Metrics for early-recognition Methods in Virtual Screening, problems and solutions.

Beschreibung des Problems

Virtual screening (VS) is the term applied to a type of problem which is conceptually the computational equivalent of (experimental) high-throughput screening, wherein a large number of samples are quickly assayed to discriminate active samples from inactive samples. Evaluating the performance of VS methods is a necessary practice that both the method developers and the end-users perform. For the developers, it is done to parametrize and validate the methods, for the end users, it is a way to select which method performs best in a given situation.

The key requirement for success in VS is that it must rank actives very early in the larger set of compounds, since only a very small proportion of the compounds ranked will actually be tested experimentally. For instance, sample repositories of pharmaceutical companies commonly contain more than 1 million compounds. Identifying a set of a few thousand compounds to screen experimentally requires that the VS method perform well within the first 0.1% of the scored compounds! Even if the VS method is excellent at retrieving all actives for any target within the first half of the data set, it is useless if the early performance is bad.

This has led to considerable research into what method or approach works best, typically by means of ‘retrospective’ evaluations, i.e. attempting to predict future, i.e. prospective, behavior by appraising techniques on known systems. Despite this there is no agreed upon theory as to how to conduct a retrospective evaluation. As a consequence, it is very difficult for an outsider to assess if methods are getting better, have stayed the same, or even worsened over time.

Bewertungen

Area under the Receiver Operating Characteristic Curve. The ROC metric is widely used across many disciplines. The popularity of ROC over other metrics lies in the fact that it is nonparametric: no assumption is made about the shape of the distribution. Also, its graphical representation gives a good feeling about the performance of the ranking.

...

Area under the Accumulation Curve. Accumulation curves are widely used to display ranking performances. However, the corresponding area under the curve, the AUAC, is not as often used partly because it is believed to be largely dependent on the ratio of actives in the set.

...

Enrichment Factor. The EF metric is simply the measure of how many more actives we find within a defined “early recognition” fraction of the ordered list relative to a random distribution.

There are several other metrics available: the area under the accumulation curve (AUAC), the average position of the actives, 2 analysis of variance (ANOVA), 16 the Z-score, 17 the enrichment factor (EF), 18 the robust initial enhancement (RIE), 7,19 and so forth.

However, it is surprising that almost none of these metrics address the “early recognition” problem specific to VS.

