

User's Guide



Version 4.3

a product of

S.A.F.E.

**Software Analysis & Forensic Engineering
Corporation**

Table of Contents

Introduction	1
CodeSuite	1
BitMatch.....	1
CodeCLOC	1
CodeCross.....	2
CodeDiff.....	2
CodeMatch	3
FileCount	3
FileIsolate	3
SourceDetective	4
Copyrights, Trademarks, Patents	5
Using CodeSuite	7
System Requirements	7
Licenses	8
The Menu and Toolbar	10
BitMatch	17
Running BitMatch	17
BitMatch Algorithms.....	19
BitMatch Basic Report	20
BitMatch Detailed Report.....	22
CodeCLOC.....	25
Running CodeCLOC.....	25
CodeCLOC Algorithm.....	27
CodeCLOC Basic Report.....	29
CodeCLOC Detailed Report	32
CodeCross	33
Running CodeCross	33
CodeCross Algorithm.....	35
CodeCross Basic Report	36
CodeCross Detailed Report.....	38
CodeDiff	41
Running CodeDiff	41
CodeDiff Algorithm.....	44
CodeDiff Basic Report	45
CodeDiff Detailed Report.....	47
CodeMatch.....	49
Running CodeMatch	49
CodeMatch Algorithms	52
CodeMatch Basic Report.....	54
CodeMatch Detailed Report	57
FileCount.....	61
Running FileCount.....	61
FileIsolate.....	63
Running FileIsolate	63

User's Guide

Statistics.....	67
Calculating Statistics.....	67
SourceDetective	69
Running SourceDetective	69
Exporting databases.....	71
Filters.....	71
HTML reports.....	78
CLOC Spreadsheets.....	79
Distribution Spreadsheets.....	83
Search Spreadsheets	86
Summary Spreadsheet	88
Languages	91
Languages Supported	91
Advanced topics.....	93
CodeSuite database format.....	93
CodeSuite filter format.....	98
Command line interface.....	100
Contacting SAFE Corporation	107
Contacting SAFE Corporation	107
Index	109

Introduction

CodeSuite

CodeSuite® is a collection of computer code analysis tools. The individual tools that comprise the suite of tools include BitMatch®, CodeCLOC®, CodeCross®, CodeDiff®, CodeMatch®, FileCount™, FileIsolate™, and SourceDetective®, all of which are described below.



BitMatch uses fast, simple algorithms to compare thousands of executable binary files in multiple directories and subdirectories to thousands of other executable binary files or source code files in order to determine which files are the most highly correlated. BitMatch is particularly useful for finding programs that have been copied, but where you only have access to the program executable binary files and not the source code.

BitMatch compares every file in one directory with every file in another directory, including all subdirectories if requested. BitMatch produces a database that can then be exported to an HTML basic report that lists the most highly correlated pairs of files. You can click on any particular pair listed in the HTML basic report see an HTML detailed report that shows the specific items in the files (strings or identifiers) that caused the high correlation.

BitMatch examines all text strings, comments, and identifier names that it can find in the executable files in order to determine copying. If a specific user message or a unique subroutine name is found in two files, there is a possibility that one was copied from the other. Note that BitMatch gives only a rough determination whether copying took place. False positives and false negatives are both possible. CodeMatch is needed to compare source code to help make a definitive determination.



CodeCLOC uses a fast, simple algorithm to calculate development progress across two different versions of software, measuring code changes, the amount of original code in your new code, and the amount of original code. CodeCLOC allows measurement of

software changes and software intellectual property changes for calculating such things as transfer pricing.

CodeCLOC compares every file in one directory with every file in another directory, including all subdirectories if requested. CodeCLOC produces a database that can then be exported to an HTML basic report that lists all files in one folder that have statements that match comments in files in the other folder. You can click on any particular file pair listed in the HTML basic report see an HTML detailed report that shows the specific lines in the files that match. CodeCLOC also produces spreadsheets showing detailed statistics about code growth from version to version.



CodeCross uses a fast, simple algorithm to compare thousands of source code files in multiple directories and subdirectories to find programming statements in one file that have been commented out of another file -- a possible sign of copying.

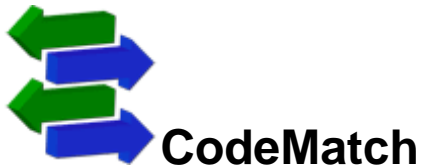
CodeCross compares every file in one directory with every file in another directory, including all subdirectories if requested. CodeCross produces a database that can then be exported to an HTML basic report that lists all files in one folder that have statements that match comments in files in the other folder. You can click on any particular file pair listed in the HTML basic report see an HTML detailed report that shows the specific lines in the files that match.



CodeDiff uses a fast, simple algorithm to compare thousands of source code files in multiple directories and subdirectories to find files that are exact matches or nearly exact matches. CodeDiff looks for identical lines in pairs of source code files. While not as sophisticated or as accurate as CodeMatch, CodeDiff runs much faster. CodeDiff is particularly useful for comparing files where it is already known that many of the files are nearly identical. CodeDiff can be run as a precursor to running CodeMatch when attempting to find source code plagiarism.

CodeDiff compares every file in one directory with every file in another directory, including all subdirectories if requested. CodeDiff produces a database that can then be exported to an HTML basic report that lists the most similar pairs of files based on matching lines of source code in the files. You can click on any particular pair listed in

the HTML basic report see an HTML detailed report that shows the specific lines in the files that are different.



CodeMatch compares thousands of source code files in multiple directories and subdirectories to determine which files are the most highly correlated. This can be used to significantly speed up the work of finding source code plagiarism, because it can direct the examiner to look closely at a small amount of code in a handful of files rather than thousands of combinations. CodeMatch is also useful for finding open source code within proprietary code, determining common authorship of two different programs, and discovering common, standard algorithms within different programs.

CodeMatch compares every file in one directory with every file in another directory, including all subdirectories if requested. CodeMatch produces a database that can then be exported to an HTML basic report that lists the most highly correlated pairs of files. You can click on any particular pair listed in the HTML basic report see an HTML detailed report that shows the specific items in the files (statements, comments, strings, identifiers, or instruction sequences) that caused the high correlation.

CodeMatch uses unique algorithms to find various different ways that source code files are correlated. These algorithms can find directly copied source code and even source code that has been modified to avoid detection.



FileCount is a utility that counts the number of files, non-blank lines, and bytes in a large set of files in a directory tree. FileCount is useful when using CodeDiff to generate statistics about a set of source code files.



Filesolate is a utility that allows files to be selectively deleted from a large group of files in an entire directory or directory tree. Filesolate is useful when examining a large number of files but only certain files are of interest and all other files can be deleted to make searches faster.



SourceDetective

SourceDetective is a utility that searches the Internet for all references to matching statements, comments, and identifiers found in a CodeSuite database. SourceDetective is used to determine whether statements, comments, and identifiers found in two sets of files are commonly used or not, depending on how many references can be found on the Internet.

Copyrights, Trademarks, Patents

Copyrights

The materials in this users guide are copyright 2005-2009 by Software Analysis and Forensic Engineering Corporation.

All written materials from SAFE Corporation regarding CodeSuite, BitMatch, CodeCLOC, CodeCross, CodeDiff, CodeMatch, FileCount, FileIsolate, and SourceDetective, including the material in this User's Guide and the source code for all versions of CodeSuite, BitMatch, CodeCLOC, CodeCross, CodeDiff, CodeMatch, FileCount, FileIsolate, and SourceDetective are the copyright of SAFE Corporation.

Trademarks

SAFE Corporation, the SAFE Corporation logo, the SAFE Corporation brand, CodeSuite, the CodeSuite logo, BitMatch, CodeCLOC, CodeCross, CodeDiff, CodeMatch, FileCount, FileIsolate, SourceDetective, and all other SAFE Corporation product names referenced herein are registered trademarks or trademarks of SAFE Corporation. All other brand and product names mentioned herein are trademarks of their respective owners.

Patents

CodeSuite is covered by U.S. patents 7,503,035 and 7,823,127 and the following pending U.S. patents: 11/467,155, 12/214,128, 12/217,711, 12/253,249, and 12/871,808.

Using CodeSuite

System Requirements

CodeSuite will run on any computer using any of the following versions of the Microsoft Windows operating system

- Windows 2000
- Windows XP
- Windows Vista
- Windows 7

Licenses

Licenses must be purchased from SAFE Corporation. Some functions of CodeSuite including FileCount, FileIsolate, generating HTML reports and spreadsheets, and filtering existing databases do not require a license.

To request licenses, open the authorization form shown below from the Help menu. Send the site code to SAFE Corporation and the number of licenses requested, along with appropriate payment. SAFE Corporation will send back an Authorization Key that must be entered into the field in the form. Press the process authorization button and the form will show the following information. Licenses are enabled for only one PC and cannot be transferred to another PC.



The image shows a Windows-style dialog box titled "CodeSuite Authorization". The dialog contains the following elements:

- Text:** "Please provide an Authorization Key to activate CodeSuite." and "To obtain a license for CodeSuite to run on this machine, email the site code below to SAFE Corporation (authorize@SAFE-corp.biz) to obtain an authorization key. Enter the authorization key below and then click the process authorization button."
- Link:** A blue hyperlink: "Click here to send the license request email to SAFE".
- Text Field:** "Your Site Code is:" followed by a text box containing "A7BE 4B53 DC5E 2FFA A6".
- Text Field:** "Please enter your Authorization Key:" followed by an empty text box.
- Form Fields:**
 - "License type:" with a dropdown menu showing "Unlimited".
 - "Licenses allocated:" with a text box showing "n/a".
 - "Licenses remaining:" with a text box showing "n/a".
 - "Days allocated:" with a text box showing "n/a".
 - "Days remaining:" with a text box showing "n/a".
 - "Languages enabled:" with a dropdown menu showing "ActionScript".
- Buttons:** "Process Authorization" and "Close".

License Type

The license can be one of three types.

- **File size based.** Used to examine a fixed amount of bytes of source code. Licenses are used up as source code is examined.
- **Time based.** Used to examine any amount of code for a fixed number of days.
- **Unlimited.** There is no limit on the number of megabytes that can be examined and there is no expiration date.

Licenses Allocated and Licenses Remaining

These fields indicate the number of licenses that were originally allocated and how many unused licenses remain. These fields are valid only for a megabyte based license. For other licenses, the fields are not applicable ("n/a").

Days Allocated and Days Remaining

These fields indicate the number of days that were originally allocated for the license and how many days remain on the license. These fields are valid only for a time based license. For other licenses, the fields are not applicable ("n/a").

Languages Enabled

This pull down list shows all of the programming languages that are enabled for analysis by the license.

See the SAFE Corporation website for license costs, as they may change.

The Menu and Toolbar

The CodeSuite menu and toolbar is shown below. Each menu selection is described in more detail below. If the menu selection can also be found on the toolbar, the corresponding toolbar icon is shown.



File Menu

The following selections are found on the File menu.



File->Open database...

This menu selection opens an existing CodeSuite database file.



File->Open filter...

This menu selection opens an existing CodeSuite filter file. For more information see the section entitled Using Filters.



File->Close all

This menu selection closes all open CodeSuite database files and filter files.



File->Export database->Filtered database

This menu selection creates a new database file by applying the open filter to the open database. For more information see the section entitled Using Filters.



File->Export database->HTML report

This menu selection converts the open CodeSuite database to a readable HTML basic report file with links to many readable HTML detailed report files. For more information, see the section entitled Creating HTML Reports.



File->Export database->Spreadsheets->CLOC Spreadsheet

This menu selection converts the open CodeCLOC database to a CLOC containing a statistical analysis of the distribution of the file scores. For more information, see the section entitled Creating CLOC Spreadsheets.



File->Export database->Spreadsheets->Distribution Spreadsheet

This menu selection converts the open CodeSuite database to a spreadsheet containing a statistical analysis of the distribution of the file scores. For more information, see the section entitled Creating Distribution Spreadsheets.



File->Export database->Spreadsheets->Search Spreadsheet

This menu selection converts the open CodeSuite database to spreadsheets containing information on the number of Internet hits for statements, comments, and identifiers. For more information, see the section entitled Creating Search Spreadsheets.



File->Export database->Spreadsheets->Summary Spreadsheet

This menu selection converts the open CodeSuite database to a spreadsheet containing a summary statistical analysis of the file scores. For more information, see the section entitled Creating Summary Spreadsheets.



File->Save filter...

Save the open filter file.



File->Save filter as...

Save the open filter file with a new name.



File->Recent databases...

Open one of the most recently opened database.



File->Recent filters...

Open one of the most recently opened filters.

File->Exit

Exit the program.

View Menu

The following selections are found on the File menu.

View->Toolbar

This menu selection toggles the toolbar on and off.

View->Status Bar

This menu selection toggles the status bar on and off.

Tools Menu



Tools->BitMatch

This menu selection brings up the BitMatch form. See the section entitled Running BitMatch for more information.



Tools->CodeCLOC

This menu selection brings up the CodeCLOC form. See the section entitled Running CodeCLOC for more information.



Tools->CodeCross

This menu selection brings up the CodeCross form. See the section entitled Running CodeCross for more information.



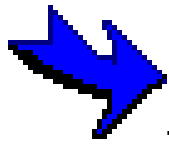
Tools->CodeDiff

This menu selection brings up the CodeDiff form. See the section entitled Running CodeDiff for more information.



Tools->CodeMatch

This menu selection brings up the CodeMatch form. See the section entitled Running CodeMatch for more information.



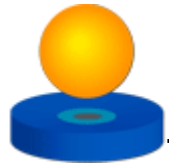
Tools->Restart...

This menu selection lets you select a CodeSuite database that was interrupted before completion. When the database is selected, CodeSuite will restart the comparison where it left off previously and will complete the comparison and generate a full database file.



Tools->FileCount

This menu selection brings up the FileCount form. See the section entitled Running FileCount for more information.



Tools->FileIsolate

This menu selection brings up the FileIsolate form. See the section entitled Running FileIsolate for more information.

Tools->Statistics

This menu selection brings up the database statistics form. See the section entitled Calculating Statistics for more information.

Tools->Filters

This menu selection brings up the filter form. See the section entitled Using Filters for more information.

Help Menu



Help->Contents

This menu selection brings up this users guide.



Help->Authorize

This menu selection brings up the authorization form for entering licenses to enable the various tools. See the section entitled Licenses for more information.

Help->About

This menu selection gives the version number and other information about CodeSuite.

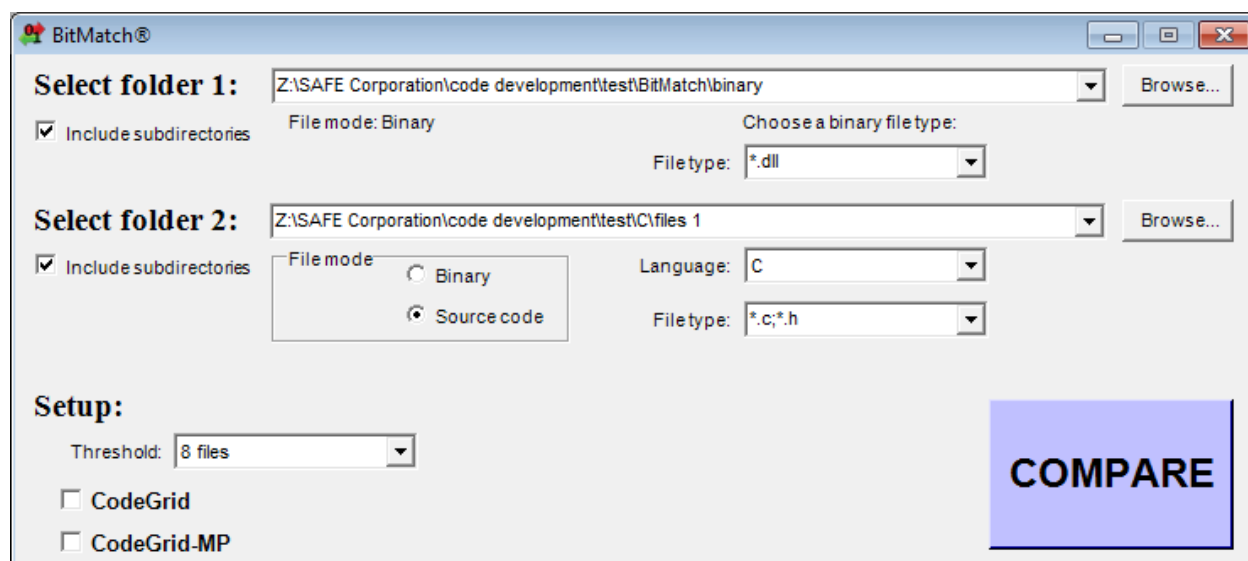
Status Bar

The status bar, located at the bottom of the CodeSuite window, is divided into four sections. From left to right, the first section shows the name of the CodeSuite database file that is open, the second section shows the name of the CodeSuite filter file that is open, the third section shows the current date, and the fourth section shows the current time.

BitMatch

Running BitMatch

BitMatch compares strings and identifier names from executable binary files to other executable binary files or source code files. Following that are step-by-step instructions for running BitMatch.



