LAMPIRAN 1. MANUAL STR_MATCHSAMPLES

Universitas Indonesia
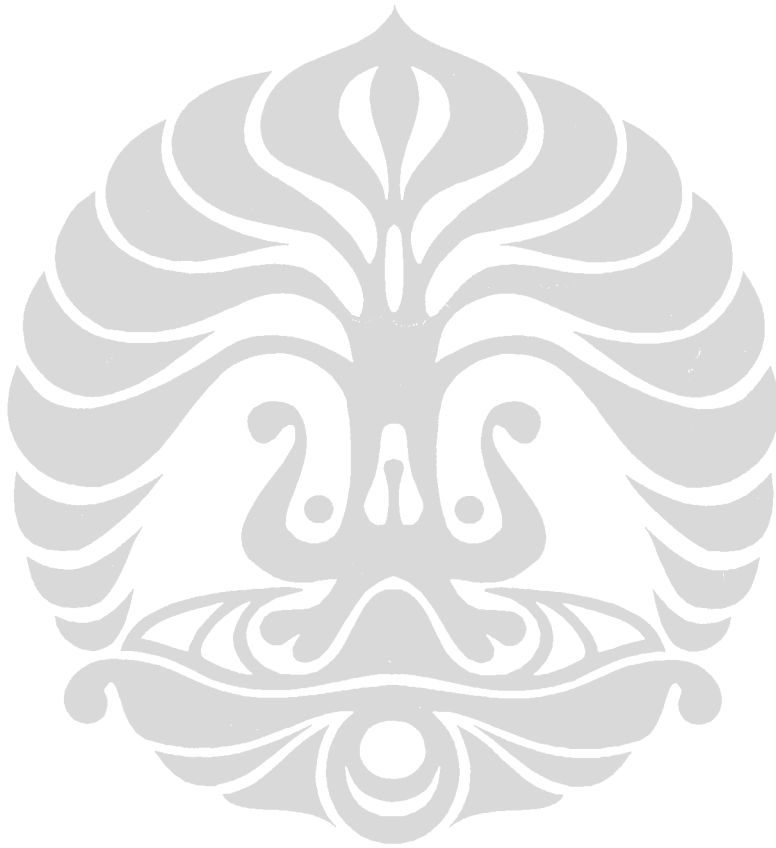
STR_MatchSamples <2-Jun-09>

A System for Matching Samples Across Excel Worksheets by Their STR Profiles

What is a 'Best Match'?

For a given "Match From" profile, the STR_MatchSamples system selects the Best Match "Match To" profiles as:
Those "Match To" profiles with the largest percentage of *loci* that share at least one well-defined allele with the "Match From" profile.
Given the same % matching loci, those "Match To" profiles with the largest percentage of *alleles* shared with "Match From".

In the absence of ambiguous/missing alleles, if A is the Best Match to B then B" will be the Best Match to A.
However, this symmetry may not hold when one or both of the profiles has ambiguous/missing alleles.

ControlPanel Commands

All of the (interesting) STR_MatchSamples commands are accessed from the ControlPanel worksheet.

There are currently six commands, each accessed via it's own Command Button:
ReadMe: Activates this ReadMe "Instruction Manual" worksheet.
Clear: Clears all status reports, working data, and report worksheets.
Add Data: Allows user to specify a particular worksheet and add it to the list of samples-to-match on Data. Shortcut: <Ctrl-A>
What data is added depends on the setting of the radio buttons in the "Use Loci Contained In" group, to the right of the Add Data button.
If "Only the initial dataset" is active, then only those loci specified in the initial Add will be used… What's more, only datasets that have those loci can be added. If one or more loci are missing, the system will notify you of the situation and stop; excess loci are ignored.
If "All datasets" is active, then all loci specified in any added dataset will be automatically used.
If "All datasets, if confirmed" is active, then you will be asked whether a "new-to-the-current-data" locus should be used.
Check Data: Insures that all profiles on worksheet Data are in standard format and have a unique identifying code. Shortcut: <Ctrl-C>
It also identifies samples that have the same profile and creates a composite identifying code for the unique profile.
Best Match: Generates the three Best Match reports (see below).
Depending on the setting of the radio buttons to the immediate right of the Best Match Command Button,
the three Best Match reports match all of the "UnMatched" (I.e., those that were not found to be "Exact Matches")
either to all profiles or just to the other "Unmatched" profiles.
Exit: Gracefully exits the STR_MatchSamples system.

