Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening Of 2D Chemical Structure Databases

John W. Raymond (john.raymond@pfizer.com)

Pfizer Global Research and Development, Ann Arbor Laboratories,
2800 Plymouth Road, Ann Arbor, Michigan 48105, USA

Peter Willett (p.willett@sheffield.ac.uk)

Krebs Institute for Biomolecular Research and Department of Information Studies,
University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Keywords Fingerprint, Graph matching, Maximum common edge subgraph, Maximum overlapping set, RASCAL, Similarity coefficient, Similarity searching, Virtual screening

Abstract This paper reports an evaluation of both graph-based and fingerprint-based measures of structural similarity, when used for virtual screening of sets of 2D molecules drawn from the MDDR and ID Alert databases. The graph-based measures employ a new maximum common edge subgraph isomorphism algorithm, called RASCAL, with several similarity coefficients described previously for quantifying the similarity between pairs of graphs. The effectiveness of these graph-based searches is compared with that resulting from similarity searches using BCI, Daylight and Unity 2D fingerprints. Our results suggest that graph-based approaches provide an effective complement to existing fingerprint-based approaches to virtual screening.

INTRODUCTION

Public databases, corporate compound collections, combinatorial synthesis programmes and commercial compound suppliers provide a rich source of data for the identification of novel bioactive molecules. This has led to much interest in the development of methods for *virtual screening*, which enables the rapid identification of that small subset of a database that has a high *a priori* probability of exhibiting the bioactivity of interest [1,2]. Similarity searching based on 2D representations of molecular structure is one of the simplest, and most common, approaches to virtual screening. It involves taking a molecule with the required activity, such as a weak lead from a high-throughput screening programme, and then searching a database to find the molecules that are most similar to it, using some measure of inter-molecular structural similarity [3]; these similar molecules (or nearest neighbours) are expected also to exhibit the activity of interest, and are thus candidates for biological screening.

Many different measures of chemical similarity have been described in the literature (see, e.g., [4-7]) but they all involve three principal components: a representation of molecular structure; a coefficient that quantifies the degree of resemblance between two such representations; and a weighting (or standardization) scheme that is used to highlight (or to normalize) different parts of the chosen representation. In this paper, we focus on just the first two of these components. Specifically, we consider two unweighted 2D structure representation: graphs based on the connection tables that describe conventional structure diagrams; and fingerprints or bit-strings (we will use the former terminology in what follows) generated from such graphs. Each of these two types of representation is processed using several different similarity coefficients.

Most similarity-based methods for virtual screening employ 2D fingerprints, typically with the Tanimoto coefficient being used to compute the degree of similarity between a pair of fingerprints. Graph-based measures, where the degree of similarity is computed as the result of a graph-isomorphism procedure, are far more demanding in computational terms than fingerprint-based approaches and have, accordingly, been far less used for virtual screening applications. However, we have recently described [8,9] an efficient algorithm, called RASCAL, for this purpose. RASCAL calculates the similarity between two chemical graphs using an approach

based on the maximum common edge subgraph (MCES). The MCES between two chemical graphs can be thought of as the largest substructure common to both molecules, and can be either a connected or a disconnected subgraph. We have previously demonstrated the efficiency of RASCAL [8]; in this paper, we report an investigation of its effectiveness when used for virtual screening and compare the results obtained with those from conventional, fingerprint-based approaches.

MEASURES OF CHEMICAL SIMILARITY

All graphs referred to in the following are assumed to be simple, undirected graphs. A graph G consists of a set of vertices $V(G)$ and a set of edges $E(G)$. The vertices in G are connected by an edge if there exists an edge $(v_i, v_j) \in E(G)$ connecting the vertices v_i and v_j in G such that $v_i \in V(G)$ and $v_j \in V(G)$. In 2D chemical graphs, the vertices of the graph correspond to the atoms of the molecule, and the edges represent the chemical bonds. The number of vertices and edges in a graph will be denoted by $|V(G)|$ and $|E(G)|$, respectively, and $|G|$ refers to the sum of $|V(G)|$ and $|E(G)|$ unless otherwise noted. G_{12} denotes the graph corresponding to the MCES between two graphs G_1 and G_2, i.e., the largest substructure common to two molecules in the present context.

Sanfeliu and Fu [10] have categorized graph distance/similarity coefficients into two classes. In *feature-based distances*, a set of features or invariants is established from a structural description of a graph, and these features are then used in a vector representation to which various distance or similarity measures can be applied. In *cost-based distances*, the distance or similarity between two graphs reflects the number of edit operations that are required in order to transform one graph into the other. In this paper, the feature-based distances and the cost-based distances correspond to the use of fingerprints and structure diagrams, respectively.

Similarity coefficients obtained using the feature-based approach are based on formulae that are functions of the relative number of bit positions that are set in each fingerprint representation of a graph (as reviewed in [3]). Here, we use the 2D fingerprints used in the Barnard Chemical Information (BCI), Daylight and Unity systems for chemical information management [11-13],

with the choice of three common approaches providing a firm basis for the evaluation of graph-based approaches to virtual screening that is the principal focus of this paper. For each fingerprint scheme, ten different association coefficients were used to calculate the degree of similarity between two chemical graphs. These association coefficients $d(G_1,G_2)$ are listed in Table 1 and mostly range from 0 to 1 where 0 denotes that the structures are maximally dissimilar (i.e., the similarity is at a minimum) and 1 denotes that the structures are maximally similar. Numerous other similarity coefficients continue to be reported in the literatures of many disciplines (see, e.g., [14-16]).

The cost-based distance between two graphs corresponds to the number of predefined edit operations necessary to transform one graph into another graph, and there have been several reports of graph edit operations including vertex (edge) deletion, edge rotation, and edge slide (see, e.g., [17-19]). Of interest to this work are the graph distance/similarity coefficients based on the MCES between two graphs [20], as the MCES between two chemical graphs provides a natural, convenient means for visualising the similarity between the corresponding molecules. It has been shown that the maximum common subgraph (MCS) and related forms such as the MCES are directly related to the cost-based distance between graphs [18, 19], and we have used several coefficients based on this concept here, as detailed in Table 2 along with their possible value ranges.

It is interesting to note that the Wallis et al. [21] formula (C1) is mathematically equivalent to the Tanimoto [3] formula (F1), and the Johnson[b] [22] measure (C5) is equivalent to the cosine [3] measure (F2) squared. In addition, C4, the normalized variation of another measure attributed to Johnson [23], is equivalent to a feature-based similarity measure (F4) proposed by Dice [3], and the Bunke and Shearer [24] measure (C2) corresponds to a feature-based formula proposed by Braun-Blanquet [16] (F9). This suggests that it may be possible to derive more effective MCES-based similarity measures using established fingerprint-based similarity coefficients; in fact, measures C3, C6, C7, and C8 are published, feature-based coefficients that have been adapted by us for use in a graph-based context.

