■ 1785

Algorithm for Predicting Compound Protein Interaction Using Tanimoto Similarity and Klekota-roth Fingerprint

Isnan Mulia*[1], Wisnu Ananta Kusuma[2], Farit Mochamad Afendi[3]
[1,2]Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, Indonesia
[3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University,
Indonesia
*Corresponding author, e-mail: isnanmulia@apps.ipb.ac.id

Abstract
This research aimed to develop a method for predicting interaction between chemical compounds contained in herbs and proteins related to particular disease. The algorithm of this method is based on binary local models algorithm, with protein similarity section is omitted. Klekota-Roth fingerprint is used for the compound's representation. In the development process of the method, three similarity functions are compared: Tanimoto, Cosine, and Dice. Youden's index is used to evaluate optimum threshold value. The result showed that Tanimoto similarity function yielded higher similarity values and higher AUC value than those of the other two functions. Moreover, the optimum threshold value obtained is 0.65. Therefore, Tanimoto similarity function and threshold value 0.65 are selected to be used on the prediction method. The average evaluation accuracy of the developed algorithm is only about 50%. The low accuracy value is allegedly caused by the only use of compound similarity on the prediction method, without including the protein similarity.

Keywords: Tanimoto similarity, Jamu, Klekota-roth fingerprint, Network pharmacology

1. Introduction

Network pharmacology is a new approach in pharmacology. This approach is a development from biological network and polypharmacology, which is used in drug discovery. Network pharmacology creates relational network between compunds contained in particular herb with proteins related with particular disease [1]. This approach has been implemented to build network pharmacology for triphala formula of ayurveda [2], reveal the molecular mechanism of Qing-Luo-Yin formula of Traditional Chinese Medicine (TCM) [3], reveal candidate drug targets related to Complex Kidney Disease (CKD) from Bu-shen-Huo-xue (BSHX) formula of TCM [4], and build C2Maps platform [5].

Jamu is one of traditional herb medicine originated from Indonesia. Jamu formula consists of natural ingredients, i.e. root, bark, leaf, fruit, and flower [6]. Just like ayurveda and TCM, the network pharmacology approach can be implemented on jamu formula. Afendi et al. [7] investigate the relationship between Indonesian herbs with jamu efficacy using biplot configuration, with data from Indonesia's National Agency of Drug and Food Control. Fitriawan et al. [8] and Ristyawan [9] developed classification system of jamu efficacy, each of which uses support vector machine and voting feature interval 5 method, based on binary data used by Afendi et al. [7]. Qomariasih [10] determined active compounds from 4 Indonesian herbs, i.e. Zingiber officinale, Blumea Balsamifera, Tinospora crispa, and Momordica charantia L., using network pharmacology and simultaneous clustering analysis, under context for medication of disease diabetes mellitus type 2. Ochieng et al. [11] developed an approach to investigate the pharmacological mechanisms of the four Indonesian herbs mentioned before. Besides on its formula, research about jamu is also performed on the plant's physical appearances. Lantana et al. [12] proposed a new method for estimating spectral reflectance from jati belanda (Guazuma ulmifolia) leaf based on digital image color, which then can be used to estimate chemical compounds contained in the leaf. Meanwhile, Karlitasari et al. [13] developed a mobile application for visualizing Indonesia medicine in 3D.

Interaction between compounds and proteins that haven't been discovered yet can be predicted using several prediction methods. Liu et al. [14] filtered highly credible negative in silico samples from compound protein interaction, to be used as training data for predicting compound protein interaction using support vector machine as the predictor. Bleakley and Yamanishi [15] proposed new inference method for predicting compound protein interaction, namely bipartite local models. The method works by predicting target proteins from given compound, and then predicting compounds which the target is the given protein. The result of the prediction is the combination of the previous two predictions.