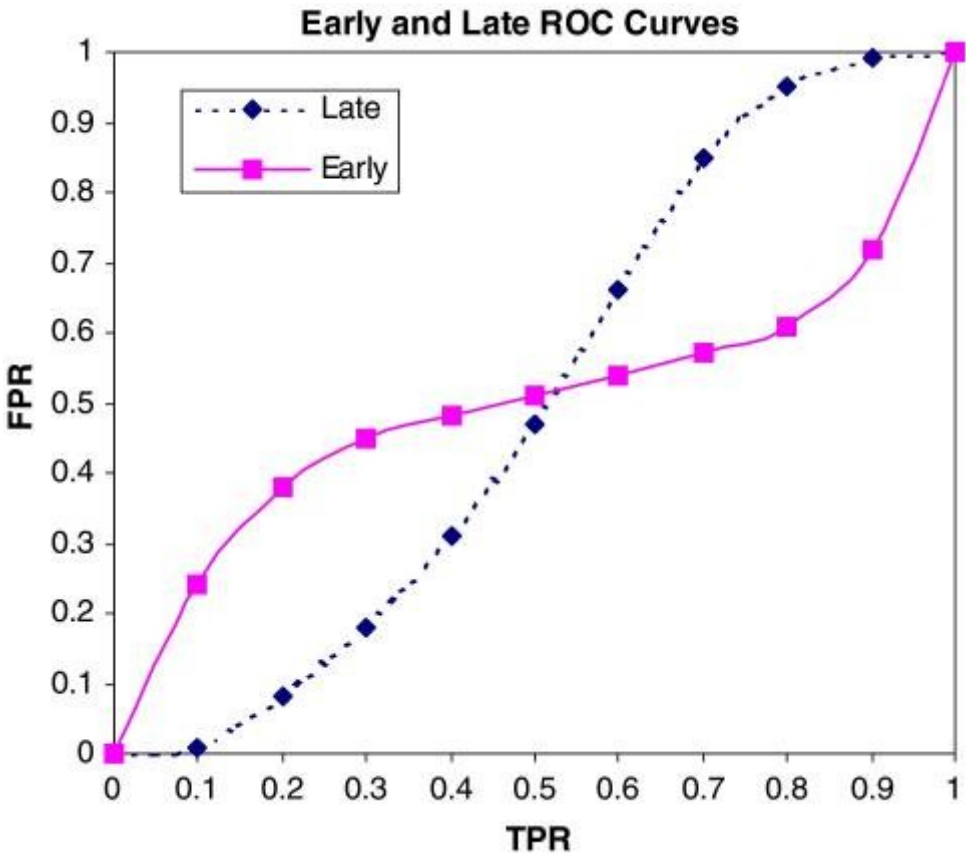

Fig. 4

Figure illustrates the supposed limitations of the AUC as a measure of performance. The graph shows two ROC curves, each with an AUC of exactly 0.5. Overall this means that an active is as equally likely to out-rank an inactive than the other way around. However, clearly in the case of the solid line a certain fraction of actives is being scored significantly higher than most inactives, while another fraction is being scored worse, i.e. it is only the average behavior that appears even-handed. Similarly, the dashed curve illustrates the case where the actives are all scored better then a certain fraction of the inactives but worse than another fraction. The solid and dashed curves are instances of bimodal score distributions for the actives and inactives respectively. Since the goal of a virtual screen is to save us the trouble of actually screening all the compounds it is entirely reasonable to prefer good ‘early’ behavior.

Ansätze

Robust Initial Enhancement. RIE, developed by Sheridan et al.,[Truchon - 19] is a metric using a continuously decreasing exponential weight as a function of rank.

...

The Boltzmann-Enhanced Discrimination of ROC (BEDROC) Metric.

So does BedROC or RIE qualify as a good metric for virtual screening? Comparing against the five criteria listed above, both are more robust than enrichment. RIE suffers from having an ill-defined numerical interpretation (i.e. how good is an RIE of 5.34?). BedROC attempts to overcome this by scaling between 0.0 and 1.0, but does this qualify as being understandable? There is no absolute, interpretable meaning to a BedROC (or RIE) number, only a relative meaning when ranking methods.

Alternative

There is no fundamental meaning to BedROC or RIE. Neither gets to the real heart of why the solid curve in Fig. represents a better method than the dashed curve. In what follows we will argue that this can be stated with respect to a set of assigned costs, assumed but never stated.

A preference for early behavior implies it is acceptable to miss a significant fraction of potential actives in favor of finding a few good leads. There is merit in this approach. Often a chemistry team can only follow up on a small number of leads.

What are not made explicit in this shift are the costs of the four components of any virtual screen: true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). The assumption behind virtual screening is that the value of a true positive similarly averaged is worth the cost of computers and modelers. This is an unproven conjecture. The assignment of a cost structure to the components of a screen is common in the field of medical diagnostics. Here the costs can be estimated with some reliability. A true positive represents the successful diagnosis of a condition that will save money when treated. A false positive means further, costly, tests will need to be performed. A false negative might cost a lot if a more severe condition develops. Finally, a true negative can be set to the cost of the test or a small saving if compared to a more expensive test. If these values are assigned to each "truth table" component (TP, FP, TN, FN), a ROC curve can be transformed into a cost curve. A small caveat is that the ROC curve deals with true and false positive rates and so to transform to real costs the expected number of actives and inactives is required, or at least the ratio of the two. Suppose we apply a cost structure to Fig. as follows:

1. TP = 8.0
2. FN = −2.0
3. FP = −0.16
4. TN = 0.02

Positive numbers are favorable, for instance the cost assigned to a true negative is the saving from not physically screening a compound. At any point in the curve the cost of progressing with all compounds higher than a given threshold t depends on

$$\text{Cost}(t) = \text{TPR} * N_a * (8.0) + (1 - \text{TPR}) * N_a * (-2.0) + \text{FPR} * N_i * (-0.16) + (1 - \text{FPR}) * N_i * (0.02$$

the False Positive Rate (FPR) and True Positive Rare (TPR):

Let us assume Na/Ni = 1/100, then:

$$\begin{aligned} \text{Cost}(t)/Ni &= (TPR * (8.0 + 2.0) - 2.0)/100 - FPR * (0.16 + 0.02) + 0.02 \\ &= 0.10 * TPR - 0.18 * FPR \end{aligned}$$

This is a simple linear scaling of the graphs in Fig. 4, as shown in Fig. 5a. As expected, the best approach is to take the method with early performance over the later performance. Notice that the late performing method is never cost effective and even the early method is only cost effective for a narrow range of rankings.

