Published paper

Holliday, J.D. and Willett, P. (2011) Representation and searching of chemical-structure information in patents. In: Current Challenges in Patent Information Retrieval. Springer , Berlin , pp. 345-357. ISBN 9783642192302

Representation and Searching of Chemical Structure Information in Patents

John D. Holliday and Peter Willett

Information School, University of Sheffield,
211 Portobello Street, Sheffield S1 4DP, UK

Abstract

This chapter describes the techniques that are used to represent and to search for molecular structures in chemical patents. There are two types of structures: specific structures that describe individual molecules; and generic structures that describe sets of structurally related molecules. Methods for representing and searching specific structures have been well established for many years, and the techniques are also applicable, albeit with substantial modification, to the processing of generic structures.

KEYWORDS.

Chemical patent, chemoinformatics, connection table, fingerprint, generic structure, linear notation, Markush structure, similarity searching, structure searching, substructure searching

X.1 Introduction

Patents are a key information resource for all types of industry, but this is particularly the case in the pharmaceutical and agrochemical industries. The main focus of these industries is to identify novel chemical molecules that exhibit useful biological activities, e.g., reducing an individual's cholesterol level or killing the insect pest of a crop (Barnard 1984; Berks 2001). Chemical patents hence need to contain not just the textual information that one would find in any type of patent, but also information about the chemical molecules of interest. These can, of course, be described by their chemical names or images, but these provide only limited searching facilities that are not sufficient to meet the requirements of modern industrial research and development. Instead, specialised types of representation and search algorithm have had to be developed to provide efficient and effective access to the structural information contained in patents. These techniques are an important component of what has come to be called *chemoinformatics* (Willett 2008), i.e., "the application of informatics methods to solve chemical problems" (Gasteiger 2006).

Two types of molecular information are encountered in chemical patents. A patent may be based on just a single specific molecule, in which case the techniques that have been developed in chemoinformatics over many years may be applied, as discussed below. However, the majority of chemical patents discuss not single molecules, but entire classes of structurally related molecules, with these classes being described by a *generic*, or *Markush*, structure. A single generic structure can represent many thousands, or even a potentially infinite number, of individual molecules, and the representational and searching techniques required are accordingly far more complex than those commonly encountered in chemoinformatics systems. In this paper, we provide an overview of the techniques that are used to handle both specific and generic chemical structures. The reader is referred to the standard texts by Leach and Gillet (2007) and by Gasteiger and Engel (2003) for further details of the techniques described below; these books also provide excellent introductions to the many aspects of chemoinformatics that are not, as yet, of direct relevance to the processing of chemical patent information.

X.2 Searching specific chemical structures

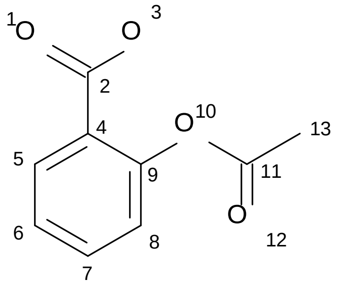
X.2.1 Representation of chemical structures

If one wishes to carry out computer-based searches of a chemical database then the molecules of interest must be encoded for searching, and we commence by describing the three main ways in which one can provide a full description of a chemical structure in machine-readable form: these are *systematic nomenclature*, *linear notations*, and *connection tables*. Before describing these, the reader should note that we consider here (and in the remainder of this chapter) only the processing of 2D chemical molecules, i.e., the planar chemical structure diagrams that are conventionally used to represent molecules in the scientific literature and that are exemplified by the structure diagram shown in Figure 1. More sophisticated techniques are required for the representation and searching of 3D chemical molecules, i.e., where one has geometric coordinate information for all of a molecule’s constituent atoms (Martin and Willett 1998).

Chemical compounds have had names associated with them ever since the days of the alchemists, but it was many years before it was realised that there was a need for systematic naming conventions to ensure that every specific molecule would have its own unique name. This name should be unique, in the sense that there should be only one possible name for a molecule, and unambiguous, in the sense that it should describe that molecule and no other; moreover, it was soon realised that the name should describe the various substructural components comprising the molecule, whereas common, non-systematic names will normally say little or nothing about a molecule's components. For example, 2-acetoxybenzoic acid is the systematic, explicit representation for the structure shown in Figure 1, which is also, and most commonly, called aspirin.

