Final Report of the AEGIS project

'PGR Duplicate Finder', a software package to assist in the identification of putative duplicates in germplasm databases.

The latest version of the Duplicate Finder can be downloaded from:
http://documents.plant.wur.nl/cgn/pgr/aegisdf/

Introduction

A critical step in the creation of the AEGIS European Collection is the identification of the candidate European Accessions and among them the so-called Most Appropriate Accessions (collectively hereinafter called MAAs). Given its labour-intensiveness, all attempts should be made to facilitate this process, allowing the Crop Working Groups (CWG) to concentrate on the choice of the MAAs rather than spend time on searching duplicates.

In the selection of the MAAs, all CWGs are confronted with the laborious search for probable duplicates. This activity has common elements for all crops, which can be formalized and automated. Software has been developed that will assist the ones responsible for the proposal of MAAs in the identification of probable duplicates. To allow for easy processing of the results this software has been implemented in an Excel environment with easy to use macros.

Ideas about the identification of duplicates have been circulating for a long time. CGN, the main participant in this project, had already, prior to the project, in its attempts to select MAAs for a number of crops created some preliminary macros to support the activity and it was clear that a more targeted development of robust tools would be very useful. In the EUROGENEBANK proposal (of the 2010 FP7 Call for proposals) IPK would develop similar tools. The JKI also has some experience with the semi-automatic identification of probable duplicates in the framework of the ECPGR *Avena* database.

Most of the scientific papers about definition and identification of genebank duplicates are (co-)written by the participants in this project.

Objective of the project, as formulated in the project proposal, therefore was: "To develop easy to use software called 'PGR Duplicate Finder' for the preliminary identification of probable duplicates on the basis of a list of passport data in the EURISCO upload format."

Material and Methods

The first step of the project was building a prototype of the Duplicate Finder. This version had minimal functionalities, and served to demonstrate to the colleagues how it would look like, and in what environment the software would be implemented.

Based on this prototype a brainstorm session with scientists involved in the identification of duplicates from CGN, IPK and JKI was organised on November 10th 2011. (For the agenda see Appendix 1.) It was participated by Christoph Germeier (JKI), Helmut Knuepffer and Markus Oppermann (IPK), and Theo van Hintum, Roel Hoekstra and Frank Menting (CGN). The PowerPoint presentations prepared for this meeting are available on request.

Based on the findings during this meeting, the CGN developers changed their initial ideas drastically, moving from sequences of if-statements, a decision tree, towards the extraction of numerical values from the fields that might contain those numbers, combining and ordering them, expecting that the probable duplicates would appear close to each other in the ordered list.

Despite the much higher efficiency of this approach as compared to the decision tree approach, performance was a big issue. Only by using Excel functions, such as string manipulation and most importantly, list ordering, it proved possible to achieve acceptable performance levels.

The first proper version of the Duplicate Finder was optimised with test data sets extracted from the ECPGR Central Crop Databases, and from EURISCO of various crops and sizes. This version was distributed to the colleagues from JKI and IPK, but also to colleagues in CGN, along with two short manuals: one for all users who want to work with the Duplicate Finder (see Appendix 3), and one for advanced users who wish to fine-tune the programme to their data set (see Appendix 4). They were all requested to test this version with their own data, and record their observations. This resulted in very many valuable comments that were all considered seriously, and most of which resulted in changes of the programming code. (For a list of feedback issues and resulting action, see Appendix 2).

Finally, after some testing of the improved version, the manuals were included in a ‘read me’ sheet in the Duplicate Finder spread sheet.

Results

The activities resulted in a speadsheet called ‘DuplicateFinder v1.0.xlsm’. It contains four visible sheets and three hidden sheets.

The visible sheets are:

Read me – A sheet with the two short manuals (as appended in appendices 3 and 4). These are included in the spreadsheet to make sure that they are always accompanying the code.

MCPD List – A sheet with the descriptors and format rules of the Multi-Crop Passport Descriptor List, on which basis the Duplicate Finder performs it searches.