ControlPanel Summary Information

Columns A and B of the ControlPanel worksheet contain the command buttons and the radio-button that specify various user-dettable options.

Columns D to G provide information on the progress of the analysis.
Status: the current condition of the 'Report' worksheets. The status of the Data worksheet is the most important.
The Data states are: Clear, Loading, Loaded, Checking and Checked. Data must first be successfully loaded, then checked.
The RepDetail, RepSummary and RepProfile states are: Clear, Processing and Done!
Last Called: The date-time when the last operation occurred that affected the status of the worksheet.
Sheet: The name of the worksheet
Contains: A short description of the eventual content of the worksheet.

Columns I and J list the number of various categories of sample listed in the Data worksheet.
Type: The match-type of the samples. The four types (well, the total plus two distinct types and a subtype) are:
Total - the total number of profiles listed in the Data worksheet.
ExactMatch - the number of unique profiles that were shared by two or more of the originally loaded profiles.
Unmatched - the number of unique profiles in the originally loaded data that were not shared.
Unmatch (able?) - the number of 'Unmatched' profiles that do not share a sample name with at least one other sample.
#: The number of samples of each type.

Columns L to O document the worksheets that have been successfully used to load profiles into the Data worksheet.
Rows: The number of rows(i.e., new samples) added from the given dataset.
Date Loaded: the datetime when the dataset was added.
Worksheet: the worksheet from which the dataset was loaded.
Directory: the directory holding the worksheet from which the dataset was loaded.
#Loci: the number of loci specified in the dataset.
Loci...: the names of all the loci in the dataset.

Ill-defined Alleles / Genotypes and How to Correct Them

To validly compare profiles, all profiles *should* be as complete as possible and every non-missing genotype unambiguously specified. While "OL" is accepted as the code for an off-ladder allele, an "OL" allele in one system may be a ">x" in another. It is up to you to resolve such differences.
If you were paying attention in STR_ConvertFormats, you corrected all ill-defined designations before you stored the converted profiles. However, do not despair: ambiguities can be resolved without leaving this system.

STR_MatchSamples regards any allele that contains any non-numeric character other than "L", "O", "X", "Y" or "?" to be ill-defined. The most common ill-defined designations are off-ladder designations such as "<" or ">" and hybrid designations from FSSi-format data when the "Designation" alleles and the i-STRess genotype did not *exactly* agree (hybrid designations list all the possible alleles separated by the four characters " or ".) When Check Data encounteres anything it doesn't recognize, it (1) paints the cell's background yellow, (2) selects the cell, (3) complains about it to you and (4) stops. The background of possibly modified desigtnations is painted blue and you are asked to resolve the issue; the background of "OL" (off-ladder) alleles is painted green and you will notified of their presence.

If you know that there are only a very few ill-defined data, your best choice is probably to (1) edit the contents of the selected cell (or just enter the correct value), (2) activate the ControlPanel worksheet and (3) re-click Check Data. Repeat until the complaints stop.
The shortcut key <Ctrl-C> (that is, the combination of the Control, Shift key and C keys) allows you to skip step (2).

When there are many ill-defined data of similar kind then it may well be more efficient to use Excel's Replace… function; this function is in the Edit menu and can be invoked with the <Ctrl-h> shortcut.
With this you can replace multiple occurances of a given string of characters throughout a selected range of cells or, if only one cell is selected or the entire sheet is selected, thoughout the entire worksheet.
If you are very confident, you can use Replace…'s Replace All command to do everthing all-at-obce but its safer to step through one-at-a-time using its Replace command. Typical use would be replacement of the hybrid construct " 0 or " with a null (nothing in the box) or space (" ").
After you correct everything that you know needs correcting, activate the ControlPanel and re-click Check Data (or <Ctrl-C>). Repeat until the complaints stop. Or you become completely frustrated, whichever comes first.

Of course, figuring out what the correct alleles / genotypes should be may be a bit of a puzzle… that's why you shoulda done it in STR_ConvertFormats when you had the pre-converted data to look at. When in doubt, replace ill-defined alleles with the "?" missing data flag.

