Chapter 4

Analysis of Data Fusion in Virtual Drug Screening

The performance of various scoring functions in virtual screening is often in consistent across different systems. The inaccuracy, inadequately predicting the true binding affinity of a ligand for a receptor, of the scoring methods is probably major weakness for virtual screening. Combining multiple scoring functions, called consensus scoring, is a popular strategy and has been shown to improve the enrichment of true positive[38,39,48]. Here, we developed a novel consensus scoring for virtual screening from data fusion in Information Retrieve (IR) systems and analyzed advantages for this novel combination way for scoring methods.

4.1 Introduction to Data Fusion

Data fusion is the process of combining data and knowledge from different sources with the aim of maximizing the useful information content. It improves reliability or discriminable capability while offering the opportunity to minimize the data retained. It can be divided the application domain of data fusion into three overlapping regions according to their characteristics: pre-processing, data alignment and correlation, post-processing. Pre-processing is mainly about data reduction. Data alignment and correlation involve interpolation and spatial or temporal correlation. Post-processing consists of combination of mathematical data, knowledge and decision making. Many researchers have focused on specific methods applied to particular problems, or particular aspects of the architecture, such as extended kalman filtering, model based approaches, wavelet decomposition, artificial neural networks and fuzzy logic. The using of data fusion could be occurred in any of the stages above:

- Raw data fusion. e.g. multiple sensors where the raw data is robustly and redundantly merged or sensors are validated.
- Feature fusion where a characteristic is extracted before fusion occurs.

- Decision fusion where measured data with or without pre-processing is combined with other data or a priori knowledge.

Practical applications of data fusion have necessarily been those areas in which the required output of an analysis may not be measured directly. Recently, Hsu et al,[49] showed that under certain conditions, IR systems can be better fused by combining the rankings of the documents returned than by the combination of their scores. We applied this idea to combine the rankings on virtual screening and found that the combination of rankings in different scoring methods indeed improved the screening performance. The analysis of data fusions on three screening targets indicated that the screening accuracy usually improved most when we adopted the combination of GEMDOCK with pharmacological preferences and GOLD with recombinant GoldScore.

4.2 Data Sets Preparations

In our previous works[19,50] and the master thesis of Tsai-Wei Shen[51], we has been tested GEMDOCK on virtual screening for thymidine kinase[50], dihydrofolate reductase (unpublished data), and estrogen receptor[19]. Various scoring functions were developed for calculating the free energy of protein-ligand binding in various approaches, such as knowledge-based[2,3], empirical, physics-based[4], and solvent-based scoring functions[5]. These scoring methods were designed with considering different kinds of interactions of physics in different approaches and it is true that these scoring methods have been applied successfully to many drug discovery projects. Because of the inadequate understanding of the present physics embedded in the protein-ligand binding process, these scoring function are still inaccurate. In order to improve the accuracy of virtual screening, we developed a novel combination concept of consensus ranking score through the data fusion technique on ranking combination and evaluated this method via analyzing the difference of performances in virtual screening.

Herpes simplex virus types 1 and 2 (HSV-1 and HSV-2) could cause painful epithelial ulcers near the mouth, on the cornea and genitals, as well as fatal encephalitis. HSV-1 TK is the center of phosphorylation of nucleosides or nucleoside analogs such as acyclovir [52,53]. Many antiviral drugs attack the replication of the viral genome with nucleoside analogs. These analogs are

activated by phosphorylation with TK and prevent DNA synthesis by the introduction of a chain-terminating nucleoside at the 3' end of the growing DNA strand. Besides antiviral drugs, these analogs have been used in a virological study of TK mutations [54] and employed extensively in gene therapy for cancer [55,56]. Therefore virtual screening for TK to exploit novel lead compounds would be of considerable value in many fields. We used HSV-1 TK as the target protein with a testing set proposed by Bissantz et al.. It included ten known active ligands of TK and 990 randomly chosen non-active compounds from the ACD. When preparing the target protein, the atom coordinates for virtual screening were taken from the crystal structure of the TK complex with the ligand deoxythymidine (PDB entry: 1kim).

