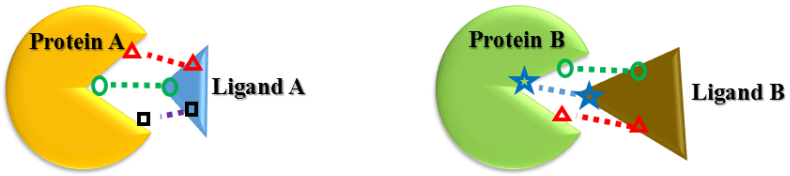
User's Manual of KGS2

Contents

Introduction

The KGS2 program is developed by Dr. Jie Liu in Dr. Renxiao Wang's group at the Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences.

KGS2 is a software patch of scoring functions, which has its objective to improve the prediction accuracy of scoring functions. Our basic assumption is that molecular systems with similar structures have similar properties, a strategy that has been applied successfully to the computation of some physicochemical properties such as partition coefficient and water solubility. Accordingly, the unknown binding affinity of a given complex can be estimated more reliably from the known binding affinity of a reference complex, which shares a similar pattern of protein-ligand **interactions** with the query complex.



known experimental binding affinity **unknown** experimental binding affinity

Figure 1. The query complex (B) and reference complex (A) share a similar pattern of protein-ligand interactions (different shapes of black marks represent different types of protein-ligand interactions)

The binding scores provided by a reasonable scoring function should correlate well with experimentally determined binding data as follows:

$$\hat{R}_{bind} = b + k \times R_{score,SF} \quad (1)$$

Here, $\hat{R}_{bind}$ denotes for the expected binding affinity of a reference protein-ligand complex (R); $R_{score,SF}$denotes for the binding score of this complex calculated by a scoring function SF; while b and$k$, respectively, are the intercept and the slope of the regression line between the binding scores and experimentally measured bindingdata of a set of protein-ligand complexes. Similarly, the expected binding affinity of a query protein-ligand complex (Q) calculated by the same scoring function is:

$$\hat{Q}_{bind} = b + k \times Q_{score,SF} \quad (2)$$

By subtracting Equation 1 from Equation 2, one has:

$$\hat{Q}_{bind} = \hat{R}_{bind} + k \times \left(Q_{score,SF} - R_{score,SF}\right) (3)$$

Replacing the expected binding affinity of $R$with the known experimental value$\left(R_{exp}\right)$, one has:

$$\hat{Q}_{bind} = R_{exp} + k \times \left(Q_{score,SF} - R_{score,SF}\right) (4)$$

Equation 4 indicates how the binding affinity of a given protein-ligand complex is computed using the known binding affinity of a proper reference complex as a starting point.

For the convenience of narration, this scoring strategy will be referred to as the KGS2 throughout this article. In principle, any scoring method may be employed to compute the required binding scores of both the reference complex and the query complex in Equation 4. Nevertheless, it is certainly more reasonable in reality to choose a capable scoring method for this purpose. The reference complex can be selected among a database of protein-ligand complexes with reliable structures and binding data. The constant k in Equation 4 can be derivedthrough a regression analysis between the experimentalbinding data and the computed binding scores by theemployed scoring method on the same database. It is introduced to scale the outcomes of scoring functions, which could be in arbitrary units, to a realistic range comparable to the experimental binding data of the reference complex.

KGS2 is distributed freely to the public. It is currently available at our group website: http://www.sioc-ccbg.ac.cn/software/KGS2. Basically, you need to register and sign a license agreement. We will then send you further instructions of how to download this program.

You may direct questions related to this program to the author at: Renxiao Wang, Ph.D.

How to use KGS2

The KGS2 program is written in ANSI C++ language and has been tested on LINUX platform. After downloading the program package, please move it to the directory where you would like the program to be ran. Then, use the program through the following three-step procedure.

