Information Research, Vol. 5 No. 2, January 2000

Textual and chemical information processing: different domains but similar algorithms

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

Abstract

This paper discusses the extent to which algorithms developed for the processing of textual databases are also applicable to the processing of chemical structure databases, and *vice versa*. Applications discussed include: an algorithm for distribution sorting that has been applied to the design of screening systems for rapid chemical substructure searching; the use of measures of inter-molecular structural similarity for the analysis of hypertext graphs; a genetic algorithm for calculating term weights for relevance feedback searching for determining whether a molecule is likely to exhibit biological activity; and the use of data fusion to combine the results of different chemical similarity searches.

Introduction

Over the years, researchers in information retrieval (IR) have developed many different techniques for the processing of databases of textual information. Established examples of such techniques include document clustering, relevance feedback, stemming algorithms and text compression, and there are many other emerging applications, such as categorisation, event detection and information filtering (Spark Jones and Willett, 1997). One such application is multimedia retrieval, where there is much current interest in extending algorithms and data structures developed for processing textual databases, for the storage and retrieval of speech, image and video data (Maybury, 1997). This paper argues that at least some of the algorithms that are used in textual information retrieval can also be applied to another type of data, *viz* the two-dimensional (2D) and three-dimensional (3D) chemical structure data that forms one of the principal components of chemical information systems (Ash *et al*., 1991). These systems were first developed principally for archival purposes, but now play an important role in research programmes to discover novel bioactive molecules for the pharmaceutical and agrochemical industries (Martin and Willett, 1998).

The University of Sheffield has been involved in research on both chemical and textual information retrieval for many years (Lynch and Willett, 1986). These studies have led us to believe that both areas have much to offer, with research in one providing a fertile source of ideas for research in the other. In some cases, the relationship is obvious, with algorithms and data structures being transferable with little or no change from one application to another; while in other cases, the relationship is less direct, involving more a general view of the sorts of information processing techniques that are required, rather than a direct transfer of technology. Here, we consider the former, more direct, type of relationship.

There are clear similarities in the ways that chemical and textual database records are characterised. The documents

in a text database are each typically indexed by some small number of keywords, in just the same way as the 2D or 3D molecular representations in a chemical database are each characterised by some small number of substructural features chosen from a much larger number of potential attributes (as discussed further in the next section of this paper). Moreover, both types of attribute follow a well-marked Zipfian distribution, with the skewed distributions that characterise the frequencies of occurrence of characters, character substrings and words in text databases being mirrored by the comparable distributions for the frequencies of chemical moieties. Thus, the overwhelming majority of all of the many millions of molecules that have ever been made contain the element carbon but even the tenth most frequent element, iodine, occurs only about one thousandth as frequently, with the great majority of the elements having vanishingly small frequencies of occurrence; similar distributions are observed for other types of chemical substructure (Lynch, 1977). These shared characteristics mean that the two types of database are amenable to efficient processing using the same type of file structure. Thus one of the first examples of what would now be referred to as text signature searching (Barton *et al*., 1974) arose from previous studies of chemical bit-string processing (Adamson *et al*., 1973), and similar comments apply to the use of an inverted file for rapid database clustering (Willett, 1981, 1982). Finally, in just the same way as a document either is, or is not, relevant to some particular user query, so a molecule is active, or is not active, in some particular biological test, thus allowing comparable performance measures to be used to assess search effectiveness in the two types of retrieval system (Edgar *et al*., 1999).

These similarities mean that it is often possible to apply similar algorithms to the two different sorts of database, as we describe in some detail below. That said, there are obvious differences, most obviously in the semantics of the representations that are used. A 2D chemical structure diagram bears a much closer relationship to the molecule it describes than does the set of words comprising a textual document, and this relationship is still stronger when, as is increasingly the case, a 3D structure with XYZ atomic co-ordinate data is available for a molecule (Martin and Willett, 1998). Both the structure diagram and the co-ordinates can be regarded as direct manifestations of the underlying wave equations that describe a molecule, and it has thus proved possible to develop powerful simulation techniques to predict the activities and properties of molecules from a knowledge of their 2D or 3D structure. Many of these molecular modelling tools have no direct textual equivalent, as the use of natural language raises a host of linguistic problems that do not arise in the chemical context (although it should be noted that linguistic parsing and recognition techniques can be used to represent and search the generic chemical structures that occur in chemical patents (Barnard *et al*., 1984; Welford *et al*., 1981)).

The remainder of this paper discusses several applications to support the belief that there may be at least some degree of overlap between the techniques used to process chemical and textual databases. The first application, which is taken from a previous paper discussing a potential relationship between these two domains (Willett, 1997), is that of chemical substructure searching; this section also introduces the basic components of chemical information systems, thus providing some of the necessary background for the more recent applications discussed in the following sections.

Design of screening systems for chemical substructure searching

Chemical information systems have historically represented molecules by means of their 2D chemical structure diagrams: these are encoded as labelled graphs in which the nodes and edges of a graph encode the atoms and bonds of a molecule (Ash *et al*., 1991). Searches for substances that contain a specific partial structure, such as a cephalosporin ring, can then be effected by graph-matching techniques that permit the identification of all occurrences of the user. s query substructure in each database structure (Barnard, 1993). The time-consuming nature of these *subgraph isomorphism* searches means that an initial *screening* stage is required that can eliminate the great bulk of the database from subsequent processing (in just the same way as a text signature search is used to reduce the amount of pattern matching that is required in serial searches of text databases). The question then arises as to what sorts of substructural characteristics should be used as indexing keys in the screening search.

