Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification

John W. Raymond,*,† Ian A. Watson,‡ and Abdelaziz Mahoui†

Discovery Informatics and Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana

Received February 3, 2009

Historically, one of the characteristic activities of the medicinal chemist has been the iterative improvement of lead compounds until a suitable therapeutic entity is achieved. Often referred to as lead optimization, this process typically takes the form of minor structural modifications to an existing lead in an attempt to ameliorate deleterious attributes while simultaneously trying to maintain or improve desirable properties. The cumulative effect of this exercise performed over the course of several decades of pharmaceutical research by thousands of trained researchers has resulted in large collections of pharmaceutically relevant chemical structures. As far as the authors are aware, this work represents the first attempt to use that data to define a framework to quantifiably catalogue and summate this information into a medicinal chemistry expert system. A method is proposed that first comprehensively mines a compendium of chemical structures compiling the structural modifications, abridges them to rectify artificially inflated support levels, and then performs an association rule mining experiment to ascribe relative confidences to each transformation. The result is a catalogue of statistically relevant structural modifications that can potentially be used in a number of pharmaceutical applications.

INTRODUCTION

In recent years, there has been a concerted effort throughout the pharmaceutical industry to increase the rate and decrease the cost of drug development. Within the field of medicinal chemistry, this has manifested itself in a number of ways including parallel synthesis and structure-based design as well as outsourcing synthetic work. While many significant technological improvements have been made, the fundamental process of transforming an initial compound of interest to a therapeutic entity has changed comparatively little. It is still an inherently stochastic and unpredictable process guided by knowledge domain experts often influenced by educated guesses and anecdotal inference. This dilemma has not gone unnoticed by the medicinal chemistry community and led Topliss[1] over 35 years ago to state, "*Historically, approaches to this problem have been rather haphazard, depending for the most part on the particular experience and intuition of the medicinal chemist...*". In response, he introduced what has come to be known as the Topliss tree, a decision tree for side-chain replacement and aromatic ring substitution based on computed electronic parameters. While the specific implementation may have fallen out of favor, the idea that chemical structure exploration is an activity that has the potential to be systematically addressed remains relevant. While it is probably not possible or even desirable to fully divorce the process from the skillful direction of individual knowledge domain experts, a systematic, algorithmic approach to lead compound advancement has the potential to help transform early stage drug discovery. The objective of this work is to provide an initial, quantitative foundation for establishing such a medicinal chemistry expert system.

One of the central tenets of modern medicinal chemistry is that many substructures and functional groups tend to exhibit similar biological and/or physical property characteristics. This principal often serves as a basis for the process of developing lead compound candidates from prototype compounds[2] by sequences of iterative, substructure modification. These correlated substructure pairs are often referred to as *bioisosteres*. In general, a bioisostere is typically defined as a pair of substructures that exhibit similar physicochemical properties while also conveying similar biological properties to a molecule, and several reviews have been published on the topic.[3−7] However, the bioisostere is more accurately labeled as a concept as this definition is functionally inadequate.

Should a bioisostere have general applicability such as the classic carboxylic acid/tetrazole pair,[8] or should it also include structure replacement pairs that are specific to individual binding site regions for unique protein binding sites employing specific binding modes?[9,10] How does one differentiate between the two? In some instances, a bioisostere may be introduced in a specific context but potentially have wider applicability.[11] How does one ascribe relative significance to different substructure surrogates or distinguish true bioisostere pairs from observations that are more appropriately attributed to random chance? Moreover, there is a need to distinguish between substructure replacements that are statistically relevant and nonobvious from those that are relevant but obvious (e.g., methyl/ethyl). If a systematic framework describing the process of lead compound optimization is to be developed, clearly a more quantitative or algorithmic characterization is required.

* Corresponding author e-mail: raymond_john_w@lilly.com.
† Discovery Informatics.
‡ Lilly Research Laboratories.

10.1021/ci9000426 CCC: $40.75
Published on Web 07/15/2009

Germane to this objective is the Drug Guru system[12] which is a qualitative medicinal chemistry expert system. The application encodes almost 200 medicinal chemistry transformation rules capable of providing suggested alternatives for structural modification based on a prototype compound. The rules are the result of literature searching and direct medicinal chemistry interaction. That subjective approach differs from the one pursued here in that there is no prioritization of the rules by probabilistic expectation or confidence, but, in common with this effort, it embodies the desire to encapsulate the medicinal chemistry knowledge base into an in silico expert system.

METHODOLOGY

Medicinal Chemistry Knowledge Base. One of the fundamental assumptions of this work is the notion of collective wisdom whereby large, diverse groups perform better than small sets of individuals at solving complex problems.[13] Our premise is that the large collections of pharmaceutically relevant compounds represent a vast, untapped resource for mining the collective knowledge of thousands of experienced medicinal chemists laboring over the course of many decades solving a multitude of pharmacological problems. The process of lead compound refinement is often analogized as a stochastic optimization problem. As such, the process of lead optimization can then be presumed to consist of both random perturbation (serendipitous exploration) as well as directed search (traditional bioisostere iteration).

Much effort has been expended in an attempt to direct medicinal chemists regarding what compounds to synthesize next given an archetype compound and an information state (i.e., biological/physical property profile via structure-based design, QSAR, etc.); however, comparatively little study has been given to the opposite perspective. What would the average medicinal chemist synthesize next given an archetype compound and an associated information state? For instance, most experienced medicinal chemists have a general comprehension of what types of structure modifications would be most appropriate in an attempt to increase solubility while simultaneously managing the probability of adversely affecting binding affinity.

To explore this, a structure data set was constructed in an attempt to represent the existing medicinal chemistry knowledge domain. The data set was constructed as a union of the Lilly corporate collection as well as purchased or in-licensed structure databases (MDDR, GVK Biosciences, Jubilant, and all marketed drugs) consisting of structures obtained from literature and patent sources. In instances where structures were known to be retrieved from patents, only those associated with an activity value were used in order to avoid the potentially aberrant effects of enumerated Markush structures. All redundant structures were removed from the merged data set which was then filtered using PipelinePilot using the following criteria: keep largest fragment; must be organic; $200 \leq$ molecular weight ≤ 650. This resulted in a data set of approximately 2.7 million structures.

No attempt was made to stratify the data set by biological activity as is typically the case in bioisostere related efforts. The fundamental objective to be addressed here is to identify the most statistically relevant substructure modifications from a conceptual perspective, omitting any particular biological or physical end-point consideration. In this regard, all aspects constituting the scope of structure modification are to be considered which includes potential biological and physical property considerations, synthetic protocols known to the average medicinal chemist, and available reagents as well as random exploration. From this perspective, it only matters that two structures are similar enough that they can be viewed as being related by simple substructural transformations, even if the underlying purpose and/or synthetic routes differ. Although not explicitly addressed here, constraining by activity class can be accomplished by a simple reimplementation of the proposed methodology using appropriately stratified data.

Molecular Structure Transformation. It is first necessary to define the context of a molecular transformation. To avoid confusion with the convention of the bioisostere, we adopt the term molecular structure transformation (MST). An MST refers to a structural transformation of a chemical compound into another distinct compound that can be characterized by the replacement of portions of the original molecule with other substructures which preserve all attachment points present in the original substructure (i.e., a molecular substructure modification, MSM). Figure 1(a) illustrates an example MST depicting two MSM occurrences.

