J Comput Aided Mol Des (2008) 22:141–146
DOI 10.1007/s10822-008-9181-z

Managing bias in ROC curves

Robert D. Clark · Daniel J. Webster-Clark

Received: 19 November 2007 / Accepted: 14 January 2008 / Published online: 7 February 2008

Abstract Two modifications to the standard use of receiver operating characteristic (ROC) curves for evaluating virtual screening methods are proposed. The first is to replace the linear plots usually used with semi-logarithmic ones (pROC plots), including when doing "area under the curve" (AUC) calculations. Doing so is a simple way to bias the statistic to favor identification of "hits" early in the recovery curve rather than late. A second suggested modification entails weighting each active based on the size of the lead series to which it belongs. Two weighting schemes are described: *arithmetic*, in which the weight for each active is inversely proportional to the size of the cluster from which it comes; and *harmonic*, in which weights are inversely proportional to the rank of each active within its class. Either scheme is able to distinguish biased from unbiased screening statistics, but the harmonically weighted AUC in particular emphasizes the ability to place representatives of each class of active early in the recovery curve.

Keywords Early recognition · ROC AUC · Virtual screening

Introduction

Since being introduced for that purpose, the receiver operating characteristic (ROC) curve has become, for many, the tool of choice for evaluating virtual screening performance in general [1–3] and docking performance in particular [4]. The ordinate in an ROC plot is the fraction of actives recovered among the top fraction x of the data set [5], just as for a simple recovery curve. It differs in that the abscissa reflects only the fraction of *inactives* recovered at that level—i.e., the false positive rate β. In a simple recovery curve, by contrast, the abscissa is the rank of the active. This somewhat subtle distinction exerts its most obvious effects on the early phase of the recovery curve, but it confers some excellent statistical properties on the overall area under the curve (AUC) that can be very useful when comparing virtual screening methods, particularly those of the high-throughput variety (vHTS) [4, 5].

Researchers would like to have the best of all worlds, recovering all actives and few (ideally, no) inactives among the top-ranking candidate ligands. In practice, there is always a trade-off between increasing the fraction of actives recovered (the true positive rate α) and reducing β. ROC analysis seeks to get beyond the tension between these desires by integrating recovery across the full range of potential false positive frequencies.

There are two potential problems with this approach as applied to date, however. One is a perceived failure to take into account the differences in inherent value between "early" and "late" hits. The improvement in ROC AUC for shifting the point at which three actives are recovered from 0.05 to 0.01 is more than offset by a single active shifting from 0.40 to 0.25, even though the former benefit represents a major advance and latter cost is of little or no practical importance. Truchon and Bayly [6] explored several weighting schemes designed to address this "early recognition problem". They concluded that the exponential weighting scheme BEDROC worked best in terms of overall discrimination and providing good "early recognition" of

R. D. Clark (✉)
Tripos Informatics Research Center, 1699 South Hanley Road, Saint Louis, MO 63144, USA
e-mail: bclark@tripos.com

D. J. Webster-Clark
Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63130, USA

actives, in their hands out-performing enrichment at specific threshold values as well as the related robust initial enhancement (RIE) method developed by Sheridan et al. [7].

A second problem relates to the uneven structural variation among the actives in test sets, particularly when complexes of closely related ligands have been used to train one of the vHTS tools being evaluated. It is generally more important in such cases to be confident that one will be able to retrieve a representative from each of several structural series than that one will be able to retrieve most or all of the individual actives [8, 9].

Methods

One of the strengths of ROC curves is that quantitative comparisons between and among them can be carried out by comparing the areas under them—the ROC AUCs. Numerically, this can be formulated as:

$$\text{ROC AUC} = \frac{1}{n}\sum_{i}^{n}(1-\beta_i) = 1 - \frac{1}{n}\sum_{i}^{n}\beta_i \qquad (1)$$

where n is the number of actives and β_i is the false positive frequency corresponding to that at which the ith known active was found, typically calculated as the fraction of decoys ranked higher than the ith active.[1] Equation 1 makes clear that the ROC AUC is directly related to the average false positive rate across all actives. In most cases, the "inactives" are not actually known to be inactive. When that is so, β_i is more accurately referred to as the *nominal* false positive frequency, and the reliability of some statistical inferences connected with AUC comparisons will suffer.