The Zipfian distribution of the occurrences of the elements has been mentioned previously, this implying a vast divergence in the discriminatory powers of searches of chemical databases that are based on elemental type. Work in Sheffield by Lynch and his co-workers in the early Seventies (see, *e.g*., Adamson *et al*., 1973) showed that improved discriminatory power could be obtained by the use of more sophisticated indexing keys that were based on variably-sized chemical *fragment substructures*, *i.e*., groups of atoms and bonds. The fragments were chosen so as to occur with approximately equal frequencies of occurrence in the database that was to be searched, an idea that now forms the basis for many current systems for 2D substructure searching (Barnard, 1993).

It was soon realised that the concept of equifrequency was of great generality, and this led to an extended investigation of the application of equifrequency ideas to the searching and processing of text databases (Lynch, 1977): here, we consider the studies that were carried out on the approach to sorting normally refererd to as *distribution sorting* (Cooper *et al.*, 1980; Cooper and Lynch, 1984). The suggested approach involves sorting a very large file in two stages: first, an initial, approximate sorting stage that sub-divides the file into an ordered series of sub-files; and then an exact sort of each of the sub-files. To those of us familiar with the old days when catalogue cards had to be sorted by hand, the initial stage is analogous to dividing the cards into the sub-files A to D, E to K, L to R and S to Z. In the present context, each of these rough pigeonholes represents a separate sub-file on disk and each of the ranges is chosen to ensure that approximately equal numbers of records are allocated to each such sub-file, with the aim of maximising the efficiency of the final, in-core sorts of these sub-files.

The 3D structure of a molecule plays a vital role in determining its biological activity. The "lock-and-key" theory suggests that a molecule may be able to act as a drug if it can fit into the active site of a protein in much the same way as a key fits into a lock (Martin and Willett, 1998), and there is thus great interest in being able to identify molecules that are appropriately "key-shaped". This is done by searching for *pharmacophores*, *i.e.*, the patterns of atoms in 3D space that are thought to be responsible for a molecule binding to an active site in a protein molecule. When our studies of pharmacophore searching started in the mid-Eighties, it was soon realised that the graph-theoretic methods developed for 2D substructure searching were also applicable here, with a 3D molecule being described by a graph in which the nodes were the atoms and the edges were the inter-atomic distances (Willett, 1995). However, the implementation of subgraph isomorphism matching on such 3D chemical graphs is still more demanding of computational resources than is 2D substructure searching, with a consequent need for the development of efficient methods of screening.

The graphs representing 2D chemical molecules are composed of atoms and bonds, and this is reflected in the compositions of the screens that are used for 2D substructure searching. It thus seemed appropriate to investigate screens for 3D searching based on the information contained in 3D chemical graphs, *i.e.*, atoms and inter-atomic distances, and we decided to evaluate screens consisting of a pair of atoms and an inter-atomic distance range. Thus, a screen might represent the presence within a molecule of an oxygen atom and a nitrogen atom separated by a distance range, *e.g.*, between 5 and 7 Ångstroms. Initial experiments demonstrated the vastly skewed distributions of distances that characterise databases of typical 3D molecules and there was thus a need to identify such inter-atomic screens so that they occurred with approximately equal frequencies of occurrence (Jakes and Willett, 1986).

The approach adopted involved varying the width of the distance range associated with each pair of atoms, so that highly populated parts of the distance distribution were associated with very narrow ranges, while less populated parts of this distribution, or very infrequently occurring pairs of atoms, were associated with more extended ranges. The algorithm that was finally developed was a simple modification of the distribution sorting algorithm described above with, in essence, the distance ranges here corresponding to the character ranges that underlie the text-sorting application (Cringean *et al.*, 1990). This proved to be both effective and efficient, and screens based on the distances between pairs of atoms now form the basis for nearly all existing systems for 3D substructure searching; indeed, the same basic methodology can be used to characterise the valence or dihedral angular relationships that exist between sets of three or four atoms, respectively, thus allowing database searches to be carried out that involve the specification of both distance and angular information. More recently, we have demonstrated that analogous procedures can be used to search databases where account is taken of the fact that most 3D molecules are not completely rigid, but are actually in a state of constant flux (so that the "keys" referred to previously might be considered as being more akin to a jelly than to a piece of rigid metal) (Willett, 1995). This further application demonstrates clearly the generality of equifrequency-based approaches for database processing.

Manipulation of hypertext graphs using measures of inter-molecular structural similarity.