There have been previous published efforts directed at in silico mining of bioisostere fragment pairs.[14−16] A common technique used in bioisostere mining approaches is to determine the maximum common substructure (MCS) between pairs of molecules identifying the corresponding substructures not present in the MCS as the candidate bioisostere replacement; thus, the interchangeable replacement of the substructures not present in the MCS constitutes an MSM. An interested reader is referred to a detailed review of the MCS in a chemical informatics context.[17] The method proposed here also follows the convention of maximal commonality for establishing an MST. The maximum commonality approach provides a convenient, algorithmic framework for both mining substructural transformations as well as asserting sufficient commonality between two molecules to satisfy the mandate that two molecules have a reasonable probability of being related by deliberate intent. Here an MST is formally defined as a transformation of one molecule (M_i) to another molecule (M_j) whereby the ratio of the number of bonds in the MCS with respect to the maximum number of bonds in M_i or M_j is at least 0.7 under the constraint that the difference in the number of bonds between M_i and M_j is less than or equal to 8 (i.e., $E(MCS)/\max(E(M_i), E(M_j)) \geq 0.7 \mid \mathrm{abs}(E(M_i)\text{-}E(M_j)) \leq 8$). These criteria significantly improve the computational efficiency of the MCS detection and help ensure that the resulting MCS adequately captures an intuitive representation of structural commonality.

Figure 1(b) demonstrates the application of the MST constraints; the second candidate MST violates both constraints. It is important to note that the proposed criterion does not put a constraint on the number of distinct MSM occurrences that constitute a valid MST. It is acknowledged that within the context of medicinal chemistry synthetic efforts that subsequently synthesized compounds often differ from the archetype lead or nearest neighbor by more than one substructure modification.

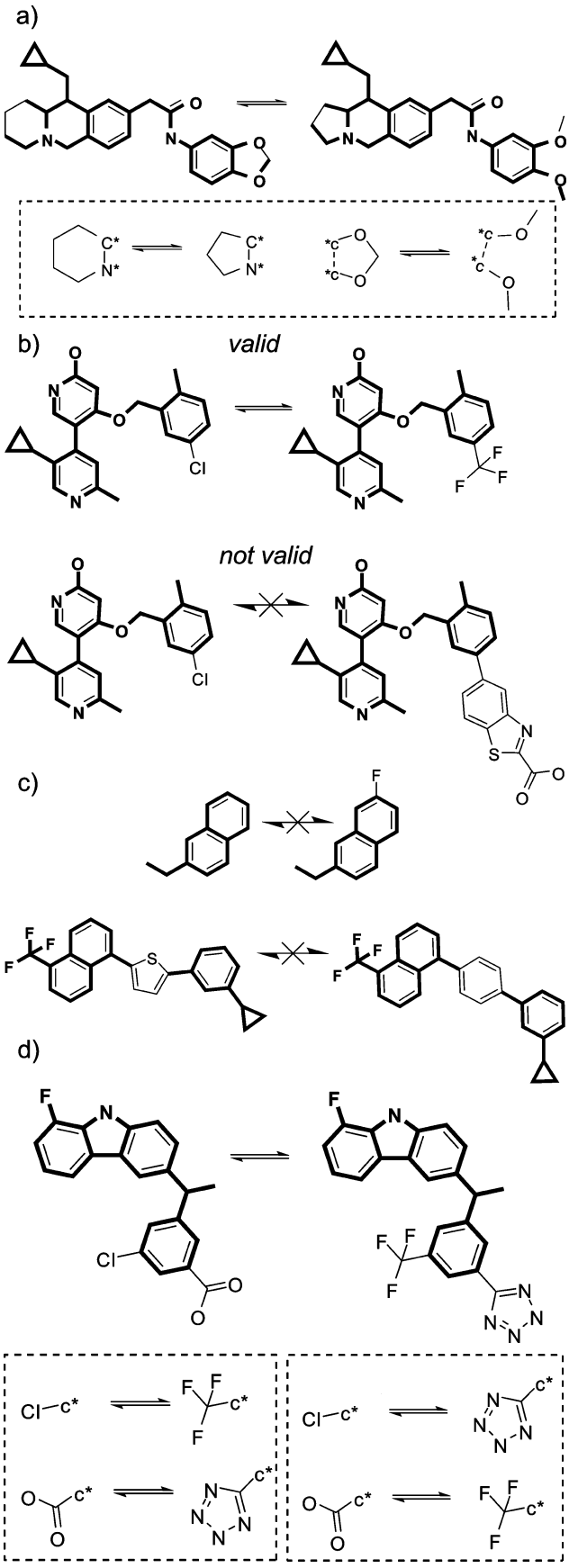


Figure 1. Example molecular structure transformations (MST). Dashed lines indicate aromatic bonds. Lower-case letters indicate aromatic atoms. (a) Example MST with two molecular substructure modifications (MSM). (b) MST constraint implementation. (c) MST determination caveats (hydrogen replacement and variable length linker replacement). (d) Symmetrically degenerate MST.

The method used here to establish the MCS between two molecules is a simplified, variant of the RASCAL algorithm.[18,19] The current implementation omits some of the pruning heuristics of the original algorithm in lieu of more efficient data structures and parallelization. There are two notable constraints that affect the type of MST that are currently identified. This includes the omission of substructure modifications where one of the substructures is a hydrogen atom. This is due to the supplemental requirement that the degrees (i.e., number of non-hydrogen neighbors) of the corresponding atoms in the MCS be equal. For instance, transformations of the type CH ↔ CF are ignored. The other limitation stems from the use of shortest path constraints during MCS determination, and, as a consequence, substructure replacements where the relative shortest path lengths between connection points differs between the two substructures are also ignored (i.e., linker replacements must have the same shortest path lengths between connection atoms). This is enforced during MCS determination by requiring that all shortest path distances between corresponding pairs of atoms (not in the same ring) in molecules M_i and M_j be equal. For the purposes of proof-of-concept, it is assumed that these will not affect the validity of the proposed methodology - only its relative completeness. Both of these limitations can be addressed in future implementations by using different control parameters during the MCS determination. Figure 1(c) illustrates both caveats.

When an MST is allowed to be composed of multiple MSM occurrences, then an additional complication must be reconciled. This is caused by degeneracy induced by symmetry in the MCS. Figure 1(d) demonstrates the problem. The phenyl ring contained in the MCS is symmetric about the axis defined by the substitution; therefore, the MST is ambiguous. In this case, it is obvious that the two MSMs are Cl/CF_3 and COOH/tetrazole; however, since the purpose of this analysis is to statistically derive MSM correlations, all potential MSM combinations constituting a valid MST must be considered. As a result it is possible for more than one valid MST to be identified per pair of chemical structures.

Evidenced by the example in Figure 1(a), there is a disparity between a graph theoretic representation of the MCS and a chemically intuitive interpretation of a substructure modification. For instance, there is a contextual difference if a replacement, *-C-* to *-O-*, exists in an acyclic side chain or within a ring system. A simple pruning strategy is used to resolve viable substructure representations from the unrefined MCS. It consists of two pruning rules implemented recursively until no more pruning is necessary. The first pruning requires that any atom in the MCS can have at most only one neighbor in M_i or M_j that is not also in the current MCS. If an atom in the MCS, a_k, is found to have more than one neighbor atom not also in the MCS, then atom a_k and all bonds incident on a_k are removed from the current MCS. The other pruning step attempts to maintain the integrity of an MSM that occurs within a ring. This step requires that if two bonds are both in the same SSSR ring[20] in either M_i or M_j and one of the bonds is in the MCS, then the other bond must also be in the MCS. If the two bonds are not both in the MCS, then the bond currently in the MCS is removed from the MCS, and any atoms which become isolated in the process are also removed from the MCS.