Report – A sheet that is used to log certain actions. On the basis of these logs, the user can correct the data sheet.

DATA – A sheet with the actual data used in the analysis, also to record the results of the analysis. It comes with a small, thousand records, dataset on *Lactuca* for demonstration purposes.

The hidden sheets are:

SimRules – A sheet to store the weights given to matches depending on the source descriptors of the data; a match between two numbers (one the accession number, the other the donor number, both with the same prefix) will have a much higher weight as two accession numbers with different prefixes.

noSndx – A sheet for fine-tuning, in which the terms that should be excluded from the matching process. Think of terms such as ‘landrace’.

MCPD Codes – A sheet where all codes used in the MCPD are listed (such as the codes for sample status or origin address).

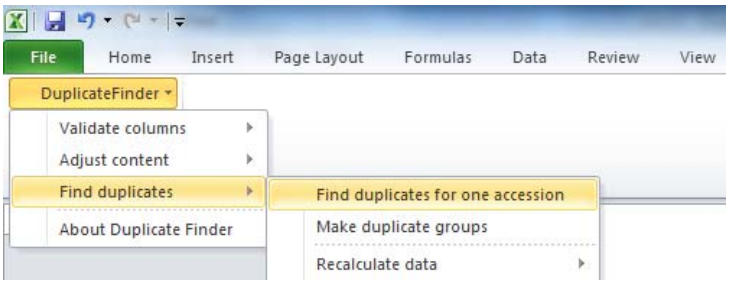
When opening the spreadsheet, a menu item ‘Add-Ins’ is added to the Excel main menu, that contains the option Duplicate Finder. Under this option a hierarchy of options appears. At the highest level there are the options:

‘Validate columns’ allowing the user to check the validity of the MCPD data set included in the DATA sheet.

‘Adjust content’ with some tools to correct the frequently occurring format errors.

‘Find Duplicates’, the main item allowing the user to identify the duplicates of one specific accession, or to create duplicate groups that group accessions with the probable duplicates.

This report is not the most appropriate place to describe all sub-options and the results of the choices, this is done in the manual (in the ‘read me’ sheet). Only the central ones will be discussed in some detail.



'Find duplicates for one accession', if this option is chosen the record that is selected at that moment (or the record that contains the selected cell) is matched with all other records. A new column is created DFSim, where all the similarity values of the selected record with all other records are recorded, and finally the sheet is sorted on that column. The accessions with the highest similarity with the selected accession are placed on top. In the example in Fig. 1, the output of searching the duplicates can be seen.

	A	C	D	G	H	M	X	Y	Z	A
1	DFSim	INSTCODE	ACCENUMB	GENUS	SPECIES	ACCENAME	DONORCODE	DONORNUMB	OTHERNUMB	ANC
2	1.00	GBR006	HRIGRU 1827	Lactuca	sativa	DETENICKY LETNI	GBR006	LS/91	:62 LAC58/73 DEUIPKGAT	
3	0.80	GBR006	HRIGRU 1807	Lactuca	sativa	BONTE CHILI	GBR006	LS/91	:DEUIPKGAT LAC212/74	
4	0.70	HUN003	RCAT033862	Lactuca	sativa	Detenicky Letni				
5	0.67	USA022	PI 491048	Lactuca	sativa	WP 062		WP 062	USA126:PI491048;WP 062	
6	0.31	NLD037	CGN04552	Lactuca	sativa	Cos Kakichisago	NLD083	91		
7	0.31	UKR021	UL1600137	Lactuca	sativa	Imhoroved Hanson				
8	0.10	GBR006	HRIGRU 4403	Lactuca	sativa	IMPERIAL WINTER RESELEC	GBR006	062/80	:PINETREE	
9	0.03	CZE122	09H5800894	Lactuca	serriola				:PI 491237;:USA 91	
10	0.03	ESP026	BGV008953	Lactuca	sativa	Lechuga			ESP004:NC053535;ESP004:BGE025345	
11	0.03	ESP026	BGV008956	Lactuca	sativa	Lechuga			ESP004:NC053538;ESP004:BGE025348	
12	0.03	NLD037	CGN22733	Lactuca	serriola					
13		ARM002	245	Lactuca	viminea					
14		ARM002	248	Lactuca	viminea					
15		AUT046	SA007	Lactuca	sativa	Gärtnerstolz				
16		AUT046	SA057	Lactuca	sativa	Parris Island Cos				
17		AUT046	SA072	Lactuca	sativa	Salata Ogulin				