The cost-based approach offers considerable flexibility with regard to the form of the calculated similarity measure. In the simple coefficients presented in Table 2, atoms and bond pairs are treated equally (i.e., $|G_{12}| = |V(G_{12})| + |E(G_{12})|$). Since $|V(G_{12})|$ can contain isolated vertices, it is reasonable to assume that $|E(G_{12})|$ should be assigned a greater weight, reflecting the greater significance of matching bond pairs. Also important is the concept of what we will refer to as *fragmentation*. If a particular MCES between two graphs is composed primarily of a single large subgraph, it is intuitive to assume that it more closely exemplifies chemical similarity than an MCES that is composed of several small, unconnected subgraphs [25]. In order to discourage excessive fragmentation in the MCES, the value of $|G_{12}|$ can be modified to account for multiple subgraph components. The coefficients in Table 2 can be customized to reflect these concerns by simply substituting the quantity

$$|V(G_{12})| + \beta \cdot \left(1 - \alpha \cdot \left(n(p, G_{12}) - 1\right)\right) \cdot |E(G_{12})|$$

for G_{12},

$$|V(G_1)| + \beta \cdot |E(G_1)|$$

for G_1, and

$$|V(G_2)| + \beta \cdot |E(G_2)|$$

for G_2. The function $n(p,G_{12})$ represents the number of unconnected subgraph components in the MCES (G_{12}) containing p or more edges. If all subgraphs have fewer than p edges, then $n(p,G_{12})$ will be assumed to be the total number of subgraph components. The constant β reflects the additional weight assigned to matched bond pairs with respect to compatible atoms, and the constant α is a penalty score for each unconnected component present in G_{12}. In preliminary studies, we have found values of p=3, $\alpha = 0.05$, and $\beta = 2.0$ seem to be effective in discerning chemical similarity, and are used in all of the experiments reported here.

Figure 1 illustrates the effect of penalizing comparisons in which G_{12} is fragmented. Using the standard definition of the G_{12}, both comparisons have approximately the same value of similarity using measure C5 from Table 2. However, when the modified definition ($C5^{frag}$) is used, a clear separation develops, and it is apparent that its use provides a more chemically intuitive notion of similarity. In fact, detailed studies of the two types of coefficient (using the evaluation methods described in the next section) demonstrate clearly the superiority of coefficients that penalise

fragmentation: it is these coefficients, accordingly, that have been used to obtain all of the results presented later in this paper.

EVALUATING THE EFFECTIVENESS OF VIRTUAL SCREENING

The dataset used in the following analysis contained 11,607 compounds from the ID Alert database that have been classified according to pharmacological activity. One hundred similarity search target structures were obtained as follows: 100 activity classes were selected at random, subject to there being at least 20 compounds with that activity; one of the compounds was then chosen at random from each of the selected activity classes; each such resulting compound was used as the target structure for a similarity search of the ID Alert file. The effectiveness of each search was determined by seeing how many of the top-ranked molecules belonged to the same activity class as the target structure, i.e., would exhibit the same activity in a real screening programme.

Edgar et al. have reviewed several measures for the effectiveness of similarity searching in chemical databases, drawing primarily upon measures described for evaluating the performance of text search engines [26]. They found that the cumulative recall and the "Goodness of Hit List" or Guner-Henry (G-H) score [27] were among the most successful of those tested for measuring the effectiveness of similarity retrieval. The additive G-H score investigated by Edgar et al. is based on the *recall* (*R*) and *precision* (*P*) of a search. Recall is defined as the fraction of the active structures that are retrieved in the search (*a*) over the total number of active structures in the database (*A*), i.e.,

$$R = \frac{a}{A},$$

with the cumulative recall being simply the recall at some fixed cut-off number of top-ranked molecules, e.g., the top-20 nearest neighbours. The precision is defined to be the fraction of the active structures that are retrieved (*a*) over the number of structures retrieved (*n*), i.e.,

$$P = \frac{a}{n}.$$

The additive G-H score is calculated as:

$$\text{G-H} = \frac{\alpha \cdot P + \beta \cdot R}{2}$$

where α and β are weights describing the relative significance of the recall and precision terms. Guner and Henry have extended the utility of the G-H score by addressing some of the weaknesses of the additive form, and proposed a modified G-H score of the form [28]:

$$\text{G-H} = \left(\frac{a(3 \cdot A + n)}{4 \cdot n \cdot A} \right) \cdot \left(1 - \frac{n - a}{N - A} \right),$$

where *N* is the total number of structures in the database. We have used the cumulative recall and the modified G-H score to evaluate the effectiveness of the various similarity measures presented previously.

Typically, similarity searching is performed by ranking the compounds in a dataset according to decreasing similarity and then selecting a specified number or percentage of the top-ranked compounds. This presents no problems when the recall at some fixed cut-off is used as a performance measure to compare different similarity measures. It can, however, have a significant impact on the value of the G-H score, and we have hence additionally used an evaluation approach based on identifying that cut-off that results in an optimal value for the G-H score.

The G-H score process is accomplished by first calculating the pair-wise similarity coefficient between the target structure and each member of the database using one of the similarity formulae in Table 1 or 2. The resulting similarity values are sorted in order of decreasing similarity so that the nearest neighbours are listed first. The G-H score is then calculated for every structure in this ranking, and the structure corresponding to the maximum G-H score in the list is determined. The position in the ranking (i_{GH}) corresponding to the maximum G-H score represents the cut-off position for the optimal search hit list, i.e., the collection of compounds from rank-1 to rank-i_{GH}. Since there are 100 activity classes, the selection process is taken over all 100 possible queries, giving 100 resultant maximal G-H scores for each similarity measure. A minimum acceptable G-H score is then specified, and all queries resulting in maximal G-H scores less than this threshold are removed from consideration: these queries will be referred to as the *discards* (*D*) since they represent ineffective target structures. The average value of recall (*R*) and precision (*P*) for the

retained queries provide a means of evaluating the relative performance of each similarity measure as does the number of discarded queries (*D*). In addition, the calculated mean cut-off similarity value can be used as a benchmark similarity threshold for future searching of datasets when the number of actives is not known.

