Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2004) 275–284

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

A method for quantifying and visualizing the diversity of QSAR models

Sergei Izrailev*, Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., 8 Clarke Drive, Cranbury, NJ 08512, USA

Received 21 June 2003; received in revised form 10 October 2003; accepted 13 October 2003

Abstract

Feature selection is one of the most commonly used and reliable methods for deriving predictive quantitative structure–activity relationships (QSAR). Many feature selection algorithms are stochastic in nature and often produce different solutions depending on the initialization conditions. Because some features may be highly correlated, models that are based on different sets of descriptors may capture essentially the same information, however, such models are difficult to recognize. Here, we introduce a measure of similarity between QSAR models that captures the correlation between the underlying features. This measure can be used in conjunction with stochastic proximity embedding (SPE) or multi-dimensional scaling (MDS) to create a meaningful visual representation of structure–activity model space and aid in the post-processing and analysis of results of feature selection calculations.

© 2003 Elsevier Inc. All rights reserved.

Keywords: Stochastic proximity embedding; Multi-dimensional scaling; Nonlinear mapping; Feature selection; Point set similarity; Quantitative structure–activity relationships; Data mining

1. Introduction

Quantitative structure–activity relationships (QSAR) are mathematical models that relate the biological activity of a series of compounds to a set of features (descriptors) derived from their chemical structure. Because many of these features may be correlated, it is often desirable to construct a QSAR model using an optimal subset that captures the relevant molecular properties and thus, yields the most predictive model. This approach also guards against over-fitting, i.e. the tendency of a learning algorithm to memorize the training patterns and lose its ability to generalize beyond the training set. Feature selection algorithms, which aim at finding such an optimal subset, range from simple deterministic greedy approaches such as forward selection [1,2] and backward elimination based on pairwise correlation [3,4], to a number of stochastic optimization techniques that include simulated annealing [5], genetic algorithms [6–9], evolutionary programming [10], artificial ants [11,12], and particle swarms [13]. Due to the combinatorial nature of the feature selection problem (there are 2^n possible combinations of n available features), these algorithms often do not find the optimal subset, but instead produce a solution that corresponds to a local minimum in the search space. Moreover, unless the feature selection algorithm is deterministic, multiple selection attempts will produce different subsets of features. Thus, it has become common practice to construct multiple models that are based on different feature subsets for the same data set by one or more selection techniques and then choose the most predictive model, or combine several predictive models in the form of an ensemble or meta-predictor [14].

An important product of feature selection is the understanding of molecular properties that contribute to predictive models. However, given multiple models, the task of identifying the models that capture essentially the same molecular properties but are based on different albeit highly correlated features, or the complementary task of identifying predictive models that capture substantially different properties, may require a lot of tedious work. Indeed, if two highly predictive models are based on hypothetical descriptor sets (A, B, C, D) and (A, B, E, F), one needs to determine whether features C and E, C and F, D and E, D and F are correlated in order to find out whether these two sets of features, and hence the two models, are equivalent or capture essentially different molecular properties. Since the number of models and features is usually non-trivial, such comparisons, when done by a human looking at tables of numbers, are labor-intensive and error prone. In this paper, we describe a new QSAR

* Corresponding author. Tel.: +1-609-409-3416; fax: +1-609-655-6930.
E-mail address: sergei.izrailev@3dp.com (S. Izrailev).

1093-3263/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2003.10.001

model visualization approach that simplifies such analysis by introducing a new measure of similarity between two feature sets of equal size. This similarity measure captures high correlation between features in the two sets and provides the basis for visual comparison of the resulting QSAR models. Since every model is defined by the set of features contributing to the model, we will use the terms "model" and "feature set" interchangeably throughout the paper.

Defining a similarity measure between different feature sets is not sufficient for understanding the underlying structure–activity model space. For n feature sets, there are $n(n-1)/2$ pair-wise similarities, which are difficult to analyze without effective visualization techniques. An intuitive solution is to represent each feature set as a point on a two- or three-dimensional map, arranged in such a way that the distances between the points on the map approximate as much as possible the (dis)similarities of the respective feature sets. Several algorithms exist for performing such an embedding, including multi-dimensional scaling (MDS) [15], nonlinear mapping (NLM) [16] and stochastic proximity embedding (SPE) [17]. Here, we chose to employ the latter due to its speed, superior scaling and simplicity of implementation. We also limited the embedding to a two-dimensional space. The result of the embedding is a set of 2D coordinates for each feature set. While the coordinates by themselves may not have any physical meaning, the distances between the points on the map reflect the similarities of the underlying models, and therefore, clusters of closely related models become immediately apparent. Moreover, specific properties of the models can be visualized by color-coding the points on the map. This technique has been previously used to visualize and analyze a variety of different types of data, including chemical libraries [17,18], ensembles of molecular conformations [17], and protein sequences [19,20]. In this work, we utilize the SPE algorithm for computing two-dimensional maps of the feature sets using the proposed similarity measure, and employ color-coding to reflect the predictive quality of the corresponding models.

First, we describe the method of calculating the similarity and representing each model as a point on a low-dimensional display map. Distances between points on this map reflect as closely as possible the similarities between the underlying feature sets. Next, we demonstrate on three classical data sets how visualization of the QSAR model space offers immediate insight into the diversity and relative quality of the respective models. Finally, we compare the new similarity measure with one that does not take into account correlation between different features.

2. Methods

2.1. Correlation-based model similarity

Let us assume that a feature selection algorithm produces m sets of k features out of n features present in the data set.

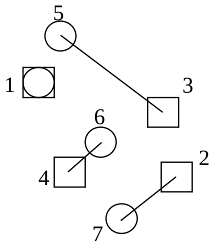

Fig. 1. Feature pair assignment for two feature sets represented by circles and squares in some hypothetical two-dimensional space where proximity is proportional to the correlation of the underlying features. Each set contains four features. Feature 1 is present in both sets, so (1,1) is the first pair. The next closest pair is (4,6), followed by (2,7) and (3,5). Note that feature 3 is better correlated to feature 6 than to 5, however, feature 6 has already been used in pair with feature 4.