Figure 1 – result of selecting the duplicates of an accession

'Make duplicate groups', when this is started, all pairwise similarities are calculated and groups are formed with the furthest neighbour algoritm, and a threshold of a similarity of 0.30. This means that all accessions in a group will have a similarity of this threshold value or larger. A column is added called DFGrp, which contains the group numbers, and the sheers is sorted accordingly. The result will be groups from which the curator can select the probable duplicates. In the example in Fig. 2, the output of creating the duplicate groups can be seen. Due to computer memory requirements there is a limit of 15,000 records that can be processed at one time. A user who has more data in the DATA sheet should split this dataset in chunks smaller than 15,000, for example on the basis of taxon.

	A	B	D	E	H	I	N	Y	Z	AA	A
1	DFGrp	DFSim	INSTCODE	ACCENUMB	GENUS	SPECIES	ACCENAME	DONORCODE	DONORNUME	OTHERNUMB	ANC
2	1		ESP026	BGV009008	Lactuca	sativa	Lechuga oreja de mulo			ESP004:NC068239;ESP004:BGE029386	
3	1		ESP026	BGV009049	Lactuca	sativa	Lechuga oreja de mula			ESP004:NC044340;ESP004:BGE030483	
4	1		ESP026	BGV009120	Lactuca	sativa	Lechuga oreja de mulo			ESP004:NC074389;ESP004:BGE032369	
5	1		ESP026	BGV009280	Lactuca	sativa	Lechuga oreja de mulo				
6	1		ESP026	BGV014123	Lactuca	sativa	Lechuga oreja de mula			ESP004:NC080422;ESP004:BGE037702	
7	1		ESP027	BGHZ-1014	Lactuca	sativa	Lechuga oreja de mulo				
8	1		ESP027	BGHZ-4329	Lactuca	sativa	Lechuga oreja de mula				
9	2		CZE122	09H5700191	Lactuca	sativa	Reine de Mai			:9-1-10/37	
10	2		NLD037	CGN04816	Lactuca	sativa	Reine de Mai; Maikonig	NLD083	474		
11	2		NLD037	CGN04818	Lactuca	sativa	Reine de Mai; Maikonig	NLD083	476		
12	2		RUS001	505900692	Lactuca	sativa	Reine de Mai				
13	2		RUS001	505901162	Lactuca	sativa	Reine de Mai				
14	2		UKR021	UL1600142	Lactuca	sativa	Reine de Mai	RUS001		UKR008:00045;UKR021:00437;RUS001:00	
15	3		CZE122	09H5700180	Lactuca	sativa	May King			:9-1-10/26	
16	3		RUS001	505901181	Lactuca	sativa	MAYKING				
17	3		USA022	PI 289035	Lactuca	sativa	MAY KING			USA126:PI289035;MAY KING;VII-10-36	
18	3		USA022	PI 536745	Lactuca	sativa	MAY KING			USA126:PI536745;MAY KING;W6 6411	
19	4		CZE122	09H5700719	Lactuca	sativa	Safir				
20	4		FRA011	LC0503	Lactuca	sativa	Saffier				
21	4		NLD037	CGN05135	Lactuca	sativa	Saffier	NLD083	1519		
22	4		NLD037	CGN11395	Lactuca	sativa	Safier	CZE061			
23	5		DEU146	LAC 1105	Lactuca	sativa				DEU146:K 10179	
24	5		GBR006	HRIGRU 1541	Lactuca	sativa	SILVESTER	GBR006	LS/67	:NUNHEMS	
25	5		NLD037	CGN04532	Lactuca	sativa	Boettners	NLD083	67		
26	5		UKR021	UL1600115	Lactuca	sativa	Merit				
27	6		DEU146	LAC 673	Lactuca	sativa				DEU146:K 8844	

