Book Section:
Birchall, K. and Gillet, V.J. (2011) Reduced graphs and their applications in chemoinformatics. In: Bajorath, J., (ed.) Chemoinformatics and Computational Chemical Biology. Methods in Molecular Biology, 672 . Humana Press , 197 - 212.

https://doi.org/10.1007/978-1-60761-839-3_8

Reduced Graphs and Their Applications in Chemoinformatics

Kristian Birchall[1] and Valerie J. Gillet[2]

[1]Department of Chemistry, University of Sheffield, Western Bank, Sheffield, S10 2TN

[2]Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP.

Summary

Reduced graphs provide summary representations of chemical structures by collapsing groups of connected atoms into single nodes while preserving the topology of the original structures. This chapter reviews the extensive work that has been carried out on reduced graphs at The University of Sheffield and includes discussion of their application to the representation and search of Markush structures in patents; the varied approaches that have been implemented for similarity searching; their use in cluster representation; the different ways in which they have been applied to extract structure-activity relationships; and their use in encoding bioisosteres.

Key Words: reduced graph, graph reduction, similarity searching, bioisosterism, structure-activity relationships, Markush structures, generic structures, database search.

Introduction

Reduced graphs (***1***) provide summary or abstract representations of chemical structures and are generated by collapsing connected atoms into single nodes, edges are then formed between the nodes according to bonds in the original structure. Reduced graphs have been used in a variety of applications in chemoinformatics ranging from the representation and search of Markush structures in chemical patents to the identification of structure-activity relationships (SARs). Many different graph reduction schemes have been devised and the optimal scheme is likely to depend on the particular application. Examples of different types of reduced graphs are shown in Figure 1. The idea of characterising chemical structures by their structural components is long established in chemical information and is implicit in most systematic chemical nomenclature: structures are fragmented into ring systems and acyclic components which are described individually with conventions used to indicate how they are connected, for example, N-(4-hydroxyphenyl) acetamide (the systematic name for paracetemol). Reduced graphs also aim to summarise structures according to their structural components, however in contrast to nomenclature systems, they retain structural information on how the components are connected in graphical form. This encoding of topology enables structural comparisons to be made which cannot be achieved through the use of nomenclature.

In this chapter, we focus on the extensive work that has been carried out on reduced graphs at The University of Sheffield for a variety of different applications. We also recognise the substantial efforts made by other groups in related methods, notably the feature trees approach by Rarey et al.**(*2, 3*)** and the extended reduced graph, ErG, by Stiefl et al.**(*4, 5*)**, and provide a brief summary of these approaches.

Reduced graphs for searching Markush structures

Reduced graphs were first used at Sheffield as a component of a search system for Markush structures **(*1, 6*)**. Markush structures (also known as generic structures) are chemical structures that involve the specification of lists of alternative substituents attached to a central core structure. They occur frequently in chemical patents where they are used to describe a large and often unlimited number of structures with the aim of protecting a whole class of compounds rather than a few specific examples. An example Markush structure is shown in Figure 2 and consists of a central core group with variable R-groups that are used to represent lists of alternative substituents (or substructures) attached to the core. Markush structures pose several difficulties for storage and retrieval. In addition to handling the large number of compounds encoded in a single representation and dealing with different ways of partitioning a structure into substructures, one of the major difficulties is the use of generic nomenclature to indicate that a substituent may be any member of a homologous series, for example, in expressions such as “R1 is an alkyl group”. Generic nomenclature presents difficulties for search since it is necessary to be able to match specific instances of a homologous series with the generic term, for example, to recognise that "methyl" is an instance of "alkyl".