Compound and protein data used as the object of this research can be connected each other if and only if there's interaction data between two entities. If the interaction data is not available, the interaction of compound and protein can be predicted using similarity data between compounds and proteins, with the help of statistical methods and machine learning. In this research, we developed a method to predict the interaction between compounds from jamu herbs and proteins corresponding with a disease, using available compounds and proteins data. The result of these predictions is expected to be used as a reference for further testing of the interaction of the compound and the protein in the laboratory.

2. The Proposed Algorithm

In this research, we proposed a new method for predicting compound protein interaction. The developed prediction method receives compound and protein which are the interaction is to be predicted as input. The algorithm of this method is developed based on binary local models algorithm developed by Bleakley and Yamanishi [15], with protein similarity section is omitted. We used compound fingerprint data [16], i.e. Klekota-Roth fingerprint [17], as a representation of every compounds used. Figure 1 shows the pseudocode of developed prediction method in design phase, while Figure 3 shows the complete pseudocode as the result of the development.

```
function predictCPI()

1.   input: inp_compound, inp_protein
2.   threshold ← t
3.   max_similarity ← 0
4.   check ← σ_compound = inp_compound AND protein = inp_protein (table_compound_protein)
5.   if (check is not empty):
6.       return 1
7.   else
8.       inp_cmpd ← π_compound, fingeprint_kr (σ_compound = inp_compound (table_compound))
9.       table_join ← table_compound C ⋈ table_compound_protein CP
10.      cmpds ←
         σ_fingerprint_kr IS NOT NULL AND protein = inp_protein AND compound <> inp_compound (table_join)
11.      sim_cmpds ← π_compound, fingeprint_kr (cmpds)
12.      for (sim_cmpd on sim_cmpds):
13.          similarity ← similarity_function(inp_cmpd, sim_cmpd)
14.          max_similarity ← max(similarity, max_similarity)
15.      end for
16.      if (max_similarity >= threshold):
17.          return 1
18.      else
19.          return 0
20.      end if
21.  end if
```

Figure 1. The pseudocode of developed prediction method in design phase

The first steps done by prediction method is determine the threshold value and set similarity value to zero. Next, the method performs a checking, whether compound protein interaction data is already available in the database or not (line 4). If the interaction data is

available, the prediction method returns true, and the prediction is not performed. Otherwise, the prediction needs to be performed.

On line 8, the prediction method retrieves detail data of the input compound from the compound table. On line 9-11, the prediction method retrieves list of compounds that target the input protein. On line 12-15, a looping is performed to determine similarity value between each compound from the list and the input compound, and compare obtained similarity value with current maximum similarity value. After the looping performed and the maximum similarity value has been obtained, this value is compared with the threshold. If similarity value is higher than threshold, the interaction between input compound and input protein is predicted to exist, and the prediction method returns true. Otherwise, the interaction between input compound and input protein is predicted to not exist, and the prediction method returns false.

3. Research Method
3.1. Datasets

Datasets used in this research were obtained from two sources. The first dataset is obtained from Qomariasih's research [10]. The second dataset consists of data that have been collected before, and is a part of a research that is currently running. The two datasets consist of compound data, protein data, and compound-protein interaction data, which is collected from several webservers, like PubChem [18] dan Uniprot [19]. Dataset 1 is used to evaluate optimum similarity function and optimum threshold, while dataset 2 is used to evaluate the prediction method.

In brief, the data preprocessing steps done are generating Klekota-Roth fingerprint data for every compounds in two datasets, choosing and checking data, and generating data for negative compound protein interaction on dataset 1. The part of compound data that is used to determine the degree of similarity between compounds is fingerprint. Fingerprint from a compound is a simplification or abstraction from compound's structure, denoted with a string of bits of certain length [16]. In this research, Klekota-Roth fingerprint [17] is used. Klekota-Roth fingerprint's string of a compound consists of 4860 bits. The generation of Klekota-Roth fingerprint is done using Chemistry Development Kit (CDK) library [20], which available on web application ChemDes [21], with SMILES data as the input. SMILES (Simplified Molecular Input Line System) is a chemical notation language that denotes molecule structure as a graph that is essentially the two-dimensional valence-oriented picture chemists draw to describe a molecule. This language is designed specifically for computer usage by chemists [22].

