promoting access to White Rose research papers

White Rose
Research Online

Universities of Leeds, Sheffield and York
http://eprints.whiterose.ac.uk/

This is an author produced version of a paper published in **QSAR & Combinatorial Science.**

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/3608/

Published paper
Willett, P. (2006) *Enhancing the effectiveness of ligand-based virtual screening using data fusion.* QSAR & Combinatorial Science, 25 (12). pp. 1143-1152.

White Rose Research Online
eprints@whiterose.ac.uk

Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion

Peter Willett

Krebs Institute for Biomolecular Sciences and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP

Abstract: Data fusion is being increasingly used to combine the outputs of different types of sensor. This paper reviews the application of the approach to ligand-based virtual screening, where the sensors to be combined are functions that score molecules in a database on their likelihood of exhibiting some required biological activity. Much of the literature to date involves the combination of multiple similarity searches, although there is also increasing interest in the combination of multiple machine learning techniques. Both approaches are reviewed here, focusing on the extent to which fusion can improve the effectiveness of searching when compared with a single screening mechanism, and on the reasons that have been suggested for the observed performance enhancement.

Keywords: Consensus scoring, data fusion, fusion rule, ligand docking, machine learning, scoring function, similarity searching, similarity-based virtual screening, structure-based virtual screening

INTRODUCTION

Discovery programmes in the agrochemical and pharmaceutical industries make extensive use of high-throughput screening (HTS) to identify potential lead compounds that could form the basis for subsequent optimisation. The cost-effectiveness of HTS means that very many more compounds can now be tested for biological activity than was possible a decade ago; even so, the sheer number of compounds available from corporate databases and vendor catalogues (let alone those that might be synthesised using combinatorial techniques) means that it is not generally possible to screen all of the molecules potentially available to a discovery programme. Instead, carefully selected subsets of the available compounds are identified using virtual screening approaches to prioritise the biological testing of the available molecules. This ensures that those molecules that have the greatest probabilities of activity are synthesised (or acquired) and then tested at as early a stage of the project as possible [1-5].

There are many different types of virtual screening in current use but they can be sub-divided into structure-based and ligand-based approaches. Structure-based virtual screening using protein-ligand docking is the method of choice when the 3D structure of the biological target is available from X-ray or NMR studies [6-8]. Ligand-based virtual screening is appropriate when there is information relating to known (or predicted) ligands; examples of this approach are machine learning methods, in which a classification rule is developed from a training-set containing known active and known inactive molecules [9-11], and similarity methods, in which molecules are ranked in order of decreasing similarity to a known active (or actives) [12-14]. All of these types of method result in a ranking of a set of molecules: the molecules at the top of the ranking are those that have achieved the highest values for some type of scoring scheme and that are thus expected to have the greatest probability of activity once they undergo actual, rather than virtual, screening.

Descriptions of novel virtual screening methods appear regularly in the literature, and this has spurred interest in comparative studies that seek to assess the relative merits of different methods when used under the same conditions (see, e.g., [15-22]). Many of these comparative studies seek to try to identify a single, best measure, using some quantitative performance criterion [13, 23]. Such comparisons help to focus attention on methods of proven robustness; however, it is most unlikely that any one method could be expected to perform equally well under all circumstances, a point made forcibly by Sheridan and Kearsley when they note that "we have come to regard looking for 'the best' way of searching chemical databases as a futile exercise. In both retrospective and prospective studies, different methods select different subsets of actives for the same biological activity and the same method might work better on some activities than others" [13]. This view is by no means restricted to chemoinformatics, as is perhaps best exemplified by the famous "no free lunch" theorem, which shows that there is no single best approach for tackling combinatorial optimisation problems [24].

If many different virtual screening methods are available, and if none of them can be expected to be consistently superior to the others, then it seems appropriate to use not one but multiple methods to prioritise a database for biological testing. This is an example of the application of *data fusion*, a technique first developed for military signal-processing purposes but now employed in a very wide range of application areas, as demonstrated by Soong's extensive bibliography [25]. This article reviews the use of data fusion methods for ligand-based virtual screening; the use of these methods for structure-based virtual screening is normally referred to as *consensus scoring*, and is reviewed in detail by Feher [26]. Here, we introduce data fusion and its application to virtual screening, review some of the applications that have been reported, including several studies carried out in our laboratory in Sheffield, and discuss the reasons why data fusion might be expected to enhance screening performance.

THE BASIC IDEA

The idea of using multiple sources of information to investigate a problem is a very old one: we use sight and sound to cross a road; we use sight, smell and taste to evaluate a restaurant meal; and we use juries, rather than individuals, to decide the outcomes of criminal trials. However, its application to the combination of digital information sources is more recent, having first come to prominence in the Eighties for signal processing in defence applications (e.g., ocean surveillance, surface-to-air defence and battlefield target identification) and then being taken up for use in a rapidly increasing range of application domains (e.g., law enforcement, remote sensing, medical diagnosis and equipment monitoring) [27-30]. A common definition of data fusion is that recommended by the US Department of Defense Joint Directors of Laboratories Data Fusion Subpanel, as quoted by Klein: "data fusion is a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from multiple sources" [30]. This definition is very broad, including a whole range of data-capture and data-manipulation activities that are normally a given when data fusion is used for virtual screening, where the principal focus is the final activity, i.e., the act of combination. The definition is, however, useful when considering other aspects of chemoinformatics for which data fusion might be used, e.g., in analytical and structure-activity applications.

