EVALUATION OF QUANTITATIVE STRUCTURE–ACTIVITY PREDICTIONS. COMPARISON OF THE PREDICTIVE POWER OF AN ARTIFICIAL INTELLIGENCE SYSTEM WITH HUMAN EXPERTS

Gilles KLOPMAN* and Istvan KOLOSSVARY☆
Department of Chemistry, Case Western Reserve University, Cleveland, Ohio 44106, USA

Received 12 February 1990

Abstract

In this paper, we present three new mathematical techniques for evaluating the predictive skills of structure–activity experts. The question addressed in this paper is how to evaluate the predictive ability of structure–activity experts in identifying the most active compounds of a set of drug candidates. The three proposed mathematical techniques are based on the Phi-square Distance, the Rank Comparison, and the Shuffle method, respectively. They have been used to evaluate the performance of a new computer system and three human experts in predicting the antibacterial potencies of a series of chemical compounds in five different biological tests. The expert system, an artificial intelligence structure–activity program called MULTICASE, performed significantly better than one of the human experts and somewhat better than the other two.

Keywords

Prediction, validation, evaluation, ranking, tie, artificial intelligence, expert system, structure–activity relationships.

1. Introduction

In designing new drugs, it is common to venture guesses as to the biological activity of the molecules planned for synthesis. These guesses are implicitly made by all chemists who design a new structure, and explicitly made by structure–activity experts who may rank for priority synthesis a set of potential candidates for biological evaluation. However, rarely is a process generated to evaluate how relevant these guesses were and very few methodologies are known to help evaluate the predictions, once the compounds are made and tested.

*To whom all correspondence should be addressed.
☆On leave from the Department of General and Analytical Chemistry, Technical University of Budapest, Szt. Gellert ter 4, 1111 Budapest, Hungary.

A number of groups have devised computer systems aimed at automating the prediction of biological activity of compounds, and it occurred to us that good validation techniques are necessary if these techniques are to be trusted and used to support human intuition. In the course of this study, we found that humans are often prone to self-indulgence about their own ability to predict. In the same vein, we found that automated techniques are not always as good as hoped when they are used to predict previously unknown facts, even when they have been validated within the domain of knowledge at the time.

Given a database of chemical structures and their associated activities (e.g. inhibitory, antibacterial, or pesticide potencies), one can divide the database randomly into a training set and a test set. Using a quantitative structure–activity relationship (QSAR) model, which in most cases is a linear combination of some molecular descriptors (such as molecular volume, topological indices derived from the connectivity graph of a molecular structure, some kind of physico-chemical properties, the presence of particular molecular substructures, etc.), a model can be developed from the training set and used to predict the activity of the compounds of the test set. This generally means that the model parameters, i.e. the linear coefficients of the selected molecular descriptors in the QSAR equation, which are determined from the training set, are then validated on the test set. The random partitioning of the database into a training set and a test set, as well as the corresponding model identification and validation, can be carried out repeatedly to check the stability of the QSAR model, i.e. to see how much the goodness of the predictions on the test set varies with the different random partitionings. Classical methods such as the sample F-test, chi-square, or non-parametric tests, as well as more recently developed techniques such as boot-strapping and cross-validation, are at the user's disposal. It is impossible to reference the very large amount of work published on this problem in the statistical literature. References [1–5], however, should help the interested reader to find relevant sources.

At this point, we wish to introduce the sharp distinction that exists between *validation* and *evaluation*. The *validation* of a QSAR model, as outlined in the previous paragraph, is based on a number of analyses and gives general credibility to the model. *Evaluation*, on the other hand, tells one how well the model is performing in a specific case, usually consisting of molecules unknown to the expert at the time of the development of the method. It should be noted that any kind of validation is meaningless without good evaluation. Indeed, if the knowledge of the activity of some molecules contributed to the *selection* of the methodology, or its *parameters*, then the activity of these molecules should not be used to validate the methodology, since this will provide no clue as to the generality of the method outside its learning domain. In a way, one may relate validation to interpolation, and evaluation to extrapolation. Our purpose here is not to discuss *validation* but to *evaluate* how well various techniques perform in a predictive mode, after they have been "validated".

In this paper, we thus attempt to find an answer to the question: How to *evaluate* the performance of a structure–activity expert, automated or human, in predicting the activities of a set of *new* chemical compounds? Or, more precisely, since one is

generally more concerned with the active compounds than with the inactive compounds, the question is how to evaluate the performance of a QSAR technique in identifying the most active compounds, i.e. the best candidates for further study. The problem is compounded by the fact that there is no advance knowledge of how many of the test compounds are in the "active" category.

An interesting analogy to the problem is weather forecasting where, for instance, one might wish to compare the ability of meteorologists to predict the sunny days of the next month. There is no advance knowledge of which days will be sunny nor, for that matter, of how many days, if any, will be sunny. In this paper, we tried to answer this kind of question by introducing new mathematical techniques for the evaluation of SAR predictions. Three evaluation methods based on the Phi-square, the Rank Comparison, and the Shuffle method will be presented and compared in an experiment where the predictions made by an expert system and by three human volunteer experts have been evaluated.