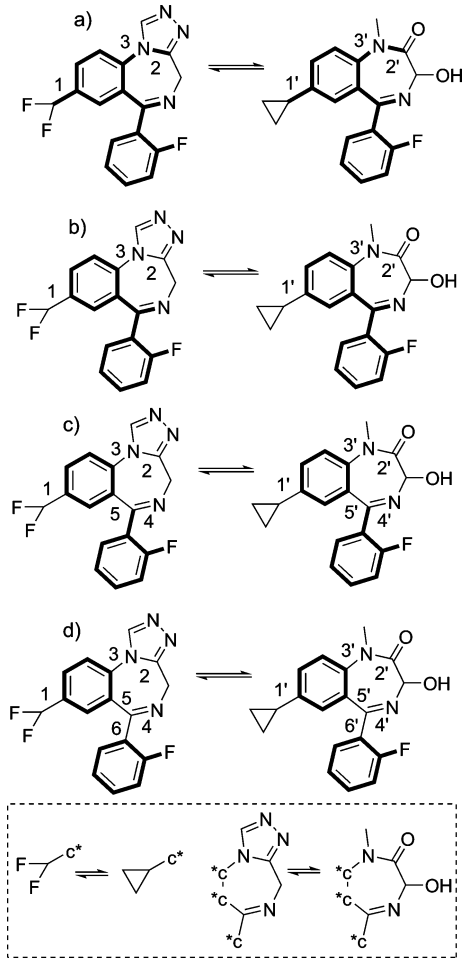


Figure 2. Recursive MCS pruning process. Final MSMs depicted in dashed-line box. Lower-case letters indicate aromatic atoms. (a) Unpruned MCS. MCS denoted in bold. (b) Bonds 1(1′), 2(2′), and 3(3′) pruned using degree constraint rule. (c) Bonds 4(4′) and 5(5′) pruned using ring constraint rule. (d) Bond 6(6′) pruned using degree constraint.

Figure 2 illustrates an application of the pruning process. The unprocessed MCS is depicted in bold in Figure 2(a). Figure 2(b) illustrates the degree pruning process whereby bonds 1(1′) and 2(2′) are removed from the MCS since they both have more than one incident bond not contained in the MCS. The removal of bond 2(2′) results in bond 3(3′) also having more than one incident bond not in the MCS; therefore, it is also removed from the MCS. Figure 2(c) demonstrates the ring pruning process. Here bonds 4(4′) and 5(5′) are culled from the MCS since they both are members of an SSSR ring, and not all members of that SSSR ring are present in the MCS (e.g., bond 2). Following the implementation of the degree and ring-based pruning, the process is repeated until no more bonds can be removed. Figure 2(d) depicts the recursive implementation of the degree pruning process following ring pruning where bond 6(6′) is pruned. Following the removal of 6(6′) no more bonds are eligible for pruning, and the MSMs can then be extracted.

This simple process has proven satisfactory for the vast majority of candidate MSMs encountered. However, as can be expected, certain anomalous instances such as caged ring structures as well as preserving known functional groups had to be resolved using customized heuristics. For instance, it is preferable for an MSM to constitute an amide to a carboxylate modification as opposed to an amine to hydroxyl.

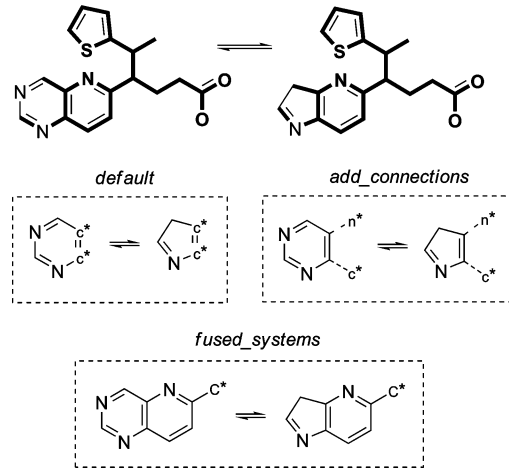


Figure 3. MSM extraction modes. MCS depicted in bold.

No algorithmic implementation can fully capture all of the nuanced subtleties that may be involved in an MST. Nor is it clear what type of MSM representation best represents a medicinal chemistry perspective. To address this limitation, three distinct MSM extraction operations were performed, *default*, *add_connections*, and *fused_systems*. All three MSM extraction operations operate on the pruned MCS. The *default* extraction consists of disconnecting the atoms and bonds in the MCS from the parent structures. The disconnected substructures comprise the candidate MSM fragment prototypes. These fragments are then labeled with any incident atoms that connect the fragment to the parent molecule. These represent the attachment points. In addition, any pair of atoms in the fragment that was bonded in the parent structure (including attachment points) is also bonded in the fragment (i.e., reconnect rings if necessary). The result of this procedure is depicted in Figure 3.

The *add_connections* option is identical to the *default* option except that it also includes all atoms in the MCS incident on the *default* attachments points. A potential advantage of this approach is that the increased specificity of the attachment environment can lead to more discriminating substructural modifications. One disadvantage to the *add_connections* approach is that it may overspecify certain MSM substructures so that a statistically interpretable sampling may not be achieved. The final extraction method implemented, *fused_systems*, is identical to the *default* method except that when a fused ring system is encountered in an MSM it extracts the ring system in its entirety rather than attempting to model a localized ring modification.

Substructure Associations. In recent years, there has been a significant body of research published regarding association rule mining in large databases. In a mathematical representation of the problem, $I = \{i_1,i_2,...,i_n\}$ represents the set of all unique items contained in database D. Database D consists of a number of transactions, T, and each transaction is a subset of items contained in I (i.e., $T \in D \mid T \subseteq I$). A transaction, T, is said to support an item set, X, if $X \subseteq T$. An association rule between two item sets, X and Y, is expressed as $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$. In the rule $X \Rightarrow Y$, X is referred to as the antecedent, and Y is the consequent. Association rule mining then consists of determining associations between item sets that are statistically supported as being interesting.

The association rule framework is illustrated using pairs of substructures in Figure 4. The set of ten fragment pairs

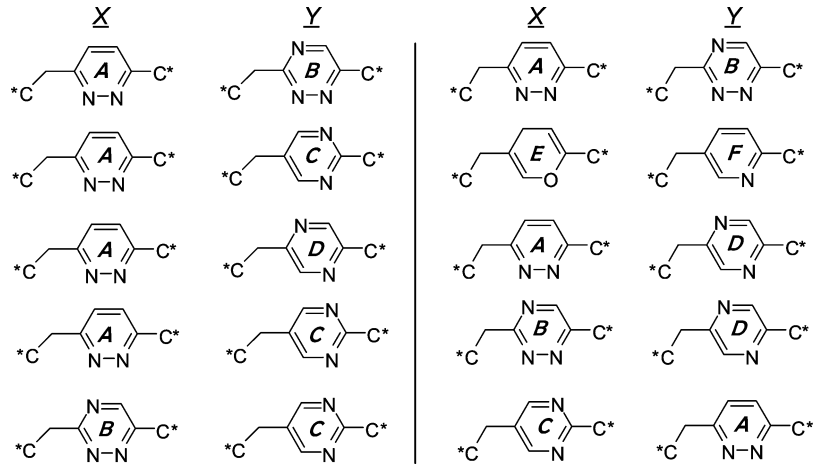


Figure 4. Example MSM association rule framework. Each (*X*,*Y*) pair is an MSM resulting from pairwise molecular structure comparisons. $I = \{A\text{-}F\}$, $|D| = 10$, and each MSM is a transaction, *T*.

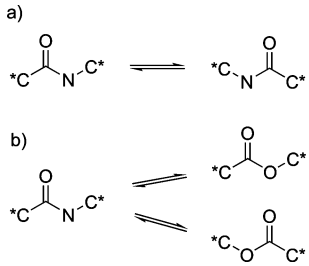


Figure 5. (a) Amide to retro-amide linker MSM and (b) amide to ester or retro-ester linker MSMs.

constitutes the database *D* (i.e., $|D| = 10$), and the six unique fragments (*A*-*F*) are the set of unique items *I*. Each (*X*,*Y*) fragment pair is an MSM linker replacement between a pair of molecules denoting a transaction, *T*, in the association rule framework. In this example, the distinction between the antecedent (*X*) and consequent (*Y*) is arbitrary since the fragment replacement is assumed to be undirected. While the association mining performed here is based on an MCS representation, the substructures comprising the MSMs could be tabulated using other fragmentation schemes.