Compound protein interaction data that are available on dataset 1 and 2 could be considered as interaction data with class label "positive", because the interaction data show that there are interactions for some compound and protein pairs. In order to make the dataset to be "balanced", in the meaning of the dataset has interaction data with label "positive" and "negative", it is required to generate interaction data with label "negative". In this context, interaction data with label "negative" can be interpreted as compound protein interaction data that is the validity is unknown. The generation of negative interaction data is only done for dataset 1.

3.2. Prediction Method

In this research, prediction method used is a simple prediction method, based on simple calculation. This method notices similarity between input compound and another compound from the dataset. It is based on central premise on medicinal chemistry that structurally similar molecules have similar biological activities [23].

The concept of the developed prediction method follow the concept of binary local models algorithm [15]. The process of predicting compound protein interaction is done from two sides, i.e. based on input compound and input protein. From the compound side, the classification rule for grouping protein data, both proteins that is targeted by input compound or proteins that is not targeted by input compound, is searched, using protein's genomic sequence data. From the protein side, the classification rule for grouping compound data, both compounds that targets input protein or compounds that is not target input protein, is searched, using compound's chemical structure data. After the two classification rule is obtained, the prediction of compound protein interaction can be predicted from two sides, and the prediction result is aggregated.

There are two important points on the development of the prediction method, i.e. similarity function and threshold value used. The similarity function is used to determine how close are the input compound and the compounds from the dataset that have interaction with the input protein. The threshold value is used as a boundary between positive and negative prediction result, for input compound and input protein. On the development stage, similarity function and threshold value have not been determined, because those values will be evaluated and determined through next steps.

3.3. Similarity Function

Similarity function is a function to measure the similarity, or the closeness, of two objects. The way to implement this function depends on the data to be used. In this research, the data that will be used with the similarity function is Klekota-Roth fingerprint, which is in the form of a string with length of 4860 bits. Therefore, the binary similarity function is selected.

The similarity of two binary strings can be described as follows. Suppose we have two binary strings **x** and **y**, each of which consists of p variables with value 0 or 1. The common association coefficients are calculated, and the result is shown on Table 1, where a, b, c, and d are the frequencies of the events (x=1 and y=1), (x=1 and y=0), (x=0 and y=1), and (x=0 and y=0), respectively, in the pair of binary vectors describing the two objects; p is the total number of variables, equal to $a+b+c+d$, which is the length of each binary vector [24].

Table 1. Frequency Table of Four Combinations for Two Possible Binary Variables [24]

	$y = 1$	$y = 0$	
$x = 1$	a	b	$a + b$
$x = 0$	c	d	$c + d$
	$a + c$	$b + d$	p

There are three similarity functions used, i.e. Tanimoto, Cosine, and Dice similarity function. This three function, according to Bajusz *et al.* [25], are the best metrics among eight metrics tested. On the evaluation process, the similarity between compounds will be calculated using three similarity functions. Afterwards, one of the three function that produces the highest similarity value for every appropriate compound pairs will be chosen, and will be implemented on the prediction method. The formula of similarity function Tanimoto, Cosine, and Dice is shown in Equation 1, 2, and 3, respectively.

$$\text{Tanimoto similarity: } S_{X,Y} = \frac{a}{a + b + c} \tag{1}$$

$$\text{Cosine similarity: } S_{X,Y} = \frac{a}{\left((a + b)(a + c)\right)^{1/2}} \tag{2}$$

$$\text{Dice similarity: } S_{X,Y} = \frac{2a}{2a + b + c} \tag{3}$$

$$\text{Sensitivity} = \frac{\text{\# of true positive}}{\text{\# of true positive} + \text{\# of false negative}} \tag{4}$$

$$\text{Specificity} = \frac{\text{\# of true negative}}{\text{\# of true negative} + \text{\# of false positive}} \tag{5}$$

$$\text{Youden's index} = \text{sensitivity} + \text{specificity} - 1 \tag{6}$$

4. Result and Discussion
4.1. Datasets

Dataset 1 consists of 55 compound data, 478 protein data, and 3059 compound protein interaction data. Meanwhile, dataset 2 consists of 12755 compound data, 2952 protein data, and 175071 compound protein interaction data. The first step of data preprocessing is generating fingerprint data for every compounds. After the generation, it is known that not all of

the compounds' fingerprint can be generated, due to the lack of SMILES data or fingerprint data generation failure using available SMILES data.