In order for the proposed analysis to be effective we must first determine an acceptable value for the G-H score cut-off. Guner and Henry proposed an idealized ranking of six archetypal search scenarios:

- Best Case: $n = a = A$. All of the active compounds are retrieved with no false positives.
- Worst Case: $a = 0$, $n = N - A$. All compounds except the actives are retrieved.
- Extreme Precision: $a = 1$, $n = 1$. Only a single, active structure is retrieved.
- Extreme Recall: $a = A$, $n = N$. The entire database is retrieved, i.e., no screening of the database at all.
- Typical Good: $R = 0.8$, $P = 0.4$. Retrieval results in high recall and medium precision.
- Typical Bad: $R = 0.5$, $P = 0.05$. Retrieval results in medium recall and low precision.

Of these six scenarios, Extreme Recall and Worst Case represent unacceptable levels of performance for a search system, and we have hence taken the G-H score corresponding to Typical Bad as a minimal cut-off, i.e., we are only interested in searches that achieve at least R=0.5 and P=0.05. Using our dataset, N=11,607, and the number of actives (A) corresponding to each of the 100 target structures ranges from 21 to 250 with a mean of 58. This results in G-H scores of 0.16, 0.13, and 0.15, respectively, and we have hence decided to take a cut-off G-H score of 0.20 as an acceptable threshold value. The appropriateness of this value is supported by an inspection of the recall and precision values corresponding to the G-H score ranking for each query, with a value of 0.20 corresponding closely to the point at which the quality of the hit list starts to deteriorate rapidly, giving increasingly large numbers of inactive molecules.

EXPERIMENTAL RESULTS

We discuss first the feature-based similarity measures, so as to provide a baseline of performance for the subsequent evaluation of the novel, cost-based measures.

Table 3 lists the results of the maximum G-H score analysis performed for all coefficients listed in Table 1 with each of the three types of fingerprint. *R* and *P* here are the mean recall and precision values and $d(G_1,G_2)$ the mean similarity at the cut-off, in each case averaged over the non-discarded queries. It appears from the data in the table that the BCI fingerprints performed slightly better than either the Unity or Daylight fingerprints for this set of target structures, in that the BCI fingerprints have the fewest discards and the highest recall (though, hardly surprisingly, the lowest precision). It is also evident from Table 3 that similarity coefficients F1 through F5 and F7 are markedly consistent in terms of recall, precision, and the number of discarded queries, the only discernable differences being in the average similarity cut-off. There is a notable difference, however, between these coefficients and coefficients F6 and F8-F10, which performed notably worse with respect to recall and/or the number of discards. F8 is particularly interesting, exhibiting both the most discards of any coefficient and very high values of recall.

Table 4 presents the corresponding results when the searches are evaluated by means of the cumulative recall for hit-lists containing 25, 50, 75, and 100 compounds. From this data, it can be seen that the coefficients exhibit a comparable degree of similarity to that displayed in Table 3. It is interesting, however, that as the size of the hit-list increases the slight advantage that the BCI fingerprints displayed in detecting active structures (Table 3) begins to diminish.

As previously mentioned, one very noticeable trend in the data presented in Tables 3 and 4 is that many of the similarity coefficients seem to perform comparably. This result is not surprising, however, when one investigates the coefficients in more detail. Table 5 presents the results of comparing each of the top-100 nearest neighbours from Table 4 using BCI fingerprints to each other. Hit-list similarity is determined using the asymmetric coefficient $c/\min\{a,b\}$ (cf F8 in Table 1). Here, *a* is the number of actives retrieved using for a particular similarity coefficient (from Table 1) when summed across all 100 searches; *b* is the number of actives retrieved using another similarity coefficient; and *c* is the number of actives common to both sets of searches.

Table 5 illustrates much of the same degeneracy between similarity coefficients presented in Tables 3 and 4. It is clear that methods F1, F4, and F5 are essentially monotonic, as are F3 and F7. In fact, only F8, F9, and F10 exhibit any significant degree of dissimilarity with any of the

other coefficients. The experimental relationships between the coefficients presented in Table 5 can be explained qualitatively by investigating the mathematical properties of some of the coefficients. For this analysis we limit ourselves to coefficients F1 through F5. First we observe that, coefficients F8 ($x/\min\{y,z\}$) and F9 ($x/\max\{y,z\}$) are mutually independent and that coefficients F1-F5 are all simple functions in three variables (x, y, and z). It is then a matter of simple algebra to transform the formulae corresponding to coefficients F1-F5 into functions whose two independent variables are simply F8 and F9. Therefore, coefficients F8 and F9 can serve as the independent variables for coefficients F1-F5, and we can observe the properties of coefficients F1-F5 as the values of F8 and F9 change.

Figure 2 illustrates the behavior of coefficients F1-F5 with respect to varying values of F8 and F9, with each plot corresponding to a fixed value of F9 and a range of values for F8. From Figure 2, it can be seen that the curves corresponding to coefficients F1, F4, and F5 are essentially parallel for all depicted values of F8 and F9. This explains the monotonicity observed in Table 5. Take, for instance, the curve corresponding to F5 in the plot for F9=0.2, which is essentially the same curve as F1 and F4 but shifted vertically on the plot. This would correspond to adding a constant to the similarity values in the ranking corresponding to coefficient F5 to produce the rankings for coefficients F1 or F4.

Note that in the plots in the bottom half of Figure 2, all of the depicted similarity coefficients produce essentially parallel lines. This explains the marked degree of degeneracy noted between many of the coefficients considered in Table 5. Since a relationship of activity between two molecular structures typically involves structures of comparable size and complexity, the differences in values between coefficients F8 and F9 are typically not great enough for a comparison to be located in either of the top two plots. Since most comparisons will, therefore, fall into plots resembling the bottom two plots, it is clear why there is such a high degree of dependency between the various similarity coefficients compared in Table 5. Unless an occasional comparison falls into a plot resembling one of the top two plots, the resulting rankings will be essentially identical between the similarity coefficients, differing only by a constant value corresponding to a vertical shift between the lines depicted in the bottom two plots of Figure 2. It does appear from Table 5, however, that similarity searching may potentially be improved by

employing one of the coefficients in F1-F7 in conjunction with coefficients F8, F9, and F10, as these coefficients seem to exhibit a relatively significant degree of variability with respect to each other.

In addition to considering the differences between similarity coefficients, we have considered the differences between the three types of fingerprint, as detailed in Table 6. Here, we have compared the top-100 nearest neighbour hit-lists obtained using different fingerprints. Inspection of the values in Table 6 suggests at least some level of variability between the three types of fingerprint, especially between BCI and the other two types. This raises the possibility that the effectiveness of searching using fingerprints might be increased by employing multiple fingerprints with the same similarity coefficient (as an alternative to the use of multiple similarity coefficients with the same fingerprint that has been reported recently [29]).