An appealing alternative to increasing the influence of early hits is to decrease the influence of late hits. A simple way to do this is to replace the linear β term in Eq. 1 with a logarithmic term, which is equivalent to integrating a semilogarithmic plot. Negation yields a "natural" statistic for which a bigger value is better, and using the common logarithm puts the results on a par with concentration-dependent values that computational chemists are accustomed to dealing with—pH, pK_a, pK_i, etc. This leads to the pROC AUC definition given in Eq. 2:

$$\text{pROC AUC} = \frac{1}{n}\sum_{i}^{n}[-\log_{10}(\beta_i)] = \frac{1}{n}\sum_{i}^{n}\log_{10}\left(\frac{1}{\beta_i}\right) \qquad (2)$$

Equation 2 makes it clear that the pROC AUC is equal to the average *stringency* at which the actives are picked out, where the stringency is defined as the negative logarithm of the corresponding false positive rate.[2] Hence a pROC AUC of 3 indicates that the average active can be successfully picked out from a background of 1000 decoys.

The ROC AUC corresponding to *random selection* is the one for which the true positive rate is expected to equal the false positive rate, so that the selectivity is nil—i.e., $\alpha = \beta$. In that case, the AUC is simply 0.50. Unlike the ROC curve, the pROC curve is unbounded on the high side (the nominal low side in Fig. 2b). Fortunately, its integral is finite nonetheless, and is given in Eq. 3, wherein "log" denotes the natural logarithm.

$$\lim_{a\to 0}\int_a^1(-\log_{10}x)\mathrm{d}x = \frac{-1}{\log 10}\lim_{a\to 0}\int_a^1(\log x)\mathrm{d}x = 0.434\lim_{a\to 0}\{x - x\log x\}|_a^1 = 0.434 \qquad (3)$$

Applying a logarithmic transform to β adds beneficial bias to an ROC curve. There are also cases where it is desirable to reduce or remove an undesirable bias in α for the data set being evaluated. In particular, if a set of known actives represents K distinct structural subsets, a generalized form of the AUC definition can be calculated, defined as:

$$\text{ROC AUC} = \frac{1}{\gamma}\sum_{j}^{K}\sum_{i}^{k_j}w_{ij}(1-\beta_{ij}) = 1 - \frac{1}{\gamma}\sum_{j}^{K}\sum_{i}^{k_j}w_{ij}\beta_{ij} \qquad (4)$$

where w_{ij} is the weight for the ith-ranked active in the jth subset and k_j is the number of actives in the jth class. The normalization factor γ is given by Eq. 4:

[1] Note that β_i can be calculated from the active and inactive ranks as $r_i - i$, where r_i is the rank of the ith-best active across *all* compounds. If ties occur, they should be resolved by assigning the average of the contested ranks: observations tied for ranks 5 and 6 each receive a rank of 5.5, for example [10].

[2] It would be convenient to have a term to denote $\log(1/\beta)$ that parallels the related concept of "potency" in medicinal chemistry. "Stringency" is suggested here because of its extensive use in biochemistry in connection with semi-log plots of interactions with and between nucleic acids. It has no conflicting meaning in the standard statistical literature, though it has been used in connection with the logarithm of risk (See [11]).

