The Influence of the DARC Project on Chemoinformatics Research at the University of Sheffield

John D. Holliday and Peter Willett[1]

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

Abstract. This paper summarises chemoinformatics research that has been carried out in Sheffield and that has been influenced by Jacques-Emile Dubois' work on the DARC project. The aspects discussed are the use of circular substructural fragments, the generation of hyperstructures, and the representation and searching of generic structures in chemical patents.

Keywords. Chemoinformatics, DARC, Generic structure, Graph theory, Hyperstructure, Markush, Substructure

INTRODUCTION

The world of chemoinformatics research has always been tight-knit, with a few groups from around the world being responsible for many of the developments that have led to the sophisticated software systems that exist today. One of the most important has been the DARC (for Documentation and Automated Research of Correlations) group founded and led by Jacques-Emile Dubois [1-5]. DARC is based on encoding the topologies of molecules, using techniques derived from the body of mathematics known as graph theory [6, 7]. This is not a novel idea now but it certainly was when Dubois' studies commenced in the early Sixties, his work culminating in EURECAS, the first commercial system for 2D substructure searching of the Chemical Abstract Service Registry Structure File [8].

The use of graph-based techniques has also been an important characteristic of the chemoinformatics research carried out in the University of Sheffield [9, 10], with much of this work following similar paths to studies conducted by the DARC group. In particular, we also have made extensive use of chemical graphs to develop structural representations that enable the use of a range of searching methods that are both effective and efficient in operation. In this paper, we summarise three areas of research that are closely related to, or that have been influenced by, some of the DARC studies: the use of circular fragment substructures for substructure searching and virtual screening; the use of chemical hyperstructures to encode the structural commonalities present in a set of molecules; and techniques for the representation and searching of the generic chemical structures occurring in chemical patents.

[1] To whom all correspondence should be addressed. Tel. +44-114-2222633; email p.willett@sheffield.ac.uk

CIRCULAR SUBSTRUCTURES

A 2D chemical structure diagram can be represented for computer processing by a labelled graph (called a connection table) in which the nodes and edges of a graph represent the atoms and bonds, respectively, of a molecule. A chemical database can hence be represented by a large number of such graphs, with searches for chemical substructures (e.g., for all molecules that contain the characteristic benzodiazepine ring system) being carried out using a subgraph isomorphism algorithm. This algorithm carries out an exhaustive comparison of the graph describing the query substructure with the graphs describing each of the database molecules [11, 12], an extremely effective procedure but one that is far too time-consuming for use on databases of non-trivial size without drastic modification. Substructure searching is computationally feasible since the subgraph-isomorphism stage (which is sometimes referred to as atom-by-atom searching) is preceded by an initial screening stage. A ***screen*** is a substructural feature, the presence of which is necessary, but not sufficient, for a molecule to contain the query substructure. These features are typically small, atom-, bond- or ring-centred molecular fragments that are algorithmically generated from a molecule's connection table when it is added to a database.

The presence of a particular fragment in a molecule is denoted by setting to "on" one or more of the bits in a ***fingerprint***, a fixed-length binary vector, so that a fingerprint provides a concise summary of all of the substructural features that are present in the corresponding molecule. An analogous fingerprint is generated to describe the query substructure, and the screen search then involves checking the fingerprint representing each database structure for the presence of the screens that are encoded in the fingerprint representing the query substructure. Only a few database molecules will normally contain all of the screens that have been assigned to a query substructure, and just these few molecules then need to be considered in the final, graph-based, atom-by-atom search. The efficiency of substructure searching system will hence be determined primarily by the extent to which the screens are able to minimise the numbers of molecules undergoing the subgraph isomorphism check, and there has been much interest in the types of substructural fragment that can be used for this purpose.

The ELCO (Environment which is Limited, Concentric and Ordered) concept lies at the heart of the DARC system and involves choosing some atom, *i*, in a molecule as the origin for a tree that is represented by successive concentric circles of atoms centred on *i*. Thus *i* is described by the atoms, *j*, that are bonded to it, with each atom *j* being described recursively by the atoms *k* bonded to it, and so on until a detailed and complete representation of the topology of the molecule has been obtained. From the resulting graph it is simple to generate fixed-radius circular substructures called FRELs (Fragment Reduced to an Environment which is Limited) that can be used for screening: in the case of EURECAS, a FREL is generated for all atoms with a connectivity of two or more, with each such atom being described by a fragment of radius two bonds [8].