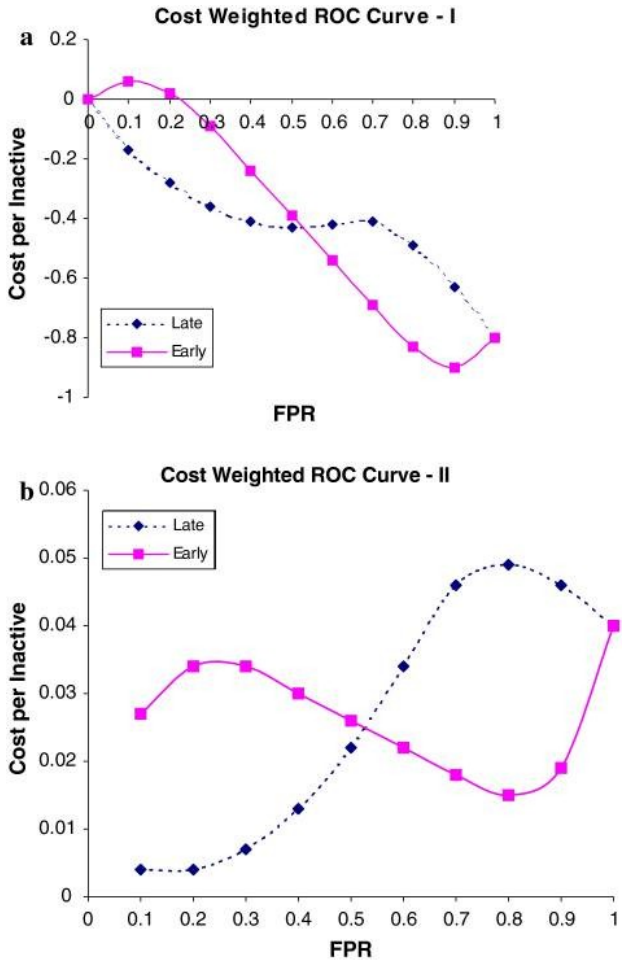

Fig. 5

Now consider a slightly difference weighting:

1. TP = 8.0
2. FN = −2.0
3. FP = −0.04
4. TN = 0.03

$$\mathrm{Cost}(t)/N_i = 0.1 * \mathrm{TPR} - 0.07 * \mathrm{FPR} + 0.01$$

Figure 5b illustrates the effect of these new weightings. By reducing the cost of a false positive by 75%, i.e. to around the savings of a true negative, both methods are always cost effective.

Suppose the cost structure of virtual screening does favors early enrichment. Can we at least say metrics such as RIE and BedROC, perhaps reformulated to be independent of extensive variables, are superior to AUC? If the early behavior shown in Fig. 4 were indeed repeated from system to system then clearly this would be the case.

