Advanced Review

Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review

Hans-Christian Ehrlich and Matthias Rarey*

The intuitive description of small and large molecules using graphs has led to an increasing interest in the application of graph concepts for describing, analyzing, and comparing small molecules as well as proteins. Graph theory is a well-studied field and many applications are present in various scientific disciplines. Recent literature describes a number of successful applications to biological problems. One of the most applied concepts aims at finding a maximal common subgraph (MCS) isomorphism between two graphs. We review exact MCS algorithms, especially designed for graphs obtained from small and large molecules, and give an overview of their successful applications. © 2011 John Wiley & Sons, Ltd. *WIREs Comput Mol Sci* 2011 1 68–79 DOI: 10.1002/wcms.5

INTRODUCTION

Chemical database systems are challenged with the task of managing a rising number of molecular entries.[1,2] Especially, the fast and efficient storage and retrieval of the database entries must be ensured. This requires a molecular description based on a sophisticated chemical model. Depending on the chemical question to be addressed, different molecular representations ranging from simple descriptions of physicochemical properties[3] over binary fingerprints[4,5] to graphs and reduced graphs[6–8] are available. Modeling molecules as labeled graphs have a long tradition[9] and is a prerequisite for most modern cheminformatic methods. The representation of molecules by graphs has two major advantages: Graphs are a very intuitive molecular representation close to our elementary chemical understanding, and they form a solid theoretical basis for computer-aided processing. Furthermore, graphs enable a database retrieval via graph isomorphism techniques, i.e., comparing molecules becomes equivalent to comparing labeled graphs. This review focuses on molecular graph comparison techniques, especially addressing the MCS problem. We introduce the graph theoretical background and summarize algorithms solving the MCS problem. Finally, we provide an overview of scientific applications that utilize MCS algorithms.

*Correspondence to: rarey@zbh.uni-hamburg.de

Center for Bioinformatics, Computational Molecular Design, Hamburg, Germany

DOI: 10.1002/wcms.5

PRELIMINARIES

Around 1860, Kekule introduced a structural formula, which is the foundation of modern chemistry. The structural formula is a graph-like representation of molecules commonly used to formulate and exchange chemical knowledge. The formula allows chemists to visualize molecules and quickly identify communalities and differences between them.

Graph Theoretical Background

A graph G is a pair $(V, E)$of vertices and edges. Each edge $e \in E$ connects two adjacent vertices $(v_1, v_2) \in V$. In a labeled graph, vertices (and/or edges) hold arbitrary labels. A graph is simple if each of its edges is undirected and unweighted. Undirected edges have no orientation between the vertices they connect. Unweighted edges have a uniform weight assigned to them. In the following, we only consider simple graphs, except when stated otherwise.

Two graphs G_1 and G_2 are isomorphic if there exists a bijective (one-to-one) mapping between the vertices of G_1 and G_2 such that two vertices in G_1 are connected by an edge, if and only if the corresponding images in G_2 are connected. An induced subgraph

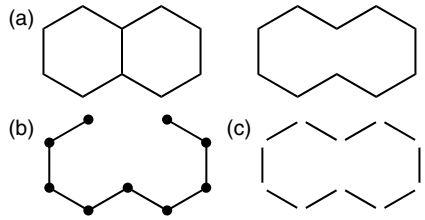

FIGURE 1 | Maximal common induced subgraph (MCIS) versus maximal common edge subgraph (MCES). (a) Molecular graph of decalin (labels not shown). (b) Molecular graph of cyclodecane (labels not shown). (c) MCIS of (a) and (b). (d) MCES of (a) and (b).

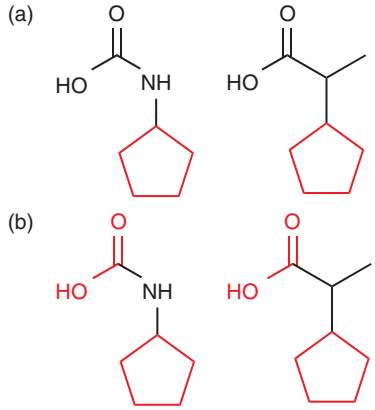

FIGURE 2 | Connected versus disconnected maximal common subgraph (MCS). (a) Connected MCS (red). (b) Disconnected MCS (red).

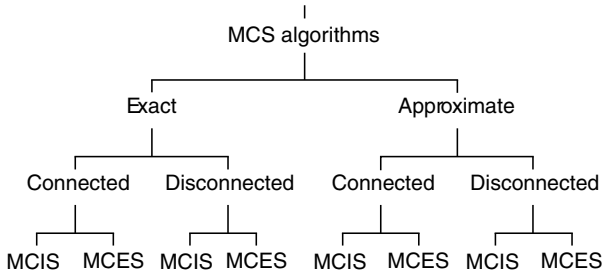

FIGURE 3 | Classification of maximal common subgraph (MCS) algorithms.

of G consists of a subset of vertices $V' \subset V$ and the subset of all edges $E' \subset E$ connecting vertices in V'. An induced subgraph isomorphism exists if G_1 is a subgraph of G_2 (or vice versa), i.e., G_1 is contained in G_2. Finally, a common induced subgraph of two graphs G_1 and G_2 is a graph G_{12} that is isomorphic to a subgraph of G_1 and a subgraph of G_2. Although there are possibly many common subgraphs between two graphs, we will focus on the largest common induced subgraph or maximal common induced subgraph (MCIS). Related to the MCIS is the maximal common edge subgraph (MCES). The MCES is a subgraph with the maximal number of edges common to both G_1 and G_2. Figure 1 shows the difference between MCIS and MCES. Note that the MCIS as well as the MCES of two graphs is not necessarily unique. We will use the term MCS to refer to both, the MCIS as well as the MCES.