The concept of a circular substructure has proved to be a powerful one that has been widely adopted (see, e.g., [13-16]). Our interest arose from a need to develop a simple procedure for generating multiple sets of screens for substructure searching, so

as to probe the effect of changes in screen-set composition on search performance. Earlier work in Sheffield and in the USA [17, 18] had demonstrated that the efficiency of screening would be maximised if the fragments chosen for inclusion in a screen set did not occur either with very high or with very low frequencies in the database that was to be screened and were statistically independent of each other; moreover the work of Adamson et al. [17] had shown the search-effectiveness of circular fragment definitions. These findings guided the screen-set selection procedure described by Willett [19], where each screen was a string of integers, in which the *i*-th integer characterised a circle of radius *i*-1 bonds centred on a specific atom.

The integers were obtained by an adaptation of the Morgan algorithm [20], which was developed to discriminate between the atoms comprising a molecule on the basis of their extended connectivity values, where the *i*-th order connectivity of an atom is calculated by summing the (*i*-1)-th order connectivities of all immediately adjacent atoms. The operation of the Morgan algorithm is illustrated in Figure 1. For screening, the initial values were not just the connectivities of an atom but integers encoding the elemental type and the pattern of pendant bonds for an atom [19], i.e., a definition similar to that used in a FREL. An upperbound radius was specified and the integers computed for circular substructures up to that size, so that a string of, e.g., six integers, would encode a central atom, and then circular substructures of radii 1, 2, 3, 4 and 5 bonds. The constituent, smaller-radius substructures could then be generated by sequential removal of the right-hand-most integer. The generation procedure was repeated for all of the molecules in a database (or a sample thereof) and the resulting sets of integer-strings then input to an algorithmic procedure that selected a set of strings that occurred approximately equifrequently and that had minimal inter-screen dependencies.

The screens sets resulting from this procedure are very simple to generate (whilst still exhibiting high screenout) and were used in two subsequent studies of screen-set behaviour. The first analysed the effect of changes in screen-set size (in terms of the number of constituent screens) on screenout performance, and showed that screenout tended to level off as the number of screens increased [21]. The second analysed the effect of changes in the size of the database that was used to generate the screen set, and showed that even quite small numbers of molecules were sufficient to generate screen sets that exhibited high screenout [22].

A commercial implementation of Morgan-based circular substructures has been developed by Scitegic Inc. as part of their Pipeline Pilot Software [23]. Two types of fragment are supported: Extended Connectivity Fingerprints (ECFPs) and Functional Connectivity Fingerprints (FCFPs). The initial code assigned to an atom is based on the number of connections, the element type, the charge, and the mass for ECFPs and on six generalised atom-types - viz., hydrogen-bond donor, hydrogen-bond acceptor, positively ionisable, negatively ionisable, aromatic and halogen - for FCFPs. This code, in combination with the bond information and with the codes of its immediate neighbour atoms is hashed to produce the next order code, which is mapped into an address space of size 2^{32}, and the process iterated until the required level of description has been achieved. The process is repeated for each heavy atom present, and the Scitegic software hence represents a molecule by list of integers,

each describing a molecular feature and each in the range -2^{31} to 2^{31}. These integers can then, if required, be mapped into a fixed-length fingerprint.

In several recent studies, we have found that the Scitegic ECFP fingerprints provide an effective tool for *virtual screening*, i.e., prioritising the biological testing of large chemical datasets so as to ensure that those molecules with the largest *a priori* probabilities of activity are assayed first in a lead discovery programme [11, 12]. Our studies have involved simulated virtual screening of the *MDL Drug Data Report* database, a publicly available file of ca. 100K molecules from the literature that have associated pharmacological activities [24]. Our studies have been of two types: virtual screening based on the similarities of database molecules to a single known active using the Tanimoto coefficient and data fusion techniques [25]; and virtual screening based on machine-learning techniques where sets of known active and known inactive molecules are used to train a scoring mechanism that can then be used to rank molecules not present in the training-set [26]. The ECFP fragments have been found to perform well in both types of application, outclassing all of the other types of 2D fingerprint that were considered in a detailed comparative study of screening effectiveness [27], and thus demonstrated the power of this approach to the characterisation of chemical substructures.