Similar procedures were used to investigate the effectiveness of the cost-based similarity measures that employ the RASCAL algorithm for the identification of the MCES, as shown in Tables 7 and 8. As noted previously, these results use the versions of the coefficients that include a penalty for fragmentation of G_{12}. It is evident from both Tables 6 and 7 that the effectiveness of the RASCAL approach is comparable to the results observed using the fingerprint methodologies.

It will be seen from Tables 7 and 8 that, with the exception of $C3^{frag}$, all of the graph-based coefficients give very similar results to each other, and this is confirmed by the data shown in Table 9. Here, we have used the asymmetric coefficient approach used in Table 5 to assess the degree of similarity between the numbers of active molecules retrieved using pairs of different coefficients. This analysis reveals that all of the coefficients have very high similarities with each other, with the sole exception of the outlier $C3^{frag}$. This is consistent with the pattern observed using the fingerprint similarities in Table 5, which underscores the importance of optimal similarity coefficient selection when implementing data fusion for enhanced similarity search effectiveness [29].

Having determined the absolute performance of the two classes of similarity measure, we now compare them directly. While Tables 3 and 7, and Tables 4 and 8, show that the current

implementation of RASCAL results in retrieval effectiveness comparable to the fingerprint methods, they provide no information as to the nature of the active compounds retrieved by the two approaches. This question is addressed in Table 10, which again uses the asymmetric similarity approach. In order to perform the comparison in an appropriate manner, only those fingerprint-based coefficients in Table 1 that had an equivalent corresponding graph-based coefficient in Table 2 were considered. Table 10 lists the asymmetric similarity in the types of active structures retrieved for each of these pairs of coefficients. It will be seen from this data that there are substantial differences between the two sets of hit-lists, with about 30% of the active molecules being different. This suggests that there is potential in using both methods in a complementary fashion when carrying out virtual screening on databases of 2D chemical structures.

To confirm the complementarity of the two similarity approaches, another set of ten virtual screening searches was applied to a different drugs file, specifically a set of 99,603 molecules from the MACCS Drug Data Report (MDDR) database. Representative molecules from ten activity classes were chosen, and searches carried out for evaluation by means of cumulative recall. To simplify the evaluation, the BCI fingerprint was selected to represent the three types of fingerprint, owing to its slightly better performance in the previous, ID Alert-based analyses. In addition, only a single similarity coefficient was used, *viz* the Kulczynski coefficient (F3 and C7), owing to its excellent performance in both the fingerprint-based and graph-based searches. The results of these ten searches are detailed in Table 11. The two types of search are seen to be relatively consistent in the numbers of active structures retrieved, with about a 20% difference in the constitution of the retrieved active structures**.** A Sign Test was used to test the hypothesis that there was a statistically significant difference in the numbers of actives retrieved by the two types of coefficient. No such difference (at the 0.05 level of statistical significance) was observed at any of the five cut-offs (25, 50, 75, 100 and 150).

One obvious way of exploiting the observed degree of complementarity is to apply data fusion (sometimes called consensus scoring) to the graph-based and fingerprint-based similarity rankings. Table 11 includes the number of actives discovered when the two rankings $F3^{(BCI)}$ and $C7^{frag}$ are fused using Tzitzikas' democratic data fusion algorithm [30]. Table 11 shows that the

fusion generated hit-lists often contain more actives than either of the original candidate rankings. Equally significant is the fact that only two out of fifty (4%) of the fused hit-lists contained fewer actives than either of the original hit-lists. This is an important observation, which suggests that data fusion can condense the information contained in a set of candidate rankings, potentially resulting in better quality hit-lists than any of the hit-lists from the candidate rankings while minimizing the risk that the fused hit-list will be of poorer quality than the original hit-lists. The enhanced performance is highlighted by the data in Table 12, which lists the total numbers of actives retrieved by the fingerprint-based, graph-based and fused coefficients. The fused totals are greater at all cut-offs, with the average percentage enhancement of 5.4% over the better of the two individual search types.

There is another way in which the two approaches could be combined. While RASCAL is very fast in operation [8], it is still slower than a fingerprint-based similarity search and one might thus consider using it to process the output from a conventional Tanimoto-based search. Graph-based coefficients are *local coefficients*, using the terminology of [31], in that they provide not only a quantitative value for the degree of resemblance between two molecules, but also an alignment of them. Thus if RASCAL is used to align the target structure and a database structure by superimposing the MCES, then the user is provided with a clear visual impression of the structural relationship between the two molecules. In such a two-stage procedure, then, the fingerprint coefficient would be used to generate the initial output ranking, and the RASCAL algorithm would be used to visualise the similarities between the target structure and the nearest neighbours from that ranking.

CONCLUSIONS

Fingerprint-based methods are very widely used for virtual screening of chemical databases where there is a need to identify bioactive compounds using the concept of 2D structural similarity. Although both efficient and effective in operation [3], fingerprint-based methods exhibit several undesirable characteristics [32, 33], and there is thus continuing interest in alternative approaches. We have recently described an algorithm, called RASCAL, that enables efficient graph-based similarity searching of large chemical databases, and here we have shown that the such searches

are effective in a simulated virtual screening environment,. Experiments with a range of similarity coefficients demonstrate that RASCAL-based searches are comparable in effectiveness to conventional, fingerprint-based similarity searching, when evaluated using cumulative recall and G-H score. It is also shown that the RASCAL and fingerprint measures identify non-identical sets of active molecules, suggesting the use of RASCAL as an effective complement to existing procedures for virtual screening of databases of 2D chemical structures.

Acknowledgments We thank the following: Pfizer (Ann Arbor) for funding; John Blankley, Alain Calvet, Eric Gifford, Christine Humblet, Sherry Marcy and David Wild (Pfizer) and Naomie Salim and John Holliday (University of Sheffield) for helpful advice; Current Drugs Limited and MDL Information Systems Inc. for provision of the ID Alert and MDDR databases; Barnard Chemical Information Ltd., Daylight Chemical Information Systems Inc., the Royal Society, Tripos Inc. and the Wolfson Foundation for software and hardware support. The Krebs Institute for Biomolecular Research is a designated centre of the Biotechnology and Biological Sciences Research Council.