The Data Worksheet

The Data worksheet contains all of the loaded data. It has three structural elements:
Title (cell A1): this lists all of the worksheets loaded via Add Data.
SampleCode (columns A and B): the "Type" and "Description" of the unique profiles.
The initial sample "Description" is a combination of the SampleCode as listed in the originating worksheet and that worksheet's name, separated by a ":". A unique identity code is assigned to all samples from the same worksheet with different profiles but identical names.
There are three possible "Types": ExactMatch, Unmatched and Unmatch (able?).
All samples start as "Unmatched". If the first part of the "Description" (that is, the SampleCode) is *not* shared by at least one other unique profile, the sample will be given the Type "Unmatched (able)?"
If two or more samples share a profile, they will be replaced by a single, unique profile with a "Description" that is a composite of all the initial "Descriptions" and the Type will be "ExactMatch".
Loci (columns D to ?): the alleles present at each locus in compact, comma-separated format.

Note: if you are attempting to find the "exact match" profiles for two or more evaluations of the same set of samples, this Data worksheet report may be all you need. It will find exactly matching profiles and SampleCodes and will clearly identify samples that exactly match neither profile nor SampleCode.

	A	B	C	D	E	F	...
1	[TEST_012006.xls]FSSI						
2							
3		SampleCode			Loci		
4	Type	Description		D8S1179	D21S11	D7S820	
5	Unmatched	AF01C:FSSI		13	?	?	
6	Unmatched	AF02C:FSSI		11,13	28,29	8,9	
...							

STR_MatchSamples <2-Jun-09>

A System for Matching Samples Across Excel Worksheets by Their STR Profiles

The RepDetail Worksheet

The RepDetail worksheet reports the 10 Best Match profiles to all or every Unmatched sample.

The Title (cell A1) is the only "global" element. The remainder of the report consists of individual sample matches, formatted in groups of three. The format elements for each sample are:

Sample From (first column of each repeat): the "Description" of the sample being matched. The number of unique, fully specified alleles in the profile is also listed ("# Alleles") as is the number of loci with at least one fully specified allele ("#Loci").

Sample To (second column of each repeat): the "Description" of the matched sample.

%Allele (third column of each repeat): the percent of alleles in common between the "From" and "To" samples.

%Loci:1+ (fourth column of each repeat): the percent of loci common to both samples that share at least one fully specified allele.

The ten matched samples are sorted in order of decreasing %Loci:1+, then by %Allele.
Ambiguous or undetermined samples (those represented as "?" in the profile) are ignored in this analysis.

This report may be of some use when evaluating profiles from related individuals. However, it's primary utility is as the source of data for the following two reports.

Details of 'Best Match' of Unmatcheds to other Unmatched samples

	A	B	C	D	E	...
3	Sample From	Sample To	%Allele	%Loci:1+		Sample From
4	AF01C:FSSI	C25B:FSSI	57	63		AF02C:FSSI
5	#Alleles: 21	C05C:FSSI	52	63		#Alleles: 24
6	#Loci: 16	C11C:FSSI	52	63		#Loci: 16
7		C15C:FSSI	52	63		
8		C17C:FSSI	52	63		
9		C01C:FSSI	48	63		
10		AF97B:FSSI	52	56		
11		C97B:FSSI	52	56		
12		AF11C:FSSI	48	56		
13		C07C:FSSI	48	56		
14						
15	AF04C:FSSI	C04C:FSSI	69	100		AF05C:FSSI
...						

The RepSummary Worksheet

The RepSummary worksheet reports the best two Best Match profiles to all or every Unmatched sample,

The Title (cell A1) is the only "global" element. The remainder of the report consists of individual sample matches, one sample per row. The format elements for each sample are:

Sample From (Column A): the "Description" of the sample being matched.

Alleles (column B) is the number of fully specified alleles in the profile.

Loci (column C) is the number of loci with at least one fully specified allele.

Best Match To (column E): the "Description" of the best Best Match sample.

%Allele (column F): the percent of alleles in common between the two samples.

%Loci:1+ (column G): the percent of loci common to both samples that share at least one fully specified allele.

Next Best Match To (column I): the "Description" of the next-to-the-best Best Match sample.

%Allele (column J): the percent of alleles in common between the two samples.

%Loci:1+ (column K): the percent of loci common to both samples that share at least one fully specified allele.