Text retrieval systems were initially based upon the Boolean retrieval model, but the systems were soon extended to permit best match searching (in which the documents are ranked in decreasing order of similarity to the query) (Spark Jones and Willett, 1997). A similar progression has occurred in chemical information systems, with the substructure searching systems described above increasingly being complemented with facilities for what is referred to as *similarity searching*. This generally involves the specification of an entire query molecule, the *target structure*, rather than the partial structure that is required for substructure searching. The target is characterised by one or more

structural descriptors that are compared with the corresponding sets of descriptors for each of the molecules in the database to find those *nearest neighbour* molecules that are most similar to the target structure. Two near-contemporaneous studies in the mid-Eighties demonstrated that counts of the numbers of fragment substructures common to a pair of molecules provided a computationally efficient, and surprisingly effective, basis for quantifying the degree of structural resemblance between the two molecules under consideration (Carhart *et al*, 1985; Willett *et al*., 1986). Specifically, the use of a simple association coefficient (usually the Tanimoto coefficient) in conjunction with the lists of screens associated with the target structure and each of the database structures provided a simple way of investigating inter-molecular structural similarities. Such fragment-based methods for similarity searching are now widely used as a complement to the established routines for substructure searching (Willett *et al*., 1998) and we have since demonstrated that these measures are also applicable to the comparison of textual, rather than chemical, graphs. Specifically, as described below, we have used these measures to determine the degree of consistency with which hypertext documents are created; other relationships between textual and chemical similarity measures are discussed by Willett (1997).

The creation of the intra-document links between the individual components of a hypertext document is a difficult, and time-consuming, task, but one in which human intervention has commonly been thought necessary if the semantic relationships that exist between the components of the document are to be made explicit. A similar view has prevailed for many years with regard to the indexing of documents in IR systems, where the existence of well-established systems for automatic indexing has not prevented the widespread use of trained library and information specialists for indexing and classifying documents prior to their incorporation in an online database. The importance of the indexing task in IR has led to many studies of *inter-indexer consistency*, *i.e*., of the extent to which agreement exists among different indexers on the sets of index terms to be assigned to individual documents. These studies (as reviewed, *e.g*., by Markey (1984)) have consistently concluded that recorded levels of consistency vary markedly, and that high levels of consistency are rarely achieved; similar examples of manual inconsistency are provided by related tasks, such as query formulation and the assessment of document relevance (Salton, 1989). The insertion of links in hypertext documents may be viewed as being analogous to the assignment of index terms to such documents, and we hence undertook a study to determine the extent to which different people produce similar link structures for the same hypertext documents (Ellis *et al*., 1994, 1996).

The hypertext documents were generated from five printed full-text documents, each a thesis, journal article or book written by a member of the Department of Information Studies at the University of Sheffield. Five copies were made of each of the chosen documents, and each of the twenty-five resulting copies was allocated to a different student volunteer from the Department. The volunteers were instructed in the use of an interactive system that allowed them to create explicit representations of links between paragraphs whose contents they decided were related, and thus to create hypertext versions of the source documents. These hypertexts were stored as graphs in which the nodes represented portions of a text (specifically paragraphs in our work but section-based or sentence-based portions could also have been used), with an edge linking a pair of nodes if the human linker had created a link between the corresponding paragraphs. Pairs of these graphs (describing the same textual document but processed by different volunteers) were then compared using a range of similarity measures, most of which were based on those used for chemical similarity searching. For example, a commonly used type of chemical fragment is the *augmented atom*, which consists of an atom and the atoms that are bonded to it, and it is possible to generate a comparable hypertext fragment consisting of a paragraph and those paragraphs that are linked to it. A measure of the similarity between two hypertext graphs can then be calculated in terms of the number of such fragments that the hypertexts have in common, and the resulting similarity is taken as a measure of the degree of inter-linker consistency.

Five hypertext versions were produced for each source document, giving a total of ten possible pairs of sets of hypertext links for that document (*i.e*., fifty possible pairs for the entire dataset). Similarity values were calculated using many different combinations of graph representation, fragment descriptor and similarity coefficient. When this was done, it was possible to draw a simple, unequivocal conclusion: that levels of inter-linker consistency are generally low, but can also be quite variable. It would hence seem that different people tend to have very different views of the semantic relationships that exist among the components of a full-text document, with the possibility that different people will tend to impose different hypertext link structures on the same source documents. This lack of agreement leads one to question the effectiveness of manually-assigned links in supporting the browsing strategies of subsequent hypertext readers, since if inter-linker consistency is found to be low, then linker/reader consistency (*i.e*., the level of agreement as to the semantic relationships that exist between the components of the text, and that therefore should explicitly be represented by links that the reader then has the opportunity to follow) can also be expected to be low. Indeed, our results provide some justification for the numerous and continuing

efforts to develop techniques for the automatic generation of hypertext links (see, *e.g.*, Crestani and Melucci, 1998) where the set of links created by a particular technique will be produced in a consistent manner, to which users might become accustomed as they browse and which might well be no worse than those created by human indxers.

A genetic algorithm for calculating relevance feedback and substructural analysis weights.