2. Methods

2.1. DEFINITIONS

The problem, as we see it, can be stated as follows. An expert studies a learning set of compounds whose structure and experimental activity is given. Once the learning set has been rationalized and possibly validated, the expert faces a set of N test compounds. The experimental activity of the N compounds is known but not made available to the expert. The expert predicts the activity of the N test compounds, which can then be ranked by ascending activity as measured and as predicted by the expert. If the measured and the predicted rankings, as well as the level of the measured and the predicted activities are identical, then a perfect prediction would have been made. However, this is seldom the case and the question is how good is the prediction and, in general, how can we measure the difference between two different rankings?

2.1.1. Phi-square Distance

We explored the possibility of using a slightly modified chi-square test [1], which we call Phi-square Distance, to measure how well the prediction of the active test compounds match the experimentally observed active compounds. The Phi-square Distance between two rankings can be calculated as follows. First, the test compounds are divided into four categories: true positives (TP), false positive (FP), false negatives (FN), and true negatives (TN). True positives is the number of active compounds which are also predicted to be active. False positives are the rest of the compounds predicted to be active, i.e. inactives misclassified as actives. Similarly, false negatives is the number of active compounds misclassified as inactives and true negatives are the inactive compounds predicted correctly to be inactive. As far as mathematics is concerned, it is a nice feature of the Phi-square Distance that the breakpoint between

actives and inactives is arbitrary, therefore allowing studies to focus on the top n ($1 < n < N$) compounds. In other terms, it allows us to evaluate what the Phi-square Distance is between the measured and the predicted ranking considering the top n compounds to be active. The breakpoint, in practice, is always determined by the expert working with the compounds, not by the person who evaluates the predictions. As far as the predicted ranking is concerned, the breakpoint between actives and inactives can be defined in two different ways. Either the measured activity of the nth compound defines the breakpoint between actives and inactives in the predicted ranking, or the first n compounds of the predicted ranking are considered to be active. These two alternatives lead to what we call the quantitative and the qualitative Phi-square Distance methods, respectively. The Phi-square Distance, PSD, can be calculated as follows:

$$PSD = \frac{TP^2}{A1 * A3} + \frac{FP^2}{A2 * A3} + \frac{FN^2}{A1 * A4} + \frac{TN^2}{A2 * A4} - 1 \qquad (1)$$

where $A1 = TP + FN$, $A2 = FP + TN$, $A3 = TP + FP$, and $A4 = FN + TN$.

It can easily be shown that in the case of chance predictions, i.e. when activity is assigned randomly to each of the test compounds, the PSD will be equal or close to zero. On the other hand, perfect prediction gives a $PSD = 1$. However, it should be noted that a perfect inverse prediction, i.e. when the predicted ranking is a perfect ranking of the test compounds in the reverse order of activity, PSD will also be equal to one. Equation (1) is the square of what is called the "phi coefficient" in ref. [1], pp. 26–27; however, we have done some algebraic manipulations in order to achieve a more readable expression. The Phi-square Distance PSD multiplied by the number of test compounds N follows the chi-square distribution with one degree of freedom. This fact allows one to calculate the probability that a particular prediction of a test set is due to pure chance [6]. Thus, a chi-square value of 3.84 indicates that there is a 5% probability of obtaining such a fit by chance. A value of 6.63 indicates only a 1% probability that such results could have been found by chance. It should be noted, however, that our objective is to measure the quality of a prediction rather than to calculate its probability of being due to chance.

It is evident that two Phi-square Distance values both equal to, say, 0.5, achieved with two test sets of different size, are not equivalent as far as the probability of being a chance "prediction" is concerned. However, the quality of those two predictions is indeed the same, and only this is what counts during the course of this study.

2.1.2. Rank Comparison method

Another evaluation technique we developed for our experiment is the Rank Comparison method. In this method, we do not use the activity values, but rather compare the measured and the predicted rankings. As with the Phi-square Distance, the first n most active test compounds are selected as active and one counts how many of

the topmost compounds (K) of the predicted rankings must be considered in order to include a fixed percentage (X) of the n active compounds of the test set. A reasonable choice of X lies between 50% and 90%. Less than 50% would be a rather weak criterion, whereas a percentage above 90% should also be avoided so that a single or a few badly mispredicted compounds will not unduly affect the outcome of the evaluation. It should be noted that with the Rank Comparison method, chance prediction never turns out to be zero. Since the chance prediction is a fixed percentage times the number of test compounds, it varies with both the strength of the percentage criterion and the size of the test set. A simple normalizing transformation is necessary to make the Rank Comparison method comparable with the Phi-square Distance, i.e. to set the chance prediction equal to zero and the perfect prediction equal to one.