References

1. Klebe, G. (Ed.), Virtual Screening: An Alternative or Complement to High Throughput Screening, Special issue of Perspect. Drug Discov. Design, 20 (2000) 1.
2. Bohm, H. J. and Schneider, G. (Eds.), Virtual Screening for Bioactive Molecules, Wiley-VCH, Weinheim, 2000.
3. Willett, P., Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 38 (1998) 983.
4. Johnson, M. and Maggiora, G. (Eds.), Concepts and Applications of Molecular Similarity, John Wiley, New York, 1990.
5. Dean, P.M. (Ed.), Molecular Similarity in Drug Design, Chapman and Hall, Glasgow, 1994.
6. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
7. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., J. Med. Chem., 39 (1996) 3049.
8. Raymond, J., Gardiner, E.J. and Willett, P., J. Chem. Inf. Comput. Sci., in press.
9. Raymond, J., Gardiner, E.J. and Willett, P., Comput. J., submitted for publication.
10. Sanfeliu, A. and Fu, K. S., IEEE Trans. Syst. Man Cybern., 13 (1983) 353.
11. Barnard Chemical Information Ltd. is at URL www.bci1.demon.co.uk
12. Daylight Chemical Information Systems Inc. is at URL www.daylight.com
13. Tripos Inc. is at URL www.tripos.com
14. Ellis, D., Furner-Hines, J. and Willett, P., Perspect. Inf. Manag., 3 (1993) 128.
15. Tulloss, R. E., In Palm, M.E. and Chapela, I.H. (Eds.), Mycology in Sustainable Development: Expanding Concepts, Parkway Publishers, Boone, North Carolina, 1997, 122.
16. Hayek, L., In Heyer, W. (Eds.), Measuring and Monitoring Miological Diversity: Standard Methods for Amphibians, Smithsonian Institution, Washington, D. C., 1994, 207.
17. Goddard, W. and Swart, H.C., Discrete Math., 161 (1996) 121.
18. Chartrand, G., Saba, F. and Zou, H., Cas. Pest. Mat., 110 (1985) 87.
19. Bunke, H., Pattern Recogn. Lett., 18 (1997) 689.

20. Skvortsova, M. I., Baskin, I. I., Stankevich, I. V., Palyulin, V. A. and Zefirov, N. S., J. Chem. Inf. Comput. Sci., 38 (1998) 785.
21. Wallis, W.D., Shoubridge, P., Kraetz, M. and Ray, D., Pattern Recogn. Lett., 22 (2001) 701.
22. Johnson, M., Naim, M., Nicholson, V. and Tsai, C., In King, R.B. and Rouvray, D.H. (Eds.), Graph Theory and Topology in Chemistry, Elsevier Science Publishers, 1987, 219.
23. Johnson, M., In Alavi, Y., Chartrand, G., Lesniak, L., Lick, D. and Wall, C. (Eds.), Graph Theory and Its Applications to Algorithms and Computer Science, John Wiley, New York, 1985, 457.
24. Bunke, H. and Shearer, K., Pattern Recogn. Lett., 19 (1998) 255.
25. Sheridan, R.P. and Miller, M.D., J. Chem. Inf. Comput. Sci., 38 (1998) 915.
26. Edgar, S.J., Holliday, J.D. and Willett, P., J. Mol. Graph. Model., 18 (2000) 343.
27. Guner, O.F. at http://www.netsci.org/Science/Cheminform/feature09.html
28. Guner, O.F., In Guner, O.F. and Henry, D.R., (Eds.), Pharmacophore Perception, Development, and Use in Drug Design, International University Line, La Jolla, CA, USA, 2000, 194.
29. Holliday, J. D., Hu, C. Y. and Willett, P., Combin. Chem. High Throughput Screen, in press.
30. Tzitzikas, Y., IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 2001.
31. Downs, G. M. and Willett, P., Rev. Comput. Chem., 7 (1995) 1.
32. Flower, D. R., J. Chem. Inf. Comput. Sci.. 38 (1998) 379.
33. Godden, J.W., Xue, L. and Bajorath, J., J. Chem. Inf. Comput. Sci., 40 (2000) 163.

Table 1. Feature-Based Similarity Coefficients. In these formulae, x = the number of bits set in both fingerprints, y = the number of bits set in the first fingerprint, z = the number of bits set in the second fingerprint, and w = the total number of bits in the bit string. In F6,

$$X = log\left(1+\frac{min\{y,z\}}{max\{y,z\}}\right)/log(2), \qquad Y = \left(log\left(2+\frac{min\{y,z\}-x}{x+1}\right)/log(2)\right)^{-\frac{1}{2}},$$

$$Z = \frac{log\left(1+\frac{x}{y}\right)\cdot log\left(1+\frac{x}{z}\right)}{log^2(2)}.$$

ID	Reference	$d(G_1,G_2)$	Range
F1	Tanimoto [14]	$\frac{x}{y+z-x}$	0 to 1
F2	Cosine [14]	$\frac{x}{\sqrt{y\cdot z}}$	0 to 1
F3	Kulczynski [14]	$\frac{x\cdot(y+z)}{2\cdot y\cdot z}$	0 to 1
F4	Dice [14]	$\frac{2x}{y+z}$	0 to 1
F5	Sokal/Sneath [14]	$\frac{x}{2y+2z-3x}$	0 to 1
F6	Tullos [15]	$X\cdot Y\cdot Z$ *	0 to 1
F7	McConnaughey [14]	$\frac{x\cdot(y+z)-y\cdot z}{y\cdot z}$	-1 to 1
F8	Asymmetric [14]	$\frac{x}{\min\{y,z\}}$	0 to 1
F9	Braun-Blanquet [16]	$\frac{x}{\max\{y,z\}}$	0 to 1
F10	Russel/Rao [14]	$\frac{x}{w}$	0 to 1

Table 2. Cost-Based Similarity Coefficients. In these formulae, $|G_{12}|$ is the size of the MCSE between two graphs of sizes of $|G_1|$ and of $|G_2|$.

ID	Reference	$d(G_1,G_2)$	Range
C1	Wallis et al. [21]	$\frac{\lvert G_{12}\rvert}{\lvert G_1\rvert+\lvert G_2\rvert-\lvert G_{12}\rvert}$	0 to 1
C2	Bunke and Shearer [24]	$\frac{\lvert G_{12}\rvert}{\max\{\lvert G_1\rvert,\lvert G_2\rvert\}}$	0 to 1
C3	Asymmetric [14]	$\frac{\lvert G_{12}\rvert}{\min\{\lvert G_1\rvert,\lvert G_2\rvert\}}$	0 to 1
C4	Normalized Johnson[a] [22,23]	$\frac{2\cdot\lvert G_{12}\rvert}{\lvert G_1\rvert+\lvert G_2\rvert}$	0 to 1
C5	Johnson[b] [22,23]	$\frac{\lvert G_{12}\rvert^2}{\lvert G_1\rvert\cdot\lvert G_2\rvert}$	0 to 1
C6	Sokal and Sneath [14]	$\frac{\lvert G_{12}\rvert}{2\lvert G_1\rvert+2\lvert G_2\rvert-3\lvert G_{12}\rvert}$	0 to 1
C7	Kulczynski [14]	$\frac{\lvert G_{12}\rvert\cdot(\lvert G_1\rvert+\lvert G_2\rvert)}{2\lvert G_1\rvert\cdot\lvert G_2\rvert}$	0 to 1
C8	McConnaughey [14]	$\frac{\lvert G_1\rvert\cdot\lvert G_{12}\rvert+\lvert G_2\rvert\cdot\lvert G_{12}\rvert-\lvert G_1\rvert\cdot\lvert G_2\rvert}{\lvert G_1\rvert\cdot\lvert G_2\rvert}$	-1 to 1