Step 1

Select the first folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 2

Specify the types of binary files in the first folder to compare. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 3

Select the second folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 4

Select whether the second folder contains binary files or source code files. If the second folder contains source code files, select the source code language.

Step 5

Specify the types of files in the second folder to compare. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 6

Select the reporting threshold from the pulldown menu. This determines how many files are reported. BitMatch reports only the most similar files. By setting the number of files to report to a large number you may get a very large database. By setting the number of files to report to a small number, the database will be smaller but it may not include all the similar files that you would like to see.

Step 7

Check the CodeGrid box to run the comparison on a grid of computers if you have a license to do so. Check the CodeSuite-MP box to run the comparison on a multicore computer if you have a license to do so. When the CodeSuite-MP box is checked, a textbox will appear allowing you to enter the number of processes to run simultaneously.

Step 8

Click on the compare button. The number of licenses, if any, that are required for this run of BitMatch will be shown. You will have the ability to cancel the BitMatch run at this point without using up licenses.

You will be then asked to name the database that will be generated. To generate readable HTML reports from the database see the section entitled Creating HTML Reports.

BitMatch Algorithms

The Algorithms

BitMatch searches binary files for all uninterrupted sequences of text characters. It then uses two CodeMatch algorithms to determine similarity between two source code files, first treating the text strings as program strings and then treating the text strings as identifiers. These algorithms are described below. When multiple files are compared, each match is given a weight and all weights are combined into a single matching score called the correlation score. The file pairs are then ranked by BitMatch score so that you can examine the most similar files.

Comment/string matching

BitMatch looks for identical comments and strings, ignoring whitespace. Comment lines and strings that contain only programming language keywords are still considered matches.

Identifier matching


BitMatch finds every instance in each file where identifiers match exactly. It eliminates programming language keywords and only reports matches for non-keyword identifiers such as variable names and function names.


BitMatch also finds every instance where an identifier in one file is part of a larger identifier in the other file. For example, the variable name "Index" in one file would partially match the variable names "NewIndex" and "Index1" in the other file. BitMatch eliminates programming language keywords and only reports matches for non-keyword identifiers such as variable names and function names.

Correlation Score

BitMatch produces a total correlation score based on the combination of above algorithms that the user chooses when running BitMatch. The minimum score is 0 while the maximum score is 100.

BitMatch Basic Report





BitMatch Basic Report
Version: 1.0.1 | Date: 03/16/08 | Time: 15:35:00

SETTINGS | RESULTS | UNCOMPARED FILES | TOTALS

SETTINGS

Compare files in folder	C:\test\BitMatch\binary <i>Including subdirectories</i>
File types	*.exe
To files in folder	C:\test\BitMatch\C <i>Including subdirectories</i>
File types	*.c;*.h
Programming language	C
Reporting file threshold	8 files

RESULTS

C:\test\BitMatch\binary\CodeSuite.exe Score	Compared to file
27	C:\test\BitMatch\C\LineCount.c
23	C:\test\BitMatch\C\CodeSuite.h
21	C:\test\BitMatch\C\CodeSuite.c

C:\test\BitMatch\binary\test1\test.exe Score	Compared to file
22	C:\test\BitMatch\C\CodeSuite.c
21	C:\test\BitMatch\C\CodeSuite.h


C:\test\BitMatch\binary\test1\test0.exe Score	Compared to file
23	C:\test\BitMatch\C\CodeSuite.c
22	C:\test\BitMatch\C\LineCount.c
19	C:\test\BitMatch\C\CodeSuite.h


TOTALS

Total number of bytes in files in folder 1 = 799957
Total number of bytes in files in folder 2 = 292584
Total run time = 50 Seconds



BitMatch Detailed Report





BitMatch Detailed Report
Version: 1.0.1 | Date: 03/16/08 | Time: 15:35:00

SETTINGS

Compare file 1:	C:\test\BitMatch\binary\LineCount\Debug\LineCount.exe
To file 2:	C:\test\BitMatch\C\LineCount.c
Links to results:	Matching Comments and Strings Matching Identifiers Score

RESULTS

Matching Comments and Strings		
File1 Line#	File2 Line#	Comment/String
5890		
9304		
9381		
9755		
9765		
10082		
10093	140	rB
10099		
10199		
10805		
10815		
14165		

14154	68	Total number of non-blank lines: %i
14157	65	Total number of Kbytes: %i
14158	64	Total number of files: %i



FILE	filepattern	folder	LineCount	name	size
stdio	string				



!CompareString	#File	(Press	.idata	arguments	CIHandle
DLineCount	Domain	Ession	files	FindFirstFile	IG_LINE
LCMapStringA	LoadLibrary	MultiByte	osinfo.c	SandleCount	
_A_SUBDIR	_finddata	_findfirst	_findnext	_LINE_LEN	
finfo	handle	InString	IsBlank	KByteCount	
SepString	stdlib				



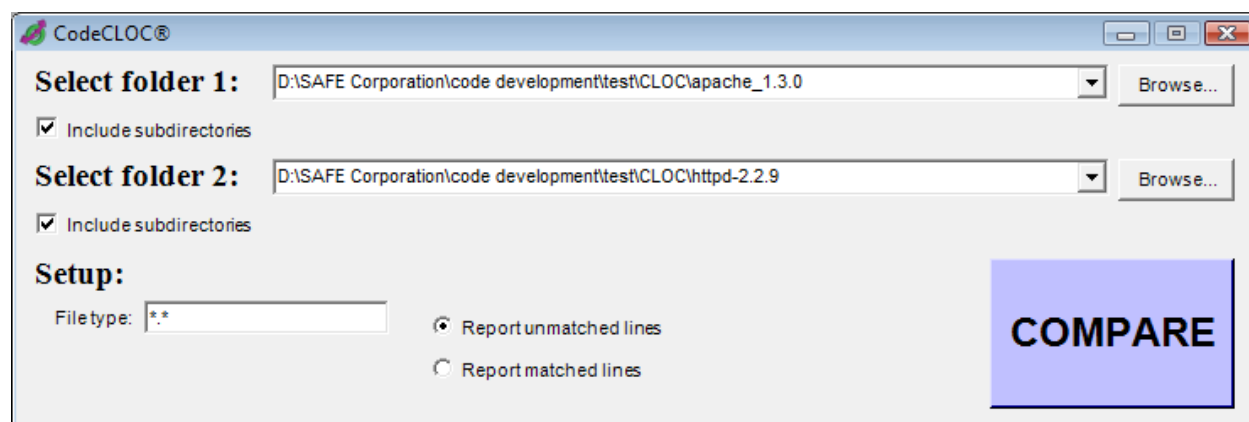
SCORE 71

CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

CodeCLOC

Running CodeCLOC

CodeCLOC compares two versions of software source code files to determine code growth. A CodeCLOC comparison is essentially a CodeDiff comparison of files with the same name such that each file is used at most once and the scores are optimized such that the overall percentage of similarity between the two sets of files is maximized. . Following that are step-by-step instructions for running CodeCLOC.



Step 1

Select the folders containing the versions of software to be compared by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories. Folder 1 contains the initial *base* version of source code files and folder 2 contains the subsequent *examined* version of source code files to compare to it to measure growth.

Step 2

Enter the file types separated by semicolons (e.g., *.cpp;*.c;*.h).

Step 3

Select one of two options for reporting lines in the database.

- **Report unmatched lines.** Report lines in either file that do not have a match in the other file. Note that if a line can be found three times in the first file and five times in the second file, two of the lines in the second file will be reported as unmatched.
- **Report matched lines.** Report lines that can be found in both files.

Step 4

Click on the compare button to start the analysis. A window will pop-up, see below, asking you to name the output file for your analysis. The number of licenses, if any, that are required for this run of CodeCLOC will be shown. You will have the ability to cancel the CodeCLOC run at this point without using up licenses.

You will be then asked to name the database that will be generated. To generate readable HTML reports from the database see the section entitled Creating HTML Reports. To generate CLOC statistics from the database see the section entitled Creating CLOC Spreadsheets.

CodeCLOC Algorithm

CodeCLOC implements the Changing Lines of Code (CLOC) method for measuring source code growth from one version of a program to another. The CLOC method counts the number of lines of code that have been added, changed, or remain unchanged. These values are then combined to express the change in software as a rate of growth. The results can also be expressed in terms of the decay of the original code, which can be useful as a measure of how much original intellectual property still exists from that original code.

Measurements

The CLOC method uses the CodeDiff and FileCount functionality to create a CodeSuite database from which a specially developed CLOC spreadsheet can be generated.

The FileCount counts files, lines of code (LOC), and number of bytes in a directory tree. CLOC requires the number of program specific files and the number of non-blank lines in the software project's directory tree. CodeDiff exhaustively compares lines of code in one set of source code files to that in another set of source code files. CLOC requires that CodeDiff is used to compare same-name files from the original version to subsequent versions of the software project. CodeCLOC also prepares a post-CodeDiff optimization such that each file is used at most once in the comparison and the file pairs are selected to maximize the total similarity scores between the selected files. The selection and optimization only occurs for multiple files in the versions that have the same name.


Typically movements of source code between files represents work being performed. Similarly a file name change represents work being performed, because file names are not generally changed from version to version unless there is a significant change to the functionality of the file. The results of the CodeDiff analysis are then exported into a CodeSuite distribution report that contains the statistical information about changes in the files and LOC.


The generated CLOC spreadsheet shows statistics about the rate of software growth. The software evolution results can be summarized using these different percentages:

- **Percent of new and modified files.** Percentage of files that were added to the examined version or modified from the base to the examined version as a percentage of the total files in the examined version.
- **Percent of new and modified files relative to base version.** Percentage of files that were added to the examined version or modified from the base to the examined version as a percentage of the total files in the base version.

- **Percent of continuing files.** Percentage of files that continued, modified or not, from the base version to the examined version as a percentage of the total files in the examined version.
- **Percent of unchanged continuing files.** Percentage of files that continued unchanged from the base version to the examined version as a percentage of the total files in the examined version.
- **Percent CLOC change.** Percentage of changed lines of code from the base version to the examined version as a percentage of the total files in the examined version.
- **Percent CLOC change relative to base version.** Percentage of changed lines of code from the base version to the examined version as a percentage of the total files in the base version.
- **Percent LOC growth.** Percentage of total lines of code in the examined version as a percentage of the total lines of code in the base version.
- **Percent unchanged continuing lines of code.** Percentage of unchanged lines of code as a percentage of the total lines of code in the examined version.

CodeCLOC Basic Report



	CodeCLOC Basic Report Version: 4.2.0 Date: 03/14/09 Time: 17:48:01
---	--

[SETTINGS](#) | [RESULTS](#) | [UNCOMPARED FILES](#) | [TOTALS](#)

SETTINGS

Compare files in folder	C:\test\C\files 1 <i>Including subdirectories</i>
File types	*.*
To files in folder	C:\test\C\files 2 <i>Including subdirectories</i>
File types	*.*
Algorithms selected	<ul style="list-style-type: none"> Ignoring case Ignoring whitespace Percentage of file pairs Report matched lines
Reporting file threshold	1 file
Reporting score threshold	0

RESULTS

C:\test\C\files 1\aaa.c Score	Compared to file
100	C:\test\C\files 2\aaa.c
80	C:\test\C\files 2\abc.c
4	C:\test\C\files 2\bpf_dump_semicolons.c
4	C:\test\C\files 2\Copy of semicolon_test.c
4	C:\test\C\files 2\semicolon_test.c
3	C:\test\C\files 2\bpf_dump_strings.c
3	C:\test\C\files 2\Copy of bpf_dump_strings.c

C:\test\C\files 1\aaa_case.c Score	Compared to file
100	C:\test\C\files 2\aaa.c
80	C:\test\C\files 2\abc.c
4	C:\test\C\files 2\bpf_dump_semicolons.c
4	C:\test\C\files 2\Copy of semicolon_test.c
4	C:\test\C\files 2\semicolon_test.c
3	C:\test\C\files 2\bpf_dump_strings.c
3	C:\test\C\files 2\Copy of bpf_dump_strings.c

C:\test\C\files 1\aaa_with_comments.c Score	Compared to file
100	C:\test\C\files 2\aaa_with_comments.c

C:\test\C\files 1\all_ints.c Score	Compared to file
100	C:\test\C\files 2\all_ints.c

C:\test\C\files 1\all_specifiers.c Score	Compared to file
100	C:\test\C\files 2\all_specifiers.c

TOTALS

Total number of bytes in files in folder 1 = 37651


Total number of bytes in files in folder 2 = 48944


Total run time = 3 Seconds



CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

CodeCLOC Detailed Report





CodeCLOC Detailed Report
Version: 1.0.0 | Date: 03/14/09 | Time: 17:48:01


SETTINGS

Compare file 1:	C:\test\C\files 1\aaa_with_comments.c
To file 2:	C:\test\C\files 2\aaa_with_comments.c
Links to results:	Matched lines Score

RESULTS

Matching Lines

File1 Line#	File2 Line#	Line
1	1	/* This is a comment*/ p = null;
2	2	private String auxonus = null; // This is a comment
5	5	This is a comment*/ p = /* This is a comment*/ null; // This is a comment

 [TOP](#)

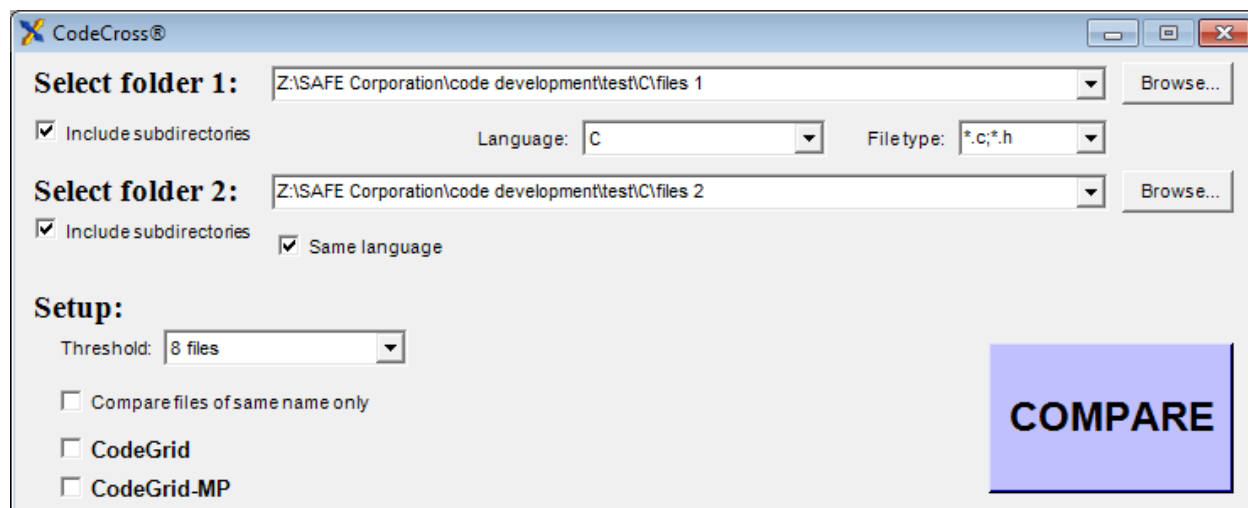
SCORE 100

CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

CodeCross

Running CodeCross

CodeCross cross-compares statements in one set of files to comments in the other set of files, and vice versa, in order to find code that has been commented out. CodeCross finds areas of source code that were used as guides to develop other source code; it finds signs of copying that CodeMatch can miss. Below is a screen shot of the CodeCross form. Following that are step-by-step instructions for running CodeCross.



The screenshot shows the CodeCross application window. It has a title bar with the CodeCross logo and standard window controls. The main area contains the following elements:

- Select folder 1:** A text field containing the path "Z:\SAFE Corporation\code development\test\C\files 1" and a "Browse..." button.
- Include subdirectories**
- Language:** A dropdown menu set to "C".
- File type:** A dropdown menu set to "*.c;*.h".
- Select folder 2:** A text field containing the path "Z:\SAFE Corporation\code development\test\C\files 2" and a "Browse..." button.
- Include subdirectories**
- Same language**
- Setup:**
 - Threshold:** A dropdown menu set to "8 files".
 - Compare files of same name only**
 - CodeGrid**
 - CodeGrid-MP**
- A large blue button labeled **COMPARE**.

Step 1

Select the first folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 2

Select a source code language from the pulldown menu.

Step 3

Select the files types to compare from the pulldown menu. You may edit the file types to include specific file types. Separate multiple file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 4

Select the second folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 5

Choose whether both sets of files are written in the same programming language. If not, uncheck the box and you will be able to select the language and file types for the second set of files. CodeCross can compare files in different programming languages.

Step 6

Choose setup options to be used for comparing files.

Select the reporting threshold from the pulldown menu. This determines how many files are reported. CodeMatch reports only the most highly correlated files. By setting the number of files to report to a large number you may get a very large database. By setting the number of files to report to a small number, the database will be smaller but it may not include all the similar files that you would like to see.