The normalized Rank Comparison Measure ($NRCM$) can be calculated as follows:

$$NRCM = \frac{(X * n/K) - (n/N)}{1 - (n/N)} = \frac{RCM - \text{Chance}}{1 - \text{Chance}}, \tag{2}$$

where K is the number of the topmost compounds in the predicted ranking that should be considered in order to include a fixed percentage (X) of the n actives out of the N test compounds. $NRCM$ is equal to zero for chance prediction, and equal to one for "perfect" prediction when $K = X * n$. However, $NRCM$ can also be negative, indicating that a tendency existed to predict inactive compounds to be active and vice versa.

2.1.3. The Shuffle method

We call the third evaluation technique the Shuffle method. It is similar to Spearman's rank correlation coefficient [7], i.e. it is based on the rank difference of the same object in two different rankings. The Shuffle method, unlike the Phi-square Distance and the Rank Comparison Measure, leads to a global index of "shuffleness", which is a measure of how much it is necessary to shuffle the measured ranking to produce the predicted ranking. Simply, the sum of the absolute differences between the ranks of the corresponding compounds in the measured and in the predicted ranking is compared to that expected to be found by random shuffling. A straightforward weighting process allows one to focus on predicting active compounds. Each rank difference is weighted by the measured activity of the corresponding compound. It can easily be shown that the sum of the absolute rank differences with random shuffling is expected to be equal to:

$$-\frac{1}{N}\sum_{i=1}^{N-1} i(i+1).$$

The weighted measure of "shuffleness" ($WSHF$) is then given by the following equation:

$$WSHF = 1 - \frac{\sum_{i=1}^{N} \underset{[\text{of compound } i]}{\text{meas. activity}} * ABS(\underset{[\text{of compound } i]}{\text{meas. rank}} - \text{pred. rank})}{\frac{1}{N}\sum_{i=1}^{N-1} \frac{i(i+1)}{2} * (\underset{[\text{of compound } N-i]}{\text{meas. activity}} + \underset{[\text{of compound } i+1]}{\text{meas. activity}})}, \quad (3)$$

where the "measured activity" terms are the weighting factors. It should be noted that in cases where the activity is measured on an inverted scale, i.e. where the most active compounds are associated with the lowest numbers on the activity scale, the weighting factors in eq. (3) should be replaced by the corresponding reciprocal activities. *WSHF* is expected to be zero for chance prediction and equal to one for perfect prediction. As for *NRCM*, *WSHF* is also negative in the case of an inverted prediction.

2.2. TIES IN THE RANKING

Each evaluation method described in the previous section is relatively simple to use and easy to automate. However, there is a problem that makes the correct evaluation of predictions much more difficult. This problem arises from the common occurrence of ties in the ranking. Indeed, it often happens that two or more compounds are associated with the same activity value, forming a tie in the ranking. Actually, automated as well as human experts may use only a few categories to rank the whole set of test compounds; for example, very active, active, and inactive. Also, sometimes the measured activity of two compounds just happens to be the same. The existence of ties in the ranking creates problems for each of the evaluation methods discussed above. Since it does not make sense to differentiate between compounds within a tie, the test compounds must not be split into actives and inactives within a tie, either in the measured or in the predicted ranking. This means that n should always point to the end of a tie in the measured ranking. For the predicted ranking, this criteron is automatically fulfilled by the quantitative Phi-square Distance method, where the splitting among actives and inactives is based on a threshold activity value, which means that a tie is always completely on one side of the threshold. However, the application of the qualitative Phi-square Distance method (where the number of measured actives is in principle equal to the number of predicted actives) presents a serious problem, since the number of measured and predicted actives might be significantly different due to different ties in the measured and in the predicted ranking.

The Rank Comparison method can handle the ties in the following way. Let x be equal to the nearest integer to $X*n$, i.e. x is the rounded $X\%$ of the n active test compounds. If the xth in the predicted ranking of the first n most active compounds falls into a tie (T), then K in eq. (2) is calculated as follows:

$$K = \underset{\text{ranked before } T}{\text{number of compounds}} + \underset{\text{tied within } T}{\text{number of compounds}} * \frac{X*n-y}{z}, \quad (4)$$

where y and z are the numbers of the first n most active test compounds ranked before T and ranked within T, respectively.

The Shuffle method is also affected by the presence of ties in the ranking. Both the measured and the predicted rank of compound i may be tied, which means that it does not make sense to compare the two ranks directly. In both the measured and the predicted ranking, the beginning and the end of the tie including compound i are determined. In the case of no ties at all, the corresponding beginnings and ends coincide, which means that eq. (3) is unaffected. If ties do exist, then the absolute rank difference in eq. (3) must be calculated as follows:

$$ABS(\text{meas. rank} - \text{pred. rank})_{\text{of compound } i} = \frac{\sum_{m=\text{beg}}^{\text{end}} \sum_{p=\text{beg}}^{\text{end}} ABS(m-p)}{(\text{end} - \text{beg} + 1)_m * (\text{end} - \text{beg} + 1)_p}, \quad (5)$$

which is the average absolute difference between the measured (m) and the predicted (p) rank of compound i taking all possible positions of this compound within its ties into account.