Both MCS types can be connected or disconnected. In a connected MCS, each vertex is reachable from every other vertex by a path through the MCS. A disconnected MCS is composed of two or more disconnected components. Figure 2 illustrates the connected and disconnected MCS for the same molecular graph.

The complete MCS algorithm classification scheme is illustrated in Figure 3.

Molecular Representation and Comparison

It is quite obvious that the atoms of a molecule can be easily represented by vertices and bonds by edges. The resulting molecular graph is often labeled to account for atom and bond properties. The degree or number of edges a vertex can have is limited by the number of covalent bonds an atom can form. Therefore, the number of edges linearly depends on the number of atoms. To represent the orientation of atoms in space, it is possible to add three-dimensional (3D) information to a molecular graph.[10,11] However, the graphs considered herein refer to the molecular topology only, except when stated otherwise.

Molecules are considered equal if a one-to-one mapping between all atoms and bonds exists, i.e., the two molecular graphs are isomorphic. To map a pair of atoms or bonds, their labels must be identical. In the case that two molecules are not exactly the same, one molecule can be a substructure of the other. Then, a subgraph isomorphism between the two molecular graphs exists. Alternatively, two molecules share a common substructure, and therefore the molecular graphs share a common subgraph.

Some problems arise when using molecular graphs. In mesomeric structures, e.g., aromatic compounds, different bond localization result in nonisomorphic labeled graphs, although the structures represent the same molecule. For stereoisomeric molecules, additional information, e.g., the relative arrangement of bonds, must be annotated to differentiate between them. Moreover, as molecules exist in potentially many tautomeric forms, the construction of their molecular graphs in a standardized form becomes especially challenging.

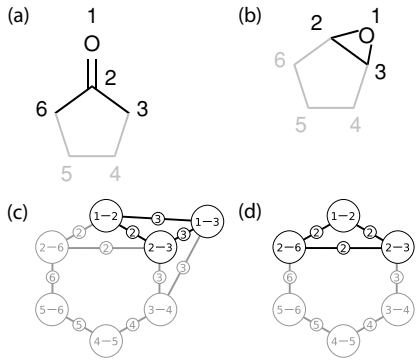

FIGURE 4 | Line graphs. Panels (a) and (b) show the molecular graph of cyclopentanone and epoxycylopentane. The highlighted subgraphs in (a) and (b) become topological equivalent in the corresponding line graphs (c) and (d). A differentiation is only possible by their vertex and edge labels.

Relation between MCIS and MCES

Most of the algorithms described in literature calculate the MCIS and only a few calculate the MCES directly. However, Whitney[12] proved that in cases without trinode/triangle subgraphs, an MCIS can be converted into an MCES. First, the molecular graphs are converted into so-called line graphs that represent the adjacency between edges, which can then in turn be converted in a compatibility graph. Figure 4 shows two molecular graphs: the corresponding line graphs and a trinode/triangle example. The details, especially the trinode/triangle problem, are discussed by Raymond and Willett.[13]

ALGORITHMS

The problem of computing an MCS between two graphs is NP-hard,[14] meaning that no polynomial time algorithm exists (unless P = NP). Nevertheless, many attempts to obtain algorithms useful in practice have been made and most of them are present in the field of computer vision and image recognition.[15] Here, we focus on recent MCS algorithms in the field of molecular science.

To obtain a clear classification of MCS algorithms, we adopt the scheme from Raymond and Willett[13] that differentiates between algorithms calculating exact or approximate, connected or disconnected, vertex-based (MCIS) or edge-based (MCES) solutions. Unfortunately, some published algorithms are not described in enough detail for a clear classification. Especially, the term MCS is often used as a synonym for both, the MCIS and MCES. Therefore, the algorithmic description leaves room for interpretation and an adequate classification becomes difficult.

Maximal Clique-based Algorithms

Calculation of an MCS between two graphs can be reduced to the problem of finding the maximal clique in a compatibility graph. A clique of a graph is a complete subgraph in which each vertex is directly connected to every other vertex. A maximal clique is, therefore, a complete subgraph with the largest possible number of vertices. Note that a graph can incorporate more than one maximal clique. A compatibility graph, also known as association graph,[16,17] modular product graph,[18] or correspondence graph,[19] of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is defined as the vertex set $V_1 \times V_2$ in which two vertices (v_1, v_2) and (u_1, u_2) are adjacent, if and only if $(v_1u_1) \in E_1$and $(v_2u_2) \in E_2$ or $(v_1u_1) \notin E_1$ and $(v_2u_2) \notin E_2$. For molecular graphs, the compatibility between two vertices or edges is additionally restricted by their labels. The labels must agree according to some compatibility criteria, e.g., the same atom types or bond orders. A maximal clique of a compatibility graph corresponds to the MCIS of the two original graphs. Figure 5 shows two molecular graphs: the compatibility graph and the correspondence between the maximal clique of the compatibility graph and the MCIS.

The approach to reduce the MCS problem to the maximal clique problem is already known for some time[18,20,21] and one of the first applications to chemical structures is described by Kuhl et al.[22] and Brint and Willett.[19] The literature describes many different clique-detection algorithms,[23–27] and Gardiner et al.[28] analyzed the performance of the most common ones. Two widely used method for arbitrary graphs are the algorithms by Bron and Kerbosch[23] and Carraghan and Pardalos.[25] Although chemical graphs are in general sparse, their compatibility graphs tend to be dense. Therefore, the general clique-detection algorithms, which do not use any chemical

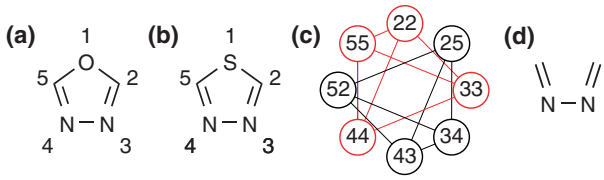

FIGURE 5 | Maximal clique in a compatibility graph. (a) Molecular graph of 1,3,4-oxadiazole. (b) Molecular graph of 1,3,4-thiadiazole. (c) Compatibility graph of (a) and (b). The graph has two maximal cliques indicated with red and black lines. (d) Maximal common subgraph (MCS) of 1,3,4-oxadiazole and 1,3,4-thiadiazole. Both cliques resemble the same MCS.