Two systematic nomenclatures are in widespread use, these being the ones developed by the International Union of Pure and Applied Chemistry (IUPAC at http://www.iupac.org) and by Chemical Abstracts Service (CAS at http://www.cas.org). IUPAC is an association of 60 national chemical societies, seeking to establish standards in nomenclature and physiochemical data measurement, while CAS is a division of the American Chemical Society and the world's largest provider of chemical information, indexing articles from more than 10,000 journals and patents from 60 national patent agencies. Systematic names continue to be widely used in the chemical literature, but are of less importance in chemoinformatics systems since they are normally converted automatically into one of the two other types of standard representation, i.e., linear notations or connection tables. A linear notation is a string of alphanumeric characters that provides a complete, albeit in some cases implicit, description of the molecule's topology. A *canonicalisation* procedure is normally invoked to ensure that there is a unique notation for each molecule. The first notation to be widely used was the Wiswesser Line Notation, which formed the basis for most industrial chemoinformatics systems in the Sixties and Seventies. Two notations are of importance in present-day systems: the SMILES (for Simplified Molecular Input Line Entry Specification) notation developed by Daylight Chemical Information Systems Inc. (Weininger 1988) and the International Chemical Identifier (or InChI), the development of which is being overseen by IUPAC. SMILES was developed for use in in-house industrial chemoinformatics systems (as is the case with much chemoinformatics software) while InChI, conversely, has been developed as an open-source, non-proprietary notation. The SMILES and the InChI for aspirin are included in Figure 1.

Notations provide a compact molecular representation, and are thus widely used for compound exchange and archival purposes. However, most chemoinformatics applications will require their conversion to a connection table representation of molecular structure. A connection table is a data structure that lists the atoms within a molecule and the bonds that link those atoms together (in many cases, only heavy atoms are included since the presence of hydrogen atoms can be deduced automatically). The table provides a complete and explicit description of a molecule’s topology, i.e., the way that it is connected together, whereas this information is normally only implicit in a linear notation. There are many ways in which the atoms and bonds can be encoded, with typical connection table formats being exemplified by those developed by MDL Information Systems Inc. (now Accelrys Inc.) (Dalby et al. 1992). A sample connection table for aspirin is shown in Figure 1 where, for example, the first line shows that atom number 1 (Oxygen) is connected by a double bond (D) to atom number 2.



InChI: 1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

Smiles: CC(=O)Oc1ccccc1C(=O)O

Name: 2-acetoxybenzoic acid

Connection Table:

```
1 O D 2
2 C D 1 S 3 S 4
3 O S 2
4 C S 2 D 5 S 9
5 C D 4 S 6
6 C S 5 D 7
7 C D 6 S 8
8 C S 7 D 9
9 C S 4 D 8 S 10
10 O S 9 S 11
11 C S 10 D 12 S 13
12 O D 11
13 C S 11
```

Fig. 1: Structure, name, InChI, SMILES and connection table for aspirin

A connection table is an example of a *graph*, a mathematical construct that describes a set of objects, called *nodes* or *vertices*, and the relationships, called *edges* or *arcs*, that exist between pairs of these objects (Diestel 2000; Wilson 1996). This means that chemoinformatics has been able to draw on the many algorithms that have been developed previously for the processing of graphs. Of particular importance in the present context are the *graph iso-*

morphism algorithms that are used to determine whether two graphs are identical and the *subgraph isomorphism* algorithms that are used to determine whether one graph is contained within another, larger graph (Gasteiger and Engel 2003; Leach and Gillet 2007).

X.2.2 Searching for specific molecules

An important search capability is structure searching: the inspection of a database to retrieve the information associated with a particular molecule (e.g., if a chemist needed to know the molecule's boiling point or to identify a synthesis for it) or to confirm the molecule's presence or absence in a database (e.g., if a chemist wanted to check whether a newly synthesised molecule was completely novel).

