into an EXCEL spreadsheet constitute "holding the data in a database"?

A. Yes and No - Just ask any IT professional

**Confused?

**So what is a database?

**What is the role of EXCEL?

3. Data-basing of Molecular Data.

Size really does matter.

Initially the quantity of molecular data would fit comfortably into an EXCEL spread-sheet ~ limits 256 columns by 64000 rows. Even when this was not the case, "work around's" could usually be found".

The data was all in one place, use of the natural row column 'flat' file structure, data all of a similar type (with limited header information) and easily transportable and used (OPENED) by other colleagues.

3. Data-basing of Molecular Data.

EXCEL can be a useful precursor but …is limited

•size,

•efficiency,

•types of information content catered for

and most importantly

•<u>security.</u>

3. Data-basing of Molecular Data.

Many current molecular methods have the potential to generate large amounts of data. With future developments, the quantity of data generated by many orders of magnitude and the need to turn data into information- and information into knowledge - will increase.

For effective and efficient use of this 'vast' and valuable data resource by many people - access control is handled by the database system and retains the highest level of security (back-up, recovery, 'journal-ing')

3. Data-basing of Molecular Data.

UPOV have been providing the UPOV ROM for several years.

30-40 types of information are included based on information and the requirements of the varied working parties and crop experts.

Data is collated from UPOV member states. Information in UPOV ROM is searchable and sent regularly, in CD form, to subscribers.

3. Data-basing of Molecular Data.

UPOV ROM uses an internationally defined standard generalised markup language SGML.

There are other popular markup languages XML ~ extensible markup language (used for the definition and exchange of information)

HTML ~ hyper text markup language (internet)

While XML is supported by many application packages, SGML needs specific code to be written in the required format.

3. Data-basing of Molecular Data.

Variety description data is being considered as an extension of the UPOV ROM. Many issues in common with molecular data are being worked through.

There are many ways in which molecular data could be stored and it is unclear how such data could be combined with admin and technical information.

End of document