information, are often slow. Raymond et al.[29,30] developed rapid similarity calculation (RASCAL), a branch-and-bound procedure that uses multiple chemically motivated heuristic strategies to improve the efficiency of clique detection on molecular graphs. RASCAL calculates the exact, disconnected MCES by converting the MCES problem into the MCIS problem.[12] In RASCAL, the two input graphs are transformed into line graphs from which the compatibility graph is constructed. A maximal clique in that compatibility graph corresponds to an MCES because each vertex resembles an edge. The RASCAL procedure was adapted to calculate the exact, disconnected MCIS and MCES on a reduced version of molecular graphs.[31] In a reduced molecular graph, a vertex no longer resembles a single atom but rather a group of atoms, e.g., a functional group. A recent developed branch-and-bound method[32] based on the clique-detection method of Carraghan and Pardalos[25] is described to detect all exact MCESs. Another branch-and-bound clique-detection algorithm[33] is an extension of Mehlhorn's algorithm[34] and uses a compatibility graph with a weakened edge definition such that two vertices (v_1, v_2) and (u_1, u_2) are adjacent, if and only if $(v_1u_1) \in E_1$ or $(v_1u_1) \notin E_1$ and $(v_2u_2) \notin E_2$. The result is an exact, disconnected MCS, in which two vertices can be adjacent in the first graph and nonadjacent in the second. Therefore, a circular structure can be matched to a linear one.

The computing time of clique-detection algorithms increases exponentially with the number of vertices and edges in the compatibility graph. The above-mentioned concept of chemical labeling, e.g., with atom types[35] or bond orders,[32] reduces the number of vertices and edges substantially. Often, only because of chemical labeling, the MCS calculation becomes feasible for an application in molecular sciences. The methods described so far calculate the disconnected MCS; however, most clique-based MCS algorithms can be modified to calculate the connected MCS.[36,37]

Backtracking Algorithms

An MCS of two graphs is usually represented by a bijective mapping of a subset of vertices of the first graph to a subset of vertices of the second. If this mapping is built-up sequentially vertex by vertex, a tree structure is defined that can be searched by classical backtracking algorithms (Figure 6 shows an example). During the traversal, a current subsolution is gradually extended to guide the search toward the final solution. When an extension does not lead to a valid or better solution, the underlying branch of the search tree gets pruned; therefore, reducing the number of backtracking iterations needed to find the MCS.

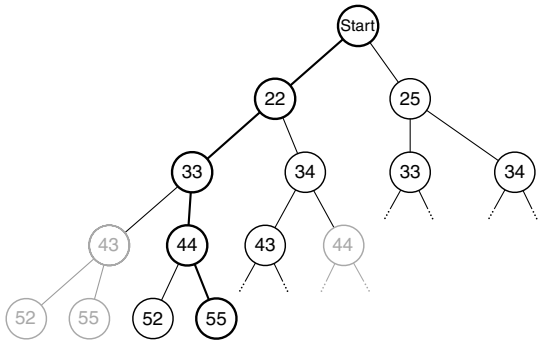

FIGURE 6 | Backtracking search tree for 1,3,4-oxadiazole and 1,3,4-thiadiazole as shown in Figure 5(a) and (b). One solution is highlighted. Cut branches are shown in gray.

Two historical backtracking methods were introduced by McGregor[38] and Ullmann.[39] McGregor[38] appears to be the first who draws a difference between an MCIS and MCES. The Ullmann[39] algorithm calculates subgraph isomorphism rather than an MCS but is worth mentioning because it was the fastest algorithm at its time and is the basis for other subgraph isomorphism algorithms.[40] Additionally, the algorithms and its variations are widely used in today's chemical substructure search systems.[41]

Most recently, Cao et al.[42] presented a backtracking procedure that calculates the exact, disconnected MCIS with control over the number of disconnected components. The algorithm works directly on the molecular graphs and makes use of multiple strategies to prune the search tree. One pruning strategy excludes extensions that exceed the allowed number of disconnected components. Another, the induced subgraph heuristic, is derived from the definition of an induced subgraph and identifies infeasible sets of vertex mappings in which the vertices do not lead to an edge-compatible vertex assignment. To further reduce the search tree, the algorithm applies a branch-and-bound strategy that uses a current suboptimal solution to calculate an estimate of the maximal possible MCIS and apply it as upper bound on the final solution. If the estimate is worse than the best solution found so far, the branch is not further explored. To find large suboptimal solutions first, the vertices are processed in decreasing order according to their number of neighbors to the current common subgraph. Therefore, the next processed vertex mostly improves the upper bound, achieving a further acceleration of the search process.

Berlo et al.[43] extended Cao et al.[42] method to calculate all exact, connected MCESs. Although the induced subgraph heuristic uses the vertex connectivity to identify vertices that do not result in a valid one-to-one mapping, it cannot be applied when calculating an MCES. However, Berlo et al.[43] use a similar vertex ordering scheme and state that their method reduces the search tree by adding multiple edges to a current solution in one step. Unfortunately, it is not proven that simultaneously adding multiple edges always gives an exact MCES. A comparison shows that Cao et al.[42] algorithms are faster up to an MCS of about 20 vertices and is then surpassed by Berlo et al.[43] method.

Dynamic Programming

Dynamic programming (DP)[44] is a long-known mathematical technique for solving multistage decision problems. The central element of DP algorithms is the hierarchical division of problems into subproblems, which are solved bottom-up storing and reusing partial solution, a technique named memorization. DP is most efficiently applied when subproblems can be solved independently from each other. Depending on the structure of subproblems, DP algorithms can achieve polynomial runtime behavior.