$$\gamma = \sum_{j}^{K} \sum_{i}^{k_j} w_{ij} \tag{5}$$

In the "ordinary" AUC calculation, weights are *uniform* and equal to unity, so $\gamma = n$. In *arithmetic weighting*, on the other hand, each class contributes an equal amount and each "hit" receives the same weight as other actives in its class regardless of its position in the ROC curve: $w_{ij} = 1/k_j$. The value of α in the ROC plot is then equal to the sum of the individual weights up to that point rather than simply the fraction of actives recovered to that point, i.e.:

$$\alpha(\beta_{ij}) = \sum_{j}^{K} \sum_{i}^{k_j} w_{ij} \alpha_{ij}(\beta_{ij}) \tag{6}$$

where $\alpha_{ij} = 0$ for actives not identified at or below the corresponding (nominal) false positive rate β_{ij} and is equal to 1 otherwise.

In *harmonic weighting*, early hits in a class receive more weight than later "hits", and larger classes receive more weight than smaller ones but not proportionately more. Rather than contributing equally, each "hit" is weighted in inverse proportion to its rank within the cluster to which it belongs: $w_{ij} = 1/i$. Similar weighting schemes are used for averaging rates, but are also often used in cases where the statistic of interest is connected to waiting for a rare event to occur or to recur. The rationale is that the first representative encountered carries all of the known information up to that point, the second representative encountered carries half of the aggregate information, and so forth. A class comprised of two actives makes an aggregate contribution of 1.5 to γ, a class comprised of three actives contributes 1.833…, etc.

Results

Figure 1 is a schematic of the artificial data set used to illustrate the effects of the various weighting schemes explored here. The three groups of points in the plot represent the three leads shown around the perimeter and their analogs, distributed in an arbitrary descriptor space defined by two structural attributes. The goal in this idealized example is to see how likely a particular vHTS method is to find the new lead shown in the center. The specific structures shown in Fig. 1a are drawn from an actual tyrosine kinase-3 inhibition project [12]. The numbers of analogs "available" for each in the analysis that follows, however, are completely artificial, as are the distribution of properties shown in Fig. 1.

Figure 1b shows how the actives cluster and the different ways in which ranks among actives might be distributed across the clusters for two different vHTS programs, e.g., two different docking engines. The red ranks in Fig. 1b come from an extremely *biased* docker, where the ranks within each cluster are themselves tightly clustered with respect to the other actives. Such bias shows up to a greater or lesser degree in cases where the protein structure used has been extracted from a complex with a ligand from the largest cluster as well as in ligand-based screening studies wherein a representative structure from that cluster is used as "bait". Such situations arise uncomfortably often in practice, since the largest cluster is apt to represent the most established structural series—which is likely to have the most solved X-ray structures and to include the most potent inhibitors.

The blue ranks in Fig. 1b, on the other hand, represent results for a (hypothetical) *unbiased* docker, the relative ranks within each cluster being spread across the full range of active scores. For purposes of illustration, it is presumed here that both vHTS methods being evaluated are effective at discriminating between actives and inactives, so the same reasonably good score profile with respect to the inactives has been assigned to both.

The unbiased blue docker will be more likely to assign a relatively high score to the novel target ligand shown in the center of Fig. 1a, so it would be good to have a performance metric that would clearly distinguish it from the biased docker.

The expected *random* scenario, in which $\alpha = \beta$ is included in the calculations below as a benchmark. It represents the null hypothesis of no discrimination, a straw man of minimal performance to which the individual ROCs can be prepared.

When the actives are uniformly weighted ($w_{ij} = 1$), identical ROC curves are obtained for the two dockers; the common curves are presented in two different ways in Fig. 2. Figure 2a shows the ROC curve as classically plotted, whereas Fig. 2b is a semilog plot.[3] For this illustration, it has been presumed that 1000 inactives have been included as decoys, that the best-scoring candidate ligands are actives, and that the best-scoring candidate comes from the largest cluster, cluster 1. There are 20 actives in total, so $\alpha_{11} = 0.05$ and $\beta_{11} = 0$. Plotting the logarithm of 0 would be problematic, but it would also be inappropriate; the data really only indicates that $\beta_{11} < 0.001$, not that it is actually equal to zero. Therefore setting $\beta_{11} = 1/N$ (where N is the number of inactives in the data set) is reasonable and conservative as well as practical. The same substitution was done when calculating all pROC AUCs.