The next step is choosing and checking data validity, mainly for the compound protein interaction data. This is done so that the amount of data to be processed is not too large and the data meet the required criteria. The criteria are as follows: the compound is registered on Pubchem [18] and has fingerprint; and the protein is registered on Uniprot [19] as one of human's (Homo sapiens) protein.

The next step is generating data with label "negative" to dataset 1. The procedure of the generation is as follows. First, enlist all compound protein pairs from the dataset that are not in the list of positive compound protein interaction. Then, choose N pairs of compound protein data from the list randomly, with N is the number of compound protein interaction data with label "positive".

After the data preprocessing, there are 55 compound data, 478 protein data, and 6066 compound protein interaction data for dataset 1. The compound protein interaction data in dataset 1 consists of 3033 compound protein interaction data with label positive and 3033 compound protein interaction data with label negative. For dataset 2, there are 6736 compound data, 1012 protein data, and 25123 compound protein interaction data.

4.2. Prediction Method

The proposed prediction method used in this research is already discussed on section 2. As mentioned before, the prediction method only uses compound similarity, and omits protein similarity. At the time of research, we haven't found appropriate metric to measure protein similarity, either structurally or functionally. On the pseudocode shown in **Error! Reference source not found.**, the similarity function and optimum threshold value used has not specified yet, because the two items will be evaluated on the next step. The similarity function and optimum threshold will be used on the program code in line 13 and 2.

4.3. Similarity Function & Threshold Evaluation Result

On this stage, the prediction algorithm is implemented in the programming language Python 3. The three similarity functions to be evaluated are also included in the code. After the code implementation, the code is executed using compound protein interaction data from dataset 1, assumed that the compound protein interaction is unknown.

After the execution and similarity values are obtained, a number of 21 threshold candidate values are generated in the range [0, 1] with the difference between values is 0.05. Afterwards, for every threshold value, the similarity functions are filtered using criteria "the similarity value is less than or equal than threshold value", aggregated, and then calculated. Some values calculated here are true negatives, false negatives, false positives, true negatives, sensitivity, specificity, Youden's index, and accuracy. The evaluation of similarity function and optimal threshold is done at the same time. It is possible, because the two evaluations use the same aggregate data. Accuracy data and area under ROC curve (AUC) values is used for evaluating similarity function. Sensitivity, specificity, and Youden's index data is used for evaluating optimal threshold. The aggregated data is attached on Supplementary File 1.

