Chemoinformatics at the University of Sheffield 2002-2014

Valerie J. Gillet, John D. Holliday and Peter Willett*

Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP.

ABSTRACT

This paper summarises work in chemoinformatics carried out in the Information School of the University of Sheffield during the period 2002-2014. Research studies are described on fingerprint-based similarity searching, data fusion, applications of reduced graphs and pharmacophore mapping, and on the School's teaching in chemoinformatics.

Keywords: similarity searching, fingerprints, reduced graphs, pharmacophore mapping, chemoinformatics education

INTRODUCTION

Work at the University of Sheffield in the field of chemoinformatics (or, as it was then called, chemical structure handling) started in 1965 with the appointment to a faculty position of Prof. Michael Lynch. He had previously been the director of basic research at Chemical Abstracts Service, where he had overseen some of the earliest work anywhere on the use of computers for the generation and searching of both bibliographic and chemical databases. On coming to Sheffield he rapidly established an active programme of teaching and research that has now continued for almost fifty years. During this time, the Information School has made significant contributions to a wide range of topics in chemoinformatics such as the design of 2D and 3D substructure search systems, the representation and searching of the generic structures in chemical patents, pharmacophore mapping and the use of similarity and diversity methods inter alia. This work has been summarised in several previous publications.[1-4]

The work of the chemoinformatics research group in Sheffield has always had a strong algorithmic and methodological focus, this reflecting our location in an informatics, rather than a chemical, academic department. We have thus drawn extensively on computational techniques from, e.g., graph theory,[5-6] cluster analysis,[7-8] image processing[9-10] and combinatorial optimisation[11-12] inter alia to design and implement a wide range of chemoinformatics applications. Lynch and Willett[1] and Bishop et al.[4] have described Sheffield work in chemoinformatics for the periods 1965-1985 and 1986-2002, respectively. In this paper we summarise some of our more recent contributions to the discipline, focusing on studies that we have carried out into similarity-based virtual screening, the

combination of rankings using data fusion, applications of reduced graph representations of chemical structures and pharmacophore mapping.

SIMILARITY-BASED VIRTUAL SCREENING

Similarity searching is one of the most popular forms of ligand-based virtual screening[13-16] and has been one of our principal research areas for many years. Indeed, the widespread use of 2D fingerprints and the Tanimoto coefficient for computing molecular similarity is arguably due in large part to one of the first operational systems for similarity searching that was developed in a Sheffield collaboration with Pfizer in the mid-Eighties.[17] Work in the Nineties focused on the development and testing of methods for 3D similarity (as summarised by Bishop et al.[4]) but our more recent studies have re-visited the implementation of similarity searching based on 2D fingerprints, focusing on the use of weighting schemes to enhance screening effectiveness and on data fusion, which is discussed in more detail in the next section.

The fingerprints used for similarity searching are typically binary in character, with the setting of bits in a fingerprint corresponding to the presence or absence of particular substructural fragments within a molecule. A more sophisticated representation is obtained when each bit is replaced by an integer or real value that reflects a fragment's specific contribution to the calculation of inter-molecular similarity, with the largest weights being assigned to the most important fragments. Fragment weights are widely used in machine learning approaches to virtual screening[18-20] but these require the availability of extensive training-sets of active and inactive molecules, whereas in similarity searching the only information that is typically available is a single bioactive reference structure. That said, there is one additional source of information that could be exploited in a similarity calculation, viz information about the frequencies with which fragments occur: either the frequencies with which they occur within individual molecules; or the frequencies with which they occur in the entire database that is being searched.

Analogous schemes have been extensively studied for the weighting of indexing terms in information retrieval, where they have been consistently shown to enhance the effectiveness of text searching. Given the relationships that exist between information retrieval and chemoinformatics[21] it hence seems reasonable to apply such schemes in the latter context. Both types of weighting had in fact been considered in Willett and Winterman's early studies of fingerprint-based similarity measures;[22] however, these involved QSAR datasets containing just a few tens of molecules, rather than the huge databases that need to be searched in modern chemoinformatics research. Arif et al. hence undertook a detailed study of these two types of weighting scheme when used with 2D fingerprints in

Frequency weighting	Inverse frequency weighting
$x_i = f_i$	$x_i = \ln\left(\frac{N}{T_i + 1}\right)$
$x_i = \ln(f_i)$	$x_i = \ln\left(\frac{N}{T_i}\right) + 1$
$x_i = \sqrt{f_i}$	$x_i = \ln\left(\frac{N + 0.5}{T_i + 0.5}\right)$
$x_i = 0.5 + 0.5\frac{f_i}{\max\{f_i\}}$	$x_i = \sqrt{\frac{\max\{T_i\}}{T_i}}$

Table 1. The fingerprint describing a molecule is considered as a vector, X, with the i-th fragment occurring f_i times ($f_i \geq 0$) in a molecule. For the frequency weights in the left-hand column, $\max\{f_i\}$ denotes the frequency for the most frequently occurring fragment in the molecule. For the inverse frequency weights in the right-hand column, the i-th fragment occurs in a total of T_i molecules ($T_i \geq 0$) in the N-molecule database and $\max\{T_i\}$ denotes the number of molecules containing the most frequently occurring fragment. In conventional, unweighted fingerprints $x_i=1$

experiments using sets of bioactive molecules from the MDL Drug Data Report (MDDR) and World of Molecular Bioactivity (WOMBAT) databases.[23-24]