The calculation of weights that describe the importance of document and query terms is an important component of best-match text IR systems. The most effective results are obtained from using relevance feedback information (Robertson and Sparck Jones, 1976; Salton and Buckley, 1990) but efforts continue to develop new weighting schemes that might further increase search effectiveness. This striving for improved performance has raised the question of what is the *upperbound*, *i.e.*, the maximum possible level, that can be achieved by a particular retrieval strategy. We have hence developed a genetic algorithm (hereafter a GA) to investigate the upperbound that can be achieved by ranked-output retrieval systems that employ weighted query terms. Indeed, the GA was developed not only for this purpose but also for calculating weights reflecting the extent to which particular types of molecular feature contribute to the biological activity of molecules (using the analogies between indexing terms and structural features, and between query relevance and biological activity that we have noted in the introduction to this paper). We will illustrate the operation of this GA by discussing the textual application (Robertson and Willett, 1996) before proceeding to an account of the results obtained in our chemical experiments (Gillet *et al*., 1998).

A GA operates on *chromosomes*, which are linearly-encoded representations of possible solutions to the problem under investigation, with each element of a chromosome describing some particular component of the encoded solution. A *fitness function* is used to calculate a numeric score that measures how "good" a solution is represented by each chromosome. The set of chromosomes at any particular stage of the processing is referred to as the current *population*, and the chromosomes in this generation are processed by *genetic operators* to yield the population that corresponds to the subsequent generation. The two principal operators are *crossover*, which combines parts of the fittest chromosomes, and *mutation*, which introduces new information at random. The chromosomes in the new generation are evaluated by means of the fitness function and the genetic operators invoked again, this process continuing until convergence is achieved, *i.e*, until there is no increase in the average fitness of each successive generation.

Each chromosome in our text GA contains a set of elements, with one element for each of the terms comprising the query and with each such element containing a weight for one of these terms. The weights for those query terms that occur in a particular document are summed, the documents are ranked in decreasing order of these sums of weights, and a cut-off is then applied to the ranking to retrieve some fixed number of the top-ranked documents. Our experiments have used standard document test collections for which full relevance data is available, and it is hence possible to determine the recall of the search by noting the number of relevant documents that have been retrieved above the cut-off. These recall values are used as the fitness function for the GA. Once termination has occurred, *i.e*., the recall values have ceased to increase, the term weights are noted for the chromosome that has the largest fitness. Thus, if the GA has successfully explored the full range of possible weights, then the final weights provide an upperbound to the retrieval performance obtainable for that query using relevance weighting of single terms.

The experiments involved the following seven document test collections: Keen (800 documents and 63 queries), Cranfield (1400 documents and 225 queries), Harding (2472 documents and 65 queries), Evans (2542 documents and 39 queries), LISA (6004 documents and 35 queries), SMART (12694 documents and 77 queries) and UKCIS (27361 documents and 182 queries). The GA was run for each query in each collection, and the final set of weights used to calculate recall values for the top-10, top-20 and top-50 documents. The F4 retrospective relevance weights of Robertson and Spark Jones (1976) were used for comparison with the GA weights since the former have been found to provide a consistently high level of performance in previous studies of relevance feedback searching.

Table 1 (below) summarises the results obtained when the top-20 documents were retrieved from a ranking; many other results are presented and discussed by Robertson and Willett (1996). An inspection of this table shows that (as would be expected) the GA weights generally perform better than the F4 weights, in terms of both average recall and numbers of queries where one was superior to the other. However, there are generally many searches where the two approaches give the same level of performance (to two decimal places in mean recall), and often at least a few queries (and a fair number in the case of the Harding collection) where the F4 search is better. Where there are differences, in either direction, between the two weights, these are overwhelmingly due to one of the weights

finding a single additional relevant document above the cut-off position in the ranking: thus, the two approaches give very similar levels of performance. A study of those few queries where the F4 weights were (unexpectedly) superior suggested that these queries were ones in which at least some of the F4 weights were negative (this corresponding to terms that occur frequently in a collection but only infrequently in the relevant documents). The initial version of the GA did not allow for the possibility of negative weights: when it was modified to allow for such occurrences, there was a substantial reduction in the (already small) number of queries where the F4 weights were more effective (Robertson and Willett, 1996). In general then, the F4 weights gave a level of performance that was only marginally inferior to those provided by the GA weights, which had been obtained by a detailed exploration of the term-weight space defined by the query terms. This being so, it seems reasonable to conclude that the F4 weights give a practicable upperbound to the performance that is achievable in relevance feedback searches of text databases.

Document	**Mean Recall**		**Number Of Queries**		
Collection	**F4**	**GA**	**Same**	**F4 Better**	**GA Better**
Keen	0.55	0.60	31	2	30
Cranfield	0.58	0.65	117	6	102
Harding	0.34	0.33	33	16	16
Evans	0.38	0.42	11	7	21
LISA	0.61	0.64	13	5	17
SMART	0.30	0.34	23	10	44
UKCIS	0.18	0.21	105	7	70

Table 1. Retrieval effectiveness in relevance feedback searches using F4 and GA weights.
[The Mean Recall portion of the table is the mean recall when averaged over all of the queries for a particular test collection using the two types of weight, while the right-hand portion shows the number of queries where the GA and F4 weights gave the same recall value (Same), and where either the F4 weights or the GA weights were better.]