The similarity between features i and j can be expressed as $q_{ij} = 1-c_{ij}^2$, where c_{ij} is their Pearson correlation coefficient calculated from the training data. Given two feature sets $S_i = \{i_1, i_2, \ldots, i_k\}$ and $S_j = \{j_1, j_2, \ldots, j_k\}$ of equal size, we match each feature in set S_i to exactly one feature in set S_j using a greedy algorithm that approximates a solution to the linear assignment problem [21]. The algorithm finds the pair of features (i,j) from sets S_i and S_j, respectively, that are the most similar among all possible pairs, i.e. with the smallest q_{ij}. Features i and j are then removed from the sets, and the process is repeated $k-1$ times until all features are removed. This process is illustrated in Fig. 1. As a result, we obtain k feature pairs with corresponding similarities $q_{ij,1}, q_{ij,2}, \ldots, q_{ij,k}$. The distance between sets S_i and S_j is then defined as

$$r_{ij} = 1 - \frac{1}{k}\sum_{l=1}^{k}\exp(-8q_{ij,l}^2) \qquad (1)$$

The exponential term is always less than or equal to 1, so r_{ij} ranges from 0 when all pairs consist of identical features to nearly 1 when all features in set S_i are uncorrelated with features in set S_j. The coefficient 8 was selected so that the exponential term would decay to a very small value (0.000335) at $q_{ij,l} = 1$. There are two factors that influenced the design of the distance function presented in Eq. (1). Firstly, if a pair of features is sufficiently uncorrelated, it does not matter exactly how small the correlation coefficient is, and such pairs should be treated almost equally. This effect is achieved by the sharp decay of the exponential function, so the influence of each individual uncorrelated pair of features on the distance between two sets is small relative to the contribution from pairs of highly correlated features. Secondly, the distance between two sets should decrease as the number of correlated feature pairs increases. Such a behavior results from the averaged sum of contributions over all feature pairs.

The algorithm for computing r_{ij}'s proceeds as follows

1. Compute the distance matrix q_{ij} for all n features.
2. Perform steps 3–10 for each pair of feature sets S_i and S_j.

3. Sort all possible pairs of points (*i*,*j*), where $i \in S_i$, $j \in S_j$ in ascending order of distance q_{ij}.
4. Mark all pairs as available. Set $k = 1$.
5. Find the first available pair (*i*,*j*) in the sorted list of pairs.
6. Set $q_{ijk} = q_{ij}$.
7. Mark all pairs that contain either feature *i* from set S_i or feature *j* from set S_j as unavailable.
8. Increment *k* by 1.
9. Repeat steps 5–8 until there are no available pairs left.
10. Compute r_{ij} according to Eq. (1).

2.2. Binary model similarity

A straightforward approach to compare two models is to check if they share any features. In this case, the feature sets are represented as bit vectors of *n* bits with the bits that correspond to the *k* selected features turned on. The similarity is then computed using the Dice coefficient

$$D_{ij} = \frac{2|\mathrm{AND}(S_i, S_j)|}{|S_i| + |S_j|} \tag{2}$$

where $|S_i|$ is the number of bits set in S_i and AND() is the bitwise "and" operation (a bit in the result is set if both of the corresponding bits in the two operands are set). Here, $|S_i| = |S_j| = k$, and the distance r_{ij} between sets S_i and S_j is computed according to

$$r_{ij} = 1 - \frac{|\mathrm{AND}(S_i, S_j)|}{k} \tag{3}$$

2.3. Stochastic proximity embedding

The procedures described above result in a symmetric matrix containing distances between feature sets, each of which is associated with a predictive QSAR model. The next step is to embed this distance matrix into a low-dimensional display map in a way that preserves the proximities (similarities) of the underlying models (in this work, we restricted output to two dimensions). The mapping is carried out using the stochastic proximity embedding algorithm [17]. SPE starts with an initial configuration, and iteratively refines it by repeatedly selecting pairs of objects (models) at random, and adjusting their coordinates so that their distances on the map d_{ij} match more closely their respective proximities r_{ij}. The algorithm proceeds as follows:

1. Initialize the coordinates $\boldsymbol{x}_i$. Select an initial learning rate λ.
2. Select a pair of points (i.e. models), *i* and *j*, at random and compute their distance $d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$. If $\mathrm{d}_{ij} \neq r_{ij}$, update the coordinates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ by

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} (\boldsymbol{x}_i - \boldsymbol{x}_j)$$

and

$$\boldsymbol{x}_j \leftarrow \boldsymbol{x}_j + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} (\boldsymbol{x}_j - \boldsymbol{x}_i)$$

where ϵ is a small number to avoid division by zero.
3. Repeat (2) for a prescribed number of steps *S*.
4. Decrease the learning rate λ by prescribed decrement $\delta\lambda$.
5. Repeat (2)–(4) for a prescribed number of cycles *C*.

The embedding was carried out using 50 cycles, 20,000 steps per cycles, and a linearly decreasing learning rate from 1.0 to 0.001. Unlike classical multi-dimensional scaling [15] and nonlinear mapping [16], SPE scales linearly with respect to sample size, and can be applied to very large data sets that are intractable by conventional embedding procedures [18] (linear scaling is not as important for the problem at hand due to the relatively small number of models examined, so other methods are also applicable). When this technique is applied to the proximities in Eq. (1), it produces a map where the distances between the points on the map reflect the similarities between the corresponding feature subsets. Furthermore, by color-coding the points by some value reflecting the quality of the underlying models, e.g. by the correlation coefficient between the predicted and the measured activities, one can obtain an intuitive visual aid for analyzing and interpreting the data.

2.4. Data sets

We demonstrate the use of this approach using the previously reported results of feature selection [13] for three well-studied data sets: antifilarial activity of antimycin analogues (AMA) [1], binding affinities of ligands to benzodiazepine/GABA$_\mathrm{A}$ receptors (BZ) [22], and inhibition of dihydrofolate reductase by pyrimidines (PYR) [23]. In ref. [13], the authors presented a number of models found by feature selection algorithms based on simulated annealing and particle swarms. The mean training Pearson correlation coefficient *R* was calculated for each model by averaging over 50 independent attempts to train a neural network QSAR model using the selected features. In addition, each model was tested 50 times using leave-one-out cross-validation and the mean cross-validated Pearson correlation coefficient R_cv was reported. These results are summarized in Tables 1–3. The notation for the descriptors can be found in the respective original references [1,22,23].

