gCOMBINE 1.1

A graphical user interface to perform structure-based Comparative Binding Energy (COMBINE) analysis

USER GUIDE

Rubén Gil-Redondo, Javier Klett
Federico Gago and Antonio Morreale

June 2010
# Table of Contents

LICENSE .................................................................................................................. 4
ORIGINAL AUTHORS .......................................................................................... 5
CHANGES ........................................................................................................... 6
INTRODUCTION ................................................................................................. 7
CONVENTIONS USED IN THIS GUIDE ............................................................. 7
INSTALLATION .................................................................................................... 8
  REQUIREMENTS ............................................................................................. 8
  OBTAINING gCOMBINE ............................................................................... 8
  UNPACKING ................................................................................................. 8
  COMPILING THE SOURCES (OPTIONAL) .................................................. 9
  SET UP THE ENVIRONMENT VARIABLES ............................................. 9
  READY TO RUN! ......................................................................................... 10
BRIEF DESCRIPTION ......................................................................................... 11
  MAIN WINDOW ......................................................................................... 11
  RESULTS WINDOW .................................................................................... 12
  FILE FORMATS .......................................................................................... 15
    Standalone COMBINE Input File ................................................................. 15
    Complexes File .......................................................................................... 16
    Data Matrix (combine.dat) ........................................................................ 16
EXAMPLES ........................................................................................................... 17
SELECTED COMBINE REFERENCES ................................................................. 18
DISCLAIMER ....................................................................................................... 19
License

License agreement.

The gCOMBINE user automatically accepts the following terms when using the program:

1) Report errors and bugs in the program to the authors (Rubén Gil-Redondo, Javier Klett, Federico Gago, Antonio Morreale) through e-mail: amorreale@cbm.uam.es

2) Do not redistribute the program. Interested users should contact the authors directly.

3) This license is for scientific non-profit and non-commercial use only. Any other use of this software for other purposes, alone or integrated into another piece of software, requires the prior consent of the authors.

4) The authors do not assume any responsibility for the licensed software and its use. The authors will not assume any liability for damages occurring through the use of the licensed software; do not guarantee the suitability of the licensed software for any application; and do not provide support.

5) Acknowledge the use of the program in scientific publications. Scientific publications where gCOMBINE was used should cite the following references:


Original Authors

This software and user guide were originally developed in 2009 by

Rubén Gil-Redondo [1]
(rgil@cbm.uam.es)

Javier Klett [1]
(jklett@cbm.uam.es)

Federico Gago [2]
(federico.gago@uah.es)

Antonio Morreale *[1]
(amorreale@cbm.uam.es)


* Corresponding author: Antonio Morreale. Unidad de Bioinformática. Centro de Biología Molecular Severo Ochoa. c/ Nicolás Cabrera 1. Campus de Cantoblanco. Madrid 28049. Spain. E-mail address: amorreale@cbm.uam.es. Tel.: 34-911964633. Fax: 34-911964420.

This work is dedicated to the memory of Ángel R. Ortiz.
Changes

Current version: 1.1 (20090707)

- Correct "ellipse" by "ellipse"
- Check the existence of combine.dat file when reading interaction matrix from file
- Solved bug when reading interaction file: energy_values.dat file was not re-generated
Introduction