The first type of weighting, frequency weighting, is based on the assumption that a fragment that occurs several times in a molecule should make a greater contribution to the overall degree of similarity than if it occurs just once, and that this contribution should be still greater if that fragment also occurs multiple times in the molecule with which it is being compared. Arif et al. considered several different ways of using the occurrence information, as detailed in the left-hand side of Table 1, and concluded that the best screening results were obtained by using the square root of the occurrence frequencies.[23] The effect of this scheme is to lessen the contribution of the more generic fragments that can occur relatively frequently within molecules, and that can thus yield high values if raw occurrence counts are used without some form of normalisation. Turning to the second type of weighting, inverse frequency weighting, the basic assumption here is that two molecules that share an infrequently occurring feature (such as a rare heterocycle) should be considered as being more similar to each other than if they share a feature (such as a benzene ring) that occurs very frequently throughout the database that is being searched. This seems inherently reasonable but Arif et al. were unable to demonstrate that taking such information into account, as detailed in the right-hand side of Table 1, was uniformly beneficial; instead the results seemed to depend on the structural diversity of the set of actives that was being sought in a similarity search.[24] Specifically, it was found that while a weight of the form $\ln(N/T_i)+1$ (the second of the inverse frequency weights in Table 1) proved to be effective in similarity searches for structurally homogeneous sets of actives, simple, unweighted binary fingerprints were more effective for the chemically more important problem of searching for diverse sets of actives.

Both of these studies used a non-binary version of the Tanimoto coefficient, which has for long been the standard similarity coefficient for use with conventional binary fingerprints.[13] However, it was clear in both cases that the effectiveness of this particular coefficient for virtual screening could, in some circumstances, be strongly affected by the precise nature of the weighting scheme that was being employed. This characteristic of the coefficient was investigated in a subsequent study by Holliday et al. who showed that a related similarity coefficient, the cosine coefficient, was less dependent on the precise nature of the weighing scheme that was being employed and that it was, accordingly, to be preferred if weighted fingerprints were to be used for similarity searching.[25]

The weighting-scheme studies summarised here are just one aspect of the work we have conducted on 2D similarity searching over the past decade: other studies have included inter alia an investigation of the extent to which fingerprint-based similarity measures can be used for scaffold-hopping applications,[26] the use of reduced graphs for similarity searching as described further below, similarity searching using Bayesian inference networks,[27] a detailed comparison of the characteristics of different similarity coefficients that can be used with binary fingerprints,[28] and our work on data fusion, as discussed in the next section.

DATA FUSION

Data fusion is the name given to a body of techniques that were first developed for signal processing in defence applications but that are now used in a very wide range of domains.[29] Thus far, its main application in chemoinformatics has been for ligand-based or structure-based virtual screening (where, in the latter case, it is often referred to as consensus scoring).[30-31] The basic idea underlying data fusion is that the use of multiple sources of evidence will result in better decision making than if only a single source of evidence is available. Thus, in the context of ligand-based virtual screening, the use of multiple screening methods is expected to increase the extent to which active molecules are clustered at the top of a database ranking, with our work on data fusion focussing on the combination of multiple similarity measures. Early studies, both in Sheffield and by the Sheridan group at Merck,[32-33] used data fusion to combine the rankings obtained when similarity searches were carried out for a reference structure using two or more similarity measures, e.g., Daylight fingerprint searches using the Tanimoto coefficient and using the cosine coefficient (a process we refer to as similarity fusion). More recent work in Sheffield has focussed on three areas: using multiple reference structures and a single similarity measure (a process we refer to as group fusion) rather than the multiple similarity measures and a single reference structure that characterise similarity fusion; evaluating the effectiveness of different fusion rules, i.e., different ways of combining multiple

similarity rankings (however produced); and finally trying to provide an underlying theoretical basis for what has always been a purely empirical approach to virtual screening.

Group fusion was first studied in a collaboration with Novartis that compared different ways in which multiple reference structures could be used to enhance ligand-based virtual screening.[34] The results were unequivocal in highlighting the effectiveness of group fusion, and this finding was confirmed in other studies that compared it with similarity fusion and with conventional similarity searching across a range of types of bioactivity data.[35-36] The advantages of group fusion were greatest when searching for structurally diverse sets of bioactive molecules, suggesting its use in the scaffold-hopping and bioisostere studies that are of particular importance in the lead-discovery stage of drug research.[37] Since the initial Sheffield studies, the group fusion approach has been widely adopted. One application, turbo similarity searching, provides a simple way of enhancing conventional similarity searching when just a single bioactive reference structure is available. The similar property principle would suggest that the nearest neighbours of the reference structure in a similarity search are likely to exhibit the same bioactivity; if we then assume that they actually do exhibit that activity then we can use them as pseudo-reference structures in a group fusion search, combining the rankings resulting from their use with that resulting from the initial reference structure.[38-39]

Thus far, we have described data fusion as involving the combination of multiple rankings of a database to produce a single, fused ranking that is the output from a similarity search. Our initial studies used simple arithmetic fusion rules that had first been described for the combination of rankings in textual information retrieval systems[40] as exemplified in Table 2. For example, the fused ranking for a database structure might be the sum of its rank positions in the individual rankings that were to be combined, or the fused ranking might be based on the sum of the similarity scores in the individual similarity searches. Chen et al.[41] conducted a detailed study of fusion rules for use in similarity fusion and group fusion, and demonstrated the general efficacy of the reciprocal rank rule (denoted by RKP in Table 2), in which the overall score for a database structure is the sum of the reciprocal ranks in each of the individual rankings that are to be fused.[42] An analysis of this rule suggests that its effectiveness may derive from the close relationship that exists between the reciprocal rank of a database structure and its probability of activity.