2.5. Implementation

All programs were implemented in the C++ programming language and are part of the DirectedDiversity® [24] software suite. All calculations were carried out on a Dell Inspiron laptop equipped with an 800 MHz Pentium III Intel processor running Windows 2000 Professional.

Table 1
Models reported for the AMA data set

No.	Features[a]	$\mu(R)$[b]	$\mu(R_{cv})$[c]
1	3,4,49	0.945	0.818
2	31,34,49	0.919	0.825
3	31,37,49	0.918	0.837
4	6,49,50	0.913	0.796
5	31,35,49	0.912	0.831
6	16,49,50	0.910	0.790
7	31,38,49	0.910	0.826
8	2,3,49	0.909	0.687
9	37,49,51	0.909	0.799
10	3,6,49	0.908	0.755
11	35,49,51	0.907	0.799
12	34,49,51	0.905	0.764
13	4,12,49	0.903	0.619

[a] Zero-based indices of features comprising the models identified by the particle-swarms and simulated annealing algorithms.
[b] Mean training R.
[c] Mean leave-one-out cross-validated R.

Table 2
Models reported for the BZ data set

No.	Features[a]	$\mu(R)$[b]	$\mu(R_{cv})$[c]
1	1,4,5,9,15,23	0.963	0.874
2	1,3,4,9,15,23	0.963	0.876
3	1,3,6,9,15,23	0.962	0.881
4	1,4,5,9,20,22	0.962	0.869
5	1,3,4,9,15,22	0.961	0.845
6	1,4,5,9,15,27	0.961	0.896
7	1,4,6,9,15,27	0.961	0.870
8	1,4,5,9,15,38	0.960	0.882
9	0,1,3,9,15,16	0.960	0.871
10	1,3,4,7,9,15	0.960	0.857
11	1,4,6,9,15,22	0.960	0.894
12	1,3,4,9,10,15	0.960	0.870
13	0,1,4,5,9,15	0.959	0.868
14	1,3,6,9,15,27	0.959	0.860
15	0,1,3,9,16,19	0.959	0.850
16	1,3,4,9,15,38	0.959	0.885
17	1,2,3,4,9,20	0.959	0.864
18	1,3,4,8,9,15	0.959	0.876
19	0,1,5,9,19,41	0.958	0.886
20	1,4,6,8,9,15	0.958	0.881
21	1,4,5,9,15,20	0.958	0.899
22	1,3,6,8,9,15	0.957	0.872
23	1,4,6,9,15,38	0.957	0.899
24	1,4,5,9,15,24	0.957	0.868
25	1,5,6,8,9,15	0.957	0.871
26	0,1,5,9,15,16	0.956	0.880
27	1,3,4,9,17,20	0.954	0.858
28	1,3,4,9,15,27	0.953	0.875
29	1,4,5,9,15,28	0.953	0.866
30	1,4,5,8,9,15	0.953	0.873
31	1,3,5,9,15,23	0.953	0.852
32	1,5,6,9,10,15	0.953	0.856
33	1,4,6,9,20,21	0.953	0.900
34	1,2,4,5,9,20	0.952	0.862
35	1,4,5,9,20,41	0.952	0.884
36	1,3,4,9,15,17	0.952	0.834
37	1,3,4,9,15,28	0.952	0.874
38	1,5,6,9,19,20	0.951	0.846
39	0,1,3,9,15,20	0.951	0.867
40	1,4,5,9,17,20	0.951	0.867
41	1,3,8,9,15,23	0.951	0.877
42	1,3,6,9,18,19	0.951	0.857
43	0,1,5,9,19,20	0.951	0.874
44	0,1,3,9,19,31	0.950	0.886
45	1,4,6,9,16,18	0.950	0.897
46	0,1,5,9,19,31	0.949	0.883
47	1,5,9,13,19,36	0.949	0.871
48	0,1,5,9,14,16	0.949	0.883
49	1,2,4,6,9,19	0.949	0.853
50	0,1,3,9,14,18	0.949	0.862
51	0,1,3,9,13,19	0.948	0.880
52	0,1,5,9,11,14	0.946	0.864
53	1,3,9,11,14,34	0.946	0.887
54	1,3,4,9,18,19	0.946	0.855
55	1,3,6,9,17,18	0.946	0.839
56	1,3,9,17,18,22	0.945	0.857
57	1,4,5,9,12,19	0.945	0.880
58	1,3,9,11,14,23	0.945	0.888
59	0,1,3,9,18,19	0.945	0.871
60	1,2,3,9,17,31	0.945	0.862
61	1,2,5,9,17,18	0.943	0.826
62	1,3,9,15,19,33	0.943	0.866
63	0,1,3,9,11,14	0.942	0.879
64	0,1,3,9,14,20	0.942	0.868
65	0,1,4,9,15,19	0.942	0.861
66	1,5,9,16,20,26	0.941	0.883
67	1,2,3,5,9,19	0.940	0.866
68	1,3,4,9,15,29	0.940	0.870
69	0,1,5,9,19,23	0.939	0.890
70	1,3,4,9,16,20	0.938	0.862
71	0,1,5,9,19,27	0.937	0.888
72	0,2,5,9,15,23	0.935	0.872
73	0,1,5,6,9,20	0.934	0.845
74	1,5,6,9,12,15	0.933	0.856
75	1,3,9,15,22,23	0.932	0.867
76	1,4,6,9,15,23	0.930	0.886
77	0,1,3,7,9,15	0.928	0.864
78	1,4,5,9,11,15	0.920	0.859
79	1,5,6,9,15,25	0.919	0.835
80	1,4,5,9,20,23	0.917	0.901
81	1,3,6,9,15,26	0.916	0.872
82	0,1,2,5,9,14	0.915	0.519

[a] Zero-based indices of features comprising the models identified by the particle-swarms and simulated annealing algorithms.
[b] Mean training R.
[c] Mean leave-one-out cross-validated R.

