Available online at www.sciencedirect.com
ScienceDirect
J. Biochem. Biophys. Methods xx (2007) xxx – xxx
JOURNAL OF biochemical and biophysical methods
www.elsevier.com/locate/jbbm

Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities

Róbert Busa-Fekete [a], Attila Kertész-Farkas [a], András Kocsor [a], Sándor Pongor [b,c,*]

[a] Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged, Aradi vértanúk tere 1., H-6720 Szeged, Hungary
[b] Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy
[c] Biological Research Centre (BRC), Temesvári krt. 62, Szeged, H-6701 Hungary

Received 12 April 2007; received in revised form 30 May 2007; accepted 24 June 2007

Abstract

Identification of problematic protein classes (domain types, protein families) that are difficult to predict from sequence is a key issue in genome annotation. ROC (Receiver Operating Characteristic) analysis is routinely used for the evaluation of protein similarities, however its results – the area under curve (AUC) values – are differentially biased for the various protein classes that are highly different in size. We show the bias can be compensated for by adjusting the length of the top list in a class-dependent fashion, so that the number of negatives within the top list will be equal to (or proportional with) the size of the positive class. Using this balanced protocol the problematic classes can be identified by their AUC values, or by a scatter diagram in which the AUC values are plotted against positive/negative ratio of the top list. The use of likelihood-ratio scoring (Kaján et al, *Bioinformatics,* **22**, 2865–2869, 2007) the bias caused by class imbalance can be further decreased.

Keywords: Receiver Operating Characteristic; ROC; AUC; Classification; Ranking; Proteins

1. Introduction

Classification of protein sequences is a fundamental exercise in genome annotation so assessing and comparing the efficiency of sequence or structure similarity measures is a crucial task. The method of current choice is ROC (Receiver Operating Characteristic) analysis that evaluates the ranking ability of a similarity measure [1,2]. Briefly, the members of a database are ranked according to their similarity to a query using a similarity score (such as calculated by BLAST [3], Smith and Waterman [4], etc.), and a similarity score is considered efficient, if the proteins belonging to the query's known class (true positives) are on the top of the list. The analysis is carried out by preparing a sensitivity vs. specificity plot, whose integral value, AUC (area under curve) is 1.00 if the true positives are all on the top of the list, and will tend to 0.5 if the ranking is random [1].

In the practice of bioinformatics there are two main variants of this method (Fig. 1). In the element-wise scenario, originally suggested by Gribskov and Robinson [5], one prepares a separate ranked list for all the queries, which will yield one AUC value for each query. The group-wise scenario can be applied if the database is a priori divided into distinct classes, such as domain types [6,7]. Here we prepare one single ranking for a protein group, in which we rank the members of the test set with respect to their maximal similarity to the positive train group. Then we calculate a single AUC for the given group.

If we wish to calculate a cumulative value for an entire database using either scenario, the usual method is to calculate the arithmetic average of the individual AUC values, what is equivalent to constructing a cumulative AUC plot (such as shown in Fig. 4 below) and calculating its integral value [6,7]. In this representation higher curves indicate better performance.

ROC analysis presupposes the existence of two classes, of which the query's known class is the positive while the rest of the

Abbreviations: ROC, Receiver Operating Characteristics; AUC, area under curve, the integral of the ROC curve; BLAST, basic local alignment search tool program of Altschul et al.; SCOP, Structural Classification of Proteins, protein domain 3D database of Murzin et al.

* Corresponding author. Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy. Tel.: +39 040 3757300; fax: +39 040 226 555.
E-mail addresses: busarobi@inf.u-szeged.hu (R. Busa-Fekete), kfa@inf.u-szeged.hu (A. Kertész-Farkas), kocsor@inf.u-szeged.hu (A. Kocsor), pongor@icgeb.org (S. Pongor).