Fusion for virtual screening is based on the idea of computing a score (approximating, directly or indirectly, to the probability of activity) for each molecule in a database by using multiple scoring functions. The multiple sets of scores are then combined to obtain a better (in some sense) set of scores than could be obtained by use of just a single function. The basic procedure is shown below, where the sets of scores that are input to the combination stage can be generated using any of the structure-based or ligand-based screening approaches that are currently available.

```
For each of the scoring functions, $I$
        Compute the score $Score(I,J)$ for each database-molecule, $J$
Combine the set of scores $\{Score(I,J)\}$ for each database-molecule to give a new fused
          score, $FScore(J)$
Rank the database in decreasing order of the $FScore(J)$ values, and apply a cut-off
          to retrieve some number of the top-ranked molecules.
```

We have already noted that this review will restrict itself to ligand-based virtual screening. Even so, the general model shown above encompasses a wide range of types of fusion that could be invoked to support a screening programme; these could hence be used to categorise the many publications on data fusion that have appeared in the literature. One mode of organisation is to consider the types of score that are input to the selection algorithm: specifically, one could input quantitative data (e.g., the similarity-coefficient values from a similarity-searching routine), ordinal data (e.g., the ranks resulting from ordering the dataset into decreasing similarity-coefficient values) or categorical data (as would be the case with the outputs from a 3D pharmacophore search or from a *k*-nearest neighbour classification routine). The last of these, a situation that is often referred to as *classifier fusion* [31, 32], is of particular importance in structure-based virtual screening, where several of the fusion rules listed by Feher [26] use voting schemes in which individual docking algorithms state that a particular molecule should, or should not, be retrieved. Ligand-based approaches have more generally used either quantitative or ordinal data, with considerable discussion as to whether the former should be converted to the latter prior to the application of the fusion rule [33-35].

Use of the ranks involves a loss of information. However, in the virtual-screening context, medicinal chemists' principal concern is to determine whether or not a molecule should be considered for further analysis, rather than the magnitudes of the scores associated with those molecules. More importantly, even if different scoring functions yield the same range of scores (e.g., zero-to-unity for the binary version of many association coefficients [12]) or if range-scaling is deliberately introduced [34]), the distribution of scores may not be the same, with the possibility of introducing bias when fusion takes place. For example, an early study involved Tanimoto-based similarity measures based on three different representations to search a set of compounds with associated cellular-uptake data [33]. Ginn *et al.* showed that one of their descriptors (based on computed physical properties) yielded similarity distributions that were different from those yielded by the other two types (2D fingerprints and 3D torsional angle descriptors), and the authors hence fused ranks, rather than similarity values [33]. Analogous problems can occur if different similarity coefficients are used. For example, the Tanimoto and cosine coefficients have both been used for similarity and diversity applications that involve 2D fingerprints. These two coefficients are defined as follows: assume that the fingerprints describing two molecules have a and b bits set and that c of these are in common. Then the Tanimoto coefficient is defined to be

$$\frac{c}{a+b-c}$$

and the cosine coefficient is defined to be

$$\frac{c}{\sqrt{ab}}.$$

When two molecules are compared, the latter coefficient will always gives a score that is equal to or greater than that resulting from the Tanimoto coefficient [36], meaning that a fusion rule based on the minimum or maximum of the scores (*vide infra*) would be biased towards one input or the other. These examples have focused on similarity-based virtual screening but similar comments would apply to the use of data fusion in machine-learning environments; for example, one might score compounds using several different substructural analysis weighting schemes (such as the R1-R4 weights described by Ormerod *et al*. [15] and by Hert *et al*. [37]) or using the outputs from binary kernel discrimination and support vector machine approaches (as described recently by Jorissen and Gilson [38]). Thus, despite the potential loss of discrimination, scores are frequently converted to ranks prior to fusion. Similar comments apply to consensus scoring, where the identification of accurate scoring functions continues to be problematic [6-8].

The first chemical applications of what we would now call data fusion were probably in spectroscopy and QSAR [39-43]; there then appeared the first applications in ligand-based virtual screening [33, 44-46], and shortly afterwards the first applications in structure-based virtual screening [47-49] (as detailed in the excellent review by Feher [26]). In what follows, the reader should assume that we are dealing with ligand-based virtual screening, unless stated otherwise; that said, many of the comments are applicable to both types of screening (and there has, indeed, been interest in combining the outputs of these very different approaches [26, 50, 51]): in what follows, data fusion should be considered as referring to ligand-based virtual screening, and consensus scoring to structure-based virtual screening.

The initial studies of data fusion all focused on similarity searching; it is probably the case that this continues to be the principal focus of study for methodological developments, although straightforward applications increasingly involve other types of screening method. The basic idea underlying the use of similarity searching for virtual screening is a very simple one that was first enunciated explicitly by Johnson and Maggiora, whose Similar Property Principle states that molecules that are structurally similar are likely to have similar properties [52]. Thus, if the Principle holds, then a database-molecule that has not been tested for biological activity but that is structurally similar to a molecule known to exhibit the activity of interest (the so-called *reference* or *target* structure) then the database-molecule is also likely to be active; moreover, this molecule is more likely to be active than another database-molecule that has a lesser degree of similarity to the reference structure. A simple screening strategy hence involves using a similarity measure to compute the degree of resemblance between the known reference structure and each of the database molecules, and then ranking these molecules in decreasing order of the computed similarities.

Two papers by a group at Merck first demonstrated the potential of data fusion for enhancing the effectiveness of similarity-based virtual screening. Previous work had established the general applicability of two small atom-bond substructural fragments, the atom pair and the topological torsion, for 2D similarity searching [13]. These fragments were further developed by using physicochemical properties to describe the atoms [44], thus providing a fuzzy matching capability, and by using inter-atomic distances, rather than through-bond distances, to encode 3D information [45]. The Merck in-house similarity searching system computes a number of these fragments for each of the molecules that are to be screened; a Dice-like coefficient is then computed for the similarity of each such molecule to the reference structure that forms the basis for a similarity search, where the Dice coefficient is

$$\frac{2c}{a+b}$$

(using the notation given previously for the Tanimoto and cosine coefficients). The user has the option to choose specific pairs of fragments as being of importance for the search and then to compute a score for these *combination descriptors* by taking the mean of the coefficient values for the two chosen fragment-types. The database is sorted into decreasing order of the resulting mean scores, and the highest ranked compounds passed on for further investigation. Alternatively, the database is sorted into decreasing order for each of the individual scores, and the two resulting ranks for each compound compared to find the lower value (i.e., nearer the top of the ranking). The database is then sorted into decreasing order of these *minimum ranks*. Searches for compounds belonging to specific pharmacological classes showed that the combination and the minimum rank approaches both performed on average as well as, or slightly better than, the better individual descriptor in each case.