3. Results and discussion

SPE maps were computed for each of the data sets using both the correlation-based and binary similarity measures. Each point on a map represents a subset of features and, therefore, the corresponding QSAR model. The distances between points on the map reflect as closely as possible the computed similarities between the corresponding models and carry the bulk of visual information, while the map axes do not have interpretable physical meaning (in certain cases, SPE may produce maps with physically meaningful axes [17]). The color of the points on the map reflects the quality

Table 3
Models reported for the PYR data set

No.	Features[a]	$\mu(R)$[b]	$\mu(R_{\mathrm{cv}})$[c]
1	0,5,10,11,19,22	0.951	0.777
2	0,2,5,10,19,22	0.951	0.786
3	5,8,11,16,19,22	0.949	0.799
4	0,4,5,10,19,22	0.948	0.771
5	5,8,10,11,19,22	0.947	0.773
6	0,1,5,10,19,22	0.946	0.790
7	0,5,10,16,19,22	0.945	0.808
8	4,5,8,10,19,22	0.944	0.781
9	0,5,10,19,21,22	0.944	0.726
10	0,5,8,10,19,22	0.944	0.804
11	0,5,10,14,19,22	0.943	0.775
12	2,5,6,16,19,22	0.941	0.793
13	5,8,10,19,20,22	0.941	0.749
14	0,5,10,19,22,25	0.941	0.726
15	0,5,10,12,19,22	0.941	0.772
16	0,5,16,19,21,22	0.940	0.742
17	2,5,6,10,19,22	0.940	0.750
18	5,8,10,18,19,22	0.940	0.754
19	0,3,5,10,19,22	0.940	0.777
20	5,8,10,13,19,22	0.939	0.771
21	0,5,6,10,19,22	0.939	0.779
22	0,4,6,10,19,22	0.939	0.748
23	5,7,10,11,19,22	0.938	0.784
24	1,4,5,10,22,25	0.938	0.618
25	5,6,10,18,19,22	0.937	0.731
26	3,5,11,16,19,22	0.937	0.792
27	2,5,8,9,19,22	0.936	0.730
28	5,8,11,19,22,25	0.936	0.632
29	3,5,6,10,19,22	0.936	0.774
30	5,11,16,19,21,22	0.936	0.732
31	5,6,16,19,21,22	0.935	0.729
32	5,6,10,19,21,22	0.935	0.702
33	0,3,5,16,19,22	0.934	0.787
34	5,8,16,19,21,22	0.934	0.748
35	3,5,6,16,19,22	0.934	0.781
36	1,5,10,22,25,26	0.933	0.626
37	0,3,10,19,22,23	0.932	0.682
38	5,10,11,19,22,25	0.932	0.686
39	1,4,10,22,23,25	0.932	0.628
40	5,10,11,19,21,22	0.930	0.717
41	1,2,5,8,19,22	0.930	0.808
42	5,6,12,16,19,22	0.930	0.771
43	1,2,4,5,19,21	0.930	0.778
44	5,8,9,19,22,25	0.929	0.659
45	0,4,16,19,22,25	0.928	0.739
46	2,3,5,6,19,22	0.927	0.787
47	4,5,10,11,19,22	0.927	0.710
48	1,2,3,5,19,22	0.926	0.795
49	1,10,19,22,25,26	0.925	0.551
50	4,5,10,19,20,22	0.925	0.728
51	2,3,5,16,19,22	0.924	0.781
52	0,1,2,16,19,21	0.920	0.785
53	3,5,8,10,21,25	0.916	0.639

[a] Zero-based indices of features comprising the models identified by the particle-swarms and simulated annealing algorithms.

[b] Mean training R.

[c] Mean leave-one-out cross-validated R.

of the corresponding models (either R or R_{cv}), as given by the color scale on the right side of the plot. We first present the results obtained with the correlation-based method. For the AMA data set, the 13 reported models shown in Fig. 2 are separated into well-defined clusters according to their feature composition and correlation between features. For example the models enclosed in circle 1 have features 31 and 49 in common and differ in highly correlated features 34, 35, 37, and 38. Likewise, the models enclosed in circle 2 have features 49 and 51 in common, and differ in the same features 34, 35, and 37. Thus, the difference between the two clusters of models is due to the difference between features 31 and 51. Also highlighted in Fig. 1 is a model discovered by the particle swarm algorithm and corresponding to features 3, 4, and 49. This model has the highest R and one of the highest R_{cv} among all models examined. It is also well separated from the other predictors with high values of both R and R_{cv}, which suggests that it captures a distinct trend in the data. In fact, the grossly under-determined nature of most QSAR data sets often results in several different but equally plausible models with comparable predictivity. Recent research has shown that one can exploit this diversity in order to minimize uncertainty and produce more stable and accurate predictors through the process of aggregation [14]. Model diversity is not obvious from the values of the descriptors, and is difficult to assert without inspecting the nonlinear map.

There are 82 models reported for the BZ data set. While we again observe grouping of the models on the map in Fig. 3, color-coding by either R or R_{cv} does not seem to produce a coherent picture. Red points are mixed with blue ones, i.e. the similarity of the features does not correlate well with the quality of the models. However, a closer view of the region enclosed by a rectangle in Fig. 3a, which encompasses mixed quality models that are similar to each other in terms of the underlying features, shed some light on this behavior, as illustrated in Fig. 4. Regions labeled 1, 2, 3, 4, and 5 contain models that share features 1, 9, and 15, also contain one of the highly correlated features 3 and 5 ($q_{3,5} = 0.04$), one of the highly correlated features 0 and 6 ($q_{0,6} = 0.06$) and differ in one remaining feature. This overall similarity due to the five common features is manifested in the close proximity of the respective models on the nonlinear map in Fig. 3. The local differences amplified in Fig. 4 stem from a single feature that varies within that region of structure–activity model space. In particular, region 1 contains two models that differ in correlated features 23 and 27 ($q_{23,27} = 0.19$), region 2 contains two models that include correlated features 25 and 26, region 3 contains two models that share feature 8 and differ in the highly correlated features 3 and 5, region 4 contains two models that differ in highly correlated features 10 and 12 ($q_{10,12} = 0.04$), and region 5 contains models that differ in correlated features 16 and 20. The models in region 2 and one of the models in region 4 have much lower values of R than the rest of the models, as indicated by the blue color. It follows that a change of feature 8 to feature 25 or 26, or replacement of feature 10 with feature 12 leads to models with inferior training R. The latter is somewhat unexpected since features 10 and 12 are strongly correlated. In fact, the two models in region 4 have very close R_{cv} values. While it has been