Given a sufficiently large database of pharmaceutically relevant structures a database of MSMs can be created by first computing all pairwise, MCS-based substructure transformations (MST). More than one MSM may occur for a given MST. An MSM abstraction differs slightly from a typical association rule transaction in that it is not unambiguously defined by the antecedent and consequent item sets. An MSM constitutes a localized molecular transformation defined by a mapping between the two substructures, annotating their relative configuration. Figure 5 depicts examples where the fragment representations are inadequate to define an MSM. In these examples, the MSM must be qualified with explicit mappings differentiating the fragment correspondences.

The association rule framework is general and can incorporate any categorical data. Therefore, it is a straightforward exercise to extend the methodology to include end point measures such as differences in biological activity and ADME properties which can facilitate association rule-based QSAR. For the purposes of methodological development, we focus solely on fragment to fragment correlations ignoring specific end-point data considerations. In addition, we consider only symmetric association rules where the order of the antecedent and consequent is immaterial. However, it is not difficult to envision scenarios where asymmetric rules might be desirable as it can be presumed that modifications involving ring closing (reducing spatial degrees of freedom) may be preferred to ring-opening in practice.

Although the literature on the topic of item set associations is extensive, only a small fraction of the published methods is applicable to the problem of resolving associations between substructures representing an MSM. Of particular interest are methods directed at the mining of rules for infrequent item sets. Unlike many other types of data, the frequency of a particular substructure in a MSM database will typically be very low relative to the number of MSMs. Furthermore, a particular MSM could still be of particular interest even if it is very infrequent. For this reason, support frequencies are not an acceptable attribution of relevance. In addition, a simple support frequency threshold does not compensate for random associations. If two fragments were more common than other fragments in the transaction database, then it is expected that the frequency of chance fragment associations is higher for the two fragments, but it does not indicate that the association is statistically relevant.

To generate relevance scores, the method proposed here involves the enumeration of candidate MSMs as well as a scoring scheme. In the case of MSMs, the enumeration of candidate rules is straightforward; however, anticipating the future inclusion of end point property data within the association rule framework, we utilize the hash-based scheme for enumerating rules described by Zhou and Yau[21] which can accommodate items sets consisting of more than one item. For the case of MSM correlation, the scoring scheme is the most important consideration.

Many measures for association rule relevance have been introduced.[22,23] Several of these measures have been adopted from the field of information retrieval and have been well studied in that context.[24] Analogous to the issues faced in similarity searching, there is no single measure that can be claimed to be superior in all contexts. After investigating several prospective measures, *hyper-lift* ($lift_{hyper}$)[25] was found to offer the most potential for prioritizing MSM rules. This is a subjective determination assessed by extensive inspection of the rules generated and is disadvantaged by the absence of the objective mechanisms for validation that are available

in other forms of relevance ranking such as precision and recall for similarity searching.[26]

$lift_{hyper}$ addresses the issue of low probability events and random associations. Other measures capable of handling low probability events have been introduced such as DuMouchel's empirical Bayes method.[27] A $lift_{hyper}$ value exceeding 1 indicates that the antecedent and consequent are positively correlated. $lift_{hyper}$ is defined as

$$lift_{hyper} = \frac{c_{xy}}{Q_\delta(C_{xy})}$$

where c_{xy} is the support count for the association of X and Y (i.e., number of MSMs where fragments X and Y co-occurred), and $Q_\delta(C_{xy})$ is the δ-quantile function for the hypergeometric probability distribution constrained by the support counts c_x and c_y for the antecedent and consequent, respectively. $Q_\delta(C_{xy})$ represents the support count for the co-occurrence of X and Y predicted by the cumulative, hypergeometric probability distribution given c_x, c_y, the number of MSMs in the database, and a specified cumulative probability threshold (δ).[25] The parameter δ is set to a conservative value of 0.99; therefore, the $lift_{hyper}$ for a rule consisting of independent item sets will randomly exceed 1 in only 1% of cases.

One difficulty that rises with the use of $lift_{hyper}$ is that it becomes undefined (i.e., infinity) when $Q_\delta(C_{xy})$ is zero. This can occur when very infrequent substructures are involved in an MSM. The fact that they occur at all in an MSM data set makes them highly relevant compared to what would be expected by random chance. These can constitute pairs of rare fragments that are truly correlated as well as infrequent fragments that happened to randomly occur together. To help differentiate these instances, we employ a secondary relevance measure called *all-confidence*, c_{all}.[28] *All-confidence* ranges from 0 to 1 and is defined as $c_{all} = c_{xy}/\max(c_x,c_y)$. c_{all} is less influenced by the size of the database being mined and provides a convenient mechanism for prioritizing MSMs where the $lift_{hyper}$ is undefined.

Referring to the example in Figure 4, the association between substructures A and C ($A \Rightarrow C$, $C \Rightarrow A$) are determined by first tabulating the support counts. A and C co-occur in 3 MSMs, so $c_{AC} = 3$. The substructures A and C occur in 7 and 4 MSMs, respectively ($c_A = 7$, $c_C = 4$). To compute $lift_{hyper}$, the denominator $Q_\delta(C_{AC})$ must be evaluated. $Q_\delta(C_{AC})$ is the inverse of the cumulative, hypergeometric probability distribution which describes the number of successes in a sequence of draws from a population without replacement. If the consequent substructure (C) is labeled as the "successful" outcome, then the number of unsuccessful outcomes is $10-4 = 6$. c_A is the sample size (i.e., number of draws). Given the c_C, $|D|$-c_C, c_A, and an example threshold δ value of 0.8, the quantile of the hypergeometric distribution yields 3 (i.e., number of "successful" outcomes predicted in c_A draws to achieve a cumulative probability of 0.8). $lift_{hyper}$ then equals 1 (3/3). c_{all} for c_{AC} is simply $3/7 = 0.43$.

Molecular Transformation Network. The lack of detailed annotation associated with compound registration is a significant impediment to gaining a better understanding of the lead optimization process within a pharmaceutical setting. It is effectively impossible to delineate the originating, lead optimization ontology, assuming that one actually exists. One alternative is to perform all-pair comparisons of molecules and assume that all pairs that fulfill specified transformation criteria represent a candidate transformation. This is the approach used by some automated bioisostere mining efforts.[14,15] It assumes that in the absence of a definitive chronological ontology that a suitable approximation can be achieved by establishing correspondences between molecular structures based on structural similarity.

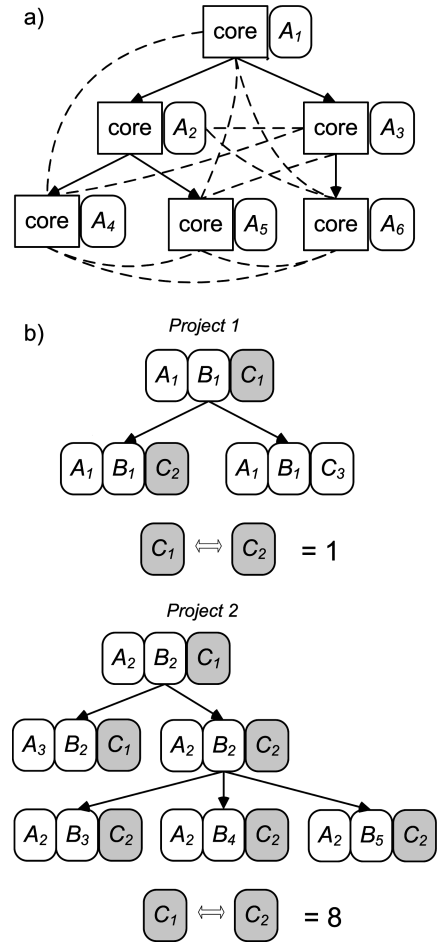


Figure 6. Inflated molecular substructure modification (MSM) support counts. (a) Example with one MSM pert MST. Solid lines indicate actual directed synthesis, and dotted lines denote artificially induced MSM occurrences. (b) Example involving MST containing more than one MSM.