Because of NP-hardness of the MCS problem, it is extremely unlikely that a DP scheme results in a polynomial-time MCS algorithm. However, the MCS problem becomes easier to solve for certain graph classes. Most importantly, the MCS between two trees can be calculated in polynomial time using DP.[45] Depending on the application, it is, therefore, worthwhile to carefully analyze the molecular graphs involved. The reduced graph descriptor feature tree[7] makes use of this concept by representing molecules as trees such that an efficient algorithm for calculating the largest common subtree rather than an exponential-time MCS algorithm can be applied.

Schietgat et al.[46] developed a DP algorithm that calculates a so-called block-and-bridge preserving exact, connected MCIS. The algorithm is based on Horvarth et al.[47] and is specially designed for outerplanar graphs. Fortunately, most molecular graphs are outerplanar. When molecules are compared, it is often not desired to assign circular substructures to noncircular ones. Therefore, during the construction of an MCS, the algorithm only matches atoms of biconnected components (blocks) and edges connecting blocks (bridges) to each other. This constraint and the fact that the considered graphs are outerplanar make a polynomial runtime for solving the MCS problem possible.

Multiple MCS Isomorphism

The algorithms described so far always search for an MCS between two graphs. Finding the MCS between multiple graphs is an interesting problem when applied to molecules and has received comparatively low attention in molecular science. Nevertheless, we want to illustrate an example for multiple MCS calculation. In cheminformatics, enumerated subgraphs are frequently used as molecular descriptors. A multiple, connected MCES can be obtained by enumerating all possible subgraphs and extracting the largest one common to all input graphs.[48] The subgraphs are retrieved from the full extended connectivity fingerprint (ECFP),[49] a common molecular descriptor. The ECFP algorithm generates circular-growing substructures using an adaptation of Morgan's algorithms[50] for canonical labeling of molecular graphs. Note that this approach is neither exact nor it can be scaled up for larger graphs.

Benchmarking

Many MCS algorithms have been developed, but very little effort has been made to create a uniform test environment to compare the algorithms with respect to runtime and their field of application. Most often, the algorithms are evaluated in a special experimental setting to show that they can outperform previous ones. The result is a number of evaluations that are hard to compare. Conte et al.[51] face the problem by assembling a diverse set of synthetical graphs that is composed of randomly connected graphs, regular and irregular meshes, and regular and irregular bounded valence graphs. Note that irregular bounded valence graphs have properties similar to molecular graphs. Three classical MCS algorithms were tested with this benchmark set: McGregor's[38] backtracking procedure, and the algorithms of Durand et al.[52] and Balas and Yu,[24] both searching for the maximal clique. None of the three algorithms showed a superior runtime behavior for all kinds of graphs. McGregor's[38] algorithm performs well when graphs are sparse and/or small. However, for irregular graphs with bounded valence, Durand's algorithm performs best. The comparison shows that an appropriate application of MCS algorithms strongly depends on the considered problem.

APPLICATIONS

MCS algorithms have a variety of applications in molecular science and the number is constantly rising. The complete scope of applications can certainly

not be covered by this article. Nevertheless, we want to give a broad overview of areas that use MCS algorithms. In the following, we provide examples for biological problems that are addressed by using MCS algorithms. The first part considers algorithms applied to molecular graphs obtained from small organic molecules such as drugs, whereas the second part provides examples for applications on graphs derived from large molecules such as proteins. The two types of graphs are fundamentally different. A small molecule graph is composed of atoms and bonds or groups of atoms and topological distances. In proteins, complete amino acids and the geometric distances between them are usually used to construct the graph. Also, the interpretation of an MCS differs for small molecules and proteins. An MCS of small molecules is often used as structural similarity measure; whereas in proteins, it resembles a common structure motif. It appears that the algorithms only find little application for ribonucleic acids (RNAs), even though common structural patterns are of major interest when studying RNA.

Small Organic Molecules

MCS algorithms can be applied to identify ligand families, predict ligand activity, or to analyze the mechanism of reactions. Although the most common application of MCS algorithms is to retrieve similarity values, the examples show in detail how small molecules are transformed into graphs and how to retrieve a measure of similarity from an MCS. Other applications of MCS algorithms involve ligand alignment,[53] the determination of quantitative structure–property relationships (QSPR),[54] and pharmacophore modeling.[55,56]

Compound Classification

A central problem when dealing with small molecules in pharmaceutical research is to group individual compounds into structurally related families or clusters. Manually grouping large databases is a tedious task and automated procedures are, therefore, often used. Automated clustering methods need a similarity metric for pairwise comparison of structures and a clustering algorithm for sorting compounds into structurally related groups. An MCS can describe common connected substructures or scaffolds as well as a set of largest common fragments or functional groups between two molecules. Stahl et al.[35] analyzed the usability of different similarity metrics obtained from disconnected MCESs for compound clustering. The motivation to use a disconnected MCES is to detect similarity between compounds that do not share a large common substructure but rather common functional groups that are disconnected. Six different MCES algorithms (Rambin,[57] an implementation of the Bron–Kerbosch algorithm,[23] Dfmax and Nmclique,[58] Pardalos,[59] Wood,[60] and Rascal[29]) and a variety of clustering methods were compared on a set of 466 compounds known to bind to nine different targets. From the number of vertices and edges that comprise the MCS, a similarity between two molecules is calculated according to[29] the following equation:

$$\mathrm{sim}(A, B) = \frac{(|V(\mathrm{MCS}_{A,B})| + |E(\mathrm{MCS}_{A,B})|)}{(|V(A)| + |V(B)|)(|E(A)| + |E(B)|)} \quad (1)$$

where $\mathrm{MCS}_{A,B}$ is the MCS between molecules A and B.