Structure searching in files of systematic nomenclature or linear notations is effected using conventional computer science techniques for single-key searching. These are typically based on hash coding, where an alphanumeric string (in this context, a systematic name or a canonicalised notation), is converted algorithmically to an integer identifier that acts as a key to the molecule's location on disk storage. A similar idea underlies the searching of connection table records; however, whereas names and notations are linear strings that can be converted into a canonical form very easily; this is not the case with connection tables and additional processing is required if hashing is to be used to enable fast structure searching. The generation of a canonical connection table requires the nodes of the chemical graph to be numbered, and there are up to $N!$ possible sets of numberings for an N-node graph. Following initial work by Gluck (1965), Morgan (1965) described an algorithm to impose a unique ordering on the nodes in a graph, and hence to generate a canonical connection table that can then be used for structure searching. With subsequent development (Freeland et al. 1979; Wipke and Dyott 1974), the resulting procedure, which is known to this day as the *Morgan algorithm*, forms the basis for all CAS databases and for many other chemoinformatics systems.

Hashing is an approximate procedure, in that different records can yield the same hashed key, a phenomenon that computer scientists refer to as a *collision*. In nomenclature and notation systems, collisions are avoided by means of a subsequent, and extremely simple, string comparison that confirms the equivalence of the query molecule and the molecule that is stored in the database that is being searched. In connection table systems, a graph isomorphism algorithm is used to confirm that a true match has been achieved, this involving an exhaustive, tree-search procedure in which nodes and edges from the graph describing the query molecule are mapped to nodes and edges of the graph describing a potentially matching database molecule. The mapping is extended till all the nodes have been mapped, in which case a match has been identified; or until nodes are found that cannot be mapped, in which case, the mapping backtracks to a previous, successful sub-mapping and a different mapping attempted. A mis-match is confirmed if no match has been obtained and if there are no further mappings available for testing. It will be realised that the mapping procedure has a time complexity that is a factorial function of the numbers of graph nodes involved in the comparison, and that the procedure can thus be very demanding of computational resources. Fortunately, various heuristics are available to expedite the identification of matches, and the use of the Morgan algorithm means that very few mis-matches need to be probed, making the overall procedure rapid in operation despite the complexity of the processing that is necessary.

X.2.3 Searching for chemical substructures

Probably the single most important facility in a chemoinformatics system is the ability to carry out a substructure search, i.e., the ability to identify all of those molecules in a database that contain a user-defined query substructure. For example, in a search for molecules with antibiotic behaviour, a user might wish to retrieve all of the molecules that contain a penicillin or cephalosporin ring system. Substructure searching is effected by checking the graph describing the query substructure for inclusion in the graphs describing each of the database molecules. This is an example of subgraph isomorphism: it involves an atom-by-atom and bond-by-bond mapping procedure that is analogous to, but more complex than, that used for a graph isomorphism search. A substructure search guarantees the retrieval of all molecules matching the search criterion: unfortunately, although it is completely effective, subgraph isomorphism is extremely inefficient since it belongs to the class of NP-complete computational problems for which no efficient algorithms are known to exist (Barnard 1993; Leach and Gillet 2007).

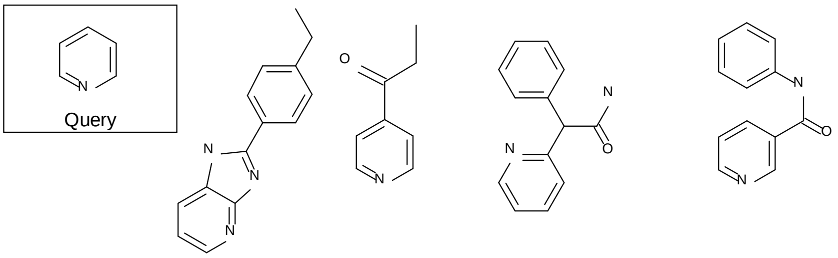


Fig. 2. Query substructure and some example hits in a search for a pyridine ring

Operational substructure searching is practicable for three reasons. First, the fact that chemical graphs are both simple (they contain relatively few nodes, most of which are of very low connectivity) and information-rich (as one can differentiate atoms and bonds by their element and bond-types, respectively). These factors serve to reduce the numbers of atom-to-atom and bond-to-bond mappings that need to be considered by a subgraph isomorphism algorithm. Second, a lot of effort has gone into the development of algorithms that can handle chemical graphs, as against graphs in general, very efficiently, with the elegant matching techniques described by Sussenguth (1965) and by Ullmann (1976) lying at the heart of current substructure searching systems. Third, and most importantly, the subgraph isomorphism search is preceded by an initial *screening search* in which each database structure is checked for the presence of features, called *screens*, that are present in the query substructure. For example, using the penicillin example mentioned above, any database structure can be eliminated from further consideration if it does not contain the fused four-membered and five-membered rings that comprise the penicillin nucleus.