Applying arithmetic weighting gives the curves shown in Fig. 3a, whereas applying harmonic weights yields the curves shown in Fig. 3b. The red line in each plot

[3] Such semilog plots will be referred to as pROC plots here, because the abscissa is labeled in terms of β and not reversed in orientation, it is scaled logarithmically.

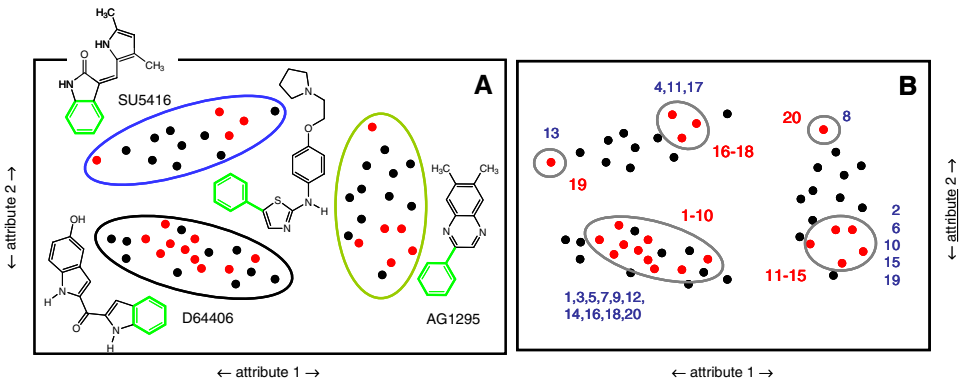

Fig. 1 Artificial data set used for ROC analyses. (**a**) Distribution of structural series in an arbitrary descriptor space. Red symbols represent active analogs and black symbols represent inactive analogs. Green highlights the key shared substructure [12]. (**b**) Clustering of actives. Ranks coming from a biased vHTS program are shown in red. Ranks for an unbiased program are shown in blue

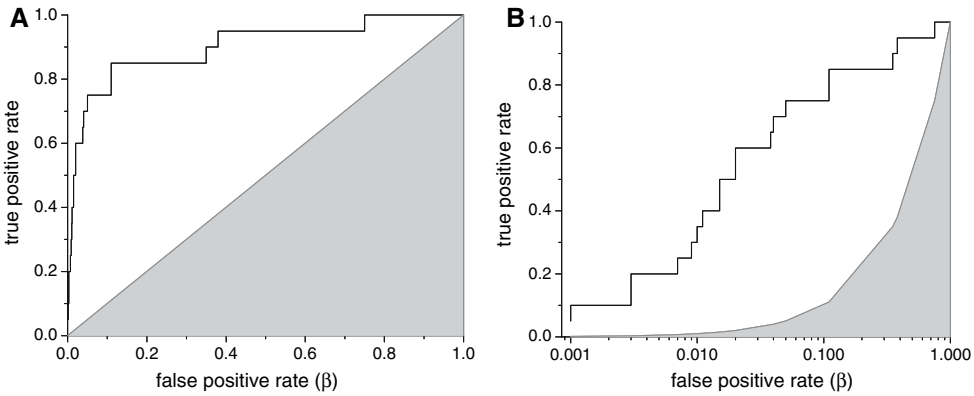

Fig. 2 Plots using uniform weighting. The upper boundary of the gray area corresponds to the curve expected under a random scenario. (**a**) Linear ROC curve. (**b**) pROC curve