The GA that was used for these relevance feedback experiments was also designed to calculate weights for *substructural analysis*, (Cramer *et al*., 1974). Here, weights are calculated that relate the presence of a molecular feature to the probability that that molecule is active in some biological test system (*cf* relating the presence of a specific index term in a document to the probability that that document is relevant to a particular query). Given some training set of compounds for which the biological activities are available, the aim of substructural analysis is to develop weights that can then be used to select new compounds for biological testing. Specifically the sum of weights is calculated for the fragment substructures present in a molecule, and then the compounds are ranked in order of decreasing sums-of-weights, so that chemical synthesis and biological assays can be focused on those compounds that have a high *a priori* probability of activity.

The use of substructural analysis methods with fragment substructures (such as those discussed in the second section of this paper) is well established, with many different types of weighting scheme having been described in the literature (Ormerod *et al*., 1989). The project to be described here used a rather different level of molecular description, specifically high-level molecular characteristics suggested by medicinal chemists at GlaxoWellcome Research and Development (our collaborators in this project) as affecting the drug-like behaviour of molecules. The features are the distribution of property values for the molecular weight, the ^2k a shape index (Kier, 1987), and the numbers of aromatic rings, rotatable bonds, hydrogen-bond donor atoms and hydrogen-bond acceptor atoms, in the molecules comprising a database. Each property was allocated a total of 20 bins; for example, the first bin for the hydrogen-bond donor feature describes those molecules in a dataset that have no donor atoms, the second bin those that have one donor atom, and so in until the 20th and last, which represents those that have 19 or more donor atoms. The molecular weight and shape index bins contained ranges of values, rather than specific integer counts,

e.g., the first two bins for molecular weight described the molecules with weights in the range 0-74.99 and 75-149.99, *etc*.

The GA described previously was used to calculate weights for each bin of each feature, with each such weight reflected the extent to which possession of that feature-value combination resulted in a molecule having some specific activity. The fitness function was then based on the occurrences of these feature-value pairs in sets of molecules for which activity data are available. For this purpose, we used sets of structures extracted from the *World Drugs Index* (or WDI, available from Derwent Information at URL http://www.derwent.co.uk) and *SPRESI* (available from Daylight Chemical Information Systems Inc. at http://www.daylight.com) databases: the former contains molecules for which biological activities have been established while the latter contains a large number of molecules that are assumed to be inactive. As with the text application, the chromosome in the GA encodes possible weights (in this case for each of the feature-value bins), and then the score for a particular molecule is the sum of the weights for those feature values that are associated with it. The fitness function of the GA is simply the number of top-ranked active molecules once the molecules have been ranked in decreasing order of these sums-of-weights.

Sets of 1000 molecules with some particular activity were chosen from the WDI and then combined with a set of 16,807 SPRESI structures. The GA was used to calculate weights for this combined dataset, and a note made of the number of actives in the top 1000 positions once convergence had occurred. The effectiveness of the weights is measured by the degree of *initial enhancement, i.e.*, the ratio of the observed number of top-1000 actives to the number of actives that would be obtained by selecting 1000 compounds at random (Gillet *et al*., 1998). These results are summarised in Table 2 and it will be seen that the GA weights give consistently better results than merely selecting compounds at random for testing, although the extent of the improvement is clearly dependent upon the specific activity under investigation. Gillet *et al*. (1998) report comparable results from extended experiments using various forms of these weights on a range of datasets; these involved not only retrospective studies (as here) but also experiments where the weights were used in a predictive manner. It was concluded that the GA provided a simple and effective way of ranking sets of compounds in order of decreasing probability of activity, thus enabling the prioritisation of compounds for synthesis and biological testing.

Drug Activity	Initial Enhancement
Hormones	8.3
Anti-cancers	7.6
Anti-microbials	7.6
Anaesthetics	4.1
Blood	3.6
Central nervous system	3.9
Psychotropics	2.5

Table 2. Effectiveness of ranking active compounds using the GA-based substructural analysis weights.
[The Drug Activity listed is that contained in the keyword activity field for each of the compounds in the WDI database.]

Combination of chemical similarity measures using data fusion

The many types of similarity measures that are available for the measurement of molecular similarity has led to comparative studies in which researchers try to identify a single, "best" measure, using some quantitative performance criterion. For example, Willett *et al*. (1986) discuss the merits of different association coefficient for

chemical similarity searching and conclude by recommending the Tanimoto coefficient for this purpose. Such comparisons, of which there are many in the chemical information literature, are limited in that they assume, normally implicitly, that there is some specific type of structural feature (similarity coefficient, weighting scheme or whatever it is that is being investigated) that is uniquely well suited to describing the type(s) of biological activity that are being sought for in a similarity search. The assumption cannot be expected to be generally valid, given the multi-faceted nature of biological activities, and this has led us to consider chemical applications of *data fusion* (Hall, 1992).

Data fusion was developed to combine inputs from different sensors, with the expectation that using multiple information sources enables more effective decisions to be made than if just a single sensor was to be employed. The methods are used in a wide range of military, surveillance, medical and production engineering applications (see, *e.g.*, Arabnia and Zhu (1998)): our interest was aroused by a paper by Belkin *et al*. (1995), in which data fusion was used to combine the results of different searches of a text database, conducted in response to a single query but employing different indexing and searching strategies. A query was processed using different strategies, each of which was used to ranking the database in order of decreasing similarity with the query. The ranks for each of the documents were then combined using one of several different fusion rules, the output of the fusion rule was taken as the document. s new similarity score and the fused lists were then re-ranked in descending order of similarity.