A screen is a substructural feature, called a *fragment*, the presence of which is necessary, but not sufficient, for a molecule to contain the query substructure. The features that are used as screens are typically small, atom-, bond- or ring-centred fragment substructures that are algorithmically generated from a connection table when a molecule is added to the database that is to be searched. A common example of a screen is the *augmented atom* fragment, which consists of an atom, and those atoms that are bonded directly to the chosen central atom. A representation of the molecule’s structure can then be obtained by generating an augmented atom fragment centred on each atom in the molecule in turn. This information is encoded for rapid searching in a fixed-length bit-string, called a *fingerprint*, whose encoded fragments hence provide a summary representation of a molecule’s structure in just the same way as a few selected keywords provide a summary representation of the full text of a document. The fingerprint representing the query can then be matched against corresponding fingerprints representing each of the molecules in the database that is to be searched. Only a very small subset of a database will normally contain all of the screens that have been assigned to a query substructure, and only this subset then needs to undergo the time-consuming subgraph isomorphism search.

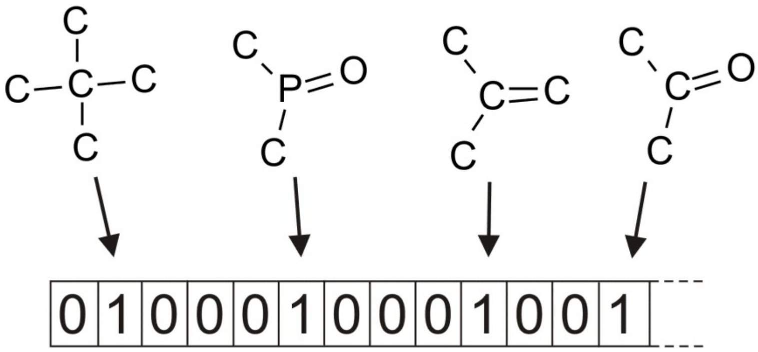


Fig. 3. Example of augmented atoms and a fingerprint

X.2.4 Similarity searching

Substructure searching provides an invaluable tool for accessing databases of chemical structures; however, it does require that the searcher is able to provide a precise definition of the substructure that is required, and this may not be possible in the early stages of a drug-discovery project, where all that is known is the identity of one or more active molecules, e.g., an existing drug from a competitor company. In such circumstances, an alternative type of searching mechanism is appropriate, called *similarity searching* (Eckert and Bajorath 2007; Willett 2009). Here, the searcher submits an entire molecule, which is normally called the *reference struc-*

ture, and the system then ranks the database in order of decreasing similarity with the reference structure, so that the molecules returned first to the searcher are those that are most closely related to it in structural terms. The underlying rationale for similarity searching is the *Similar Property Principle* (Johnson and Maggiora 1990), which states that molecules that have similar structures will have similar properties. Hence, if the reference structure has some interesting property, such as reducing a person's susceptibility to angina, then structurally similar molecules are also likely to exhibit this characteristic.

There are many different ways in which inter-molecular structural similarity can be quantified, with the most common similarity measures being based on the comparison of molecular fingerprints to identify the numbers of fragments common to a pair of molecules. This provides a very simple, but surprisingly, effective way of identifying structural relationships, as exemplified by the molecules shown in Figure 4. However, we shall not discuss similarity searching any further here, since similarity-based approaches have not, to date, been considered in much detail for searching the generic structures that form the principal focus of this chapter. This may, of course, change in the future as techniques for searching chemical patents become more widely used and as more sophisticated searching methods become necessary for effective database access. For example, Fliri et al. (2009, 2010) have recently described the use of fingerprint-based similarity methods to search sets of molecules randomly enumerated from Markush structures (see Section X.3.4).