Summary of Two 'Best Match'es to other Unmatched samples

	A	B	C	D	E	F	G	H	I	J	K
3	Sample	#Alleles	#Loci		Best Match	%Allele	%Loci:1+		Next Best	%Allele	%Loci:1+
4	C11C:FSSI	28	16		AF11C:FSSI	75.0	100.0		AF12C:FSSI	53.6	81.3
5	AF07C:FSSI	27	16		C07C:FSSI	70.4	100.0		AF82B:FSSI	55.6	81.3
6	C12C:FSSI	27	16		AF12C:FSSI	70.4	100.0		AF20C:FSSI	59.3	87.5
7	AF11C:FSSI	30	16		C11C:FSSI	70.0	100.0		C16C:FSSI	53.3	81.3
8	AF03C:FSSI	26	16		C03C:FSSI	69.2	100.0		C09C:FSSI	50.0	68.8
...											

The RepProfile Worksheet

The RepProfile worksheet provides a detailed analysis of the differences between every Unmatched sample and its Best Match.

The Title (cell A1) is the only "global" element. The remainder of the report consists of pairs of profiles that are each other's Best Match. A blank line separates each sample pair. The format elements for each pair are:

Type (Column A): the "Type" of the samples (see Data, above).

Description (Column A): the "Description" of the samples (see Data, above).

Loci (columns D on) contain the compact-format profiles of the paired samples. The profiles are color-coded according to how well they match.

a) Loci with exactly the same, fully-specified alleles,
b) Loci with ill-defined alleles,
c) Loci with fully-specified alleles and that share at least one allele but not all alleles and
d) Loci with fully-specified alleles but that share no allele.

Profiles for Unmatched and its 'Best Match' to other Unmatched samples

	A	B	C	D	E	F	G	H	I	J	K
3							Loci				
4	Type	Description		D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539
5	UnMatched (able?)	AF01C:FSSI		13	?	?	?	16,17	6,9	11	11
6	UnMatched (able?)	C25B:FSSI		13,16	30.2,32	10	?	15,17	6,9	12	9,10
7											
8	UnMatched (able?)	AF02C:FSSI		11,13	28,29	8,9	11	14,15	9.3	11	9,12
9	UnMatched (able?)	C02C:FSSI		11,13	29,31.2	9,12	11	15,18	9.3	10,11	10,12
...											

Want to change the color codes used in the RepProfile worksheet?

Change Colors

Just select each of the above profile description cells, row 116 column 2 to row 119 column 2 (I.e, the ones starting with "a)" to "d)") and use the Fill Color tool to set the background color to your desire.
Alternatively, after selecting the desired cell use the formatting tool on the Patterns tab of the Cells… command under the Format menu - or right-click and use the tool on the Patterns tab of the Format Cells command.
After you have modified one or more of the cells, click the Change Colors button. The chosen colors will be applied to the example (to the right of the description cells) so that you may review your choices.
If you don't like the effect, change the colors and click the Change Colors button once again. Stop when you are happy with the colors.

The Other Worksheets

The ControlPanel provides access to the various command, displays the current status of the analysis and lists the profile worksheets that have been loaded.

The Introduction and ReadMe worksheets provide some guidance into the use of this system (but you obviously already have already figured that out...)