2.3. A WORKING EXAMPLE

To become familiar with the rather abstract definition of the evaluation methods introduced in the previous sections, let us show a simple working example. In table 1, the measured and the predicted ranking of ten compounds are listed (imaginary data).

Table 1

Measured and predicted ranking of ten imaginary compounds

	Measured ranking compound	activity	Predicted ranking compound	activity
Actives	A	25	A	28
	B	19	C	22
	C	18	E	22
	D	18	F	22
			G	22
	E	11	B	18
	F	9		
	G	9	H	8
Inactives	H	4	D	5
	I	1	I	3
	J	1	J	3

Table 1 (continued)

Quantitative Phi-square Distance:

$N = 10$, $n = 4$, threshold activity = 18,
$TP = 3$ (A, B, C), $FP = 3$ (E, F, G), $FN = 1$ (D), $TN = 3$ (H, I, J),
$A1 = 4$, $A2 = 6$, $A3 = 6$, $A4 = 4$,

$$PSD = \frac{3*3}{4*6} + \frac{3*3}{6*6} + \frac{1*1}{4*4} + \frac{3*3}{6*4} - 1 = 0.0625.$$

Qualitative Phi-square Distance:

$N = 10$, $n = 4$, predicted actives (A, C, E, F, G) (G is tied with C&E&F!),
$TP = 2$ (A, C), $FP = 3$ (E, F, G), $FN = 2$ (B, D), $TN = 3$ (H, I, J),
$A1 = 4$, $A2 = 6$, $A3 = 5$, $A4 = 5$,

$$PSD = \frac{2*2}{4*5} + \frac{3*3}{6*5} + \frac{2*2}{4*5} + \frac{3*3}{6*5} - 1 = 0.000.$$

Rank Comparison method:

$X = 50\%$, $N = 10$, $n = 4$,
$K = 1 + 4(0.5 * 4 - 1)/1 = 5$,

$$NRCM = \frac{0.5*4\ /\ 5 - 4\ /\ 10}{1 - 4\ /\ 10} = 0.000.$$

$X = 75\%$, $N = 10$, $n = 4$,
$K = 6$,

$$NRCM = \frac{0.75*4\ /\ 6 - 4\ /\ 10}{1 - 4\ /\ 10} = 0.167.$$

Shuffle method:

The numerator for eq. (3) is the sum of the following terms:
$25 * 0 = 0$ (A),
$19 * 4 = 76$ (B),
$18 * (1 + 0 + 1 + 2 + 2 + 1 + 0 + 1)/(3 - 2 + 1)/(5 - 2 + 1) = 18$ (C),
$18 * (5 + 4)/2 = 81$ (D),
$13 * (3 + 2 + 1 + 0)/4 = 19.5$ (E),
$9 * (4 + 3 + 2 + 1)/4 = 22.5$ (F),
$9 * (5 + 4 + 3 + 2)/4 = 31.5$ (G),
$4 * 1 = 4$ (H),
$1 * 0 = 0$ (I), and
$1 * 0 = 0$ (J).

The denominator for eq. (3) is equal to $0.1(1(1 + 19) + 3(4+18) + 6(9 + 18) + 10(9 + 13) + 15(13 + 9) + 21(18 + 9) + 28(18 + 4) + 36(19 + 1) + 45(25 + 1)) = 387.1$, and *WSHF* is thus equal to:

$$WSHF = 1 - (0 + 76 + 18 + 81 + 19.5 + 22.5 + 31.5 + 4 + 0 + 0)/387.1 = 0.348.$$

From a casual inspection of table 1, the measured and the predicted rankings do not appear to be too much different. However, focusing on the top four active compounds (A, B, C, D), the predicted ranking turns out to be not much better than chance prediction (see the *PSD* and the *NRCM* results). Only the Shuffle method gives a somewhat higher value (0.348), which means that the overall prediction is better than the prediction of the top four compounds.

2.4. DATA

An experiment has been run to evaluate the predictive power of our new artificial intelligence structure–activity technique, which is called MULTICASE [8]. MULTICASE is the recent and totally redesigned version of the CASE (Computer Automated Structure Evaluation) program [9]. In the experiment, the predictions of the antibacterial potencies of a series of compounds in five different tests by MULTICASE and three human experts have been compared. Several hundred compounds were tested for their antibacterial potency in five different standard tests (Gram Negative Mics (AP1), Gram Positive Mics (AP2), DNA Gyrase Inhibition (AP3), Mean Subcutaneous Protective Dose in Mice (AP4), and Mean Oral Protective Dose in Mice (AP5)) by a cooperating pharmaceutical company. Of these, sixty-nine were selected by one of the experts to become the learning set and fifty-three were to be used as a test set. No information beyond the molecular structures of the learning set was given. No prior or additional information was to be used by the experts. The actual number of test compounds in the DNA Gyrase test and in the Mice tests is less – forty-four and thirty-four, respectively – because the experimental activity of some of the test compounds had not been measured in these tests. The breakpoint between actives and inactives was determined in all of the five tests by one of the human experts, C2, based on the standards of the pharmaceutical company.