APPLICATIONS OF DATA FUSION

Work in Sheffield

The Merck studies described in the previous section were closely followed by two from Ginn *et al*. [33, 46]. These studies again focused on the potential benefits to be gained by using similarity measures based on multiple structure representations, and were occasioned by work on the use of data fusion in information retrieval (IR) systems, specifically on the combination of the rankings produced by different retrieval mechanisms when applied to databases of textual documents. An early IR investigation by Belkin *et al*. [53] combined the results of multiple text-database searches that had been conducted in response to a single user query but that employed different indexing and searching strategies. Each such strategy yielded a ranking of the text database that was being searched and the set of rankings was then combined using simple arithmetic fusion rules, such as taking the largest rank (MAX), the smallest rank (MIN, as in the Merck study mentioned previously) or the sum of ranks (SUM); this work soon led to many other studies and data fusion is now a standard approach in IR (as reviewed by Croft [54] and by Hsu and Taksa [55]).

The first study by Ginn *et al*. was conducted as part of an evaluation of the EVA descriptor for similarity applications. The EVA (for EigenVAlue) descriptor is derived from IR- and Raman-range molecular vibrations that are typically obtained through the application of a classical normal co-ordinate analysis to an appropriately energy-minimised 3D structure. Similarity searches were conducted using Unity 2D fingerprints and EVA, and the resulting rankings, individual and fused, were then used for simulated property prediction of logP values. The detailed experiments that were carried out showed that fused searches could yield improved predictions in some cases [46]. The second study was more wide-ranging, using three very different datasets and several different types of structural descriptor, both 2D and 3D, for each dataset [33]. The experiments showed that the simple SUM fusion rule resulted in an average level of search performance that was at least as good as the best individual measure: since the latter often varied unpredictably from one search to another search, it was concluded that the use of a fusion rule would generally provide a more consistent level of search performance than would a single similarity measure. That said, some of the experiments involved varying both the similarity coefficient and the structure representation, so that it was not possible to identify the precise reason for the observed performance enhancements. Since these initial studies by Ginn *et al*., there have been several subsequent projects in our laboratory that have sought to establish the general applicability of data fusion, as summarised below.

The measurement of inter-molecular structural similarity based on 2D fingerprints has been studied for many years (see, e.g., [16, 17, 36, 44-46, 56]). Much of this work has used the familiar Tanimoto coefficient to compare pairs of fingerprints, but two recent studies compared a total of 22 different coefficients that could be used for similarity-based virtual screening [57, 58]. Whilst many of the coefficients were shown to yield comparable results, some differences were apparent. Research was hence undertaken to determine whether data fusion could further improve search performance as compared to the use of just the Tanimoto coefficient; thus, whereas previous studies had combined multiple representations of the reference structure, the aim here was to use a fixed representation but with multiple similarity coefficients. Salim *et al*. hence carried out extensive simulated virtual screening experiments on the *MDL Drug Data Report* database (MDDR, available from MDL Information Systems Inc. at http://www.mdli.com) [58]. The experiments used 13 different coefficients to search for molecules characterised by three different types of 2D fingerprint and belonging to seven bioactivity classes of current pharmaceutical interest. The searches involved all the individual coefficients, fused searches using all possible pairs of coefficients, fused searches containing all possible triples of coefficients etc. (analogous combinatorial studies of consensus scoring functions for ligand docking have been reported by Yang *et al.* [35] and Oda *et al.* [59]). Analysis of the extensive results showed that combinations of between two and four coefficients could improve screening performance over searches using just the industry-standard Tanimoto coefficient. However, the results were extremely inconsistent, with no one combination of coefficients providing a consistently high level of performance, and with the best-performing combination for one biological target often performing poorly in searches for a different target. This lack of consistency has been observed in other screening studies [60-62].

It was disappointing to find that it was not consistently possible to identify some single combination that could be expected to enhance the effectiveness of screening in all circumstances. However, a subsequent detailed analysis by Holliday *et al.* revealed that this was due in large part to the marked biases (which could be either positive or negative) that many coefficients have for the retrieval of molecules of a particular size (as reflected in the numbers of bits set in their fingerprints) [63]. The effect of molecular size on the performance of the Tanimoto coefficient for similarity and diversity applications had been noted previously [64, 65]; the studies by Salim *et al.* and Holliday *et al.* demonstrated the generality of this behaviour and its effect on the performance of data fusion based on multiple coefficients.