There are also three normally hidden worksheets: Macro which holds the Excel Macro code which does the analyses, Work which is used for a number of transient calculations and DateTime which is used to set the date and time display formats.
The first column of Work also lists the alphanumeric characters that can be present in a "valid" allele specification. If you want/need to view any of these worksheets, you can 'unhide' them using Format > Sheet > Unhide > sheetname.
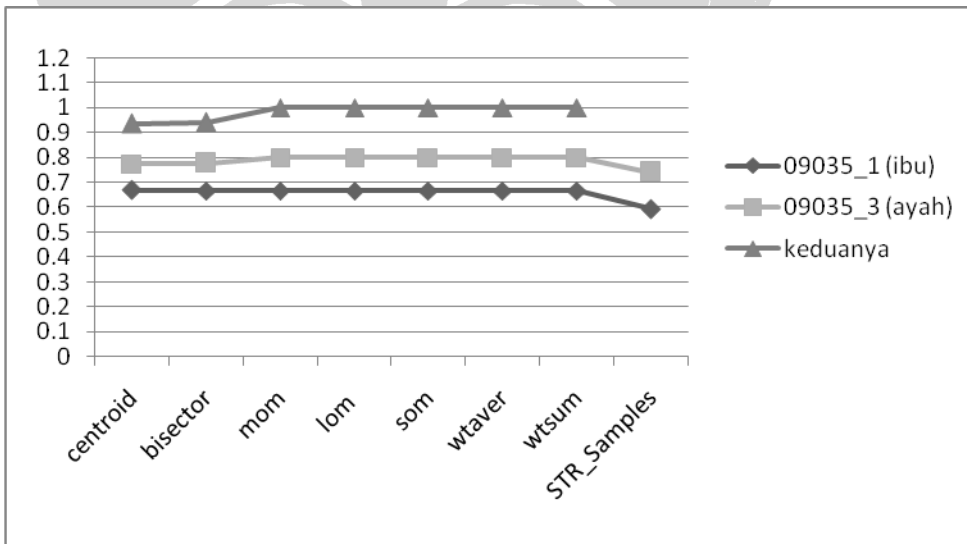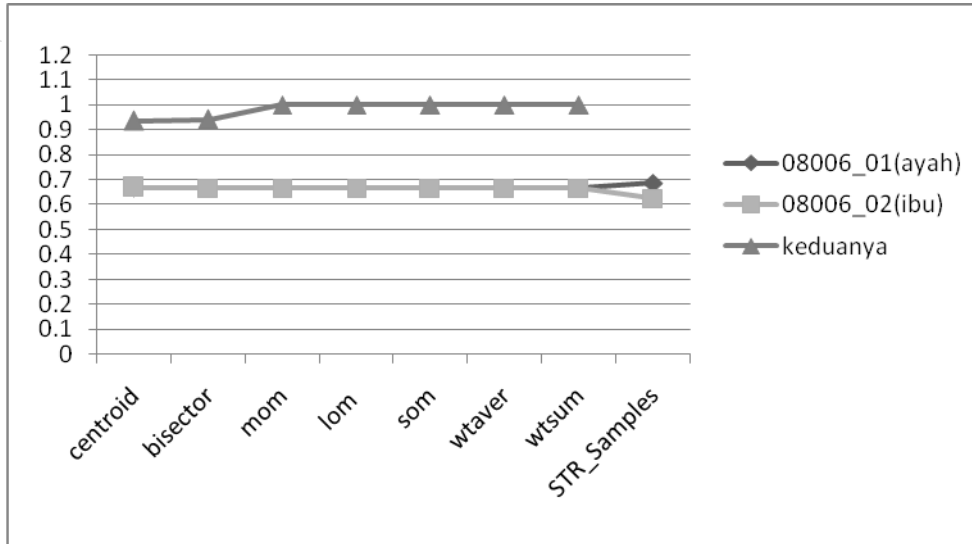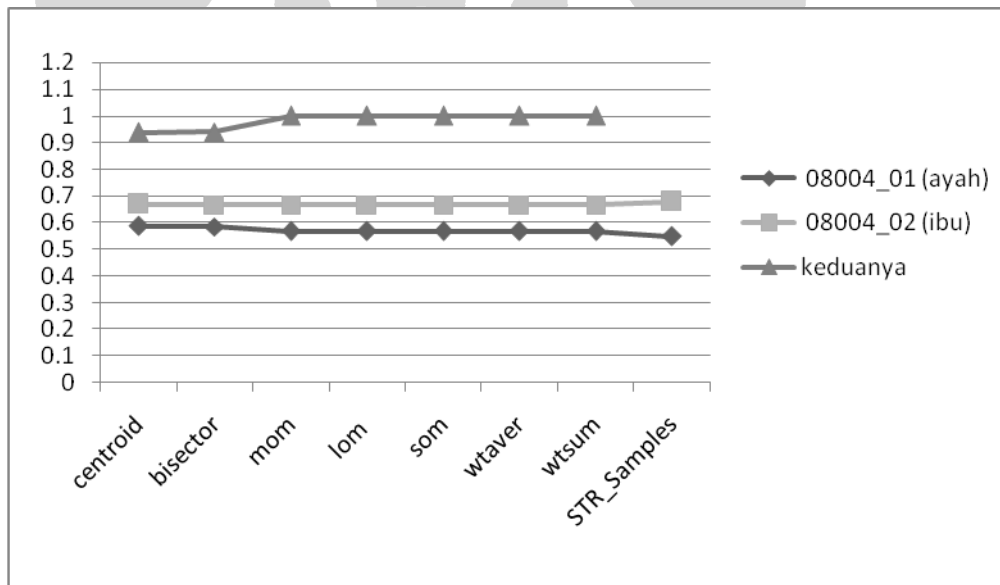
LAMPIRAN 2. CONTOH HASIL UJI COBA