Ginn *et al*. (1999) have described the application of these ideas in the context of matching a target structure against a database, using several different measures of chemical similarity as summarised in Figure 1.

1. Execute a similarity search of a chemical database for some particular target structure using two, or more, different measures of inter-molecular similarity.
2. Note the rank position, r_i, of each database structure in the ranking resulting from use of the *i*-th similarity measure.
3. Combine the various rankings using a fusion rule to give a new combined score for each database structure
4. Rank the resulting combined scores, and then use this ranking to calculate a quantitative measure of the effectiveness of the search for the chosen target structure.

Figure 1. Combination of similarity rankings using data fusion

The fusion rules used were those identified by Belkin *et al*. (1995) and shown in Figure 2, and the combined scores output by the fusion rule are then used to re-order the database structures to give the final ranked output. It will be seen from Figure 2 that the MIN and MAX 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 "poor" similarity measure amongst those that are being combined. The SUM rule, where each database structure is assigned the sum of all the rank positions at which it occurs in the input lists, is expected to be more stable against the presence of a single poor or noisy input ranking, and this was generally found to be the rule of choice in our experiments (Ginn *et al*., 1999).

Name	Fusion Rule
MIN	minimum (r_1, r_2 ,&r_i &r_n)
MAX	maximum (r_1, r_2,&r_i &r_n)
SUM	$\sum_{i=1}^{n} r_i$

Figure 2: Fusion rules for combining *n* ranked lists, where r_i denotes the rank position of a specific database structure in the *i*-th ($1 \leq i \leq n$) ranked list.

Our experiments involved combining searches of several different datasets using several different similarity measures in each case. The dataset considered here contained 75 compounds selected by Kahn (1998) in a

discussion of various types of structural descriptor, with each of the compounds belonging to one of 14 well-defined biological activity classes (such as angiotensin-converting enzyme inhibitors and HIV-1 protease inhibitors). Six different types of similarity measures were used in the experiments, as detailed by Ginn *et al*. (1999); for the present, we note merely that they encoded information about the steric, electrostatic and hydrophobic characteristics of molecules (similarity measures denoted by the symbols "F" and "J"), about the 3D arrangement of pharmacophore points in molecules (denoted by the symbols "3" or "T") and about the occurrences of chains of up to 7 non-hydrogen atoms (denoted by the symbols "2" or "N"). The SUM rule was used to generate all possible combinations of rankings from these six similarity measures. Each of the members of the dataset was used as the target structure for a similarity search and Table 3 details the mean numbers of actives (*i.e*., molecules with the same activity as the target structure) found in the top-10 nearest neighbours when averaged over all 75 searches. The values of *c* at the top of the table denote the number of similarity measures that were fused (so that, *e.g*., *c*=1 represents the original measures and *c*=2 represents the fusion of a pair of the original measures) and a shaded element indicates a fused combination that is better than the best original individual measure (which was one of the 3D pharmacophore measures).

***c*=1**		***c*=2**		***c*=3**		***c*=4**		***c*=5**		***c*=6**	
2>	0.80	23	1.10	23F	1.28	23FJ	1.52	23FJN	1.45	23FJNT	1.43
3	1.12	2F	1.04	23J	1.39	23FN	1.23	23FJT	1.69		
F	0.89	2J	1.01	23N	1.04	23FT	1.43	23FNT	1.36		
J	1.08	2N	0.68	23T	1.24	23JN	1.31	23JNT	1.43		
N	0.63	2T	0.95	2FJ	1.35	23JT	1.45	2FJNT	1.43		
T	0.69	3F	1.09	2FN	1.08	23NT	1.25	3FJNT	1.51		
		3J	1.25	2FT	1.28	2FJN	1.28				
		3N	1.00	2JN	1.03	2FJT	1.53				
		3T	1.32	2JT	1.10	2FNT	1.28				
		FJ	1.20	2NT	0.95	2JNT	1.17				
		FN	0.91	3FJ	1.40	3FJN	1.35				
		FT	1.11	3FN	1.19	3FJT	1.55				
		JN	0.89	3FT	1.33	3FNT	1.41				
		JT	0.93	3JN	1.25	3JNT	1.36				
		NT	0.85	3JT	1.45	FJNT	1.32				
				3NT	1.20						
				FJN	1.11						
				FJT	1.21						
				FNT	1.11						
				JNT	1.12						

Table 3. Mean number of actives found in the top-10 positions for chemical similarity searches when combining

various numbers, *c*, of different similarity measures.
[The pale-green shading indicates a fused result at least as good as the best original similarity measure.]

It will be seen that very many of the fused combinations in Figure 3 are shaded, thus supporting the idea that improvements in effectiveness can be achieved by using more than just a single similarity measure. The table also shows that the fraction of the combinations that are shaded increases in line with *c*, so that all combinations with c^3 4 perform at least as well as the best of the individual similarity measures. That said, the best result overall was obtained with 23FJT (rather than with 23FJNT, the combination involving all of the individual measures): thus, while simply fusing as many individual measures as are available appears to work very well, superior results may be obtained (for this dataset at least) from fusing a subset of the individual measures.

