J Mol Model (2003) 9:66–75
DOI 10.1007/s00894-002-0110-0

ORIGINAL PAPER

S. F. Badreddin Abolmaali · Claude Ostermann · Andreas Zell

The Compressed Feature Matrix–a novel descriptor for adaptive similarity search

Received: 20 June 2002 / Accepted: 12 November 2002 / Published online: 5 February 2003

Abstract The Compressed Feature Matrix (CFM) is a new molecular descriptor for adaptive similarity searching. Depending on the requirements, it is based on a distance or geometry matrix. Thus, the CFM permits topological and three-dimensional comparisons of molecules. In contrast to the common distance matrix, the CFM is based on features instead of atoms. Each kind of these features may be weighted separately, depending on its (estimated) contribution to the biological effect of the molecule. In this work, we show that the CFM allows us to adapt similarity evaluations to particular ligands as well as to classification requirements. The CFM method is analyzed regarding correctness, adaptivity and speed. Applying the basic setting of feature weights, the similarity evaluations using the CFM on the one hand and the Tanimoto coefficient together with MACCS Keys on the other yield similar results. However, in contrast to the latter method, the CFM even permits us to focus on small parts of molecules to serve as a basis for similarity. Accordingly, we have achieved striking results not only by readjusting the feature weights with regard to the scaffold but also to the side chain of the respective target. The results of the latter run turned out to be rather independent of the molecular scaffold. Hence, the CFM is suitable not only for common similarity evaluation, but also for techniques such as lead or scaffold hopping.

Electronic Supplementary Material Supplementary material is available for this article if you access the article at http://dx.doi.org/10.1007/s00894-002-0110-0. A link in the frame on the left on that page takes you directly to the supplementary material.

S. F. B. Abolmaali (✉) · A. Zell
Department of Computer Science,
University of Tuebingen, Sand 1,
72076 Tübingen, Germany
e-mail: abolmaali@informatik.uni-tuebingen.de
Tel.: +49-7071-2978979

C. Ostermann
ALTANA Pharma AG,
Byk Gulden Straße 2, 78467 Konstanz, Germany

Keywords Similarity · Descriptor · Computer chemistry features · Scaffold hopping

Abbreviations *CFM:* Compressed Feature Matrix · *HTS:* high throughput screening · *MCS:* maximum common substructure · *col./cols:* column/columns · *MAO-A:* monoamine oxidase A

Introduction

Similarity searching is an essential task in pharmaceutical research, especially in high throughput screening (HTS) analysis, scaffold/lead hopping and lead structure optimization. Commonly, methods used for the evaluation of molecular similarity are divided into two main groups, according to whether they are based on topological features or on three-dimensional structures. Furthermore, there are two basically different techniques for the comparison of molecules. [1] On the one hand, some descriptors are used to compare molecules by pairs, such as molecular shape similarity descriptors [1, 2] and the maximum common substructure (MCS). [1, 3] On the other hand, there are various types of descriptors that are calculated independently for each molecule, e.g. BCUT descriptors, [1, 4, 5] autocorrelation descriptors [1, 6] and substructure descriptors [1] such as hashed fingerprints, [7] molecular holograms [8] and atom pairs. [9]

Most of the descriptors and methods mentioned focus on different aspects of the molecules to be compared. The large number of approaches shows that a major question concerning similarity searching is on what basis should molecular similarity be evaluated? In many pharmaceutical applications ligands are rated similar if they exhibit similar biological effects. Unfortunately, the 3D structure or at least structural elements of the particular receptor are often unknown. Then, the molecules tested are usually compared to known active ingredients. However, in these cases, problem-focused results can only be obtained if those molecular features and structures that are thought to be responsible for the effect are suitably emphasized

(therefore, limiting the description of a molecule to its atomic level is not sufficient in most cases). However, the effects investigated, and thus the important features, would normally vary according to the particular problem. With respect to these constraints, we have developed a novel descriptor for similarity evaluation, called the Compressed Feature Matrix (CFM), [10, 11] that on the one hand, describes molecules based on a user-defined set of features and, on the other hand, permits the discrete weighting of these features.

Regarding its structure, the CFM is closely related to both the distance matrix [12] and the geometry matrix. [1] However, in contrast to these, the CFM is based on features instead of atoms, which also applies to other similarity descriptors, e.g. the feature tree descriptor, [13] as well as to substructure descriptors. [1, 7, 8] The CFM is also in principle not restricted to a single set of features. This opens up the possibility of masking atoms and atomic groups that express negligible features. Therefore, the CFM requires less memory space than related matrix descriptors. Software packages such as PETRA [14] provide various methods for the calculation of molecular properties that might be used as features. With regard to the procedure of comparing molecules, the CFM, to some degree, resembles the atom-pair descriptor. However, while the specific weighting of features is an essential property of the CFM, the atom-pair descriptor does not make any assumptions about the importance of different types of functional groups or ring systems. [9]

Irrespective of the method of similarity searching used, the resulting numerical output depends on three main components: the representation of the molecules, the particular weighting scheme and the selected similarity measure. [1, 15] Different weighting schemes are reported [16, 17, 18, 19] that refer either to the descriptors used (if more than one are applied) or to the structural elements of the molecules analyzed. In this work, the structure of the CFM (i.e. the molecular representation) as well as the appendant method of similarity evaluation (the similarity measure), including the weighting of the structural features, are described. We show that the CFM facilitates similarity searching, adaptive to classification requirements and to the characteristics of particular sets of ligands. As a benchmark we use the Tanimoto coefficient applied to MACCS Keys. The latter method of similarity searching is provided by various software packages, e.g. MOE, [20] SUBSET [21] and ISIS/Base. [22]

Materials and methods

CFM structure

As a basis for the construction of a CFM, a feature set must be defined that fits the requirements of the particular problem. In this work, we refer to a set of twelve feature types: terminal carbon atoms (cat), hydrogen donor and acceptor qualities (don, acc), positive and negative charges (pos, neg), radicals (rad) and rings comprising from three to eight atoms (tri, qua, pen, hex, hep, oct).

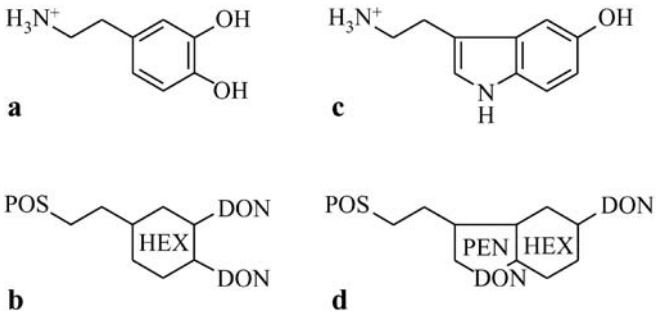

Fig. 1 Chemical structures (**a**) and (**c**) and feature graphs (**b**) and (**d**) of dopamine and serotonine

	DON	DON	POS	HEX
DON	0	3	6	1
DON	3	0	7	1
POS	6	7	0	3
HEX	1	1	3	0

a

	DON	DON	POS	PEN	HEX
DON	0	5	5	0	1
DON	5	0	7	3	1
POS	5	7	0	3	4
PEN	0	3	3	0	0
HEX	1	1	4	0	0

b

Fig. 2 The topological CFMs of dopamine (**a**) and serotonine (**b**)

In the following, feature types are displayed in lower case letters while particular features are upper case.