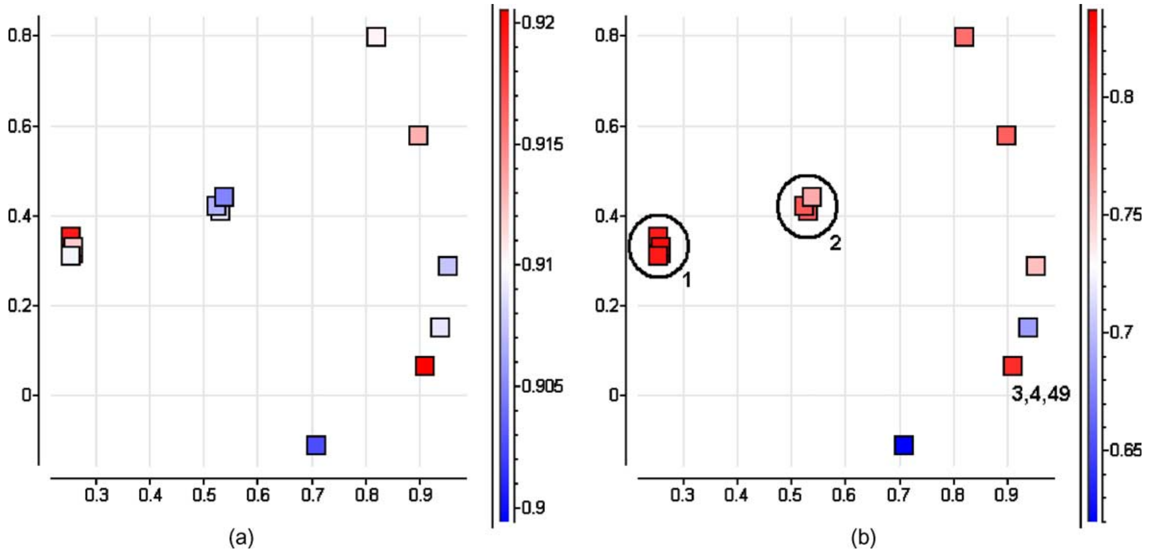

Fig. 2. A map of 13 models found by feature selection algorithms for the AMA data. Each square represents a model that corresponds to a certain feature subset. The models are colored by the mean (a) training and (b) leave-one-out cross-validated Pearson correlation coefficient, as shown on the color scale to the right of each plot.

shown that R and R_{cv} are rather poorly correlated [13], the reason why a substitution of a feature with an equivalent one impacts R is unclear.

The map depicting the 55 models for the PYR data set is shown in Fig. 5. Here, there is a visible trend for the quality of the models to change gradually across the model space. The largest population of high quality models is located near the center of the map, both in terms of R and R_{cv}. However, once again, the maps reflect poor correlation between the training R and cross-validated R_{cv}. In particular, the map in Fig. 5b with coloring by R_{cv} suggests that good predictive models are actually more diverse than one would expect from looking at the map in Fig. 5a colored by R: more predictive models (points that have more red color) are concentrated near the center of the plot in Fig. 5a, while in Fig. 5b more predictive models cover a much larger area of the plot.

Note that the range of the color scale on the right side of the plot reflecting the colors that correspond to specific values of R and R_{cv} does not necessarily cover the whole range of values in the data and can be selected manually to produce the greatest contrast in order to distinguish more effectively models of different quality. For example in Fig. 3b the color scale is set in such a way that allows distinguishing between models with R_{cv} values between about 0.83 and

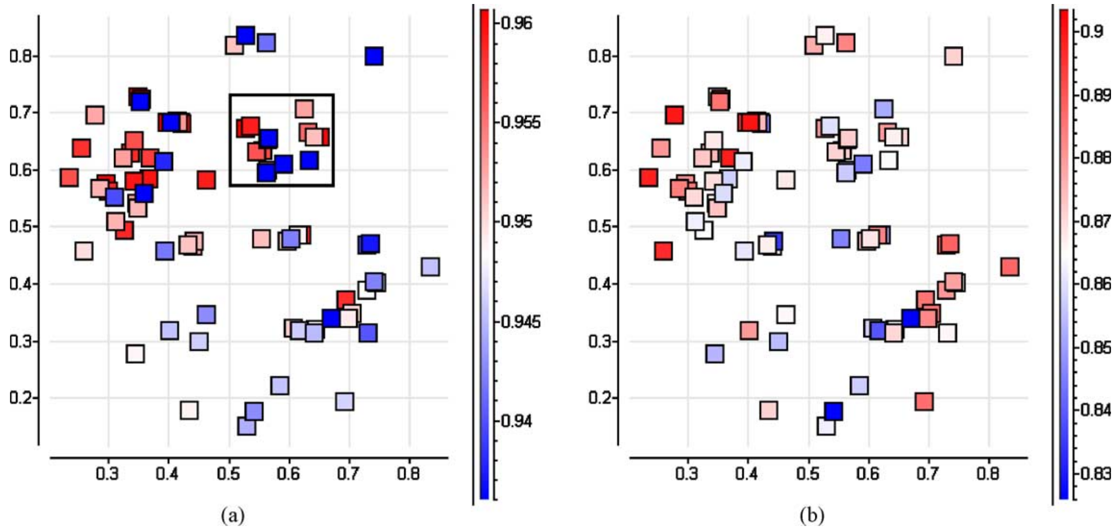

Fig. 3. A map of 82 models found by feature selection algorithms for the BZ data. Each square represents a model that corresponds to a certain feature subset. The models are colored by the mean (a) training and (b) leave-one-out cross-validated Pearson correlation coefficient, as shown on the color scale to the right of each plot. A close-up of the area inside the rectangle is presented in Fig. 4.