gCOMBINE is a Java-written graphical user interface (GUI) for performing COMparative BINding Energy (COMBINE) analysis on a set of ligand-receptor complexes with the aim of deriving highly informative quantitative structure-activity relationships. The essence of the method is to decompose the ligand-receptor interaction energies in a series of terms, explore the origins of the variance within the set using Principal Component Analysis (PCA), and then assign weights to selected ligand-residue interactions using Partial Least Squares (PLS) analysis. The GUI allows plenty of interactivity and provides multiple plots representing the energy descriptors entering the analysis, scores, loadings, experimental vs. predicted regression lines, and the evolution of parameters such as $r^2$ (correlation coefficient), $q^2$ (cross-validated $r^2$), standard deviation of the error in the predictions and absolute average error as the number of extracted Latent Variables (LV) increases. Other representative features include a sigmoidal dielectric function for electrostatic energy calculations, alternative crossvalidation options (leave-N-out and random groups), definition of the applicability domain (through the drawing of a confidence ellipse), and the possibility to carry out several tasks that were formerly performed externally, such as variable removal, optional truncation of positive interaction energy values, and generation of ready-to-use PDB files containing information related to the importance for activity of individual protein residues that can be displayed and color-coded using a standard molecular graphics program such as PyMOL. It is expected that this user-friendly tool will expand the applicability of the COMBINE analysis method and encourage more groups to use it in their drug design research programs.

This release includes the gCOMBINE software, which is composed mainly of the command-line COMBINE program, its graphical user interface based on Java Swing and the needed external libraries. All the requirements and installation instructions are to be found in the "Installation" section of this guide.

This guide starts with an overview of gCOMBINE (current section), and then goes on with the instructions for obtaining, installing and configuring a working gCOMBINE system. Finally, two tutorials provide the basics for developing a 3D-QSAR analysis project using gCOMBINE.

You are invited to ask the gCOMBINE team for any problem you may find during the installation or usage of this software. Though gCOMBINE is provided as-is, the gCOMBINE team will do as much as possible to help users handle the software. If you find any error in the gCOMBINE software (including the documentation), you are encouraged to notify the gCOMBINE team about it, so that it can be corrected in future releases and notified to current users. You will find the gCOMBINE contact information at the end of this "Introduction".

Currently, gCOMBINE has only been tested on Linux x86 platforms (either 32-bit or 64-bit) and Windows XP systems; however, since the source code is freely distributed, you could probably compile it to run in other Unix-like systems and other Windows versions.

Conventions used in this guide

The following typographical conventions apply along this guide:

- **Italic**
  - File and directory names/extensions, URLs, executable files, command options, and emphasis.
- **Constant width**
  - Configuration options, environment variables, file contents, and sample code, user names.
- **Constant width bold**
  - Commands to be typed verbatim on the screen. Computer names.
- **Constant width italic**
  - Replaceable content in code and command-line information.
Installation

The list of software and hardware requirements for running a gCOMBINE system appears below. An overview of the installation process follows. Then, this "Installation" section shows you how to obtain the gCOMBINE distribution and install it.

Requirements

For execution:
- Linux operating system running on x86.32bit- or x86.64bit- processors. gCOMBINE has specifically been tested on Fedora Core 1/x86.32bit and CentOS 5/x86.64bit machines. It would probably also run on other Unix systems. It also runs on Windows XP systems through the cygwin1.dll library
- Java(TM) SE Runtime Environment 1.6.0 or newer

For compilation:
- GNU Make
- Bourne-Again shell (bash), C shell (csh), and TENEX C shell (tcsh) interpreters
- GCC 3.3 or newer / GNU Fortran (gfortran or g77) 3.3 or newer
- Cygwin (on Windows systems)
- Java(TM) SE Development Kit 1.6.0 or newer

Obtaining gCOMBINE

The software is freely distributed for academic and research purposes. The gCOMBINE.zip file can be downloaded from http://ub.cbm.uam.es/gCOMBINE.