Using the feature set described, nonterminal carbon atoms do not occur in the CFM. This is valid because the lengths of carbon chains are considered within the distance values of the feature pairs. Furthermore, heterocycles are not specified explicitly because they are implicitly expressed by the feature of the respective cycle plus the feature(s) of the heteroatom(s). An example of this is shown in Fig. 1c, d where the chemical structure and the feature graph of serotonine are displayed. Here, the pyrrole ring of serotonine is represented by the features PEN (since it is a five membered ring) and DON (which is the feature of the comprised nitrogen atom). Within the corresponding CFM, the fact that the nitrogen atom is a member of the ring is indicated by the distance value of zero between the two participating features (Fig. 2b, row 1 (DON), column 4 (PEN)).

Corresponding to its structural relationship to the distance matrix the CFM **C** is defined as the concatenation

$$\mathbf{C} := \begin{pmatrix} \mathbf{f} \\ \mathbf{D} \end{pmatrix} \tag{1}$$

where the row vector $\mathbf{f} := (F_k)_{k=1}^{n}$ contains the features and where $\mathbf{D} := (d_{ij})_{i,j=1}^{n}$ is the respective distance or geometry matrix based on these features. Therefore, according to the particular problem, the matrix may be based on either topological or Euclidean distances.

As an example, Fig. 1 shows the chemical structure (a) and the feature graph (b) of dopamine. Its CFM is shown in Fig. 2a.

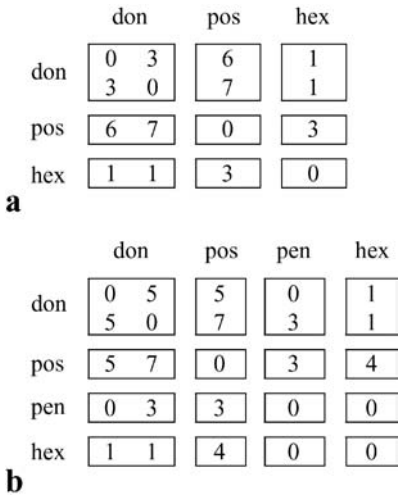

Fig. 3 The topological submatrices of the CFMs of dopamine (**a**) and serotonine (**b**)

Similarity evaluation

Since similarity evaluation is performed by the comparison of CFMs, two molecules are considered identical if their CFMs are identical, irrespective of whether they comprise the same atoms and atomic groups. In contrast, different kinds and numbers of features as well as varying topologies diminish the degree of similarity. The question that this section deals with is in what way and to what extent do the differences between two CFMs influence the degree of similarity between two molecules? As an example, the successive steps of the similarity evaluation between dopamine and serotonine are represented. At first, both CFMs are split into submatrices. In this context, a submatrix holds the distances between all members of two particular kinds of features, e.g. don/don or cat/hex. Figure 3 shows the submatrices of dopamine (a) and serotonine (b).

Subsequently, the entries of corresponding submatrices are linearly aligned in such a way that the sum of the deviations between the entries is minimal. Since the CFM is a symmetric matrix, only the upper triangular matrix is considered. The submatrix alignment of dopamine and serotonine is given in Table 1.

Obviously, there are some corresponding entries in both ligands while some other entries have no counterpart. In the following, the effects of these interrelations are described in the context of similarity evaluation.

Each matrix entry e of a CFM stands for a substructure that comprises two features connected by a chain of a certain length. Aligned entries stand for similar substructures (that comprise the same features) within both molecules. As a rule, the more corresponding entries that occur in the submatrix alignment and the smaller the respective deviations are, the more similar are the molecules compared. Since the importance of the different kinds of features depends on the particular problem, the contribution of matching substructures to molecular similarity is extended by the product of the discrete weighting factors w of both features. Thus, the similarity s^+ between two molecules M and M' is at first defined as

$$s^+ = \sum_{i=1}^{n} \frac{w^x \cdot w^y}{\left(e_i^{\mathrm{M}} - e_i^{\mathrm{M}'}\right)^2 + 1} \tag{2}$$

Here, n is the number of corresponding entries in M and M', w^x and w^y are the weights of the feature types x and y defining the particular submatrices and e^{M} and $e^{\mathrm{M}'}$ are the aligned entry values within these submatrices.

To achieve a higher emphasis on the occurrence of matching substructures than on the deviations, one may alternatively use the absolute deviation instead of its square:

$$s^+ = \sum_{i=1}^{n} \frac{w^x \cdot w^y}{\left|e_i^{\mathrm{M}} - e_i^{\mathrm{M}'}\right| + 1} \tag{3}$$

In contrast to the set of corresponding substructures, there are some entries that lack a counterpart in the other molecule. These features decrease the degree of similarity. Therefore, their respective weights v are summed up to the penalty term s^-,

$$s^- = \sum_{j=1}^{m} v_j^z \tag{4}$$

Here, m is the number of unpaired features and v^z is the penalty weight of the respective feature type z. In most cases, the values of the weighting factors w and v would be different for the same kind of feature, because the presence of a certain kind of feature might be more meaningful to receptor binding than its absence, and vice versa.

The overall similarity s between two molecules is then expressed as

$$s = s^+ - s^- \tag{5}$$

Because of its structure and the algorithm of similarity evaluation, the CFM does not depend on any kind of numbering and it is invariant against translation, rotation and the center of gravity.

As an example, dopamine and serotonine are compared using the following set of weighting factors w/v : don: 5/0, pos: 5/0, pen: 2/0, hex: 10/0. With these values the similarity between dopamine and serotonine equals 5+12.5+25+50+50+25–0=167.5 (using Eq. (2) for the calculation of s^+). While a single scalar is not meaningful by itself, a target-specific scale is defined. Thereby, the upper limit of the scale is determined by comparing the target to itself. The resulting value stands for identical CFMs and thus for 100% similarity. In general, the similarity of any tested molecule to a known ligand is equal to or less than 100%. However, since the comparison results consist of positive and negative terms, similarity values less than zero might occur. Therefore, the results are normalized to the open interval $]-\infty; 100]$.

The similarity of dopamine (which serves as the target in this example) to itself is 25+25+25+50+50+50=225. Accordingly, the normalized similarity (normS) between dopamine and serotonine is 74.4. Hereby, emphasis was placed on the occurrence of the feature types hex, pos and don, assuming that these features are essential for receptor binding. Obviously, other values of w and v might cause totally different results.

A similarity evaluation based on only two molecules has little significance, particularly since it may be influenced arbitrarily. In fact, the advantage of our method becomes evident on the evaluation of large data sets. Used in this way, the result of a

Table 1 Submatrix alignment of dopamine and serotonine

Feature types	don,don	don,pos		don,pen		don,hex		pos,pen	pos,hex	pen,hex
Dopamine	3	6	7	–	–	1	1	–	3	–
Serotonine	5	5	7	0	3	1	1	3	4	0
$\lvert e^{\mathrm{M}} - e^{\mathrm{M}'} \rvert$[a]	2	1	0	–	–	0	0	–	1	–

[a]The term $\lvert e^{\mathrm{M}} - e^{\mathrm{M}'} \rvert$ specifies the deviation between corresponding submatrix entries of the two molecules M and M'

similarity search is a list of the database molecules ordered according to their likeness to the target structure.