Thus far, we have assumed that similarity-based data fusion involves combining the rankings (or similarities) that result from searching a database with a single bioactive reference structure but with multiple similarity measures, an approach that Whittle *et al.* refer to as *similarity fusion* [34]. The alternative, *group fusion* approach involves combining the rankings (or similarities) that result from searching a database with a single similarity measure (e.g., 2D fingerprints and the Tanimoto coefficient) but with multiple bioactive reference structures. The idea of combining structural information from multiple molecules is by no means new [66-71]; drawing on earlier work by Xue *et al.* [69] and Schuffenhauer *et al.* [70], Whittle *et al.* [34] and Hert *et al.* [72] studied the search-effectiveness of group fusion, in comparison with both conventional similarity searching and similarity fusion. They found that better results were obtained from using similarity scores, rather than rank positions (a not unexpected finding since ranks are commonly used in similarity to alleviate problems resulting from the different similarity distributions engendered by different similarity measures (*vide supra*)) and that the MAX fusion rule gave better results than the SUM rule. Extensive searches of the MDDR database showed clearly the benefits obtainable from group fusion. In particular, it was found that picking as few as ten active reference structures and combining them using group fusion enabled searches to be carried out that were comparable to even the very best from amongst many hundreds of conventional similarity searches using individual reference structures. Further searches were carried out using pharmacological activity classes that had been chosen to reflect a range of structural diversities [34, 37]. These experiments demonstrated that the benefits of group fusion are greatest when the sought actives are structurally diverse; conventional similarity searching or similarity fusion, conversely, are most effective when the actives are strongly clustered in structural space. Similarity fusion and group fusion would thus appear to be complementary in character.

Hert *et al*. have also described a modification of conventional similarity searching that makes use of group fusion [37, 73]. Given a bioactive reference structure, the top-ranked structures resulting from a similarity search are expected to have a high probability of activity as a consequence of the Similar Property Principle (*vide supra*); Hert *et al*. made the assumption that such molecules *are* indeed active and that they can hence be used as the reference structures for further similarity searches that can then be combined using group fusion; similar ideas have been used previously in IR [74] and bioinformatics [75]. Extensive searches of the MDDR database demonstrated that this activity-assumption resulted in searches that were nearly always superior to conventional similarity searching (where just the initial reference structure is used) in its ability to identify active molecules, with some of the increases in performance being quite marked.

Some recent studies

A data fusion system has two principal components: the functions that are used to score each of the molecules in the database that is to be screened; and the fusion rule that is used to combine the sets of resulting scores. Simple variations in these two components hence permit a very large number of different types of fusion to take place: for example, one could generate scores (or ranks) using different machine-learning tools, such as substructural analysis, a support vector machine and binary kernel discrimination *inter alia*; and one could combine scores (or ranks) by summing them, by taking the maximum or by taking the geometric mean *inter alia*. Given this range of possibilities it is hardly surprising that many reports are now appearing on the use of data fusion for virtual screening: we discuss below several recent papers that demonstrate the current state of the art in data fusion, and show how it is starting to be used as a standard component of systems for virtual screening.

The early comparison of fusion rules by Ginn *et al*. [33] found that the SUM rule, when applied to ranks, yielded a consistently high level of performance, and one that was generally superior to those obtained from use of the MIN or MAX rules (*vide supra*). The latter two rules represent the assignment of extreme ranks to database structures and it is thus hardly surprising that both can be highly sensitive to the presence of a single outlier screening method amongst those that are being combined, whereas the SUM rule is expected to be more stable to the presence of an outlier or of noisy input rankings. This fusion rule, normally with ranks but sometimes with scores, has hence been the method of choice for several years: it is, however, by no means the only type of rule that might be employed. Thus, Feher lists no less than nine types of fusion rule that have been used for consensus scoring, these including [26]: voting procedures (each scoring function returns a yes/no vote for a molecule and the overall decision is based on the number of votes); arithmetic procedures based on ranks or scores (such as the SUM, MAX and MIN rules); weighted arithmetic procedures (where ranks or scores are weighted on the basis of the presumed effectiveness of the associated scoring function); and statistical procedures (such as MLR or PLS, these requiring the availability of training data to compute the various parameters involved [76]).

A detailed comparison of six fusion rules for similarity searching of GPCR assay data using 2D and 3D descriptors has been reported by Baber *et al*. [62]. The rules included voting, SUM, and regression procedures, and their extensive comparison suggested that the two best methods were summed ranks and logistic regression; the latter gave the best virtual screening performance but involved a training stage in which the parameters of the logistic model needed to be determined for each type of GPCR and each type of scoring function. The performance of SUM was only marginally inferior and this, of course, can be used directly in the absence of training data; accordingly Baber *et al.* concluded that SUM was the fusion rule of choice for lead-discovery or scaffold-hopping searches, but that logistic regression should be used at a later stage in a project, e.g., during the optimization phase when considerable amounts of assay data will have become available. A detailed comparison of rules for classifier fusion by Kittler *et al.* ascribed the superiority of SUM, when compared with a range of alternative arithmetic procedures (minimum, maximum, median, product and majority voting), to its robustness in the face of errors in estimating the probability of membership for each of the existing categories [31].

Raymond *et al*. have recently introduced a new fusion rule, called *conditional probability,* that is based on estimating the probability that a database molecule, *J*, is active given a similarity score *Score*(*I*,*J*) with respect to an active reference structure, *I* [61]. The overall score for *J* is then obtained as the product of the probabilities computed for each of the individual scoring functions, and ranking of the database in terms of the decreasing product scores hence represents a ranking in decreasing probability of activity; a similar approach for ranking documents in IR has been reported by Manmatha *et al*. [77]. Conditional probability is rather more complex than sum-of-scores for two reasons. First it is necessary to establish the nature of the correlation between probability of activity and similarity score for each of the scoring methods that is to be combined, this involving a training stage (as with several other studies, e.g., [62, 76, 78]). Second, combining the individual probabilities by means of a product function assumes that the scores and rankings produced by the different scoring methods are statistically independent, and this is most unlikely to be the case in practice. However, the assumption does not appear to lessen the effectiveness of the procedure since experiments with a range of descriptors (including 2D fingerprints, 2D and 3D maximum common substructure procedures, and two shape-matching procedures) suggested that conditional probability performed as well as or better than summing the ranks.