0165-022X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbbm.2007.06.003

database is the often much larger negative group, in other terms the positive/negative ratio (p/n) is often negligibly small. In order to handle this class-imbalance problem and to get a dataset of manageable sizes it is customary to cut the top list at some point, and according to Gribskov and Robinson [5] this can be done by limiting the top list so as to include n negatives (where n is usually taken as some plausible number like 10, 50 etc.). These are the so-called ROCn (e.g. ROC10, ROC50) values that were originally proposed for the element-wise scenarios but they are also frequently used in the group-wise scenarios.

ROC analysis is considered reliable because it includes both specificity and sensitivity, so a method with a high AUC value can be expected to be a robust in a variety of conditions. While we generally agree with this view, we noticed that the ROCn values can be differentially biased in various classes within a database. The result of this differential bias is that one cannot directly compare AUC values between different classes, although AUC values obtained on the same class with different methods are comparable. This is a major problem, since discovery of new protein or gene groups is perhaps the most important tasks in genome research, so there is a need to assess and compare, without class-imbalance artifacts, the predictability of protein groups. Doubtlessly, this assessment could be done by developing and fine-tuning classifier algorithms for each of the new candidate groups, which is, on the other hand too time consuming in view of the amounts of data to be analyzed. Our motivation was to use ROC analysis directly for the purpose. As the positive/negative ratio within a top list is known to influence the AUC values, we were looking for methods where this ratio can be controlled and monitored.

Here we propose a simple method wherein the number of the selected negatives included into the analysis is equal to (or proportional with) the number of the positive sample. This balanced ROC analysis provides less biased AUC values, which can be used to identify difficult groups within a database.

2. Materials and methods

2.1. Datasets and algorithm

The SCOP95 dataset used for modeling consisted of protein domain sequences taken from the SCOP database v.1.69, filtered for 95% identity [8]. For the comparison we used the Smith–Waterman algorithm [4] as coded in the EMBOSS program package [9] and applied predefined classification tasks for superfamily prediction. The classification tasks and the Smith–Waterman data were taken from record PBS0001 of the Protein Classification Benchmark collection [10], so the positive set contained members of s superfamily, one family being the +test set and the rest of the families were the +training set. We performed the evaluation using either the element-wise or the group-wise scenario. In each case we used supervised subdivision [12].

2.2. ROC calculation

Calculation of AUC was carried out as described [6,7,10,11], using top lists including a given number of negative samples.

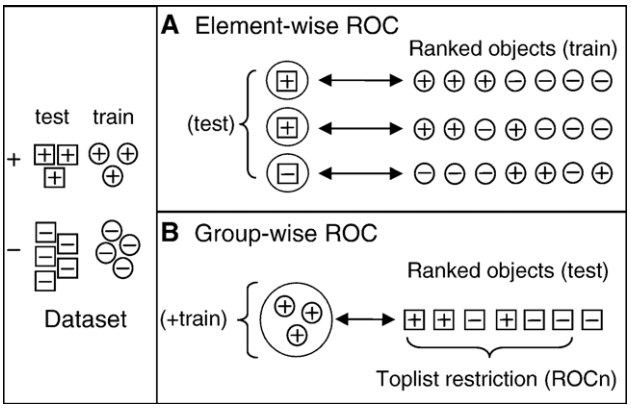

Fig. 1. Scheme of the two scenarios for calculating ROC. In the element-wise scenario (A), each query is compared to a dataset of + and − train examples. A ROC curve is prepared for each query and the integrals (AUC values) are combined to give the final result for a group of queries. In the group-wise scenario (B) the queries of the test set are ranked according to their similarity to the train group, and the ROC AUC value calculated from this ranking will be used to characterize the group.

The number of negatives (n) was expressed in multiples of the positive samples (p) participating in the restricted ranking. This normalized number N_r is thus related to the positive/negative ratio, so at $N_r = 1.0$, $p/n \leq 1.0$. p/n be lower than 1.0 for difficult classes, since in these cases there are positive samples that do not show up in the top lists, especially when less sensitive methods comparison methods are used.