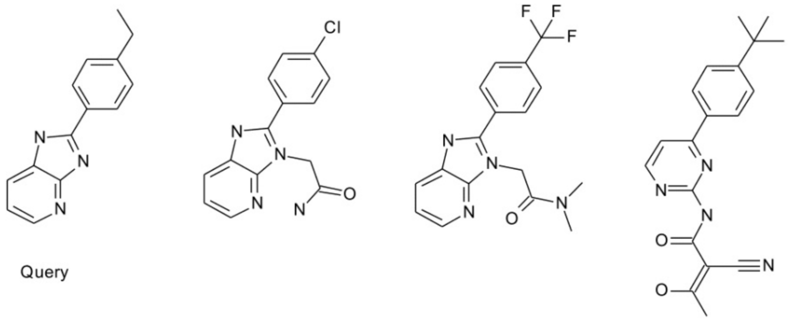


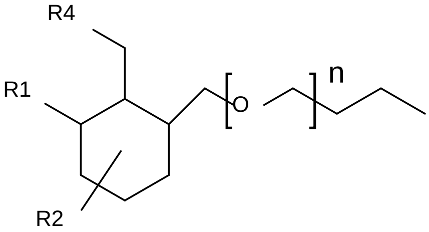
Fig. 4. Example of output from a similarity search

X.3 Searching generic chemical structures

X.3.1 Markush structure representation

In order to ensure complete coverage of the scope of invention, and hence protect the inventor’s property rights, patent documents tend to extend beyond the realm of specific description but, instead, describe the invention using broader terms. Those features which reflect the novelty of the invention are described in full and unambiguous terms, whilst other features, although fundamental to the invention, may be optional or alternative in nature. An example of the latter feature might be a new refrigerator for which the internal light might be described using a vague term such as “device for illuminating the interior”. The same is true of chemical patents in which features of the compound which are fundamental to the novelty of its operation are described using specific terms, and those for which alternatives may be substituted are described generically. The result of this treatment is a single description which can represent a potentially vast number of specific molecules, many (or even most) of which will have never been synthesised or tested.

The logical and linguistic terminology that exists in the chemical patent literature has been described in detail by Dethlefsen et al. (1991), leading to a classification of the structural variations which exist. These authors identified four types of structural variation, which are exemplified in Figure 5. Substituent variation involves the (possibly optional) set of alternative components which may be attached at a fixed point of substitution (e.g., R1 in the figure); position variation involves the alternative positions of attachment between two components of the molecule (e.g., R2). Frequency variation involves the repetition of a component either within a linear sequence or as an attachment to a ring system (e.g., *n*, indicating the presence of between 1 and 2 occurrences of the –O-CH_2- substructure); and homology variation involves the use of terminology which is itself generic in nature and which defines the component as being a member of a family of related chemical substituents (e.g., R4 in the figure indicating an alkyl group member containing 1, 2 or 3 carbon atoms).



R1 is optionally F, Cl or Br

R2 is OH or CH_3

R4 is C 1-3 alkyl

n = 1-2

Fig. 5: Examples of structure variation in generic chemical structures

Figure 5 illustrates a relatively simple generic structure, but repeated nesting of alternative components within parent components is a common feature in chemical patents, leading to a complex and often confusing structure. Enumeration of all of these the specific molecules is rarely an option due to storage requirements and computational costs. Therefore, an alternative method of computer representation is required. The basic structure adopted by current commercial systems (Berks 2001) is a logical tree in which the invariant core of the structure, the graphical component in Figure 5 for example, becomes the root. The various optional and alternative components become the branches of the tree, and the logical and connectional relationships are maintained within the representation (Barnard et al. 1982), as exemplified in Figure 6.

The logical tree encodes all of the linkages, potential or actual, within the set of molecules covered by a Markush structure, and it can hence be regarded as a form of connection table, albeit one that is far more complex that that used to describe a single specific molecule.