Work in Sheffield and elsewhere[30-31] has shown clearly that fusion-based screening is often comparable with, or superior to, the best of the individual similarity searches that are being combined and that screening effectiveness is much more consistent from search to search than when just a single similarity method is available. Is there an underlying theoretical rationale for this empirical finding? This question was investigated by Whittle et al., who developed and tested an analytical model of

Fusion rule	Formula
MAX	$\max\{S_1 \ldots S \ldots S_n\}$
SUM	$\frac{1}{n}\sum_{i=1}^{n} S_i$
MNZ	$p\sum_{i=1}^{n} S_i$
RKP	$\sum_{i=1}^{p} \frac{1}{R_i}$

Table 2. Examples of fusion rules that can be used to combine the results associated with a specific database structure in multiple similarity searches. In these rules, it is assumed that the structure achieves a similarity score of S_i in the i-th ($1 \leq i \leq n$) of the n similarity searches (or a ranking of R_i if the database structures are sorted into decreasing order of the similarity values for the i-th search). It is also assumed that p ($1 \leq p \leq n$) is the number of times that the structure is ranked above a user-defined threshold, e.g., it occurs in the top-1% of the ranking.

fusion-based similarity searching.[43-44] The model successfully predicted that group fusion would generally be superior to similarity fusion and that certain fusion rules would be superior to others. However, the model was very complex, requiring a large amount of training data that would not normally be available in the early stages of a drug programme when similarly-based virtual screening is normally employed and that ruled out its intended use for predicting the utility of new types of fusion rule. A more successful analytic study was reported by Holliday et al..[45] They demonstrated that a power law distribution could be used to predict to a fair degree of accuracy the numbers of database structures common to one similarity search, to two similarity searches, to three etc. They also demonstrated that the proportion of actives increased rapidly in these sets of common structures: the probability of activity of a database structure hence increases in line with its frequency of retrieval in multiple similarity searches, thus providing a simple empirical justification for the use of fusion methods in virtual screening.

REDUCED GRAPHS

Reduced graphs provide summary representations of chemical structures by collapsing groups of bonded atoms into nodes, the nodes are then connected according to the arrangement of the fragments in the original structures. They were first used by the Sheffield group as one component of a search system for Markush structures, also known as generic chemical structures.[46] A Markush structure typically consists of a core structure to which variable substituents are attached. The substituents can be specific substructures, such as methyl or ethyl, or they can be defined using generic nomenclature, such as alkyl group. Markush structures are used in chemical patents to allow a collection of related compounds to be protected rather than just a single specific compound. In the Markush search

system, reduced graphs are used to provide a level of search that is intermediate between a fingerprint-based screening step and a full atom-by-atom search. The generic chemical structures are reduced to simpler graphs by collapsing ring systems into ring nodes and acyclic chains into non-ring nodes. Each node is then further characterised according to properties such as the number and type of atoms it represents. Reduced graphs offer an intermediate level of search since they retain topological information, unlike fingerprints, while being much smaller than the original structures from which they are derived. They also allow substituents represented by generic nomenclature such as acyclic alkyl group consisting of 1-4 carbon atoms to be compared against specific substructures such as methyl or ethyl.

Since 2002, the Sheffield group has investigated the use of reduced graphs in a number of applications including similarity searching, analysis of HTS data, clustering and identification of structure-activity relationships. Our approach to representing structures as reduced graphs is close in concept to the Feature Trees developed by Rarey et al.[47] and the ErGs presented by Steifl et al.[48] Gillet et al. investigated a variety of graph reduction schemes for similarity searching and found the Ar/F(4) reduced graphs shown in Figure 1 to be most effective for scaffold hopping.[49] In this graph reduction scheme, a chemical structure is described as a set of ring, feature and linker nodes. Ring nodes are defined as aromatic or aliphatic. Acyclic chains are fragmented into feature and linker nodes by identifying carbon atoms which are not doubly or triply bonded to heteroatoms, so called isolating carbons; connected isolating carbon atoms form linker nodes with the remaining fragments defining feature nodes. Feature nodes and ring nodes are further characterised as hydrogen bond donor, hydrogen bond acceptor, or both donor and acceptor, as appropriate. The characterisation of nodes is typically done using user-defined SMARTS descriptions.

Different methods for comparing two reduced graphs have also been developed. Gillet et al. represented the reduced graphs as binary fingerprints analogous to Daylight fingerprints.[49] The resulting fingerprint is typically much sparser than the fingerprint generated from the original structure due to the much smaller number of nodes in a reduced graph, compared to atoms in the chemical structure. Improved performance was obtained by representing the reduced graphs as node-pair descriptors,[50] which are similar in concept to the more familiar atom-pair descriptors developed by Carhart et al.[51] The Sheffield work was further developed by Harper et al., who used an edit distance method to determine the similarity between two reduced graphs in which the minimum cost of converting one reduced graph to the other is calculated by considering operations such as mutation, insertion and deletion of nodes.[52] This approach can be useful when a small change in structure gives rise to a relatively large change in the reduced graph, such as the substitution of a carbon to a heteroatom in the middle of a carbon chain. Birchall et al. then trained a genetic algorithm (GA) to optimise the relative costs associated with different node operations.[53] For example, the cost of