For the balanced ROC (BaROC) protocol we selected the top list in such a way that it contained as many negative samples as the number of samples in the entire positive set. This corresponds to $N_r = 1.0$, so the p/n values are between zero and 1.0. It is equally possible to use a value of $N_r = 2.0$ is i.e. to use twice as many negatives as there are positives, in that case the p/n values will be between zero and 0.5. One can easily show that for any N_r, $p/n \leq 1/N_r$.

We mention that the BaROC protocol – same as the ROCn principle suggested by Gribskov and Robinson [5]— is based on toplist truncation, so the AUC value expected for random ranking is lower than 0.5.

2.3. Likelihood-ratio scoring

In the practice of protein sequence classification the input variable of ROC analysis is a sequence similarity score, such as a BLAST or Smith–Waterman score. Recently it has been proposed that a simple likelihood-ratio approximant, LR, is a more efficient ranking indicator [12]. The LR score of a query is calculated as

$$\mathrm{LR} = \frac{S^{+}_{\max}}{S^{-}_{\max}} \tag{1}$$

where $S^{+}_{\max}$ and $S^{-}_{\max}$ are the top similarities obtained between the query and members of the positive and negative groups, respectively. In contrast to a simple similarity score $S^{+}_{\max}$, LR also contains information on the negative class. A similar

Please cite this article as: Busa-Fekete R et al. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. J Biochem Biophys Methods (2007), doi:10.1016/j.jbbm.2007.06.003

Table 1
Dependence of the AUC value n the size of the negative set

No. of negatives in the top list	AUC	p/n
10	0.900	0.100000
50	0.500	0.040000
Full negative (5990)	0.833	0.000835

Globin-like proteins, a.1.1. in SCOP95. See text for details.

scoring measure, an approximant of the posterior probability p is used by the IBk program of the WEKA package [13].

$$p = \frac{S^{+}_{\max}}{S^{+}_{\max} + S^{-}_{\max}}. \quad (2)$$

Since the likelihood ratio is defined as $p/(1-p)$, it is easy to show that these two indices have in fact the same meaning and thus lead to identical ranking. In our comparisons we used the LR score, in addition to simple scoring that uses only $S^{+}_{\max}$. We point out that LR scoring is meaningful only in the group-wise scenario, the ranking of the element-wise scenario does not change upon the transformation to LR, so the ranking remains the same as with simple scoring.

3. Results

The heterogeneity of protein groups is a fundamental problem of protein classification. Protein groups vary in terms of the number of group-members, the length and variability of the sequences, the separation from the nearest non-member sequences etc. So we can expect that using a top list of a given length may influence the ROCn value.

Let's take the globin superfamily a.1.1 of the SCOP95 data set as an example, and use the Smith–Waterman algorithm to compare sequence similarities. This superfamily has 103 members, and we will define a classification task wherein family a.1.1.1 (Globine-like, 5 members) will be the test group. We will calculate AUC in a group-wise scenario, using 10, 50 negatives (which corresponds to the generally used ROC10 and ROC50 scenarios), or the full dataset in the analysis. The results in Table 1 indeed show that the values are strongly dependent on the number of negatives taken into the consideration.

The phenomenon is shown graphically for the entire SCOP95 dataset in which the values represent the average of 246 classification tasks (Fig. 2). In this case we calculated ROC AUC in three different ways, i.e. using the element-wise scenario, the group-wise scenario with and without likelihood-ratio scoring. The values are plotted against the size of the negative set expressed in multiples of the positive set. This was carried out by preparing the ranked list in the usual way and truncating it from the top when the necessary number of negatives was reached. It is conspicuous that the AUC values calculated by any of the three methods show a steeply increasing tendency if the number of negatives is below the size of the positive group. The methods using simple scoring (i.e. when the Smith–Waterman score was used for ROC analysis either in the group-wise or in the element-wise scenario) continue an increasing tendency above this threshold, but apparent that the curve of likelihood-scoring (in a group-wise scenario) does not, its