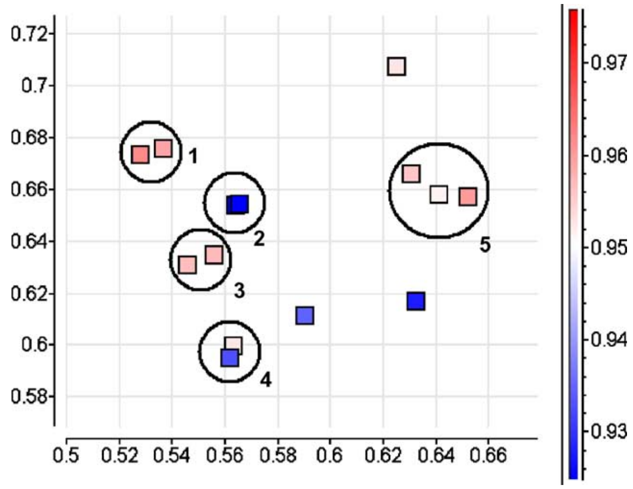

Fig. 4. A close-up of the rectangular area shown in Fig. 3a.

0.9, even though the smallest R_{cv} value in the data is 0.519 (see Table 2).

SPE maps of the model space calculated using the binary similarity measure are quite different both qualitatively and quantitatively. Fig. 6 shows the SPE maps calculated for the AMA data set using the correlation-based and binary similarity measures. While the general topology of the model spaces is almost the same, clusters 1 and 2 are not nearly as tight on the map that is based on the binary similarity measure because this approach does not recognize that the one feature that varies within each cluster is highly correlated from one model to another. Similar behavior can be observed on the maps for the BZ data set presented in Fig. 7. Models enclosed in circle A on the map in Fig. 7a computed using the correlation-based similarity are shown as blue squares on the map in Fig. 7b calculated with the binary similarity measure. These models form a very tight cluster on the correlation-based map due to the highly correlated features that they contain. However, on the binary map these models end up in three different clusters circled in Fig. 7b, each of which is formed by models that share five out of six features. Models in region 1 share features 1, 3, 4, 9, and 15; models in region 2 share features 1, 4, 5, 9, and 15; and models in region 3 share features 1, 4, 5, 9, and 20. On the other hand, points located in region 1 circled on the map in Fig. 7b are shown as red circles on the map in Fig. 7a. These points form a cluster on the map based on the binary similarity measure because they all share five features and differ in the remaining one. When placed on the map that takes into account feature correlation, these models end up spread around according to the correlation of their features with those in the other models.

The PYR data set is especially interesting because out of the 351 pairs of 27 features there are only five pairs that are highly correlated. In the extreme case, where all features are completely uncorrelated, the two similarity measures become essentially the same because then only pairs of identical features contribute to the similarity measure (see Eq. (1)). The PYR data set is very close to this limit, and therefore, the SPE maps calculated by the two methods are generally very similar (Fig. 8). However, even in this case, the mapping of the models that contain correlated features is drastically different. The same six models are shown as circles on both maps. Because each of these models contains one or two features that are highly correlated to another feature in other models, all six models are perceived to be sufficiently similar to each other and are located in the same region of the correlation-based SPE map (Fig. 8a). In contrast, the binary similarity measure completely missed

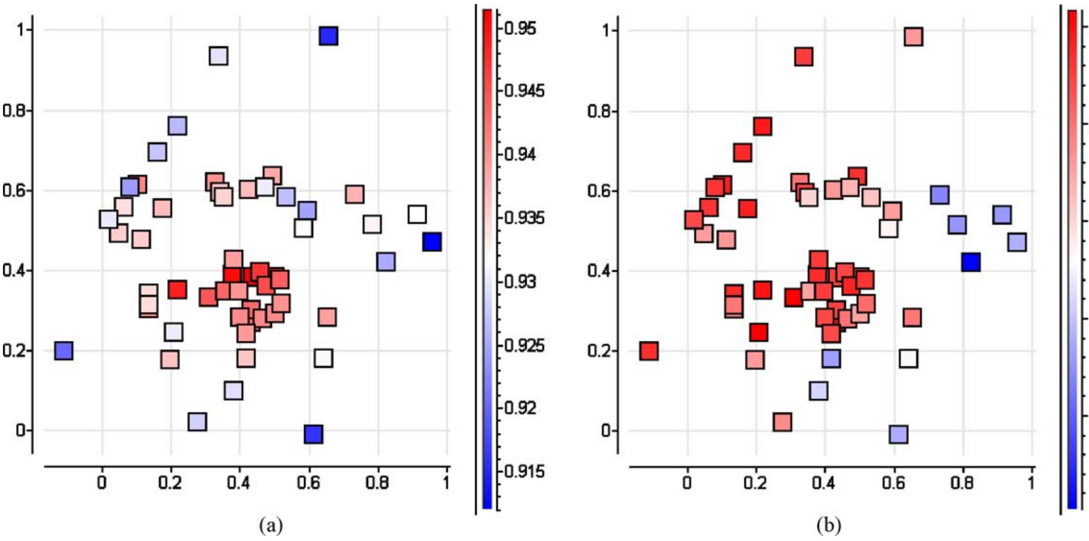

Fig. 5. A map of 53 models found by feature selection algorithms for the PYR data. Each square represents a model that corresponds to a certain feature subset. The models are colored by the mean (a) training and (b) leave-one-out cross-validated Pearson correlation coefficient, as shown on the color scale to the right of each plot.