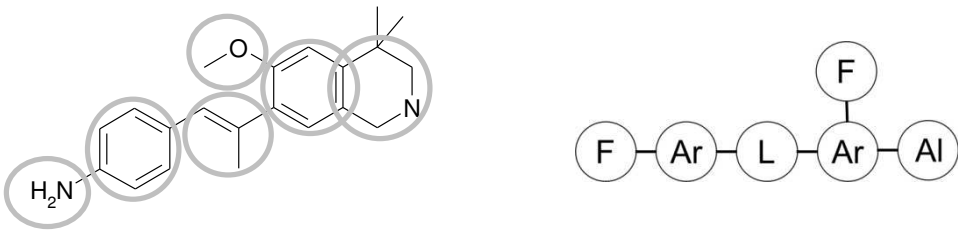


Figure 1. The chemical structure is reduced to aliphatic (Al), aromatic (Ar), linker (L) and feature (F) nodes. These are further characterised as donor, acceptor and donor and acceptor as appropriate.

substituting a donor node by a negatively ionisable node may be weighted differently to the cost of substituting it by a linker node. The GA was trained to identify costs that were applicable across several activity classes as well as on specific activity classes. Class specific weights were shown to both improve retrieval performance and provide clues on the underlying structure-activity relationships.

Following the earlier work by Takahashi,[54] Barker et al. used graph matching techniques to determine the similarity between two reduced graph representations.[6] The reduced graphs were represented as fully connected graphs with every node connected to every other node and the edges labelled according to the shortest bond distance between the two nodes as measured in the original graph. Two reduced graphs were then compared using a maximum common subgraph algorithm. The graph matching approach proved to be more effective in both recall of actives and the diversity of the actives retrieved than the fingerprinting methods, and clearly demonstrated the scaffold hopping capabilities of the reduced graphs.

Gardiner et al. developed a tool based on reduced graphs for browsing the output of a conventional 2D clustering method.[55] Clustering is often used as a way of organising large sets of compounds for review, for example, the results from a high-throughput screen where the aim is to extract structure-activity information from clusters enriched with active compounds. The most commonly used clustering techniques are based on traditional 2D fingerprints; however, it can be difficult to decipher the structural commonalities that are present within the resulting clusters. In Gardiner's work, the compounds within a cluster are represented as reduced graphs and a maximum common substructure (MCS) algorithm is applied iteratively to identify one, or more, reduced graphs that represent the compounds in the cluster. The reduced graphs can enable key functionalities of the compounds to be easily identified as well as multiple series present within the same cluster and related clusters can be found by comparing representatives from different clusters.

Reduced graphs have also been used to derive structure-activity relationship models. For example, Birchall et al. developed an evolutionary algorithm to grow reduced graph queries (subgraphs) that discriminate between actives and inactives in high throughput screening data.[56] The algorithm was demonstrated on datasets extracted from the MDDR where it was shown to give good classification rates with the resulting reduced graph queries encoding structure-activity information that was readily interpreted by chemists. The approach was extended to allow multiple structure-activity relationships to be extracted from a single activity class.[57] This was achieved by introducing an objective that scored each query on its uniqueness. A query was deemed to have high uniqueness if it retrieved actives that were not retrieved by other queries in the population so that each query represented a different set of active compounds.

The extent to which reduced graphs are able to identify bioisoteres has also been investigated.[58] Bioisosteres are structural fragments that can be exchanged without significantly altering a molecule's biological activity. The fragments may be quite different in structure, for example, tetrazole and carboxylic acid, so that traditional 2D fingerprints are unlikely to be effective at finding fragments that can exhibit similar behaviour. Bioisosteres were extracted from the BIOSTER database [59] and encoded as reduced graphs using a reductions scheme similar to the Ar/F(4) scheme described above. Many cases were found where the bioisosteric groups were represented either by nodes of the same type or closely related nodes (e.g. aromatic donor and aromatic donor and acceptor) demonstrating that reduced graphs do enable equivalences to be found between structurally distinct fragments. However, attempts at including this information on bioisosteric equivalences in reduced graph-based similarity searching led to only modest improvements in performance. This is due to the bioisoteric equivalences being just as likely to occur in inactive structures as they are in actives, rather than being a problem inherent in the reduced graph approach itself.

PHARMACOPHORE MAPPING

A pharmacophore describes the three-dimensional arrangement of electronic features that enables a small molecule to bind to a specific biological target and trigger or block its biological response.[60] Typical features that can give rise to inter-molecular interactions are hydrogen bond donors, hydrogen bond acceptors, ionic, aromatic and hydrophobic groups, and these are often referred to as pharmacophoric features. A pharmacophore can be used to search a database of small molecules to identify any that have pharmacophoric features that map onto pharmacophore and therefore could potentially bind to the target. Pharmacophore mapping is the process of determining a pharmacophore and can be structure-based or ligand-based.[61-62] In structure-based methods, the three-dimensional structure of the binding site is used to suggest locations for complementary features in a small

molecule. Ligand-based methods are used when the three dimensional structure of the target is unknown and an attempt is made to elucidate the requirements for binding from a series of known active molecules. In ligand-based pharamacophore elucidation, the aim is to superimpose the active compounds such that their pharmacophoric features are aligned. This is typically a challenging problem, especially for flexible ligands, since the active conformations are unknown and also because there are usually many features in the molecules that could be overlaid but not all are essential for binding. Therefore, except in the simplest of cases, it is unreasonable to expect that the correct solution can be found unambiguously. Various scoring functions have been devised to guide the pharmacophore generation process and typically a set of potential pharmacophore hypotheses are generated which are then validated externally by, for example, visual inspection and testing against held-back actives that were not used to generate the pharmacophore.