From 6066 compound protein interaction data, it is known that the highest accuracy value is achieved by similarity function Tanimoto on threshold candidate value of 0.65, with value of 0.583. On the evaluation of Youden's index, the highest Youden's index value is also achieved by similarity function Tanimoto on threshold candidate value of 0.65, with value of 0.166. ROC curve for three similarity functions is then drawn and AUC value for each similarity functions is calculated, using the trapezoidal approximation. The ROC curve for three similarity functions is shown on Figure 2. After the AUC value obtained, it is known that Tanimoto similarity function has the highest AUC value among the three similarity functions, with value of 0.555. From this result, it can be inferred that the optimum similarity value and the optimum threshold value is Tanimoto similarity function and 0.65. After this step, the design of the prediction method is as shown on Figure 3.
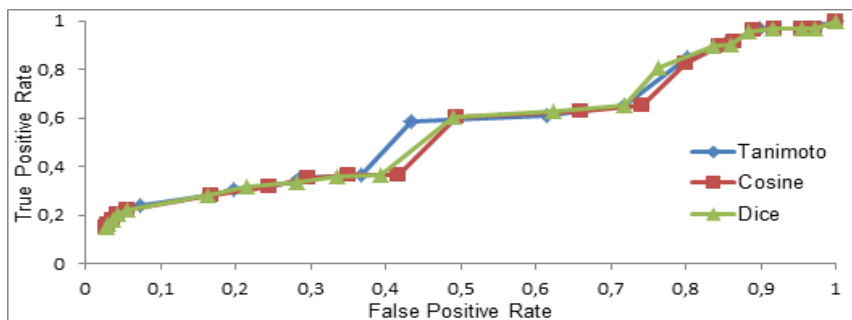
Figure 2. ROC curve for three similarity functions

```
function predictCPI()

1.    input: inp_compound, inp_protein
2.    threshold ← 0.65
3.    max_similarity ← 0
4.    check ← σ_{compound = inp_compound AND protein = inp_protein}(table_compound_protein)
5.    if (check is not empty):
6.         return 1
7.    else
8.         inp_cmpd ← π_{compound, fingeprint_kr} (σ_{compound = inp_compound}(table_compound))
9.         table_join ← table_compound C ⋈ table_compound_protein CP
10.        cmpds ←
           σ_{fingerprint_kr IS NOT NULL AND protein = inp_protein AND compound <> inp_compound}(table_join)
11.        sim_cmpds ← π_{compound, fingeprint_kr} (cmpds)
12.        for (sim_cmpd on sim_cmpds):
13.             similarity ← tanimoto_similarity(inp_cmpd, sim_cmpd)
14.             max_similarity ← max(similarity, max_similarity)
15.        end for
16.        if (max_similarity >= threshold):
17.             return 1
18.        else
19.             return 0
20.        end if
21.   end if
```

Figure 3. The complete pseudocode of developed prediction method

4.4. Prediction Method Evaluation Result

In this stage, the amount of data used to evaluate the prediction method from every subset is 22611 compound protein interaction data. The value is a rounding from 90% of the amount of compound protein interaction data from dataset 2. The generation of the sub datasets is done 10 times, where in every repetition an amount of 22611 interaction data is chosen randomly, and stored in different tables.

After the sub datasets is generated, the evaluation of prediction method is done. On the evaluation result, it is known that the minimum accuracy obtained is 50.374% and the maximum accuracy obtained is 50.683%, whereas the average accuracy is 50.495%. The low accuracy value is allegedly caused by the prediction method that only use the compound similarity to predict the compound protein interaction, without including the protein similarity. The evaluation result of the prediction method is shown in Table 2.

Table 2. Result of Prediction Method Evaluation

Sub Dataset	# of True Positive	# of False Negative	Sensitivity	Accuracy
1	11395	11216	0.50396	0.504
2	11426	11185	0.50533	0.5053
3	11431	11180	0.50555	0.5056
4	11435	11176	0.50573	0.5057
5	11448	11163	0.5063	0.5063
6	11460	11151	0.50683	0.5068
7	11390	11221	0.50374	0.5037
8	11401	11210	0.50422	0.5042
9	11400	11211	0.50418	0.5042
10	11390	11221	0.50374	0.5037
			Minimum Accuracy	0.50374
			Maximum Accuracy	0.50683
			Average Accuracy	0.504958

5. Conclusion

In this research, we developed a method for predicting the interaction between compound and protein. The prediction method uses Tanimoto similarity function and threshold value of 0.65. The evaluation of the prediction method yield average prediction accuracy value of 50,5%. The result of this research is not reliable enough. Therefore, we suggest some points for the next research, as follows:

a. To use machine learning approaches to build prediction algorithm, so that the prediction algorithm can capture the pattern of compound protein interaction data that is being predicted.

b. To include the protein similarity measurement in the calculation to predict the compound protein interaction, so the prediction result is viewed not only from the compound's side.