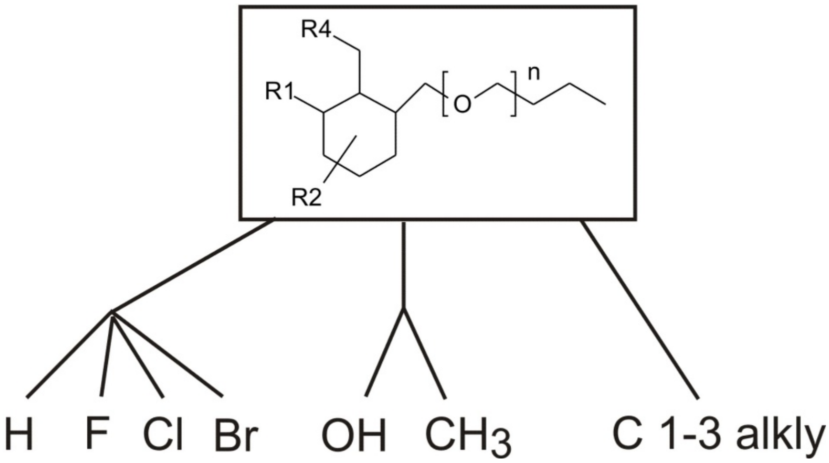


Fig. 6: Tree representation of a generic structure

X.3.2 Representational transparency

The representation of the components themselves in the tree depends on whether they are specific or generic in nature, the latter being the instance of homology variation. Specific components can be represented by a connection table, or even a line notation, whereas components relating to a chemical family, or homologous series, require alternative means. In the latter case, the representation is usually a single node which may be labelled according to the family group, and which is usually qualified by further attributes such as the number of carbon atoms or number of rings present. In the Markush DARC system, which originated from a collaboration between Derwent Publications and the French Patent Office INPI, (now called the Merged Markush Service, MMS, and produced by Thomson Reuters) these are termed “Superatoms”, whilst the MARPAT system produced by CAS uses “Hierarchical Generic Groups”.

Whichever method is employed, there remains the problem of *transparency* between the two types of representation, i.e., the lack of a common representation across components. During a search operation, whether for a structure or for a substructure, the aim is to identify mappings between the components of the query structure and those of the database structure. This operation is complicated by the requirement to map features which are specific in one representation with those which may be generic in others, a one-to-many mapping, or even features which are generic in both. In order to overcome this transparency problem, a common representation is usually sought so that the mapping becomes like-for-like. The enumeration of all possible specific members of the homologous series is again usually not an option, so a more appropriate step is the aggregation of specific components into their respective generic nodes. In the Sheffield Generic

Structures Project (Lynch and Holliday 1996), several aggregation methods were investigated, leading to a transparent representation called a *reduced graph* (Gillet et al. 1987). Figure 7 illustrates an example of such a graph in which aggregation is based on the ring (R) or non-ring nature of the features, and on further subdividing the non-ring features into those which are all carbon (C) and those which are non-carbon (Z).

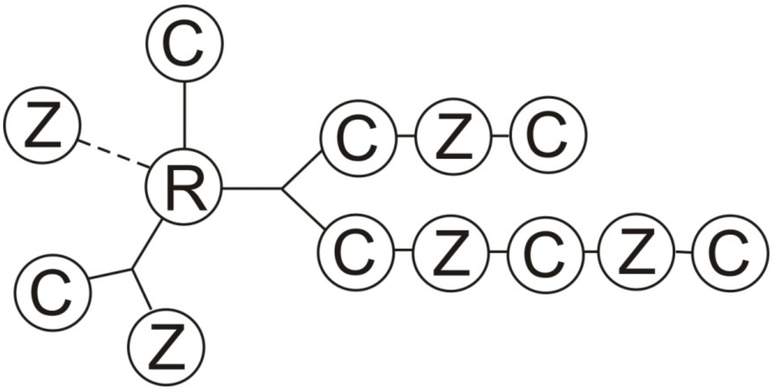


Fig. 7: Reduced graph representation of the generic structure of Figure 6 (optional connections are indicated by a dotted line)

Since we now have a common representation, one-to-one mapping can be carried out between the query and database structure. The final, and now less complex, stage is to map the constituent features of the matching query node and the database node. These are still likely to contain generic and/or specific components, but the operation is now more localised and much simpler and can be implemented using a modified version of Ullmann’s subgraph isomorphism algorithm (Holliday and Lynch 1995)

X.3.3 Fragmentation codes and screening

Early structure-based retrieval systems operated almost exclusively on the basis of fragmentation codes in which the structural components were described using a series of fragment descriptors that were analogous in principle to the fragments used for screening substructure searches of databases of specific molecules. The most notable fragmentation codes were the Derwent Chemical Code used by Derwent Publications Ltd. (Simmons 1984), the DuPont/IFI code (Kaback 1984) and the GREMAS code from International Documentation in Chemistry (Suhr 1984). The GREMAS system was highly effective and it was later possible to generate the codes automatically from the structure representation (Rössler and Kolb 1970).