CHEMICAL HYPERSTRUCTURES

An important concept in the DARC system is that of a ***hyperstructure***, a molecule-like object that is formed by the superimposition of sets of molecules and that stores areas of structural commonality only once. The connectivity of a group of molecules is thus encoded with minimal redundancy, with each individual structure being contained as a substructure of the complete hyperstructure. A hyperstructure represents the logical union (Boolean OR) of a set of structures (as compared to the maximum common substructure – or MCS – that represents their logical intersection, or Boolean AND). Let the hyperstructure for a set of structures S(1), S(2)….S(N) be denoted by H. Then the generation of H can be described by the following pseudocode:

```
H := S(1)
FOR X := 2 TO N DO
        H:= H È S(X)
```

Each iteration of the main loop involves two steps. First, a comparison step matching the X-th structure to the current hyperstructure by means of a graph-fitting procedure that maximises the degree of overlap between H and S(X) using an MCS (or MCS-like) algorithm; second, an updating step that first eliminates from S(X) those substructural features that are already present in H and then adds the remaining substructural features in S(X) to H.

The hyperstructure concept was developed by the DARC group to facilitate structure-activity correlation studies, which typically involve relatively small numbers of structurally related molecules. We have found that it can also be used to organise large files of structurally disparate molecules, an application area that Dubois identified as being less developed [5]. Our starting point was a paper by Vladutz and Gould [28], who noted that the elimination of duplicate substructural moieties means that a hyperstructure should require less storage than its constituent molecules,

resulting in some degree of data compression, and might yield increased substructure-searching speeds, since the repeating units need be searched only once. However, hyperstructures are far more complex than the graphs describing the individual constituent molecules, with the result that the discovery of the optimal mappings of these constituent molecules is computationally demanding. This is particularly the case when, as in our projected database-searching application, there are many, structurally disparate molecules to be processed; indeed, much of our early work focused on determining whether it was feasible to generate hyperstructures in such circumstances. Three approaches were studied [29, 30]. The first, based on the use of a maximum overlap set (MOS) algorithm, which is analogous to an MCS algorithm, proved to be too demanding of computational resources for practical use. The second was much faster but less precise. The third approach was based on a genetic algorithm (as described further below); this proved to be by far the most appropriate and enabled a detailed examination of the utility of the hyperstructure concept for substructure searching. It was found that whilst there was some potential for increasing search speeds, this was likely to be far outweighed by the complexities of the procedures required to generate and to search large hyperstructures [31]. However, more recently, we have considered an alternative use for a hyperstructure: its construction from a training-set of known active and known inactive molecules to suggest topological pharmacophores and qualitative structure-activity relationships [32]. The operation of the hyperstructure-generation procedure is illustrated in Figure 2.

The basic idea is a simple one, with related ideas having been studied by other workers [33-35]. Assume that a training-set is available, containing molecules that have been found to be either active or inactive in a particular bioassay of interest; then a hyperstructure can be generated in which each node is weighted according to the activities of each of the molecules that have been mapped to that node during the generation of the hyperstructure. Specifically, each node and edge in the resultant hyperstructure has two weights associated with it, one for activity and one for inactivity. Since the hyperstructure is initialised with a single molecule all the hyperstructure nodes and edges are initialised to one or zero according to the activity of the starting-point molecule. Subsequent mappings to these nodes and edges cause the respective activity or inactivity frequency counters to be incremented, depending on the (in)activity of the input training-set molecule that is being mapped. Any new nodes and edges that are appended to the hyperstructure will also have their activity and inactivity weights initialised to one or zero accordingly. The resulting integer counts can then be used to compute a range of weights; the simplest is obtained merely by subtracting the inactive count of the node or edge from the active count (other, more sophisticated approaches are possible [32]): a positive, negative or zero outcome indicates that a feature is weighted to be predominantly active, predominantly inactive or neutral.

Experiments with several sets of molecules from the National Cancer Institute [36] and IDAlert [37] databases showed that a hyperstructure generated in this way could encode a significant amount of the structure-activity information present in a training-set, with distributions of activity and inactivity weights that were significantly different from those obtained in randomisation experiments using scrambled activity data [32]. This implies that the activity and inactivity weights for the nodes and edges can be used to indicate those parts of the hyperstructure that are positively or

negatively associated with activity, and we have studied two ways in which this information can be exploited.