The procedure of similarity evaluation is conducted in two steps. At first, a target structure is determined, and the weighting factors w and v are adjusted manually, normally with regard to those features that are (thought to be) responsible for its biological effects as e.g. receptor binding or toxicity. Furthermore, other problem-specific classification requirements (charge, hydrophobicity, size etc.) may affect the weighting of the features. As a rule, important feature types are assigned values larger than one, while unfavorable features may be penalized by negative values. In the second step, a query database is searched for the target with these weighting factors applied. Such a run is successful if those structures are ranked highest that are most similar to the target structure, referring to the particular problem. This may be verified by adding known active ingredients to the database that are similar to the target regarding structural and binding properties. If these control structures are classified as expected, other high ranking compounds may be suitable for further chemical and pharmaceutical investigation. If not, the feature weights are readjusted for another search.

Similarity searching with CFMs is based on the optimal alignment of corresponding submatrices. The more local similarities that occur between the two molecules to be compared, the higher the positive similarity term s^+. If the submatrices of the two compounds are of about equal size, the matching entries (and thus the value of s^+) clearly result from a high global similarity. However, large database molecules may comprise submatrices that contain significantly more entries than the corresponding submatrices of the target. In these cases, high values of s^+ may be achieved even if the global analogy of the ligands compared is comparatively low. One way to solve this problem of potential "false positives" is to determine negative feature weights for unfavorable features (see earlier). In addition to this, the similarity results of those compounds that comprise more features than the target structure may optionally be standardized to the number of features of the latter. The resulting similarity value $\bar{s}$ is defined as

$$\bar{s} = \frac{s \cdot |F|^{\mathrm{T}}}{|F|^{\mathrm{D}}} \tag{6}$$

where $|F|^{\mathrm{T}}$ and $|F|^{\mathrm{D}}$ are the numbers of features occurring in the target and in the database molecule, respectively.

Software

The concept of the CFM is realized by the software COFEA (**Co**mpressed **Fea**ture Matrix) which is implemented in Java 2, JDK version 1.30. Since Java is a platform-independent programming language, the software runs under different operating systems.

COFEA provides two main independent program modules. The first one parses MDL Mol files [23, 24] and transforms them into CFM files, storing the molecules as Compressed Feature Matrices. This is especially significant for large amounts of data because the CFM data format takes about ten times less storage capacity than the MDL Mol file format. The second program module performs the similarity search, using CFM files as its input data format. There are several ways to influence the searching process. As mentioned earlier, the most important parameters are the feature weights assigned to atoms and atomic groups. Furthermore, the similarity values of compounds that comprise more features than the target structure may optionally be standardized to the size (i.e. the number of features) of the target. Both of these properties directly affect the quality of similarity evaluation. In addition to these, there are other parameters regarding computing time and storage requirements. Thus, COFEA permits the determination upper and lower limits for the number of each kind of feature to occur in the database molecules, i.e. the maximum and minimum allowed number of elements per feature group. If a database molecule does not fit these specifications, it will be precluded from further consideration. The effect of this kind of preselection on computing time was determined using a Windows-based computer with 384 MB RAM and an 850 MHz CPU.

Data sets

In this work, the software COFEA and with it the concept of the CFM are analyzed regarding correctness, adaptivity (including the handling of database molecules that are larger than the target ligand) and speed. Therefore, we used three data sets that represent different subsets of two reference databases. The first reference database contains 72 available MAO-A inhibitors, the second one comprises 8,655 active ingredients of different activity classes. (Both databases were composed by Michael Bieler.) The basis for the latter was a data set supplied by Tocris Cookson Ltd [25] that comprises 826 biologically active ligands, classified into 112 activity classes (regarding the different (sub-) types of the respective receptors). This database was enlarged to its final size by adding structures from literature data and commercial databases. The new compounds were selected on the basis of 2D Unity fingerprints [26] with the SYBYL tool SELECTOR, [27] using various ligands of different activity classes as targets. The precondition for the selection of a new database entry was a minimum Tanimoto coefficient of 0.8. The three test data sets that were used for the evaluation of the CFM were selected as follows.

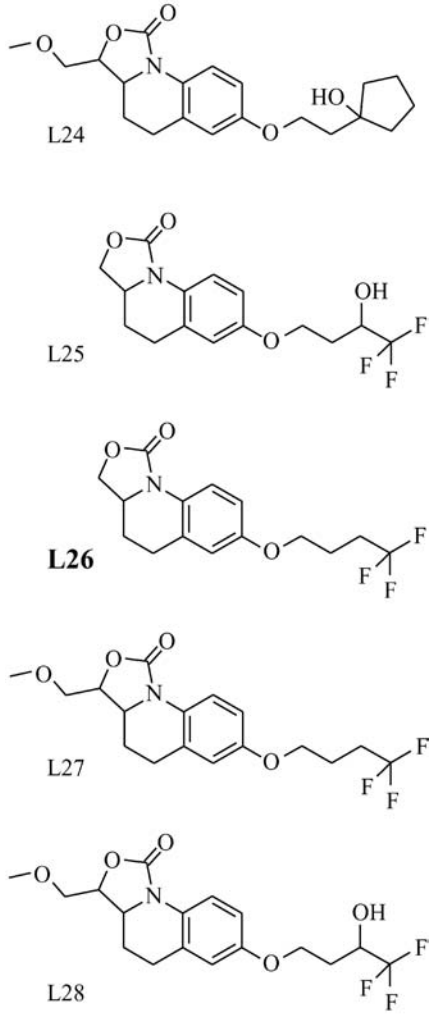

Fig. 4 The group of five MAO-A inhibitors (G1_L26) comprising the same scaffold as L26