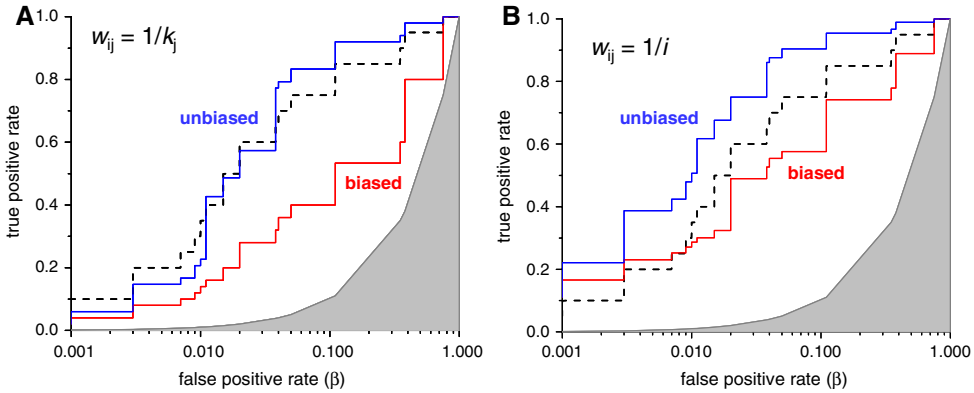

Fig. 3 pROC curves under different weighting schemes. Blue lines represent results for unbiased vHTS and red lines represent results for biased vHTS. Dotted lines correspond to uniform weighting. (**a**) Arithmetic weighting. (**b**) Harmonic weighting

corresponds to the pROC curve for the biased docker and the blue line corresponds to the pROC curve for the unbiased docker. The curve obtained for uniform weighting is included for reference in both panels of Fig. 3 as a dotted line.

Arithmetic weighting does not shift the center of the unbiased curve from that seen for uniform weighting because the ranks for all clusters are more or less evenly distributed around a false positive rate of 0.02—i.e., you would expect to find half of your activity among ligands scoring in the top 2% of the data set. The low- and high-scoring actives belong to the larger classes, however, which accounts for the blue curve in Fig. 3a being more tightly focused around that center. The biased profile, in contrast, is shifted to the right, reflecting the fact that nearly 10% of the decoys would need to be physically screened to recover half of the "activity"—which, in this case, reflects the number of chemical series recovered rather than simply the number of actives.

Because early hits are relatively more influential, harmonic weighting (Fig. 3b) shifts both pROC curves up and to the left with respect to their position under arithmetic weighting, but not enough to bring the biased curve above the "normal" pROC obtained using uniform weighting.

Table 1 shows the AUC statistics obtained in each case, and compares them to the values expected for random

Table 1 Effect of bias and weighting scheme on AUC values

Weighting	ROC AUC			pROC AUC		
	Unbiased	Biased	Random	Unbiased	Biased	Random
Uniform	0.903	0.903	0.500	1.664	1.664	0.434
Arithmetic	0.937	0.728	0.500	1.688	1.025	0.434
Harmonic	0.962	0.836	(0.633)[a]	2.059	1.484	(0.570)[a]

[a] The values set off in parentheses are specific for the distribution of cluster sizes used in the particular example described; the others are general

selection. As expected, the common log transform heightens the discrimination between the biased and unbiased vHTS methods.

Discussion

ROC analysis was originally formulated to evaluate the ability to resolve closely spaced signals of comparable intensity, a situation that differs from virtual screening in at least two key ways. Firstly, the active and inactive populations encountered in vHTS are very disparate in size as well as in distribution, with inactives typically out-numbering actives by 1000:1 or more.