The similarity calculation is extended by two correction terms. The first penalizes a different relative topological arrangement between the three largest functional groups. The second raises the similarity index if the largest MCES fragment comprises more than 70% of one molecule indicating a common scaffold. A combination of the RASCAL–MCES algorithm with the average linkage unweighted pair group method with arithmetic mean (UPGMA)[61] cluster method most accurately separates compounds into their distinct classes while creating relatively pure clusters and the least number of singletons.

Compound Activity Prediction

A key step in finding new drugs is the identification of chemical compounds that shows activity in specific biological processes. Virtual screening techniques try to identify potential active compounds by large-scale *in silico* activity prediction with the aim to reduce the number of molecules that need to be experimentally tested. The similarity principle in drug design states that, statistically, structurally similar compounds tend to have similar activity.[62] Rarey and Dixon[7] identify molecules that belong to the same class of inhibitors on a set of 972 molecules[63] from the MACCS Drug Data Report database[64] and predict binding geometries on 58 molecules[65] taken from the Brookhaven Protein Data Bank.[66] Molecules are reduced to acyclic graphs, feature trees, in which vertices represent functional groups and edges resemble the relative topological arrangement of groups. A vertex describes steric features, e.g., van der Waals volume, and chemical features such as possible interactions a group can form with a potential receptor. A weighted average over the similarity values of vertex matches in the exact, disconnected MCIS of two feature trees is used as the similarity between two

molecules. Predictions of ligand-binding geometry have an average root mean square derivation (RMSD) under 4 Å in 61% of the 58 molecules. In a virtual screening experiment, enrichment factors at 1% of the 972 molecules screened are similar when using feature trees and daylight fingerprints. Nevertheless, 50% of the top-ranking molecules obtained using feature trees differ from the ones found using daylight fingerprints.

Schietgat et al.[46] successfully perform a compound activity prediction on the National Cancer Institute (NCI) dataset containing about 70,000 active and inactive compounds to treat human tumor cells. The molecules are transformed into binary strings that encode the occurrence of frequent substructure patterns in a set of active compounds. A substructure mining algorithm[47] in combination with the block-and-bridge preserving MCIS algorithm (see above) calculates all patterns present in a ligand, and the presence of a most frequent pattern is indicated by setting the correspond bit. Finally, a support vector machine classifies the compounds based on their binary description. A 10-fold cross-validation shows a prediction performance comparable to other methods based on subgraph isomorphism,[67] fingerprints, or kernels.

Scaffold Hopping

A problem of similarity-based methods for virtual screening is their tendency to only identify compounds that are structurally very similar to the original active molecules. However, it is of special pharmaceutical interest to find novel molecules that are built from different molecular scaffolds while preserving activity against the same target protein. Different scaffold hopping methods successfully address this task.[68] Barker et al.[31] investigate the scaffold-hopping ability of different MCS-based molecular similarity measures. The molecular graph is transformed into a reduced graph with vertices resembling functional groups and edges representing the topological distances between them. A reduced graph is similar to a pharmacophoric description, which describes the structural features essential for the biological activity. Barker et al.[31] adapted the similarity formula from Eq. (1) and studied the influence on the similarity value when using either an MCIS or an MCES between two reduced graphs. A comparison with daylight fingerprints[4] in a simulated virtual screening experiment on a filtered version of the MDL Drug Data Report indicates a similar enrichment ability at 1% of the database screened and only small differences when using the MCIS or MCES, respectively. The method retrieves about the same number of unique scaffolds, but the scaffolds are complementary in terms of diversity to those found using daylight fingerprints.

Reaction Mapping

Understanding the mechanism of enzymatically catalyzed reactions is of major interest when studying metabolic pathways of the cell. A chemical reaction transforms the reactant molecule to the product by deleting existing bonds and forming new ones. These reaction centers can be experimentally identified but only in a small scale. The works of Korner and coworkers[69,70] are aimed at automatically determining reaction centers in high-throughput applications. Each reactant and product is modeled as a molecular graph in which vertices are atoms and edges are bonds. An edge also holds a weight that corresponds to the stability of the bond. A weighted MCES (wMCES) that maximizes the sum of common edge weights is used to determine the set of bonds that are most likely conserved when a reactant is changed to the product. All bonds not part of the wMCES are either broken or formed during the reaction. The sets of conserved bonds and reaction bonds allow an identification of a bijective atomic mapping between reactant and product. For the experiment, the RASCAL algorithm was modified to calculate the wMCES and its automated application most often results in a correct mapping of over 8000 manually mapped chemical reactions obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG)[71] and BioPath database.[72]

Quantitative Structure–Activity Relationships

The recently exponential growth in the number of publications presenting quantitative structure–activity relationship (QSAR) and QSPR shows the importance of an accurate prediction of compound activity/property in modern chemistry and biochemistry. The concept of QSAR/QSPR is to transform chemical knowledge and intuition into mathematically derived equations that correlate the structure with a known activity/property. With such a model, it is possible to search any number of compounds, even the ones that are not yet synthesized, for the desired activity/property. Cuadrado et al.[73] derive a QSAR model to predict the blood–brain barrier permeability from a known set of 136 active compounds. The QSAR model describes each ligand as a vector of similarity values against all actives. In principle, any measure that describes the similarity between two ligands can be used. For a detailed review on molecular similarity measures, see Refs 74 and 75. Cuadrado et al.[73] derive similarity values by approximating the

van der Waals surface area[76] of an extended MCIS (EMCIS). The EMCIS contains information about the position of substituents that are not part of the original MCIS. The model is trained and tested using leave-one-out validation to guarantee a high prediction performance. An independent test indicates a prediction performance similar to previous QSAR models and other approaches based on 3D methods or neural networks.