Other datasets studied by Ginn *et al*. (1999) were: 8178 molecules from the Starlist file (available from BioByte Corp. at URL http://clogp.pomona.edu) for which experimental values of the octanol/water partition coefficient, an important parameter in statistical methods for the prediction of biological activity, were available; several sets of 3,500 molecules from the WDI database for which biological activity data were available; and 136 biological dyes used to stain cells so as to visualise various organelles. In all cases, it was found that use of a fusion rule such as SUM generally resulted in a level of performance (however quantified) that was at least as good as (and often noticeably better than) the best individual measure. The best individual measure often varies from one target structure to another in an unpredictable manner, and the use of a fusion rule will thus generally provide a more consistent level of searching performance than will a single measure of chemical similarity. With the increasing number of such measures available that can be implemented with a reasonable degree of efficiency (Willett *et al*., 1998), it would seem appropriate to consider the use of more than one of them for similarity searching of chemical structure databases.

Conclusions

This paper has presented several examples of algorithms and data structures that are applicable to the processing of both textual and chemical databases. These examples suggest that each area has much to offer to the other, with the transfer of methodology occurring from chemical to textual applications, and *vice versa*.

The substructure searching methods discussed in the second section provide an excellent example of this transfer of ideas. Here, an approach that was first developed for chemical retrieval (specifically in the context of 2D substructure searching) was soon shown to be applicable to many applications in the general area of textual databases. One such application, distribution sorting, resulted in an algorithm that was then adapted for use in a rather different area of chemical retrieval (3D substructure searching). In fact, although not discussed here, the idea of using distance and angular information for database searching has now been extended to the computationally demanding task of searching for patterns in the 3D structures of proteins, where our search methods have led to the discovery of many previously unknown structural relationships between proteins (see, *e.g*., Artymiuk *et al*., 1996, 1997). In the case of the hypertext-comparison project, the similarity measures discussed above were originally developed for calculating the similarities between pairs of 2D molecules, while the GA for the calculation of weights was designed from the start for use in both relevance feedback and substructural analysis. Finally, our use of data fusion for combining chemical similarity measures was a simple application of work done by others in the textual domain. Given the range of applications we have already been able to identify, we trust that other researchers will be encouraged to investigate the many similarities that exist between chemical and textual database processing.

Acknowledgements. I thank all of my colleagues, past and present, for their contributions to the research that has been summarised here, and the many organisations that have provided financial support for my research over the years. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

- Adamson, G.W., Cowell, J., Lynch, M.F., McLure, A.H.W., Town, W.G. and Yapp, A.M. (1973) "Strategic considerations in the design of screening systems for substructure searches of chemical structure files." *Journal of Chemical Documentation*, **13**, 153-157.