The first approach was based on the observation that much of a typical hyperstructure consists of nodes and edges that have low valued weights and that thus have little association with (in)activity. A threshold is set so as to highlight just those parts of the hyperstructure that have a high positive or high negative weight, and hence to highlight features that might be expected to occur preferentially in active or in inactive molecules. The former features might then be considered as a topological pharmacophore that could be used as the query in a substructure search to identify other, previously untested molecules possessing this combination of features. The second approach involves consideration of the individual molecules in the training-set molecule. This is extracted from the hyperstructure and its nodes and edges colour-coded according to the activity and inactivity weights in the hyperstructure, e.g., red for active and blue for inactive, thus facilitating the identification of particularly active or inactive substructures. Alternatively, a new test-set molecule can be mapped to the hyperstructure and its nodes and edges coded according to the weights for those nodes and edges to which it is mapped. In either case, a simple visualisation is achieved that identifies the "hotspots" in a molecule that may be of particular importance in determining that molecule's biological activity [32].

More recently, we have extended the concept of a hyperstructure to take account of not just the topologies but also the geometries of molecules, specifically, we have described an algorithm for aligning multiple 3D structures [38] that forms one of the components of the GALAHD (Genetic Algorithm with Linear Assignment for Hypermolecule Alignment of Datasets) software for automated pharmacophore detection [39].

GENERIC STRUCTURES IN CHEMICAL PATENTS

The requirement that a chemical patent covers the widest definition of a family of compounds, in order to protect the interests of the claimant, has meant that the descriptions used in patent claims are often complex and varied. A suitable database system is required to store and search this patent information and the representation used is the *generic* structure. Generic chemical structures, often called Markush structures, introduce a level of complexity in chemical structure handling which arises from the need to cover a potentially large, possibly even infinite, number of compounds under the one representation. Dethlefsen [40] classifies the types of variation found in generic structures into four types: substituent variation, where the substituent at a given position is defined as a list of alternatives (e.g., phenyl substituted in para position by Cl, F or Br); position variation, where the position of attachment of substructural fragments is not distinct (e.g., dichlorobenzene); frequency variation, where a substructural fragment may occur with variable frequency (e.g., $CH_3(CH_2)_{1\text{-}3}OH$); and homology variation, where a substituent is defined as a class of structural homologues (e.g. n-alkyl). Clearly, the extensive use of all four types of variation, added to the fact that such instances are often nested to many levels within the representation, means that a generic structure database system requires the development of novel algorithms and representations in order to deal with

these high levels of complexity. An example of a generic structure illustrating these four types of variation is shown in Figure 3.

The need for a comprehensive storage and search system was identified by Prof. Michael Lynch in the late 1970s, and in 1979 the Sheffield Generic Structures Project began [41]. This was the first attempt at a structure-based approach, the only existing systems at that time being based on fragment codes [42,43]. By the late 1980s, the Sheffield system was nearly complete and comprised an input language, GENSAL (for Generic Structure Language) [44,45], a computer representation, the Extended Connection Table Representation (or ECTR) [46], a fragment screening system [47] and an intermediate screening stage utilising a reduced-graph representation [48]. The fragment descriptors used in the screening stage were a subset of the CAS Online screen dictionary [49] and included sequences of atoms and bonds, as well as augmented atoms, the latter effectively a circular substructure, or FREL, with a radius of one bond.

Two commercial systems appeared at about this time, one from the Chemical Abstracts Service (CAS) and now known as MARPAT [50], the other, Markush DARC [51], being an extension of the generic DARC system from Telesystemes. Generic DARC was developed as part of the DARC system, originated by Dubois, and provided generic query generation for searches on files of specific (i.e. non-variant) structures. The extension into a full Markush system came about through a collaboration between Telesystemes, Derwent Publications and the French Patent Office (INPI). Markush DARC extended the capability of Generic DARC to allow storage and searching of full generic structures using a graphical interface. This was facilitated by the use of ***Superatoms***, effectively an extension of the periodic table in which a single atom represented a generic group expression (or homologous series), e.g., CHK for an alkyl. The early system was limited in that it did not include all of the types of structural variation found in generic structures, and it did not allow for transparency between Superatoms and their equivalent specific structural components, e.g., ethyl would not match alkyl, and it was left to the encoder to define the possible specific instances of a homologous series. These limitations were later overcome, so as to allow complete transparency between specific and non-specific components [52].