Reduced graphs were developed in the Sheffield Generic Chemical Structures Project **(*7*)** to provide an additional level of search that is intermediate in complexity between the traditional fragment screening and atom-by-atom search methods that were developed for specific structures, and to provide an effective way of dealing with generic nomenclature **(*6*)**. In the Sheffield project, homologous series are represented by parameter lists which indicate the structural features that characterise the series such as: the number and type of rings present and the presence or absence of heteroatoms etc. Most of the substituents that are expressed as homologous series in patents can be classified as ring or non-ring (for example, aryl, heterocycle, alkyl, alkene etc) and can therefore be represented as single nodes in a ring/non-ring (R/N) graph reduction scheme. The reduced graph representation of the generic structure is also shown in Figure 2; the reduced graph is rooted on the central ring node which is derived from the core structure and contains alternative nodes indicated by the branched edge labelled "OR" and an optional node indicated by the dashed edge. In the example shown, the partitioning of the generic structure into partial structures corresponds with the node definitions. However, in other cases, a single reduced graph node might span different partial structures in the generic structure.

The searching of Markush structures is carried out at three levels. The first level is a fragmentation search in which fragments are generated from both the structural fragments and the parameter lists used to represent the generic nomenclature: the fragments are organised as those which MUST be present in the generic structure and those that MAY be present since they occur in alternative substructures. The reduced graph search is based on graph matching procedures and is considerably faster than graph matching at the atom and bond level due to the relatively small size of the reduced graphs. The final search is an atom-by-atom search modified to deal with the generic nomenclature. The three search methods are applied in sequence: for a given query, those database compounds which pass the fragment stage are passed to the reduced graph search and finally those compounds remaining after the reduced graph search are subjected to the most time-consuming atom-by-atom search. Although the Sheffield search system did not become a public system in its own right, it undoubtedly had a major influence on the Markush DARC system of Derwent Information Limited (***8***) and the MARPAT system of Chemical Abstracts Service (***9, 10***).

Reduced graphs for similarity searching

Since the advent of similarity searching in the 1980s much effort has been expended on developing new descriptors with the aim of identifying compounds that share the same activity. The first similarity searching procedures were developed using fragment bitstrings that were devised for substructure search (***11, 12***). These proved to be remarkably successful, although this good performance was, in part, due to the nature of the datasets on which they were evaluated, which often consisted of series of structural analogues. A more recent focus in similarity searching has been the identification of compounds exhibiting the same activity but belonging to different lead series; a technique that has become known as scaffold hopping (***13***). Such compounds offer important advantages over structural analogues: there is the potential to move away from the patent space of the query compound; and they provide the possibility of exploring more than one lead in parallel with clear advantages should one series fail due to poor ADME properties or difficult chemistry.

Various graph reduction schemes have been developed for similarity searching. In this context the challenge is to reduce structures so that their pharmacophoric features are highlighted to enable compounds that share the same activity but belong to different chemical series to be perceived as similar. Figure 3 shows a series of compounds that are active at opioid receptors. The similarities of each of codeine, heroin and methadone to morphine are shown based on Daylight fingerprints (***14***) (a conventional 2D fingerprint) and the Tanimoto coefficient. The obvious 2D structural similarities of codeine and heroin to morphine are reflected in the high scores. However, methadone scores poorly despite having similar activity. The shaded spheres indicate a mapping between the structures that is based on their common functional groups and reveals similarities between methadone and the other compounds which are not evident using conventional 2D fingerprints. When used for similarity searching the aim of the graph reduction approach is to recognise such mappings so that the resulting reduced graphs can be thought of as topological pharmacophores.

Varying the level of specificity

Different graph reduction and node labelling schemes have been devised that vary in the level of specificity that is encoded and therefore in the degree of discrimination that is achieved between different structures. Figure 4 shows four levels of node specificity for a reduction scheme based on three node types: Rings, Features and Linkers. In this scheme, linkers and features are distinguished using the concept of isolated carbons which are acyclic carbon atoms that are not doubly or triply bonded to a heteroatom (***14***). Connected isolated carbon atoms form linker nodes with the remaining connected acyclic components defining feature nodes. Non-hydrogen bonding terminal atoms are removed, as indicated by the exclusion of the terminal methyl groups in structure A, Figure 5. The different levels in the hierarchy are derived by further describing the nodes according to the properties of their constituent atoms in terms of aromaticity and hydrogen bonding character. As the level of detail encoded within the nodes is increased the number of unique reduced graphs that are represented in a database increases, see Figure 5. In experiments on the World Drug Index database, Gillet et al. determined that reduced graphs at level four in the hierarchy were most effective in discriminating between actives and inactives (***15***).

