BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 24 2012, pages 3334–3335
doi:10.1093/bioinformatics/bts625

Data and text mining

Advance Access publication October 18, 2012

CDRUG: a web server for predicting anticancer activity of chemical compounds

Gong-Hua Li[1] and Jing-Fei Huang[1,2,*]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences and [2]Kunming Institute of Zoology-Chinese University of Hongkong Joint Research Center for Bio-resources and Human Disease Mechanisms, Kunming, Yunnan 650223, China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Cancer is the leading cause of death worldwide. Screening anticancer candidates from tens of millions of chemical compounds is expensive and time-consuming. A rapid and user-friendly web server, known as CDRUG, is described here to predict the anticancer activity of chemical compounds. In CDRUG, a hybrid score was developed to measure the similarity of different compounds. The performance analysis shows that CDRUG has the area under curve of 0.878, indicating that CDRUG is effective to distinguish active and inactive compounds.

Availability: The CDRUG web server and the standard-alone version are freely available at http://bsb.kiz.ac.cn/CDRUG/.

Contact: huangjf@mail.kiz.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 16, 2012; revised on September 26, 2012; accepted on October 14, 2012

1 INTRODUCTION

Cancer causes millions of deaths per year. Screening candidates from natural or synthetic compounds is an efficient way for cancer drug discovery (Shoemaker, 2006). In the past decades, tens of millions of chemical compounds have been deposited in the public database (Wang *et al.*, 2012). Discovering anticancer compounds from this huge database through experimental methods is expensive and time-consuming (Chabner and Roberts, 2005). Although the National Cancer Institute (NCI) has been screening the anticancer compounds for tens of years, only 1% (~70 000) compounds were tested for screening anticancer compounds (Shoemaker, 2006). Therefore, a rapid and effectively computational method is required for prediction of anticancer activity of chemical compounds.

Here, we constructed a web server, termed Cancer Drug (CDRUG), to predict the anticancer activity of given compound(s). CDRUG uses a novel molecular description method (relative frequency-weighted fingerprint) to implement the compound 'fingerprints'. Then, a hybrid score was calculated to measure the similarity between the query and the active compounds. Finally, a confidence level (*P*-value) is calculated to predict whether the query compound(s) have, or do not have, the activity of anticancer.

*To whom correspondence should be addressed.

2 METHODS

To predict the anticancer activity of chemical compounds, we first constructed a benchmark dataset, which contains two subsets, active and inactive dataset. The active dataset consists of 8565 anticancer compounds, whereas the inactive dataset includes 9804 compounds. All of the datasets are from NCI-60 Developmental Therapeutics Program (DTP) project (Shoemaker, 2006). The detailed method of constructing benchmark dataset can be found in Part I of Supplementary Material.

Second, we used a novel molecular description method, termed as relative frequency-weighted fingerprint (*RFW_FP*), to calculate the compound 'fingerprints'. *RFW_FP* was calculated as follows:

$$RFW\_FP(i) = Bit(i) \times \left(\frac{F_{\text{active}}(i)}{F_{\text{inactive}}(i)}\right)^{\alpha} \quad (1)$$

where *i* represents *i*th Daylight pattern or fingerprint. In Daylight theory, each compound contains more than one and less than 1024 patterns or fingerprints. *RFW_FP(i)* is *i*th relative frequency-weighted fingerprint. *Bit(i)* is calculated by Pybel (O'Boyle *et al.*, 2008), a python wrapper of Openbabel (O'Boyle *et al.*, 2011); if the compound has *i*th fingerprint, $Bit(i) = 1$, else $Bit(i) = 0$. $F_{\text{active}}(i)$ and $F_{\text{inactive}}(i)$ are the frequency of *i*th fingerprint in the active and inactive compounds, respectively. α is the amplifying factor. In this study, α was optimized as 4.0 (Supplementary Fig. S2).

Third, the relative frequency-weighted Tanimoto coefficient (*RFW_TC*) between two compounds was calculated as follows:

$$RFW\_TC(m,n) = \frac{S_{mn}}{S_m + S_n + S_{mn}} \quad (2)$$

where *RFW_TC(m,n)* is the relative frequency-weighted Tanimoto coefficient between two compounds (*m* and *n*). S_m *and* S_n are the sum of relative frequency-weighted fingerprints in compound *m* and *n*, respectively. S_{mn} *is* the sum of the common relative frequency-weighted fingerprints between two compounds.

Fourth, a hybrid score (HSCORE), based on both the *RFW_TC* and the MinMax Kernel (*KMM*) (Swamidass *et al.*, 2005), was calculated to measure the similarity between two chemical compounds. *HSCORE* was calculated as follows:

$$HSCORE(m,n) = RFW\_TC(m,n) \times KMM(m,n) \quad (3)$$

where *HSCORE(m,n)* is the hybrid score between two compounds (m and n). *KMM(m,n)* is the MinMax kernel (Swamidass *et al.*, 2005) between two compounds and is calculated by jCompoundMapper (Hinselmann *et al.*, 2011), a Java API for molecular kernel theory (Swamidass *et al.*, 2005).