Step 7

Choose whether to only compare files if they have the same name. This will speed up the comparison significantly because far fewer combinations of files are compared.

Step 8

Check the CodeGrid box to run the comparison on a grid of computers if you have a license to do so. Check the CodeSuite-MP box to run the comparison on a multicore computer if you have a license to do so. When the CodeSuite-MP box is checked, a textbox will appear allowing you to enter the number of processes to run simultaneously.

Step 9

Click on the compare button. The number of licenses, if any, that are required for this run of CodeMatch will be shown. You will have the ability to cancel the CodeMatch run at this point without using up licenses.

You will be then asked to name the database that will be generated. To generate readable HTML reports from the database see the section entitled Creating HTML Reports.


CodeCross Algorithm


CodeCross compares statements in one file to comments and strings in another file and calculates the number of complete matches as a percentage of the total number of statements, comments, and strings.

CodeCross Score

CodeCross produces a score that is a combined percentage of statements that match comments and strings and a percentage of comments and strings that match statements. The minimum score is 0 while the maximum score is 100.

CodeCross Basic Report





CodeCross Basic Report
 Version: 1.1.0 | Date: 12/29/08 | Time:
 19:47:18

[SETTINGS](#) | [RESULTS](#) | [UNCOMPARED FILES](#) | [TOTALS](#)

SETTINGS

Compare files in folder	C:\test\CodeCross\files 1 <i>Including subdirectories</i>
File types	*.c;*.h
Programming language	C
To files in folder	C:\test\CodeCross\files 2 <i>Including subdirectories</i>
File types	*.c;*.h
Programming language	C
Reporting file threshold	8 files

RESULTS

C:\code development\test\CodeCross\files 1\bpf_dump_strings.c Score	Compared to file
--	------------------

71	C:\test\CodeCross\files 2\bpf_dump_identifiers.c
71	C:\test\CodeCross\files 2\bpf_dump_mod.c
12	C:\test\CodeCross\files 2\aaa_commented.c
2	C:\test\CodeCross\files 2\W32NReg_commented.c

C:\code development\test\CodeCross\files 1\bpf_image.c Score	Compared to file
68	C:\test\CodeCross\files 2\bpf_dump_strings.c
2	C:\test\CodeCross\files 2\W32NReg_commented.c

C:\CodeCross\files 1\bpf_image_commented.c Score	Compared to file
38	C:\test\CodeCross\files 2\bpf_dump_strings.c
24	C:\test\CodeCross\files 2\W32NReg_commented.c

TOTALS


Total number of bytes in files in folder 1 = 33829


Total number of bytes in files in folder 2 = 25147

Total run time = 2 Seconds



CodeCross Detailed Report





CodeCross Detailed Report
Version: 1.1.0 | Date: 12/29/08 | Time:
19:47:18

SETTINGS

Compare file 1:	C:\test\CodeCross\files 1\aaa_case.c
To file 2:	C:\test\CodeCross\files 2\aaa_commented.c
Links to Results:	Matching Statements to Comments Matching Comments to Statements Score

RESULTS

Matching Statements to Comments		
File1 Line#	File2 Line#	Statement
1	1 4	P = Null;
2	2 5	Private String Auxonus = Null;

Matching Comments to Statements

File1 Line#	File2 Line#	Comment/String
3	6	* The Regents of the University of California. All rights reserved.
5	8	* Redistribution and use in source and binary forms, with or without
6	9	* modification, are permitted provided that: (1) source code distributions
7	10	* retain the above copyright notice and this paragraph in its entirety, (2)
8	11	* distributions including binary code include the above copyright notice and
9	12	* this paragraph in its entirety in the documentation or other materials
10	13	* provided with the distribution, and (3) all advertising materials mentioning



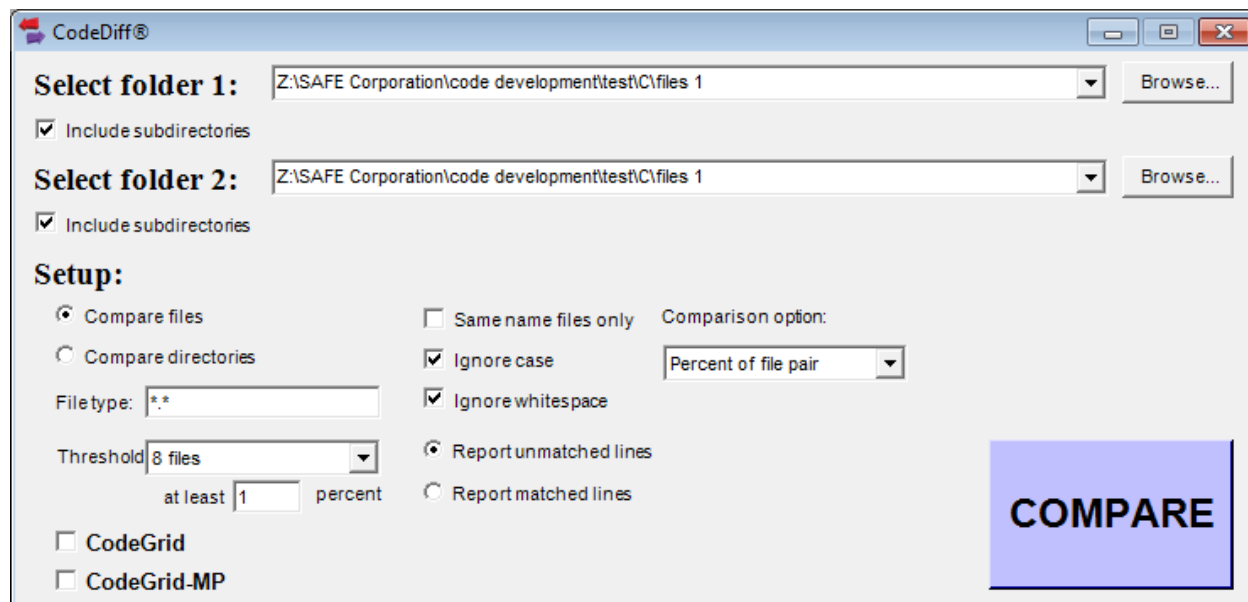
SCORE 100

CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

CodeDiff

Running CodeDiff

CodeDiff compares files on a line by line basis to determine the percentage similarity. Below is a screen shot of the CodeDiff form. Following that are step-by-step instructions for running CodeDiff.



The screenshot shows the CodeDiff application window. It has two 'Select folder' sections, each with a text field containing 'Z:\SAFE Corporation\code development\test\C\files 1' and a 'Browse...' button. Below each folder selection is a checked checkbox for 'Include subdirectories'. The 'Setup:' section contains several options: 'Compare files' (selected), 'Compare directories' (unselected), 'Filetype: *.*', 'Threshold: 8 files' (with a dropdown arrow), 'at least 1 percent', 'Same name files only' (unselected), 'Ignore case' (checked), 'Ignore whitespace' (checked), 'Comparison option: Percent of file pair' (dropdown), 'Report unmatched lines' (selected), and 'Report matched lines' (unselected). At the bottom left are checkboxes for 'CodeGrid' and 'CodeGrid-MP'. A large blue 'COMPARE' button is on the right.

Step 1

Select the first folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 2

Select the second folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 3

Choose setup options to be used for comparing files.

You have a choice of two options.

- **Compare files.** CodeDiff will compare each pair of files in the directories specified and subdirectories if that option is selected. This option is used for comparing all combinations of files in all directories to find those that are similar.

- **Compare directories.** CodeDiff will compare only files with identical names in directories that have the same name and are in the same place in the directory tree. CodeDiff will also list all files in one directory that have no corresponding files in the other directory. This option is used for comparing entire directory trees to find similar files and missing files.

You have a choice of any combination of the following three options.

- **Same name files only.** CodeDiff will only compare files if they have the same exact name (not case sensitive). This option is not available when comparing directories, because only files of the same name are compared when comparing directories.
- **Ignore case.** CodeDiff will consider two lines matching if they are identical in all other respects even if the letter cases are different.
- **Ignore whitespace.** Before performing a comparison on any lines, CodeDiff will reduce all sequences of whitespace characters (space or tab) to a single space.

You have a choice of two options for reporting lines in the database.

- **Report unmatched lines.** Report lines in either file that do not have a match in the other file. Note that if a line can be found three times in the first file and five times in the second file, two of the lines in the second file will be reported as unmatched.
- **Report matched lines.** Report lines that can be found in both files.

You have a choice of three comparison options.

- **Percentage of file pair.** The percentage generated by CodeDiff will be the percentage of lines in the two files that match with respect to the total number of lines in both files.
- **Percentage of first file.** The percentage generated by CodeDiff will be the percentage of lines in the first file that match a line in the second file with respect to the total number of lines in the first file.
- **Percentage of second file.** The percentage generated by CodeDiff will be the percentage of lines in the second file that match a line in the first file with respect to the total number of lines in the second file.

Step 4

Specify the types of files to compare. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 5

Select the reporting threshold from the pulldown menu. This determines how many files are reported. CodeDiff reports only the most similar files. By setting the number of files to report to a large number you may get a very large database. By setting the number of files to report to a small number, the database will be smaller but it may not include all the similar files that you would like to see.

Specify the minimum percentage matching to report. CodeDiff reports the percentage of lines that are identical between files, ranging from 0 to 100. If you specify 0 as the minimum threshold, even files that do not have any matching lines will be reported. If you specify 100 as the minimum threshold, only files whose lines all match exactly will be reported.

Step 6

Check the CodeGrid box to run the comparison on a grid of computers if you have a license to do so. Check the CodeSuite-MP box to run the comparison on a multicore computer if you have a license to do so. When the CodeSuite-MP box is checked, a textbox will appear allowing you to enter the number of processes to run simultaneously.

Step 7

Click on the compare button. The number of licenses, if any, that are required for this run of CodeDiff will be shown. You will have the ability to cancel the CodeDiff run at this point without using up licenses.

You will be then asked to name the database that will be generated. To generate readable HTML reports from the database see the section entitled Creating HTML Reports.

CodeDiff Algorithm

CodeDiff compares each line of code in two sets of files and calculates the number of lines of code that match completely as a percentage of the total number of lines of code. The order of the lines is not considered so if a file were compared to an identical copy where the statements were all in a different order, this would still result in a 100% match.

If CodeDiff is set to ignore case, lines are considered matches even if the letters have different cases.

If CodeDiff is set to ignore whitespace, all sequences of whitespace (spaces and tabs) are converted to a single space before the comparison is performed.

If CodeDiff is set to generate the percentage of file pairs, it will generate the percentage of lines in the two files that match with respect to the total number of lines in both files. If CodeDiff is set to generate the percentage of the first file, it will generate the percentage of lines in the first file that match a line in the second file with respect to the total number of lines in the first file.

Similarity Score

CodeDiff produces a similarity score that is a percentage of matching line within the files. The minimum score is 0 while the maximum score is 100.

CodeDiff Basic Report



CodeDiff Basic Report

Version: 4.2.0 | Date: 03/14/09 | Time: 17:48:01


SETTINGS | RESULTS | UNCOMPARED FILES | TOTALS

SETTINGS


Compare files in folder	C:\test\C\files 1 <i>Including subdirectories</i>
File types	*.*
To files in folder	C:\test\C\files 2 <i>Including subdirectories</i>
File types	*.*
Algorithms selected	<ul style="list-style-type: none">• Ignoring case• Ignoring whitespace• Percentage of file pairs• Report matched lines
Reporting file threshold	8 files
Reporting score threshold	1


RESULTS

C:\test\C\files 1\aaa.c Score	Compared to file
----------------------------------	------------------

100	C:\test\C\files 2\aaa.c
80	C:\test\C\files 2\abc.c
3	C:\test\C\files 2\Copy of bpf_dump_strings.c
C:\test\C\files 1\aaa_case.c Score	
100	Compared to file C:\test\C\files 2\aaa.c
80	C:\test\C\files 2\abc.c
3	C:\test\C\files 2\bpf_dump_strings.c
C:\test\C\files 1\all_specifiers.c Score	
100	Compared to file C:\test\C\files 2\all_specifiers.c
TOTALS	
Total number of bytes in files in folder 1 = 37651	
Total number of bytes in files in folder 2 = 48944	
Total run time = 3 Seconds	
 TOP	
CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation	

CodeDiff Detailed Report





CodeDiff Detailed Report
Version: 4.2.0 | Date: 03/14/09 | Time: 17:48:01


SETTINGS

Compare file 1:	C:\test\C\files 1\aaa_with_comments.c
To file 2:	C:\test\C\files 2\aaa_with_comments.c
Links to results:	Matched lines Score

RESULTS

Matching Lines

File1 Line#	File2 Line#	Line
1	1	/* This is a comment*/ p = null;
2	2	private String auxonus = null; // This is a comment
3	3	p = /* This is a comment*/ null; /* This is a comment*/
4	4	/* Yes,
5	5	This is a comment*/ p = /* This is a comment*/ null; // This is a comment

 [TOP](#)

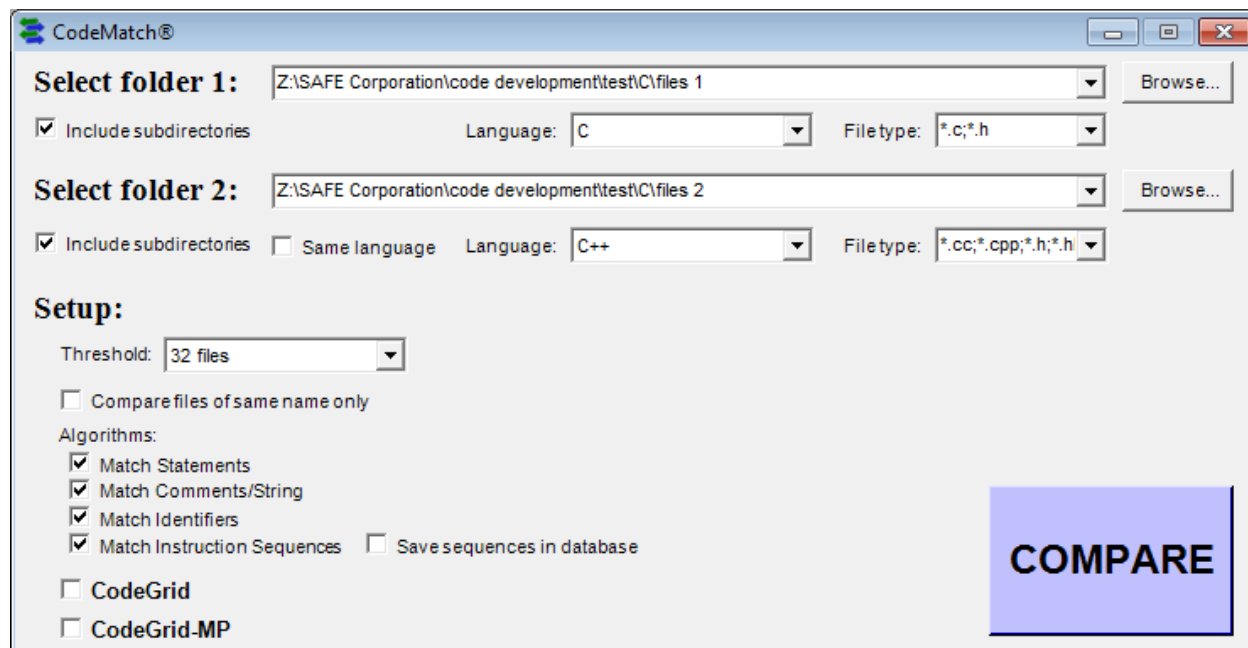
SCORE 100

CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

CodeMatch

Running CodeMatch

CodeMatch compares files using a set of algorithms to determine their correlation. Below is a screen shot of the CodeMatch form. Following that are step-by-step instructions for running CodeMatch.



The screenshot shows the CodeMatch application window with the following settings:

- Select folder 1:** Z:\SAFE Corporation\code development\test\C\files 1
- Include subdirectories
- Language: C
- Filetype: *.c;*.h
- Select folder 2:** Z:\SAFE Corporation\code development\test\C\files 2
- Include subdirectories
- Same language
- Language: C++
- Filetype: *.cc;*.cpp;*.h;*.h
- Setup:**
 - Threshold: 32 files
 - Compare files of same name only
 - Algorithms:
 - Match Statements
 - Match Comments/String
 - Match Identifiers
 - Match Instruction Sequences
 - Save sequences in database
 - CodeGrid
 - CodeGrid-MP

A large blue button labeled **COMPARE** is located at the bottom right of the window.

Step 1

Select the first folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 2

Select a source code language from the pulldown menu.

Step 3

Select the files types to compare from the pulldown menu. You may edit the file types to include specific file types. Separate multiple file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 4

Select the second folder for comparison by clicking on the browse button or entering the path in the text field. Check the box to include files in all subdirectories.

Step 5

Choose whether both sets of files are written in the same programming language. If not, uncheck the box and you will be able to select the language and file types for the second set of files. CodeMatch can compare files in different programming languages.