Fundamental to DARC systems is the use of FRELs in the topological screening stage and these were enhanced in Markush DARC by the development of FRELs derived from Superatom and real atom components [52] and generic FRELs [53]. Both the Markush DARC and the Sheffield systems investigated the generation of topological fragments from homologous series. Benichou and Klimczak mention the possibility of generating FRELs from Superatoms [52]; a similar approach was tested in Sheffield using a program called TOPOGRAM [54], but this was found to have limits for some of the more complex classes. The Sheffield and Telesystemes DARC teams, together with Derwent Publications, collaborated from 1990-1992 on a project to produce a transparent screening strategy based on reduced graphs. The Sheffield project had a working input system and a reduced graph representation which was partitioned in a similar way to DARC Superatoms. The aim of the project was to translate the Markush DARC data format into an all-Superatom representation via Sheffield's ECTR and reduced graph algorithms. The result was a simplified graph on which reduced-graph screening could be carried out.

More recently, many of the concepts underlying the generic structure systems developed in Sheffield and elsewhere have been used in the design of virtual compound libraries [55]. A virtual library consists of a core structure surrounded by any number of variable substituent groups, each group containing possibly thousands of alternative monomers. A library shares many of the features of a generic structure and many of the algorithms and representations described here can hence be used with only minor modifications and enhancements.

CONCLUSIONS

Jacques-Emile Dubois was one of the pioneers of chemoinformatics. He was instrumental in establishing the power and versatility of graph-based methods for the processing of information about the topologies of chemical compounds, and his publications have encouraged many others to contribute to this ever-growing field. Many of the studies that we have carried out in Sheffield have been influenced by his writings (as summarised in this paper), and we are pleased we have been able to make our small contribution to this celebration of Jacques-Emile's life and work.

REFERENCES

1. Dubois, J.-E., French national policy for chemical information and the DARC system as a potential tool of this policy, *J. Chem. Doc*., **1973**, *13*, p. 8.
2. Dubois, J.-E., *Computer Representation and Manipulation of Chemical Information*, W.T. Wipke, S.R. Heller, R.J. Feldmann, E. Hyde (eds), Wiley, New York, **1974**, p. 239-264.
3. Dubois, J.-E., *Chemical Applications of Graph Theory*, A.T. Balaban (ed), Academic Press, London, 1976, p. 333-370.
4. Dubois, J.-E., Sobel, Y., DARC system for documentation and artificial intelligence in chemistry, *J. Chem. Inf. Comput. Sci*., **1985**, *25*, p. 326.
5. Dubois, J.-E., *The History and Heritage of Scientific and Technological Systems*, W.B. Rayward, M.E. Bowden (eds), Information Today, Medford NJ, **2004**, p. 149-167.
6. Wilson, R., *Introduction to Graph Theory*; 4th ed., Longman: Harlow, **1996**.
7. Diestel, R. *Graph Theory*, Springer-Verlag: New York, **2000**.
8. Attias, R., DARC substructure search system: a new approach to chemical information, *J. Chem. Inf. Comput. Sci*., **1983**, *23*, p. 102.
9. Lynch, M.F., Willett, P., Information retrieval research in the Department of Information Studies, University of Sheffield: 1965-1985, *J. Inf. Sci.*, **1987**, *13*, p. 221.
10. Bishop, N., Gillet, V.J., Holliday, J.D., Willett, P., Chemoinformatics research at the University of Sheffield: a history and citation analysis, *J. Inf. Sci.*, **2003**, *29*, p. 249.
11. Leach, A.R., Gillet V.J., *An Introduction to Chemoinformatics*, Kluwer, Dordrecht, **2003**.
12. Gasteiger, J., Engel, T. (eds), *Chemoinformatics*, Wiley-VCH, Weinheim, *2003*.
13. Randic, M., Fragment search in acyclic structures, *J. Chem. Inf. Comput. Sci*., **1978**, *18*, p. 101.
14. Schubert, W., Ugi, I., Constitutional symmetry and unique description of molecules, *J. Amer. Chem. Soc.*, **1978**, *100*, p. 37.
15. Bremser, W., HOSE – a novel substructural code, *Anal. Chim. Acta*, **1978**, *103*, p. 355.
16. Bender, A., Mussa, H.Y., Glen, R.C., Reiling, S., MolSim: searching using atom environments, information-based feature selection and a naive Bayesian classifier, *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 170-178
17. Adamson, G.W., Cowell, J., Lynch, M.F., McLure, A.H.W., Town, W.G., Yapp, A.M., Strategic considerations in the design of a screening system for substructure searches of chemical structure files. *J. Chem. Doc*., **1973**, *13*, p. 153.
18. Hodes, L., Selection of descriptors according to discrimination and redundancy. Application to chemical substructure searching, *J. Chem. Inf. Comput. Sci*., **1976**, *16*, p. 88.
19. Willett, P., A screen set generation algorithm, *J. Chem. Inf. Comput. Sci*., **1979,** *19*, p. 159.