Zhang and Muegge report a study of group fusion applied to scaffold-hopping searches of seven MDDR activity classes using 2D and 3D descriptors [21]. In each case, searches were carried out using multiple actives as the reference structure, and the resulting sets of scores combined using six different fusion rules: average of the ranks or of the Tanimoto scores; weighted voting based on the ranks or the scores; a consensus of the previous four rules; and maximum of the ranks or of the scores. The results were very variable, with the best overall performance being the four-rule consensus followed by the maximum of the Tanimoto similarities (which had been shown previously to perform well in the study by Whittle *et al*. [34]), and with the best group-fusion searches sometimes out-performing flexible ligand docking. Analogous detailed comparisons of different consensus rules for structure-based virtual screening have been reported by Yang *et al*. [35] and by Oda *et al*. [59].

Finally in this section, it is worth noting three studies that have considered machine-learning, rather than similarity-searching, methods. Jorissen and Gilson have discussed a virtual screening system that has been developed at the Centre for Advanced Research in Biotechnology and that is based on a support vector machine (SVM) [38]. Their experiments involved fusing the SVM rankings with those from a BKD routine using the SUM rule, and showed that the fused rankings were comparable with or superior to the better of the individual scoring functions. However, it should be noted that the SVM used 50 calculated physicochemical descriptors obtained with a variable-selection routine whereas the BKD routine used a hashed 2D fingerprint, so that it is not clear whether the performance enhancement arose from the multiple scoring functions or from the multiple representations that were used. Jorissen and Gilson's study involved the fusion of rankings, but many machine-learning methods function as classifiers, categorising input molecules as being either predicted active or predicted inactive. In such cases, simple voting schemes may be used to fuse the outputs of the individual classifiers, e.g., retrieve those molecules that are predicted to be active by at least two of the classifiers. We note two such recent studies. Plewczynski *et al*. report the use of support vector machines, random forests, neural networks, *k*-nearest neighbour classification, trend vectors, naive Bayesian classification and decision trees to categorise sets of ligands for five pharmaceutically important biological targets [79]. Their study focused on the performance of the individual methods but they also investigated how voting might affect performance. They found that consensus voting could indeed bring about substantial increases in precision (i.e., a reduction in the number of false positives), but that (hardly surprisingly) this was often accompanied by comparably large decreases in recall. Finally, Givehchi and Schneider have reported the combination of the outputs from three different types of artificial neural network based on seven different types of 2D descriptor to disambiguate G-protein coupled receptor (GPCR) molecules from non-GPCR molecules [80]. The outputs from the trained neural networks, either binary outputs or actual classification scores, were combined using a jury voting procedure, with a resulting noticeable improvement in predictive performance.

REASONS FOR THE EFFECTIVENESS OF DATA FUSION

There is considerable evidence for the belief that data fusion is an effective tool for virtual screening, with positive results having been achieved ever since the initial experiments by the Merck and Sheffield groups. By effective, we mean here the ability to retrieve active molecules from the database that is being searched than would be the case in a conventional database search. That said, it is important to define what is meant by "conventional". In some of the early studies by Ginn *et al.* [33], SUM fusion of ranks gave better results than even the best of the individual similarity searches that were fused; more generally, however, experiments have shown that fusion results are comparable to the best individual scoring function or better than the average function. Importantly, the best individual function tends to vary from search to search whereas an effective fusion rule is robust to changes in reference structure, database and biological target, ensuring a consistent level of search performance. Cases in structure-based screening where consensus scoring is deleterious have been reviewed by Feher [26].

If several search methods are available, each with their different characteristics, then it might seem reasonable to combine them all in a search. Sheridan and Kearsley have argued strongly for such an approach on the basis of their experiences at Merck [13], and it is also the conclusion of Kogej *et al.* from an extended analysis of similarity searches of AstraZeneca assay data that used a voting fusion rule to combine the results of searches based on nine different types of 2D fingerprint [81]. Others, however, have argued for the combination of smaller numbers of searches. In particular, the theoretical analysis by Wang and Wang (*vide infra*) suggests that little benefit is to be gained from using more than three or four scoring functions, a conclusion that is in line with several practical studies of both structure-based and ligand-based fusion [35, 58, 59].

The variations in performance that have been noted in several of the studies cited here have led to interest in the factors that determine whether or when data fusion can be expected to work. This has been the subject of debate in the IR and pattern recognition context for some years [31, 55, 82-85] but the first such report in the chemoinformatics area was a simulation study by Wang and Wang [86]. Each of a set of 5000 hypothetical compounds was assigned a random number representing its experimental binding affinity from a Normal distribution; the 100 molecules with the largest assigned affinities were deemed to be the active molecules for the simulation. These numbers were then perturbed by another random number drawn from a Normal distribution of lower variance to represent the affinity predicted by a scoring function; this second stage was repeated ten times for each hypothetical molecule and the resulting set of scores fused using one of three consensus rules (mean of the scores, mean of the ranks and a voting procedure). The simulation demonstrated that performance increased rapidly with an increase in the number of scoring functions (although the latter increase levelled off once three or four functions had been included in the consensus) and that the consensus performance was superior to any individual scoring function. The latter result was explained by the simple statistical fact that the mean of repeated samplings will tend to be closer to the true value than any individual sampling: in other words, multiple rankings will better reproduce the ideal ranking (i.e., the ranking in decreasing order of experimental affinity) than will any individual ranking. Their analysis also demonstrated that this effect would provide diminishing benefits once more than about four rankings were included in a consensus. Wang and Wang thus provide an elegant explanation of the observed behaviour but their simulation has been criticised by both Baber *et al.* [62] and Verdonk *et al.* [87] as the simulation assumes that the scoring functions that are being combined are of comparable effectiveness and are independent; neither of these assumptions are likely to be the case in practice.