Secondly, most of the information needed to resolve two signals is received late in the signal acquisition and is a matter of adding detail. In lead identification, on the other hand, it is the early "hits" that carry the most critical information [6, 7], particularly when it comes to identifying structurally distinctive actives. This is the reason that discussions of docking performance often involve separate plots for the early and later parts of the recovery curve [13], or use semilog plots instead to improve visualization [14–16]. As it happens, the logarithmic transform is a classic way to ameliorate disparities in population size and make variances more uniform. It also serves to shift emphasis from the later points in the ROC curve to the early ones. Taken together, these considerations argue that the bias introduced by using pROC AUCs to assess vHTS performance should be a net positive. Moreover, the cited benefits are obtained without requiring an extraneous parameter such as the exponential prefactor α used in RIE and BEDROC, and without the attendant problems with metric variance and saturation seen for the latter metrics [6].

The zero-point continuity correction involved in assigning a β value of $1/N$ to actives that out-score all inactives makes the calculated pROC AUC dependent on the number of inactives in the data set. Note, however, that this is a conservative effect, in that limiting the number of inactives can only underestimate the "true" pROC AUC expected in the limit where all possible actives are being compared to an infinite number of decoys. In addition, it has the desirable side effect of making the limited resolution of the performance statistic explicit rather than burying it in a secondary statistic. In fact, the continuous variable β is most robustly estimated from the discontinuous ranks for *both* direct ROC and pROC analyses by applying a more extreme transform: adding 0.5 to *each* rank, then dividing by $N + 1$ rather than by N [10, 17]. Calculating β this way has its largest effect at the left side of the plots and has almost the same numerical effect on the AUC calculations as the more localized zero-point correction described here. Its superiority in principle is offset, however, by the fact that the plots it produces would be less readily interpretable. Putting the entire discontinuity penalty onto zero-point observations, as has been done here, is an approximation that is more consistent with the WYSWYG principle.

One virtue of the ROC AUC is that it is bounded above and below by 1 and 0, respectively. The pROC AUC is technically unbounded above, though the logarithmic transform and zero-point correction mitigate this in practice. The more important bound, however, is set by the frequency of *false* false positives—i.e., the frequency with which actives appear among the nominally inactive decoys. HTS experience indicates that this frequency is around 0.1%, so a pROC AUC much above 3.5 (a stringency of 0.03%) is grounds for suspecting over-fitting or bad decoy selection rather than superior performance.

Unfortunately, not all biases are so benign. vHTS analyses are rarely based on statistically balanced data sets. Instead, they make use of sets of actives that are structurally "clumpy". Such clumpiness is due in part to incidental aspects of drug development, in part to details of patent law, and in part to the fact that changes in chemical structure are intrinsically discontinuous. These factors often produce rather arbitrarily uneven distributions of actives across structural series. When that is the case, vHTS methods that pick out most of the actives found in larger classes and miss smaller classes will "look" better than those that successfully identify a few actives from each class. Such bias is especially perverse in those cases where class size is inversely related to the opportunity it presents—as is often the case. As a practical matter, all actives within a cluster should eventually be found anyway (at least in principle) as a result of follow-up similarity searching. Hence they are not really "missed" if they do not turn up among the early hits, so penalizing a vHTS method that only picks out some ligands from each cluster is neither necessary nor appropriate.

This is the logic behind evaluating performance in terms of "hit" lead *series* rather than in terms of how well individual actives are recovered [8, 9]. This approach is hindered somewhat by the limited number of diverse, well-characterized lead series available, at least in the

public domain. It can also be somewhat sensitive to the specific clustering method and threshold similarities used to define the chemical series, though the practical impact of such sensitivity should be minor provided the same criteria are used for (virtual) lead follow-up. One convenient aspect of arithmetic weighting is that the expected value for the AUC is the same as one would get by including one ligand drawn at random from each class in the test set. It is a much more precise estimate of that AUC, however, since it represents all possible combinations rather than a single sample.

Not all clusters are created equal, however, especially when one moves beyond validation studies carried out on well-characterized data sets. There is a real trade-off between redundancy among vHTS "hits" and the opportunities presented by novel chemistries. As in "real" HTS, the most valuable hit is often not the first in a chemical series but the *second*, because the latter establishes the potential for finding a useful structure-activity relationship (SAR). The third hit is of value for the same reason, but less so. This is the logic behind applying harmonic weighting to the (p)ROC.