Unpacking

Unzip gCOMBINE.zip files into a folder, i.e. /usr/local/gCOMBINE. From now on, we will refer to this directory with the environment variable $GCOMBINE_HOME (in this case, it will point to /usr/local/gCOMBINE). Under $GCOMBINE_HOME you can see the distribution directory structure:

```
bin: folder for executable files and libraries
    | gCOMBINE.jar: the gCOMBINE GUI program
    | lib: folder for needed external libraries
    |     | AbsoluteLayout.jar: a Netbeans useful library
    |     | appframework-1.0.3.jar: Swing Application Framework
    |     | cygwin1.dll: Cygwin main library
    |     | jcommon-1.0.14.jar: JCommon
    |     | jfreechart-1.0.11.jar: JFreeChart
    |     | swing-worker-1.1.jar: SwingWorker
    | linux: folder for linux specific executables
    |     | combine.exe: standalone COMBINE executable for Linux system
    |     | gCOMBINE.csh: script that starts the gCOMBINE in a Linux system
    | windows: folder for windows specific executables
    |     | combine.exe: standalone COMBINE executable for Windows system
    |     | gCOMBINE.bat: script that starts the gCOMBINE in a Linux system

docs: documentation folder
examples: files with some COMBINE examples
src: source code folder
    | combine: COMBINE Fortran source code folder
    | gui: COMBINE Java GUI source code folder
```

Note that the gCOMBINE program is composed of two subprograms:
- combine.exe: this is the command-line COMBINE program
- gCOMBINE.jar: this is the graphical user interface (GUI) for combine.exe
The external libraries are covered by the GNU General Public License (GPL) in the case of:

- cygwin1.dll (http://www.cygwin.com)

and by the GNU Lesser General Public Licence (LGPL) in the cases of

- appframework-1.0.3.jar (https://appframework.dev.java.net)
- jcommon-1.0.14.jar (http://www.jfree.org/jcommon)
- jfreechart-1.0.11.jar (http://www.jfree.org/jfreechart)
- swing-worker-1.1.jar (https://swingworker.dev.java.net).

**Compiling the sources (optional)**

The gCOMBINE package contains executable files for Linux (i686 32 bits platform) and Windows (XP SP3 32 bits platform) systems. If these executables do not work for you or if you prefer to set up the compilation options for your machine, then you can go to the $GCOMBINE_HOME/src/combine folder and compile it after modifying the Makefile file if needed (FC variable to set the Fortran compiler and FFLAGS variable to select the compilation options). Then, you can compile simply by typing:

```
make
```

Then you should copy the combine.exe file that is generated to the suitable $GCOMBINE_HOME/bin subfolder.

In Windows systems the Cygwin software (http://www.cygwin.com) should be used to compile COMBINE. If you do this then it is better to update the cygwin1.dll library in $GCOMBINE_HOME/bin/lib by replacing it with the cygwin1.dll file from your own Cygwin distribution.

Since the GUI for combine.exe is written in Java language, there is no need to recompile because it is platform-independent. Afterwards, you have the source files into $GCOMBINE_HOME/src/gui if you want to compile it after a modification.

**Set up the environment variables**

In order to run gCOMBINE you need to set up two environment variables:

- GCOMBINE_HOME: absolute path for the $GCOMBINE_HOME folder
- COMBINE_EXE: absolute path to the combine.exe program for the system

The usual way to do this on Linux systems is:

```
setenv GCOMBINE_HOME absolute-path-gCOMBINE-folder
setenv COMBINE_EXE $GCOMBINE_HOME/bin/system-folder/combine.exe
```

where you should set the correct values for absolute-path-gCOMBINE-folder ($GCOMBINE_HOME) and system-folder (linux or windows). A common practice is to add the last two sentences into the .cshrc file (or .bashrc file in its correct syntax) in order to set the two environment variables for all new sessions.

If you are under a Windows systems, the way to set up the environmental variables is to go to “System Properties” (menu Properties into “My Computer” icon), Advanced tab and “Environmental Variables” button. There you can create the variables by clicking the “New” button and filling in the two fields for each variable.

Note: if you are under a Windows systems and you plan to use the standalone COMBINE executable, then remember to add the absolute path to cygwin1.dll library in your PATH environment variable (this needed step is automatically done when you run gCOMBINE with gCOMBINE.bat).
**Ready to run!**

OK, now you are ready to launch gCOMBINE. Use the gCOMBINE.csh file if you are under a Linux system or the gCOMBINE.bat file if you are under a Windows system.
Brief Description

Main Window

1. From this menu you can clean the main window ("New Model"), load a model with their results ("Load Model") or save the current model with their results ("Save Model"). Take into account that only model parameters and results are saved, not the input files. From this menu you can also close the application ("Close").