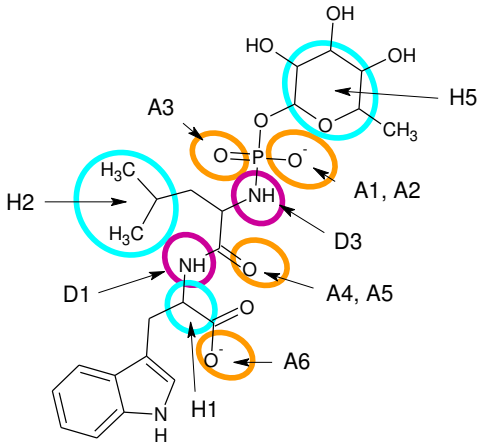
Sheffield's interest in ligand-based pharmacophore mapping commenced with studies in the late Eighties on the use of clique detection algorithms for the identification of 3D maximum common substructures,[63] with the Bron-Kerbosch algorithm highlighted by this work being subsequently incorporated in the DISCO program developed by Martin et al. at Abbott Laboratories.[64] GASP was developed later in collaboration with the Wellcome Research Laboratories, with the program subsequently being distributed by Certara (previously Tripos Inc.)[65] It uses a genetic algorithm to explore the conformational space of the ligands simultaneously with different mappings between their features. The hypotheses are scored using a weighted sum fitness function comprising the conformational energies of the ligands, the volume of the overlay and the goodness of fit of the individual molecules to the pharmacophore. In 2002, Patel et al. evaluated the effectiveness of the most commonly used pharmacophore mapping programs at that time, namely GASP, CATALYST and DISCO, on a test set derived from X-ray crystal structures of protein-ligand complexes.[66] Five proteins were selected that each had crystal structures with different ligands bound. A pharmacophore was derived for each protein by superimposing the ligands according to the protein binding site. Each program was then evaluated on its ability to reproduce the manually derived pharmacophores, in the absence of knowledge of the proteins. This study highlighted several limitations in these first generation programs and they have now been superseded by a second generation of pharmacophore mapping programs.

To address limitations in GASP, the Sheffield group looked at modifying the genetic algorithm to a multiobjective evolutionary algorithm where the three objectives are handled independently rather than being combined into a single weighted function;[67] the rationale being that the objectives in GASP can be competing, for example, a better fit to a hypothesis may be achievable if the molecules are assumed to adopt higher energy conformations. In the multiobjective algorithm, Pareto ranking is used to find a set of optimal solutions that represent different compromises in the objectives, without

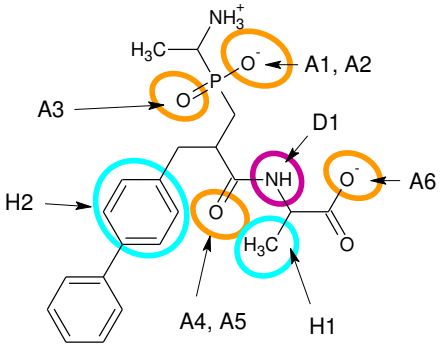
the need to assign relative weights. Cottrell et al. then extended this approach to permit partial matches between a ligand and the pharmacophore hypotheses, overcoming the requirement in GASP that all ligands match all features of the pharmacophore, which limited its practical use to sets of two or three ligands.[68] In subsequent work, Gardiner et al. biased the conformations explored during the search using torsion angle distributions in Mogul which are derived from the Cambridge Crystallographic Database.[12] Figure 2 shows an alignment produced by the algorithm (bottom right) for a set of four neprilysin ligands that have been extracted from X-ray crystal structures 1rh1, 1dmt, 1r1j and 1y8j. The alignment obtained by superimposing the X-ray structures is shown bottom left.

Taylor et al. subsequently developed a novel deterministic approach for the generation of overlays based on bitstring operations, where the bitstrings describe triplets of features within each ligand.[69] New objective functions were also developed for scoring the overlays with the potential solutions being Pareto-ranked. A diverse subset of the best scoring solutions is selected and these can be refined, by relaxing the ligand conformations and also looking for alternative ligand conformations for a given set of feature matches, before being mapped onto a low dimensional space which aids visualisation of the final set of hypotheses. The algorithm was tested on ten protein-ligand families with the number of ligands in each set varying from two to sixteen and performed well finding a solution close to the known solution for eight of the ten cases. The performance was less good for cases where the known pharmacophore consisted of fewer than three full pharmacophore features, however, a good solution was found for one of these cases by imposing an order on the way the ligands were overlaid.

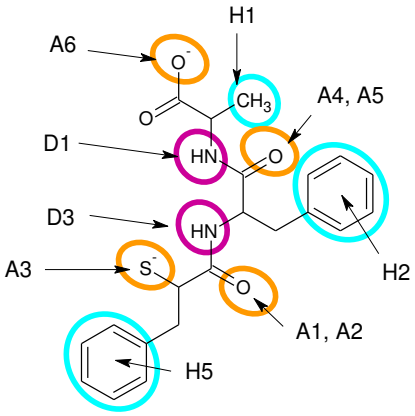
Other work on pharmacophore mapping at Sheffield involved the development of a novel approach to matching flexible 3D molecules, based on an algorithm from computer vision research, that forms one of the principal building blocks in the GALAHAD program distributed by Tripos Inc.[9, 70]