Table 3. Maximum G-H Score Retrieval Results for Fingerprint-Based Similarity Coefficients. All the mean values are taken over the set of (100–D) non-discarded queries

ID	*R*	*P*	$d(G_1,G_2)$	*D*	ID	*R*	*P*	$d(G_1,G_2)$	*D*
$F1^{(BCI)}$	0.084	0.77	0.70	40	$F6^{(BCI)}$	0.067	0.81	0.69	41
$F1^{(Unity)}$	0.066	0.85	0.73	44	$F6^{(Unity)}$	0.061	0.88	0.72	46
$F1^{(Daylight)}$	0.061	0.83	0.67	45	$F6^{(Daylight)}$	0.065	0.86	0.63	44
$F2^{(BCI)}$	0.082	0.76	0.82	40	$F7^{(BCI)}$	0.083	0.75	0.66	40
$F2^{(Unity)}$	0.064	0.85	0.85	44	$F7^{(Unity)}$	0.062	0.84	0.70	43
$F2^{(Daylight)}$	0.055	0.85	0.80	45	$F7^{(Daylight)}$	0.056	0.84	0.62	45
$F3^{(BCI)}$	0.084	0.76	0.83	41	$F8^{(BCI)}$	0.208	0.45	0.91	73
$F3^{(Unity)}$	0.062	0.84	0.85	43	$F8^{(Unity)}$	0.100	0.63	0.95	70
$F3^{(Daylight)}$	0.056	0.84	0.81	45	$F8^{(Daylight)}$	0.141	0.56	0.94	68
$F4^{(BCI)}$	0.084	0.77	0.82	40	$F9^{(BCI)}$	0.078	0.81	0.80	41
$F4^{(Unity)}$	0.066	0.85	0.84	44	$F9^{(Unity)}$	0.053	0.91	0.83	48
$F4^{(Daylight)}$	0.061	0.83	0.79	45	$F9^{(Daylight)}$	0.065	0.88	0.75	45
$F5^{(BCI)}$	0.083	0.77	0.56	40	$F10^{(BCI)}$	0.069	0.77	0.09	58
$F5^{(Unity)}$	0.066	0.85	0.60	44	$F10^{(Unity)}$	0.059	0.81	0.21	61
$F5^{(Daylight)}$	0.061	0.83	0.53	45	$F10^{(Daylight)}$	0.061	0.88	0.12	58

Table 4. Ranking Results for Fingerprint-Based Similarity Coefficients. The values listed are the mean recall for all 100 queries for hit-lists of sizes 25, 50, 75, and 100 compounds.

ID	R_{25}	R_{50}	R_{75}	R_{100}	ID	R_{25}	R_{50}	R_{75}	R_{100}
F1$^{(BCI)}$	0.085	0.122	0.141	0.154	F6$^{(BCI)}$	0.081	0.116	0.134	0.147
F1$^{(Unity)}$	0.075	0.108	0.126	0.142	F6 $^{(Unity)}$	0.072	0.099	0.121	0.135
F1$^{(Daylight)}$	0.078	0.109	0.127	0.142	F6$^{(Daylight)}$	0.076	0.107	0.126	0.138
F2$^{(BCI)}$	0.086	0.122	0.142	0.155	F7$^{(BCI)}$	0.085	0.124	0.141	0.154
F2 $^{(Unity)}$	0.075	0.108	0.128	0.144	F7 $^{(Unity)}$	0.075	0.107	0.127	0.141
F2$^{(Daylight)}$	0.077	0.107	0.128	0.141	F7$^{(Daylight)}$	0.075	0.103	0.124	0.137
F3$^{(BCI)}$	0.085	0.124	0.141	0.155	F8$^{(BCI)}$	0.037	0.062	0.078	0.094
F3$^{(Unity)}$	0.075	0.107	0.127	0.141	F8$^{(Unity)}$	0.033	0.054	0.067	0.080
F3$^{(Daylight)}$	0.075	0.103	0.124	0.137	F8$^{(Daylight)}$	0.028	0.043	0.058	0.071
F4$^{(BCI)}$	0.085	0.122	0.141	0.154	F9$^{(BCI)}$	0.079	0.110	0.124	0.139
F4$^{(Unity)}$	0.075	0.108	0.126	0.142	F9 $^{(Unity)}$	0.068	0.095	0.113	0.127
F4$^{(Daylight)}$	0.078	0.109	0.127	0.142	F9$^{(Daylight)}$	0.071	0.105	0.122	0.135
F5$^{(BCI)}$	0.085	0.122	0.141	0.154	F10$^{(BCI)}$	0.054	0.077	0.098	0.107
F5 $^{(Unity)}$	0.075	0.108	0.126	0.142	F10$^{(Unity)}$	0.047	0.066	0.079	0.086
F5$^{(Daylight)}$	0.078	0.109	0.127	0.142	F10$^{(Daylight)}$	0.052	0.075	0.092	0.099

Table 5. Comparison of Mean Hit-List Similarity between BCI Similarity Coefficients.

Similarity is calculated for the top-100 hit-lists for all 100 queries using the asymmetric coefficient $c/\min\{a,b\}$ where a is the number of actives retrieved using one similarity measure over all hit-lists, b is the number of actives retrieved using the other similarity measure over all hit-lists, and c is the number of actives common to both similarity measures over all hit-lists). If min{a,b}=0 for a particular query, then the asymmetric similarity is set to one.