- Arabnia, H.R. and Zhu, D., editors (1998). *Proceedings of the International Conference on Multisource-Multisensor Information Fusion, Fusion. 98*. CSREA Press.
- Artymiuk, P.J., Poirrette, A.R., Rice, D.W. and Willett, P. (1996) "Biotin carboxylase comes into the fold." *Nature Structural Biology*, **3**, 128-132.
- Artymiuk, P.J., Poirrette, A.R., Rice, D.W. and Willett, P. (1997) "A polymerase I palm in adenylyl cyclase?" *Nature*, **388**, 1997, 33-34.
- Ash, J.E., Warr, W.A. and Willett, P., editors (1991) *Chemical Structure Systems*. Chichester: Ellis Horwood.
- Barnard, J.M. (1993) "Substructure searching methods: old and new." *Journal of Chemical Information and Computer Sciences*, **33**, 532 - 538.
- Barnard, J.M., Lynch, M.F. and Welford, S.M. (1984) "Computer storage and retrieval of generic chemical structures in patents. Part 6. An interpreter program for the generic structure language GENSAL." *Journal of Chemical Information and Computer Sciences*, **24**, 66-70.
- Barton, I.J., Creasey, S.E., Lynch, M.F. and Snell, M.J. (1974) "An information-theoretic approach to text-searching in direct-access systems." *Communications of the Association for Computing Machinery*, **17**, 345-350.
- Belkin, N.J., Kantor, P., Fox, E.A. and Shaw, J.B. (1995) "Combining the evidence of multiple query representations for information retrieval." *Information Processing and Management*, **31**, 431-448.
- Carhart, R.E., Smith, D.H., Venkataraghavan, R. (1985) "Atom pairs as molecular features in structure-activity studies: definition and applications." *Journal of Chemical Information and Computer Sciences*, *25*, 64-73.
- Cooper, D., Dicker, M.E. and Lynch, M.F. (1980) "Sorting of textual databases: a variety generation approach to distribution sorting." *Information Processing and Management*, **16**, 49-56.
- Cooper, D. and Lynch, M.F. (1984) "The use of binary search trees in external distribution sorting." *Information Processing and Management*, **20**, 547-557.
- Cramer, R.D., Redl, G. and Berkoff, C.E. (1974) "Substructural analysis. A novel approach to the problem of drug design." *Journal of Medicinal Chemistry*, **17**, 533-535.
- Crestani, F. and Melucci, M. (1998) "A case study of automatic authoring: From a textbook to a hyper-textbook." *Data and Knowledge Engineering*, **27**, 1-30.
- Cringean, J.K., Pepperrell, C.A., Poirrette, A.R. and Willett, P. (1990) "Selection of screens for three-dimensional substructure searching." *Tetrahedron Computer Methodology*, **3**, 37-46.
- Edgar, S.J., Holliday, J.D. and Willett, P. (1999) "Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures." In preparation.
- Ellis, D., Furner-Hines, J. and Willett, P. (1994) "On the creation of hypertext links in full-text documents: Measurement of inter-linker consistency." *Journal of Documentation*, **50**, 67-98.
- Ellis, D., Furner-Hines, J. and Willett, P. (1996) "On the creation of hypertext links in full-text documents: Measurement of retrieval effectiveness." *Journal of the American Society for Information Science*, **47**, 287-300.
- Gillet, V.J., Willett, P. & Bradshaw, J. (1998) "Identification of biological activity profiles using substructural analysis and genetic algorithms." *Journal of Chemical Information and Computer Sciences*, **38**, 165-179.
- Ginn, C.M.R., Willett, P. and Bradshaw, J. (1999) "Combination of molecular similarity measures using data fusion." *Perspectives in Drug Discovery and Design*, submitted for publication.
- Hall, D.L. (1992) *Mathematical Techniques in Multisensor Data Fusion*. Northwood, MA: Artech House.
- Jakes, S.E. and Willett, P. (1986) "Pharmacophoric pattern matching in files of 3-D chemical structures: selection of inter-atomic distance screens." *Journal of Molecular Graphics*, **4**, 12-20.
- Kahn, S.D. (1998) "Combinatorial libraries: structure activity analysis, in: Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F. and Schreiner, P.R., editors, *Encyclopedia of Computational Chemistry*. Chichester: John Wiley. Vol 1, pp. 417-425.
- Kier, L.B. (1987) "Indexes of molecular shape from chemical graphs." *Medicinal Research Reviews*, **7**, 417-440.
- Lynch, M.F. (1977) "Variety generation - a re-interpretation of Shannon. s mathematical theory of communication and its implications for information science." *Journal of the American Society for Information Science*, **28**, 19-25.
- Lynch, M.F. and Willett, P. (1987) "Information retrieval research in the Department of Information Studies, University of Sheffield." *Journal of Information Science*, **13**, 221-234.
- Markey, K. (1984) "Inter-indexer consistency tests: A literature review and report of a test of consistency in indexing visual materials." *Library and Information Science Research*, **6**, 155-177.
- Martin. Y.C. and Willett, P., editors (1998) *Designing Bioactive Molecules: Three-Dimensional Techniques*

and Applications. Washington: American Chemical Society.
- Maybury, M.T., editor (1997) *Intelligent Multimedia Information Retrieval*. Cambridge MA: MIT Press.
- Ormerod, A., Willett, P. and Bawden, D. (1989) "Comparison of fragment weighting schemes for substructural analysis." *Quantitative Structure-Activity Relationships*, **8**, 115-129.
- Robertson, A.M. and Willett, P. (1996) "An upperbound to the performance of ranked-output searching: optimal weighting of query terms using a genetic algorithm." *Journal of Documentation*, **52**, 1996, 405-420.
- Robertson, S.E. and Sparck Jones, K. (1976) "Relevance weighting of search terms." *Journal of the American Society for Information Science*, **27**, 129-146.
- Salton, G. (1989) *Automatic Text Processing*. Reading, MA: Addison-Wesley.
- Salton, G. and Buckley, C. (1990) "Improving retrieval performance by relevance feedback." *Journal of the American Society for Information Science*, **41**, 288-297.
- Sparck Jones, K. and Willett, P., editors (1997) *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann.
- Welford, S.M., Lynch, M.F. and Barnard, J.M. (1981) "Computer storage and retrieval of generic chemical structures in patents. Part 3. Chemical structure grammars and their role in the manipulation of chemical structures." *Journal of Chemical Information and Computer Sciences*, **21**, 161-168.
- Willett, P. (1981) "A fast procedure for the calculation of similarity coefficients in automatic classification." *Information Processing and Management*, **17**, 53-60.
- Willett, P. (1982) "The calculation of inter-molecular similarity coefficients using an inverted file algorithm." *Analytica Chimica Acta*, **138**, 339-342.
- Willett, P. (1995) "Searching for pharmacophoric patterns in databases of three-dimensional chemical structures." *Journal of Molecular Recognition*, **8**, 290-303.
- Willett, P. (1997) "Information retrieval research in the University of Sheffield", *ACM SIGIR Forum*, **31**(2), 7-13.
- Willett, P., Barnard, J.M. and Downs, G.M. (1998) "Chemical similarity searching." *Journal of Chemical Information and Computer Sciences*, **38**, 983-996.
- Willett, P., Winterman, V., Bawden, D. (1986) "Implementation of nearest neighbour searching in an online chemical structure search system **26**, 36-41.

How to cite this paper:

Willett, Peter (2000) "Textual and chemical information processing: different domains but similar algorithms" *Information Research*, **5**(2) Available at: http://informationr.net/ir/5-2/paper69.html

---

Articles citing this paper, according to Google Scholar

Contents **Home**