Figure 2 – result of creating duplicate groups

Recommendations

Duplicate Finder v1.0 has been tested extensively, and can form the basis of a duplicate search. The user is however always the one who decides. For this he might have to do additional searches, ordering and other manipulations. But the Duplicate Finder creates a comfortable starting point.

Duplicate Finder v1.0 is truly the version 1.0. A lot could be improved. CGN has invested much more time than originally budgeted, but could have inverted even more time. The user is invited to give feedback and bug reports to CGN, and CGN will try to fix bugs and make smaller improvements on the basis of that feedback, resulting in v1.1, v1.2 etc. It will however not be possible to create new versions of the software without a new project.

Bibliography

Germeier, C.U., L. Frese, S. Bücken (2003) “Concepts and data models for treatment of duplicate groups and sharing of responsibilities in genetic resources information systems." Genetic Resources and Crop Evolution 50(7): 693-705.

van Hintum, T. J. L. and H. Knüpffer (1995). "Duplication within and between germplasm collections. I. Identifying duplication on the basis of passport data." Genetic Resources and Crop Evolution 42(2): 127-133.

van Hintum, T. J. L. and D. L. Visser (1995). "Duplication within and between germplasm collections. II. Duplication in four European barley collections." Genetic Resources and Crop Evolution 42(2): 135-145.

van Hintum, T. J. L., I. W. Boukema and D.L.Visser. (1996). "Reduction of duplication in a Brassica oleracea germplasm collection." Genetic Resources and Crop Evolution 43(4): 343-349.

van Hintum, T. J. L. (2000). "Duplication within and between germplasm collections. III. A quantitative model." Genetic Resources and Crop Evolution 47(5): 507-513.

van Treuren, R., A. Magda, R. Hoekstra and T.J.L. van Hintum (2004). "Genetic and economic aspects of marker-assisted reduction of redundancy from a wild potato germplasm collection." Genetic Resources and Crop Evolution 51(3): 277-290.

van Treuren, R., J. M. M. Engels, R. Hoekstra and T.J.L. van Hintum (2009). "Optimization of the composition of crop collections for ex situ conservation." Plant Genetic Resources: Characterisation and Utilisation 7(2): 185-193.

Appendices

1- Agenda Duplicate Finder brainstorm session
2- Feedback and improvements of the Duplicate Finder (DF)
3- Short manual DuplicateFinder
4- Fine-tuning the DuplicateFinder

APPENDIX 1

Agenda Duplicate Finder brainstorm session

Location: PRI, Wageningen, The Netherlands

9:00 Theo – welcome, discussion of the agenda, and introduction of the project 'Duplicate Finder'

9:15 Frank – the first attempts at CGN to help curators select MOS for the European collections

9:30 Roel and Theo – explanation and demonstration of the things done so far in the project

10:15 biological break

10:30 Christoph – work on the Avena database regarding the handling of duplicates

11:15 Helmut – work on identifying duplicates in barley done at IPK (KWIK etc.)

12:00 discussion regarding the presentations

12:30 lunch at 'Restaurant van de Toekomst' (= restaurant of the future = a scientific experiment studying your behaviour in restaurants !)

14:00 discussion regarding

- the general approach of Duplicate Finder
- the algorithms to use in Duplicate Finder
- the next steps and involvement of IPK and JKI in the project

16:00 visit to CGN genebank (Roel) and demo of GENIS, Genis web, etc. (Frank)

APPENDIX 2

Feedback and improvements of the Duplicate Finder (DF)

Below are listed the observations of the testers of the Duplicate Finder (DF), and the improvements made on the basis thereof. The testers were : C. Germeier (JKI), H. Knüpffer and M. Oppermann (IPK) and R. van Treuren and F. Menting (CGN).