3. Results

In table 2(a) to 2(e), the comparison of the rankings of the test compounds in the five tests is presented. In each table, the leftmost column is the measured ranking (ME), followed by the predicted ranking by MULTICASE (MC) and by the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns. The tables are split into two parts, separating the measured actives from the measured inactives, as well as the predicted actives from the predicted inactives for all of the experts, based on the breakpoint defined by the company standards. The three human experts are leading chemists of the cooperating pharmaceutical company. However, their preknowledge of the results was somewhat different. Indeed, both C1 and C2 had been working for several years with antibacterial agents and are experts in that area. C1 compiled the data set and selected the learning and the test set, C2 had been

Table 2(a)

Comparison of the rankings of the test compounds in the Gram Negative Mics Test. The leftmost column is the measured ranking (ME), followed by the predicted ranking of MULTICASE (MC) and of the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns

	ME	MC	C1	C2	C3
Actives	41	41	43	41	41
	50	50	42	29	29
	29	42	37	43	43
	32	15	44	26	14
	43		41	15	26
	49	48	50	44	30
	46	45	32		52
	48	29	49	50	45
	14	43	48	32	15
	18	49	38	49	19
	42	14	40	46	
		11	39	48	48
Inactives	11	17	45	14	42
	10	37		42	11
	17	20	29	11	10
	22	26	11	10	22
	37	30	22	17	38
	20	38	26	22	34
	26	40	30	37	28
	30	39	52	30	25
	38	31	15	38	51
	40	08	31	40	24
	52	21	47	52	50
	34	24	46	34	46
	28	16	18	39	17
	39	27	34	45	40
	45	44	21	25	39
	15	47	24	31	08
	25	19	16	35	21
	31	13	27	51	27
	35	33	23	08	47
	51	18	14	21	07
	08	10	10	24	12
	21	52	17	16	18
	24	23	28	27	37
	16	09	51	19	31
	27	32	19	12	16
	44	46	12	18	23
	47	22	20	28	00
	23	34	25	47	33
	09	28	08	23	01
	19	25	07	09	32
	07	35	13	13	49
	00	51	33	33	20
	13	07	00	20	35
	33	00	02	07	44
	02	02	35	01	09
	01	01	09	00	13
	04	04	05	02	02
	06	06	03	04	04
	12	12	36	06	06
	05	05	01	36	05
	03	03	06	05	03
	36	36	04	03	36

Table 2(b)

Comparison of the rankings of the test compounds in the Gram Positive Mics Test. The leftmost column is the measured ranking (ME), followed by the predicted ranking of MULTICASE (MC) and of the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns

	ME	MC	C1	C2	C3
	42	42	42	42	41
	43	38	43	43	14
	41	37	50	41	26
	14	49	44	50	17
	38	50	41	26	34
	31	32	38		52
	37	40	37	14	28
	49	39	49	38	51
	50	41	32	31	
	26	31	48	37	21
Actives	30	30	45	49	22
	46	17		30	24
			31	46	29
	17	43	26	17	25
Inactives	32	24	46	32	42
	40	14	40	40	43
	48	26	22	48	38
	21	48	24	22	37
	08	45	34	24	50
	22	13	39	34	30
	24	22	52	39	46
	34	44	30	44	40
	35	18	15	15	48
	39	23	11	27	08
	44	15	27	45	39
	15	10	29	16	15
	11	12	28	23	11
	27	52	47	28	27
	29	35	51	21	10
	18	21	33	08	19
	45	27	14	35	23
	16	28	17	19	12
	25	47	21	52	31
	20	46	18	51	45
	10	08	19	11	49
	19	34	23	29	32
	23	11	08	18	35
	12	29	25	25	44
	13	16	10	20	18
	52	25	12	10	16
	28	20	13	12	20
	47	19	16	13	13
	07	07	20	47	47
	09	09	07	09	07
	51	51	02	07	09
	02	02	03	02	02
	05	05	36	03	05
	06	06	35	33	06
	04	04	09	00	04
	03	03	06	05	03
	33	33	00	06	33
	00	00	05	04	00
	01	01	04	01	01
	36	36	01	36	36

Table 2(c)

Comparison of the rankings of the test compounds in the DNA Gyrase Test. The leftmost column is the measured ranking (ME), followed by the predicted ranking of MULTICASE (MC) and of the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns

	ME	MC	C1	C2	C3
	38	38	38	44	43
	40	40	40	41	26
	42	44	42	43	29
	44	41	44		30
	41	43	43	38	15
	43	21	39	40	34
	21	26	41	42	11
	26	39	32	26	
	29	27	31	29	41
	39	25	11	39	10
Actives	30	24	47	30	22
	32			32	12
		29	21	15	38
	15	32	26	27	40
Inactives	19	30	29	31	42
	20	20	30	34	21
	27	31	15	46	39
	31	34	27	48	19
	34	46	34	14	27
	46	11	48	25	46
	48	13	28	11	14
	14	16	22	10	28
	25	42	45	28	24
	11	15	33	22	47
	10	19	24	24	02
	28	48	19	47	45
	13	14	20	12	00
	18	10	46	45	31
	22	28	14	21	48
	23	18	25	19	25
	24	22	10	20	23
	47	23	13	13	44
	07	47	18	23	32
	08	07	23	07	20
	09	08	07	08	13
	16	09	08	16	18
	12	12	16	33	07
	02	02	12	18	08
	45	45	09	09	09
	33	33	02	02	16
	00	00	00	05	33
	05	05	05	06	05
	06	06	06	00	06
	36	36	36	36	36
	01	01	01	01	01

Table 2(d)

Comparison of the rankings of the test compounds in the Mean Subcutaneous Protective Dose in Mice Test. The leftmost column is the measured ranking (ME), followed by the predicted ranking of MULTI-CASE (MC) and of the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns

	ME	MC	C1	C2	C3
Actives	32	50	32	41	41
	29	32	41	32	50
	41	29	49	29	11
	49	41	50	49	43
	50	49	11	50	34
	11	37	43	11	26
	43	00	34	43	17
			48	34	14
Inactives	34	48	46	26	
	26	11	37	48	22
	48	43	31	46	24
	25	34	42	42	29
	30	26			48
	46	25	29	25	30
	10	30	26	30	46
	15	46	30	10	15
	37	10	15	15	42
	31	15	17	37	38
	42	31	18	31	27
	17	42	22	17	32
	18	17	40	22	49
	22	18	24	24	25
	40	22	14	14	10
	24	40	38	00	37
	20	24	39	08	31
	14	20	35	18	18
	00	14	25	40	40
	38	38	10	38	20
	08	08	08	39	00
	39	39	21	13	08
	13	13	27	35	39
	21	21	20	16	13
	35	35	00	20	21
	16	16	13	21	35
	27	27	16	27	16

Table 2(e)

Comparison of the rankings of the test compounds in the Mean Oral Protective Dose in Mice Test. The leftmost column is the measured ranking (ME), followed by the predicted ranking of MULTICASE (MC) and of the three human experts, C1, C2, and C3, respectively. The ties in the rankings are marked by vertical bars alternating on the left- and on the right-hand side of the columns

	ME	MC	C1	C2	C3
Actives	49	49	41	49	41
	32	32	50	29	34
	46	29		41	43
	29	37	49	50	17
	41	00	32		26
	50	41	46	32	14
	11	50	34	46	24
	34	48	43	11	46
	43		48	34	29
	48	11	37	43	50
		46	42	48	48
Inactives	22	34	29	24	22
	37	43	11	30	37
	18	22	22	17	30
	42	18	24	22	40
	00	42	30	37	15
	24	24	40	42	39
	30	30	15	31	25
	40	40	17	25	38
	15	15	31	26	11
	17	17	39	00	08
	31	31	38	15	10
	39	39	26	18	
	08	08	14	40	49
	25	25	18	39	32
	38	38	00	08	18
	13	13	08	38	42
	20	20	25	13	00
	26	26	10	20	31
	35	35	21	35	13
	10	10	27	10	20
	14	14	13	14	35
	16	16	20	16	16
	21	21	35	21	21
	27	27	16	27	27

Table 3

Evaluation of the rankings of the test compounds in different tests. In each test, the first line is the evaluation of the MULTICASE ranking (MC), the second, third, and fourth lines are that of the human experts, C1, C2, and C3, respectively. In each test, the first column is the quantitative Chi-square Distance, the second column is the qualitative Chi-square Distance, the third, fourth, and fifth columns are the Rank Comparison Measures with X = 50, 75, and 90%, respectively, and finally, the last column is the result of the Shuffle evaluation. N is the total number of test compounds and n is the number of the active test compounds.