Variations on this basic approach have since been described which include the definition of additional nodes types such as positively and negatively ionisable groups. Flexibility in the definition of node types is generally achieved through the use of user-defined SMARTS definitions for various groups such as hydrogen bond donors and acceptors.

Comparing reduced graphs using fingerprints

Various approaches have been devised to enable the similarity between a pair of molecules to be calculated based on their reduced graph representations. In analogy with the use of fragment bitstrings to compare chemical graphs, a similar approach has been taken to represent reduced graphs as binary vectors. For example, a mapping of node types to atoms not in the usual organic set, such as the transition metals, allows the reduced graphs to be represents as SMILES strings, as shown in Figure 6, and the Daylight fingerprinting routines to be used to generate path-based fingerprints from reduced graphs (***15***). While this approach provided a convenient way of comparing reduced graphs, the different characteristics of reduced graphs, relative to the structures from which they are derived, are such that the resulting fingerprints are suboptimal for quantifying the similarity between reduced graphs. For example, reduced graphs consist of fewer nodes than their corresponding chemical graphs so that the resulting fingerprints can be quite sparse and small changes in a chemical structure, such as the insertion of a heteroatom into an acyclic chain, can result in a quite different set of nodes and therefore fingerprint.

Improved performance was obtained by representing the reduced graphs as node-pair descriptors (***16***), which are similar in concept to the more familiar atom-pair descriptors developed by Carhart et al. (***11***). For example, Harper and colleagues developed fingerprints based on node-edge pairs in which additional bits are set, for example, to encode branch points so that more of the topology of the reduced graph is represented and to encode paths of length one shorter than the actual length to introduce a fuzziness to the fingerprint (***17***). The "fuzzy bits" enable the similarity between RGs that differ by the insertion or deletion of a single node to be perceived, which would otherwise give rise to a set of node-edge pairs of different lengths.

Harper et al. also developed an edit distance method to quantify the similarity between reduced graphs which is based on the cost of converting one reduced graph to the other by considering mutation, insertion and deletion of nodes. The edit distance technique is well known in computational biology where it is used for sequence comparisons with similarity related to the number of operations required to change one sequence to another. In the context of reduced graphs, edit distance is well suited to dealing with the problem of small changes in chemical structure leading to different patterns of nodes, for example, by the insertion of a heteroatom into a carbon chain. Furthermore, different weights can be assigned to different node operations to reflect similarities in node types. For example, in Harper's work the substitution of a "donor" to a "donor & acceptor" node was assigned a low cost, whereas, the substitution of a "donor" to a "negatively ionisable group" was assigned a high cost. Harper showed that combining the edit distance similarity measure with a node-pair fingerprinting method improved the performance of the reduced graphs in similarity searches compared to the path-based fingerprints. The edit distance method is illustrated in Figure 7: the left hand side shows the minimum cost of converting reduced graph B into A based on the matrix of substitution costs and the insertion/deletion costs shown on the right.

The costs assigned to the individual node operations by Harper were based on intuition. Subsequently, Birchall et al. (***18***) used a genetic algorithm to identify optimised sets of weights that gave improved performance over a variety of activity classes extracted from the MDL/Symyx Drug Data Report (MDDR) database (***19***). They also generated sets of weights optimised on specific activity classes and showed that class specific weights could not only improve retrieval performance but could also provide some clues on the underlying structure-activity relationship.

Comparing reduced graphs using graph matching procedures

By definition, reduced graphs contain fewer nodes and edges than the chemical graphs from which they are derived, making them more amenable to graph matching procedures. Takahashi et al. described an early approach to the use of graph matching techniques to compare reduced graph representations, albeit based on a very small number of compounds (***20***). They considered a set of five structurally diverse antihistamines and a set of six antipsychotropic agents, and in both cases, some of the structural similarities were found. In more recent work, Barker et al. represented the reduced graph as a fully connected graph in which the edges represent bond distances in the original chemical graph and used maximum common subgraph techniques to calculate the similarity between pairs of reduced graphs using much larger datasets (***21***). They demonstrated improved performance of the reduced graph relative to Daylight fingerprints both in terms of the recall of actives and in the diversity of the actives retrieved thus suggesting that reduced graphs might be beneficial in scaffold hopping applications.