Step 6

Choose setup options to be used for comparing files.

Select the reporting threshold from the pulldown menu. This determines how many files are reported. CodeMatch reports only the most highly correlated files. By setting the number of files to report to a large number you may get a very large database. By setting the number of files to report to a small number, the database will be smaller but it may not include all the similar files that you would like to see.

Step 7

Choose whether to only compare files if they have the same name. This will speed up the comparison significantly because far fewer combinations of files are compared.

Step 8

Choose setup options and algorithms to be used for comparing files. You have a choice of any combination of four algorithms. For more information on each of these algorithms, see the section entitled CodeMatch Algorithms.

- **Statement Matching**
- **Comment Matching**
- **Identifier Matching**
- **Instruction Sequence Matching**

Note that when you select instruction sequence matching, you will have the option to record the instruction sequences in the resulting database. Putting the sequences in the database makes it easier to determine whether the sequences are interesting, but will produce a significantly larger database.

Step 9

Check the CodeGrid box to run the comparison on a grid of computers if you have a license to do so. Check the CodeSuite-MP box to run the comparison on a multicore computer if you have a license to do so. When the CodeSuite-MP box is checked, a

textbox will appear allowing you to enter the number of processes to run simultaneously.

Step 10

Click on the compare button. The number of licenses, if any, that are required for this run of CodeMatch will be shown. You will have the ability to cancel the CodeMatch run at this point without using up licenses.

You will be then asked to name the database that will be generated. To generate readable HTML reports from the database see the section entitled Creating HTML Reports.

CodeMatch Algorithms

The Algorithms

CodeMatch uses several algorithms to determine similarity between two source code files. These algorithms are described below. When multiple files are compared, each match is given a weight and all weights are combined into a single matching score called the CodeMatch score. The file pairs are then ranked by CodeMatch score so that you can examine the most similar files.

Statement matching

CodeMatch looks for identical program statements (i.e., functional source code), ignoring whitespace and eliminating comments and strings. Statements that contain only programming language keywords are not considered matching. For statements to be considered matches, they must contain at least one identifier (non-keyword) such as a variable name or function name.

Comment/string matching

CodeMatch looks for identical comments and strings, ignoring whitespace. Comment lines and strings that contain only programming language keywords are still considered matches.

Instruction sequence matching

CodeMatch looks for sequences of instructions that match. CodeMatch notes the longest such sequence in each pair of files. A sequence matches if the initial programming language statement on each line is identical, regardless of what follows it. Even if variable names are altered in one file, CodeMatch will report similarities in the files. The following shows an example of two identical instruction sequences in C:

```
// File 1
if (x == 5)
{
    // Loop on j here
    for (j = 0; j < Index; j++)
        printf("x = %i", j);
}
else
    break; // Here's the break

// File 2
if (xyz < 2)
    for (jjj = 0; jjj < i; jjj++)
```

```
    {
        printf("Hello world\n");
    }
else
    break;
```

Identifier matching


CodeMatch finds every instance in each file where identifiers match exactly. It eliminates programming language keywords and only reports matches for non-keyword identifiers such as variable names and function names.


CodeMatch also finds every instance where an identifier in one file is part of a larger identifier in the other file. For example, the variable name "Index" in one file would partially match the variable names "NewIndex" and "Index1" in the other file. CodeMatch eliminates programming language keywords and only reports matches for non-keyword identifiers such as variable names and function names.

Correlation Score

CodeMatch produces a total correlation score based on the combination of above algorithms that the user chooses when running CodeMatch. The minimum score is 0 while the maximum score is 100.

CodeMatch Basic Report





CodeMatch Basic Report
 Version: 5.3.1 | Date: 08/28/08 | Time: 11:33:11

SETTINGS | RESULTS | UNCOMPARED FILES | TOTALS

SETTINGS

Compare files in folder	C:\test\C\files 1 <i>Including subdirectories</i>
File types	*.c;*.h
Programming language	C
To files in folder	C:\test\C\files 2 <i>Including subdirectories</i>
File types	*.c;*.h
Programming language	C
Algorithms selected	<ul style="list-style-type: none"> Statement Matching Comment Matching Identifier Matching Instruction Sequence Matching
Reporting file threshold	8 files

RESULTS

C:\test\C\files 1\aaa.c Score	Compared to file
82	C:\test\C\files 2\aaa.c
82	C:\test\C\files 2\abc.c
71	C:\test\C\files 2\aaa_with_comments.c
15	C:\test\C\files 2\.svn\bpf_image.c
9	C:\test\C\files 2\all_specifiers.c
9	C:\test\C\files 2\bpf_dump_semicolons.c
9	C:\test\C\files 2\bpf_dump_strings.c
9	C:\test\C\files 2\semicolon_test.c

C:\test\C\files 1\aaa_case.c Score	Compared to file
82	C:\test\C\files 2\aaa.c
82	C:\test\C\files 2\abc.c
71	C:\test\C\files 2\aaa_with_comments.c
15	C:\test\C\files 2\.svn\W32NReg.c
13	C:\test\C\files 2\.svn\bpf_image.c

C:\test\C\files 1\aaa_whitespace.c Score	Compared to file
82	C:\test\C\files 2\aaa.c
82	C:\test\C\files 2\abc.c
15	C:\test\C\files 2\.svn\bpf_image.c
9	C:\test\C\files 2\all_specifiers.c

9	C:\test\C\files 2\bpf_dump_strings.c
9	C:\test\C\files 2\semicolon_test.c

C:\test\C\files 1\aaa_with_comments.c Score	Compared to file
87	C:\test\C\files 2\aaa_with_comments.c
71	C:\test\C\files 2\aaa.c
69	C:\test\C\files 2\abc.c
15	C:\test\C\files 2\.svn\bpf_image.c
8	C:\test\C\files 2\all_specifiers.c
8	C:\test\C\files 2\W32NReg.c

C:\test\C\files 1\all_ints.c Score	Compared to file
82	C:\test\C\files 2\all_ints.c
17	C:\test\C\files 2\all_specifiers.c
11	C:\test\C\files 2\.svn\W32NReg (no comments).c
11	C:\test\C\files 2\.svn\W32NReg.c

TOTALS

Total number of bytes in files in folder 1 = 37651
Total number of bytes in files in folder 2 = 37627
Total run time = 16 Seconds



CodeMatch Detailed Report

S.A.F.E.



CodeMatch Detailed Report

Version: 5.3.1 | Date: 08/28/08 | Time: 11:33:11

SETTINGS

Compare file 1:	C:\test\C\files 1\bpf_image.c
To file 2:	C:\test\C\files 2\.svn\bpf_image.c
Links to results:	Matching Statements Matching Comments and Strings Matching Instruction Sequences Matching Identifiers Partially Matching Identifiers Score

RESULTS

Matching Statements

File1 Line#	File2 Line#	Statement
22	22	#include <windows.h>
23	23	#include <sys/types.h>
35	35	char *fmt, *op
36	36	static char image[256]

37	37	char operand[64]
39	39	v = p->k
40	40	switch (p->code) {
199 204 209 214	199	case BPF_ALU BPF_OR BPF_X:
254	254 259 264 269 270	case BPF_ALU BPF_NEG:



Matching Comments and Strings

File1 Line#	File2 Line#	Comment/String
2	2	* Copyright (c) 1990, 1991, 1992, 1994, 1995, 1996
3	3	* The Regents of the University of California. All rights reserved.
5	5	* Redistribution and use in source and binary forms, with or without
6	6	* modification, are permitted provided that: (1) source code distributions
7	7	* retain the above copyright notice and this paragraph in its entirety, (2)
8	8	* distributions including binary code include the above copyright notice and
9	9	* this paragraph in its entirety in the documentation or other materials
10	10	* provided with the distribution, and (3) all advertising materials mentioning
11	11	* features or use of this software display the following acknowledgement:



Matching Instruction Sequences

File1 Line#	File2 Line#	Number of matching instructions
22	22	202
43	129	71
46	51	64
46	56	60
46	61	56
46	66	52
46	71	48
46	76	44
46	81	40
46	86	36
46	91	32



256	64	BPF_A	BPF_ABS	BPF_ADD	BPF_B
BPF_CLASS	BPF_DIV	BPF_H	bpf_image	BPF_IMM	BPF_JA
BPF_JEQ	BPF_JGE	BPF_JGT	BPF_JMP	BPF_JSET	BPF_LDX
BPF_LEN	BPF_LSH	BPF_MEM	BPF_MISC	BPF_MSH	BPF_OP
BPF_OR	BPF_RET	BPF_RSH	BPF_ST	BPF_STX	BPF_TXA
BPF_W	BPF_X	code	fmt	image	jt
op	operand	stdio	string	sys	



Partially Matching Identifiers

File1 Identifiers

0x00FF	BPF_ALU	bpf_filter	BPF_IMM	BPF_IN	BPF_LEN	BPF_MEMWORDS	BPF_RETURN
BPF_STMT	BPF_SUB	EXTRACT_LONG	INT	netlong	types	UCHAR	W32N_htonl
winsock							

File2 Identifiers

0x0004	0x0005	_stdcall	_TEXT	_W32N_ADA	_WAdapter0	_WAdapter1	_WAdapter2
dwDataLen	DWORD	dwType	ERR_IMPLIED	ERR_SUCCESS	H_LOCAL	hAdapter	hClassNet
KEY_READ	LONG	pAdapterInfo	PCHAR	PW_ADAPTER	QueryValue	TChar	VER_WIN32
W0Adapter	W0Window	W0Windows	W32N_Adapt	W32N_NET	WINCARDS	wsprintf	



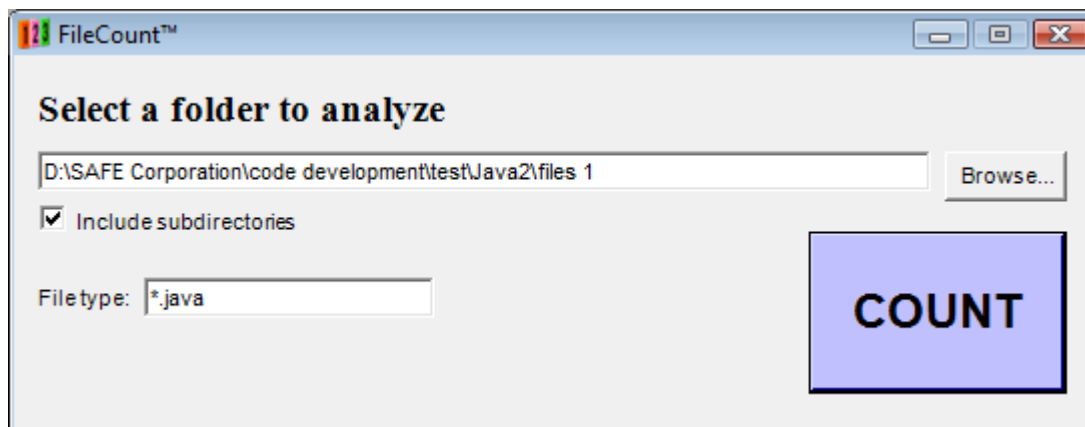
SCORE 100

CodeSuite copyright 2003-2010 by Software Analysis and Forensic Engineering Corporation

FileCount

Running FileCount

FileCount is a utility that counts the number of files, non-blank lines, and bytes in a large set of files in a directory tree. FileCount is useful when using CodeDiff to generate statistics about a set of source code files.



Step 1

Select the folder where the files are that need to be counted by clicking on the browse button or entering the path in the text field. Check the box to include all subdirectories.

Step 2

Type in the file types. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 3

Press the count button. FileCount will then search the directory and all subdirectories, if specified, counting all of the files that meet the file type, and counting the total number of non-blank lines and bytes. When complete, a dialog box will appear with these counts.

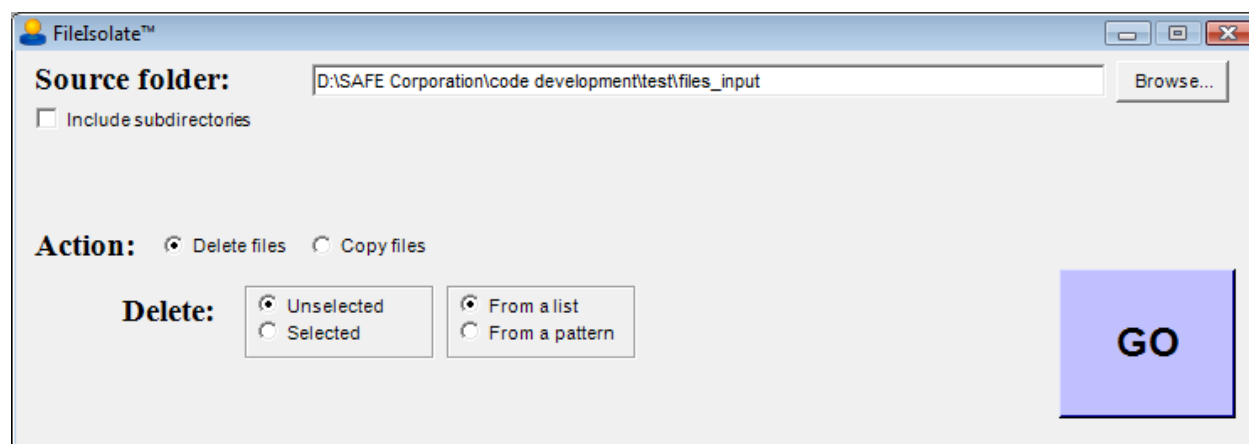
Filesolate

Running Filesolate

Filesolate allows files and file types to be selectively copied or deleted from an entire directory tree.

Deleting files

Below is a screen shot of the Filesolate form where the option is selected to delete files. Following that are step-by-step instructions for running Filesolate to delete files.



Step 1

Choose the **Delete files** action.

Step 2

Select the folder where the files are that need to be deleted by clicking on the browse button or entering the path in the text field. Check the box to include all subdirectories.

Step 3

Choose options for deleting files.

- **Unselected files.** Choose this option to delete all files and file types that are not selected.
- **Selected files.** Choose this option to delete all files and file types that are selected.

Choose options for selecting files.

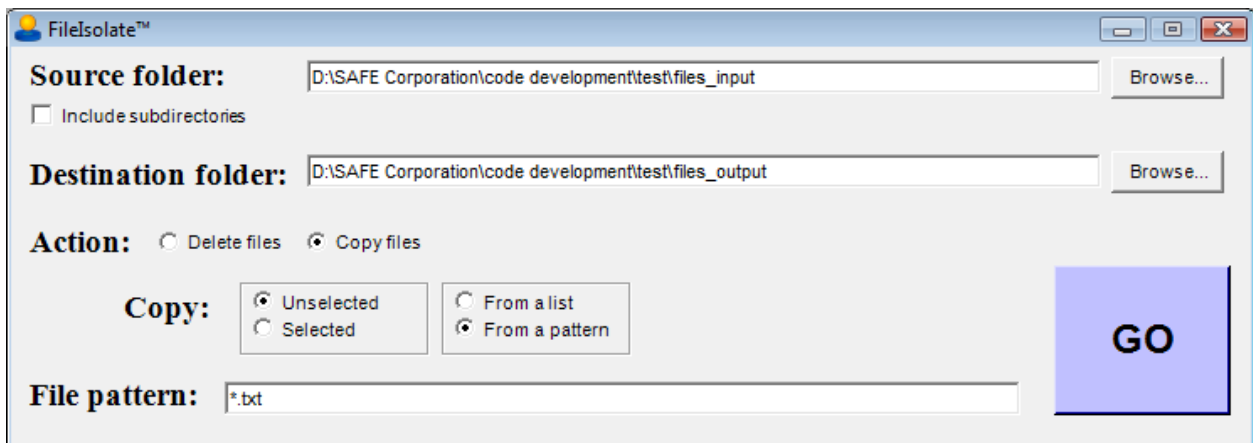
- **From a list of files.** Choose this option to select all files that are named in a text file. You will be prompted for the file containing the list of files. All files that have a name in this list will be selected.
- **From a file pattern.** Choose this option to select files whose names fit a pattern. A field will appear that allows you to type in file patterns. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 4

Press the go button. Filesolate will then search the directory and all subdirectories, if specified. Filesolate will delete all selected files or delete all files that were not selected, depending on the options specified.

Copying files

Below is a screen shot of the Filesolate form where the option is selected to copy files. Following that are step-by-step instructions for running Filesolate to copy files.



The screenshot shows the Filesolate™ application window. The 'Source folder' field contains 'D:\SAFE Corporation\code development\test\files_input' and the 'Destination folder' field contains 'D:\SAFE Corporation\code development\test\files_output'. Both fields have 'Browse...' buttons. The 'Action' section has two radio buttons: 'Delete files' (unselected) and 'Copy files' (selected). Under the 'Copy' section, there are two groups of radio buttons: 'Unselected' (selected) and 'Selected' (unselected), and 'From a list' (unselected) and 'From a pattern' (selected). The 'File pattern' field contains '*.txt'. A large blue 'GO' button is located on the right side of the form.

Step 1

Choose the **Copy files** action.