Finally, for each query chemical compounds, the maximum *HSCORE* between the query and the active dataset (8565 compounds) was calculated. Then the *P*-value, based on the maximum *HSCORE,* was calculated. Because the maximum *HSCORE* is not >1.0, and because the

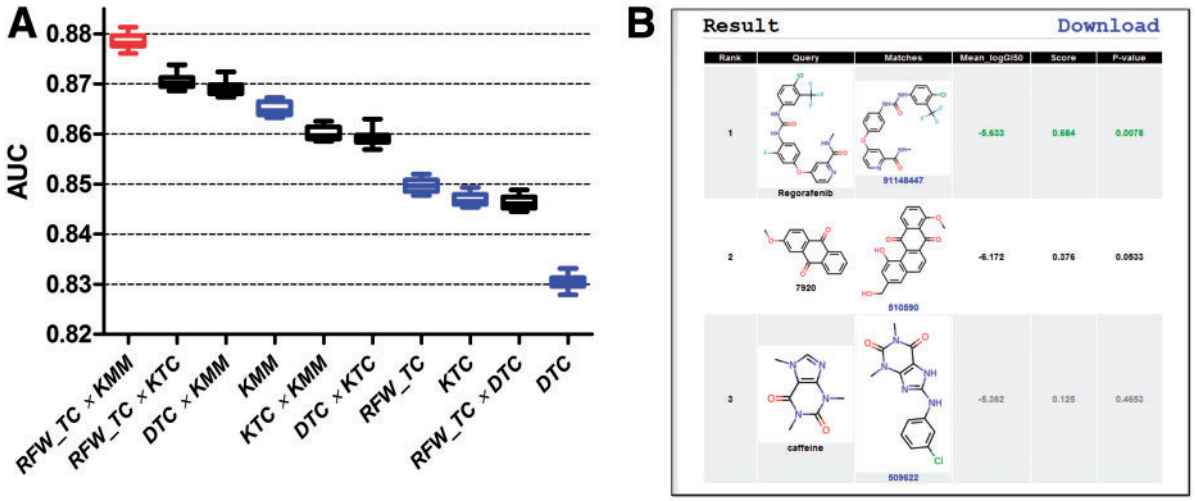

Fig. 1. Overview of CDRUG. (**A**) Comparison of different methods. Non-hybrid methods are coloured by blue. CDRUG (hybrid method of $RFW\_TC \times KMM$) is coloured by red. KMM = MinMax kernel; KTC = Tanimoto kernel; DTC = Daylight Tanimoto coefficient; RFW_TC = relative frequency-weighted Tanimoto coefficient. (**B**) The output page of CDRUG. Highly possible, possible and less possible results are coloured by green, black and grey, respectively

maximum *HSCORE* of the inactive compounds have a generalized extreme value distribution (Supplementary Fig. S3), we can calculate the *P*-value as follows:

$$p(x) = F(1.0; \mu, \sigma, \xi) - F(x; \mu, \sigma, \xi) \quad (4)$$

where $p(x)$ is the *P*-value at the maximum *HSCORE* of x; $F(x;\mu,\sigma,\xi)$ is the cumulative function of generalized extreme value distribution. Using the maximum likelihood method ('fgev' function in R 'evd' package), we estimated the location parameter μ of 0.094, the scale parameter σ of 0.063 and the shape parameter ξ of 0.302.

To implement the performance of CDRUG, 20 runs of 5-fold cross-validation method (Part II in Supplementary Material) were used to calculate the area under the curve (AUC). Four non-hybrid methods (including Daylight Tanimoto coefficient, *RFW_TC*, Tanimoto Kernel and MinMax kernel) and six hybrid methods were used to test the performance of the CDRUG (Part V of Supplementary Material).

3 DESCRIPTION OF CDRUG

CDRUG uses a novel weighted method (*RFW_FP*) to implement the molecular fingerprints, and it then uses a hybrid score to measure the compound similarity. The result shows that CDRUG outperforms other methods ($P < 10^{-13}$, *t*-test) (Fig. 1A and Supplementary Figs S4 and S5). When non-hybrid methods were used, MinMax kernel obtained the best performance with the AUC of 0.865. But when the hybrid methods were used, the AUC of CDRUG increased to 0.878. CDRUG can hit ~65% positive results at the false-positive rate of 0.05 (Supplementary Figs S4 and S5). These results indicated that the CDRUG is effective to predict anticancer activity of the chemical compounds.

CDRUG is rapid. It accepts one or more compounds to implement a prediction. A query with 1–20 compounds requires ~35 s, whereas a query with 1000 compounds only requires ~4 min. Therefore, CDRUG can be applied to both case-study and large-scale prediction.

CDRUG is user-friendly. The only requirement of CDRUG is the SMILE(s) (Weininger, 1988) of the query compound(s). The result contains the query compound(s), the matched compound(s), average GI50 value of the matched compound(s), the maximum *HSCORE* and the *P*-value. CDRUG also predicts whether the query compound(s) have or do not have the activity of anticancer. The results are categorized as highly possible, possible and less possible depending on the *P*-value, and they are coloured by green, black and grey, respectively (Fig. 1B).

4 CONCLUSION

CDRUG web server provides an effective, rapid and user-friendly interface to predict anticancer activity of chemical compounds.

Funding: National Basic Research Program of China (2009CB941300); the National Natural Science Foundation of China (31123005); Chinese Academy of Sciences (2007211311091 to J.F.H.).

Conflict of Interest: none declared.

REFERENCES

Chabner,B.A. and Roberts,T.G. (2005) Timeline—chemotherapy and the war on cancer. *Nat Rev. Cancer*, **5**, 65–72.

Hinselmann,G. *et al.* (2011) jCompoundMapper: an open source Java library and command-line tool for chemical fingerprints. *J. Cheminform.*, **3**, 3.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem Cent. J.*, **2**, 5.

Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat Rev. Cancer*, **6**, 813–823.

Swamidass,S.J. *et al.* (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, **21**, I359–I368.

Wang,Y.L. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.

Weininger,D. (1988) Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.