As with specific structure searching, graph-based generic systems, such as MARPAT and Markush DARC, also require an initial fragment-based screening stage in order to reduce the number of compounds being sent to more computer intensive search strategies. In addition to the standard screens used at CAS for searching specific molecules, the MARPAT system uses generic group screens in which the components are reduced to their Hierarchical Generic Groups. The Markush DARC system also extended their existing specific search screens with the addition of Fuzzy FRELs (where a FREL is a circular fragment that can be considered as a larger version of the augmented atom discussed previously; some of these fuzzy FRELs were defined in terms of Superatoms and others reflected specific local variations. In the system developed at Sheffield, the approach was to generate specific fragment descriptors from the generic components (Holliday et al. 1993). Two types of screen were developed: those from the invariant components of the molecule, i.e. those alternatives which are common to all molecules covered by the generic; and those which would be optional depending on the individual specific molecules being considered at any point. In Figure 5, for instance, a screen denoting a halogen would be common to all molecules, with a logical "bubble-up" of all screens from the branches of the tree to its root maintaining the logical relationships between screens (Downs et al. 1989).

X.3.4 Recent Developments

More recently, there has been renewed interest in Markush structures; in part due to the increased computer power which was not available when the current systems first evolved. One area of interest is the application of Oracle relational database systems for storing and searching Markush structures (Barnard and Wright 2009; Csepregi et al. 2009). Many of the new developments do not, however, deal with all types of structure variation, and rely on the same philosophy of extending current systems for handling specific chemical structures.

Two other areas of interest are the automatic extraction of structural information from the patent documents (Valko and Johnson 2009; Zimmermann et al. 2005) and enumeration of specific compounds from the Markush structure. Chemical patent documents contain structures for the specific claim as well as a selection of examples. Although these usually represent a very small proportion of the possibly infinite number of compounds represented by the Markush structure, they are clearly a rich source of information and are indexed accordingly. A further source of structural information comes from the translation of nomenclatural terms identified in the document, as in the SureChem database and search system (at http://www.surechem.org). Full enumeration of all represented compounds is not possible for most structures due to the combinatorial complexity. However, as noted previously, sets of randomly enumerated specifics have been used for similarity searching, enabling rapid patent analysis and virtual library creation (Fliri et al. 2009; Fliri et al. 2010).

X.4 Conclusions

The structures of chemical molecules are an important component of the information contained in chemical patents. Individual molecules can be searched using well-established techniques from chemoinformatics, and substantial enhancements to these techniques have allowed them to be used for the representation and searching of the generic chemical structures in patents, which can describe very large numbers of structurally related molecules. In this chapter, we have summarised the techniques that are currently available for structure and substructure searching of both specific and generic structures. There are, however, many problems that remain to be addressed. Most importantly, the very generic descriptions that are sometimes used in patents mean that very large hit-lists can result even in response to quite specific structural queries: it is hence likely that there will be much interest in the future in the use of similarity-based procedures to rank search-outputs so that attention can be focused on just the top-ranked structures and patents.

References

Barnard JM (ed) (1984) Computer handling of generic chemical structures. Gower, Aldershot.

Barnard JM (1993) Substructure searching methods - old and new. J Chem Inf Comp Sci 33: 532-538.

Barnard JM, Wright PM (2009) Towards in-house searching of Markush structures from patents. World Pat Inf 31: 97-103.

Barnard JM, Lynch MF et al (1982) Computer storage and retrieval of generic structures in chemical patents. Part 4. An extended connection table representation for generic structures. J Chem Inf Comp Sci 22: 160-164.

Berks AH (2001) Current state of the art of Markush topological search systems. World Pat Inf 23: 5-13.

Csepregi S, Mate N et al. Csizmadia F (2009) Representation, searching & enumeration of Markush structures – from molecules towards patents. http://www.chemaxon.com/library/scientific-presentations/calculator-plugins/representation-searching-enumeration-of-markush-structures-from-molecules-towards-patents-2009-update/. Accessed 14 September 2010.

Dalby A, Nourse JG et al (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Comp Sci 22: 244-255.

Dethlefsen W, Lynch MF et al (1991) Computer storage and retrieval of generic chemical structures in patents, Part 11. Theoretical aspects of the use of structure languages in a retrieval system. J Chem Inf Comp Sci 31: 233-253.