Universitas Indonesia
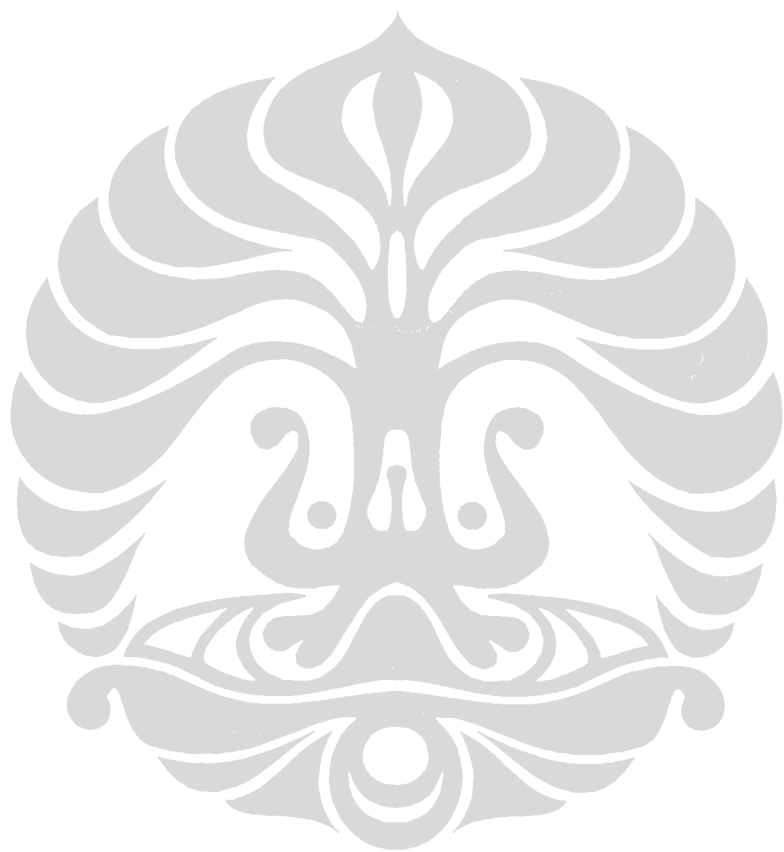
09038_2 (anak)

09035_2 (anak)