Ribonucleic Acid

Structure comparison is one of the central tools used for function prediction of novel RNAs. Often, a sequence comparison is sufficient for finding related RNAs with known function from which a function prediction can be obtained. For some major RNA families, such as transfer RNA and ribosomal RNA, sequence comparison fails due to the low-sequence similarity between family members. Fortunately, these families show a highly conserved fold. Therefore, a direct comparison of the secondary structure can reveal similarities not present on the sequence level. Chao[32] compared RNA structures of different complexity and searched for the presents of iron response elements (IREs) in the untranslated region of human messenger RNA (mRNA). The mRNA structure is modeled as a graph, with nucleotides forming the vertices and edges resembling either covalent bonds or hydrogen contacts between nucleotides. The search for IREs in human mRNAs yield 26 genes from which six are known to contain IREs. The comparison of structures within different RNA families results in the proposal of an extended vertex-encoding scheme. Instead of labeling each vertex with the corresponding nucleotide symbol, vertices are labeled according to their secondary structure. The scheme is useful when only the RNA structure, regardless of its sequence, should be compared.

Proteins

The two main application fields of MCS algorithms when studying proteins are protein alignment that identifies global structural similarities between proteins and pattern analysis in which the major interest lies in local similarities.

Structural Alignment

Understanding the function and architecture of proteins is a central problem in molecular biology. The 3D protein fold mainly determines the protein function, stability, and general behavior. Therefore, the structural comparison of proteins can give valuable insights into the nature of proteins. Jain and Lappe[77] compare protein structures by approximately solving the contact map overlap (CMO) problem[78] in which a protein structure is modeled as a contact graph and the MCS of two proteins describes the similarity between them. A contact graph consists of protein residues and two residues are connected by an edge if their distance in space is small enough. To obtain a solution to the CMO problem, the approximate MCES algorithm softassign[79,80] maximizes the number of common contacts between two proteins, and a self-developed DP strategy ensures that the order of residues forming the solution is the same in both proteins. The algorithm computes almost optimal matches on a CMO test set compiled by Strickland et al.[81] and shows running times around minutes for proteins up to 1500 residues in size. The results indicate that the method is faster than other CMO algorithms, the runtime scales well with increasing protein size, and that the algorithm is most efficiently applied when comparing structurally similar proteins.

Structural Pattern Analysis

The identification of substructures or motifs in proteins that are related to a specific function or fold generally leads to a hypothesis about the evolutionary origin or conducted function of the protein. The analysis of complete databases for frequently occurring motifs is an opportunity to identify conserved substructures. Caboche et al.[33] analyzed the NORINE database[82] for structural commonalities between nonribosomal peptides (NRPs). In contrast to regular proteins, their structure can be partially or fully cyclic, branched, or even polycyclic. The NORINE database provides about 700 NRPs as molecular graphs. A vertex of a graph corresponds to a monomer and an edge to a chemical bond between two monomers. The database is successfully analyzed for family-specific structural features in NRPs, and an example application is given to predict the product of NRP-producing proteins from the protein structure.

Artymiuk et al.[83] study common folding motifs between adenylyl cyclase and DNA polymerase 1 and between biotin carboxylase and adenosine diphosphate (ADP)-forming peptide synthetases in detail. The molecular graph representation of the analyzed proteins makes use of the fact that the spatial arrangement of secondary structure elements (SSEs) describes the protein's fold. Although SSEs are approximately linear structures, they are modeled by a vector drawn along their major axis. A molecular graph representing the protein structure is then composed such that each vertex holds an SSE vector and each edge describes a geometric relationship

between a pair of them. The Bron and Kerbosch[23] algorithm is modified to account for edge labels when searching for the maximal clique. The resulting disconnected MCIS gives the structural relationship between two protein folds. Artymiuk et al.[83] revealed common folding modes between the proteins that indicate similar function between adenylyl cyclase and DNA polymerase 1 and homology between the families of biotin carboxylase and ADP-forming peptide synthetases.

CONCLUSION

The aim of this review is to provide an overview of current algorithms that solve the MCS problem for molecular graphs and to show their general applications in the field of molecular science. Most algorithms address MCS problem by solving the maximal clique problem on a compatibility graph. However, two of the most recently presented algorithms[42,43] are backtracking procedures. Because of its good performance on molecular graphs, RASCAL based on clique detection belongs to the widely applied implementations.

One major application of MCS algorithms is their use to determine similarity between small organic compounds. In contrast to fingerprint methods, the MCS captures topological relations between atoms or functional groups. It, therefore, results in a similarity concept well reflecting the synthetic chemists understanding of molecular relationships. The additional topological information can be of high relevance when searching for alternative ligands with similar biological activity. Applied to proteins, MCS algorithms can accomplish the detection of global and local similarities.

The special kind of MCS used to address a problem is of central importance for the application as well as for algorithm design. We hope that future publications use the proposed classification scheme and give a clear description of the algorithm's intended application field. New MCS algorithms need to be compared with existing methods in a reproducible environment, preferably on a generalized test set or at least on a large number of varying graphs. The number of test sets available, especially those that resemble the properties of molecular graphs, is small; therefore, we encourage further research.

Because of NP-hardness, algorithms solving the MCS problem in general will likely stay exponential in runtime requirement. Nevertheless, for a specific application, there are typically lots of options to optimize. For performance, modeling vertex and edge compatibility is critical. Moreover, some graph classes such as trees and planar graphs allow much faster methods for MCS calculation. For future applications of MCS in molecular science, time spent for carefully modeling the problem as an MCS problem is, therefore, well invested.

We hope that this review provides an entry point into the current state of MCS algorithms and gives an insight into the broad range of applications in molecular science. A number of problems in molecular science can be solved with MCS-based approaches and, therefore, we encourage exploring new fields for their broader application.