2. From here you can see information about gCOMBINE.

3. This is the tab where the model parameters are configured.

4. The tab showing the model results after running COMBINE.

5. Location of the COMBINE executable.

6. Working folder for current model.

7. Run COMBINE button.

8. Optional name for the model.

9. Optional description for the model.


12. Y-randomization (“scrambling”) can be performed or not, by selecting yes/no in this combo box. (Note that including scrambling in the analysis can increase the run-time considerably).

13. Scaling of variables can be done either through the whole set of variables or by blocks ("Block Unscaled Weights").

14. The user can choose “Calculate” or “Read from file” from this combobox. This last option is useful when a model has already been built and the user wants to carry out a new analysis on the same set of compound but using different parameters.

15. The number of latent variables (from 5 to 10) can be chosen by the user.

16. The user can choose to perform either “Leave N out” or “Random groups” for the cross-validation test; it is also possible to choose the number of elements that are left out or included in each random group.
17. Different dielectric models can be selected by the user for the electrostatic energy computations: “Uniform dielectric constant”, “Goodford’s images”, “Distance dependent dielectric constant”, “Poisson-Boltzmann from .dph files” or “Sigmoidal”.
18. The value of the dielectric constant.
19. If the user has decided to incorporate some external variables, these can be added by first selecting their in this combobox and then introducing their values in the columns that appear in the table below (EV1, EV2...).
20. A data pre-treatment option by means of which the user can set to zero any positive van der Waals and/or electrostatic energy values.
21. By selecting a cut-off value on this box, variables with a standard deviation less than the value introduced by the user will be removed from the analysis.
22. A new line for a new complex will be included in the table.
23. The selected complex will be removed.
24. All compounds and their activities, etc... can be loaded from a text file. See “File Formats” section for details.
25. Once the user has introduced all compounds, the table can be saved in the previous format.
26. This is the table of compounds for the model. You can choose the Type ("Training", “Test”, or “Not used”), sort the columns (by clicking on its header) or give a name for the extra variables (also by clicking on its header).

**Results Window**

gCOMBINE provides a summary with the values for correlation coefficients on both training and test sets (if provided), errors between predicted and experimental values on training and test sets (if provided) and for each principal component. This data is presented in a table and is also graphically displayed in a plot through each principal component.

![Results Window](image)

Also predicted versus experimental values are plotted for each latent variable. The straight line (the diagonal) included in this plot as a reference helps the user observe the deviations between experimental and predicted values.
To understand the impact of the original variables on the predictive ability of the model, the PLS regression coefficients are plotted. They are separated into two different plots, the upper one corresponding to the van der Waals interactions and the lower one corresponding to the electrostatic interaction plus the additional variables if provided.

Plots with the values of the $i^{th}$ principal component versus loadings values of the $(i+1)^{th}$ principal component are provided. They show relations between variables, since variables lying together far from the origin denote possible correlations between them. This loading corresponds to the decomposition of the X-matrix. The relation between the original variables and the new orthogonal latent variables can be unveiled by plotting the contributions of the calculated energy descriptors to two successive principal components (loading plot).
Plots with score values of the $i^{\text{th}}$ principal component versus loadings values of the $(i+1)^{\text{th}}$ principal component are also provided. Likewise, the complexes can be plotted in the space defined by two successive latent variables (score plot). The values of the scores can be understood as the values of the compound into the new variable space, the principal component space, and since this new space is normalized with zero mean, compounds lying far from the origin have values significantly away from the mean, and can therefore be acting as outliers. The automatic generation of a confidence ellipse, as defined by Hotelling’s $T^2$ for the set of points, is included in the plot to aid in the identification of an approximate applicability domain.
The variables entering the analysis, i.e. the original energy values calculated in the force field plus EV-variables values, if provided, are plotted. This graph can help identify complexes in which modeling errors are giving rise to anomalous energy descriptors and also allow the user to visualize the amount of dispersion in a particular ligand-residue interaction within the set of complexes.