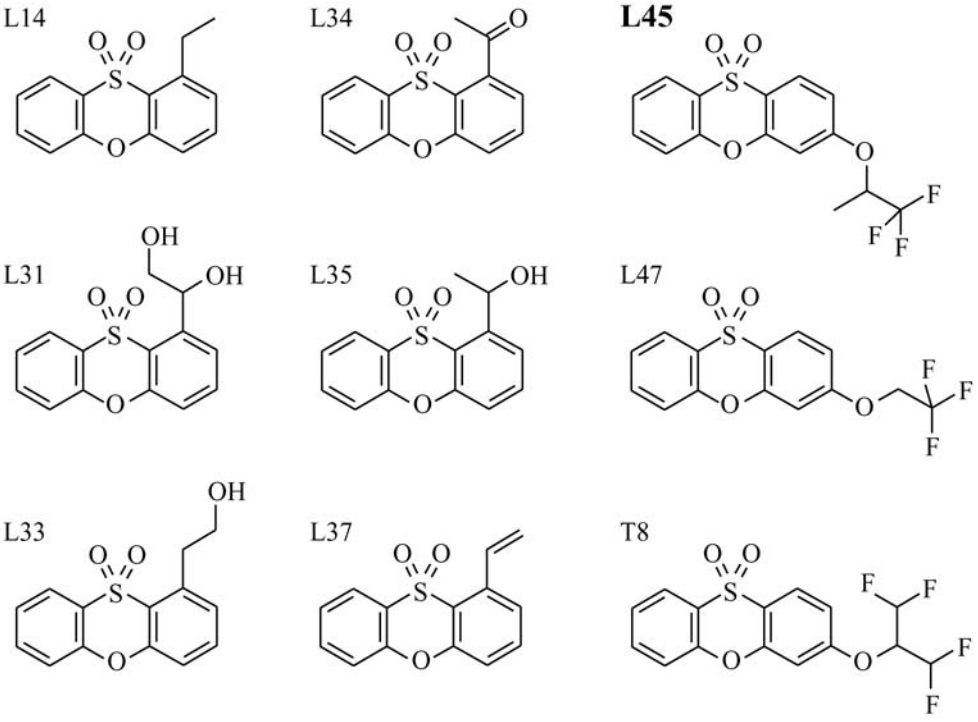

Fig. 5 The group of nine MAO-A inhibitors (G2_L45) comprising the same scaffold as L45

Test data set A serves as a basis for the evaluation of adaptivity. It contains the whole database of MAO-A inhibitors, as well as 488 non-MAO-A inhibitors with molecular weights between 75 and 295. The two targets that are used in this context are the MAO-A inhibitors L26 (MW=299) and L45 (MW=333), each of them representing a group of five and nine MAO-A inhibitors (G1_L26 and G2_L45), respectively (Figs. 4 and 5). The elements of each of these groups have particular cyclic scaffolds in common, yet with different side chains.

To illustrate the handling of molecules that are larger than the respective target, 300 compounds between 300 and 500 Da plus the 72 MAO-A inhibitors were combined in test data set L. In this context, the size of a molecule concerns the number of its features $|F|$. Finally, test data set S is used for the evaluation of computing time. It represents a subset of 8,460 compounds selected from the second reference database. The molecular weights of these structures range from 33 to 800. The SDfiles of the three test data sets described are available in the supplementary material.

Results

Adaptivity

At first, the adaptivity of the CFM-based algorithm of similarity searching was evaluated. Therefore, the MAO-A inhibitor L26 was compared to test data set A three times, each time with another sense of similarity and thus using different sets of weighting factors. In the first run, the positive weighting factors w^f were 1 and the negative weighting factors v^f were zero for all feature types. In this basic setting, all features are assumed to be equally important, which means that the search is performed without any problem specifications and without penalty terms ($s=s^+$). In the following, only those weighting parameters will be specified that differ from their basic values. The second set of weights was adjusted with the objective of finding structures containing the same (or at least similar) cyclic scaffold as the target L26. The scaffold is composed of the features types acc, pen and hex, and accordingly the weights of these features were adjusted. However, while the feature types pen and hex only occur in the scaffold, acc is the main feature type of the side chain. Therefore, emphasis was especially placed on pen and hex, setting the feature weights as follows: $w^{acc}=2$; $w^{pen}=10$; $w^{hex}=10$. In the third run, test data set A was searched for compounds similar to the inhibitor L26, especially concerning its side chain. The latter is, on the one hand, predominantly composed of hydrogen bond acceptors, on the other hand, it is connected to the scaffold by the feature type hex. Accordingly, the weighting factors for acc and hex were readjusted ($w^{acc}=10$; $w^{hex}=5$). For each of these runs, the 50 high-scoring compounds are shown in Table 2 (cols.2–4). The complete results can be seen in the supplementary material. Finally, as a benchmark, the search was performed using MACCS Keys as the molecular fingerprints and the Tanimoto coefficient as the similarity metric (Table 2, col. 1; supplementary material). For this run, we used the software program MOE. [20]

With the latter method, 28 of a total of 72 MAO-A inhibitors are found within the first 50 compounds of test data set A (Table 2, col. 1). A similar result is achieved using the CFM-based searching method with the basic setting of weighting factors. Here, 31 of the 50 high scoring structures are MAO-A inhibitors (Table 2, col. 2). The readjustment of the feature weights towards the cyclic backbone of target L26 yields a totally different result (Table 2, col. 3). Only the five inhibitors that

Table 2 Results of four different similarity evaluations of target L26 within test data set 'A'. The compounds that fit the problem specifications are bold face