REFERENCES

1. Irwin JJ, Shoichet BK. ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005, 45:177–182.
2. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. The NCBI Biosystems database. *Nucleic Acids Res* 2009, 38:492–496.
3. Xue L, Godden JW, Bajorath J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J Cheml Inf Comput Sci* 2000, 40:1227–1234.
4. *Daylight Theory Manual*. Daylight Chemical Information Systems Inc.
5. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002, 42:1273–1280.
6. Takahashi Y, Sukekawa M, Sasaki S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J Chem Inf Comput Sci* 1992, 32:639–643.
7. Rarey M, Dixon JS. Feature trees: a new molecular similarity measure based on tree matching. *J Comput-Aided Mol Des* 1998, 12:471–490.
8. Harper G, Bravi GS, Pickett SD, Hussain J, Green DVS. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J Chem Inf Comput Sci* 2004, 44:2145–2156.
9. Trinajstic N. *Chemical Graph Theory*. 2nd ed. New Directions in Civil Engineering. CRC Press; 1992.
10. Gund P. Three-dimensional pharmacophoric pattern searching. *Prog Mol Subcell Biol* 1977, 5:117–143.

11. Gund P. Pharmacophoric pattern searching and receptor mapping. *Ann Rep Med Chem* 1979, 14:299–308.
12. Whitney H. Congruent graphs and the connectivity of graphs. *Am J Math* 1932, 54:150–168.
13. Raymond JW, Willett P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput-Aided Mol Des* 2002, 16:521–533.
14. Garey MR. *Computers and Intractability*. New York: W. H. Freeman and Company; 1979.
15. Bunke H. Graph Matching: theoretical foundations, algorithms, and applications In: *International Conference on Vision Interface*. Quebec, Canada: Montreal; 2000, 82–88.
16. Pelillo M, Siddiqi K, Zucker SW. Matching hierarchical structures using association graphs. *IEEE Trans Pattern Anal Machine Intell* 1999, 21:1105–1120.
17. Yang B, Snyder WE, Bilbro GL. Matching oversegmented 3D images to models using association graphs. *Image Vis Comput* 1989, 7:135–143.
18. Barrow HG, Burstall RM. Subgraph isomorphism, matching relational structures and maximal cliques. *Inf Process Lett* 1976, 4:83–84.
19. Brint A, Willett P. Algorithms for the identification of 3-dimensional maximal common substructures. *J Chem Inf Comput Sci* 1987, 27:152–158.
20. Levi G. A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo* 1973, 9:341–352.
21. Cone M, Venkataraghavan R, McLafferty F. Molecular structure comparison program for the identification of maximal common substructures. *J Am Chem Soc* 1977, 99:7668–7671.
22. Kuhl F, Crippen G, Friesen D. A combinatorial algorithm for calculating ligand-binding. *J Comput Chem* 1984, 5:24–34.
23. Bron C, Kerbosch J. Finding all cliques of an undirected graph. *Commun of the ACM* 1973, 16:575–577.
24. Balas E, Yu CS. Finding a maximum clique in an arbitrary graph. *SIAM J Comput* 1986, 15:1054–1068.
25. Carraghan R, Pardalos P. An exact algorithm for the maximum clique problem. *Oper Res Lett* 1990, 9:375–382.
26. Shindo M, Tomita E. A simple algorithm for finding a maximum clique and its worst-case time complexity. *Syst Comput Japan* 1990, 21:1–13.
27. Babel L. Finding maximum cliques in arbitrary and in special graphs. *Computing* 1991, 46:321–341.
28. Gardiner EJ, Artymiuk PJ, Willett P. Clique-detection algorithms for matching three-dimensional molecular structures. *J Mol Graph Model* 1997, 15:245–253.
29. Raymond JW, Gardiner EJ, Willett P. RASCAL: calculation of graph similarity using maximum common edge subgraphs. *The Computer Journal* 2002, 45:631–644.
30. Raymond JW, Gardiner EJ, Willett P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comput Sc* 2002, 42:305–316.
31. Barker EJ, Buttar D, Cosgrove DA, Gardiner EJ, Kitts P, Willett P, Gillet VJ. Scaffold hopping using clique detection applied to reduced graphs. *J Chem Inf Model* 2006, 46:503–511.
32. Chao S-Y. Maximum common substructure extraction in RNA secondary structures using clique detection approach. *World Acad Sci, Eng Technol* 2008, 45:219–228.
33. Caboche S, Pupin M, Leclere V, Jacques P, Kucherov G. Structural pattern matching of nonribosomal peptides. *BMC Struct Biol* 2009, 9:15.
34. Mehlhorn K. Data structures and algorithms 2: graph algorithms and NP-completeness. In: *Monographs in Theoretical Computer Science. An EATCS Series*. Vol. 2. Springer; London, UK 1984.
35. Stahl M, Mauser H, Tsui M, Taylor NR. A robust clustering method for chemical structures. *J Med Chem* 2005, 48:4358–4366.
36. Jauffret P, Tonnelier C, Hanser T, Kaufmann G. Machine learning of generic reactions: 2. toward an advanced computer representation of chemical reactions. *Tetrahedron Comput Methodol* 1990, 3:335–349.
37. Koch I. Enumerating all connected maximal common subgraphs in two graphs. *Theor Comput Sci* 2001, 250:1–30.
38. McGregor JJ. Backtrack search algorithms and the maximal common subgraph problem. *Softw: Pract Exp* 1982, 12:23–34.
39. Ullmann JR. An algorithm for subgraph isomorphism. *J Assoc Comput Machinery* 1976, 23:31–42.
40. Wong AKC, Akinniyi FA. An algorithm for the largest common subgraph isomorphism using the implicit net. *Proc IEEE Syst, Man, and Cybern* 1983, 1:197–201.
41. Barnard J. Substructure searching methods — old and new. *Journal of Chem Inf Comput Sci* 1993, 33:532–538.
42. Cao Y, Jiang T, Girke T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* 2008, 24:366–374.
43. Berlo RJPv, Groot MJLd, Reinders MJT, Ridder Dd. Efficient calculation of compound similarity based on maximum common subgraphs and its application to prediction of gene transcript levels. In: *Information & Communication Theory Group, Technical Report.Delft, the Netherlands: Delft University of Technology*; 2009.
44. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*. Cambridge, MA: MIT Press; 2001.