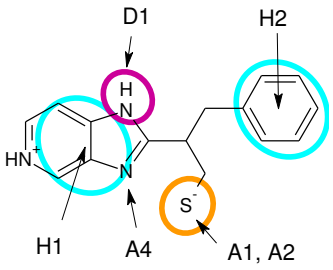
1dmt



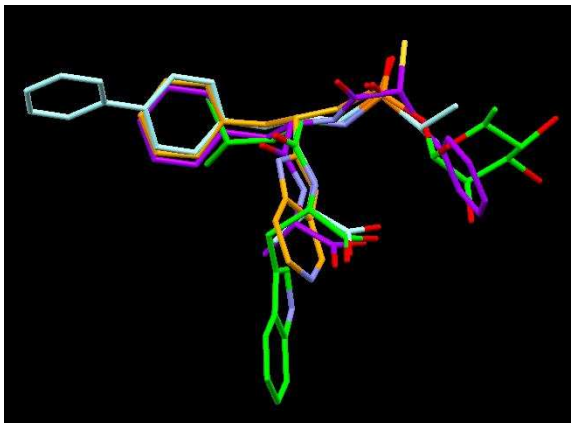
1r1h



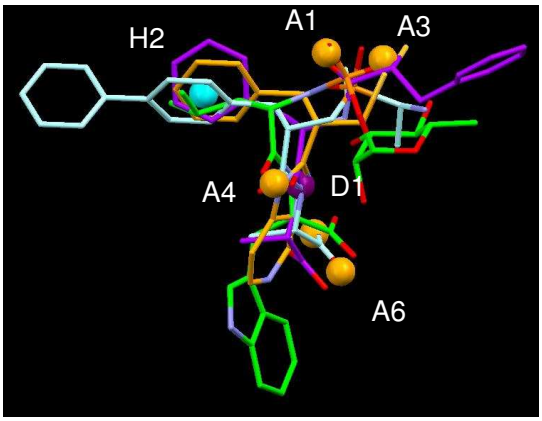
1r1j



1y8j



Xray Alignment



MOGA alignment

Figure 2. The bottom right image shows an alignment produced by the multiobjective pharmacophore method for the four neprilysin ligands extracted from X-ray crystal structures and shown above. The bottom left image shows the X-ray alignment which was generated by superimposing the protein-ligands complexes using the protein active sites. Hydrophoboc features are shown in cyan, donor features in purple and acceptor features in orange.

OTHER CONTRIBUTIONS

The previous sections have described work undertaken in Sheffield in four of our major areas of study. However, we have also made other contributions, e.g., in the application of chemoinformatics techniques to the representation and searching of biological macromolecules such as protein and RNA structures,[71-75] in the analysis of matched molecular pairs[76] and in charting the historical development of chemoinformatics and of its associated literature.[77-80]

We have also sought to influence the development of the field by means of conferences and educational programmes. Thus, in 1998 we were invited by the Chemical Structure Association Trust (CSAT) and the Molecular Graphics and Modelling Society (MGMS) to host a meeting on Computational Approaches to the Design and Analysis of Combinatorial Libraries. The success of the conference has resulted in a continuing series of conferences under the general title of the Joint Sheffield Conference on Chemoinformatics. The conference is held every three years, in the year preceding the triennial Noordwijkerhout International Conference on Chemical Structures, with the most recent Sheffield conference being held in July 2013. Sponsorship from the CSAT and the MGMS, and from a host of chemical software and database companies, has enabled us over the years to support the attendance of large numbers of doctoral students at the meetings, which often provide them with an initial opportunity to present a poster describing their research to an international audience.

This paper has focused on research in Sheffield, but we have also been active in teaching the subject; indeed, arguably the very first textbook in the field was based on material taught in a one-semester module that Lynch developed shortly after coming to Sheffield.[81] The first full Masters programme in Chemoinformatics was launched in Sheffield in 2000 supported by funding from the UK's Engineering and Physical Sciences Research Council and ran until 2008. The programme included substantial engagement with industry, with most students undertaking industrial placement dissertation projects where they carried out research of interest to the host organisation. Some of these projects resulted in publications; for example, the comparison of pharmacophore programs by Patel et al. described earlier began as an MSc dissertation project[66] and a bibliometric study of the Journal of Chemical Information and Modeling, the core journal in chemoinformatics.[80] The lack of an up-to-date textbook led to Gillet co-authoring An Introduction to Chemoinformatics[82] (first edition was published in 2003), which became a core text for this programme as well as for the Masters courses at University of Manchester Institute of Science and Technology (now University of Manchester) and Indiana University.[83] The MSc was mainly taken by new chemistry graduates, many of whom went onto careers in Chemoinformatics. The MGMS also approached Sheffield to run a short course to address the need to up-skill chemists and computational chemists already working in

the industry who had not received formal Chemoinformatics training. The course A Practical Introduction to Chemoinformatics is run annually, attracting up to twenty delegates from around the globe.

It will thus be seen that the Information School has made substantial contributions to many aspects of chemoinformatics. The work is widely recognised: thus the ca. 140 articles published by the authors in the period 2002-2014 have already attracted ca. 3300 citations in the Thomson Reuters Web of Science database, and the authors' contributions to chemical information science made them the recipients of the 2013 Jason Farradane Award, where they were noted as providing "the most widely recognized and well-established research and teaching base in the field." Our work continues, with recent publications covering topics as diverse as the use of 2D similarity measures for the registration of orphan drugs,[84] the identification of jumping emerging patterns for the generation of toxicity alerts,[85-86] and the use of reaction vectors for de novo design.[87-88]