Step 2

Select the source folder where the files are that need to be copied by clicking on the browse button or entering the path in the text field. Check the box to include all subdirectories.

Step 3

Select the destination folder where the files are to be copied by clicking on the browse button or entering the path in the text field. If the destination folder does not exist, it will be created.

Step 4

Choose options for copying files.

- **Unselected files.** Choose this option to copy all files and file types that are not selected.
- **Selected files.** Choose this option to copy all files and file types that are selected.

Choose options for selecting files.

- **From a list of files.** Choose this option to select all files that are named in a text file. You will be prompted for the file containing the list of files. All files that have a name in this list will be selected.
- **From a file pattern.** Choose this option to select files whose names fit a pattern. A field will appear that allows you to type in file patterns. Separate different file types with a semicolon. Use the * and ? wildcard characters if needed.

Step 5

Press the go button. Filelsolate will then search the directory and all subdirectories, if specified. Filelsolate will copy all selected files or copy all files that were not selected, depending on the options specified.

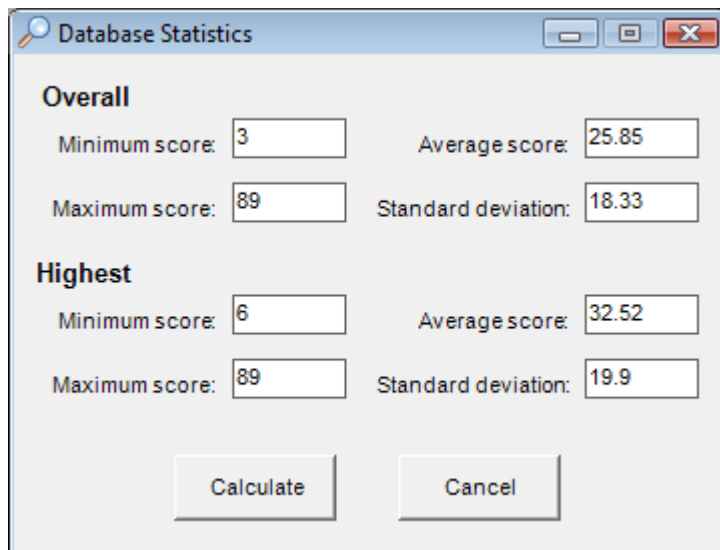
Statistics

Calculating Statistics

The following statistics can be calculated from a CodeSuite database:

- Minimum score of all of the scores in the database.
- Maximum score of all of the scores in the database.
- Average score of all of the scores in the database.
- Standard deviation of all of the scores in the database.
- Minimum score of the highest scoring file pairs for each file in folder 1 in the database.
- Maximum score of the highest scoring file pairs for each file in folder 1 in the database.
- Average score of the highest scoring file pairs for each file in folder 1 in the database.
- Standard deviation of all of the highest scoring file pairs for each file in folder 1 in the database.

Below is a screen shot of the database statistics form. Simply open a database and press the Calculate button.



The screenshot shows a window titled "Database Statistics" with a search icon in the top-left corner and standard window controls (minimize, maximize, close) in the top-right. The window is divided into two sections: "Overall" and "Highest".

Overall

Minimum score:	<input type="text" value="3"/>	Average score:	<input type="text" value="25.85"/>
Maximum score:	<input type="text" value="89"/>	Standard deviation:	<input type="text" value="18.33"/>

Highest

Minimum score:	<input type="text" value="6"/>	Average score:	<input type="text" value="32.52"/>
Maximum score:	<input type="text" value="89"/>	Standard deviation:	<input type="text" value="19.9"/>

At the bottom of the window are two buttons: "Calculate" and "Cancel".

SourceDetective

Running SourceDetective

SourceDetective considers each statement, comment, and identifier in the database and searches the Internet for references to each one. The number of times a statement, comment, or identifier is found in the search is then inserted into a new copy of the database, leaving the original database intact. Spreadsheets can then be generated to show the number of "hits" for each element. For more information, see the section entitled Creating Spreadsheets. Databases can be filtered to remove elements that have large hit numbers, meaning they are commonly found in other programs or documents. For more information, see the section entitled Using Filters.



Step 1

Choose options for searching the Internet.

- **Search for statements.** Choose this option to search the Internet for all statements in the current database. Note that when running SourceDetective on a CodeDiff database, this is the only option that produces results.
- **Search for comments/strings.** Choose this option to search the Internet for all comments and strings in the current database.
- **Search for identifiers.** Choose this option to search the Internet for all identifiers in the current database.

Step 2

Press the go button to begin the search. One search will be performed for each option selected and the number of "hits" will be inserted into the current database.

Exporting databases

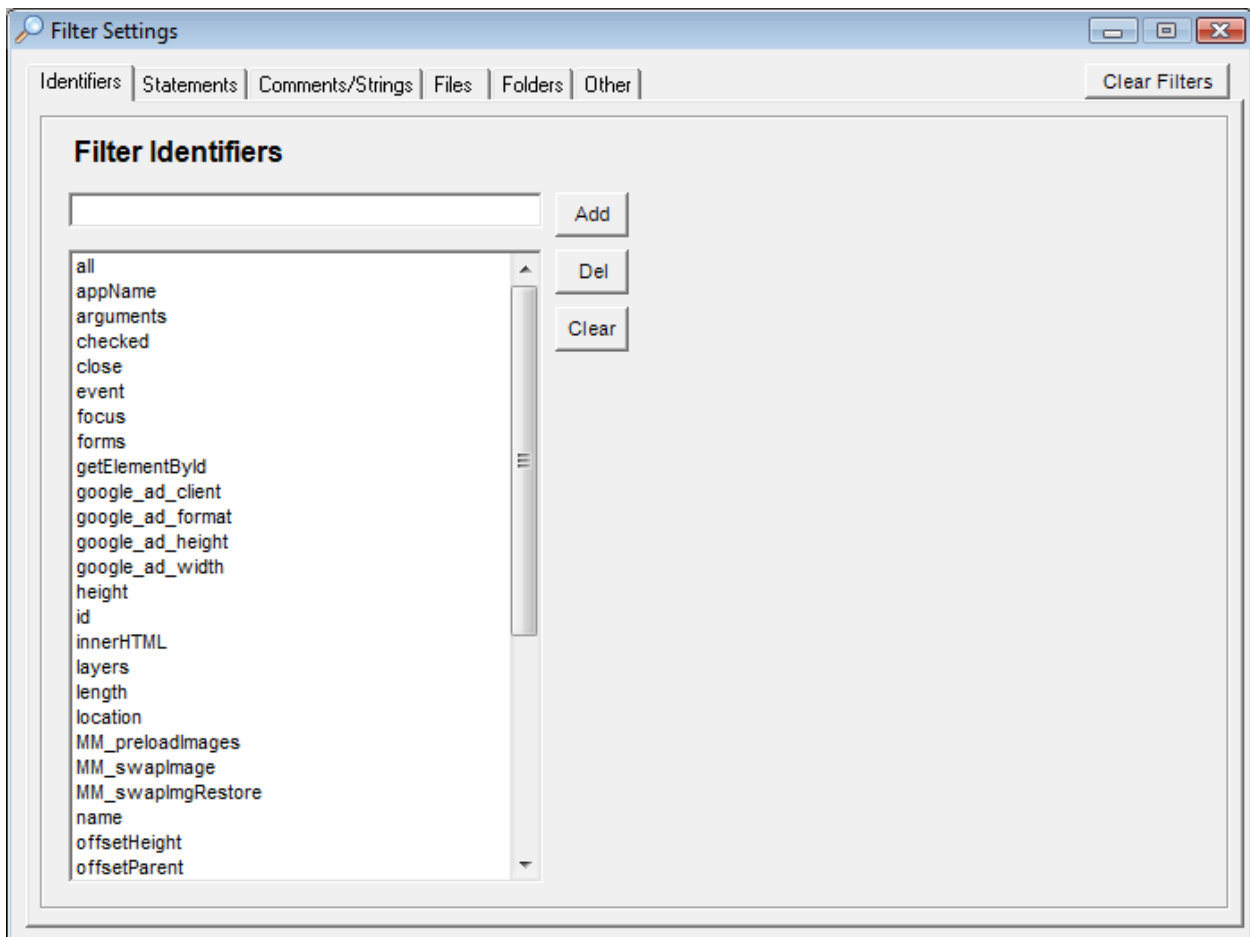
Filters

Filters are used to filter information out of a CodeSuite database that is not relevant to the analysis. For example, you may realize that certain files did not need to be included in the analysis and so you can filter them out. Or you may realize that certain identifier names or comments are very common and they can be ignored in the analysis. Filtering allows you to remove these elements from the analysis without needing to run the program again. When filtering a database, a new filtered database is created, leaving the original database intact.

The filter form has several tabbed forms that are shown and described below.

Identifier Filters

The identifier filter form is shown below.

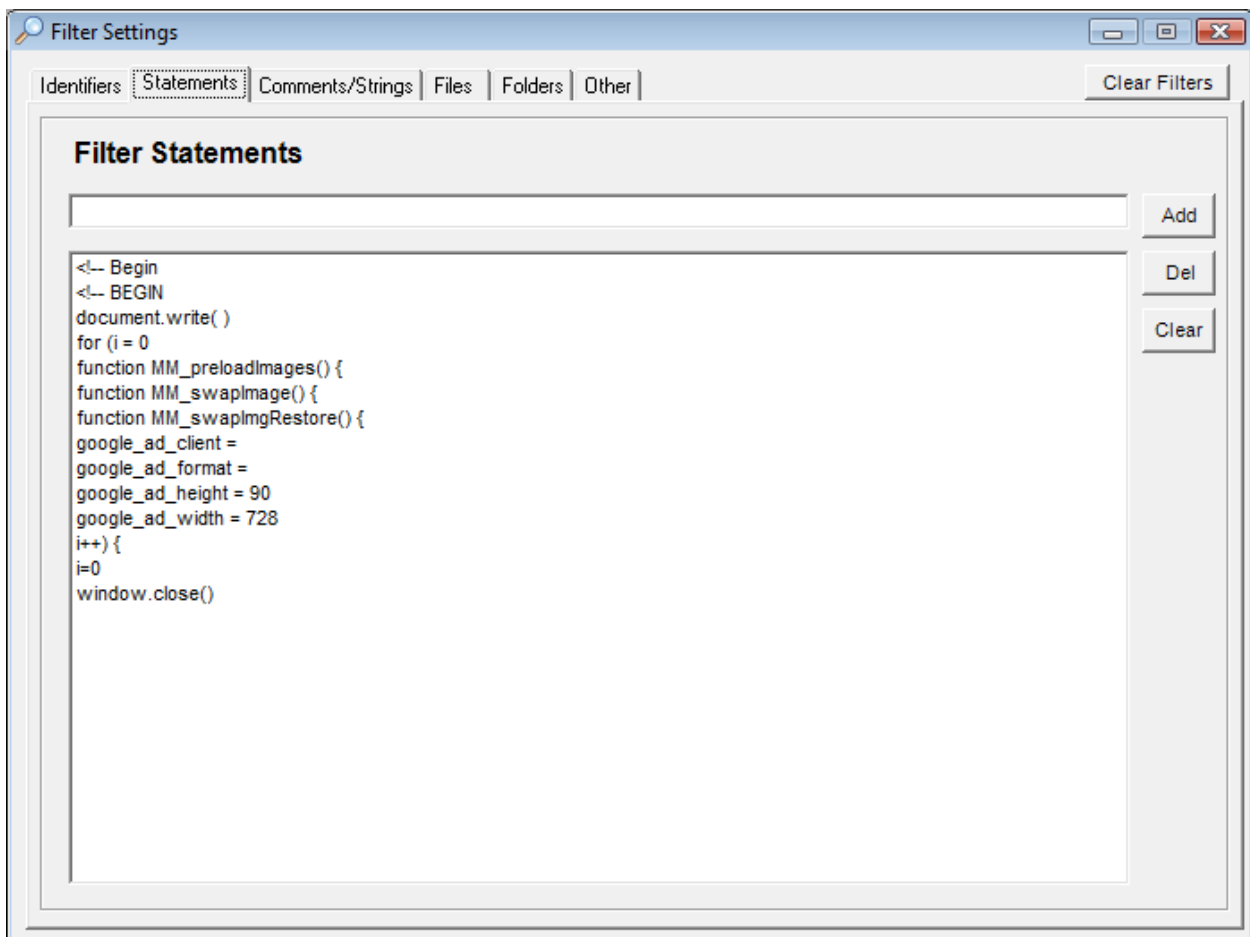


The identifier filter is used to filter out specific identifiers from a CodeSuite database. Enter a specific identifier in the field at the top and click on the Add button to add the identifier to the list of identifiers to be filtered. Select one or several identifiers from the list and click on the Del button to remove the identifiers from the list. Click on the Clear button to clear the list. When the database is filtered, the identifiers on the list will be removed from the database and the file pair scores will be adjusted accordingly.

Note that identifiers are only produced in databases generated by BitMatch or CodeMatch but not those generated by CodeDiff, so identifier filtering will have no effect on a CodeDiff generated database.

Statement Filters

The statement filter form is shown below.

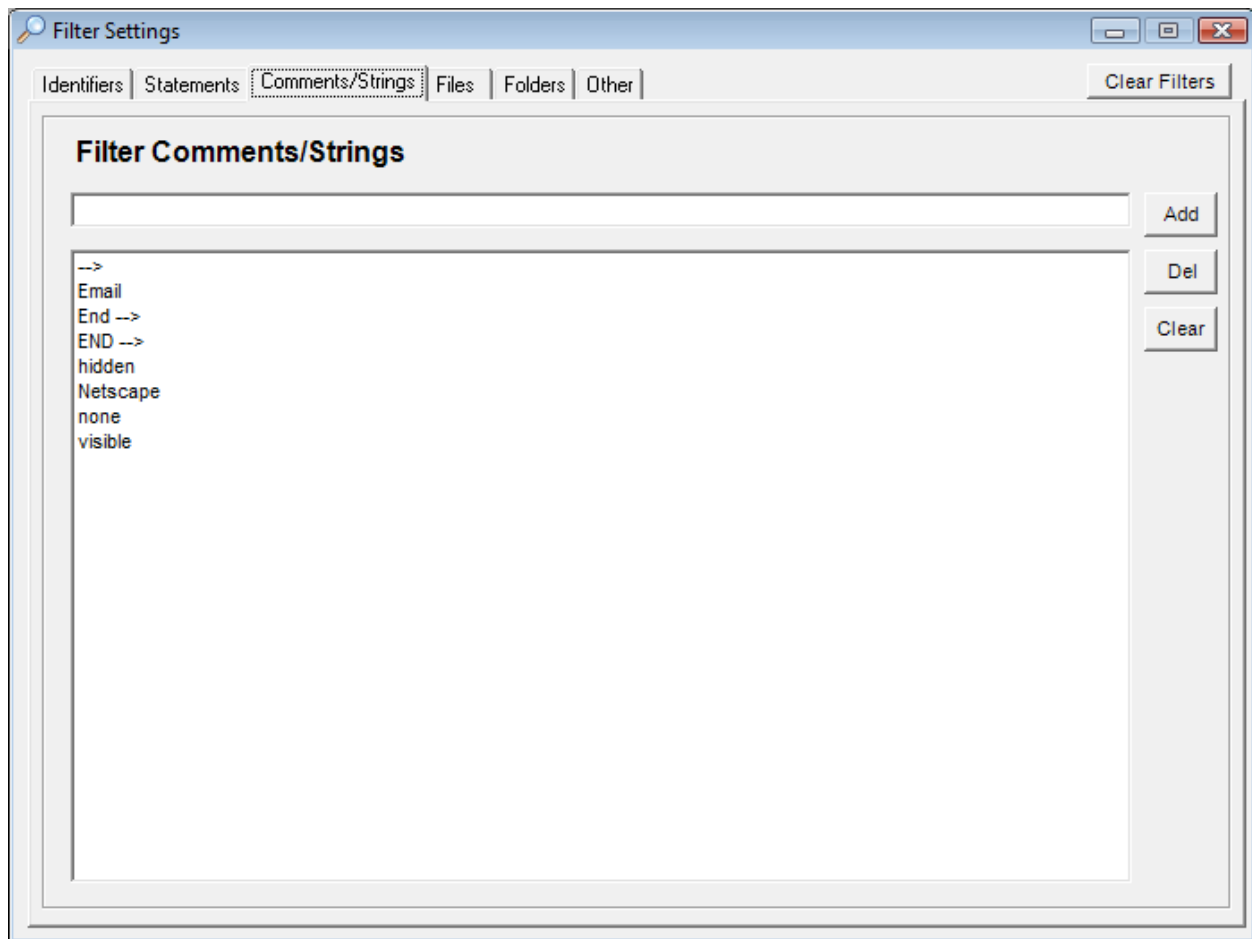


The statement filter is used to filter out specific statements from a CodeSuite database. Enter a specific statement in the field at the top and click on the Add button to add the statement to the list of statement to be filtered. Select one or several statements from the list and click on the Del button to remove the statements from the list. Click on the

Clear button to clear the list. When the database is filtered, the statements in the list will be removed from the database and the file pair scores will be adjusted accordingly. Statement filtering always ignores whitespace.