However, there is a fundamental flaw associated with this approach. It ignores the practical realities involved in medicinal chemistry resource allocation. This presents the potential for inflated support counts for substructure pairs associated with over-represented compound series. Figure 6(a) illustrates a contrived example project consisting of six compounds. The synthesis effort consists of replacing the portion of the molecule indicated by A_i. The solid lines (with arrow heads) represent the true synthetic ontology of the project indicating the structural observations motivating subsequent synthesis. The dotted lines denote the pairwise correlations enforced by performing all pairwise comparisons. Since these artifact correlations scale by $O(n^2)$ where n is the number of analogues in a synthesis effort, using support counts as a surrogate for relevance nonlinearly, overweights synthesis efforts that have experienced greater resource allocation.

This effect is further compounded when more than one MSM is allowed per MST. Figure 6(b) illustrates this scenario. It depicts two hypothetical synthesis efforts. Project 1 represents a minimally resourced effort, and Project 2 typifies a larger project with more medicinal chemistry activity. In this diagram, A_i, B_j, and C_k indicate small portions of the molecule that have undergone substructural modification. In both projects, the discovery of the desirable C_1 to C_2 transformation was observed once. However, as an artifact of an all-pairs approach on a larger set of analogues, the second project overweights its support count by 8-fold. Thus, a naïve, pairwise approach can result in misleading values of relevance.

The result of an all-pairs comparison of molecules in a database can be represented as a molecular transformation network where each molecule represents a node in the network. A link exists between two nodes if the associated molecule pairs fulfill the established criteria for a molecular transformation. This results in a network with too many links. The approach taken here is to reduce the number of links in the transformation network while maintaining reachability of the nodes. To accomplish this, techniques from the field of computational geometry are exploited.

Gabriel[29] and proximity[30] graphs represent distance relationships between objects. For the molecular transformation network, the distance, d_{ij}, between each node (molecule) is assigned a value $d_{ij} = 1 - E(MCS)/(E(M_i) + E(M_j) - E(MCS))$, where $E(MCS)$, $E(M_i)$, and $E(M_j)$ are the number of bonds in the MCS, M_i, and M_j, respectively. This distance is one minus the Tanimoto similarity which obeys the metric triangle inequality.[31] A Gabriel graph, G_G, consists of the same node set as the parent graph G. The distinction between G_G and G lies in the edge set. An edge exists between two nodes i and j in G_G if and only if $d_{ij} \leq \sqrt{((d_{ik})^2 + (d_{jk})^2)}$ in G for all nodes k where $k \neq i$ and $k \neq j$. The proximity graph (G_P) is also defined on the node set of G; however, it has a more restrictive definition of the edge set and is given by $d_{ij} \leq \max(d_{ik}, d_{jk})$.

Figure 7 illustrates the properties of Gabriel and proximity graphs. The graph in Figure 7(a) depicts a random graph with 25 nodes and a uniform edge probability of 0.5. The respective Gabriel and proximity graph transforms assume the edges linking the nodes are weighted by 2D Euclidean distance. Both edge reduction techniques can significantly reduce the density of the parent graph while also maintaining much of the relative reachability between nodes. In general, the number of edges follow the trend $E(G_P) \leq E(G_G) \leq E(G)$.

The proximity graph transformation has been used previously used in structure−activity relationship (SAR) visualization;[32] however, we have adopted the Gabriel graph as the network reduction transformation operation. The reason for this is that the proximity graph is significantly more restrictive with respect to a node's neighborhood connectivity. Since the network represents an assumed relationship between two molecules based solely on structural commonality, it is possible for anomalous links to arise whereby a molecule may be more similar to another molecule yet synthetic intent may dictate that another slightly less similar neighbor may be a more appropriate link. To address this effect, the Gabriel graph was used as the adjacency criterion.

Figure 8 depicts an example of Gabriel graph reduction of a molecular transformation network. For the purposes of the example, the molecular structures in the NCI Open Database Compound collection were passed through the MSM rule generating process, and a small component of ten molecular structures was selected. Due to the mutual similarity between the molecules, the component was a complete graph whereby every node is connected to every other node (Figure 8(a)). Figure 8(b) demonstrates the significant reduction in the number of links in the component following the Gabriel graph procedure. The primary effect of this procedure is not that it exactly replicates the most plausible medicinal chemistry ontology but rather that it prevents large clusters of analogous compounds from overwhelming the information content in MSM space for the entire data set.

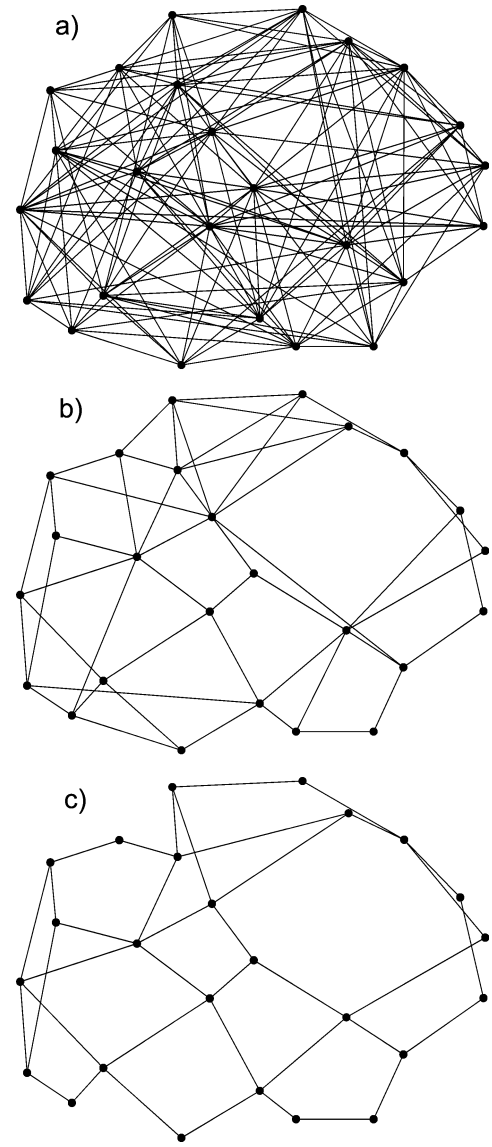


Figure 7. Depictions[33] of example network transformations: (a) random graph G, (b) Gabriel graph G_G, and (c) proximity graph G_P.