i
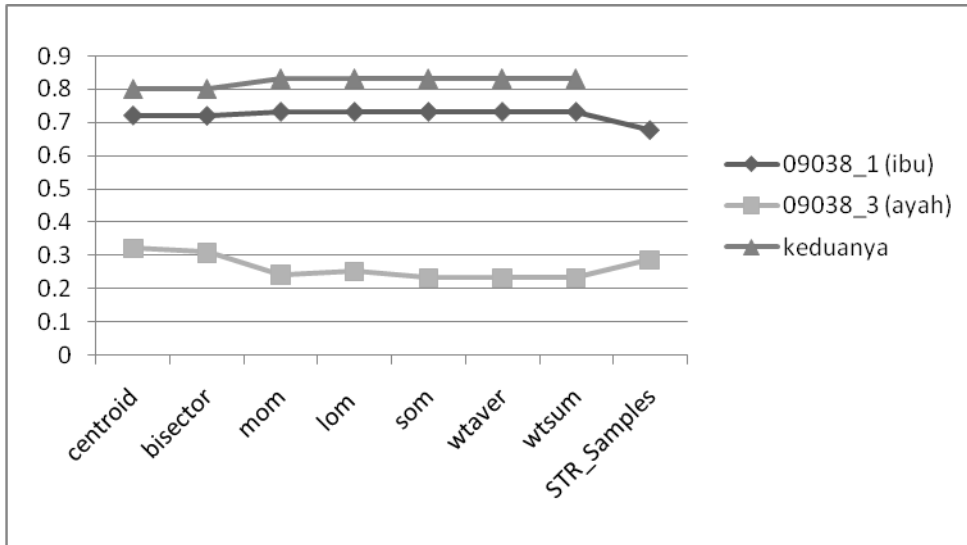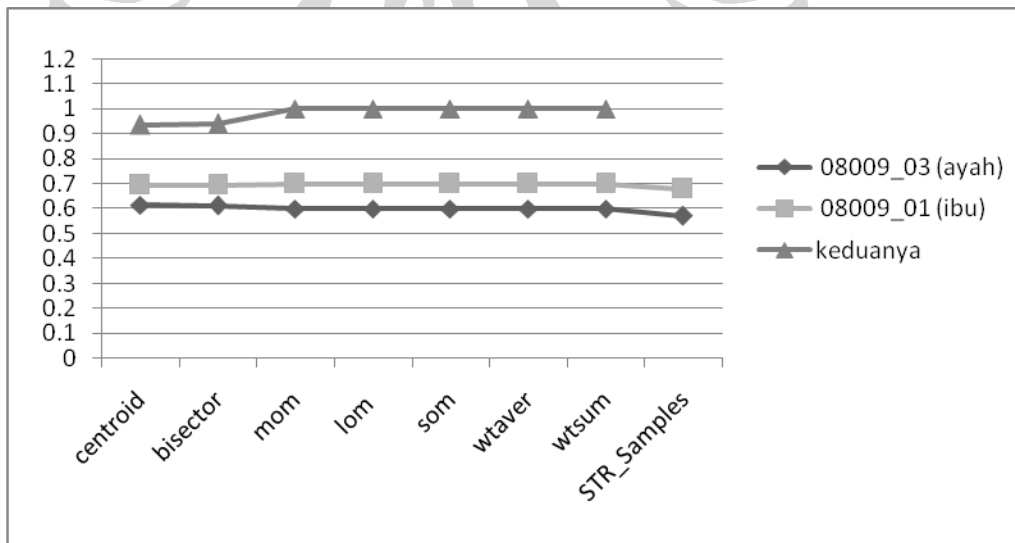
08006_03 (anak)

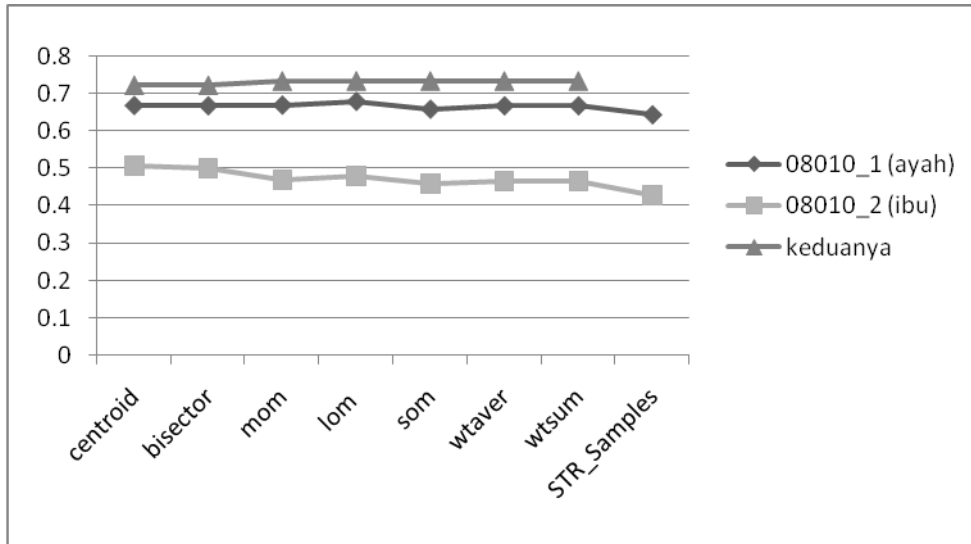**08004_03 (anak)**

08002_C (anak)

08009_02 (anak)

08010_3 (anak)