There have been several studies of what is required for successful fusion, with three recent papers focussing on criteria for successful consensus scoring [35, 59, 62]. Yang *et al.* describe a combinatorial study of scoring functions for ligand docking [35]. They start by noting that the results of previous studies of consensus scoring seem to depend on the fusion rule and on the number and the nature of the scoring functions used; they then go on to describe two criteria by which the performance of a consensus can be evaluated. These criteria are: the ratio of the effectiveness (however defined) of the best scoring function in a consensus to that of the worst; and the rank-score graph, where a normalised version of the score produced by a scoring function is plotted against the ranks of the compounds when ordered by those scores. From a combinatorial study of all of the 31 combinations possible with five different scoring functions, Yang *et al.* concluded that the best results were obtained with just two of the functions that performed well on their own and that exhibited different rank-score graphs; they also noted that the fusion of ranks performed at least as well as the fusion of scores. These findings are analogous to previous reports in the IR literature: for example, Ng and Kantor have suggested the use of the Kendall rank correlation coefficient to determine the degree of correlation between the rankings from two scoring functions, and hence the extent to which it would be worth including them in a consensus [83] (see also Hsu and Tahesi [55] and Beitzel *et al.* [85]). Yang *et al.* suggest that the rank-score graph could be used to select effective combinations of scoring functions in the absence of actual performance data.

The requirement that the individual components of an effective consensus should be effective in their own right [35] seems an entirely reasonable one, but one that is not fully supported by the combinatorial study of Oda *et al.* [59]. This was on an impressive scale with 511 combinations, all those possible from nine scoring functions, being tested with up to nine different consensus rules and 220 ligand-protein systems. The extensive results were far from consistent, but the authors were able to suggest that scoring functions that were ineffective on their own could be effective in a consensus when they compensated for shortcomings in other functions in the consensus; similar results are reported in Givehchi and Schneider's work on the combination of neural network classifiers [80]. It should be noted that while this conclusion contradicts the first criterion of Yang *et al.*, it is in agreement with their second, i.e., that the components of a successful consensus should be different in nature.

Reference has already been made to the comparison of consensus rules by Baber *et al.* [62]. Their study also investigated in some detail how and why consensus scoring works. They start by describing the assumptions inherent in the simulation study by Wang and Wang (*vide supra*) and question the extent to which these assumptions might be appropriate in an experimental context. They then note that the Tanimoto similarities between pairs of active molecules in their experiments were consistently larger than between pairs of inactive molecules (as would be expected if the Similar Property Principle holds for the compounds in the four internal GPCR discovery programmes considered by Baber *et al.*) The actives are hence more tightly clustered than are the inactives: when multiple scoring functions are used they are likely to repeatedly select many actives but not necessarily the same inactives. This suggestion (which is in agreement with work in IR by Lee [82]) was confirmed by an analysis using the Kendall Coefficient of Concordance [88], which showed that rankings of the actives by six different scoring functions were much more closely correlated than the rankings of the inactives.

The studies reported thus far, both in this section and earlier, hence suggest that the different rankings (whether in ligand-based or in structure-based virtual screening) should be statistically independent but that at the same time there should be at least some level of correlation between them if positive reinforcement of the positions of the actives is to take place. That said, the studies have been far from unequivocal in their findings, and Whittle *et al.* have hence recently reported a rigorous theoretical approach to the modelling of data fusion in the context of virtual screening [89]. The principal focus of their work is similarity fusion using pairs of fingerprint-based similarity coefficients, but extensions are reported to encompass more than two similarity coefficients, to encompass multiple structure representations, and to encompass group fusion. The analysis also takes explicit account of a possible limitation in the analyses by Yang *et al*. and by Baber *et al*. discussed previously. Both of these reports compared the behaviour of different scoring functions across entire ranked databases, whereas practical applications of consensus scoring normally involve just some small fraction of a database (e.g., the top-5% of the ranked molecules). It is the difference in behaviour in this part of the ranking, rather than the entire ranking, that is likely to be of importance in the identification of scoring functions that are sufficiently disparate in character to enhance screening performance.

The theoretical model of Whittle *et al*. shows that the origin of performance enhancement for simple fusion rules can be traced to a combination of differences between the retrieved active (i.e., true positives) and retrieved inactive (i.e., false positives) similarity distributions and the geometrical difference between the regions of these multivariate distributions that the chosen fusion rule is able to access. For pair-wise similarity fusion, a simple analytical model demonstrates some conditions for which both the SUM and MAX rules have at least the potential to enhance screening performance when compared with conventional similarity searching. Indeed, an upperbound analysis shows that this enhancement should be obtainable on a routine basis given sufficiently large amounts of training data; however, as Baber *et al*. have noted previously, this is typically not available at an early stage of a discovery programme (which is where similarity searching is most commonly used).

More generally, the analysis reveals that the operation of a data fusion system is far more complex than previously realised, involving subtle interactions between multiple factors that, taken together, severely complicate attempts to predict the effect of fusion on search performance. Thus, the combination of just two lists of similarity values (e.g., fusing the results of searches using the Tanimoto coefficient and the cosine coefficient) depends on eight distinct distributions: the retrieved–active and retrieved-inactive distributions of both lists for both matched and unmatched compounds, where a matched (or unmatched) compound is that retrieved by both (or by just one) of the two similarity searches that are being fused. Moreover, different fusion rules will be differentially affected by the interactions between these eight factors. It is hence not surprising that the behaviour of data fusion is frequently found to be inconsistent for different datasets. Whittle *et al*. demonstrate that positive fusion can be obtained for two similarity searches if the bivariate distribution of the retrieved active and retrieved inactive similarity scores are different (which supports the views of Baber *et al*. and Yang *et al*. regarding the need for different rankings). However, this behaviour may occur only for some parts of a database ranking, and performance enhancement will result only if this bivariate distribution matches the appropriate integration region for the chosen fusion rule more closely than do either of the individual search regions.