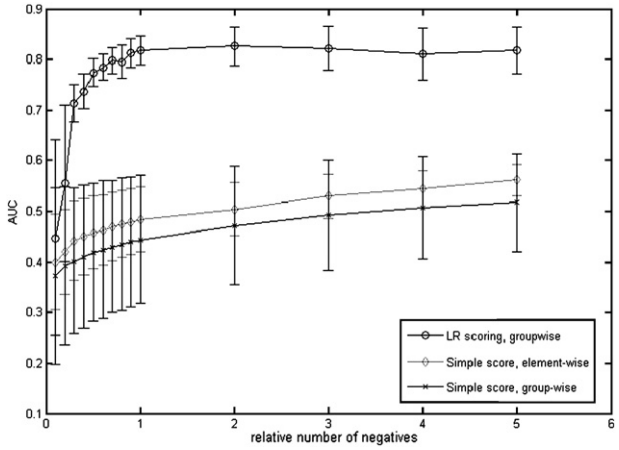

Fig. 2. Dependence of the ROC AUC values on the size of the negative set. Average AUCs were calculated for all the 246 classifications tasks within the SCOP95 dataset. The error bars indicate the average standard deviations. The curves represent different methods of calculation as indicated in the inset and described in Materials and methods. Note that the group-wise scenario with likelihood-ratio scoring gives values that are independent of the size of the negative set while the results of the others show an increasing tendency and have higher standard deviation values (indicated by the error bars).

value is seemingly independent of the number of negative samples in the top list. In addition, the latter method has substantially lower standard deviation values than the other two methods. In principle, the comparison of the scoring methods should not depend on the dataset, and in fact we see that the ranking order of the methods is consistently the same as we reach values of $N_r > 1$. On the other hand, the curves are crossing each other below $N_r = 1.00$, which shows that the comparison of the methods is not consistent if

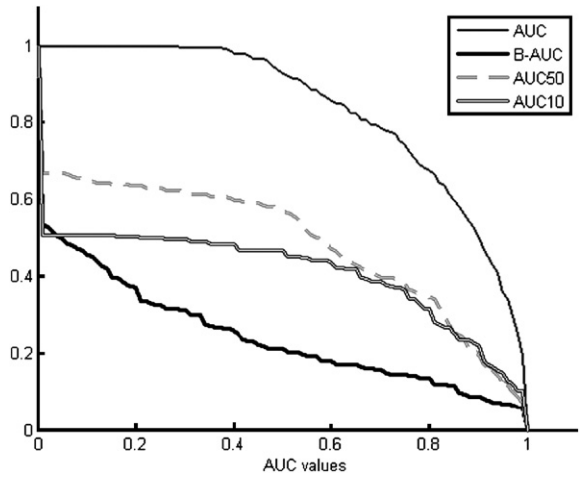

Fig. 3. Cumulative AUC curves for various calculation/scenarios. The calculations were done on the super families of the SCOP95 database (PCB0001, see Materials and methods), using various strategies for top list-restriction. The cumulative AUC curves plot the number of queries or groups (Y-axis) that exceed the AUC value indicated on the X axis. For uniformity, we normalized the Y values to 1.00 by dividing them by the total number of queries or groups, respectively. AUC indicates calculation on the entire dataset, AUC50 and AUC10 indicate calculations based on truncated toplists as suggested by Gribskov and Robinson [5], B-AUC indicates calculation according to the present BAROC protocol.In this representation the curves running higher indicate a better performance, which means that AUC on the full dataset gives higher scores than all the other methods, and the balanced ROC gives the lowest scores.

Please cite this article as: Busa-Fekete R et al. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. J Biochem Biophys Methods (2007), doi:10.1016/j.jbbm.2007.06.003