Searching method	Tanimoto		CFM		CFM		CFM	
	MACCS Keys		w^{f}=1; v^{f}=0		w^{acc}=2; w^{pen}=10; w^{hex}=10		w^{acc}=10; w^{hex}=5	
Focus	None		None		Scaffold		Side chain	
No.	Name	*s*	Name	normS	Name	normS	Name	normS
1	**M_Inh_L26**	100.0	**M_Inh_L26**	100.0	**M_Inh_L26**	100.0	**M_Inh_L26**	100.0
2	**M_Inh_L25**	90.9	**M_Inh_L25**	100.0	**M_Inh_L27**	100.0	**M_Inh_L25**	100.0
3	**M_Inh_L27**	86.2	**M_Inh_L27**	100.0	**M_Inh_L28**	100.0	**M_Inh_L27**	100.0
4	**M_Inh_L28**	80.7	**M_Inh_L28**	100.0	**M_Inh_L25**	100.0	**M_Inh_L28**	100.0
5	**M_Inh_L24**	73.4	**M_Inh_L10**	80.4	**M_Inh_L24**	78.4	**M_Inh_L10**	87.3
6	**M_Inh_L42**	70.3	**M_Inh_L42**	68.2	**SOL_1849**	74.9	**M_Inh_L42**	71.3
7	**M_Inh_L43**	70.3	**M_Inh_T9**	68.2	**SOL_1353**	70.1	**M_Inh_T9**	71.3
8	**M_Inh_L48**	69.8	**M_Inh_L43**	64.4	**SOL_1472**	68.9	**M_Inh_L43**	67.4
9	SOL_1353	69.8	**M_Inh_L44**	62.8	**SOL_986**	67.8	**M_Inh_L48**	64.7
10	**M_Inh_T3**	67.2	**M_Inh_L48**	62.8	SOL_1363	66.2	**M_Inh_L44**	64.7
11	**M_Inh_T9**	67.2	SOL_1849	61.3	SOL_1969	65.9	**M_Inh_L45**	60.7
12	**M_Inh_L13**	66.2	SOL_1363	60.4	M_Inh_L53	64.6	**M_Inh_L47**	60.7
13	**M_Inh_L9**	66.2	**M_Inh_L24**	58.4	**SOL_1219**	62.4	**M_Inh_T8**	59.5
14	SOL_849	64.9	**M_Inh_T5**	58.2	M_Inh_L9	62.2	SOL_1849	59.0
15	**M_Inh_L10**	64.2	**M_Inh_L47**	55.3	M_Inh_T13	62.1	M_Inh_T5	58.2
16	**M_Inh_L49**	62.5	**M_Inh_L45**	55.3	**SOL_1184**	62.1	SOL_1363	56.6
17	**M_Inh_L44**	62.1	**M_Inh_L53**	55.0	SOL_1522	61.7	M_Inh_L46	52.7
18	**M_Inh_T2**	61.8	**M_Inh_L49**	54.9	SOL_2084	61.4	M_Inh_L49	52.7
19	**M_Inh_L11**	61.5	**M_Inh_L46**	54.9	SOL_1850	61.0	SOL_1844	52.7
20	**M_Inh_L12**	61.5	**M_Inh_L9**	54.6	M_Inh_L11	60.7	SOL_1349	52.6
21	**M_Inh_L46**	60.6	**M_Inh_T8**	54.5	M_Inh_L13	60.3	M_Inh_L11	52.4
22	**M_Inh_L8**	60.6	**M_Inh_L11**	54.3	**SOL_711**	59.7	M_Inh_L9	52.2
23	SOL_1566	58.6	**M_Inh_T13**	52.8	M_Inh_T12	59.2	SOL_1729	51.2
24	SOL_1883	58.2	SOL_1850	52.4	M_Inh_L42	58.5	SOL_1850	50.5
25	**M_Inh_L41**	56.7	**M_Inh_L13**	52.3	SOL_753	58.5	**SOL_1825**	49.4
26	**M_Inh_T10**	56.5	SOL_1844	52.3	M_Inh_T9	58.5	M_Inh_L13	49.3
27	SOL_1522	55.9	SOL_1349	50.9	M_Inh_L10	58.2	**SOL_1754**	48.1
28	**M_Inh_L30**	55.7	SOL_1729	50.4	M_Inh_T10	58.2	**SOL_1135**	47.5
29	SOL_522	55.7	SOL_1284	49.2	M_Inh_T5	58.1	**SOL_862**	47.5
30	**M_Inh_T11**	55.2	SOL_753	49.0	SOL_1935	58.0	M_Inh_L24	47.4
31	SOL_1082	55.0	SOL_1935	48.3	**SOL_1448**	57.6	SOL_753	46.6
32	**M_Inh_T5**	54.8	**M_Inh_T10**	48.1	SOL_1902	57.5	M_Inh_L53	46.0
33	SOL_2082	54.6	SOL_1472	47.3	M_Inh_L8	56.5	SOL_1456	45.3
34	SOL_1097	53.6	SOL_1969	47.3	M_Inh_L12	56.5	M_Inh_T13	44.9
35	**M_Inh_L40**	52.9	SOL_1825	46.7	M_Inh_T3	56.5	SOL_1958	44.3
36	**M_Inh_L50**	52.9	SOL_1353	46.7	M_Inh_T2	56.5	SOL_1935	44.0
37	SOL_814	52.5	SOL_1456	46.3	M_Inh_L46	56.3	SOL_1284	43.7
38	**M_Inh_L38**	52.4	**M_Inh_L19**	46.3	M_Inh_L49	56.3	SOL_1764	43.3
39	SOL_1334	52.2	**M_Inh_L52**	46.3	M_Inh_L43	56.2	**SOL_1519**	43.2
40	SOL_1829	52.2	SOL_1829	46.2	SOL_1284	56.0	SOL_1656	43.2
41	SOL_1877	51.6	SOL_1322	46.0	M_Inh_L44	55.3	M_Inh_T10	42.9
42	SOL_1740	51.3	SOL_862	45.4	M_Inh_L48	55.3	SOL_2082	42.8
43	SOL_2066	51.3	SOL_1135	45.4	M_Inh_L54	55.0	SOL_1485	42.1
44	SOL_1519	50.9	**M_Inh_L41**	45.3	M_Inh_L41	54.8	SOL_1322	41.6
45	SOL_1006	50.9	SOL_1902	45.2	SOL_1391	54.2	SOL_1403	41.4
46	SOL_1590	50.9	**M_Inh_L8**	45.1	M_Inh_L19	53.8	M_Inh_L41	41.3
47	SOL_1018	50.8	**M_Inh_L12**	45.1	M_Inh_L52	53.8	SOL_1433	41.3
48	SOL_1181	50.0	**M_Inh_T3**	45.1	M_Inh_L59	53.7	SOL_1472	41.0
49	SOL_1427	50.0	**M_Inh_T2**	45.1	SOL_2067	53.2	M_Inh_T3	40.9
50	SOL_1764	50.0	**M_Inh_T12**	44.9	SOL_2020	53.0	M_Inh_T2	40.9

contain the same scaffold as the target (i.e. the members of G1_L26) are found at the very top of the ranking. They are closely followed by 8 (of 10) other compounds of test data set A that comprise identical ring systems (with differently positioned hydrogen bond acceptors). Finally, emphasis was placed on the side chain of the ligand L26. Test data set A contains thirteen MAO-A inhibitors with the same/similar side chains, which are all grouped at the top of the respective ranking. In addition, five non-MAO-A inhibitors that contain related side chains were detected. Although it was not the primary goal of the last two runs to find as many MAO-A inhibitors as possible, there were 30 and 26 hits, respectively, within the 50 high-scoring compounds.

The test described above was rerun, this time with the MAO-A inhibitor L45 as the target structure. The results and the weighting parameters determined are shown in Table 3 and in the supplementary material.

Table 3 The results of four different similarity evaluations of target L45 within test data set 'A'. The compounds that fit the problem specifications are bold face