### File Formats

#### Standalone COMBINE Input File

<table>
<thead>
<tr>
<th>Line Number</th>
<th>Fortran 77 Format</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1           | (3i5)             | Y-randomization?: 0 – No, 1 – Yes  
Scaling?: 0 – No, 1 – Yes, 2 – Block Unscaled Weight  
Interaction Matrix?: 0 – Calculate, 1 – Read from file |
| 2           | (2i5,f10.5)       | Number of Latent Variables?: an integer number between 5 and 10  
Type of Dielectric Model?: 0 – Uniform dielectric constant, 1 – Goodford’s images method, 2 – Distance-dependent dielectric constant, 3 – Poisson-Boltzmann from .dph files, 4 – Sigmoidal  
Dielectric Constant?: a real number greater than or equal to 0.0 (kcal/mol) |
| 3           | (2i5)             | Use external variables?: 0 – No, 1 – Yes  
Number of External Variables?: an integer number between 0 and 10 |
| 4           | (2i5)             | Number of training complexes?: an integer value greater than 0  
Number of test complexes?: an integer value greater than or equal to 0 |
| 5           | (i5,f10.5)        | Zeroing positive values?: 0 – None, 1 – Van der Waals, 2 – Electrostatics, 3 – All  
Std. dev. cutoff?: a real value greater than or equal to 0.0 (kcal/mol) |
| 6           | (2i5)             | Validation Method?: 0 – Leave N Out, 1 – Random Groups  
Number of validation elements?: an integer number greater than 0 |
“Number of training complexes” lines
(a8,3x,a4,11f10.5)
- File name (without extension) for the training complex
- Ligand residue name*
- Pharmacological activity
The last 10 columns are for each extra variable

“Number of test complexes” lines
(a8,3x,a4,11f10.5)
- File name (without extension) for the test complex
- Ligand residue name*
- Pharmacological activity
The last 10 columns are for each extra variable

Rest of lines
(a8,3x,a4,11f10.5)
Same format as training/test complexes but for complexes not to be used in model derivation (these lines will be ignored by COMBINE)

* PAR if the ligand is made up of more than one residue. Then gCOMBINE automatically recognizes these residues as belonging to the ligand because (1) they do not belong to the protein (i.e. their residue names are not amino acids), (2) they are not water molecules (WAT or HOH residue names), and (3) they are not found in the built-in database of common cofactors and prosthetic groups.

Complexes File
This type of file is used by gCOMBINE when the user wants to load/save the list of complexes. All the lines have the same format; values for each line are separated by blank spaces (1 or more). The meaning for each column is:

<table>
<thead>
<tr>
<th>Column</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 – Training complex, 1 – Test complex, 2 – Complex not used</td>
</tr>
<tr>
<td>2</td>
<td>File name (without extension) for the complex (up to 8 chars)</td>
</tr>
<tr>
<td>3</td>
<td>Pharmacological activity (a real number)</td>
</tr>
<tr>
<td>4-13</td>
<td>Values (real numbers) for the extra variables (is used)</td>
</tr>
</tbody>
</table>

Data Matrix (combine.dat)