08020_3 (anak)

iv
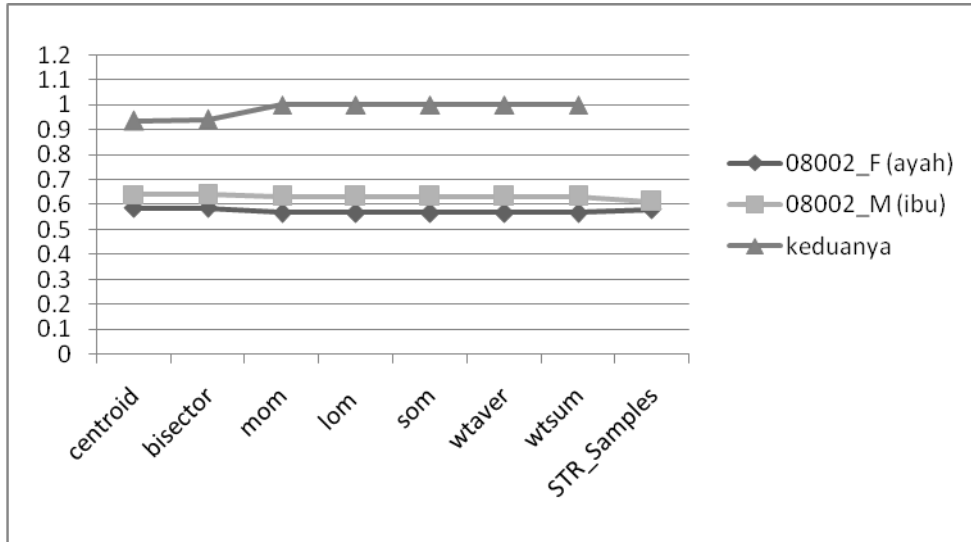
LAMPIRAN 3. CONTOH HASIL PENCARIAN DALAM BASISDATA

Universitas Indonesia

08002_C (anak)

resultDatabase

Legend:
- 08002_M
- 08002_F
- JT02
- 08126643665

Nilai Kemiripan

Nilai Kemiripan Tiap 2 Profile

08002_M: 0.641
08002_F: 0.587
JT02: 0.533
08126643665: 0.533

Lihat Detail

08004_03 (anak)

resultDatabase

Legend:
- 08004_02
- 08004_01
- 09035_3
- JT64

Nilai Kemiripan

Nilai Kemiripan Tiap 2 Profile

08004_02: 0.668
08004_01: 0.587
09035_3: 0.563
JT64: 0.56

Lihat Detail

i

08006_03 (anak)

08009_02 (anak)

08010_3 (anak)

08020_3 (anak)
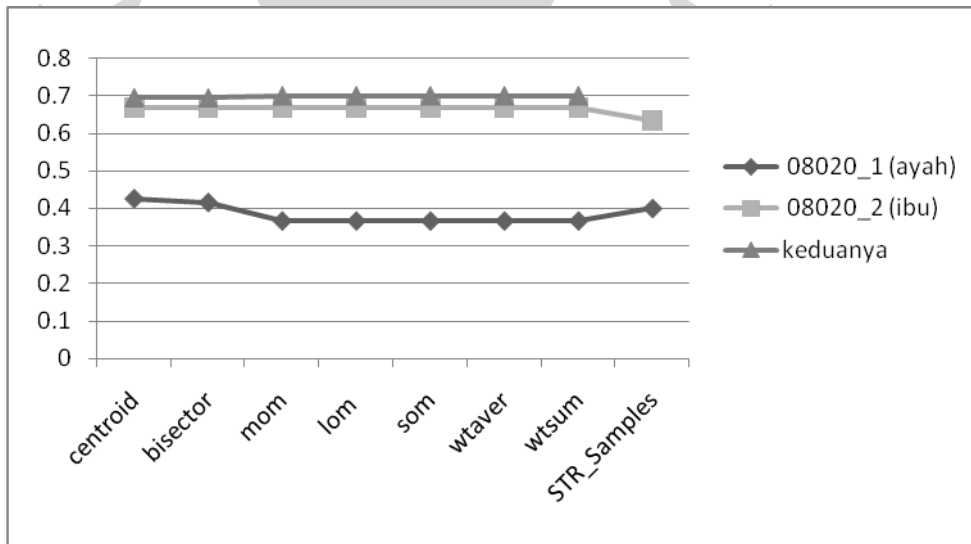
iii

resultDatabase

Nilai Kemiripan

Nilai Kemiripan Tiap 2 Profile

09035_3: 0.775
09035_1: 0.668
08010_1: 0.641
JT38: 0.614

Lihat Detail

09035_2 (anak)

resultDatabase

Nilai Kemiripan

Nilai Kemiripan Tiap 2 Profile

09038_1: 0.722
JT79: 0.536
JT18: 0.533
08122093904: 0.533

Lihat Detail

09038_2 (anak)

iv