	Qnt *CSD*	Qlt *CSD*	*NRCM* X = 50%	*NRCM* X = 75%	*NRCM* X = 90%	*WSHF*
		AP1: Gram Negative Mics Test (N = 53, n = 11)				
MC	0.146	0.028	0.201	0.069	0.028	0.548
C1	0.042	0.216	0.412	0.167	0.160	0.577
C2	0.066	0.064	0.153	0.103	0.089	0.577
C3	0.052	0.024	0.118	0.050	0.011	0.432
		AP2: Gram Positive Mics Test (N = 53, n = 12)				
MC	0.324	0.324	0.569	0.538	0.462	0.496
C1	0.251	0.259	0.546	0.392	0.221	0.771
C2	0.356	0.282	0.660	0.369	0.308	0.797
C3	0.022	0.000	0.082	0.117	0.136	0.304
		AP3: DNA Gyrase Inhibition Test (N = 44, n = 12)				
MC	0.347	0.521	0.625	0.656	0.390	0.588
C1	0.421	0.342	1.000	0.509	0.390	0.749
C2	0.195	0.145	0.312	0.214	0.187	0.440
C3	0.085	0.091	0.214	0.163	0.042	0.201
		AP4: Mean Subcutaneous Protective Dose (N = 34, n = 7)				
MC	0.410	0.410	0.669	0.637	0.594	0.646
C1	0.289	0.289	0.370	0.370	0.370	0.504
C2	0.117	0.475	0.530	0.493	0.481	0.707
C3	0.163	0.163	0.370	0.148	0.020	0.203
		AP5: Mean Oral Protective Dose (N = 34, n = 10)				
MC	0.308	0.000	0.595	0.390	0.080	0.653
C1	0.150	0.514	0.764	0.722	0.356	0.636
C2	0.320	0.673	0.871	0.732	0.692	0.759
C3	0.043	0.034	0.128	0.102	0.039	−0.206

intimately involved with the synthesis of the compounds. C3 is the only expert with apparently no a priori knowledge about the set of compounds used in the experiment.

The evaluation results are presented in table 3. Each ranking in each test is evaluated in six different ways. These are the quantitive Phi-square Distance, the qualitative Phi-square Distance, and the Rank Comparison Measure with $X = 50$, 75, and 90%, each focusing on the active test compounds, and finally, the Shuffle method. The total number of the test compounds (N) and the number of the active test compounds (n) is also shown for each test in table 3.

4. Discussion

4.1. DISCUSSION OF THE METHODOLOGIES

If the measured potency threshold is greater than any of the predicted potencies, i.e. when there are neither true nor false positives, even a perfect ranking gives an undefined *PSD* value when using the quantitative Phi-square Distance. It should be noted, however, that this is more of an advantage than a disadvantage of the method. Indeed, the quantitative Phi-square Distance characteristics are such that even if the ranking itself is perfect, if none of the active compounds are predicted to be active, then this prediction is definitely a bad prediction.

In the theoretical section of this paper, it was mentioned that the qualitative Phi-square Distance method, where the number of measured actives is in principle equal to the number of predicted actives, had a serious drawback in practice. This is vividly demonstrated in, for example, the Oral Protective Dose case where the number of active test compounds is ten (see table 3). It can be seen that the qualitative Phi-square Distance is equal to zero for the MULTICASE prediction, which is obviously nonsense. The reason for the strange result can be found in table 2(e), where it is seen that the tenth compound in the MC ranking is tied with all of the remaining compounds. This means that the whole test set is considered to contain only active compounds, i.e. there are neither false nor true negatives causing the undefined *PSD* result.

The Rank Comparison method evaluates the rankings regardless of the predicted potency levels. The use of ranks eliminates the problems which would otherwise occur when attempting to relate potencies for each of the scales used for predictions. This is not, however, necessarily an advantage over the quantitative Phi-square Distance method for, as we stated earlier in this section, the evaluation of the rankings does not make too much sense when nothing can be said about the predicted potency levels. Furthermore, even when restricting ourselves to the ranks, there are still problems with the Rank Comparison method. Indeed, we find that there are too many degrees of freedom involved with the choice of $X\%$. It is clear that $X\%$ should fall somewhere between 50% and 90%, but there is no clue as to what the optimum value ought to be. We cannot even say that $X = 75\%$ or $X = 90\%$ is always a sharper criterion than $X = 50\%$ (see, for instance, in table 3 the C3 prediction in the Gram Positive Mics test or the C1 prediction in the Subcutaneous Protective Dose test). In addition, we find,

for example, that the C1 prediction of the active test compounds in the DNA Gyrase test is better than that of MC with $X = 50\%$ (1.000 versus 0.625), but the opposite is true with $X = 75\%$ (0.509 versus 0.656) and with $X = 90\%$, C1 and MC are both found to be equal to 0.390.

Thus, in general, the numerical value of the Rank Comparison Measure is a rather unpredictable function of $X\%$. This makes the utility of the method quite questionable. This problem is even more serious when one considers that with a "perfect" prediction, *NRCM* is always equal to one, regardless of $X\%$ (see eq. (2)). This means that the level of "perfectness" is not taken into account in the Rank Comparison method. In other words, the fact that it is easier to predict "perfectly" with $X = 50\%$ than with $X = 90\%$ is completely ignored. However, it is possible that the comparison of two different rankings on different levels of $X\%$, which may be a rather complicated procedure, is the correct way of using the Rank Comparison method.