Comment/String Filters

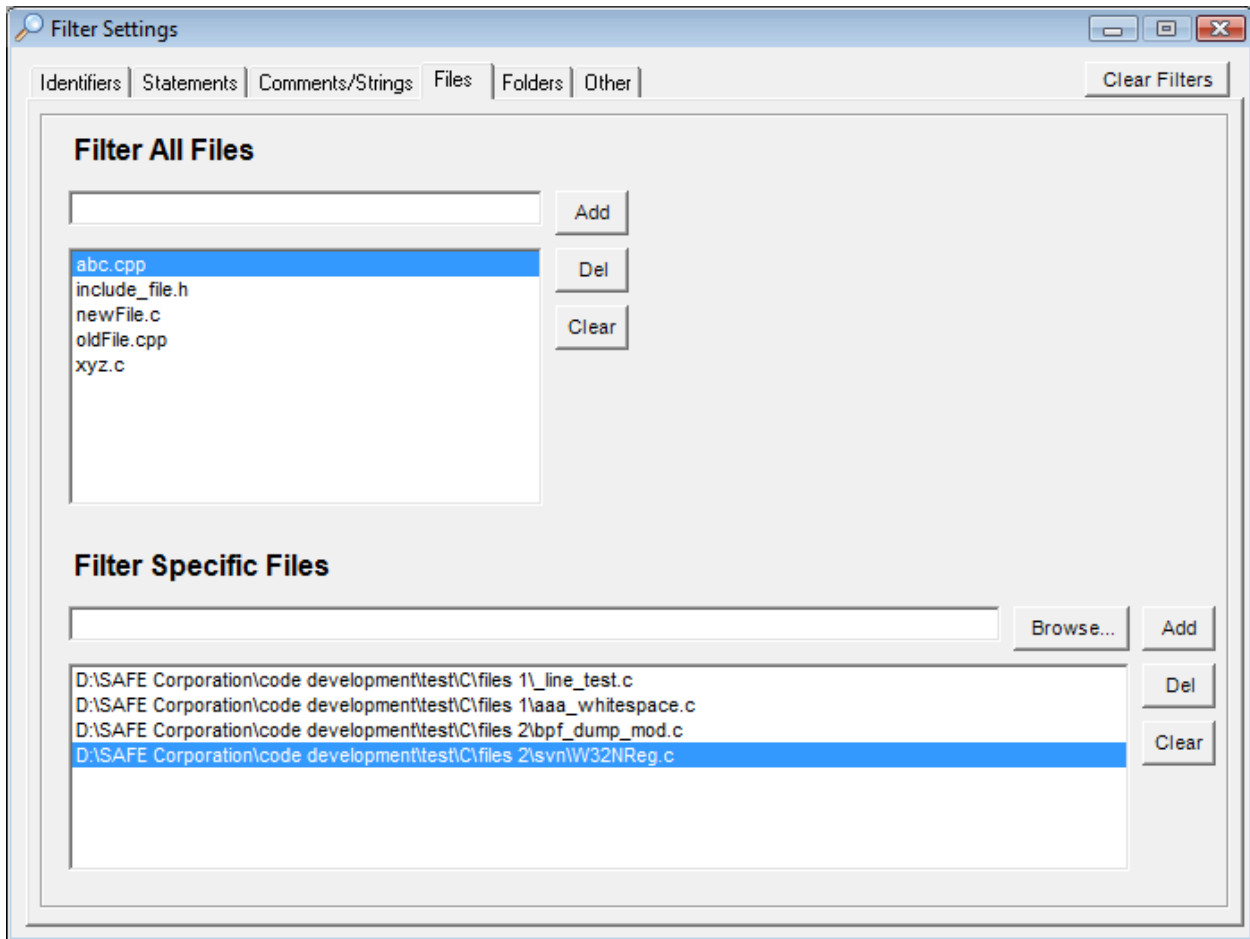
The comment/string filter form is shown below.



The comment/string filter is used to filter out specific comments and strings from a CodeSuite database. Enter a specific comment or string in the field at the top and click on the Add button to add it to the list to be filtered. Select one or several items from the list and click on the Del button to remove them from the list. Click on the Clear button to clear the list. When the database is filtered, the comments and strings in the list will be removed from the database and the file pair scores will be adjusted accordingly. Comment/string filtering always ignores whitespace.

File Filters

The file filter form is shown below.



Filter All Files

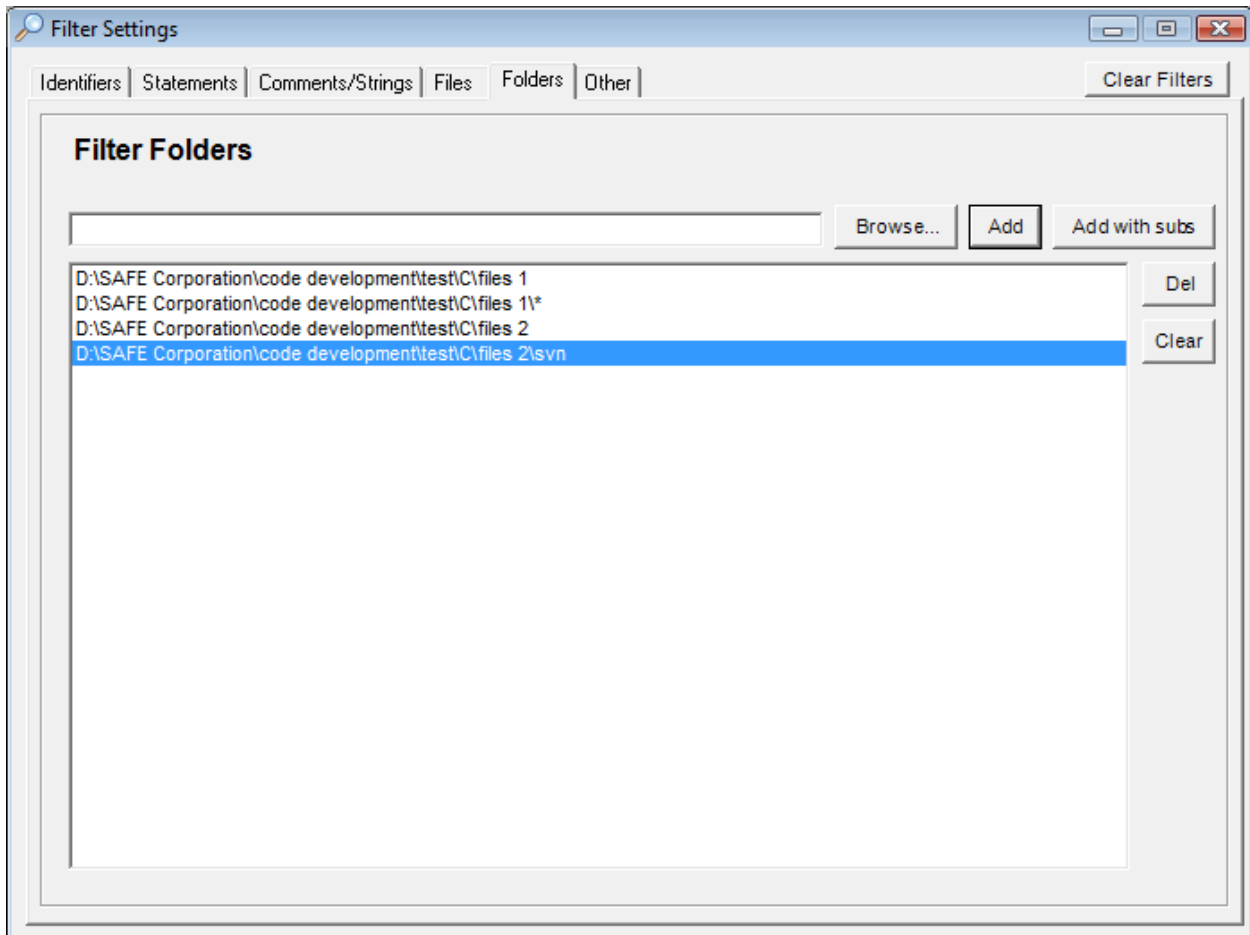
Filtering all files is used to filter out file names from a CodeSuite database. Enter a specific file name in the field at the top and click on the Add button to add the file to the list of files to be filtered. Select one or several files from the list and click on the Del button to remove the files from the list. Click on the Clear button to clear the list. When the database is filtered, the files in the list will be removed from the database, as if those files had never been compared. Note that you can use the * and ? wildcard characters.

Filter Specific Files

Filtering specific files is used to filter out specific files in specific paths from a CodeSuite database. Enter a specific file name and file path in the field at the top or click on the Browse button to select a specific file name and file path to be added to the filter. Click on the Add button to add the file to the list of files to be filtered. Select one or several files from the list and click on the Del button to remove the files from the list. Click on the Clear button to clear the list. When the database is filtered, the specific files in the list will be removed from the database, as if those files had never been compared.

Folder Filters

The folder filter form is shown below.



The folder filter is used to filter out specific folders from a CodeSuite database. Enter a specific folder in the field at the top or click on the Browse button to select a specific folder to be added to the filter. Click on the Add button to select the folder to be added to the list of folders to filter. Click on the Add with subs button to select a folder and all subfolders (specified by the trailing *) to be added to the list of folders to filter. Select one or several folders from the list and click on the Del button to remove the folders from the list. Click on the Clear button to clear the list.

Other Filters

The other filter form is shown below.

Filter Settings

Identifiers | Statements | Comments/Strings | Files | Folders | Other | Clear Filters

Filter Thresholds

Minimum Score: Maximum Files:

Maximum Score:

Filter Hits

Minimum Hits:

Maximum Hits:

Filter Statements Yes No

Filter Comments/Strings Yes No

Filter Sequences Yes No

Filter Identifiers Yes No

Filter Partial Identifiers Yes No

Filter Thresholds

The other filter form allows a reduction of the number of file pairs in the database. When the database is filtered, all file pairs that fall outside the maximum and minimum scores will be removed from the database. Also only the maximum number of file pairs with the highest scores will remain in the database. Leaving a field blank will result in no threshold filter.

Filter Hits

The other filter form allows elements to be filtered out based on the number of hits found on the Internet using SourceDetective. When the database is filtered, all matching elements with hit values that fall outside the maximum and minimum hit values will be removed from the database and the scores will be modified appropriately. Leaving a field blank will result in no hit filter.

Filter Statements

The other filter form allows statements to be filtered out of the database. Selecting yes will eliminate statements from the database and modify the scores appropriately. For CodeDiff and CodeCLOC, this will filter out lines in the database.

Filter Comments/Strings

The other filter form allows comments and strings to be filtered out of the database. Selecting yes will eliminate comments and strings from the database and modify the scores appropriately.

Filter Sequences

The other filter form allows instruction sequences to be filtered out of the database. Selecting yes will eliminate instruction sequence from the database and modify the scores appropriately.

Filter Identifiers

The other filter form allows identifiers to be filtered out of the database. Selecting yes will eliminate identifiers from the database and modify the scores appropriately.

Filter Partial Identifiers

The other filter form allows all partial identifiers to be filtered out of the database. Selecting yes will eliminate all partial identifiers from the database and modify the scores appropriately.

Clear Filters

Pressing this button clears all filters on each filter form.

HTML reports

A CodeSuite database can be automatically turned into HTML reports for easy reading and presentation of results. A basic report is generated that shows file pairs and their scores. By clicking on the score, a detailed HTML report will come up. These detailed reports are kept in subfolders. The detailed reports give more information about how the score was determined, showing specific similarities or differences between the files. The file names are given at the top of the report and include a hyperlink that, when clicked, allows the file to be brought up in a viewer or editor. The back and next buttons on the detailed reports allow you to navigate the detailed reports without going back to the basic report.

For information on BitMatch reports, see the sections entitled BitMatch Basic Report and BitMatch Detailed Report.

For information on CodeCLOC reports, see the sections entitled CodeCLOC Basic Report and CodeCLOC Detailed Report.

For information on CodeCross reports, see the sections entitled CodeCross Basic Report and CodeCross Detailed Report.

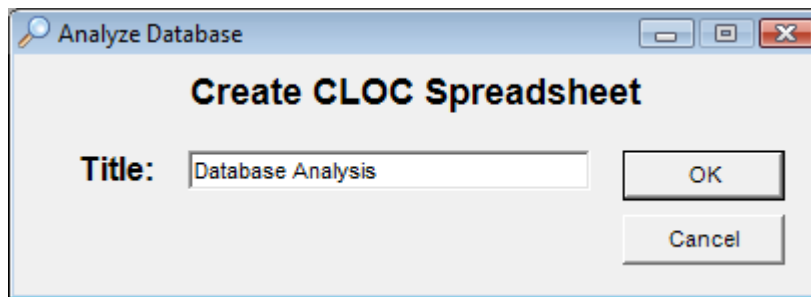
For information on CodeDiff reports, see the sections entitled CodeDiff Basic Report and CodeDiff Detailed Report.

For information on CodeMatch reports, see the sections entitled CodeMatch Basic Report and CodeMatch Detailed Report.

CLOC Spreadsheets

A CLOC spreadsheet shows the Changing Lines Of Code (CLOC) measurement produced by CodeCLOC) among the two sets of files belonging to two versions of a program. A CodeCLOC comparison is essentially a CodeDiff comparison of files with the same name such that each file is used at most once and the scores are optimized such that the overall percentage of similarity between the two sets of files is maximized. Note that perfect optimization would require compute time that grows unacceptably long for even small groups of files, so a local optimization is performed in a reasonable amount of time. In some cases, CLOC measurements on the same sets of files may result in very slightly different numbers.

Below is the form you use to create a CLOC spreadsheet from the File menu.



The title that you enter will appear at the top of spreadsheet. Press the OK button and you will be asked to name the spreadsheet file.

Example CLOC Spreadsheet

Below is an example of a spreadsheet created from a CodeCLOC database.

	A	B
1	CodeCLOC CLOC Spreadsheet	
2		
3	Database:	C:\SAFE Corporation\code development\test\C\results\CodeCLOC.cdb
4	Database Analysis	
5	Analysis date:	4/27/2010
6		
7	Version 1 directory:	C:\SAFE Corporation\code development\test\C\files1
8		including subdirectories
9	Version 2 directory:	C:\SAFE Corporation\code development\test\C\files2
10		

User's Guide

11		
12	File types:	*.cpp;*.c;*.h
13		
14	Total files in first set [TF0]	150
15	Total lines of code in the first set [LOC0]	64829
16	Total bytes in the first set [TB0]	2246044
17	Total files in the second set [TF1]	809
18	Total lines of code in the second set [LOC1]	291038
19	Total bytes in the second set [TB1]	10741835
20		
21	Change Summary	
22	Files	
23	New files [NF]	739
24	Modified continuing files [MCF]	70
25	Lines	
26	New and modified lines of code [CLOC]	275635
27		
28	Continuation Summary	
29	Files	
30	Continuing files [CF]	70
31	Unchanged continuing files [UCF]	0
32	Lines	
33	LOC in continuing files [LOCinCF]	49313
34	CLOC in continuing files [CLOCinCF]	33910
35	Unchanged continuing lines of code [ULOC]	15403
36		
37	Summary Statistics	
38	Files	
39	Percent of new and modified files ((NF+MCF)/TF1)	100
40	Percent of new and modified files relative to base version ((NF+MCF)/TF0)	539
41	Percent of continuing files (CF/TF1)	9
42	Percent of unchanged continuing files (UCF/TF1)	0
43	Lines	
44	Percent CLOC change (CLOC/LOC1)	95
45	Percent CLOC change relative to base version (CLOC/LOC0)	425
46	Percent LOC growth (LOC/LOC0)	449
47	Percent unchanged continuing lines of	5

code (ULOC/LOC1)	
------------------	--

The top line shows that the spreadsheet was created from a CodeSuite database generated by CodeCLOC. The third line gives the name of the database file. The fourth line is the title that the user placed in the spreadsheet form shown above. The analysis date shows the date that the spreadsheet was created. The folders are the ones that were compared to create the database including subdirectories if noted.

The quantities in the spreadsheet are described below. The two versions being compared consist of the initial *base* version of code and a subsequent *examined* version that is being compared to it.

- **Total files [TF0]:** The total number of files in the base version.
- **Total lines of code [LOC0]:** The total lines of code (blank lines ignored) in the base version.
- **Total bytes [TB0]:** The total bytes of code in the base version.
- **Total files [TF1]:** The total number of files in the examined version.
- **Total lines of code [LOC1]:** The total lines of code (blank lines ignored) in the examined version.
- **Total bytes [TB1]:** The total bytes of code in the examined version.
- **New files [NF]:** The number of files that exist in the examined version but did not exist in the base version.
- **Modified continuing files [MCF]:** The number of continuing files (files that exist in both the examined version and the base version) with new lines of code.
- **New and modified lines of code [CLOC]:** The number of lines of code in the examined version that are new or have been modified from the base version.
- **Continuing files [CF]:** The number of files in the examined version that were also in the base version.
- **Unchanged continuing files [UCF]:** The number of continuing files that have not changed from the base version to the examined version.
- **Lines of code in continuing files [LOCinCF]:** The number of lines of code in the continuing files.
- **New and modified lines of code in continuing files [CLOCinCF]:** The number of lines of code in the continuing files that are new or have been modified from the base version.
- **Unchanged continuing lines of code [ULOC]:** The number of lines of code in continuing files that have not changed.