Experimental Procedure. An all-pairs network was constructed for each MSM extraction method (i.e., *default*, *add_connections*, and *fused_systems*) by performing an MCS comparison between each pair of molecules in the data set of 2,691,787 molecular structures, covering approximately 3.6×10^{12} comparisons per extraction simulation.[34] For each simulation, all pairs of molecules fulfilling the MST structural

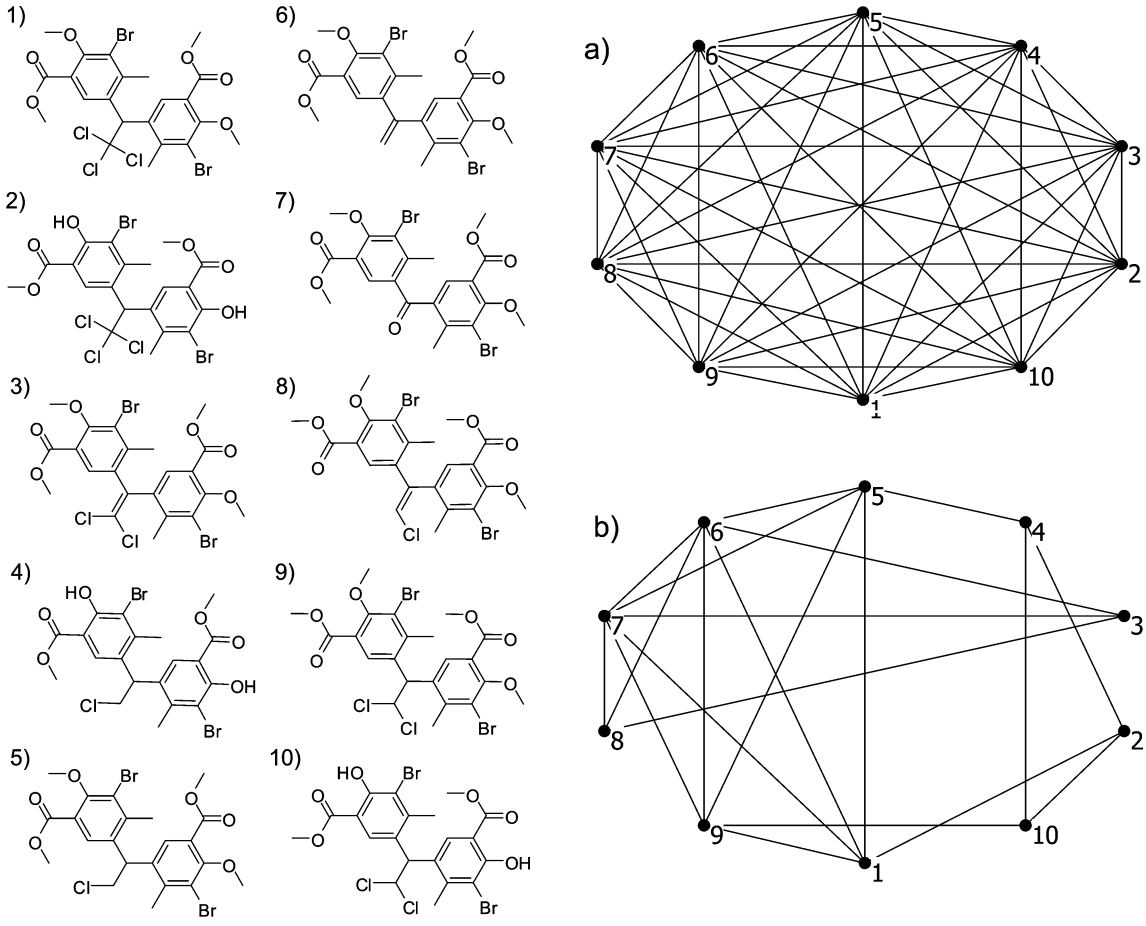


Figure 8. Example MST network. The network was constructed by performing all pairwise comparisons of structures in the NCI Open Database Compound collection: (a) original network using the MCS-based criteria and (b) Gabriel graph representation of the network.

commonality criteria were determined. These MST molecule to molecule networks were then reduced to their Gabriel graph counterparts, reducing on average the number of links in the network by 82%. Following the determination of the Gabriel graph reduction, all MSMs comprising each MST (i.e., network link) were identified and tabulated. This resulted in more MSMs than MSTs since it is possible for each MST to be comprised of more than one MSM. The database of MSMs was then evaluated using the association rule mining scheme.

In order to be considered for rule generation, both of the fragments in the MSM were required to have a minimum item support level of 5 occurrences, and the minimum support level for the co-occurrence of the antecedent and consequent was also set to 5. All transactions not containing a fragment fulfilling the minimum item support level were attributed to noise and were removed from the database. Due to the nature of an algorithmic fragmentation scheme, anomalous fragments can be generated during the MCS process, and a prefiltering based on minimum support counts eliminates most of them and provides efficiency improvements. Following rule generation, the number of unique fragments represented in each rule set was determined.

The results of the three extraction simulations were then combined into a master database to serve as the basis for an expert medicinal chemistry system. The separate rule sets were joined by identifying each unique MSM rule as defined by its corresponding fragment pair. It is common for the same MSM rule to be generated independently by the different MCS extraction schemes. Whenever a rule was present in more than one rule set, the rule with the highest $lift_{hyper}$ was added to master rule set. This is a common data fusion technique.[35] For duplicate rule occurrences where at least one rule occurrence had an undefined $lift_{hyper}$ (*i.e.*, infinite) and at least one occurrence had a real valued $lift_{hyper}$, the rule possessing the highest real valued $lift_{hyper}$ was added to the master rule set. In the event that a rule could not be selected solely on the $lift_{hyper}$ criterion, the rule with the highest c_{all} value was used to as the deciding criterion for inclusion in the master rule set.

