Comparative Molecular Binding Energy Analysis of HIV-1 Protease Inhibitors Using Genetic Algorithm-Based Partial Least Squares Method

Yen-Chih Chen[1], Jinn-Moon Yang[2], Chi-Hung Tsai[1], and Cheng-Yan Kao[1]

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{r91051, d90008, cykao}@csie.ntu.edu.tw

[2] Department of Biological Science and Technology & Institute of Bioinformatics,
National Chiao Tung University, Hsinchu, Taiwan
moon@cc.nctu.edu.tw

1 Introduction

Comparative molecular binding energy analysis (COMBINE) [1] is a helpful approach for estimation of binding affinity of congeneric ligands that bind to a common receptor. The essence of COMBINE is that the ligand-receptor interaction energies are decomposed into residue-based energy contributions, and then the partial least squares (PLS) analysis is applied to correlate energy features with biological activity. However, the predictive performance of PLS model drops with the increase of number of noisy variables. With regard to this problem genetic algorithm (GA) combined with PLS approach (GAPLS) [2] for feature selection has demonstrated the improvement on the prediction and interpretation of model. Therefore, the purpose of this paper is to derive a more accurate and more efficient GAPLS in COMBINE by introducing a number of successive refinements.

2 Methodology

The structure-activity data of forty-eight inhibitors of human immunodeficiency virus type I (HIV-1) protease analyze by Perez *et al.* [3] was used as a new data set for our methodology. According to COMBINE, the calculated ligand-receptor interaction energies in the refined complexes were partitioned on a per residue basis. The 48-by-396 data matrix was built with columns representing each of the interaction energy features and rows representing each inhibitor in the data set, and then only the 50 features with highest Mahalanobis distance are treated as significant features and retained to the further GAPLS analysis.

GAPLS is a sophisticated hybrid approach that combines GA as an efficient feature selection method with PLS as a robust statistical method. In accordance with the framework of GAPLS, a number of successive refinements which can be summarized as follows:

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 385–386, 2004.

(i) Weighted standard deviation error of predictions (*WSDEP*) is used as objective function to measure the internal predictability with respect to the selected features.

$$WSDEP = \sqrt{\frac{\sum\left(y_i - y_{pred,i}\right)^2}{N - lv - 1}}\left(\frac{100}{95}\right)^{lv} , \tag{1}$$

where y_i and $y_{pred,i}$ are the observed and predicted inhibitory activities belong to inhibitor *i*, *N* is the total number of samples and *lv* is the number of latent variables.

(ii) An extra bit *lv*, representing the number of latent variables, is appended to the original chromosome and expect GAPLS model to efficiently solve the problem of the optimum number of latent variables in PLS.

(iii) Post-ranking function is added to identify the most suitable chromosome with the least number of features and the best objective in the final population.

3 Results and Discussion

In order to demonstrate the performances of our model GAPLS and GAPLS$_{\text{exp}}$, using 32 and 48 inhibitors in the training set, we quoted the models C$_{\text{delphi}}$ and C$_{\text{expanded}}$ from Perez *et al.* [3]. All the statistics of the models with respect to the predictability are listed in Table 1. GAPLS and GAPLS$_{\text{exp}}$ provide not only a number of the most significant energy features but also the excellent predictability. Moreover, the selected amino acid residues are in a good agreement with the important binding sites in HIV-1 protease, and pinpoint the regions of significance in three-dimensional space where the actual ligand-receptor interactions involved.

Table 1. Comparison of the different regression models

Model	*Samples*	*Features*	*lv*	r^2	q^2	$SDEP_c$	$SDEP_{ex}$
C$_{\text{delphi}}$	32	47	2	0.90	0.73	0.69	0.59
C$_{\text{expanded}}$	48	54	2	0.91	0.81	0.66	-
GAPLS	32	16	1	0.87	0.87	0.48	0.49
GAPLS$_{\text{exp}}$	48	15	2	0.92	0.91	0.46	-

References

1. Ortiz, A.R., Pisabarro, M.T., Gago, F., Wade, R.C.: Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. J. Med. Chem. 38 (1995) 2681-2691
2. Hasegawa, K., Kimura, T., Funatsu, K.: GA Strategy for Variable Selection in QSAR Studies: Enhancement of Comparative Molecular Binding Energy Analysis by GA-Based PLS Method. Quant. Struct.-Act. Relat. 18 (1999) 262-272
3. Perez, C., Pastor, M., Ortiz, A.R., Gago, F.: Comparative Binding Energy Analysis of HIV-1 Protease Inhibitors: Incorporation of Solvent Effects and Validation as a Powerful Tool in Receptor-Based Drug Design. J. Med. Chem. 41 (1998) 836-852