20. Morgan, H.L., The generation of a unique machine description for chemical structures: a technique developed at Chemical Abstracts Service, *J. Chem. Doc*., **1965**, *5*, p. 107.
21. Gannon, M.T., Willett, P., Sampling considerations in the selection of fragment screens for chemical substructure search systems, *J. Chem. Inf. Comput. Sci*., **1979,** *19*, p. 251.
22. Willett, P., The effect of screen set size on retrieval from chemical substructure search systems, *J. Chem. Inf. Comput. Sci*., **1979,** *19*, p. 253.
23. The Pipeline Pilot software is available from Scitegic Inc. at http://www.scitegic.com
24. The *MDL Drug Data Report* database is available from MDL Information Systems Inc. at http://www.mdli.com
25. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A., Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information, *J. Med. Chem*, **2005**, *48*, p. 7049.
26. Chen, B., Harrison, R.F., Pasupa, K., Wilton, D.J., Willett, P., Wood, D.J., Lewell, X.Q., Virtual screening using binary kernel discrimination: effect of noisy training data and the optimisation of performance, *J. Chem. Inf. Comput. Sci*., **2006**, *46*, p. 478.
27. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A., Topological descriptors for similarity-based virtual screening using multiple bioactive reference structures, *Org. Biomol. Chem.*, **2004**, *2*, p. 3256.
28. Vladutz, G., Gould, S.R., *Chemical Structures. The International Language of Chemistry*, W.A. Warr (ed), Springer Verlag, Berlin, **1988**, p. 371-384.
29. Brown, R.D., Downs, G.M., Willett, P., Cook, A.P.F., A hyperstructure model for chemical structure handling: generation and atom-by-atom searching of hyperstructures, *J. Chem. Inf. Comput. Sci*., **1992**, *32*, p. 522.
30. Brown, R.D., Jones, G.J., Willett, P., Glen, R.C., Matching two-dimensional chemical graphs using genetic algorithms, *J. Chem. Inf. Comput. Sci*., **1994**, *34*, p. 63.
31. Brown, R.D., Downs, G.M., Jones, G., Willett, P., A hyperstructure model for chemical structure handling: techniques for substructure searching, *J. Chem. Inf. Comput. Sci*., **1994**, *34*, p. 47.
32. Brown, N., Willett, P., Wilton, D.J., Lewis, R.A., Generation and display of activity-weighted chemical hyperstructures, *J. Chem. Inf. Comput. Sci*., **2003**, *43*, p. 288.
33. Simon, Z., Chirica, A., Holban, S., Ciubotaru, D., Mihala, G.I., *Minimum Steric Difference. The MTD Method for QSAR Studies*, Research Studies Press, Letchworth, **1994.**
34. Downs, G.M., Gill, G.S., Willett, P., Walsh, P.T., Automated descriptor selection and hyperstructure generation to assist SAR studies, *SAR QSAR Environment. Res.* **1995**, *3*, p. 253.
35. Palyulin, V.A., Radchenko, E.V., Zefirov, N.A., Molecular Field Topology Analysis method in QSAR studies of organic compounds, *J. Chem. Inf. Comput. Sci*. **2000**, *40*, p. 659.
36. The ID Alert database is available from Current Drugs Limited at http://www.current-drugs.com/
37. The NCI AIDS database is available from the National Institutes of Health at http://dtp.nci.nih.gov/
38. Richmond, N.J., Willett, P., Clark, R.D., Alignment of three-dimensional molecules using an image recognition algorithm, *J. Mol. Graph. Model.*, **2004**, *23*, p. 199.
39. The GALAHD software is available from Tripos Inc. at http://www.tripos.com
40. Dethlefsen, W., Lynch, M.F., Gillet, V.J., Downs, G.M., Holliday. J.D., Barnard, J.M., Computer storage and retrieval of generic chemical structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, p. 233.
41. Lynch, M.F., Barnard, J.M., Welford, S.M., Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy, *J. Chem. Inf. Comput. Sci.* **1981**, *21*, p. 148.
42. Norton, P., Central Patents Index (CPI) as a source of information for the pharmaceutical chemist, *Drug. Inf. J.* **1982**, *16*, p. 208.
43. Rossler, S., Kolb, A., The GREMAS system – an integral part of the IDC system for chemical documentation, *J. Chem. Doc.* **1980**, *10*, p. 128.
44. Barnard, J.M., Lynch, M.F., Welford, S.M., Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL: a formal language for the description of generic chemical structures, *J. Chem. Inf. Comput. Sci.* **1981**, *21*, p. 151.
45. Barnard, J.M., Lynch, M.F., Welford, S.M., Computer storage and retrieval of generic chemical structures in patents. 6. An interpreter program for the generic structure language GENSAL, *J. Chem. Inf. Comput. Sci.* **1984**, *24*, p. 66.