<table>
<thead>
<tr>
<th>Line Number</th>
<th>Fortran 77 Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(a)</td>
<td>Brief description of the set of complexes to be studied</td>
</tr>
<tr>
<td>2</td>
<td>(i5)</td>
<td>Number of variables + number of ligand pieces in the .top file</td>
</tr>
<tr>
<td>3</td>
<td>(i5)</td>
<td>Number of ligands in the training set + number of ligands in the test set</td>
</tr>
<tr>
<td>4</td>
<td>(i5)</td>
<td>Ligand number</td>
</tr>
<tr>
<td>5</td>
<td>(a8)</td>
<td>Ligand name</td>
</tr>
<tr>
<td>“number of residues” lines</td>
<td>(f12.6,3x,a4,i4)</td>
<td>Van der Waals energy values for residue i</td>
</tr>
<tr>
<td>“number of residues” lines</td>
<td>(f12.6,3x,a4,i4)</td>
<td>Electrostatic energy values for residue i</td>
</tr>
<tr>
<td>“number of external variables” lines</td>
<td>(f12.6,3x,a2,i2)</td>
<td>Value for extra variable i</td>
</tr>
<tr>
<td>one line</td>
<td>(f12.6,3x,a4)</td>
<td>Value for experimental activity</td>
</tr>
</tbody>
</table>

Repeat from line 4 for each compound
Examples

Two gCOMBINE examples are provided in the \$GCOMBINE_HOME/examples subfolder. They are called example1 and example2. Each one contains a set of ".top" and ".crd" files for each complex, and a parameters file (combine.in) that can be loaded for a quick setup of model parameters. In order to run one of the examples, follow the next steps:

1) Start gCOMBINE.
2) Click on the “Select” button near the “Working Folder” box to select the examples folder you want to use (\$GCOMBINE_HOME/examples/example1 or \$GCOMBINE_HOME/examples/example2).
3) A “Name” and a “Description” for the model can be given.
4) Load the parameters file combine.in by clicking the “Load Parameters” button.
5) Press the “RUN COMBINE” button and wait for a few minutes until the results for this model are produced and displayed.
Selected COMBINE References

Ortiz, A.R.; Pisabarro, M.T.; Gago, F. and Wade, R.C.
Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis

Ortiz, A.R.; Pisabarro, M.T.; Gago, F. and Wade, R.C.
Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis: Application to Human Synovial Fluid Phospholipase A2 Inhibitors

Wade, R.C., Ortiz, A.R. and Gago, F.
Comparative Binding Energy Analysis

Perez, C.; Pastor, M.; Ortiz, A.R.; Gago, F.
Comparative binding energy (COMBINE) analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design

Lozano, J.J.; Pastor, M.; Cruciani, G.; Gaedt, K.; Centeno, N.B.; Gago, F.; Sanz, F.
3D-QSAR methods on the basis of ligand-receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands

Tomic S; Nilsson L.; Wade R.C.
Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis

Wang T; Wade R.C.
Comparative Binding Energy (COMBINE) Analysis of Influenza Neuraminidase-Inhibitor Complexes

Rodriguez-Barrios F, Gago F.
Chemometrical Identification of Mutations in HIV-1 Reverse Transcriptase Conferring Resistance or Enhanced Sensitivity to Arylsulfonylbenzonitriltes.

Gil-Redondo, R.; Klett, J; Gago, F. and Morreale, A.
gCOMBINE: A graphical user interface to perform structure-based Comparative Binding Energy (COMBINE) analysis
Proteins (2009) (submitted)
Disclaimer

This software is copyrighted by Rubén Gil-Redondo, Javier Klett, Federico Gago and Antonio Morreale. The following terms apply to all files and documents associated with the software unless explicitly disclaimed in individual files. Do not redistribute the program. Interested users should contact directly to the authors. gCOMBINE is for scientific non-profit and non-commercial use only. Any other use of this software for other purposes, alone or integrated into other software, requires the prior consent of the authors.

IN NO EVENT SHALL THE AUTHORS OR DISTRIBUTORS BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OF THIS SOFTWARE, ITS DOCUMENTATION, OR ANY DERIVATIVES THEREOF, EVEN IF THE AUTHORS HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

THE AUTHORS AND DISTRIBUTORS SPECIFICALLY DISCLAIM ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. THIS SOFTWARE IS PROVIDED ON AN "AS IS" BASIS, AND THE AUTHORS AND DISTRIBUTORS HAVE NO OBLIGATION TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.