Estrogens such as 17β-estradiol are steroid hormones as key mediators of female reproductive glands and they also exert their actions on other systems. For example, estrogens contribute to the maintenance of bone tissue through a process involving bone resorption and bone formation[57]. Hormone replacement therapies have been used for the treatment of vasomotor symptoms related to the menopause and for prevention of osteoporosis [58,59]. Compounds mimic estrogen in some tissues while antagonizing its action in others are named selective estrogen receptor modulators (SERMs) [60]. Many SERMs such as tamoxifen and raloxifene, are currently on the market for the treatment of hormone-dependent breast cancer [61] and prevention and treatment of osteoporosis [62], respectively. But there are often several intolerable side effects such as benign and malignant lesions of the uterus when patients take the treatment with SERMs for a long term. Therefore, the search for proper SERMs among both existing and new drugs has been a challenging task in recent years [63,64]. We have applied GEMDOCK on virtual screening against ERα with a testing set proposed by Bissantz et al. It was composed of ten known antagonists of ERα, ten known agonists of ERα[65]and 990 randomly selected compounds from ACD (Available Chemicals Directory). When preparing the target proteins, the atom coordinates for virtual screening were taken from the crystal structure of ERα complex with the ligand 4-hydroxytamoxifen (PDB entry: 3ert) for screening antagonists and with the ligand 17β-estradiol (PDB entry: 1gwr) for screening agonists. The atom coordinates of each ligand were sequentially taken from the database.

Dihydrofolate reductase (DHFR) catalyzes the reduction of 7,8-dihydrofolate or folate to 5,6,7,8-tetrahydrofolate (THF) in an NADPH-dependent pathway. THF is an essential cofactor for other enzymes involving one-carbon-transfer reactions necessary for the biosynthesis of numerous amino acids and purines. THF also acts as a cofactor for thymidylate synthase, which

is responsible for the methylation of deoxyuridylate to thymidylate, a key component for synthesis of DNA. DHFR is found in cells of all living organisms, where it maintains the intracellular level of THF. Therefore, the inhibition of DHFR activity reduces the intracellular pool of THF resulting in inhibition of DNA synthesis and leading to cell death. Based on this mechanism, human DHFR (hDHFR) has become a major drug target in anticancer therapy. It is also a target for inhibition of bacterial, fungal, and protozoal DHFRs to treat human infectious diseases by many implicated microorganisms [66,67]. With the wide use of these antifolate drugs, the resistance of DHFRs in human or other microorganisms is widespread. Therefore, it is urgent to search for new targets or new effective inhibitors to deal with the problem [68,69]. We used hDHFR as the target protein for virtual screening with a testing set prepared by ourselves. The testing set was composed of ten known active ligands of hDHFR and 990 randomly selected compounds from the MDL Drug Data Report (MDDR). When preparing the target protein, the atom coordinates for virtual screening were taken from the crystal structure of the DHFR complex (PDB entry: 1hfr). The atom coordinates of each ligand were sequentially taken from the database.

In these virtual screening, GEMDOCK used settings as Table7 and the pharmacological preference of each target was mining from their known ligand sets with steps in chapter 3. And GOLD was in default library screening settings for these targets.

4.3 Methods of Data Fusion

It has been reported that fusion among different scoring methods would improve the performance and the performance would superior to individual ones. In this study, we attend to fuse four scoring methods via their ranking and scoring. The ranking fusion function for data fusion of ranking score is described as following:

$$R(x) = \frac{\sum_{i=1}^{n} x'_n}{n} \tag{15}$$

x is new rank, x' represents original rank in virtual screening and n is total numbers of combined methods. In order to fairly compare with scoring combination and correctly combine scoring

from different methods, we have to normalize the scores of different methods and the normalization function in this study works as:

$$s_i = \frac{s'_i - s'_{\min}}{s'_{\max} - s'_{\min}} \quad (16)$$

s_i is normalized score and s'_I is original score. s'_{max} is maximum value in screening list and s'_{min} is minimum value. We combine normalized scores of different scoring functions and put them in order. We accessed the performances of linear combinations of ranking and scoring via a widely used evaluation indexes in virtual screening, such as enrichment factor and false positive rate. In order to further evaluate the quality and accuracy of fusion ranking for screening, we averaged these factors of each combination and the formulas of these factors have been mentioned in section 3.3.3 in detail.