REFERENCES

[1] M. F. Lynch, P. Willett, *J. Inf. Sci.* **1987**, *13*, 221-234.
[2] M. F. Lynch, P. Willett, *Chim. Oggi* **1990**, *8*, 55-63.
[3] P. Willett, *J. Med. Chem.* **2005**, *48*, 4183-4199.
[4] N. Bishop, V. J. Gillet, J. D. Holliday, P. Willett, *J. Inf. Sci.* **2003**, *29*, 249-267.
[5] J. W. Raymond, E. J. Gardiner, P. Willett, *Comput. J.* **2002**, *45*, 631-644.
[6] E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, V. J. Gillet, P. Kitts, P. Willett, *J. Chem. Inf. Model.* **2006**, *46*, 503-511.
[7] T. Varin, R. Bureau, C. Mueller, P. Willett, *J. Mol. Graphics Modell.* **2009**, *28*, 187-195.
[8] C.-W. Chu, J. D. Holliday, P. Willett, *Bioorg. Med. Chem.* **2012**, *20*, 5366-5371.
[9] N. J. Richmond, P. Willett, R. D. Clark, *J. Mol. Graphics Modell.* **2004**, *23*, 199-209.
[10] R. L. Martin, E. Gardiner, V. J. Gillet, S. Senger, *J. Chem. Inf. Model.* **2012**, *52*, 757-769.
[11] V. J. Gillet, *Methods in Molecular Biology* **2004**, *275*, 335-354.
[12] E. J. Gardiner, D. A. Cosgrove, R. Taylor, V. J. Gillet, *J. Chem. Inf. Model.* **2009**, *49*, 2761-2773.
[13] P. Willett, *Drug Discovery Today* **2006**, *11*, 1046-1053.
[14] P. Willett, *Annu. Rev. Inf. Sci. Technol.* **2009**, *43*, 3-71.
[15] D. Stumpfe, J. Bajorath, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260-282.
[16] G. M. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2014**, *57*, 3186-3204.
[17] P. Willett, V. Winterman, D. Bawden, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36-41.
[18] B. B. Goldman, W. P. Walters, *Annu. Rep. Comput. Chem.* **2006**, *2*, 127-140.
[19] J. L. Melville, E. K. Burke, J. D. Hirst, *Comb. Chem. High Throughput Screening* **2009**, *12*, 332-343.
[20] J. B. O. Mitchell, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 468-481.
[21] P. Willett, *Information Research* **2000**, *5*, http://InformationR.net/ir/5-2/infres52.html.
[22] P. Willett, V. Winterman, *Quant. Struct.-Act. Relat.* **1986**, *5*, 18-25.
[23] S. M. Arif, J. D. Holliday, P. Willett, *J. Comput.-Aided Mol. Des.* **2009**, *23*, 655-668.
[24] S. M. Arif, J. D. Holliday, P. Willett, *J. Chem. Inf. Model.* **2010**, *50*, 1340-1349.
[25] J. D. Holliday, P. Willett, H. Xiang, *Int. J. Chemoinf. Chem. Eng.* **2012**, *2*, 28-41.
[26] E. J. Gardiner, J. D. Holliday, C. O'Dowd, P. Willett, *Future Med. Chem.* **2011**, *3*, 405-414.
[27] B. Chen, C. Mueller, P. Willett, *J. Cheminf.* **2009**, *1*.

[28] R. Todeschini, V. Consonni, H. Xiang, J. D. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901.
[29] B. V. Dasarathy, *Inf. Fusion* **2010**, *11*, 299-300.
[30] P. Willett, *QSAR Comb. Sci.* **2006**, *25*, 1143-1152.
[31] P. Willett, *J. Chem. Inf. Model.* **2013**, *53*, 1-10.
[32] 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.
[33] C. M. R. Ginn, D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23-37.
[34] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.
[35] M. Whittle, V. J. Gillet, P. Willett, A. Alex, J. Loesel, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840-1848.
[36] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Model.* **2006**, *46*, 462-470.
[37] S. R. Langdon, P. Ertl, N. Brown, *Mol. Inf.* **2010**, *29*, 366-385.
[38] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Med. Chem.* **2005**, *48*, 7049-7054.
[39] E. J. Gardiner, V. J. Gillet, M. Haranczyk, J. Hert, J. D. Holliday, N. Malim, Y. Patel, P. Willett, *Stat. Anal. Data Mining* **2009**, *2*, 103-114.
[40] N. J. Belkin, P. Kantor, E. A. Fox, J. B. Shaw, *Inf. Process. Manage.* **1995**, *31*, 431-448.
[41] B. Chen, C. Mueller, P. Willett, *Mol. Inf.* **2010**, *29*, 533-541.
[42] G. V. Cormack, C. L. A. Clarke, S. Buettcher, *Proc. 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval* **2009**, 758-759.
[43] M. Whittle, V. J. Gillet, P. Willett, J. Loesel, *J. Chem. Inf. Model.* **2006**, *46*, 2193-2205.
[44] M. Whittle, V. J. Gillet, P. Willett, J. Loesel, *J. Chem. Inf. Model.* **2006**, *46*, 2206-2219.
[45] Ed.J. D. Holliday, E. Kanoulas, N. Malin, P. Willett, in *J. Cheminf., Vol. 3*, **2011**.
[46] V. J. Gillet, G. M. Downs, A. Ling, M. F. Lynch, P. Venkataram, J. V. Wood, W. Dethlefsen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126-137.
[47] M. Rarey, J. S. Dixon, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471-490.
[48] N. Stiefl, I. A. Watson, K. Baumann, A. Zaliani, *J. Chem. Inf. Model.* **2006**, *46*, 208-220.
[49] V. J. Gillet, P. Willett, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338-345.
[50] E. J. Barker, E. J. Gardiner, V. J. Gillet, P. Kitts, J. Morris, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346-356.
[51] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
[52] G. Harper, G. S. Bravi, S. D. Pickett, J. Hussain, D. V. S. Green, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145-2156.
[53] K. Birchall, V. J. Gillet, G. Harper, S. D. Pickett, *J. Chem. Inf. Model.* **2006**, *46*, 577-586.
[54] Y. Takahashi, M. Sukekawa, S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639-643.
[55] E. J. Gardiner, V. J. Gillet, P. Willett, D. A. Cosgrove, *J. Chem. Inf. Model.* **2007**, *47*, 354-366.
[56] K. Birchall, V. J. Gillet, G. Harper, S. D. Pickett, *J. Chem. Inf. Model.* **2008**, *48*, 1543-1557.
[57] K. Birchall, V. J. Gillet, G. Harper, S. D. Pickett, *J. Chem. Inf. Model.* **2008**, *48*, 1558-1570.
[58] K. Birchall, V. J. Gillet, P. Willett, P. Ducrot, C. Luttmann, *J. Chem. Inf. Model.* **2009**, *49*, 1330-1346.
[59] I. Ujváry, *Pestic. Sci.* **1997**, *51*, 92-95.
[60] J. H. van Drie, *Internet Electron. J. Mol. Des.* **2007**, *6*, 271-279.
[61] O. Güner (Ed.), *Pharmacophore Perception, Development and Use in Drug Design*, International University Line, La Jolla CA, **2000**.
[62] A. R. Leach, V. J. Gillet, R. A. Lewis, R. Taylor, *J. Med. Chem.* **2010**, *53*, 539-558.
[63] A. T. Brint, P. Willett, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152-158.