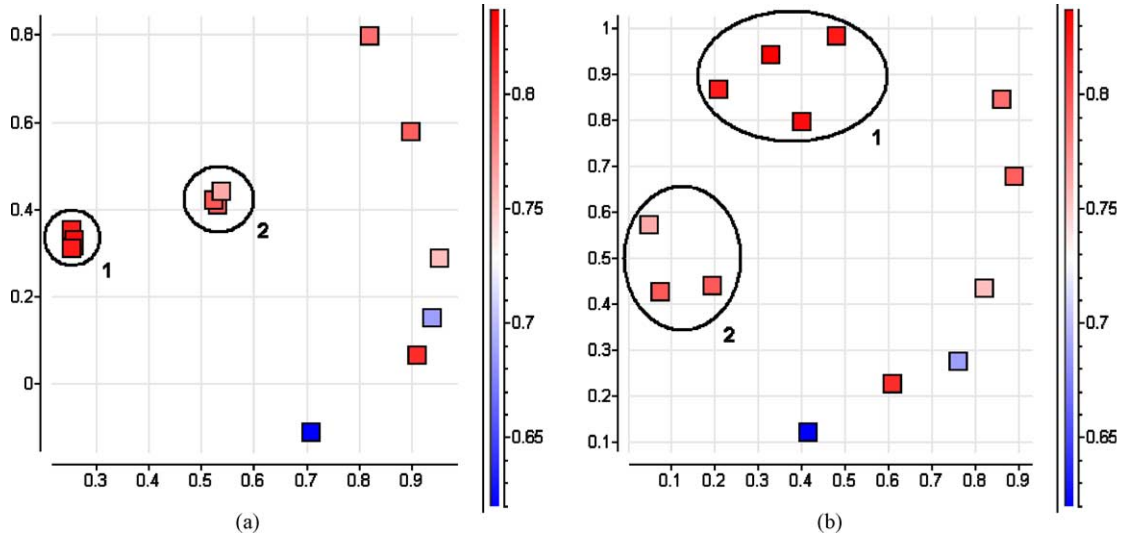

Fig. 6. Maps of the models for the AMA data set computed with (a) the correlation-based and (b) the binary similarity measure. While the overall clusters are preserved, the inter-model distances are greatly affected. The models are colored by the mean leave-one-out cross-validated Pearson correlation coefficient, as shown on the color scale to the right of each plot.

the close relationships between the models and as a result, they were placed in two remote groups. Since the distance between models on the map should reflect their similarity, such an inconsistency proves that the binary metric is insufficient.

We would like to stress that the correlation-based similarity measure reflects similarity not between points in multi-dimensional space, but rather between point sets. The specific form of the similarity measure that we proposed is just one possible way to measure point set similarity (in this case, point sets are feature subsets) and provides a simple way of computing the similarity between models in conjunction with a feature correlation matrix. Perhaps, the most widely known point set similarity measure is the Hausdorff distance that is defined as the largest distance from any point in one set to its nearest neighbor in the other set. However, the Hausdorff distance is not applicable in this case because it does not have the desirable averaging properties and it is extremely sensitive to outliers.

In the proposed similarity measure, all features are treated equally without regard to whether or not they contribute to the predictivity of the model. This provides a means of describing the similarity of the feature sets independently of the feature selection algorithm or the measure used to evaluate the model quality, such as R or $R_{\rm cv}$. While the feature set similarity is represented on the 2D map by the distance between points, the relative model quality is reflected by the color of the points. The contribution of a feature to the

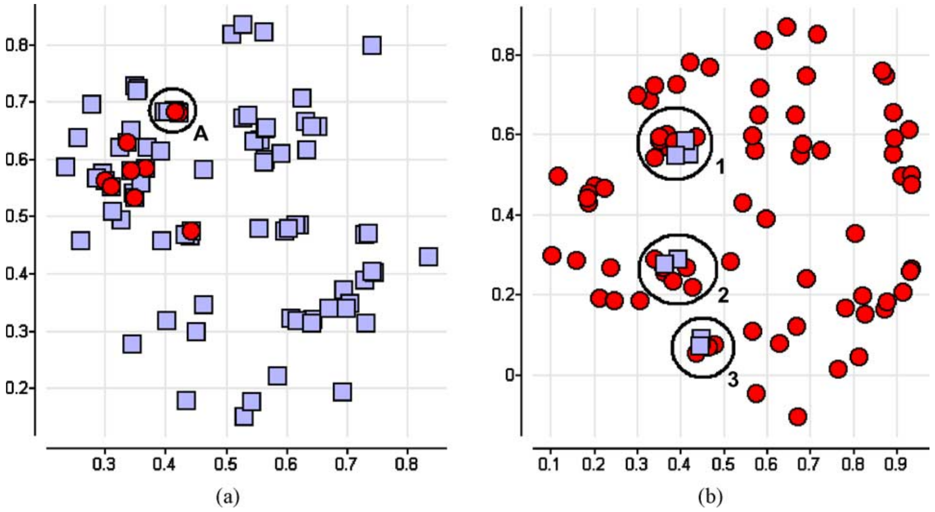

Fig. 7. Maps of the models for the BZ data set computed with (a) the correlation-based and (b) the binary similarity measure. Models located in region A on map (a) are shown as cyan squares on map (b). Models located in region 1 circled on map (b) are shown as red circles on map (a).

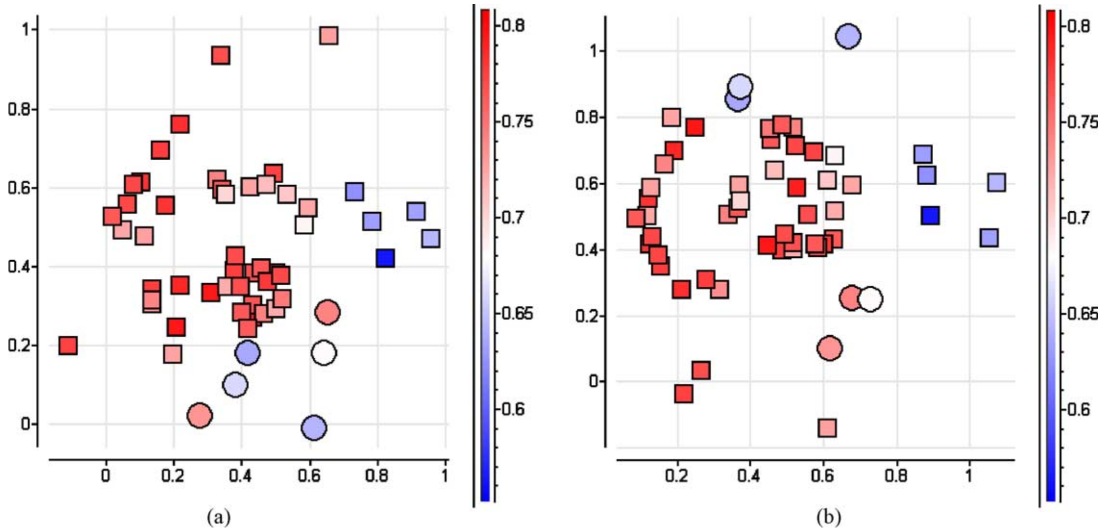

Fig. 8. Maps of the models for the PYR data set computed with (a) the correlation-based and (b) the binary similarity measure. The same six models are shown as circles on both maps. The models are colored by the mean leave-one-out cross-validated Pearson correlation coefficient, as shown on the color scale to the right of each plot.

predictivity of any given QSAR model may be evaluated by the feature selection algorithm, however, it is unclear whether this information can be meaningfully extracted from the feature selection results, comprised in general of only the feature sets and the corresponding model quality, and incorporated into the feature set similarity measure.