Whittle *et al*. suggest that SUM is likely to yield better results than does MAX for similarity fusion, that group fusion is much more likely to offer performance benefits than does similarity fusion, and that MAX is more appropriate than SUM for group fusion; all of these predictions are in line with experimental results in previous studies of data fusion. Given the evident success of the model it is unfortunate that the principal conclusion is that data fusion is too complex a procedure to enable a simple prediction of the behaviour that can be expected in any specific circumstances.

CONCLUSIONS

Over the last few years, data fusion has become accepted as a simple way of enhancing the performance of existing systems for ligand-based virtual screening, by combining the results of two or more screening methods. In some cases, the fused search may be better than even the best individual screening method; more generally, when averaged over large numbers of searches, the fused search provides a high level of consistency that is better than that obtainable from any individual screening method. Whilst the practical advantages of data fusion are now well understood, there is still much dispute as to the reasons for the observed behaviour, with a recent analysis suggesting that it may not be feasible to provide reliable predictions as to when fusion will be beneficial in a practical context. This is not to say that alternative types of theoretical model may not be more successful in elucidating the key criteria for performance enhancement; in the interim, it is perhaps best merely to note that data fusion is a simple, computationally efficient technique that can often enhance the performance of existing systems for ligand-based virtual screening.

Acknowledgements. My thanks are due to Val Gillet, Clare Ginn, Jerome Hert, John Holliday, Naomie Salim and Martin Whittle for their contributions to the work on data fusion that has been carried out in this laboratory, and Martin Whittle for helpful comments on this paper.

REFERENCES

1. H.-J. Bohm, G. Schneider, Eds., *Virtual Screening for Bioactive Molecules*; Wiley-VCH, Weinheim, **2000**.
2. G. Klebe, Ed., *Virtual Screening: an Alternative or Complement to High Throughput Screening*; Kluwer, Dordrecht, **2000**.
3. T.I. Oprea, *Molecules* **2002**, *7*, 51-62.

4. J. Bajorath, *Nature Rev. Drug Discov.* **2002**, *1*, 882-894
5. F.L. Stahura, J. Bajorath, *Combin. Chem. High-Through. Screen.* **2004**, *7*, 259-269.
6. I. Halperin, B. Ma, H. Wolfson, R. Nussinov, *Proteins* **2002**, *47*, 409-443.
7. P.D. Lynn, *Drug Discov. Today* **2002**, *7*, 1047-1055.
8. R.D. Taylor, P.J. Jewsbury, J.W. Essex, *J. Comput.-Aid. Mol. Design* **2002**, *16*, 151-166.
9. R.D. Cramer, G. Redl, C.E. Berkoff, *J. Med. Chem.* **1974**, *17*, 533-535.
10. R. Burbridge, M. Trotter, B. Buxton, S. Holden, *Comput. Chem.* **2001**, *26*, 5-14.
11. G. Harper, J. Bradshaw, J.C. Gittins, D.V.S. Green, A.R. Leach, *J. Chem. Inf. Comput. Sci.* **2001,** *41***,** 1295-1300**.**
12. P. Willett, J.M. Barnard, G.M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
13. R.P. Sheridan, S.K. Kearsley, *Drug Discov. Today* **2002**, *7*, 903-911.
14. A. Bender, R.C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.
15. A. Ormerod, P. Willett, D. Bawden, *Quant. Struct.-Activ. Relat.* **1989**, *8*, 115-129.
16. R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1-9
17. X. Chen, C.H. Reynolds, *J. Chem. Inf. Comput. Sci***. 1998,** *42*, 1407-1414.
18. G. Cruciani, M. Pastor, R. Mannhold, *J. Med. Chem.* **2002**, *45*, 2685-2694.
19. E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, *Proteins* **2004**, *57*, 225-242.
20. J.C. Cole, C.W. Murray, J.W. Nissink, R.D. Taylor, R. Taylor, *Proteins* **2005**, *60*, 325-332.
21. Q. Zhang, I. Muegge, *J. Med. Chem.* **2006**, *49*, 1536-1548.
22. G.L. Warren, C.W. Andrews, A.-M.Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, M.S. Head, *J. Med. Chem.*, in the press.
23. S.J. Edgar, J.D. Holliday, P. Willett, *J. Mol. Graph. Model.* **2000**, *18*, 343-357.
24. D.H. Wolpert, W.G. Macready, *IEEE Trans. Evolut. Comput.* **1997**, *1*, 67-82.
25. R. Soong, *The Data Fusion Bibliography*, at URL http://www.zonalatina.com/datafusion.doc
26. M. Feher, *Drug Discov. Today* **2006**, *11*, 421-428.
27. E. Waltz, J. Llinas, *Multisensor Data Fusion*, Artech House, Norwood MA, **1990**.
28. D.L. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood MA, **1992**.
29. I.R. Goodman, R.P.S. Mahler, H.T. Nguyen, *Mathematics of Data Fusion*, Kluwer, Norwell MA, **1997**.
30. L.A. Klein, *Sensor and Data Fusion Concepts and Applications*, SPIE Optical Engineering Press, Bellingham WA, 2nd edition, **1999**.
31. J. Kittler, J.M. Hatel, R.P. Duin, J. Matas, *IEEE Trans. Patt. Anal. Mach. Intell.*, **1998**, *20*, 226-239.
32. B.F. Buxton, W.B. Langdon, S.J. Barrett, *Measure. Control* **2001**, *34*, 229-234.
33. C.M.R. Ginn, P. Willett, J. Bradshaw, *Perspect. Drug Discov. Design* **2000**, *20*, 1-16.
34. M. Whittle, V.J. Gillet, P. Willett, A. Alex, J. Losel, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840-1848.
35. J.-M. Yang, Y.-F. Chen, T.-W. Shen, B.S. Kristal, D.F. Hsu, *J. Chem. Inf. Model.* **2005**, *45*, 1134-1146.
36. J.D. Holliday, S.S. Ranade, P. Willett, *Quant. Struct.-Activ. Relat.* **1995**, *14*, 501-506.
37. J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Model.* **2006**, *46*, 462-470.
38. R.N. Jorissen, M.K. Gilson, *J. Chem. Inf. Model.* **2005**, *45*, 549-561.
39. T.Clerc, F. Erni, F., *Topics Curr. Chem.* **1973**, *39*, 91-107.
40. D.L. Duewer, R.D. Clark, *J. Chemomet.* **1991**, *5*, 503-521.
41. Ajay, *Chemomet. Intell. Lab. Systems* **1994**, *24*, 19-30.
42. H. Masui, M. Yoshida, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 294-298.
43. S. So, M. Karplus, *J. Med. Chem.* **1996**, *39*, 1521-1530
44. S.K. Kearsley, S. Sallamack, E.M. Fluder, J.D. Andose, R.T. Mosley, R.P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118-127.
45. R.P. Sheridan, M.D. Miller, D.J. Underwood, S.K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128-136.