$F2^{(BCI)}$	0.98								
$F3^{(BCI)}$	0.96	0.98							
$F4^{(BCI)}$	1.0	0.98	0.96						
$F5^{(BCI)}$	1.0	0.98	0.96	1.0					
$F6^{(BCI)}$	0.94	0.93	0.91	0.94	0.94				
$F7^{(BCI)}$	0.95	0.97	1.0	0.95	0.95	0.90			
$F8^{(BCI)}$	0.84	0.85	0.89	0.83	0.83	0.75	0.89		
$F9^{(BCI)}$	0.92	0.90	0.88	0.92	0.92	0.97	0.87	0.69	
$F10^{(BCI)}$	0.83	0.84	0.86	0.83	0.83	0.78	0.86	0.74	0.73
	$F1^{(BCI)}$	$F2^{(BCI)}$	$F3^{(BCI)}$	$F4^{(BCI)}$	$F5^{(BCI)}$	$F6^{(BCI)}$	$F7^{(BCI)}$	$F8^{(BCI)}$	$F9^{(BCI)}$

Table 6. Comparison of Mean Hit-List Similarity between the Fingerprint-based Similarity Coefficients (using the asymmetric coefficient approach described in the caption to Table 5).

	BCI-Unity	BCI-Daylight	Unity-Daylight
F1	0.76	0.79	0.81
F2	0.76	0.79	0.81
F3	0.77	0.80	0.81
F4	0.76	0.79	0.81
F5	0.76	0.79	0.81
F6	0.74	0.77	0.78
F7	0.76	0.80	0.81
F8	0.61	0.69	0.79
F9	0.73	0.74	0.76
F10	0.78	0.78	0.84

Table 7. Maximum G-H Score Retrieval Results for Graph-Based Coefficients. All the mean values are taken over the set of (100–D) non-discarded queries

ID	***R***	***P***	$d(G_1,G_2)$	***D***
$C1^{frag}$	0.083	0.82	0.75	41
$C2^{frag}$	0.082	0.84	0.83	46
$C3^{frag}$	0.072	0.83	0.91	44
$C4^{frag}$	0.083	0.82	0.85	41
$C5^{frag}$	0.083	0.83	0.73	41
$C6^{frag}$	0.083	0.82	0.61	41
$C7^{frag}$	0.082	0.84	0.86	41
$C8^{frag}$	0.081	0.84	0.71	41

Table 8. Ranking Results for Graph-Based Similarity Coefficients. The values listed are the mean recall for all 100 queries for hit-lists of sizes 25, 50, 75, and 100 compounds.

ID	R_{25}	R_{50}	R_{75}	R_{100}
$C1^{frag}$	0.087	0.119	0.134	0.148
$C2^{frag}$	0.080	0.105	0.120	0.127
$C3^{frag}$	0.061	0.089	0.105	0.119
$C4^{frag}$	0.087	0.119	0.134	0.148
$C5^{frag}$	0.088	0.119	0.138	0.151
$C6^{frag}$	0.087	0.119	0.134	0.148
$C7^{frag}$	0.087	0.123	0.140	0.152
$C8^{frag}$	0.087	0.123	0.141	0.152

Table 9. Comparison of Average Hit-List Similarity between Graph-Based Similarity Coefficients (using the asymmetric coefficient approach described in the caption to Table 5).

$C2^{frag}$	0.94							
$C3^{frag}$	0.80	0.65						
$C4^{frag}$	1.00	0.94	0.80					
$C5^{frag}$	0.99	0.93	0.83	0.99				
$C6^{frag}$	1.00	0.94	0.80	1.00	0.99			
$C7^{frag}$	0.98	0.92	0.85	0.98	0.99	0.98		
$C8^{frag}$	0.98	0.92	0.85	0.98	0.99	0.98	1.00	
	$C1^{frag}$	$C2^{frag}$	$C3^{frag}$	$C4^{frag}$	$C5^{frag}$	$C6^{frag}$	$C7^{frag}$	$C8^{frag}$

Table 10. Comparison of Average Hit-List Similarity between the Fingerprint-Based and Graph-Based Similarity Coefficients (using the asymmetric coefficient approach described in the caption to Table 5).

	RASCAL-BCI	RASCAL-Unity	RASCAL-Daylight
$C1^{frag}$ / F1	0.69	0.70	0.67
$C2^{frag}$ / F9	0.65	0.64	0.60
$C3^{frag}$ / F8	0.61	0.68	0.70
$C4^{frag}$ / F4	0.69	0.70	0.67
$C6^{frag}$ / F5	0.69	0.70	0.67
$C7^{frag}$ / F3	0.68	0.71	0.69
$C8^{frag}$ / F7	0.68	0.71	0.69

Table 11. Similarity Searches Of The MDDR Database

				Active Compounds Retrieved										Hit-List Similarity (Asymmetric)	
Activity Class	**MDDR ID**	**Number in Activity Class**	**Similarity Coefficient**	a_{25}	a_{25}^{fused}	a_{50}	a_{50}^{fused}	a_{75}	a_{75}^{fused}	a_{100}	a_{100}^{fused}	a_{150}	a_{150}^{fused}	**100 Compound Hit-List**	**150 Compound Hit-List**
Acetylcholine Esterase Inhibitor	9221	679	$F3^{(BCI)}$	14	13	17	15	17	16	17	18	19	23	0.75	0.68
			$C7^{frag}$	13		15		15		16		21			
Prolylendopeptidase Inhibitor	9248	300	$F3^{(BCI)}$	25	25	32	41	39	50	40	56	44	72	0.68	0.66
			$C7^{frag}$	24		40		46		51		56			
Lipid Peroxidation Inhibitor	12453	609	$F3^{(BCI)}$	11	13	13	13	13	15	13	16	14	17	1.00	1.00
			$C7^{frag}$	12		15		15		15		16			
Excitatory Amino Acid Inhibitor	12454	216	$F3^{(BCI)}$	20	14	22	22	24	24	24	28	26	30	0.92	0.84
			$C7^{frag}$	9		15		20		25		26			
CCK Antagonist	42711	456	$F3^{(BCI)}$	23	24	39	49	59	70	71	82	83	93	0.65	0.75
			$C7^{frag}$	24		45		60		72		98			
H+/K+ ATPase Inhibitor	54112	698	$F3^{(BCI)}$	22	23	35	37	47	52	62	64	92	91	0.57	0.74
			$C7^{frag}$	22		36		48		58		77			
IL-1 Inhibitor	2450	364	$F3^{(BCI)}$	17	21	27	27	31	31	32	34	35	41	0.88	0.83
			$C7^{frag}$	18		28		35		39		51			
Dopamine (D1) Agonist	11124	65	$F3^{(BCI)}$	9	10	11	11	11	12	11	13	11	13	0.91	0.91
			$C7^{frag}$	11		11		11		11		11			
Xanthine	27120	112	$F3^{(BCI)}$	10	11	19	18	25	25	36	30	44	38	0.58	0.70
			$C7^{frag}$	8		16		24		33		37			
Leukotriene B4 Antagonist	27214	285	$F3^{(BCI)}$	10	10	10	11	10	11	11	12	13	12	0.91	0.85
			$C7^{frag}$	10		10		11		13		13			