A serious problem arises with the Shuffle method if there is a large difference between the potency level of the very few top test compounds and the potency level of the rest of the active compounds in the test set, which is often the case. According to the weighting process in eq. (3), this means that the ranking of those few top test compounds dominates the Shuffle evaluation result. The problem is that in this case the partition of the test set into actives and inactives might be extremely skewed, which jeopardizes the reliable evaluation of the predicted ranking.

4.2. COMPARISON OF THE PREDICTIONS OF THE EXPERTS

According to the quantitative *PSD* evaluation, MULTICASE is superior to the human experts in the AP1 and the AP4 tests (see table 3). In AP2, AP3, and AP5, MC performed significantly better than C3 and on a comparable level with the other two, C1 and C2. Adding the results of the five tests gives an overall evaluation of the predictions of the experts. The results are shown in table 4 and lead to our overall conclusion that MULTICASE performed significantly better than one of the human experts and somewhat better than the other two. As a matter of fact, the average Phi-square value for MULTICASE, i.e. 1.535/5 = 0.31, when multiplied by the average number of molecules in the test sets, gives a chi-square value of 13.4, far exceeding the 99% confidence level (chi-square = 6.63) usually considered indicative of a good fit.

The qualitative *PSD* results are not reliable because the long ties in the predicted rankings bias the threshold between actives and inactives. This problem was discussed earlier in the previous section. We do not rank the experts by the *NRCM* results either, for as was also discussed earlier in this section, the choice of $X\%$ is a rather arbitrary parameter of the Rank Comparison method. However, a qualitative look at the *NRCM* results in table 3 confirms the global observation that MULTICASE generally performs better than C3 and on a comparable level with the other two human experts, C1 and C2. According to the Shuffle evaluation, C1 and C2 are superior to MC in the Gram tests and C1, C2, and MC performed on a comparable level in the AP3 to AP5 tests.

Table 4

Overall performance of the experts based on the quantitative Chi-square Distance. The entries here are the quantitative Chi-square Distance results taken from table 3. MC is MULTICASE; C1, C2, and C3 are the human experts. AP1–AP5 are the different tests as indicated in table 3. The five results of each expert are added in the bottom line

	MC	C1	C2	C3
AP1	0.146	0.042	0.066	0.052
AP2	0.324	0.251	0.356	0.022
AP3	0.347	0.421	0.195	0.085
AP4	0.410	0.289	0.117	0.163
AP5	0.308	0.150	0.320	0.043
Sum	1.535	1.153	1.054	0.365

C3 is in each test inferior to any of the other experts. However, the problem of the hegemony of the very few top compounds, which was also discussed earlier in this section, endangers the reliability of the Shuffle evaluation results. Indeed, for each test, the potency level of the top three to five test compounds was more than ten times higher than the potency level of the rest of the active compounds in the test set.

4. Conclusions

In this paper, we have presented three mathematical techniques for evaluating the quality of SAR predictions. Unlike any kind of correlation method which gives equal weight to each compound, irrespective of the level of potency, each methodology in this paper takes into account the fact that we are more concerned with the active compounds than with the inactive compounds. Overall, we suggest the use of the quantitative Phi-square Distance method, which appeared to be superior to the other presented techniques for evaluating the quality of activity predictions. This is the only method which does not have any conceptual uncertainties, such as $X\%$ in the Rank Comparison method. Only the quantitative Phi-square Distance can be used with data where ranking ties exist without the need of introducing (more or less) arbitrary parameters. Finally, only this technique takes the predicted level of potency fully into account.

We also conclude that MULTICASE performed significanly better than one of the human experts (C3) and somewhat better than the other two, C1 and C2. Considering that MULTICASE and C3 were the only experts without prior knowledge other than the molecular structures and the potencies of the compounds in the learning set, this is a particularly good result.

Acknowledgement

Helpful discussions with Dr. Bernard Ycart are gratefully acknowledged.

References

[1] A.R. Baggaley, *Intermediate Correlational Methods* (Wiley, 1964).
[2] M.L. Puri (ed.), *Nonparametric Techniques in Statistical Inference* (Cambridge University Press, 1970).
[3] W.W. Daniel, *Applied Nonparametric Statistics* (Houghton Mifflin Co., Boston, 1978).
[4] S. Wold, Technometrics 20, 4(1978)397.
[5] R.D. Cramer, III, J.D. Bunce, D.E. Patterson and I.E. Frank, Quant. Struct.-Act. Relat. 7(1988)18.
[6] D.E. Bailey, *Probability and Statistics Models for Research* (Wiley, 1971).
[7] W.H. Berger (ed.), *CRC Handbook of Tables for Probability and Statistics* (The Chemical Rubber Co., Cleveland, OH, 1966), pp. 329–330.
[8] G. Klopman, MULTICASE: A hierarchical computer automated structure evaluation program, in press.
[9] G. Klopman, J. Amer. Chem. Soc. 106(1984)7315.