1) **For the "Make duplicate groups" task the maximum possible number of records is unclear. The message on the out-of-memory error, advising to reduce the number of records, pops up after about a quarter of an hour.**
    - Despite having another 2GB of free RAM available, Excel can only use about 630MB of memory. Closing other applications has therefore no effect. An early test has been included, to check if the software can perform the grouping with the current number of records. Buffer clearing algorithms have been included, but have only partial effect. To clear Excel buffers it may help to save, close and re-open the file after having copied/deleted data in the sheet. A comment has been included in the manual accordingly.
    - Three arrays in the clustering module have been declared as INTEGER instead of LONG, saving memory.
    - For grouping, the maximum number of records appears to be exactly 15000. The calculations take 5 hours and 45 minutes on a fast computer. 96% of the time is used for the clustering algorithm. Calculation time increases exponential with the number of records, grouping the demo set of 1000 lettuce records takes only 12 seconds.
    - For the "Find duplicates for one accession" task, high numbers of records are no problem (tested on 164 000 accessions of wheat).

2) **After the out-of-memory error message, the cursor remained in the "egg timer" shape.**
    - DF adjusted, problem solved.

3) **Concerning the limited number of records for the "Make duplicate groups" task, how should a curator subdivide his data into meaningful subsets?**
    - The curator knows his crop and should be able to subdivide his records in a meaningful way, e.g. splitting wild and cultivated germplasm. Smaller subsets need less calculation time!
    - Most of the crops are rather small. Only the main cereals will have problems. If the curator really needs to run subsets >15000 then the current DF in Excel is not suited.

4) **The MCPD export from EURISCO contains LATITUDED and LONGITUDED (decimal degrees), which are rejected by the Duplicate Finder.**
    - This could have been avoided by renaming the headers into LATITUDE and LONGITUDE, which makes them recognised by DF, and subsequently from the

"Adjust content" menu run "LON to MCPD format" and "LAT to MCPD format", which would have converted the decimal values to degrees, minutes and seconds. A reference to these "Adjust content" options is added to the manual.

5) **The "GENUS remove trailing blanks" task would be useful for many other columns.**
   - The task has been expanded to "remove trailing blanks from one column", meaning that the user can remove trailing blanks from any column of his choice.

6) **The EURISCO MCPD export has an incorrect date format for incomplete dates (e.g. 20110000 instead of 2011----). An attempt to use the DF function of transforming into correct MCPD date led to deletion of the date values, instead of reformatting.**
   - This was a bug in the DF that has been corrected: problem solved.
   - Dates where month and day have been exchanged (e.g. 19753105) will not be reformatted and remain marked as faulty. Presumably the complete subset from a specific source needs to be reformatted, which will include dates that seem fine (e.g. one day later: 19750106).

7) **The functionality of some menu options is unclear.**
   - The "Recalculate data" option has been placed one level deeper in the menu to clearly separate it from the two duplicate finding options. It has an additional "Read me" button explaining its functionality.

8) **A recursive operation would be useful in many cases (search additional duplicates with the information from the previous ones found).**
   - This is an advanced feature which cannot be implemented in the framework of the current project. It may be considered for a later version.

9) **The DF seems to group only a part of the complete data set.**
   - Many accessions cannot be linked to others and remain solitary. Groups containing only one member do not get a separate group number.
   - Possibilities for fine tuning the software is described in the manual, however clustering will never group all accessions, since some simply lack the information to base grouping on.

10) **For groups that seem to be largely based on the Soundex algorithm (on ACCENAME) the grouping is not always satisfactory.**
   - Excluding common names like (landrace or seabeet) from getting a soundex by including them in the (hidden) sheet NoSndx, followed by recreating the (hidden) DATA3 sheet using the "Recalculate data" option could improve this situation. This is explained in the Manual for Fine-tuning.
   - Obviously it is impossible for the algorithm to link, for example, Meikoningin with May Queen. The results of DF tool only serve as a starting point in the search for duplicates, and will always have to be checked and improved by the curator!