Searching method	Tanimoto MACCS Keys		CFM w^{f}=1; v^{f}=0		CFM w^{hex}=10		CFM w^{acc}=10; w^{hex}=5	
Focus	None		None		Scaffold		Side chain	
No.	Name	*s*	Name	normS	Name	normS	Name	normS
1	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0
2	**M_Inh_T8**	95.4	**M_Inh_L47**	81.8	**M_Inh_L47**	93.5	**M_Inh_L47**	97.4
3	**M_Inh_L47**	91.3	**M_Inh_T8**	80.9	**M_Inh_T8**	93.4	**M_Inh_T8**	95.9
4	**M_Inh_L34**	86.4	SOL_1849	65.9	**M_Inh_L34**	77.8	**M_Inh_L28**	65.5
5	**M_Inh_L35**	86.4	SOL_2082	60.7	**M_Inh_L14**	72.9	**M_Inh_L27**	65.5
6	**M_Inh_L14**	82.2	**M_Inh_L28**	59.7	**M_Inh_L35**	72.9	SOL_1349	64.8
7	**M_Inh_L37**	80.0	**M_Inh_L27**	59.7	**M_Inh_L37**	72.9	**M_Inh_L42**	63.5
8	**M_Inh_L33**	71.2	SOL_1349	58.1	**M_Inh_L31**	69.2	**M_Inh_T9**	63.5
9	**M_Inh_L31**	69.8	SOL_1850	56.7	**M_Inh_L33**	69.2	SOL_1849	63.5
10	**M_Inh_L23**	62.1	SOL_1709	53.2	***SOL_1201***	69.1	SOL_1844	60.6
11	SOL_836	53.1	**M_Inh_L42**	53.1	*SOL_2082*	68.1	***SOL_1825***	59.5
12	SOL_1439	48.1	**M_Inh_T9**	53.1	***SOL_1320***	68.0	SOL_2082	56.5
13	SOL_2085	47.4	SOL_1882	50.9	SOL_1697	67.9	SOL_1273	56.3
14	SOL_934	45.9	SOL_1273	50.9	***SOL_1243***	66.9	SOL_1882	56.3
15	SOL_1729	42.3	SOL_1433	49.0	SOL_1145	66.2	SOL_1363	55.4
16	SOL_1273	41.5	SOL_1764	48.5	SOL_1967	64.0	**M_Inh_L43**	55.3
17	SOL_1462	40.9	SOL_1472	48.5	***SOL_1903***	63.6	**M_Inh_L10**	54.1
18	**M_Inh_L25**	40.0	**M_Inh_L11**	48.0	SOL_1709	63.4	M_Inh_T5	53.9
19	**M_Inh_L26**	38.8	SOL_1729	48.0	SOL_1609	62.7	**M_Inh_L48**	53.9
20	**M_Inh_L27**	38.4	**M_Inh_T10**	48.0	SOL_1472	61.2	**M_Inh_L44**	53.9
21	SOL_1419	38.2	**M_Inh_L23**	47.3	***SOL_925***	60.4	**M_Inh_L25**	53.3
22	**M_Inh_L28**	38.2	**M_Inh_T13**	47.3	SOL_1391	60.2	**M_Inh_L26**	53.3
23	SOL_1201	38.0	**M_Inh_L49**	47.2	M_Inh_L61	60.0	SOL_1729	52.6
24	SOL_1156	37.7	**M_Inh_L46**	47.2	***SOL_607***	58.1	SOL_1850	52.1
25	SOL_2082	37.7	**M_Inh_L44**	46.2	SOL_800	58.1	SOL_1456	52.1
26	SOL_1135	37.5	**M_Inh_L48**	46.2	SOL_1489	57.9	M_Inh_L46	51.1
27	SOL_1387	37.5	SOL_753	45.7	*SOL_1082*	57.7	M_Inh_L49	51.1
28	SOL_1991	37.0	**M_Inh_L26**	45.2	*SOL_1796*	57.1	SOL_1433	49.6
29	SOL_1814	36.9	**M_Inh_L25**	45.2	SOL_1548	54.6	SOL_1709	49.3
30	SOL_1825	36.8	SOL_1485	44.4	SOL_1960	54.1	SOL_1322	48.9
31	SOL_862	36.8	**M_Inh_L10**	44.2	*SOL_1259*	53.9	M_Inh_L11	48.0
32	SOL_1815	36.4	SOL_1363	44.0	SOL_1522	53.5	M_Inh_T13	47.1
33	SOL_1882	36.0	**M_Inh_T5**	43.5	SOL_1265	53.3	SOL_1656	45.6
34	SOL_1958	35.4	SOL_1697	43.1	*SOL_1301*	53.2	SOL_753	44.5
35	SOL_1693	35.0	SOL_1825	42.7	SOL_2026	53.1	SOL_1485	44.5
36	**M_Inh_L42**	34.2	**M_Inh_L8**	42.5	SOL_1184	52.7	***SOL_862***	44.4
37	**M_Inh_L3**	33.9	SOL_1844	42.4	SOL_694	52.4	***SOL_1135***	44.4
38	**M_Inh_L46**	33.8	SOL_1944	42.4	*SOL_1866*	52.2	SOL_1958	44.4
39	**M_Inh_L24**	33.3	SOL_1958	42.4	SOL_1580	49.7	SOL_1764	44.3
40	**M_Inh_T9**	32.9	SOL_1829	42.2	SOL_852	47.9	SOL_1472	44.2
41	**M_Inh_L43**	32.5	SOL_1935	42.2	M_Inh_L27	47.8	M_Inh_L13	44.0
42	SOL_1170	32.4	**M_Inh_L19**	42.2	M_Inh_L28	47.8	*SOL_953*	43.8
43	SOL_848	31.3	**M_Inh_L34**	41.7	SOL_1849	47.6	SOL_1793	43.7
44	SOL_991	31.3	SOL_1145	41.3	SOL_874	46.8	M_Inh_L9	43.6
45	**M_Inh_L49**	31.1	**M_Inh_L9**	41.2	SOL_1126	46.7	M_Inh_L53	43.0
46	**M_Inh_T13**	31.0	**M_Inh_L13**	40.6	SOL_878	46.7	M_Inh_T10	42.5
47	SOL_953	31.0	**M_Inh_L43**	40.5	SOL_1398	46.5	SOL_1235	41.9
48	SOL_1559	30.9	SOL_1322	40.5	M_Inh_T13	46.3	M_Inh_L24	41.7
49	SOL_715	30.9	SOL_508	40.5	SOL_600	45.1	M_Inh_L19	41.6
50	SOL_1414	30.8	SOL_1489	40.2	M_Inh_L53	44.8	M_Inh_L52	41.3

Using the Tanimoto coefficient with MACCS Keys, 22 MAO-A inhibitors are found within the 50 database molecules that are most similar to the target (Table 3, col. 1). Thereby, the two groups G2_L45 (ranks 1 to 9) and G1_L26 (ranks 18 to 22) are separated according to their different scaffolds. In contrast to this, the CFM-based method with the basic weightings is predominantly influenced by hydrogen bond acceptors. This is because seven of the eleven comprised features of L45 are of the type acc. The ranking of the 50 high-scoring compounds contains 25 MAO-A inhibitors (Table 3, col. 2). Turning the focus to the scaffold of target L45, the nine members of group G2_L45 are ranked highest. The only six non-MAO-A inhibitors of test data set A that contain the same cyclic scaffold as the target structure are found between ranks 10 and 24 (Table 3, col. 3). Regarding the side chain

Table 4 The alterations of the searching results of target L45 with changing feature weights. The compounds that fit the problem specifications are bold face