Clustering

The reduced graph approach has also been used for various clustering applications. Clustering is widely used to present sets of compounds to chemists for review, for example, typically the results from a high-throughput screening exercise will be clustered and clusters that are enriched in active compounds will be examined in an attempt to extract structure-activity information. The most commonly used clustering techniques are based on traditional 2D fingerprints that are derived from the chemical structures themselves, however, when using such fingerprints, it may be difficult to decipher the structural commonalities that are present within a cluster. Harper et al. used reduced graphs to cluster high throughput screening data **(*17*)**. Each molecule is represented by several *motifs* which include the reduced graph, near neighbours of the reduced graph in which single nodes are deleted or changed, and Bemis and Murcko frameworks **(*22*)**. Molecules that share a common motif are clustered together and the clusters are sorted with large clusters consisting predominantly of active compounds being presented to the user first. The reduced graphs and frameworks allow the structural characteristics of the compounds to be easily seen, in contrast to clustering based on conventional fingerprints.

In related work, Gardiner et al. have used reduced graphs to identify cluster representatives **(*23*)**. Here a dataset is clustered using conventional 2D fingerprints, the members of a cluster are then represented as reduced graphs and a maximum common subgraph (MCS) algorithm is applied iteratively in order to obtain one or more reduced graph cluster representatives. The reduced graphs offer two advantages for this application: first, their small size means that the MCS comparisons can be run in real-time; and second, the cluster representatives can be mapped back to the original structures that they represent, allowing the chemists to interpret the key functionalities required for activity. The method also enables multiple series present within the same cluster to be identified as well as related clusters by comparing representatives from different clusters.

Reduced graphs for identifying SARs

Reduced graphs have been used in conjunction with recursive partitioning in order to derive structure-activity relationship models. As proof of principle, an SAR model was developed for angiotensin II receptor antagonists and compared with the known literature (***16***). A fingerprint representation of the reduced graph was used to determine the splitting criteria in a decision tree based on a training set of 100 actives and 2000 inactives extracted from the MDDR database (***19***). A portion of the resulting tree is shown in Figure 8 with the shaded box highly enriched in actives and containing 70 of the active compounds. The splits in the tree are based on the presence/absence of node-edge pairs: Ar_d-2-Ar_n represents an aromatic ring containing a hydrogen bond donor separated by two edges from and aromatic ring with no hydrogen bonding character; Ar_n-1-Ar_n represents two aromatic rings with no hydrogen bonding characteristics separated by a single edge. These two node-edge pairs can be combined to represent the substructure A which compares well with the 2D SAR model for angiotensin II receptor antagonists described by Bradbury et al. (***24***). The approach was subsequently used in a procedure to select compounds for screening against a kinase inhibition assay with a hit rate of around 7% reported.

A disadvantage of the use of fingerprint representations to represent SARs is the loss of information on how the node-edge pairs are connected. For example, substructure A in Figure 8 represents one way in which the node-edge pairs could be combined, however, there are other arrangements of rings that are also consistent with the same set of node-edge pairs, for example, substructure B. More recently, Birchall et al. have developed an evolutionary algorithm (EA) to grow reduced graph queries (subgraphs) with the aim of discriminating between actives and inactives in high throughput screening data (***25***). The reduced graph queries are encoded as SMARTs strings (such as that shown in Figure 9) and allow a more detailed description of the structure-activity relationship to be developed. For example, a query can consist of any number of connected (or even disconnected nodes). Moreover, the use of atom primitives in the SMARTs language (such as OR and NOT logic) enables the range of substructures that can be captured in a single expression to be extended. For example, a series of alternative node types can be specified at a given location in a subgraph to allow expressions such as "non-feature ring node OR acceptor ring node". The SMARTS expressions are mapped to tree-based chromosomes with the primitives tagged to nodes as shown in Figure 9. Tree-based evolutionary operators have been developed to enable new trees to be evolved through the exchange of subtrees between chromosomes and various mutation operators.