11) It is meaningless to evaluate accession numbers without considering the institution codes. The same would apply for donor numbers without donor codes and collecting numbers without collector codes.

- As explained in the manual, the (hidden) sheet SimRules illustrates how the DF deals with these issues. The corresponding institution codes are not used, but numbers without a Prefix get a lower similarity value. Furthermore, the similarity value is multiplied with a factor (0 – 1), which is low for small numbers and high for long numbers. This should prevent over estimation of similarities.
- Since the DF aims at identifying *potential* duplicates, to be further screened by the experts, we think it is important to bring together any records with possibly informative matches.

12) Equal numbers without a prefix (even with length 4) in for example COLLNUMB and OTHERNUMB hardly contribute to the similarity.

- See also comment 11. This is a dilemma. Giving higher ratings may be desirable in one data set and not in another. We have adjusted the rates on numerical matches trying to fix this issue, however the dilemma remains.

13) Institute codes included in the OTHERNUMB column cause unrealistic similarities.

- In the sheet DATA3 the Institute codes are automatically removed from the PREFIX column, which solves the problem.

14) In TestsetReduced.xlsm (data from MS-Access) improper data were deleted. Nevertheless runtime error 13 occurs.

- The file was re-tested by the developers. An error value in the NUMB column of the hidden sheet DATA3 occurred. The record where the error occurred differed depending on the language of the Windows system (English or Dutch). All error values in the sheet are now deleted automatically, meaning that the calculation of the similarities is not being disrupted anymore. The deleted information will be copied in the Prefix column.

15) Provide the DF tool also for older Excel versions.

- It was converted to and tested with Excel 2000 and functions well, when working with smaller number of records (e.g. 5000). The Excel memory problem (topic 1) is even a bigger problem in lower Excel versions. This is for instance shown by errors in the macros using the application.transpose functionality. This would require specific workarounds for lower Excel versions.
- The conversion to a lower version can be performed by the user or a colleague. If necessary the developers will provide a converted version with the warning that the DF was not properly tested for this version.

16) For the less experienced user, the import of Eurisco data should be explained in the documentation, also that Macros need to be activated.

- This will be taken into consideration.

17) Excel seems to be a little-suited platform for programming the DF tool, due to the inbuilt data type conversions, which are difficult to control by the user. For the import of EURISCO data one should convert all columns into

“text” format, to avoid uncontrollable conversions of text and numbers into dates. It would have been better to implement the functionalities in another programming environment (e.g. MS-Access) and to make it even independent of the underlying database system.

- Excel is spread sheet software and not intended for database use. However, since most users are familiar with Excel, because of its user-friendliness and its extensive calculation functionality, and the DF is only a first step in the search for duplicates, we think it is best suited for this task. The user can continue the search with changing environment.

APPENDIX 3

short manual

DuplicateFinder

DuplicateFinder is a tool to support you in your search for duplicates and most appropriate accessions, it doesn't do the job for you. It will still be difficult and time consuming to find the duplicates. We just hope DuplicateFinder will help you in this effort and save you a little time.

Introduction

DuplicateFinder is a little piece of software to help identify potential duplicate PGR samples on the basis of passport data. It is simple to use, and will help anyone analysing a Central Crop Database, or a local PGR documentation system by creating groups of accessions that are likely to contain the potential duplicates. In the end the user will have to decide about which accessions (s)he thinks are actual duplicates.

The software comes in a MS Excel environment: it is a spreadsheet with macro's. The user only has to copy the passport data in the spreadsheet, and run the macro's either (1) to identify the potential duplicates of a selected accession or (2) to create groups of material which can contain the duplicates.

Restrictions and limitations

1 – The passport data have to be formatted according to the Multi Crop Passport Descriptor (MCPD) list. Or, actually, the software currently only uses the descriptors ACCENUMB, ACCENAME, COLLNUMB, DONORNUMB and OTHERNUMB, so these fields have to have the right headers in the sheet so that the software can recognise them.