We would also like to note that the SPE algorithm as described here attempts to preserve distances between all pairs of points, which is usually not possible unless the dimensionality of the data is less or equal to the dimensionality of the target space (2D in this case) [17]. Thus, there are always points for which the distance constraints are violated. Such violations may manifest themselves, for example by the fact that the range of distances on the map may be larger than the range of distances in the original distance matrix. Embedding into three-dimensional space normally produces significantly fewer distortions and can be more useful for interactive analysis. In addition, application of a variation of the SPE algorithm that only preserves distances smaller than some predefined cutoff value [17] may provide additional level of detail for groups of closely related models.

The algorithm described above is currently limited to models based on equal number of selected features. Extension of this approach to sets of features of different cardinality presents a challenge and is the subject of an ongoing investigation.

4. Conclusions

The introduced QSAR model similarity measure provides the basis for a new approach to the visualization and analysis of QSAR models obtained by feature selection algorithms. Because it takes into account the correlation between the underlying features, this approach captures and presents in an intuitive form close relationships between models that would otherwise be difficult to discover. The new similarity measure also makes possible further analysis of the model space using a variety of similarity, diversity and clustering algorithms.

Acknowledgements

We are grateful to Dr. Walter Cedeño for helpful discussions, and to Dr. F. Raymond Salemme of 3-Dimensional Pharmaceuticals, Inc. for his support of this work.

References

[1] D.L. Selwood, D.J. Livingstone, J.C.W. Comley, A.B. O'Dowd, A.T. Hudson, P. Jackson, K.S. Jandu, V.S. Rose, J.N. Stables, Structure–activity relationships of antifilarial antymycin analogues: a multivariate pattern recognition study, J. Med. Chem. 33 (1990) 136–142.

[2] D.C. Whitley, M.G. Ford, D.J. Livingstone, Unsupervised forward selection: a method for eliminating redundant variables, J. Chem. Inf. Comput. Sci. 40 (2000) 1160–1168.

[3] O. Kikuchi, Systematic QSAR procedures with quantum chemical descriptors, Quant. Struct–Act. Relat. 6 (1987) 179–184.

[4] D.J. Livingstone, E. Rahr, Corchop—an interactive routine for the dimension reduction of large QSAR data sets, Quant. Struct–Act. Relat. 8 (1989) 103–108.

[5] J.M. Sutter, S.L. Dixon, P.C. Jurs, Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing, J. Chem. Inf. Comput. Sci. 35 (1995) 77–84.

[6] D.R. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.

[7] S. So, M. Karplus, Evolutionary optimization in quantitative structure–activity relationship: an application of genetic neural networks, J. Med. Chem. 39 (1996) 1521–1530.

[8] A. Yasri, D. Hartsough, Toward an optimal procedure for variable selection and QSAR model building, J. Chem. Inf. Comput. Sci. 41 (2001) 1218–1227.

[9] K. Hasegawa, Y. Miyashita, K. Funatsu, GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.

[10] B.T. Luke, Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 1279–1287.

[11] S. Izrailev, D.K. Agrafiotis, A new method for building regression tree models for QSAR based on artificial ant colony systems, J. Chem. Inf. Comput. Sci. 41 (2001) 176–180.

[12] S. Izrailev, D.K. Agrafiotis, Variable selection for QSAR by artificial ant colony systems, SAR QSAR Environ. Res. 13 (2001) 417–423.

[13] D.K. Agrafiotis, W. Cedeño, Feature selection for structure–activity correlation using binary particle swarms, J. Med. Chem. 45 (2002) 1098–1107.

[14] D.K. Agrafiotis, W. Cedeño, V.S. Lobanov, On the use of neural network ensembles in QSAR and QSPR, J. Chem. Inf. Comput. Sci. 42 (2002) 903–911.

[15] I. Borg, P.J.F. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, New York, 1997.

[16] J.W. Sammon, IEEE Trans. Comput. C-18 (1969) 401–409.

[17] D.K. Agrafiotis, H. Xu, A self-organizing principle for learning nonlinear manifolds, Proc. Natl. Acad. Sci. 99 (2002) 15869–15872.

[18] D.K. Agrafiotis, Stochastic proximity embedding, J. Comput. Chem. 24 (2003) 1215–1221.

[19] D.K. Agrafiotis, A new method for analyzing protein sequence relationships based on Sammon maps, Protein Sci. 6 (1997) 287.

[20] M. Farnum, H. Xu, D.K. Agrafiotis, Exploring the nonlinear geometry of sequence homology, Protein Sci. 12 (2003) 1604–1612.

[21] D. Coppersmith, G.B. Sorkin, Constructive bounds and exact expectations for the random assignment problem, Random Struct. Alg. 15 (1999) 113–144.

[22] D.J. Maddalena, G.A.R. Johnson, Prediction of receptor properties and binding affinity of ligands to benzodiazepine/GABA_A receptors using artificial neural networks, J. Med. Chem. 38 (1995) 715–724.

[23] J.D. Hirst, R.D. King, M.J.E. Sternberg, Quantitative structure–activity relationships: neural networks and inductive logic programming compared against statistical methods: I, the inhibition of dihydrofolate reductase by pyrimidines, J. Comput.-Aided Mol. Design 8 (1994) 405–420.

[24] D.K. Agrafiotis, R.F. Bone, F.R. Salemme, R.M. Soll, United States Patents 5,463,564 (1995); 5,574,656 (1996); 5,684,711 (1997); and 5,901,069 (1999).