A chromosome is evaluated by parsing the tree to generate a SMARTS query which is then searched across a training set of actives and inactives, also represented as reduced graphs. Fitness is measured using the F-measure which is the harmonic mean of precision (P), the ratio of actives to total compounds retrieved, and recall (R), the fraction of the actives retrieved, as follows: $F= 2PR/(P+R)$. The EA has been configured to evolve reduced graph queries that maximise the F-measure.

When applied to various activity classes extracted from the MDDR database (***19***) the EA was able to evolve reduced graph queries that give good classification rates and which encode structure-activity information that is readily interpreted by chemists. The approach was subsequently extended to first, explore trade-offs in recall and precision and second, to allow multiple SARs to be extracted from a single activity class (***26***). The rationale for exploring the trade-off between precision and recall is that the optimum balance between these two objectives may depend on the application. For example, when seeking a structure activity model it may be of interest to evolve a query with high precision at the expense of relatively low recall. Conversely, when evolving a query to be used in virtual screening it may be more appropriate to choose a query that has higher recall but lower precision or even to choose a query that returns the same number of hits as the screening capacity. By treating recall and precision as independent objectives in a multiobjective optimisation procedure, a range of solutions are found which vary from high recall-low precision queries through to low recall-high precision queries. Multiple queries are evolved through the introduction of a third objective, called uniqueness, which compares each query with all others in the population. A query receives a high uniqueness score if the actives that it retrieves are not found by other queries in the population. This enabled multiple SARs to be derived where each SAR described a different set of active compounds. The combination of these complementary SARs allows for improved recall and precision as well as increasing the level of detail in the overall SAR description of a given activity class.

Reduced graphs for encoding bioisosteres

Bioisosteres are structural fragments that can be exchanged without significant change to a molecule's biological activity. Since bioisosteres may be quite different in structure, e.g. tetrazole and carboxylic acid, it is challenging for conventional similarity measures to reflect their functional similarity. Graph reduction approaches are an attractive means of dealing with such equivalences as they allow several different structures to be encoded as the same node type. Birchall et al (***27***) investigated how bioisostere information could be exploited in similarity searching using a graph-matching approach. Bioisosteres extracted from the BIOSTER database (***28***) were often found to be encoded by the same node type, supporting the applicability of the reduced graph encoding. However, there were also many cases where the bioisosteric fragments were not encoded as the same node type or even by the same number of nodes. The graph reduction and matching schemes were then modified to recognise and permit matches between instances of bioisosteric fragments, enhancing the similarity between molecules containing such fragments. Similarity searches in the WOMBAT database (***29***) found that although this approach clearly demonstrates scaffold hopping potential, there is a significant trade-off in terms of the number of inactives that are also retrieved. The issue here arises from the fact that bioisosteric equivalences are often dependent on the specific context in which they are considered, both in terms of the intra-molecular environment and the extra-molecular environment, something that is perhaps too complex for broad generalisation based on the available data. By altering the rules used for graph partitioning, node type assignment and node type matching, reduced graphs provide the flexibility to allow the recognition of increasingly structurally distinct equivalences. However, this must be balanced against the degree of information loss inherent in graph reduction that may lead to the recognition of unreasonable equivalences. The key is in deciding what constitutes a reasonable equivalence.