2 – The fact that only these five fields are used, implies that other fields which might contain clues about duplication, such as taxonomic and origin location fields, are not used. Duplicates that can be identified only on these fields will not be identified by the software!

3 – The first time the software is used, it will take some time (depending on the hardware and size of the dataset, a few minutes max) to prepare the data structures used in the background for the identification of duplicates (hidden sheet 'DATA3').

4 - The creation of potential duplication groups can take much longer (up to a few hours for large data sets), and can only be run on data sets with 15,000 accessions or less. If the data set is larger, the user has to restrict the number of records to be processed for each run and thus, might have to form smaller sub-sets of data in order to stay within the above mentioned number limit. When large numbers of records are to be processed it can be useful to clear the memory buffers by saving, closing and re-opening the file.

5 – The software can change the format of your data sheet (layout, colour, size, etc.). It will never change the data values, except in the macros under the ‘Adjust content’ option!

Steps

1 – Open DuplicateFinder

Open the spreadsheet DuplicateFinder.xlsm in such a way that macro’s can be run. This might require some changes in the security setting of Excel, but most likely only involves clicking ‘accept’ at some stage. (This manual is assuming you are using Excel 2010 or later, if you are using an older version of Excel, you might have to look for the macros mentioned below since the menu structure might be slightly different – but everything will function also in earlier versions.)

2 – Copy your data

Make sure the data you want to analyse are in a spreadsheet, ready to be copied in the sheet ‘DATA’ of DuplicateFinder. Delete the sample data in the sheet ‘DATA’ and copy your data in.

3 – Check format of your data

If you are not sure about the formatting, check the sheet ‘MCPD List’ of DuplicateFinder. It lists the descriptors of the Multi Crop Passport Descriptor (MCPD) list. Make sure your data in the sheet ‘DATA’ have at least proper headings for the descriptors ACCENUMB, ACCENAME, COLLNUMB, DONORNUMB and OTHERNUMB. If you like you can have DuplicateFinder check the format by running the macro ‘ValidateAllColumns’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Validate columns’ and ‘Validate all columns’). In the sheet ‘Report’ you will see what columns were found or missing, what values were missing, wrongly formatted or wrongly coded. In the data sheet the recognized columns will have bold headers, and the wrong values in these columns will be given a red background. The macros provided under the option ‘Adjust Content’ will help you reformatting. If you have corrected headers or fields and want to check again, you can restrict the check to the column you changed (run the macro ‘ValidateOneColumn’ by choosing the option ‘Validate one column’). If you want to change the formatting of some of the MCPD columns, check the menu option ‘Adjust content’, but be aware that these macros do change the content of the data!

4 – Find potential duplicates of one accession

Select a cell in the record of the accession you want to match with the others, and run the macro ‘FindDuplicatesOneAcc’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Find duplicates’ and ‘Find duplicates for one accession’). For the first search the software might need some time to prepare the required data structures (few minutes maximum). The software will create two new columns DFSim and DFIDno, if they were not created before. DFIDno will contain temporary unique ID numbers of each accession – you can ignore it, it will be hidden after the calculations have been finished. The column DFSim will show the similarity between the selected accession (starting with the highest similarity value at the top of the sheet) and others that are displayed in decreasing order of similarity.

5 – Create potential duplication groups

If you run the macro ‘MakeDuplicateGroups’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Find duplicates’ and ‘Make duplicate groups’), DuplicateFinder will create groups with similar accessions. Since each accession needs to be compared with each other accession, this might take a while (up to a few hours, depending on hardware and number of records). The result of all these calculations will be placed in a new column called DFGrp, where similar accessions will be given the same group number. Accessions that were not clustered with others will have no group number.