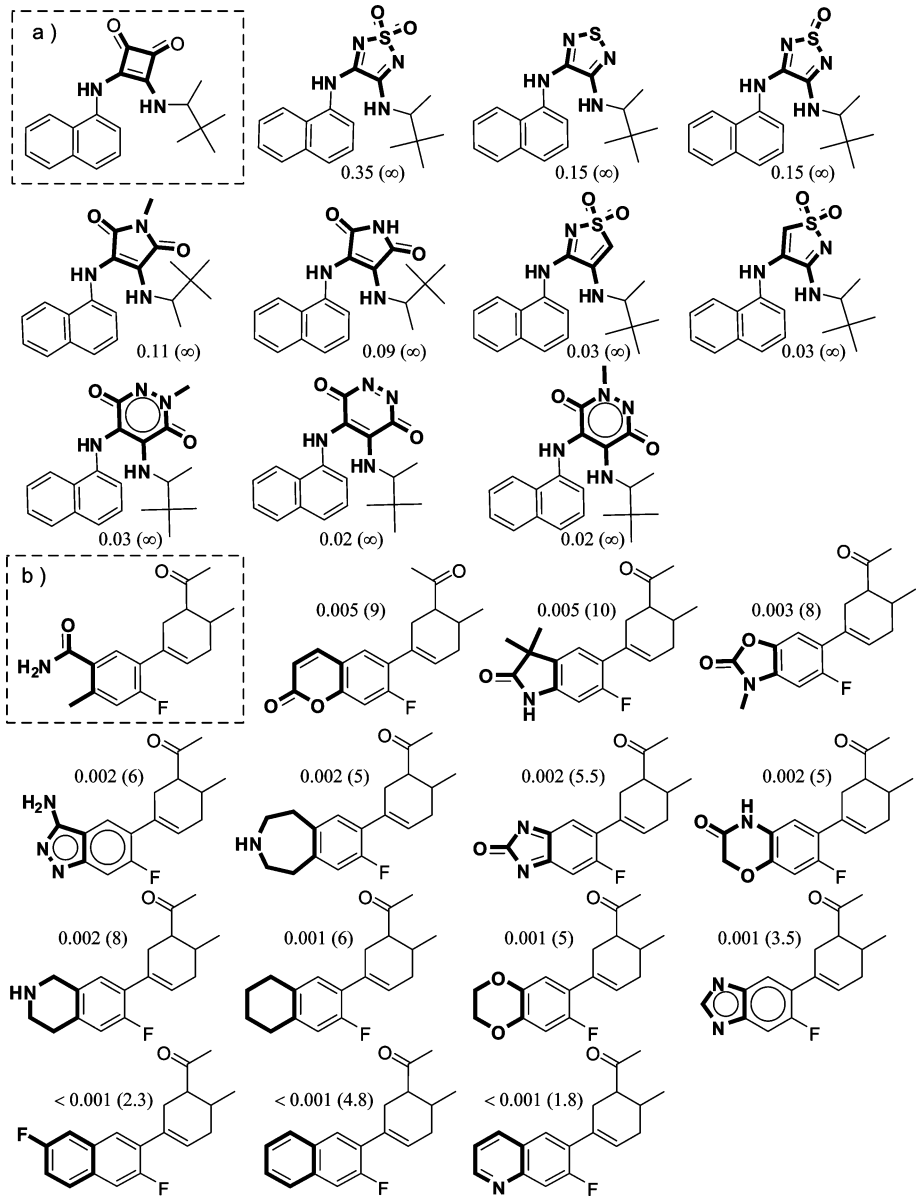
DISCUSSION

Table 1 summarizes the results of the three extraction and subsequent rule generation simulations. The number of rules generated where the $lift_{hyper}$ exceeded 1 are 404,582, comprising 95,454 unique fragments. It has been found that while $lift_{hyper}$ provides an effective measure for filtering MSM association rule by probabilistic relevance, it is not necessarily an optimal measure for MSM rule prioritization. It has been found that c_{all} often provides a more desirable prioritization of the generated association rules ($lift_{hyper} > 1$). The primary reason for this is that $lift_{hyper}$ is capable of discerning rare MSMs that are correlated as well as common MSMs that are also correlated, but it lacks the ability to distinguish between them. All-confidence, c_{all}, is better able to distinguish between these types of associations. It also has the tendency to rank MSM rules with higher absolute support counts in the Gabriel graph higher which tends to

Table 1. Results of the Three Extraction and Rule Generation Simulations and Subsequent Rule Set Merging Process

extraction method	no. of MST network links	no. of MST Gabriel graph links	no. of MSM	no. of rules ($lift_{hyper} > 1$)	no. of unique fragments in rules
default	72,225,088	12,965,103	14,617,265	163,463	34,092
add_connections	71,631,443	12,870,079	14,476,879	229,269	61,589
fused_systems	70,289,291	12,499,650	14,052,315	153,364	33,789
master rule set:				404,582	95,454

Chart 1. (a) Diaminocyclobutene MSM Query[a] and (b) Amide and Methyl Substitution Ring Closure MSM Query



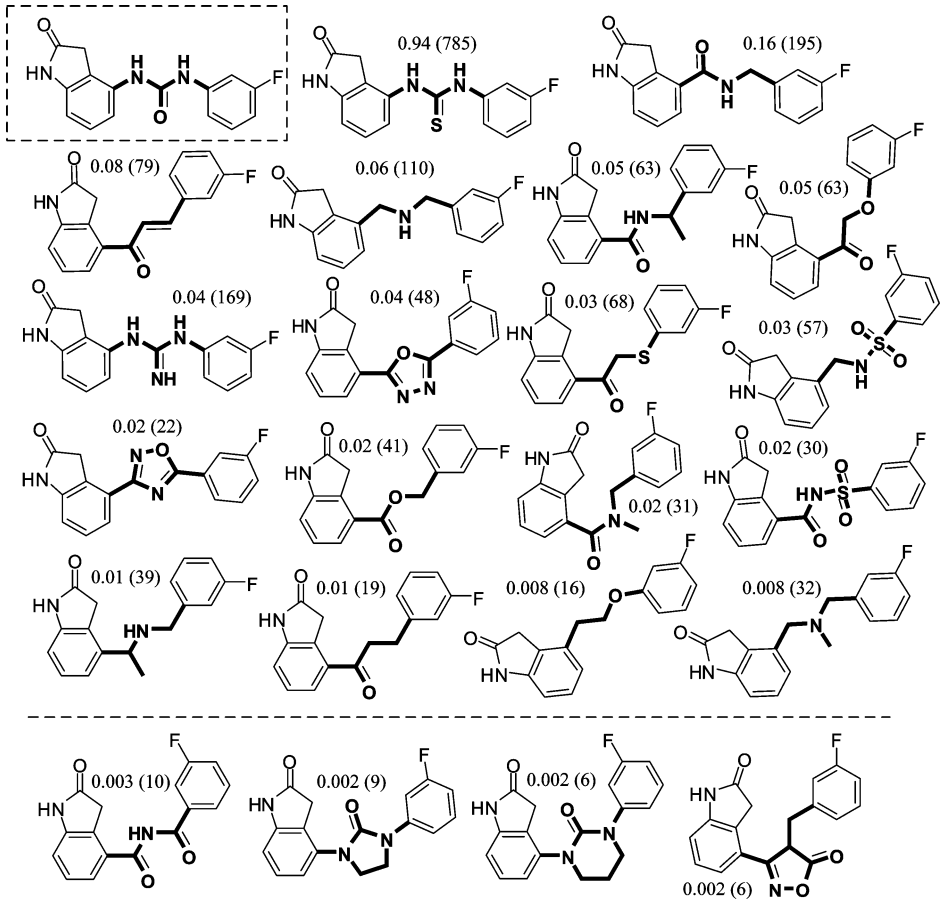
[a] MSM replacements ranked by c_{all}. $lift_{hyper}$ is given in parentheses.

provide a more intuitive presentation for visual inspection. It is important to note that the context of any prioritization score is only relevant to other MSM rules which share a substructure. Comparing relevance scores between MSM rules that do not share a fragment is not meaningful.

The primary reason that the number of generated rules and unique fragments does not appear to be reduced as much as one might expect when the three extraction simulations are merged is due primarily to the lack of overlap between the *add_connections* simulation with the *default* and *fused_systems* simulations. This is because the *add_connections* simulation by definition adds additional connection points during the extraction process, whereas the *default* and *fused_systems* simulation can have at most one connection point attached to any atom in an extracted fragment. The *add_connections* fragment may have multiple connection points incident on any atom in the extracted fragment, making it impossible to be isomorphic with any fragment in the *default* and *fused_systems* extractions.

Chart 1 illustrates the MSM prioritizations for two example queries. The first example demonstrates a linker replacement, and the second example represents a ring closure query. In the ring closure example, the MSM representing the transformation of the amide and methyl substitutions to a benzene

Chart 2. Urea Linker MSM Query[a]



[a] Top ranked MSMs are depicted above the dotted line in order of c_{all} prioritization. Due to space considerations, not all MSMs are depicted. Those below the dotted line represent a sampling of lower ranked MSMs. Note that due to the symmetry of the urea query fragment about the carbonyl, all asymmetric replacement fragments can be mapped in two ways. For concision, only one of the mappings is depicted.

ring ($c_{all} < 0.001$) is almost five times as frequent in the Gabriel graph as the two top ranked MSMs ($c_{all} = 0.005$), illustrating the difference between using frequency counts and measures of statistical relevance. Chart 2 depicts the top ranked MSMs associated with a urea linker query. Note that since the urea substructure is symmetric, all asymmetric replacements possess two mappings; however, for conciseness, only one mapping for each asymmetric substructure is depicted.

CONCLUSION

An algorithmic framework that condenses the historical knowledge base of chemical structure modification within a pharmaceutical context has been presented and implemented. The proposed framework has the ability to estimate the perceived "reasonableness" of a structure modification to a lead compound relative to how well it agrees with the aggregate medicinal chemistry community. This methodology has many potential uses from a drug discovery perspective. Some of the more straightforward applications include providing insights to medicinal chemistry teams during the lead optimization process as well helping address intellectual property considerations. The method can also be readily extended to address physical and biological end-point considerations (i.e., rule-based QSAR).