- **Percent of new and modified files:** $((NF+MCF)/TF1)$
- **Percent of new and modified files relative to base version:** $((NF+MCF)/TF0)$
- **Percent of continuing files:** $(CF/TF1)$
- **Percent of unchanged continuing files:** $(UCF/TF1)$
- **Percent CLOC change:** $(CLOC/LOC1)$
- **Percent CLOC change relative to base version:** $(CLOC/LOC0)$
- **Percent LOC growth:** $(LOC1/LOC0)$
- **Percent unchanged continuing lines of code:** $(ULOC/LOC1)$

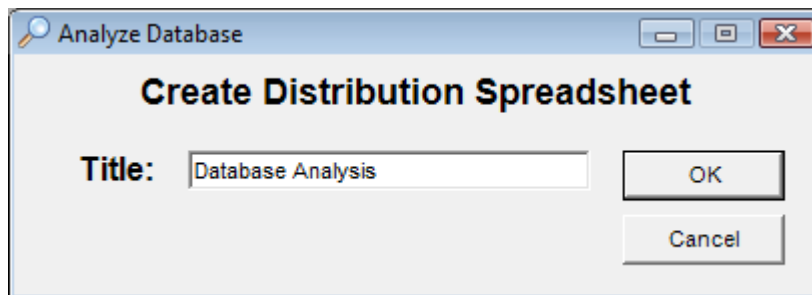
Distribution Spreadsheets

A distribution spreadsheet shows the distribution of scores (typically similarity scores from CodeDiff) among the two sets of files compared. It is important to note that for each file in the first set of compared files, a distribution spreadsheet only considers the similarity scores for the most similar file in the second set of files compared. So for each file in the first set of files, the spreadsheet will only consider the most similar file in the second set, even other less similar files are noted in the database. Also note that while each file in the first set of files is considered exactly once, some files in the second set of files may be considered multiple times or not at all.

Consider set A consisting of files A1 and A2. Consider set B consisting of files B1 and B2. Now consider the table below of similarity scores. The file with the most similarity to file A1 is file B1 and the file with the most similarity to file A2 is also file B1. So file B1 will be counted twice in the distribution spreadsheet.

Set A	Set B	Similarity Score
File A1	File B1	90
File A1	File B2	20
File A2	File B1	67
File A2	File B2	12

Below is the form you use to create a distribution spreadsheet from the File menu.



The title that you enter will appear at the top of spreadsheet. Press the OK button and you will be asked to name the spreadsheet file.

Example Distribution Spreadsheet

Below is an example of a spreadsheet created from a CodeDiff database.

	A	B	C	D	E	F
1	CodeDiff Results Distribution					
2	Database Analysis					

User's Guide

3	Database	C:\SAFE Corporation\code development\test\C\results\CodeMatch1.cdb				
4	Run date	12/1/2007				
5	Analysis date	12/2/2007				
6						
7	Folder 1	C:\SAFE Corporation\code development\test\C\files1				
8		including subdirectories				
9	Folder 2	C:\SAFE Corporation\code development\test\C\files2				
10						
11						
12	Algorithm	Ignoring case				
13	Algorithm	Ignoring whitespace				
14	Algorithm	Percentage of file pairs				
15	File threshold	8				
16	Score threshold	1				
17						
18	Total files in folder 1	16				
19						
20	File pair comparisons					
21			Folder 1		Folder 2	
22	Match percentage	Number of files	Number of lines	Number of bytes	Number of lines	Number of bytes
23	20	1	5	252	5	230
24	28	1	291	7725	227	5010
25	75	2	8	224	8	228
26	99	1	56	1997	52	1890
27	100	11	960	27288	960	27282
28						
29	Totals	16	1320	37486	1252	34640
30	Total changed	5	216		170	
31	Percent changed	31.25	16.37		13.58	

The top line shows that the spreadsheet was created from a CodeSuite database generated by CodeDiff. The second line is the title that the user placed in the spreadsheet form shown above. The run date shows the date that CodeDiff was run while the analysis date shows the date that the spreadsheet was created. The folders are the ones that were compared to create the database.

The algorithms show all algorithms that were selected for the CodeDiff run. In this case, letter case and whitespace were ignored and the percentage given is the percentage of lines in the first file that were matched in the second file. The file threshold shows that

the top 8 files were reported. The score threshold shows that files needed at least 1% similarity to be reported. Note that the database may have been subsequently filtered and that this fact is not reflected in the spreadsheet.

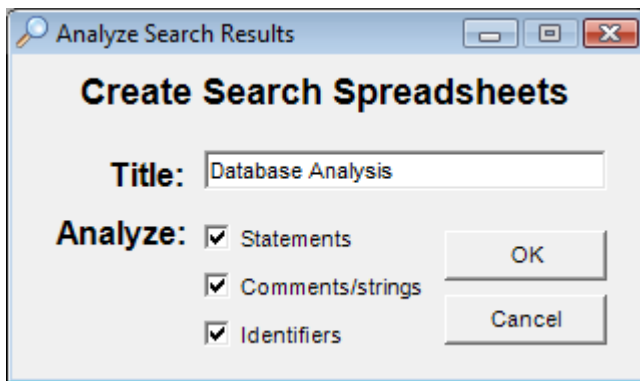
The total number of files in the first folder (including all subfolders if that was selected for the CodeDiff run) is shown as 16.

Rows 21 and 22 are the table header. Column A shows a percentage score while column B shows the total number of files. Columns C and D show the total number of lines and bytes in files in folder 1 that had this particular score. Columns E and F show the total number of lines and bytes in files in folder 2 that had this particular score. Note that only the highest percentage score is considered for this analysis. In other words, if file A in folder 1 is matched 97% by file X in folder 2, 93% by file Y in folder 2, and 37% by file Z in folder 2, only file X in folder 2 is considered in the analysis.

Search Spreadsheets

A search spreadsheet shows the number of times a code element (i.e., statement, comment, string, or identifier) can be found when searching the Internet using SourceDetective.

Below is the form you use to create a search spreadsheet from the File menu. Before a search spreadsheet can be created, SourceDetective must be run on the database to search the Internet for code elements such as statements, comments, and identifiers.



The title that you enter will appear at the top of spreadsheet. Press the OK button and you will be asked to name the spreadsheet file.

Example Search Spreadsheet

Below is an example of a spreadsheet created from a CodeMatch database.

	A	B
1	CodeMatch Internet Search Results	
2	Database Analysis	
3	Database	C:\SAFE\code development\test\C\results\CodeMatch1.cdb
4	Run date	12/1/2007
5	Analysis date	12/2/2007
6		
7	Folder 1	C:\SAFE Corporation\code development\test\C\files1
8		including subdirectories
9	Folder 2	C:\SAFE Corporation\code development\test\C\files2
10		
11		

12	Statements	Search Score
13	#include <Assert.h>	2147483647
14	#include <time.h>	2147483647
15	&hKeyAdapter	5
16	(void)sprintf(image	29
17	bpf_dump(struct bpf_program *p, int option)	37
18	case BPF_ALU BPF_LSH BPF_K:	100
19	case BPF_LD BPF_B BPF_ABS:	155
20	char xchar	398
21	hKeyClassNet,	1
22	if(nPlatformVersion == 0x0004)	0
23	int number222	0
24	NULL,	226000000
25	switch (p->code) {	57
26	xyz abc	110000

The top line shows that the spreadsheet was created from a CodeSuite database generated by CodeMatch. The second line is the title that the user placed in the spreadsheet form shown above. The run date shows the date that CodeMatch was run while the analysis date shows the date that the spreadsheet was created. The folders are the ones that were compared to create the database.

The next line shows that this is a listing of statements. The left column lists statements in alphabetical order. The right column gives the number of hits on the search engine for this statement.

Summary Spreadsheet

Below is the form you use to create a summary spreadsheet from the File menu.

The screenshot shows a dialog box titled 'Analyze Database' with the main heading 'Create Summary Spreadsheet'. It contains two input fields: 'Title:' with the text 'Database Analysis' and 'Interval:' with the text '10'. There are 'OK' and 'Cancel' buttons to the right of the input fields.

The title that you enter will appear at the top of spreadsheet. The interval is used to divide up file pair scores for the spreadsheet. Press the OK button and you will be asked to name the spreadsheet file.

Example CodeMatch Spreadsheet

Below is an example of a spreadsheet created from a CodeMatch database.

	A	B	C	D	E	F	G	H
1	CodeMatch Results Summary							
2	Database Analysis							
3	Run date	7/4/2006						
4	Analysis date	8/1/2006						
5								
6	Algorithm	Statement Matching						
7	Algorithm	Comment/String Matching						
8	Algorithm	Identifier Matching						
9	Algorithm	Instruction Sequence Matching						
10	File threshold	4						
11								
12	Total files in folder 1	22						
13	Total file pairs	83						
14								
15	Scores		(0-9)	(10-19)	(20-29)	(30-39)	(40-49)	(50-59)
16	Numbers		17	15	26	21	0	4
17	Percentage		20	18	31	25	0	5

The top line shows that the spreadsheet was created from a CodeSuite database generated by CodeMatch. The second line is the title that the user placed in the spreadsheet form shown above. The run date shows the date that CodeMatch was run while the analysis date shows the date that the spreadsheet was created.

The algorithms show all algorithms that were selected for the CodeMatch run. In this case, the statement matching, comment matching, identifier matching, and instruction sequence matching algorithms were all selected. The threshold shows that the top four files that were correlated to the files in folder 1 were reported. Note that the database may have been subsequently filtered and that this fact is not reflected in the spreadsheet.

The total number of files in the first folder (including all subfolders if that was selected for the CodeMatch run) is shown as 22. The total number of files pairs that are reported in the database is 83.

Row 15 is the table header. Columns C through H show intervals of match scores (correlation scores). For example column C represents match scores between 0 and 9 inclusively.

Row 16 shows the number of file pairs in each match score interval while row 17 shows the percentage of file pairs in each match score interval.

Languages

Languages Supported

The following programming languages are currently supported by BitMatch and CodeMatch:

- ABAP
- ASM-M68k
- BASIC
- C
- C++
- C#
- Delphi
- Flash ActionScript
- Fortran
- FoxPro
- Java
- JavaScript
- LISP
- MASM assembly for Intel processors
- MATLAB
- Pascal
- Perl
- PHP
- PowerBuilder
- Python
- RealBasic
- Ruby
- SQL
- Verilog
- VHDL
- Visual Basic

Check the SAFE Corporation website for new language modules, available at no charge, as they become available. If the language you need is not available, contact SAFE Corporation about creating it for a nominal fee.

Advanced topics

CodeSuite database format

SAFE Corporation wants third party developers to create software to analyze CodeSuite database files and generate reports and statistics from them. In light of that desire, below is an example of a CodeSuite database file with all sections explained by a comment. Comments are lines beginning with a # symbol.

Tag Definitions

The CodeSuite database tags and their meanings are given in the table below.

Tag	Definition
#	First character in a comment
<Program>	Program name
<Version>	Program version
<Date>	Date file was created
<Time>	Time file was created
<LComment>	Left-justified comment to be placed in the output file
<Describe>	Insert program description
<CComment>	Right-justified comment to be placed in the output file
<Filters>	Beginning of filters used to create this database (see Filters section)
</Filters>	End of filters used to create this database
<Folder1>	Input folder 1
<Subs1>	Compare subdirectories for folder 1? T or F
<Folder2>	Input folder 2
<Subs2>	Compare subdirectories for folder 2? T or F
<Language>	Programming language
<FileType>	Filetype
<SameName>	Compare files of same name only? T or F
<Dirs>	Compare directories? T or F
<Algorithm>	Algorithm name
<FileThresh>	File number threshold to report
<ScoreThresh>	Score threshold to report
<Filter>	Filter name
<Dir1>	Input file 1 path relative to InFolder1
<Dir2>	Input file 2 path relative to InFolder2
<File1>	Input file 1 name
<NumWords1>	Number of words in file 1 (for DocMate)
<NumLines1>	Number of lines in file 1
<Size1>	Number of bytes in file 1
<File2>	Input file 2 name
<NumWords2>	Number of words in file 2 (for DocMate)
<NumLines2>	Number of lines in file 2

<Size2>	Number of bytes in file 2
<StatementScore>	Statement match score
<CommentScore>	Comment match score
<SequenceScore>	Sequence match score
<IdentifierScore>	Identifier match score
<SubIdentifierScore>	Partial identifier match score
<Score>	Match score
<NoScore>	File was not compared -- no score
<Differences>	Beginning of line differences
</Differences>	End of line differences
<Statements>	Beginning of statement comparison
</Statements>	End of statements comparison
<Comments>	Beginning of comment comparison
</Comments>	End of comment comparison
<Lines1>	Lines in file 1
<Lines2>	Lines in file 2
<Line>	Instruction/comment line
<Sequences>	Beginning of instruction sequences
<InSeq>	Instruction sequence
<Instr>	Instruction in sequence
</Sequences>	End of instruction sequences
<IDs>	Beginning of identifiers
<ID>	Identifiers
</IDs>	End of identifiers
<PIDs>	Beginning of partial identifiers
<PID1>	Partial identifiers in file 1
<PID2>	Partial identifiers in file 2
</PIDs>	End of partial identifiers
<BingHits>	Number of hits on Bing search engine for preceding elements
<YahooHits>	Number of hits on Yahoo search engine for preceding elements
<Folder1Bytes>	Total bytes in all files examined in folder 1
<Folder1Lines>	Total lines in all files examined in folder 1
<Folder1Files>	Total number of files examined in folder 1
<Folder2Bytes>	Total bytes in all files examined in folder 2
<Folder2Lines>	Total lines in all files examined in folder 2
<Folder2Files>	Total number of files examined in folder 2
<ExTime>	Time to execute program

Example CodeSuite database file

#Each line begins with a tag, ends with a newline
 #Comments begin with #, end with a newline

```
<Program>CodeMatch
<Version>version
<Date>date
<Time>time
```

```
<CComment>Centered comment
<LComment>Left justified comment
```

```

<Folder1>input_folder
<Subs1>T
<Folder2>compare_folder
<Subs2>F
<Language>programming_language
<CaseSensitive>T
<FileType>filetypes
<SameName>T
<Dirs>T
#Note: <Dirs> tag specifies that comparison is being done on a directory
basis
<Algorithm>algorithm_name
<Algorithm>algorithm_name
<Algorithm>algorithm_name
<FileThresh>number
<ScoreThresh>number

#The following filters (from the filter file) were used to create this
database
<Filters>
<Filter>filter
<Filter>filter
</Filters>

#For each file in the first set of files being compared:
<Dir1>input_file_1_path_relative_to_Folder1
<Dir2>input_file_2_path_relative_to_Folder2
<File1>filename
<NumLines1>number_of_lines
<Size1>size_in_bytes
<File2>filename
<NumLines2>number_of_lines
<Size2>size_in_bytes
#Note: <Dir1> and <Dir2> tags may or may not be repeated if they are
unchanged from subsequent files

#Show differences for CodeDiff
<Differences>
<Line>line
<YahooHits>129375
<Lines1>line_number line_number
<Lines1>line_number line_number
<Line>line
<YahooHits>673
<Lines2>line_number line_number line_number
</Differences>

<Statements>
<Line>instruction
<YahooHits>75
<Lines1>line_number line_number
<Lines1>line_number line_number
<Lines2>line_number line_number line_number
<Lines2>line_number
<Line>instruction
<YahooHits>12

```

User's Guide

```
<Lines1>line_number line_number
<Lines2>line_number line_number line_number
<Line>instruction
<YahooHits>9348753
<Lines1>line_number line_number line_number
<Lines2>line_number line_number line_number
</Statements>
<StatementScore>statement_correlation_score
# Note: If there is no statement score
# there were no statements to compare or
# statements were not compared in this run.

<Comments>
<Line>comment
<YahooHits>2341
<Lines1>line_number line_number line_number
<Lines2>line_number line_number line_number
<Line>comment
<YahooHits>444
<Lines1>line_number line_number
<Lines2>line_number line_number line_number line_number line_number
<Lines2>line_number
</Comments>
<CommentScore>comment_correlation_score
# Note: If there is no comment score
# there were no comments to compare or
# comments were not compared in this run.

#Instruction listing using <Instr> tag is optional
<Sequences>
<InSeq>line_number line_number sequence_length_number
<Instr>Instruction 1
<Instr>Instruction 2
<Instr>Instruction 3
<Instr>Instruction n<InSeq>line_number line_number sequence_length_number
<InSeq>line_number line_number sequence_length_number
<InSeq>line_number line_number sequence_length_number
</Sequences>
<SequenceScore>sequence_correlation_score
# Note: If there is no sequence score
# there were no sequences to compare or
# sequences were not compared in this run.

<IDs>
<ID>identifier_string identifier_string identifier_string identifier_string
<YahooHits>12 888 3 90023
<ID>identifier_string
<YahooHits>129
</IDs>
<PIDs>
<PID1>partial_identifier_string partial_identifier_string
<PID1>partial_identifier_string partial_identifier_string
<PID1>partial_identifier_string
<PID2>partial_identifier_string partial_identifier_string
<PID2>partial_identifier_string partial_identifier_string
<PID2>partial_identifier_string
</PIDs>
```

```
<IdentifierScore>identifier_correlation_score
# Note: If there is no identifier score
# there were no identifiers to compare or
# identifier matching was not selected for this run.

<Score>file_pair_correlation_score

#For each uncomparing file (a file that was never compared to anything):
<Dir1>input_file_1_path_relative_to_Folder1
<File1>filename
<NumLines1>number_of_lines
<Size1>size_in_bytes
<NoScore>

#Summary
<Folder1Bytes>number
<Folder1Lines>number
<Folder1Files>number
<Folder2Bytes>number
<Folder2Lines>number
<Folder2Files>number
<ExTime>execution time
```

CodeSuite filter format

SAFE Corporation wants third party developers to create software to filter CodeSuite database files. In light of that desire, below is an example of a CodeSuite filter file with all sections explained by a comment. Comments are lines beginning with a # symbol.