45. Gupta A, Nishimura N. Finding largest subtrees and smallest supertrees. *Algorithmica* 1998, 21:183–210.
46. Schietgat L, Ramon J, Bruynooghe M, Blockeel H. An efficiently computable graph-based metric for the classification of small molecules. In: *Proceedings of the 11th International Conference on Discovery Science*. Berlin, Heidelberg: Springer-Verlag; 2008, 197–209.
47. Horvarth T, Ramon J, Wrobel S. Frequent subgraph mining in outerplanar graphs. In: *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY: ACM; 2006, 197–206.
48. Hassan M, Brown RD, Varma-O'brien S, Rogers D. Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* 2006, 10:283–299.
49. SciTegic. *Pipeline Pilot — Basic Chemistry Collection User Guide*. Telesis Court, San Diego, CA: ; 2006, 92121–4779.
50. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc* 1965, 5:107–113.
51. Conte D, Guidobaldi C, Sansone C. A comparison of three maximum common subgraph algorithms on a large database of labeled graphs. In: *Graph Based Representations In Pattern Recognition*. Berlin: Springer-Verlag; 2003, 130–141.
52. Durand PJ, Pasari R, Baker JW, Tsai C-C. An efficient algorithm for similarity analysis of molecules. *Internet J Chem* 1999, 2.
53. Thorner DA, Willett P, Wright PM, Taylor R. Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs. *J Comput-Aided Mol Des* 1997, 11:163–174.
54. Cuissart B, Touffet F, Cremilleux B, Bureau R, Rault S. The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J Chem Inf Comput Sci* 2002, 42:1043–1052.
55. Martin YC, Bures MG, Danaher EA, DeLazzer J, Lico I, Pavlik PA. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *Jf Comput-Aided Mol Des* 1993, 7:83–102.
56. Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today* 2008, 13:23–29.
57. Available at: http://www.twisted-helices.com/computing/rambin/rambin.html (Accessed November 11, 2010).
58. Available at: ftp://dimacs.rutgers.edu/pub/challenge/graph/solvers/ (Accessed November 11, 2010).
59. Pardalos PM, Xue J. The maximum clique problem. *J Glob Opt* 1992, 4:301–308.
60. Wood DR. An algorithm for finding a maximum clique in a graph. *Oper Res Lett* 1997, 21:211–217.
61. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 1958, 28:1409–1438.
62. Johnson EG, Maggiora GM, *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 1990.
63. Briem H, Kuntz ID. Molecular similarity based on DOCK-generated fingerprints. *J Med Chem* 1996, 39:3401–3408.
64. *MACCS Drug Data Report (MDDR)*. San Leandro, CA: MDL Information Systems Inc.
65. Lemmen C, Lengauer T, Klebe G. FLEXS: a method for fast flexible ligand superposition. *J Med Chem* 1998, 41:4502–4520.
66. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000, 28:235–242.
67. Bringmann B. Don'be afraid of simpler patterns. In: 10th *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Berlin: Springer; 2006, 55–66.
68. Schneider G, Schneider P, Renner S. Scaffold-hopping: how far can you jump? *QSAR Comb Sci* 2006, 25:1162–1171.
69. Korner R, Apostolakis J. Automatic determination of reaction mappings and reaction center information. 1. the imaginary transition state energy approach. *J Chem Inf Model* 2008, 48:1181–1189.
70. Apostolakis J, Sacher O, Korner R, Gasteiger J. Automatic determination of reaction mappings and reaction center information. 2. validation on a biochemical reaction database. *J Chem Inf Model* 2008, 48:1190–1198.
71. Kanehisa M. The KEGG database. *Novartis Foundation Symp* 2002, 247:91–101; discussion 101–103, 119–128, 244–252.
72. Reitz M, Sacher O, Tarkhov A, Trumbach D, Gasteiger J. Enabling the exploration of biochemical pathways. *Org Biomol Chem* 2004, 2:3226–3237.
73. Cuadrado MU, Ruiz IL, Gomez-Nieto MA. QSAR models based on isomorphic and nonisomorphic data fusion for predicting the blood brain barrier permeability. *J Comput Chem* 2007, 28:1252–1260.
74. Maggiora GM, Shanmugasundaram V. Molecular similarity measures. *Methods in Mol Biol (Clifton, N.J.)* 2004, 275:1–50.

75. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci* 1998, 38:983–996.
76. Gutman I, Kortvelyesi T. Wiener indices and molecular surfaces. *Zeitschrift fur Naturforschung* 1995, 50a:669–671.
77. Jain BJ, Lappe M. Joining softassign and dynamic programming for the contact map overlap problem. In proceedings of the 1st international conference on Bioinformatics research and development. *BIRD* 2007. Berlin: Springer Heidelberg, 410–424.
78. Godzik A, Skolnick J. Flexible algorithm for direct multiple alignment of protein structures and sequences. *Comput Appl Biosci : CABIOS* 1994, 10:587–596.
79. Gold S, Rangarajan A. A graduated assignment algorithm for graph matching. *Pattern Anal and Machine Intell, IEEE, Trans on Pattern Anal and Machine Intell*, 1996, 18:377–388.
80. Ishii S, Sato MA. Doubly constrained network for combinatorial optimization. *Neurocomputing* 2002, 43:239–257.
81. Strickland DM, Barnes E, Sokol JS. Optimal protein structure alignment using maximum cliques. *Oper Res* 2005, 53:389–402.
82. Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res* 2008, 36:326–331.
83. Artymiuk P, Spriggs R, Willett P. Graph theoretic methods for the analysis of structural relationships in biological macromolecules. *J Am Soc Inf Sci Technol* 2005, 56:518–528.