It is believed that the proposed framework may offer additional value by helping to provide insights into the way medicinal chemistry teams operate in aggregate and potentially offer strategic improvements to the process of iterative lead optimization. Accepting the analogy that lead development is an optimization process, it can be assumed that lead optimization is comprised of two governing forces, directed search (intellectual and technological insights) as well as purely random exploration (serendipity). The relative role of the two competing forces for successful project outcomes is not obvious. The process of lead optimization is an inherently stochastic process, governed to a large extent by random events (from the perspective of our ability to satisfactorily hedge against them). The approach presented here codifies much of the directed search component of lead optimization through the prioritization of "relevant" structure modifications.

It has been observed that in activities that are comprised of both skill and chance, human participants disproportionately attribute the observation of successful outcomes to skill thereby attenuating the role of random influences. This effect is known as the "illusion of control".[36,37] The proposed, prioritization strategy provides an exploratory foundation for investigating the relative roles of random exploration and directed chemical structure modification, so that operational improvements may be made to the process of lead optimization. It is believed that project specific learnings can

potentially be augmented with more general medicinal chemistry experience leading to improved tactical decision making.

ACKNOWLEDGMENT

The authors extend their gratitude to the reviewers for their helpful comments and recommendations, and we thank our colleague, Jeffrey Sutherland, for helpful suggestions for improving the manuscript.

Supporting Information Available: The first generation MSM data set has been incorporated into an early prototype hypothesis generation utility, Molecular Morphing Map (M3), and a brief overview of the tool. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15* (10), 1006–1011.
(2) Sneader, W. *Drug Prototypes and Their Exploitation*; John Wiley & Sons: Chichester, 1996; p 788.
(3) Chen, X.; Wang, W. The Use of Bioisosteric Groups in Lead Optimization. *Annu. Rep. Med. Chem.* **2003**, *38*, 333–346.
(4) Wermuth, C. G. Molecular Variations Based on Isosteric Replacements. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, 1996; pp 203−236.
(5) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147–3176.
(6) Olesen, P. H. The Use of Bioisosteric Groups in Lead Optimization. *Curr. Opin. Drug Discovery Dev.* **2001**, *4* (4), 471–478.
(7) Fujita, T. Similarities in Bioanalogous Structural Transformation Patterns among Various Bioactive Compounds Series. *Biosci., Biotechnol., Biochem.* **1996**, *60*, 557–566.
(8) Zych, A. J.; Herr, R. J. Tetrazoles as Carboxylic Acid Bioisosteres in Drug Discovery. *PharmaChem* **2007**, *6* (4), 21–24.
(9) Rosen, T.; Nagel, A. A.; Rizzi, J. P.; Ives, J. L.; Daffeh, J. B.; Ganong, A. H.; Guarino, K.; Heym, J.; McLean, A.; Nowakowski, J. T.; Schmidt, A. W.; Seeger, T. F.; Siok, C. J.; Vincent, L. A. Thiazole as a Carbonyl Bioisostere. A Novel Class of Highly Potent and Selective 5-HT3 Receptor Antagonists. *J. Med. Chem.* **1990**, *33* (10), 2715–20.
(10) Bailey, T. R.; Diana, G. D.; Mallamo, J. P.; Vescio, N.; Draper, T. L.; Carabateas, P. M.; Long, M. A.; Giranda, V. L.; Dutko, F. J.; Pevear, D. C. An Evaluation of the Antirhinoviral Activity of Acylfuran Replacements for 3-Methylisoxazoles. Are 2-Acetylfurans Bioisosteres for 3-Methylisoxazoles. *J. Med. Chem.* **1994**, *37* (24), 4177–84.
(11) Fludzinski, P.; Evrard, D. A.; Bloomquist, W. E.; Lacefield, W. B.; Pfeifer, W.; Jones, N. D.; Deeter, J. B.; Cohen, M. L. Indazoles as Indole Bioisosteres: Synthesis and Evaluation of the Tropanyl Ester and Amide of Indazole-3-Carboxylate as Antagonists at the Serotonin 5HT3 Receptor. *J. Med. Chem.* **1987**, *30* (9), 1535–7.
(12) Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: A Computer Software Program for Drug Design Using Medicinal Chemistry Rules. *Bioorg. Med. Chem.* **2006**, *14* (20), 7011–7022.
(13) Hong, L.; Page, S. E. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (46), 16385–16389.
(14) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(15) Southall, N. T. Ajay Kinase Patent Space Visualization Using Chemical Replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.
(16) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of Target-Specific Bioisosteric Fragments from Ligand-Protein Crystallographic Data. *J. Comput.-Aided Mol. Des.* **2006**, *20* (6), 385–394.
(17) Raymond, J.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching Of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 521–533.
(18) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comp. J.* **2002**, *45* (6), 631–644.
(19) Raymond, J.; Gardiner, E.; Willett, P. Heuristics for Rapid Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305–316.
(20) Downs, G. M. Ring Perception. In *Encyclopedia of Computational Chemistry*; Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Shreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; Vol. 4, pp 2509−2515.
(21) Zhou, L.; Yau, S. Efficient Association Rule Mining among both Frequent and Infrequent Items. *Comput. Math. Appl.* **2007**, *54*, 737–749.
(22) Tan, P. N.; Kumar, V.; Srivastava, J. Selecting the Right Objective Measure for Association Analysis. *Inf. Syst.* **2004**, *29*, 293–313.
(23) Wu, T.; Chen, Y.; Han, J. Association Mining in Large Databases: A Re-examination of Its Measures. In *Lecture Notes Computer Science*; Springer: Berlin/Heidelberg, Germany, 2007; Vol. 4702, pp 621−628.
(24) Ellis, D.; Furner-Hines, J.; Willett, P. Measuring the Degree of Similarity between Objects in Text Retrieval Systems. *Perspect. Inform. Manage.* **1993**, *3* (2), 128–149.
(25) Hahsler, M.; Hornik, K. New Probabilistic Interest Measures for Association Rules. *Intell. Data Anal.* **2007**, *11* (5), 437–455.
(26) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343–357.
(27) DuMouchel, W.; Pregibon , D. Empirical Bayes Screening for Multi-Item Association. In *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM: San Francisco, CA, 2001; pp 67−76.
(28) Omiecinski, E. R. Alternative Interest Measures for Mining Associations in Databases. *IEEE Trans. Knowl. Data Eng.* **2003**, *15* (1), 57–69.
(29) Gabriel, K. R.; Sokal, R. R. A New Statistical Approach to Geographic Variation Analysis. *Syst. Zool.* **1969**, *18*, 259–270.
(30) Jaromczysk, J. W.; Touissaint, G. T. Relative Neighborhood Graphs and Their Relatives. *Proc. IEEE* **1992**, *80* (9), 1502–1517.
(31) Lipkus, A. H. A Proof of the Triangle Inequality for the Tanimoto Distance. *J. Math. Chem.* **1999**, *26* (1−3), 263–265.
(32) Johnson, M. Structure-Activity Maps for Visualizing the Graph Variables Arising in Drug Design. *J. Biopharm. Stat.* **1993**, *3* (2), 203–236.
(33) *NetDraw: Graph Visualization Software*; Borgatti, S. P.; Harvard: Analytic Technologies: 2002.
(34) IBM BladeCenter HS21 Cluster; Xeon dual core 3.0 GHz; 1072 cores.
(35) Ginn, C. M.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1–16.
(36) Langer, E. J. The Illusion of Control. *J. Pers. Soc. Psychol.* **1975**, *32* (2), 311–328.
(37) Langer, E. J.; Roth, J. Heads I Win, Tails It's Chance: The Illusion of Control as a Function of the Sequence of Outcomes in a Purely Chance Task. *J. Pers. Soc. Psychol.* **1975**, *6* (6), 951–955.

CI9000426