Related Approaches

The intention of this chapter has been to summarise the extensive work carried out on reduced graphs at Sheffield, however, in acknowledgement of the significant contributions made by other groups we briefly summarise the closely related approaches of feature trees and ErG. The feature tree, developed by Rarey and Dixon, also seeks to generalise chemical structures by emphasising their functional features (***2***). A ring/non-ring reduction similar to that in Figure 1a is carried out except that a separate node is assigned to each non-terminal acyclic atom. The resulting structure is a tree (ie it does not contain any cycles) which allows significant improvements in speed when comparing two feature trees due to the greater efficiency of tree-matching algorithms relative to graph-matching. Each node in the tree is "labelled" with a range of *features* derived from its constituent atom(s) such as their volume and molecular interaction capabilities. Calculating the similarity between two trees is based on first finding a match of sub-trees and then using a weighted combination of the feature similarities of the matching nodes. Feature trees of a lower specificity can be derived by collapsing sub-trees into single nodes to give rise to a hierarchy of representations which allows similarities to be determined at varying levels of specificity. Feature trees have been applied to a number of applications including: similarity searching based on a single query (***2***); similarity searching based on multiple queries by combining the queries into a multiple feature tree model called MTree (***30***); and fast similarity searching in very large combinatorial libraries (***3***).

The extended reduced graph (ErG) approach developed by Stiefl et al. (***31***) is similar to the reduced graph but includes a number of extensions. For example, charged and hydrophobic features are encoded explicitly and rings are encoded as ring centroids with substituted ring atoms encoded as separate nodes. The nodes are connected according to the shortest paths in the chemical graph. Although the ErG is a more complex graph than the reduced graph, positional information is better conserved and inter-feature distances in the original molecule tend to be more accurately represented. Furthermore, separation of the ring features from the ring itself permits similarity to be reflected between rings of different feature types. For example, while the reduced graph encodes pyrrole and phenyl rings as different node types, the ErG approach represents pyrrole as an aromatic node joined to a donor node which retains some commonality with the single aromatic node resulting from a phenyl group. The ErG can be encoded in a fingerprint, similar to those developed for reduced graphs, however, Stiefl et al. use a hologram approach where each bit encodes a count of the fragment frequency rather than binary presence or absence. Some fuzziness in matching is permitted by incrementing the bits for paths that are both longer and shorter than the path-length. In simulated virtual screening experiments across a range of activity classes, ErG was found to be comparable to Daylight fingerprints of chemical graphs, in terms of enrichments, however, they were found to be more effective for scaffold hopping since a greater diversity of structural classes were found. Stiefl and Zaliani (***32***) also describe an extension of ErG in which a weighting scheme is used to increase the significance of specified features. They demonstrated improved performance compared to the unweighted method, however, this approach is dependent on the availability of experimental data to identify the significant features.

Conclusions

Reduced graphs provide flexible ways of generalising molecular structures while retaining the topology of the original structures. They have proved to be useful for a number of different applications with the optimal graph reduction scheme being dependent on the particular application. For example, for the representation and search of Markush structures a simple ring/non-ring reduction permits the encoding of generic nomenclature expressions into single nodes which enables one of the difficulties of handling these structures to be overcome. In applications that aim to identify structure-activity relationship, more complex graph reduction schemes are usually required so that pharmacophoric groups can be identified. It is usually possible to allow the definitions of pharmacophoric features to be determined at run time through the use of SMARTS representations of features such as hydrogen bond donors, hydrogen bond acceptors and ionisable groups. This allows different properties to be emphasised in different applications. Reduced graphs enable similarities to be perceived between heterogeneous compounds which is beneficial for scaffold hopping applications and for the capture of SARs from structurally diverse compounds. Furthermore, the small size of the reduced graphs relative to the structures from which they are derived permits the use of graph matching algorithms so that mappings between structures can be generated which assists in interpreting the results of similarity and SAR analyses.