[64] Y. C. Martin, M. G. Bures, E. A. Danaher, J. Delazzer, I. Lico, P. A. Pavlik, *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83-102.
[65] G. Jones, P. Willett, R. C. Glen, *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532-549.
[66] Y. Patel, V. J. Gillet, G. Bravi, A. R. Leach, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653-681.
[67] S. J. Cottrell, V. J. Gillet, R. Taylor, D. J. Wilton, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 665-682.
[68] S. J. Cottrell, V. J. Gillet, R. Taylor, *J. Comput.-Aided Mol. Des.* **2006**, *20*, 735-749.
[69] R. Taylor, J. C. Cole, D. A. Cosgrove, E. J. Gardiner, V. J. Gillet, O. Korb, *J. Comput.-Aided Mol. Des.* **2012**, *26*, 451-472.
[70] N. J. Richmond, C. Abrams, P. R. N. Wolohan, E. Abrahamian, P. Willett, R. D. Clark, *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567-587.
[71] A.-M. Harrison, D. R. South, P. Willett, P. J. Artymiuk, *J. Comput.-Aided Mol. Des.* **2003**, *17*, 537-549.
[72] R. V. Spriggs, P. J. Artymiuk, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 412-421.
[73] P. J. Artymiuk, R. V. Spriggs, P. Willett, *J. Am. Soc. Inf. Sci. Technol.* **2005**, *56*, 518-528.
[74] N. Nadzirin, E. J. Gardiner, P. Willett, P. J. Artymiuk, M. Firdaus-Raih, *Nucleic Acids Res.* **2012**, *40*, W380-W386.
[75] M. Firdaus-Raih, H. Y. Hamdani, N. Nadzirin, E. I. Ramlan, P. Willett, P. J. Artymiuk, *Nucleic Acids Res.* **2014**, *42*, W382-W388.
[76] G. Papadatos, M. Alkarouri, V. J. Gillet, P. Willett, V. Kadirkamanathan, C. N. Luscombe, G. Bravi, N. J. Richmond, S. D. Pickett, J. Hussain, J. M. Pritchard, A. W. J. Cooper, S. J. F. Macdonald, *J. Chem. Inf. Model.* **2010**, *50*, 1872-1886.
[77] P. Willett, *J. Mol. Graphics Modell.* **2007**, *26*, 602-606.
[78] P. Willett, *J. Inf. Sci.* **2008**, *34*, 477-499.
[79] P. Willett, *Aslib Proceedings* **2008**, *60*, 4-17.
[80] R. Al Jishi, P. Willett, *J. Chem. Inf. Model.* **2010**, *50*, 1915-1923.
[81] M. F. Lynch, J. M. Harrison, W. G. Town, J. E. Ash, *Computer Handling of Chemical Structure Information*, Macdonald, London, **1971**.
[82] A. R. Leach, V. J. Gillet, *An Introduction to Chemoinformatics*, 2nd edition, Kluwer, Dordrecht, **2007**.
[83] D. J. Wild, G. D. Wiggins, *Drug Discovery Today* **2006**, *11*, 436-439.
[84] P. Franco, N. Porta, J. D. Holliday, P. Willett, *J. Cheminf.* **2014**, *6:5*.
[85] R. Sherhod, V. J. Gillet, P. N. Judson, J. D. Vessey, *J. Chem. Inf. Model.* **2012**, *52*, 3074-3087.
[86] R. Sherhod, P. N. Judson, T. Hanser, J. D. Vessey, S. J. Webb, V. J. Gillet, *J. Chem. Inf. Model.* **2014**, *54*, 1864-1879.
[87] H. Patel, M. J. Bodkin, B. Chen, V. J. Gillet, *J. Chem. Inf. Model.* **2009**, *49*, 1163-1184.
[88] D. Hristozov, M. Bodkin, B. Chen, H. Patel, V. J. Gillet, in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design, Vol. 1076* (Ed.: R. J. Bienstock), **2011**, pp. 29-43.