Uncompress the package

You can do this in a Linux shell as:
tar -xvf KGS2.tar

You will get a directory named as "KGS2" under your working directory. Under that directory, there are several subdirectories:

"bin/": executable binary code
"example/": examples for practicing KGS2
"manual/": user manual in doc
"parameter/": parameter files
"reference_complex/": the general-set v2014 used for using KGS2, 10605 protein-ligand complexes in total

General synopsis for running KGS2

The basic function of KGS2 is to search the reference complexes of a query complex from the specified data set. All of the parameters needed to run KGS2 are assembled in an input

parameter file. You are supposed to edit this file to meet your own purpose.

```
#########################################################################
#                              KGS2/SIMILARITY                          #
#########################################################################
###
FUNCTION        SIMILARITY
###
### specify the input and output files ------------------------------
###
#
QUERY_RECEPTOR_PDB_FILE    ../reference_complex/1a42_protein.pdb
QUERY_LIGAND_MOL2_FILE     ../reference_complex/1a42_ligand.mol2
#
OUTPUT_TABLE_FILE    ./similarity.table
###
### specify the path of the data set for complexes' similarity search
###
REFERENCE_COMPLEXES_DIRECTORY       ../reference_complex
###
```

Figure 2. The input parameter file of KGS2

To run KGS2, simply use this file as input:

KGS2input_parameter_file

The very first parameter in the input parameter file is FUNCTION, which should be set as "SIMILARITY". This tells the program to perform KGS2 computation. You are not supposed to change this. Other parameters specified in the input file will be explained below. You can find an example input file under the "bin/" directory.

Note that all of the lines started with a "#" sign in the input file will be considered as a comment line and is neglected by the program.

Parameters for setting input files

The first few parameters in the parameter file defines the input structural files. KGS2 needs the three-dimensional structure of the given protein-ligand complex to calculate its reference complexes. The structure could be either experimentally determined or modeled by a docking program. Since most today's molecular docking programs keep the protein structure rigid while docking the ligand molecules, for the sake of efficient computation, KGS2 requires the protein and the ligand molecules to be stored in two separate files. The protein is required to be stored in a PDB file, and the ligand molecules should be stored in a Mol2 file.

The parameter QUERY_RECEPTOR_PDB_FILE specifies the path and the name of the PDB file that stores the query protein molecule. To prepare this PDB file: (1) Remove any ligand molecule or other organic cofactors. (2) If a metal ion resides in the binding site and is believed to be important for ligand binding, keep it as part of the protein. KGS2 doesn’t consider any kind of metal ions in computation. According to PDB convention, a metal ion should be described by a line started by "HETATM". (3) For water molecules, you may keep them in the PDB file (also in the HETATM section). However, KGS2 will not consider them in computation. (4) Remember to add hydrogen atoms. KGS2 only needs polar hydrogen atoms on the protein in computation. But adding all of the hydrogen atoms (polar and non-polar) will not hurt.

The parameter QUERY_LIGAND_MOL2_FILE specifies the path and the name of the Mol2 file that stores the structures of the query ligand molecules to be searched. One thing should be kept in mind is: the ligand molecules must be pre-docked into the binding pocket of the query protein. KGS2 will not do docking for you --- it only searches the reference protein-ligand complexes of given protein-ligand complexes. Also, please make sure that the docked ligand molecules are saved in the same coordinate system as the protein molecule.

Since the Mol2 format is defined by Tripos, naturally we recommend SYBYL for preparing all Mol2 files. Other molecular modeling software may support the Mol2 format as well. There are also some programs, such as Babel, which are designed for converting different formats. However, our experience is that such conversion is not always flawless. To prepare the ligand molecules correctly: (1) Please try your best to correctly set the atom types and bond types according to the Tripos conventions. (2) All hydrogen atoms (polar and non-polar) need to be added to the ligand molecules. (3) Atomic charges are not necessary for KGS2 computation.