Careful

DuplicateFinder is designed to identify potential duplicates, not to provide an environment to edit data. However, if you do decide to change the data, and want to continue searching duplicates or creating groups, you need to recalculate the similarities between the accessions. For this purpose you should run the macro ‘RecalculateData’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Find duplicates’ and ‘Recalculate data’). Be aware that the grouping you calculated before is not changed, unless you recalculate it.

Tips for use

- First play some time with the software, using the 1000 sample records that come with the spreadsheet, and see the possibilities. Be aware that everything takes longer if there are more records – the time to create groups is roughly quadratic to the number of accessions – twice as many accessions takes four times as long.

- Create your column(s) to store the results of your inspection of the output. For example, you can create a column GROUP (or something similar) to store the groups you accepted or identified yourself, or you can create a column STATUS, to indicate if an accession is a ‘Most Appropriate’ or a ‘Probable Duplicate’. Based on the results of running the macro’s you can fill and change the values in these columns.

- If the number of accessions in your data set exceeds 15,000, you should select a homogeneous group (one taxon, only cultivars, etc.) to run the ‘MakeDuplicateGroups’ macro since it cannot handle more accessions and run DF separately for each ‘homogeneous group’.

- If you want to run the ‘MakeDuplicateGroups’ macro, and the number of accessions is high, consider starting it before going home or attending a meeting. Make sure that the ‘Power savings options’ of your computer doesn’t prevent it from continuing to work when you leave the room.

Acknowledgements

DuplicateFinder was developed in the framework of, and with financial support of the AEGIS initiative of the European Cooperative Programme for Plant Genetic Resources (ECPGR) by the Centre for Genetic Resources, The Netherlands (CGN - Roel Hoekstra,

Theo van Hintum, Frank Menting and other staff) with support of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK - Helmut Knüpffer) and Julius Kühn Institute (JKI - Christoph Germeier), both in Germany.

FEEDBACK

If you have problems using the DF, or find bugs that we didn't find, please contact the developers: Roel Hoekstra (roel.hoekstr@wur.nl) or Theo van Hintum (theo.vanhintum@wur.nl).

APPENDIX 4

fine-tuning the

DuplicateFinder

DuplicateFinder is a tool to support you in your search for duplicates and most appropriate accessions, it doesn't do the job for you. It will still be difficult and time consuming to find the duplicates. We just hope DuplicateFinder will help you in this effort and save you a little time.

Introduction

Before you continue, please first read the 'short manual'. For those who are experienced in VBA, this manual gives some possibilities for fine tuning the software.

Exclude specific ACCENAME content from the search process

A soundex is created on the column ACCENAME. Amongst others, this soundex is used to match accessions. Sometimes the content of this field is too unspecific (e.g. MESTNYI or landrace) and should be excluded from the matching process. Therefore, unhide the sheet noSndx (right-click a sheet-tab, chose unhide) and in the column noSOUNDEX you can add names that should be ignored.

For advanced VBA users only:

Parameters for the calculation of the Similarities

The macro 'DoCalcSimilarities' in the module 'CalcSimlarities' calculates the similarities between accessions, using the information stored in the (hidden) sheet DATA3. The hidden sheet 'SimRules' gives an overview of the parameters used in the calculations. Adapting the parameters in that sheet has no effect. The parameters need to be changed in the macro itself.

Make Duplicate Groups

In the macro 'ClusterDuplicationGroups' in the module 'Cluster' the threshold for grouping is set to 30% (SimilarityThreshold = 0.3). You may prefer a higher threshold (e.g. 0.5), this will result in smaller groups since (clusters of) accessions have to be more similar to be joined.

For further clustering of (clusters of) accessions default the MINIMUM similarity option is used, which will create relative small groups. Optionally you can choose MAXIMUM (--> large groups) or AVERAGE similarity (--> medium sized groups) by deactivating the default (put a ' in front of the line) and activate your choice (by removing the ' in the front of the specific line) in de Do-loop.

NB: the hidden sheet ‘MCPD Codes’ is used by the macro’s to check which DESCRIPTORS / SAMPSTAT / COLLSRC / ORIGCTY / INSTCODE / STORAGE values are allowed.