4.4 Screening Accuracy Analysis

Result statistics of ranking combination of HSV-1 thymidine kinase (TK), human dihydrofolate reductase (DHFR) and estrogen receptor α (ER) on virtual screening tests are shown as Figure 11. In Figure 11, characteristics of ABCD represent different scoring functions such as original scoring (GEMDOCK-none) and pharmacophor-based scoring function (GEMDOCK-both) of GEMDOCK and original GoldScore (GOLD-ori) and recombinant scoring terms in GoldScore (GOLD-new) of GOLD. The recombinant score of GOLD integrates the terms of external *vdw* and external *H-bond* of GoldScore and this recombinant removed the penalty of internal *vdw* which made poor performances of GOLD in our test. We evaluated the performance and accuracy of virtual screening by average enrichment factor and average false positive rate. As shown in Figure 11A and 11B, four individual methods performed poor enrichments on TK. The enrichment were 6.20, 28.16, 10.34 and 7.09 for GEMDOCK without pharmacological preferences (none), GEMDOCK with pharmacological preferences (both), GOLD with GoldScore (ori) and GOLD with recombinant GoldScore (new), respectively. When method fusions carried out with combining a pair of methods one by one, the accuracies improved from 12.95 to 27.20 in average of overall enrichment and the average of false positive rates dropped from 5.65% to 2.97% (Table 13). Fusing three selected methods could further improve the whole enrichment to 35.36 in average and the overall false positive rate fell in average value of 1.94%.

With the increase of fused method, the average accuracy would promote and perform in averages of enrichment and false positive rate. Table 12, 13 and 14 shows similar trends of these promotions in average enrichments and false positive rates, except three combinations and four combinations in ER (Table15). From Figure 11E and 11F, the individual accuracies of GEMDOCK-none and GOLD-ori were 34.88 and 34.06 and the other two methods, GEMDOCK-both and GOLD-new, had good performances with 92.19 and 75.15, respectively. As shown in Figure 11E and 11F, combinations with GEMDOCK-none or GOLD-ori were in comparatively poor performances than others. It was a possibly reason that highly false positives leaded to highly noise and this phenomenon interfered and exceeded the correction ability of fusion. The data fusion result of DHFR was shown in Figure 11C and 11D. The accuracy of primary methods was best among three virtual screening experiments (in average of 69.27 against 12.95 (TK) and 59.07 (ER)) and the fusion of methods improved most in this case. The average of four methods combination on DHFR had the highest value (87.19) in enrichment factor. This phenomenon indicated data fusion indeed works well in different conditions. Data fusion could improve the quality of screening whether the primary methods performed well or not. Although Table 13, 14 and 15 showed the average accuracy level improving with number of fused methods, the unusual value of data fusion was shown in the extreme improve of combining a pair of methods. Comparing with Figure 11A, 11C and 11E, the maximum of promotion always occur in the combination of a pair of methods. The best composition in TK was GEMDOCK-both and GOLD-ori and it provided 1.9 fold improvement over best primary method (GEMDOCK-both). In DHFR, the best composition appeared on the combination of GEMDOCK-both and GOLD-new and the enrichment factor was 91.65 and a quite low false positive rate at 0.09%. In ER, the combination of GEMDOCK-both and GOLD-new presented an outstanding enrichment factor of 96.52. Comparing statistics on Table 13, 14 and 15 and overview of fusion shown in Figure 11, the results indicated that data fusion of ranking was useful for improving the accuracy of virtual screening.