The parameter REFERENCE_COMPLEXES_DIRECTORY specifies the path of protein-ligand complexes to be searched. The KGS2 program would compute similarity between the query complex and each complex from this directory. In this directory, a file of complexes' list.

```
8.24    1ydb_protein.pdb        1ydb_ligand.mol2
8.06    3dbu_protein.pdb        3dbu_ligand.mol2
8.64    1g45_protein.pdb        1g45_ligand.mol2
8.05    3oys_protein.pdb        3oys_ligand.mol2
7.99    3oyq_protein.pdb        3oyq_ligand.mol2
6.40    2nno_protein.pdb        2nno_ligand.mol2
7.20    3k2f_protein.pdb        3k2f_ligand.mol2
7.30    3s8x_protein.pdb        3s8x_ligand.mol2
6.60    1kwr_protein.pdb        1kwr_ligand.mol2
6.74    3oim_protein.pdb        3oim_ligand.mol2
7.37    1cnx_protein.pdb        1cnx_ligand.mol2
```

Figure 3. The list file of reference complex

The 1st column: the experimental binding affinity of reference complex;
The 2st column: the protein file of reference complex;
The 3st column: the ligand file of reference complex.

All of the residues within 10 angstrom from any part of the ligand molecule are defined as pocket residues and will be considered in complexes' similarity calculation.

In any case, KGS2 will use the ligand molecule saved in the QUERY_LIGAND_MOL2_FILE for defining binding pocket. Of course the downside of this approach is that it may not define the binding pocket as precisely as the one using a proper reference molecule.

Parameters for setting output files

The next parameter defines the output file that stores the results of KGS2 computation. The basic output is defined by the OUTPUT_TABLE_FILE parameter, which specifies a file tabulating the results of each calculated complexes' similarity result. Every following line contains the information of a pair of complexes. The meaning of each column is:

1	1a42_protein.pdb	39	0.951219	1bnq_protein.pdb
2	1a42_protein.pdb	32	0.727273	1cil_protein.pdb
3	1a42_protein.pdb	29	0.659091	1cin_protein.pdb
4	1a42_protein.pdb	29	0.644444	1i90_protein.pdb
5	1a42_protein.pdb	28	0.608696	1bnu_protein.pdb
6	1a42_protein.pdb	29	0.591837	1bnt_protein.pdb
7	1a42_protein.pdb	29	0.568627	1bnm_protein.pdb
8	1a42_protein.pdb	29	0.568627	1bnv_protein.pdb
9	1a42_protein.pdb	25	0.568182	1cim_protein.pdb
10	1a42_protein.pdb	26	0.520000	1bnn_protein.pdb
11	1a42_protein.pdb	30	0.517241	1i8z_protein.pdb

Figure 4. The output file of KGS2

The 1st column: rank of the reference complex. All the reference complexes are ranked in a decreasing order by their similarity score;

The 2nd column: the name of query complex;

The 3rd column: match number;

The 4th column: similarity score (Tanimoto coefficient 0~1);

The 5th column: the name of reference complex;

This table is a standard space-parsed text file, you can use any spreadsheet program, such as Excel, to load this table.

The shortcut to run KGS2

The standard way for running KGS2, which has been described above, is suitable for computing multiple reference complexes against a given query complex. Sometimes the user just wants to compare one protein-ligand complex against its query protein-ligand complex and get a quick feedback of their similarity score. KGS2 provides a shortcut for this purpose (the synopsis is a little different from the previous versions of KGS2: a flag of "-similarity" is required now):

KGS2 –similarity the_query_protein_PDB_file the_query_ligand_Mol2_file the_reference_protein_PDB_file the_ reference_ligand_Mol2_file

The results will be printed on the screen. KGS2 will not create the file named by the OUTPUT_TABLE_FILE parameter.

References

[1] Kota, Kasahara.;Matsuyuki, Shirota.; Kengo, Kinoshita. Comprehensive Classification and Diversity Assessment of Atomic Contacts in Protein−Small Ligand Interactions. J. Chem. Inf. Model., 2013, 53, 241–248

[2] Tiejun, Cheng.;Zhihai Liu,; Renxiao Wang. A knowledge-guided strategy for improving the accuracy of scoring functions in binding affinity prediction. BMC Bioinformatics 2010, 11, 193

[3] Renxiao, Wang.;Luhua, Lai.; Shaomeng, Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. Journal of Computer-Aided Molecular Design, 2002, 16, 11–26

[4] Stefano, Forli.; Arthur J. Olson. A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking. J. Med. Chem. 2012, 55, 623−638

[5] Howard, J.F.; Paul, L. Pocket Similarty: Are Carbons Enough? J. Chem. Inf. Model. 2010, 50, 1466-1475.