Conclusions

pROC curves (semilog plots) clearly improve qualitative discrimination between vHTS results with respect to the linear ROC plots generally used. It seems reasonable that the extension to quantitative analysis—pROC AUCs—would be similarly beneficial, particularly given the fact that the transform involved is a simple, non-parametric one that preserves many of the good statistical properties of the linear form.

As is often the case in docking and scoring problems, no one α weighting method is likely to be "best" for all applications [18]. The inherent properties of any given vHTS method will usually affect discrimination within and between clusters differently, leading to a general, characteristic sensitivity to bias that can readily be assessed by comparing uniformly and arithmetically weighted AUCs. Flexibility and general promiscuity in the target protein, however, is likely to be at least as important, so it will be prudent to carry out such analyses on the target of interest or one closely related to it. Harmonic weighting is very well suited to comparisons of performance on specific data sets, but its sensitivity to the distribution of cluster sizes among the actives makes direct comparisons between systems problematic. It would not make sense to average harmonic AUCs across a diverse set of targets, for example, whereas averaging arithmetically weighted AUCs will yield valid statistics.

Such considerations underlie the somewhat artificial data sets used here to illustrate how large the differences between biased and unbiased methods can be. The contrast is likely to be less for any particular combination of docking method, target protein, decoys and actives considered. It will also depend on how complementary the method used to cluster the actives is to how docking is done, the key point being not choosing the "right" method but applying the one that will actually be used in lead follow-up. No one particular circumstance could "prove" that the methods described here will be useful, nor could any handful of applications. Their theoretical justification is rooted in first principles, but their practical value will only become evident after they have been applied to a wide range of real-world problems.

Acknowledgements The authors appreciate the helpful suggestions provided by Peter Willett of the University of Sheffield and by our anonymous reviewers, and wish to thank Ajay Jain (UCSF) and Anthony Nicholls (OpenEye) for organizing the American Chemical Society Symposium that led to this special issue of the Journal.

References

1. Jain AN (2000) J Comput-Aided Mol Des 14:199–213
2. Cuissart B, Touffet F, Cremilleux B, Bureau R, Raul S (2002) J Chem Inf Comput Sci 42:1043–1052
3. Jain AN (2004) J Med Chem 47:947–961
4. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) J Med Chem 48:2534–2547
5. Egan JP (1975) Signal detection theory and ROC analysis. Academic Press, New York
6. Truchon J-F, Bayly CI (2007) J Chem Inf Model 47:488–508
7. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) J Chem Inf Comput Sci 41:1395–1406
8. Good AC, Hermsmeier MA, Hindle SA (2004) J Comput-Aided Mol Des 18:529–536
9. Good AC, Oprea TI (2008) J Comput Aided Mol Des 22. doi: 10.1007/s10822-007-9167-2
10. Daniel WW (1978) Applied nonparametric statistics. Houghton-Mifflin Co., Boston
11. Hamilton JT, Viscusi WK (1999) Calculating Risks? The Spatial and Political Dimensions of Hazardous Waste Policy. MIT Press, Boston
12. Furet P, Bold G, Meyer T, Roesel J, Guagnano V (2006) J Med Chem 49:4451–4454
13. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) J Chem Inf Model 47:1504–1519
14. Halgren TA, Murphy RB, Friesner RB, Beard HS, Frye LL, Pollard WT, Banks JL (2004) J Med Chem 47:1750–1759
15. Schellhammer I, Rarey M (2007) J Comput Aided Mol Des 21:223–238
16. Shepphird JK, Clark RD (2006) J Comput Aided Mol Des 20:763–771
17. Snedecor GW, Cochran WG (1989) Statistical Methods, 8th edn. Iowa State Press, Ames IA
18. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) PROTEINS 60:325–332