Figure 12 shows the relationships between ranks and scores for three screening targets. The scoring showing in Figure 12 was normalized through equation 16. Figure 12 shows that the scoring gradient of GEMDOCK is larger than GOLD. The differences are because characteristics in scoring functions of GMEDOCK are different from GOLD in their mimic. Analyzing the relation between Figure 12 and Figure 12, it revealed a possible mode of the fusion performance. Researches of data fusion in information retrieval (IR) reported that the fusion performance

corresponds to the divergence of individual performance. From this hint in information retrieval, we tried to find the method pair with most variation from Figure 12 and identified the corresponding relation with accuracy improvement for Figure 11. Taking Figure 12A as an example the most variation pair was GEMDOCK-none and GOLD-new, but the best accuracy combination was GEMDOCK-both and GOLD-ori. This phenomenon could not fit with the characterization in IR. When taking the false positive rate of Figure 11 into account, a possible rule appeared. It was found that combining the pair with lowest false positive rate would follow the best performance in the enrichment. In Figure 11A and 11B, GEMDOCK-both and GOLD-ori had lowest false positive rates (1.43% and 5.04%) among primary methods and the combination of these methods had best enrichment (54.44) and lowest false positive rate (0.61%) among whole dataset (see Table 13). Similar phenomenon occurred in the rest two targets (DHFR and ER). In DHFR and ER, both GEMDOCK-both and GOLD-new had lowest false positive rates and their combination also brought the best performance in DHFR (enrichment 91.65, see Table 14 and Figure 11C and 11D) and ER (enrichment 96.52, see Table 15 and Figure 11E and 11F), respectively.

Consensus scoring is the popular strategy for solution the scoring inaccuracy problem in virtual screening[70]. The procedure of consensus scoring is to combine the top rankings of different methods[71,72] and the process of consensus scoring combines the scoring of individual methods. In order to compare data fusion of ranking with consensus of scoring, the analysis of ranking combination and scoring combination shows in Figure 13. Figure 13A and 13B show the performance of ranking combination and scoring combination. As shown in Figure 13, the ranking combinations are superior to scoring combinations in screening accuracy. The best enrichment for TK in ranking combination was 54.44 and the best of scoring combination was 37.96 (see Table 16). The best enrichment of DHFR with ranking and scoring combinations were 91.65 and 88.13, respectively (see Table 17). In ER, the best enrichment of ranking combination was 96.52 and the best one of scoring combination was 94.85 (see Table 18). Table 16, 17 and 18 summarized the statistics of scoring combinations in the same targets and these statistics revealed the accuracy of consensus scoring would promote with increase of scoring methods. Figure 13C and 13D indicate the screening accuracy with different combination methods and the ranking combinations remained better than scoring ones. Figure 13E and 13F show a little different with TK and DHFR. In ER some primary scoring methods already had well performance and the others did not. When combining the methods with highly differences, somehow scoring

combinations would work better than ranking combinations, such as combinations of AB, CD, BC, ABC, BCD and ABCD in Figure 13E. This phenomenon was similar with ranking fusion, but scoring combination would be better in such case.

From the study in data fusion on three screening cases, it proves that data fusion works in combining ranking of screening sets. When we associated individual ranks into different combinations, the accuracies of these combinations were better than each individual. If we chose the most divergent pair according to their rank-score curves (Figure 12), we could obtain the best performance such as Figure 11 shown. The best performance in TK is combined GEM-both and GOLD-ori; in DHFR and ER, we obtained best performance by combining GEM-both and GOLD-new. In this study we found that integrating ranks of two methods performed best and the performance decreased with more method integrated (e.g. Table 14, the best enrichment dropping from 91.65 to 87.19). This phenomenon was also found in data fusion in IR[49]. The comparisons shown in Figure 13 are fusions via scoring and ranking combination. The result proved that data fusion by ranking combination would indeed improve the enrichment factor and reduce false positive rate in virtual screening. In the recent future, we would make more evaluation for this novel consensus scoring concept and apply to our virtual screening method for improving the accuracy and reducing the false positive rate.