46. Barnard, J.M., Lynch, M.F., Welford, S.M., Computer storage and retrieval of generic chemical structures in patents. 4. An extended connection table representation (ECTR) for structures, *J. Chem. Inf. Comput. Sci.* **1982**, *22*, p. 160.
47. Welford, S.M., Lynch, M.F., Barnard, J.M., Computer storage and retrieval of generic chemical structures in patents. 5. Algorithmic generation of fragment descriptors for generic structure screening, *J. Chem. Inf. Comput. Sci.* **1984**, *24*, p. 57.
48. Gillet, V.J., Downs, G.M., Ling, A., Lynch, M.F., Venkataram, P., Wood, J.V., Dethlefsen, W., 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. Comput. Sci.* **1987**, *27*, p. 126.
49. Dittmar, P.G., Farmer, N.A., Fisanick, W., Haines, R.C., Miller, J.A., Koch, B., The CAS ONLINE Search System. I. General system design and selection, generation and use of search screens, *J. Chem. Inf. Comput. Sci.* **1983**, *23*, p. 93.
50. Fisanick, W., The Chemical Abstracts Service generic chemical (Markush) storage and retrieval capability. Part 1. Basic concepts, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, p. 145.
51. Shenton, K., Norton, P., Fearns, E.A., *Chemical Structures. The International Language of Chemistry*, W.A. Warr (ed), Springer Verlag, Berlin, **1988**, p 169-178.
52. Benichou, P., Klimczak, C., Handling genericity in chemical structures using the Markush DARC software, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, p. 43.
53. Dubois, J.-E., Panaye, A., Attias, R., DARC system: notions of defined and generic structures. Filiation and coding of FREL substructure classes, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, p. 74.
54. Welford, S.M., Lynch, M.F., Barnard, J.M., Computer storage and retrieval of generic chemical structures in patents. 3. Chemical grammars and their role in the manipulation of chemical structures, *J. Chem. Inf. Comput. Sci.* **1981**, *21*, p. 161.
55. Barnard, J.M., Downs, G.M., von Scholley-Pfab, A., Brown, R.D., Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries, *J. Mol. Graph. Model.* **2000**, *18*, p. 452.