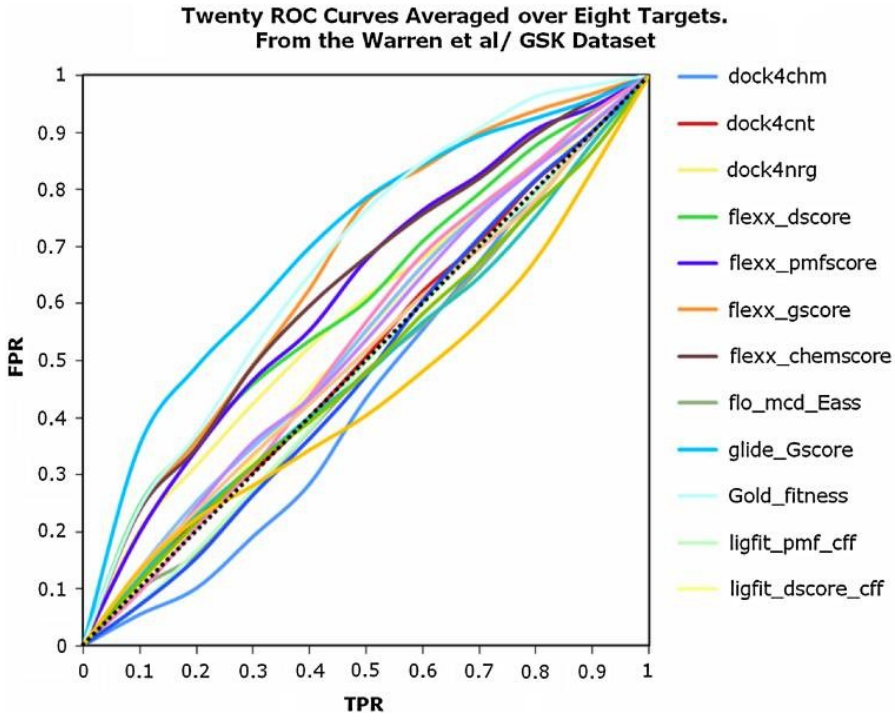

Fig. 6

In Fig. 6 we show data from Warren et al. for twenty docking procedures averaged over all eight targets in the study. Examination of these curves reveals nothing that resembles the biphasic nature anticipated from Fig. 4. Individual curves might

occasionally suggest biphasic behavior but there is little evidence for this in targetaveraged ROC curves.

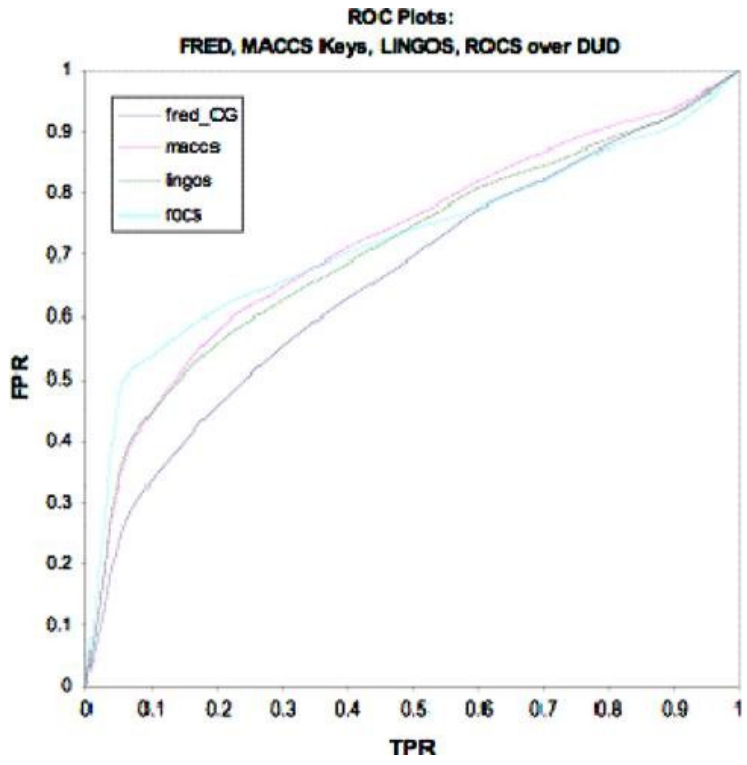

Figure 7

Figure 7 shows similar curves for the four methods in Fig. 1b averaged across the DUD set. The curves in Fig. 7 are smoother because the averaging across forty targets in DUD is more extensive than the eight from GSK and show even less evidence of biphasic behavior. There are two possibilities for these observations. Either the individual curves are not biphasic or the averaging dilutes this characteristic.

Conclusions

In the current report we have examined the evaluation techniques applied to virtual screening, including new approaches such as RIE and BedROC. Despite the great importance of a virtual screening in a drug development process, still there are no well-established and generally accepted approaches for method evaluation in this

field, which is always the case for a young and developing industry. Also an attempt was made to link an evaluation of a method effectiveness to a real cost structure when developing a new drug. This approach demonstrates all benefits of using virtual screening, it translates abstract numerical metrics into a real cost structure that allows not only researchers that conduct the screening to evaluate the method effectiveness , but also allows people who are not engaged in this field to do so.