Tag definitions

The CodeSuite filter tags and their meanings are given in the table below.

Tag	Definition
#	Filter file comment character
<Statement>	Statements to filter
<Comment>	Comments to filter
<Identifier>	Identifiers to filter
<File>	Files to filter
<Folder>	Folders to filter
<Threshold>	Filter threshold (possible parameters are given below)
MinScore	Minimum score threshold value
MaxScore	Maximum score threshold value
MaxFile	Maximum file number threshold value
<AllStatements>	Filter all statements
<AllComments>	Filter all comments
<AllSequences>	Filter all sequences
<AllIdentifiers>	Filter all identifiers
<PartialIDs>	Filter all partial identifiers

Example CodeSuite filter file

```
#Comments begin with #, end with a newline
#Note that all filters are case sensitive
```

```
#List of statements to filter out
<Statement>Statement1
<Statement>Statement2
<Statement>Statement3
```

```
#List of comments to filter out
<Comment>Comment1
<Comment>Comment2
<Comment>Comment3
<Comment>Comment4
```

```
#Filter out all sequences
<AllSequences>
```

```
#List of identifiers to filter out
<Identifier>Identifier1
<Identifier>Identifier2
```

```
<Identifier>Identifier3
<Identifier>Identifier4
<Identifier>Identifier5

#Filter out all partial identifiers
<PartialIDs>

#List of files to filter out
<File>*\file1
<File>*\file2
<File>*\file3
<File>D:\CodeSuite\Code Development\CodeSuite UI\CodeDiff.frx
<File>D:\CodeSuite\Code Development\CodeSuite UI\CodeSuite.vbp
<File>D:\CodeSuite\Code Development\CodeSuite UI\PENS04.ICO

#List of folders to filter out
<Folder>D:\CodeSuite\documents
<Folder>D:\Congregation Beth David\Security

#Threshold filters
<Threshold>MinScore 10
<Threshold>MaxScore 98
<Threshold>FileNum 20
```

Command line interface

Running CodeSuite

CodeSuite can be invoked from the command line. This allows easy integration with other programs and also allow multiple comparisons to be run one after the other in a batch mode.

In order to run CodeSuite from the command line, your PATH environment must point to the CodeSuite location (see below for instructions).

To run a series of CodeSuite functions from a command file, type

```
cscl cfile
```

where:

cfile a command file listing one or more lines of arguments and options

To run a CodeSuite comparison, type

```
cscl -PGn [-option] [-option] dbase dir1 dir2 patt1 patt2 [lang1 lang2]
```

where:

dbase	name of the output database file
dir1	the first directory of files to compare
dir2	the second directory of files to compare
patt1	file patterns in dir1
patt2	file patterns in dir2
lang1	language of files in dir1 (BitMatch, CodeCross, CodeMatch only)
lang2	language of files in dir2 (BitMatch, CodeCross, CodeMatch only)

options:

h	print this description
PGn	which function to run (must be the first option)
	n = 0 for CodeMatch
	n = 1 for CodeDiff
	n = 2 for BitMatch
	n = 3 for CodeCross
	n = 4 for CodeCLOC

	n = 5 for DocMate
RSn	recursively examine subdirectories n = 0 for neither n = 1 for directory1 only n = 2 for directory2 only n = 3 for both (default)
SCn	directory contains source code (BitMatch, CodeCross, CodeMatch only) n = 0 for neither directory (BitMatch default) n = 1 for directory1 only n = 2 for directory2 only n = 3 for both (CodeMatch default)
FTn	n = number of files to report (default = 8)
SNn	compare files with the same name n = 0 for full comparison (default) n = 1 for same name comparison
CSn	compare statements (CodeMatch only) n = 0 for no comparison n = 1 for comparison (default)
CCn	compare comments (CodeMatch only) n = 0 for no comparison n = 1 for comparison (default)
CI n	compare identifiers (CodeMatch, DocMate only) n = 0 for no comparison n = 1 for comparison (default)
CQn	compare instruction sequences (CodeMatch, DocMate only) n = 0 for no comparison n = 1 for comparison without listing sequences in the database (default) n = 2 for comparison with listing sequences in the database
QTn	n = minimum sequence to report (CodeMatch, DocMate only, default = 10)
IWn	ignore whitespace (CodeDiff only) n = 0 to consider n = 1 to ignore (default)
ICn	ignore case (CodeDiff only) n = 0 to consider n = 1 to ignore (default)
PMn	percentage calculation (CodeDiff only) n = 0 for percent of total lines of both files (default) n = 1 for percent of lines of first file n = 2 for percent of lines of second file

- CD compare directories (CodeDiff only)
 n = 0 to not compare directories (default)
 n = 1 to compare directories
- PTn n = minimum percent to report (CodeDiff only, default = 0)
- RMn report matching lines in the database (CodeDiff only)
 n = 0 to report lines in both files that do not match (default)
 n = 1 to report lines in both files that do match

To filter a CodeSuite database, type

```
cscl -PG10 dbase_in filter dbase_out
```

where:

- dbase_in name of the input database file to filter
- filter name of the filter file
- dbase_out name of the output filtered database file

To create HTML reports from a CodeSuite database, type

```
cscl -PGn dbase report
```

where:

- dbase name of the database file
- report name of the HTML report file to generate

options:

- PGn format of HTML report to generate
 n = 20 for basic report and detailed reports
 n = 21 for basic report only
 n = 22 for basic report for online use

To create spreadsheets from a CodeSuite database, type

```
cscl -PGn [-option] [-option] dbase spreadsheet
```

where:

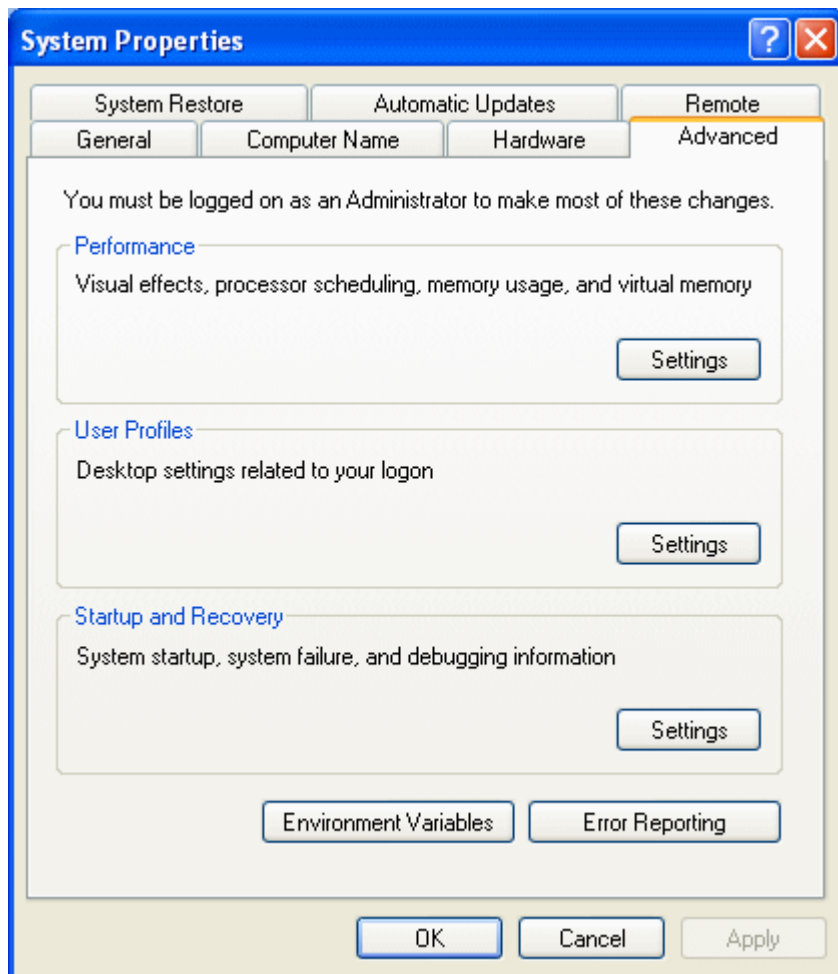
- dbase name of the input database file to filter
- spreadsheet name of the spreadsheet file(s) to generate

options:

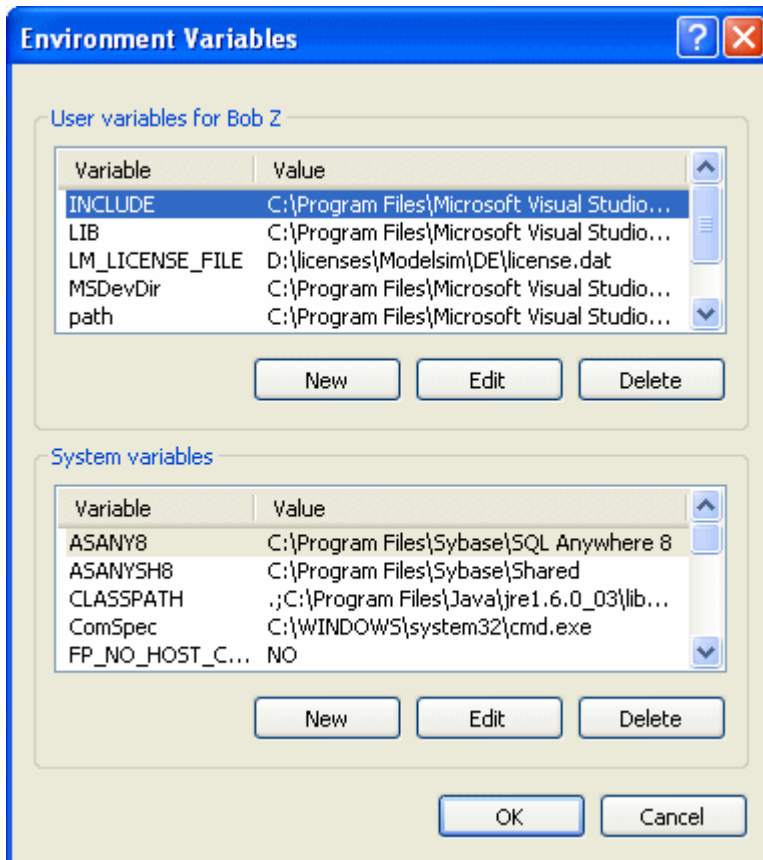
- PGn which spreadsheet(s) to generate (must be the first option)
 n = 30 for distribution spreadsheet
 n = 31 for search spreadsheets
 n = 32 for summary spreadsheet
- INn interval (summary spreadsheet only, default = 10)
- CSn consider statements (search spreadsheet only)
 n = 0 for no consideration
 n = 1 for consideration (default)
- CCn consider comments (search spreadsheet only)
 n = 0 for no consideration
 n = 1 for consideration (default)
- CIn consider identifiers or words (search spreadsheet only)
 n = 0 for no consideration
 n = 1 for consideration (default)

Setting up the PATH environment

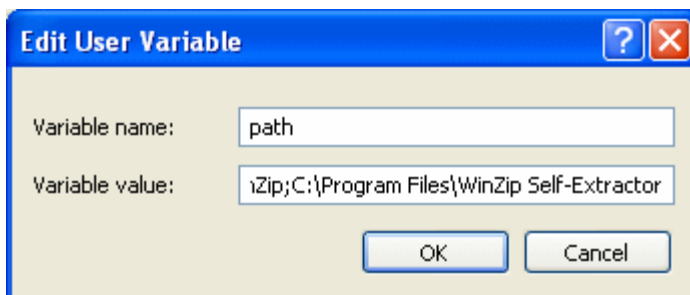
On your computer open Control Panel->Performance and Maintenance->System or right-click on My Computer and choose "Properties". In the box that opens, click the "Advanced" tab to bring up the dialog box shown below.



Next, click the "Environment Variables" button in the lower left corner to bring up the dialog box shown below.



In the top window, highlight the path variable and click on the "Edit" button below the window to bring up the dialog box shown below.



At the end of the path value, type a semicolon followed by the full path to CodeSuite on your computer. If you installed to the default directory, the path is "C:\Program Files\SAFE\CodeSuite" (without the quotes). Click "OK" on this dialog box and the previous two dialog boxes and you are done.

Contacting SAFE Corporation

Contacting SAFE Corporation



Software Analysis and Forensic Engineering Corporation
20863 Stevens Creek Blvd.
Suite 456
Cupertino, CA 95014
www.SAFE-Corp.biz

Tel. (408) 517-1194
Fax (408) 741-5231
Email: Support@SAFE-Corp.biz

Index

A		
ABAP	91	
ActionScript	91	
ASM-M68k	91	
Authorization Key	8	
B		
BASIC	91	
BitMatch	10, 17	
BitMatch Algorithms	19	
BitMatch Basic Report	20	
BitMatch Detailed Report	22	
C		
C 91		
C#	91	
C++	91	
Case	41, 44	
CLOC	27	
CLOC spreadsheet	79	
CodeCLOC	10, 25	
CodeCLOC Algorithm	27	
CodeCLOC Basic Report	29	
CodeCLOC Detailed Report	32	
CodeCross	10, 33, 35	
CodeCross Algorithm	35	
CodeCross Basic Report	36	
CodeCross Detailed Report	38	
CodeCross Score	35	
CodeDiff	10, 41, 44	
CodeDiff Algorithm	44	
CodeDiff Basic Report	45	
CodeDiff Detailed Report	47	
CodeGrid	17, 33, 41, 49	
CodeMatch	10, 49, 52	
CodeMatch Algorithms	49, 52	
CodeMatch Basic Report	54	
CodeMatch Detailed Report	57	
CodeMatch Optimization	49	
CodeSuite Database	41, 49, 71, 78	
CodeSuite database format	93	
CodeSuite filter format	98	
CodeSuite-MP	17, 33, 41, 49	
Command line interface	100	
Comment/String Filters		71
Comment/String Matching		49, 52
Copyrights		5
Correlation Score		52
D		
Database		41, 49, 71, 78
Delphi		91
Distribution Spreadsheet		83
F		
File Filters		71
File Menu		10
FileCount		10, 61
FileIsolate		10, 63
Filters		10, 71
Flash		91
Folder Filters		71
Fortran		91
FoxPro		91
H		
Help Menu		10
Hit Filters		71
HTML reports		78
I		
Identifier Filters		71
Identifier Matching		49, 52
Ignore case		41
Ignore whitespace		41
Instruction Sequence Matching		49, 52
Internet		69
J		
Java		91
JavaScript		91
L		
Languages Enabled		8
Languages Supported		91
License Type		8
Megabyte based		8
Time based		8
Unlimited		8
Licenses		8
Allocated		8
Remaining		8

User's Guide

LISP	91	Similarity Score	44
LOC.....	27	SourceDetective	10, 69
M		SQL.....	91
MASM	91	Statement Filters	71
MATLAB.....	91	Statement Matching	49, 52
Menu	10	Statistics.....	10, 67
P		Status Bar	10
Pascal	91	Summary Spreadsheet.....	88
Patents.....	5	System Requirements	7
Percentage.....	44	T	
Percentage options	41	Threshold Filters	71
Percentage of file pair.....	41	Toolbar	10
Percentage of first file	41	Tools Menu	10
Percentage of second file	41	Trademarks	5
Perl.....	91	V	
PHP	91	Verilog	91
PowerBuilder.....	91	VHDL.....	91
Python	91	View Menu	10
R		Visual Basic.....	91
RealBasic	91	W	
Restart.....	10	Whitespace	41, 44, 52
Ruby	91	Wildcard	41, 49
S		Windows 2000.....	7
SAFE Corporation	5	Windows 7.....	7
Search Spreadsheet.....	86	Windows Vista.....	7
Sequence Filters	71	Windows XP	7