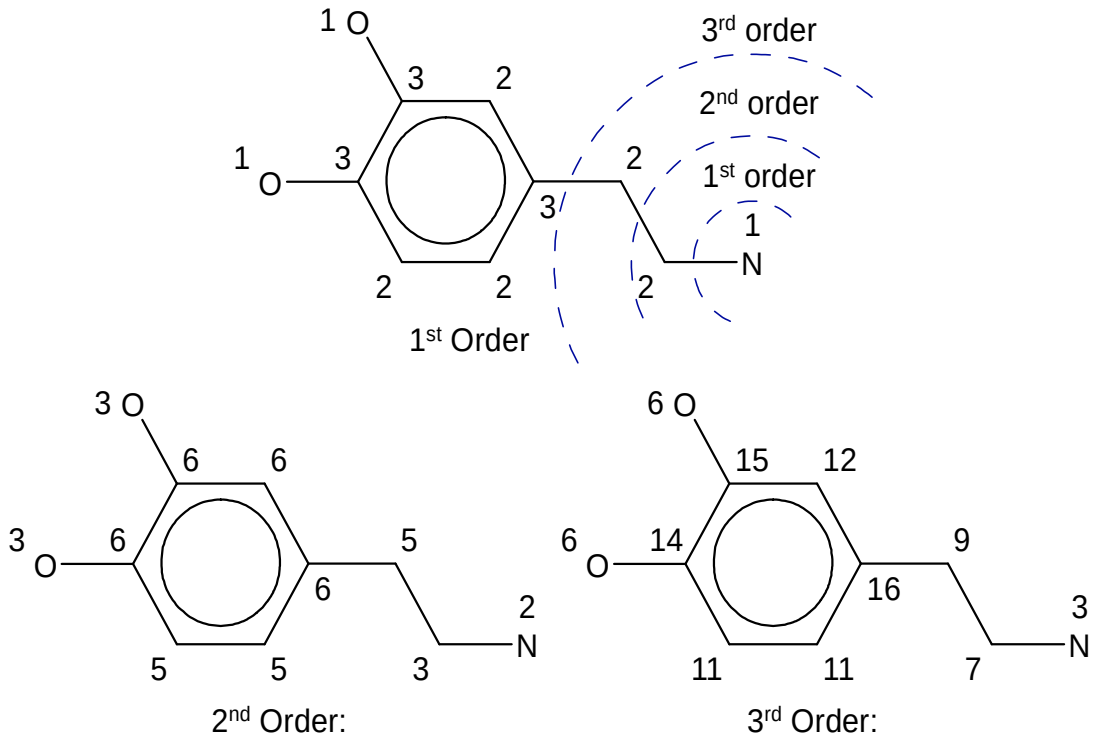


Figure 1. Operation of the Morgan algorithm. The numbers of connected non-hydrogen atoms for each atom (the connectivity) are shown under 1st order. The *i*-th order connectivity (*i*=2 or 3 here) for each atom is then obtained by summing the (*i*-1)-th order connectivities for the immediately adjacent atoms.

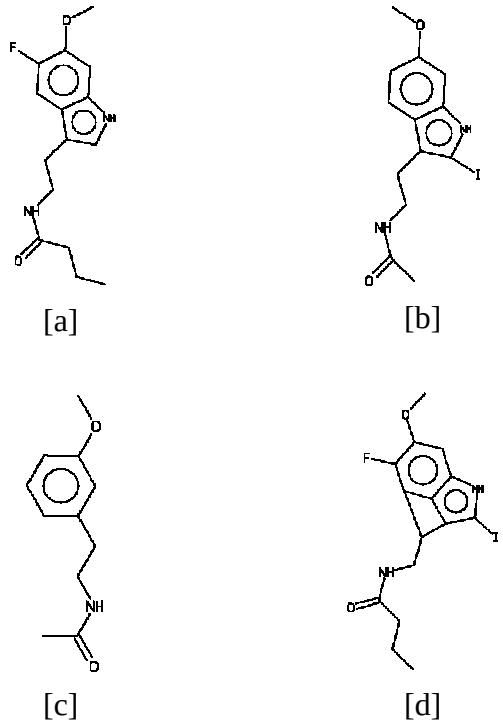
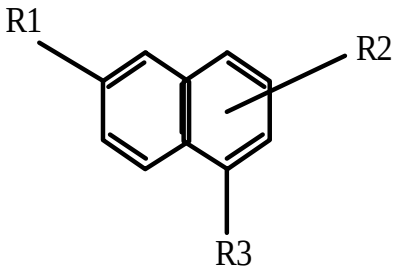


Figure 2: A chemical hyperstructure [d] generated from the molecules [a], [b] and [c] by progresive superposition of structure diagrams using a genetic algorithm as described by Brown et al. [30, 32].



R1 is H, Cl or $(CH_2)_nCH_3$
n is 2 to 4
R2 is F or Cl
R3 is 1-3 carbon alkyl, an oxygen-containing ring
or an electron withdrawing group.

Figure 3: Example of a generic, or Markush, chemical structure, in which the representation describes a large number of different specific molecules.