References

[1] Hopkins AL. Network pharmacology. *Nat Biotechnol.* 2007; 25(10): 1110–1

[2] Chandran U, Mehendale N, Tillu G, Patwardhan B. *Network pharmacology: An emerging tool for natural product drug discovery and evidence-based Ayurveda.* Proc Indian Natl Sci Acad. 2015; 81(3): 561–8.

[3] Zhang B, Wang X, Li S. An Integrative Platform of TCM Network Pharmacology and Its Application on a Herbal Formula, Qing-Luo-Yin. *Evidence-Based Complement Altern Med.* 2013; 1–12.

[4] Shi S, Cai Y, Cai X, et al. A Network Pharmacology Approach to Understanding the Mechanisms of Action of Traditional Medicine: Bushenhuoxue Formula for Treatment of Chronic Kidney Disease. de Carvalho DP, *editor. PLoS One.* 2014; 9(3): e89123.

[5] Huang H, Wu X, Pandey R, et al. C2Maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics.* 2012; 13(Suppl 6): S17.

[6] Riswan S, Sangat-Roemantyo H. Jamu as Traditional Medicine in Java, Indonesia. *South Pacific Study.* 2002; 23(1): 1–10.

[7] Afendi FM, Darusman LK, Hirai A, et al. *System Biology Approach for Elucidating the Relationship Between Indonesian Herbal Plants and the Efficacy of Jamu.* In: 2010 IEEE International Conference on Data Mining Workshops. IEEE. 2010. p. 661–8.

[8] Fitriawan A, Kusuma WA, Heryanto R. *A Classification System for Jamu Efficacy Based on Formula Using Support Vector Machine.* In: 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS). Bali: IEEE. 2013. p. 291–5.

[9] Ristyawan Y. Sistem Klasifikasi Khasiat Formula Jamu dengan Metode Voting Feature Interval 5. Undergraduate Thesis. Bogor: Bogor Agricultural University; 2014.

[10] Qomariasih N. Analisis Gerombol Simultan dan Jejaring Farmakologi pada Penentuan Senyawa Aktif Jamu Anti Diabetes Tipe 2. Thesis. Bogor: Bogor Agricultural University; 2015.

[11] Ochieng PJ, Kusuma WA, Rafi M, Sumaryada T. Deciphering the Action Mechanism of Indonesia Herbal Decoction in the Treatment of Type II Diabetes Using A Network Pharmacology Approach. *Int J Pharm Pharm Sci.* 2017; 9(3): 243.

[12] Lantana DA, Herdiyeni Y, Batubara I. Estimated Spectral Reflectances of Jati Belanda Leaf Using Fourier Transformation. *TELKOMNIKA Indonesian Journal of Electrical Engineering and Computer Science.* 2014; 12(9): 6958–6964.

[13] Karlitasari L, Ismanto, Putra AP. Android-Based 3D Visualization of Indonesia Medicinal Plants. *TELKOMNIKA Indonesian Journal of Electrical Engineering and Computer Science.* 2017; 5(1): 229–233.

[14] Liu H, Sun J, Guan J, Zheng J, Zhou S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*. 2015; 31(12): i221–9.

[15] Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*. 2009; 25(18): 2397–403.

[16] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. *Methods*. 2015; 71(C): 58–63.

[17] Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics*. 2008; 24(21): 2518–25.

[18] Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. *Drug Discov Today*. 2010; 15(23–24): 1052–7.

[19] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2012; 40(D1): D71–5.

[20] [Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci*. 2003; 43(2): 493–500.

[21] Dong J, Cao D-S, Miao H-Y, et al. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J Cheminform*. 2015; 7(1): 60.

[22] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988; 28(1): 31–6.

[23] Martin YC, Kofron JL, Traphagen LM. Do Structurally Similar Molecules Have Similar Biological Activity? *J Med Chem*. 2002; 45(19): 4350–8.

[24] Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J Chem Inf Model*. 2012; 52(11): 2884–901.

[25] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform*. 2015; 7(1): 20.

[26] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ Anaesthesia, Crit Care Pain*. 2008; 8(6): 221–3.

[27] Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. *Crit Care*. 2004; 8(6): 508.