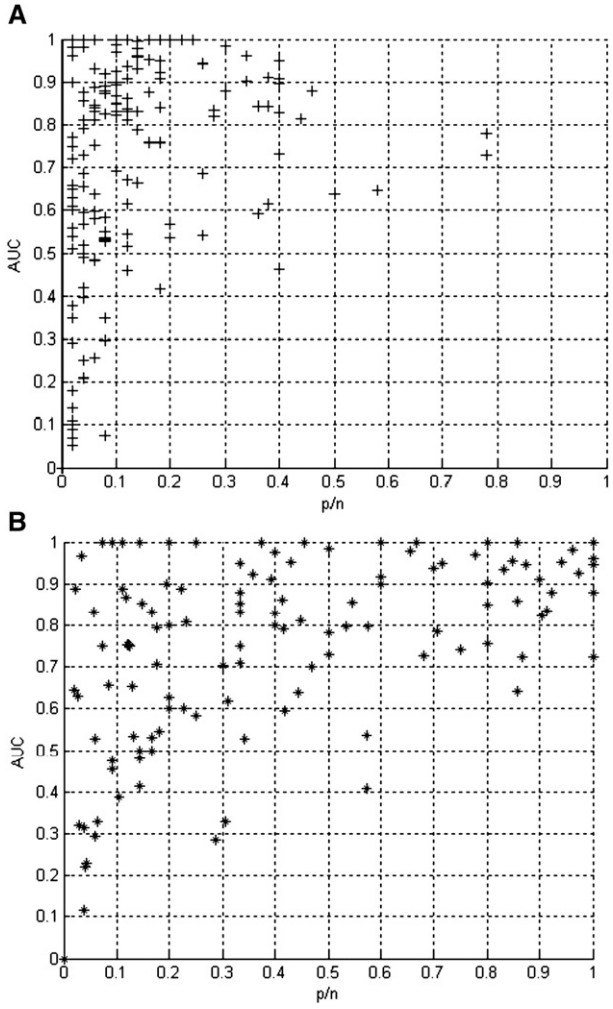

Fig. 4. Dependence of AUC values on the positive/negative ratio. The AUC values were calculated for the 246 groups of the SCOP95 database using a group-wise scenario with supervised cross validation as described [6,7,10,11]. The top panel shows the calculation for AUC50, in which the top list of each group contains exactly 50 negative samples. The lower panel shows the balanced ROC, in which the top list contains as many negatives as there are positives in the calculation. (N_r=1.00). The data points are more spread.

shorter top lists are used. Taken together, the results confirm the dependence of the values on the number of the negative samples. And since fixing the same size of the negative set for all of the classes will place the different classes at different points within this curve, we can appreciate that there will be a differential bias within the various groups.

One of the underlying reasons is that the top lists of a given length may contain a strongly varying number of positives and negatives, so the positive/negative ratio changes as we consider longer and longer top lists. On the other hand, selecting a number of negatives that is proportional to the size of the positive group will adjust the positive/negative ratio to a level roughly balanced with respect to the group size. With this strategy, the experimenter can be more certain that the AUC differences observed between classes are not class-imbalance artifacts but indicators of quality differences between classes. We propose to use a number of negatives that is equal to or double of the number of the positives participating in the ranking. Fig. 3 shows that the balanced ROC analysis calculated in this manner is a more stringent test than AUC50 or AUCC10, at least for the SCOP95 dataset in which there are many relatively small groups.

The balancing effect is clearly shown in Fig. 4 where we plotted the AUC value for 246 groups against the *p*/*n* ratio. In the case of ROC50, the *p*/*n* ratio is way below one (A, top). In the case of the balanced ROC (B, bottom), *p*/*n* gets near one for a number of the groups, but it can be lower in the case of distant similarities since in those cases not all the positives will show up in the top list. In the present dataset we see many such cases since the SCOP dataset is difficult for a sequence comparison method such as Smith Waterman. Simply put, regions in the BAROC scatter diagram allows one to distinguish various classes within a database. High AUC, high *p*/*n* ratio indicates tight groups, with high similarities between the members and well separated from the rest. High AUC low *p*/*n* is characteristic of less well-separated or less tight groups in which the within-group similarity is not sufficiently strong. Finally, low AUC, low *p*/*n* ratio is indicative of problem groups that are difficult to predict.