No.	CFM w^{hex}=1		CFM w^{hex}=4		CFM w^{hex}=7		CFM w^{hex}=10	
	Name	(normS)	Name	(normS)	Name	(normS)	Name	(normS)
1	M_Inh_L45	100.0	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0
2	**M_Inh_L47**	81.8	**M_Inh_L47**	89.0	**M_Inh_L47**	91.8	**M_Inh_L47**	93.5
3	**M_Inh_T8**	80.9	**M_Inh_T8**	88.7	**M_Inh_T8**	91.7	**M_Inh_T8**	93.4
4	SOL_1849	65.9	SOL_2082	64.9	**M_Inh_L34**	72.4	**M_Inh_L34**	77.8
5	SOL_2082	60.7	**M_Inh_L34**	63.1	SOL_2082	66.9	**M_Inh_L35**	72.9
6	M_Inh_L28	59.7	SOL_1709	60.2	**M_Inh_L37**	66.5	**M_Inh_L14**	72.9
7	M_Inh_L27	59.7	SOL_1697	57.7	**M_Inh_L35**	66.5	**M_Inh_L37**	72.9
8	SOL_1349	58.1	SOL_1472	56.8	**M_Inh_L14**	66.5	**M_Inh_L31**	69.2
9	SOL_1850	56.7	SOL_1849	55.9	SOL_1697	64.1	**M_Inh_L33**	69.2
10	SOL_1709	53.2	**M_Inh_L35**	55.8	SOL_1709	62.3	**SOL_1201**	69.1
11	M_Inh_L42	53.1	**M_Inh_L37**	55.8	SOL_1145	62.1	SOL_2082	68.1
12	M_Inh_T9	53.1	**M_Inh_L14**	55.8	**M_Inh_L31**	62.1	**SOL_1320**	68.0
13	SOL_1882	50.9	SOL_1145	55.5	**M_Inh_L33**	62.1	SOL_1697	67.9
14	SOL_1273	50.9	M_Inh_L28	54.8	***SOL_1201***	62.0	**SOL_1243**	66.9
15	SOL_1433	49.0	M_Inh_L27	54.8	**SOL_1320**	60.8	SOL_1145	66.2
16	SOL_1764	48.5	SOL_1967	52.4	SOL_1472	59.7	SOL_1967	64.0
17	SOL_1472	48.5	**M_Inh_L33**	50.6	SOL_1967	59.6	**SOL_1903**	63.6
18	M_Inh_L11	48.0	**M_Inh_L31**	50.6	**SOL_1243**	59.5	SOL_1709	63.4
19	SOL_1729	48.0	SOL_1489	50.3	M_Inh_L61	56.3	SOL_1609	62.7
20	M_Inh_T10	48.0	***SOL_1201***	50.3	**SOL_1903**	55.4	SOL_1472	61.2
21	M_Inh_L23	47.3	M_Inh_L61	49.6	SOL_1489	55.1	**SOL_925**	60.4
22	M_Inh_T13	47.3	M_Inh_T13	49.4	SOL_1391	54.7	SOL_1391	60.2
23	M_Inh_L49	47.2	**SOL_1320**	49.3	SOL_1609	54.3	M_Inh_L61	60.0
24	M_Inh_L46	47.2	SOL_1850	49.2	SOL_1082	51.9	**SOL_607**	58.1
25	M_Inh_L44	46.2	**SOL_1243**	47.7	**SOL_925**	51.7	SOL_800	58.1

of L45, the 13 expected MAO-A inhibitors plus one of the similar non-MAO-A inhibitors are found between ranks 1 and 22 (Table 3, col. 4). In the last two runs, the numbers of high scoring MAO-A inhibitors were 14 and 25, respectively.

In those runs performed with the basic settings of weighting parameters, the resulting rankings reflect the similarity between the target structure and the database molecules, evenly regarding the whole compounds. To turn one's attention to a specific part of the target structure requires readjustment of the weighting factors according to the appropriate problem specification. As an example, Table 4 displays the alterations that arise in the searching results of target L45 when emphasis is successively placed on its scaffold. For this purpose, four different values of the weighting factor w^{hex} (1, 4, 7 and 10) are taken into consideration.

At each step, the contribution of the hydrogen bond acceptors decreases, while the influence of the ring system increases. This effect not only accounts for the MAO-A inhibitors, but also for other compounds that show a similar scaffold. Accordingly, the overall number of high scoring MAO-A inhibitors must decrease, because only the members of the group G2_L45 have the scaffold specified. Accordingly, raising the value of w^{hex} beyond 10 results in an even larger distance between the G2_L45 group and the other MAO-A inhibitors.

Large database compounds

In this section, the effects of the negative feature weights, as well as of standardizing the similarity values of large database molecules to the size of the target, are shown for the MAO-A inhibitor L45 using test data set L as a reference. The results of three different runs are compared. For the first, the weighting parameters are adjusted with regard to the scaffold of the inhibitor L45. According to the composition of the reference data set, the weights (w^{acc}=2; w^{hex}=10) reflect the structure of the target more distinctly than was necessary in the previous section. In the second run, compounds with more or less than three six-membered rings are penalized by adding the negative weighting factor v^{hex}=-80. Finally, the similarity values of the larger compounds were standardized to the size of the target. The complete results of each run are given in the supplementary material. Table 5 shows the rankings of the 20 high-scoring compounds.

In all these examples, the nine members of group G2_L45 are found within the ranges shown. This is also (rather exclusively) true for the only four non-MAO-A inhibitors that comprise the queried cyclic backbone. Thus, the three results differ mainly in the sequence of the 13 compounds (whereby the best agglomeration is achieved using the standardization method). Therefore, it remains up to the user which method he would estimate to be most suitable for the particular problem specification.

Table 5 The effects of negative feature weights and of standardizing the similarity values of large database molecules. The compounds that fit the problem specifications are bold face

Searching method	CFM w^{acc}=2; w^{hex}=20		CFM w^{acc}=2; w^{hex}=20; v^{hex}=−80		CFM w^{acc}=2; w^{hex}=20; v^{hex}=−80	
Standardization	none		none		$s \cdot \|F\|^{L45} \div \|F\|^{D_a}$	
No.	Name	normS	Name	normS	Name	norm $\bar{S}$
1	**M_InhL45**	100.0	**M_Inh_L45**	100.0	**M_Inh_L45**	100.0
2	**M_Inh_L47**	96.1	**M_Inh_L47**	96.1	**M_Inh_L47**	96.1
3	**M_Inh_T8**	96.0	**M_Inh_T8**	96.0	**M_Inh_T8**	96.0
4	**SOL_4509**	94.9	**SOL_4130**	92.8	**M_Inh_L34**	86.5
5	**SOL_4130**	92.8	**SOL_4509**	90.3	**M_Inh_L35**	83.4
6	SOL_6552	91.2	**SOL_4881**	88.6	**M_Inh_L37**	83.4
7	**SOL_3251**	91.1	**M_Inh_L34**	86.5	**M_Inh_L14**	83.4
8	**SOL_4881**	88.6	**SOL_3251**	86.4	**M_Inh_L33**	81.0
9	**M_Inh_L34**	86.5	SOL_6734	84.4	**M_Inh_L31**	81.0
10	SOL_2820	85.0	**M_Inh_L37**	83.4	SOL_2820	80.3
11	SOL_6734	84.4	**M_Inh_L14**	83.4	**SOL_4130**	78.5
12	SOL_7061	83.6	**M_Inh_L35**	83.4	**SOL_4509**	76.4
13	**M_Inh_L14**	83.4	SOL_6552	81.9	SOL_3135	73.7
14	**M_Inh_L35**	83.4	**M_Inh_L33**	81.0	**SOL_3251**	73.1
15	**M_Inh_L37**	83.4	**M_Inh_L31**	81.0	SOL_2691	72.4
16	SOL_3645	82.0	SOL_2820	80.3	SOL_3645	70.8
17	SOL_4769	81.4	SOL_2639	80.1	SOL_2376	70.6
18	**M_Inh_L33**	81.0	SOL_7061	79.0	SOL_3338	70.5
19	**M_Inh_L31**	81.0	SOL_3645	77.3	SOL_4769	70.3
20	SOL_2639	80.1	SOL_4769	76.7	SOL_3632	69.9

[a] $|F|^{L45}$ and $|F|^{D}$ are the numbers of features occurring in the target and in the database molecule, respectively

Table 6 Computation times and numbers of evaluated molecules resulting from different restriction patterns