Table 12. Total Numbers Of Actives Retrieved by Graph-Based, Fingerprint-Based and Fused Similarity Coefficients. The Percentage Enhancement is defined as

$$100 \times \frac{a(Fused) - \max\{a(\mathrm{F3}^{(\mathrm{BCI})}), a(\mathrm{C7}^{\mathrm{frag}})\}}{\max\{a(\mathrm{F3}^{(\mathrm{BCI})}), a(\mathrm{C7}^{\mathrm{frag}})\}}$$

where $a(X)$ is the number of active molecules retrieved by similarity coefficient X

	Total Active Compounds Retrieved				
	a_{25}	a_{50}	a_{75}	a_{100}	a_{150}
$F3^{(BCI)}$	161	225	276	317	381
$C7^{frag}$	151	231	285	333	406
Fused Search	164	244	306	353	430
Percentage Enhancement	1.9	5.6	7.4	6.0	5.9

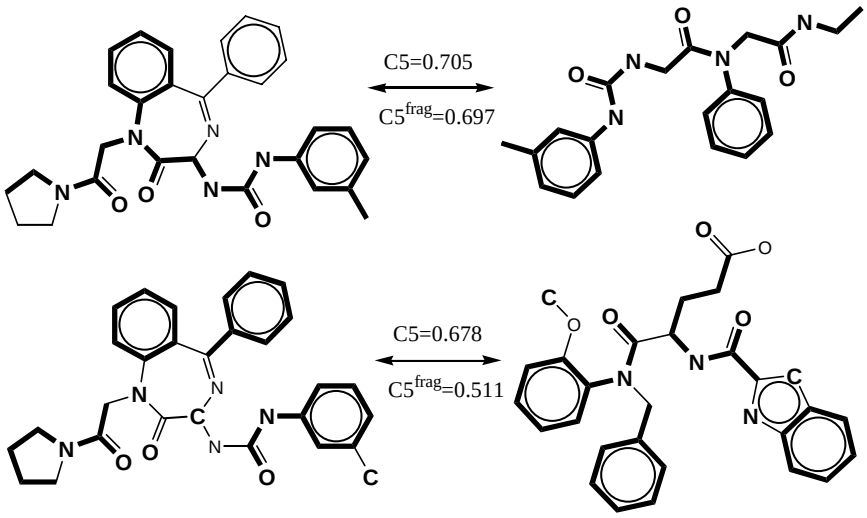

Figure 1. Difference in C5 Similarity Values Resulting from the Modified Definition of the MCES. The atoms and bonds corresponding to the respective MCES between each pair of molecules are highlighted in boldface.

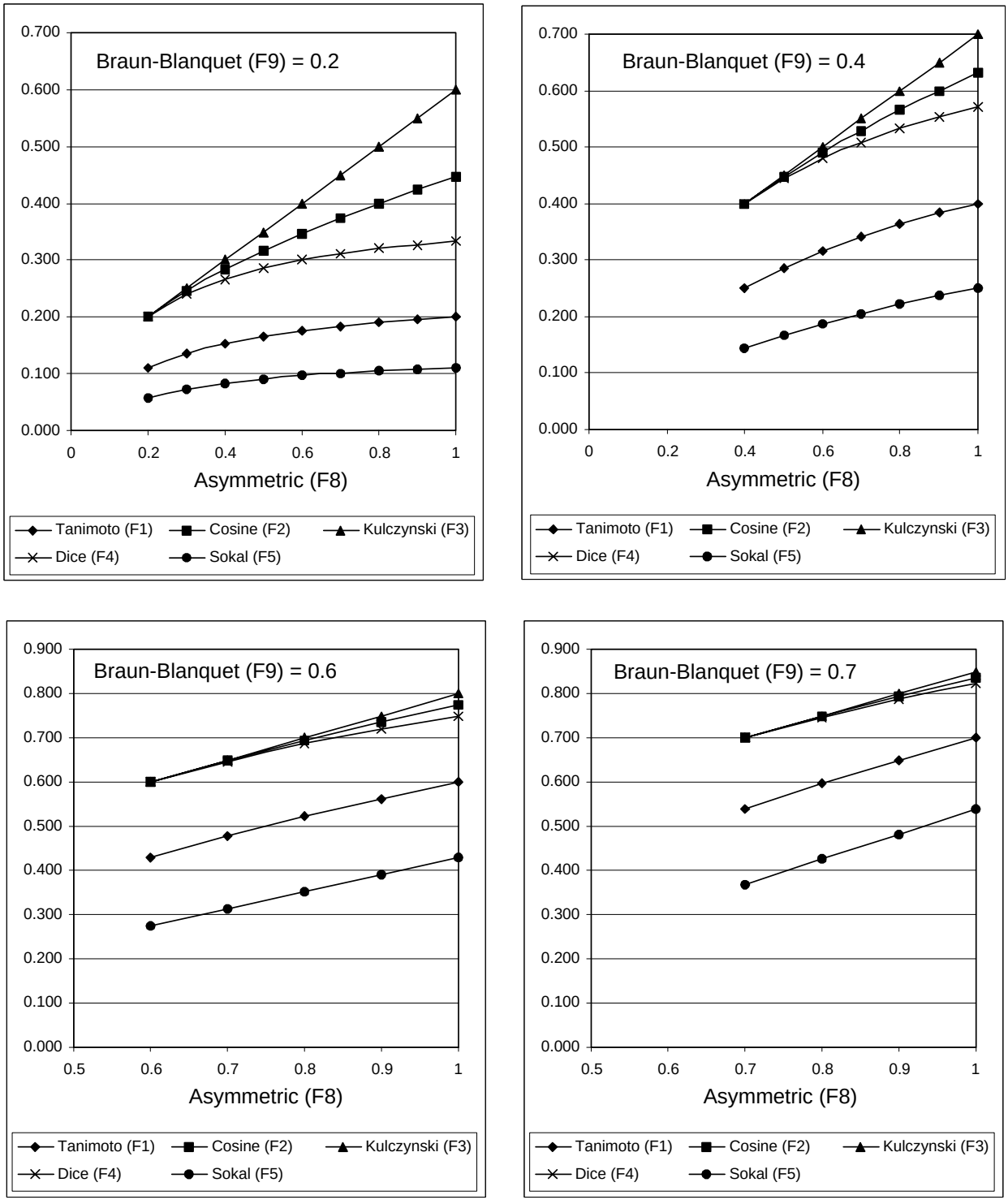

Figure 2. Relative Behavior of Similarity Coefficients