4. Discussion

Generally speaking, for statistical evaluation such as a ROC AUC calculation, it is recommended to have the same number of positive and negative samples [1,2]. This condition is never met in bioinformatics databases, we have much less positives than negatives. A correct solution would be to randomly select equal numbers from the two classes and to construct many classifiers for all cases, which is clearly too time consuming given the number of classes and data. Instead, bioinformaticians resort to the truncation of the top lists, which – as we saw in Table 1 – does not necessarily helps one to decide whether or not a low AUC value is indicative of a problematic object class or it is a class-imbalance artifact. As a way out, we propose the balanced ROC scenario which adjusts the number of negatives to the level of the size of the positive set. In this manner, classes, where the positives are within the "balanced" top list, will have a high *p*/*n* value as well as a high AUC. These classes are the easy cases for which the comparison measure (in our case the Smith–Waterman similarity score) works efficiently. In the case of difficult classes, we will have fewer positives in the top list. So we can detect the problem classes, using a scatter plot like the one shown in Fig. 4, where they will be in the region of low AUC and low *p*/*n* values. In this manner we can make use of the reliability of AUC calculation without having to construct a classifier for all the groups. Likelihood-ratio scoring (Eq. (1)) provides a further tool for getting rid of the class-imbalance bias, and it can be efficiently used in the BAROC scenario.

Finally we mention that the use of the method was illustrated here on 246 sequence classification tasks taken from the SCOP database. We tested the method on the other sequence and 3D classification tasks included in the Protein Classification Benchmark Collection [10], with identical results (data not shown) which gives us hope that this protocol will be applicable to other classification methods that rank the objects according to a variable that characterizes class-membership.

Please cite this article as: Busa-Fekete R et al. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. J Biochem Biophys Methods (2007), doi:10.1016/j.jbbm.2007.06.003

5. Simplified description of the method

We propose a ROC analysis protocol that makes it possible to single out classes in a database that are likely to be difficult to predict. The method, termed Balanced ROC (BAROC), consists of calculating an AUC value for a ranked top list which truncated so as to contain as many (or twice as many) negative objects as there are positive objects in the entire analysis. In this manner each class will be analyzed with a top list whose length depends on the size of the class. The difficult groups can then be identified by their low AUC values and/or their low positive/negative ratio within the top list. The identification is helped by a scatter plot of AUC vs. positive/negative ratio, as well as by the use of a likelihood-ratio scoring scheme (Eq. (1)), that can be efficiently used in the BAROC protocol instead of the simple similarity scores.

Acknowledgements

Work at ICGEB was supported in part by grants from the Ministero dell'Università e della Ricerca (D.D. 2187, FIRB 2003 (art. 8), "Laboratorio Internazionale di Bioinformatica"). A. Kocsor was supported by the János Bolyai fellowship of the Hungarian Academy of Sciences

References

[1] Egan JP. Signal detection theory and ROC analysis. New York; 1975.
[2] Duda RO, Hart PE, Stork DG. Pattern classification. New York: John Wiley & Sons; 2000.
[3] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[4] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7.
[5] Gribskov M, Robinson NL. Use of Receiver Operating Characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 1996;20:25–33.
[6] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Comput Biol 2000;7:95–114.
[7] Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol 2003;10:857–68.
[8] Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–9.
[9] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–7.
[10] Sonego P, Pacurar M, Dhir S, Kertész Farkas A, Kocsor A, Gáspár Z, et al. A protein classification benchmark collection for machine learning. Nucleic Acids Res 2007;35:232–6.
[11] A. Kertész Farkas, S. Dhir, P. Sonego, M. Pacurar, S. Netotea, H. Mijveen, A. Kuzniar, J.A.M. Leunissen, A. Kocsor, and S. Pongor, Benchmarking protein classification algorithms by supervised crossvalidation. J Biochem Biophys Meth (in press).
[12] Kaján L, Kertész-Farkas A, Franklin D, Ivanova N, Kocsor A, Pongor S. Application of a simple likelihood ratio approximant to protein sequence classification. Bioinformatics 2006;22:2865–9.
[13] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Second Edition. San Francisco, CA: Morgan Kaufmann; 2005.

Please cite this article as: Busa-Fekete R et al. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. J Biochem Biophys Methods (2007), doi:10.1016/j.jbbm.2007.06.003