Diestel R (2000) Graph theory. Springer-Verlag, New York.

Downs GM, Gillet VJ et al (1989) Computer storage and retrieval of generic chemical structures in patents, Part 10. Assignment and logical bubble-up of ring screens for structurally explicit generics. J Chem Inf Comp Sci 29: 215-224.

Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches. Drug Discov Today 12: 225-233.

Fliri A, Moysan E, et al (2009) Methods for processing generic chemical structure representations. US Patent 2009/0132464.

Fliri A, Moysan E, Nolte M (2010) Method for creating virtual compound libraries within Markush structure patent claims. WO Patent 2010/065144 A2.

Freeland R, Funk S et al (1979) The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula. J Chem Inf Comp Sci 19: 94-98.

Gasteiger, J (2006) The central role of chemoinformatics. Chemomet Intell Lab Syst 82: 200-209.

Gasteiger J, Engel T (eds) (2003) Chemoinformatics: A textbook. Wiley-VCH, Winheim.

Gillet VJ, Downs GM et al (1987) Computer-storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical-structure retrieval. J Chem Inf Comp Sci 27: 126-137.

Gluck DJ (1965) A chemical structure storage and search system developed at DuPont. J Chem Doc 5: 43-51.

Holliday J D, Downs GM et al (1993) Computer storage and retrieval of generic chemical structures in patents, Part 15. Generation of topological fragment descriptors from nontopological representation of generic structure components. J Chem Inf Comp Sci 33: 369-377.

Holliday JD, Lynch MF (1995) Computer storage and retrieval of generic chemical structures in patents. Part 16. The refined search: an algorithm for matching components of generic chemical structures at the atom-bond level. J Chem Inf Comp Sci 35: 1-7.

Johnson MA, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. John Wiley, New York.

Kaback SM (1984) The IFI/Plenum chemical indexing system. In: Barnard JM, (ed) Computer handling of generic chemical structures. Gower, Aldershot.

Leach AR, Gillet VJ (2007) An Introduction to Chemoinformatics. Kluwer, Dordrecht.

Lynch MF, Holliday JD (1996) The Sheffield Generic Structures Project - a retrospective review. J Chem Inf Comp Sci 36: 930-936.

Martin YC, Willett P (eds) (1998) Designing bioactive molecules: Three-dimensional techniques and applications. American Chemical Society, Washington.

Morgan H (1965) The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. J Chem Doc 5: 107-113.

Rössler S, Kolb A (1970) The GREMAS System, an integral part of the IDC system for chemical documentation. J Chem Doc 10: 128-134.

Simmons ES (1984) Central Patents Index Chemical Code: a user’s viewpoint. J Chem Inf Comput Sci 24: 10-15.

Sussenguth EH (1965) A graph-theoretic algorithm for matching chemical structures. J Chem Doc 5: 36-43.

Suhr C, von Harsdorf E, Dethlefsen W (1984) Derwent’s CPI and IDC’s GREMAS: remarks on their relative retrieval power with regard to Markush structures. In: Barnard JM (ed) Computer handling of generic chemical structures. Gower, Aldershot.

Ullmann JR (1976) An algorithm for subgraph isomorphism. J ACM 23: 31-42.

Valko AT, Johnson AP (2009) CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition. J Chem Inf Model 49: 780–787.

Weininger D (1988) SMILES, a chemical language and information-system.1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28: 31-36.

Willett P (2008) From chemical documentation to chemoinformatics: fifty years of chemical information science. J Inf Sci 34: 477-499.

Willett P (2009) Similarity methods in chemoinformatics. Ann Rev Inf Sci Technol 43: 3-71.

Williams AJ, Yerin A (2009) Automated identification and conversion of chemical names to structure-searchable information. In: Banville DL (ed) Chemical Information Mining. CRC Press, Boca Raton.

Wilson R (1996) Introduction to graph theory. Longman, Harlow.

Wipke WT, Dyott TM (1974) Stereochemically unique naming algorithm. J Am Chem Soc 96: 4825-4834.

Zimmermann M, Thi LTB, Hofmann M. (2005) Combating illiteracy in chemistry: towards computer-based chemical structure reconstruction. ERCIM News 60: 40-41.