Restriction pattern[a]	Time (ms)	Number of eval. molecules
$\|F\|_{f_i}^{L35} \pm \infty$	7060	8460
$\|F\|_{f_i}^{L35} \pm 4$	4938	6958
$\|F\|_{f_i}^{L35} \pm 3$	3836	5870
$\|F\|_{f_i}^{L35} \pm 2$	2343	3860
$\|F\|_{f_i}^{L35} \pm 1$	891	1449

[a] $|F|_{f_i}^{L35}$ (with i=0 ... n) is the number of features $|F|$ of type f_i occurring in the target L35; n is the number of different feature types of the predefined feature set. The restriction pattern $|F|_{f_i}^{L35} \pm x$ precludes all database molecules that do not match the condition $|F|_{f_i}^{L35} - x \leq |F|_{f_i}^{D} \leq |F|_{f_i}^{L35} + x$ for all feature types f_i. Here, D is the CFM of the respective database molecule

Speed

Besides correctness and adaptivity, COFEA was analyzed regarding computing time. Therefore, we used a data set containing 8,460 compounds with molecular weights from 33 to 800, as well as the MAO-A inhibitor L35 (Fig. 5, center) as a target structure. Since the duration of similarity searching significantly depends on the number of compounds evaluated, the effect of precluding unsuitable molecules prior to the actual search was quantified. Therefore, the search was performed with five different restriction patterns. The results are shown in Table 6. All of the patterns are symmetrical, i.e. the allowed positive and negative deviations are the same for all feature types. However, any kind of pattern may be used according to the particular problem.

An extrapolation of the searching time required for the whole data set (no preselection; Table 6, first pattern) yields an estimated value of less than 85 s per 100,000 molecules. The more stringent the restriction patterns, the lower are the average calculation times.

Discussion

The CFM was introduced as a novel feature-based descriptor that enables problem-specific similarity evaluation. For testing our method, we used the Tanimoto coefficient together with MACCS Keys as a benchmark. The analysis described showed that the two approaches yield similar results if the basic settings of weighting factors are applied to the CFM. In addition, we obtained striking results concerning the adaptivity of similarity evaluation. The specific weighting of the features even allows us to focus on small particular structures that are independent of the molecular scaffold. This characteristic makes the CFM suitable for techniques such as scaffold or lead hopping.

Concerning computing time, the CFM-based similarity search proved to be suitably fast for interactive use. In combination with adequate restriction patterns, the searching speed as well as the specificity of the results may be increased.

With regard to the process of similarity evaluation, the CFM is related to the atom-pair descriptor. For both descriptors, similarity evaluation is based on comparison of substructures representing atom pairs and feature pairs, respectively. However, in contrast to the CFM, the atom-pair descriptor only correlates atom pairs that show exactly the same interjacent distances. This is valid because within that descriptor all atoms, and thus all correlations between atoms, of a molecule are included. In contrast to this, the CFM is not restricted to equal distances of correlated feature pairs. This proves to be advantageous in at least two aspects. On the one hand, negligible features may be omitted. Therefore, the comparison of two molecules takes significantly less calculation steps than is the case with a descriptor that uses all atoms for a proper description of the molecules. On the other hand, Euclidean distances instead of topological distances may be used, enabling similarity evaluation on a three-dimensional level. The latter will be the subject of further investigation.

Similarity evaluation based on the CFM model yielded significant results, although the feature set used in this work neither discriminates between aromatic and aliphatic ring systems, nor is there an independent feature for structures that may be either hydrogen-bond donors or acceptors. Further investigation will be done to evaluate the relevance of using varying feature sets. Another goal is to automate the adjustment of the feature weights.

Supplementary material

The SDfiles of the three test data sets as well as the complete evaluation results are available in the supplementary material.

Acknowledgments This work was performed within the SOL-project (**S**earch and **O**ptimization of **L**ead structures) which is supported by the german federal ministry of education and research, bmb+f. We thank the coordinator of the SOL-project, Prof. Dr. Johann Gasteiger, and Michael Bieler, who composed the reference databases.

References

1. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, p 311
2. Hopfinger AJ, Burke BJ (1990) Molecular shape analysis: a formalism to quantitatively establish spatial molecular similarity. In: Johnson MA, Maggiora GM (eds) Concepts and applications of molecular similarity. Wiley, New York, pp 173–209
3. Scsibrany H, Varmuza K (1992) Topological similarity of molecules based on maximum common substructures. In: Ziessow D (ed) Software Development in Chemistry, Proceedings of the 7th CIC-Workshop "Computers in Chemistry". Berlin/Gosen, Germany
4. Burden FR (1989) J Chem Inf Comput Sci 29:225–227
5. Pearlman RS, Smith KM (1998) Novel software tools for chemical diversity. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in drug design, vol 2. Kluwer/ESCOM, Dordrecht, The Netherlands, pp 339–353
6. Moreau G, Broto P (1980) Nouv J Chim 4:359–360
7. Ihlenfeld WD, Gasteiger J (1994) J Comput Chem 15:793–813
8. Hurst T, Heritage TW (1997) HQSAR. A highly predictive QSAR technique based on molecular holograms. 213th ACS National Meeting, San Francisco, Calif.
9. Carhart RE, Smith DH, Venkataraghavan R (1985) J Chem Inf Comput Sci 25:64–73
10. Abolmaali SFB, Zell A (2001) Proceedings of the 15th Molecular Modeling Workshop, Darmstadt, Germany
11. http://www.pc.chemie.tu-darmstadt.de/mmws/2001/vortraege/BadreddinAbolmaali.pdf
12. Trinajstic N (1992) Chemical graph theory. CRC, Boca Raton, Fla., p 322
13. Rarey M, Dixon JS (1998) J Comp-Aided Mol Design 12:471–490
14. PETRA, Version 2.6 (1998) Computer Chemie Centrum, Erlangen, Germany http://www2.ccc.uni-erlangen.de/software/petra/intro.phtml
15. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Comput Sci 38:983–996
16. Cramer RD, Redl G, Berkoff CE (1974) J Med Chem 17:533–535
17. Fisanick W, Cross KP, Rusinko A (1992) J Chem Inf Comput Sci 32:664–674
18. Bath PA, Morris CA, Willett P (1993) J Chemomet 7:543–550
19. Turner DB, Willett P, Ferguson AM (1995) SAR QSAR Environ Res 3:101–130
20. MOE, Release 2001.01 (2001) Chemical Computing Group Inc, Montreal, Canada, http://www.chemcomp.com/fdept/prodinfo.htm
21. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) J Chem Inf Comput Sci 41:702–712 http://cactus.nci.nih.gov/SUBSET/
22. ISIS/Base, Version 2.1.4 (1998) MDL Information Systems, Inc, San Leandro, USA, http://www.mdli.com/products/isis-base.html
23. CT file formats, http://www.mdli.com/downloads/ctfile/ctfile_subs.html
24. Dalby A, Nourse JG, Hounshell WG, Gushurst AKI, Grier DL, Leland BA, Laufer J (1992) J Chem Inf Comput Sci 32:244–255
25. http://www.tocris.com/
26. UNITY Chemical Information Software, Version 2.5. (2001) Tripos Inc, 1699 S Hanley Rd, St Louis, MO 63144
27. SYBYL 6.6, Ligand-based design manual (1999) Tripos Inc, St Louis, pp 21–36