References

1. Gillet, V. J., Downs, G. M., Ling, A., Lynch, M. F., Venkataram, P., Wood, J. V., and Dethlefsen, W. (1987) Computer-storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical-structure retrieval, *Journal of Chemical Information and Computer Sciences* ***27***, 126-137.
2. Rarey, M., and Dixon, J. S. (1998) Feature trees: A new molecular similarity measure based on tree matching, *Journal of Computer-Aided Molecular Design* ***12***, 471-490.
3. Rarey, M., and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces, *Journal of Computer-Aided Molecular Design* ***15***, 497-520.
4. Stiefl, N., Watson, I. A., Baumann, K., and Zaliani, A. (2006) ErG: 2D pharmacophore descriptions for scaffold hopping, *Journal of Chemical Information and Modeling* ***46***, 208-220.
5. Stiefl, N., and Zaliani, A. (2006) A knowledge-based weighting approach to ligand-based virtual screening, *Journal of Chemical Information and Modeling* ***46***, 587-596.
6. Gillet, V. J., Downs, G. M., Holliday, J. D., Lynch, M. F., and Dethlefsen, W. (1991) Computer-storage and retrieval of generic chemical structures in patents. 13. Reduced-graph generation, *Journal of Chemical Information and Computer Sciences* ***31***, 260-270.
7. Lynch, M. F., and Holliday, J. D. (1996) The Sheffield Generic Structures Project - A retrospective review, *Journal of Chemical Information and Computer Sciences* ***36***, 930-936.
8. Shenton, K., Nortin, P., and Fearns, E. A. (1988) Generic Searching of Patent Information, in *Chemical Structures - The International Language of Chemistry* (Warr, W., Ed.), pp 169-178, Springer, Berlin.
9. Fisanick, W. (1990) The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. Part 1. Basic Concepts, *Journal of Chemical Information and Computer Sciences* ***30***, 145-154.
10. Ebe, T., Sanderson, K. A., and Wilson, P. S. (1991) The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. Part 2. The MARPAT File, *Journal of Chemical Information and Computer Sciences* ***31***, 31-36.
11. Carhart, R. E., Smith, D. H., and 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.
12. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of nearest-neighbor searching in an online chemical structure search system, *Journal of Chemical Information and Computer Sciences* ***26***, 36-41.
13. Brown, N., and Jacoby, E. (2006) On scaffolds and hopping in medicinal chemistry, *Mini-Reviews in Medicinal Chemistry* ***6***, 1217-1229.
14. Daylight. Daylight Chemical Information Systems, Inc., 120 Vantis - Suite 550, Aliso Viejo, CA 92656, USA. www.daylight.com at http://www.daylight.com.
15. Gillet, V. J., Willett, P., and Bradshaw, J. (2003) Similarity searching using reduced graphs, *Journal of Chemical Information and Computer Sciences* ***43***, 338-345.
16. Barker, E. J., Gardiner, E. J., Gillet, V. J., Kitts, P., and Morris, J. (2003) Further development of reduced graphs for identifying bioactive compounds, *Journal of Chemical Information and Computer Sciences* ***43***, 346-356.
17. Harper, G., Bravi, G. S., Pickett, S. D., Hussain, J., and Green, D. V. S. (2004) The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data, *Journal of Chemical Information and Computer Sciences* ***44***, 2145-2156.
18. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2006) Training similarity measures for specific activities: Application to reduced graphs, *Journal of Chemical Information and Modeling* ***46***, 577-586.
19. MDDR. Symyx Technologies Inc, 2440 Camino Ramon, Suite 300, San Ramon, CA 94583. http://www.symyx.com.

20. Takahashi, Y., Sukekawa, M., and Sasaki, S. (1992) Automatic identification of molecular similarity using reduced graph representation of chemcial structure, *Journal of Chemical Information and Computer Sciences* ***32***, 639-643.

21. Barker, E. J., Cosgrove, D. A., Gardiner, E. J., Gillet, V. J., Kitts, P., and Willett, P. (2006) Scaffold-Hopping Using Clique Detection Applied to Reduced Graphs, *Journal of Chemical Information and Modeling* ***46***, 503-511.

22. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs.1. Molecular frameworks, *Journal of Medicinal Chemistry* ***39***, 2887-2893.

23. Gardiner, E. J., Gillet, V. J., Willett, P., and Cosgrove, D. A. (2007) Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs, *Journal of Chemical Information and Modeling* ***47***, 354-366.

24. Bradbury, R. H., Allott, C. P., Dennis, M., Fisher, E., Major, J. S., Masek, B. B., Oldham, A. A., Pearce, R. J., Rankine, N., Revill, J. M., Roberts, D. A., and Russell, S. T. (1992) New nonpeptide angiotensin-II receptor antagonists .2. Synthesis, biological properties, and structure-activity relationships of 2-alkyl-4-(biphenylmethoxy)quinoline derivatives, *Journal of Medicinal Chemistry* ***35***, 4027-4038.

25. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2008) Evolving interpretable structure - Activity relationships. 1. Reduced graph queries, *Journal of Chemical Information and Modeling* ***48***, 1543-1557.

26. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2008) Evolving interpretable structure - Activity relationship models. 2. Using multiobjective optimization to derive multiple models, *Journal of Chemical Information and Modeling* ***48***, 1558-1570.

27. Birchall, K., Gillet, V. J., Willett, P., Ducrot, P., and Luttmann, C. (2009) Use of Reduced Graphs To Encode Bioisosterism for Similarity-Based Virtual Screening, *Journal of Chemical Information and Modeling* ***49***, 1330-1346.

28. Ujvary, I. (1997) BIOSTER: A database of structurally analogous compounds, *Pesticide Science* ***51***, 92-95.

29. WOMBAT. Sunset Molecular. Available at http://www.sunsetmolecular.com/.

30. Hessler, G., Zimmermann, M., Matter, H., Evers, A., Naumann, T., Lengauer, T., and Rarey, M. (2005) Multiple-ligand-based virtual screening: Methods and applications of the MTree approach, *Journal of Medicinal Chemistry* ***48***, 6575-6584.

31. Stiefl, N., Watson, I. A., Baumann, K., and Zaliani, A. (2006) ErG: 2D Pharmacophore Descriptions for Scaffold Hopping, *Journal of Chemical Information and Modeling* ***46***, 208-220.

32. Stiefl, N., and Zaliani, A. (2006) A Knowledge-based Weighting Approach to Ligand-based Virtual Screening, *Journal of Chemical Information and Modeling* ***46***, 587-596.

Figure Legends

Figure 1. Different graph reduction schemes. a) A ring/non-ring reduction where a fused ring system is reduced to a single node. b*)* A ring/non-ring reduction where each smallest ring is treated as an individual node. c) A carbon/heteroatom reduction. d) A homeomorphic reduction in which atoms of degree two are removed. The node types are denoted as follows: R: ring; N: non-ring; C: carbon; and H: heteroatom.

Figure 2. A Markush structure and its reduced graph representation based on a ring/non-ring reduction scheme.

Figure 3. The similarities of codeine, heroin and methadone are shown to morphine based on Daylight fingerprints and the Tanimoto coefficient.

Figure 4. A hierarchy of reduced graphs.

Figure 5. The reduced graph for compound A at each level in the hierarchy in Figure 4 is shown together with a series of related compounds: B to F. At level one, all compounds are represented by the same reduced graph, at level 2, compounds A to E share the same reduced graph through to level 4 where only compounds A to C share the same reduced graph. The discrimination between structures is dependent on the level of descriptions encoded within the reduced graph.

Figure 6. A reduced graph represented as a SMILES string. Note that terminal, non-hydrogen bonding atoms have been removed when forming the reduced graph and that the fused ring nodes are represented by the "=" symbol.

Figure 7. The edit distance cost of converting the pattern of nodes in A to B is 3 based on the substitution cost matrix and insertion/deletion costs shown on the right.

Figure 8. A decision tree generated for angiotensin II antagonists based on reduced graph representations. Substructure A is consistent with the node-edge pairs in the shaded box and is consistent with the known SAR. However, a limitation of the use of node-edge pairs for this

application is that these two node-edges pairs are also present in other substructures, such as, B which may not be relevant to the SAR.

Figure 9. A reduced graph query is shown as a SMARTS string in the centre. The left-hand side shows how the SMARTS string is mapped to a tree-based chromosome. The SMARTS primitives are tagged to nodes in the chromosome: D1 indicates degree 1; AND indicates a disconnected node (shown as "." in the SMARTS); RF indicates a ring fusion which is represented by a double bond in the SMARTS string. Two nodes are grouped to indicate that they represent alternative nodes. The right-hand side shows a molecule that matches the query with the nodes corresponding to the query highlighted.