46. C.M.R. Ginn, D.B. Turner, P. Willett, A.M. Ferguson, T.W. Heritage, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23-37.
47. P.S. Charifsen, J.J. Corkery, M.A. Murcko, W.P. Walters, *J. Med. Chem.* **1999**, *42*, 5100-5109.
48. M. Stahl, M. Rarey, *J. Med. Chem*. **2001**, *44*, 1035-1042.
49. R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, J.B. Matthew, *J. Mol. Graph. Model.* **2002**, *20*, 281-295.
50. J.Mestres, R.M.A. Knegtel, *Perspect. Drug Discov. Design* **2000**, *20*, 191-207.
51. X. Fradera, R.M. Knegtel, J. Mestres, *Proteins* **2000**, *40*, 623-636.
52. M.A. Johnson, G.M. Maggiora, Eds., *Concepts and Applications of Molecular Similarity*, Wiley, New York, **1990**.
53. N.J. Belkin, P. Kantor, E.A. Fox, J.B. Shaw, *Inf. Proc. Manag.* **1995**, *31*, 431-448.
54. W.B. Croft, Combining approaches to information retrieval. In: W.B. Croft, Ed., *Advances in Information Retrieval*, Kluwer, Boston, 2000, pp. 1-36.
55. D.F. Hsu, I. Taksa, *Inf. Retriev.* **2005**, *8*, 449-480.
56. P. Willett, V. Winterman, D. Bawden, *J. Chem. Inf. Comput. Sci.*, **1986**, *26*, 36-41.
57. J.D. Holliday, C.-Y. Hu, P. Willett, *Combin. Chem. High-Through. Screen.* **2002**, *5*, 155-166.
58. N. Salim, J.D. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435-442.
59. A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, S. Hirono, *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 380-391.
60. M. Whittle, P. Willett, W. Klaffke, P. van Noort, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449-457.
61. J.W. Raymond, M. Jalaie, M.P. Bradley, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601-609.
62. J.C. Baber, W.A. Shirley, Y. Gao, M. Feher, *J. Chem. Inf. Model*. **2006**, *46*, 277-288.
63. J.D. Holliday, N. Salim, M. Whittle, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819-828.
64. D.R. Flower, *J. Chem. Inf. Comput. Sci.* **1988**, *38*, 379-386.
65. M.A. Fligner, J.S. Verducci, P.E. Blower, *Technometrics* **2002**, *44*, 110-119
66. N.E. Shemetulskis, D. Weininger, C.J. Blankey, J.J. Yang, C. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862-871.
67. R.P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456-1469.
68. S.B. Singh, *J. Med. Chem*. **2001**, *44*, 1564-1575.
69. L. Xue, F.L. Stahura, J.W. Godden, J.Bajorath, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 746-753.
70. A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 391-405.
71. L. Xue, J.W. Godden, F.L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1218-1225.
72. J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.
73. J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Med. Chem.* **2005**, *48*, 7049-7054.
74. W.B. Croft, T.J. Lucia, J.K. Cringean, P. Willett, *Inf. Proc. Manag.* **1989**, *25*, 599-614.
75. S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, *Nucleic Acids Res*. **1997**, *25*, 3389-3402.
76. G.E. Terp, B.N. Johansen, I.T. Christensen, F.S. Jorgensen, *J. Med. Chem*. **2001**, *44*, 2333-2343.
77. R. Manmatha, T. Rath, F. Feng, *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* **2001**, *24*, 267-275.
78. S. Betzi, K. Suhre, B. Chetrit, F. Guerlesquin, X. Morelli, *J. Chem. Inf. Model.*, **2006**, *46*, 1704-1712.
79. D. Plewczynski, S.A.H. Spieser, U. Koch, *J. Chem. Inf. Model*. **2006**, *46*, 1098-1106.
80. A. Givehchi, G. Schneider, *Mol. Divers*. **2005**, *9*, 371-383.

81. T. Kogej, O. Engkvist, N. Blomberg, S. Muresan, *J. Chem. Inf. Model*. **2006**, *46*, 1201-1213.
82. J.H. Lee, *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* **1997**, *20*, 267-276.
83. K.B. Ng, P.B. Kantor, *J. Amer. Soc. Inf. Sci.* **2000**, *51*, 1177-1189.
84. N.S.V. Rao, *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 904-909.
85. S.M. Beitzel, E.C. Jensen, A. Chowdhury, D. Grossman, N. Goharian, O. Frieder, *J. Amer. Soc. Inf. Sci. Tech*. **2004**, *55*, 859-868.
86. R. Wang, S. Wang, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422-1426.
87. M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T.M. Mooij, C.W. Murray, R.D. Taylor, P. Watson, *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 793-806.
88. S. Siegal, N.J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*; McGraw Hill, New York, **1988**.
89. M. Whittle, V.J. Gillet, P. Willett, J. Losel, *J. Chem. Inf. Model.* **2006**, *46*, in the press.