ELSEVIER

Synergy between combinatorial chemistry and *de novo* design

Andrew R. Leach, Richard A. Bryce,[1] and Alan J. Robinson[2]

Glaxo Wellcome Medicines Research Centre, Hertfordshire, United Kingdom

Traditional de novo *design algorithms are able to generate many thousands of ligand structures that meet the constraints of a protein structure, but these structures are often not synthetically tractable. In this article, we describe how concepts from structure-based* de novo *design can be used to explore the search space in library design. A key feature of the approach is the requirement that specific templates are included within the designed structures. Each template corresponds to the "central core" of a combinatorial library. The template is positioned within an acyclic chain whose length and bond orders are systematically varied, and the conformational space of each structure that results (core plus chain) is explored to determine whether it is able to link together two or more strongly interacting functional groups or pharmacophores located within a protein binding site. This fragment connection algorithm provides "generic" 3D molecules in the sense that the linking part (minus the template) is built from an all-carbon chain whose synthesis may not be easily achieved. Thus, in the second phase, 2D queries are derived from the molecular skeletons and used to identify possible reagents from a database. Each potential reagent is checked to ensure that it is compatible with the conformation of its parent 3D conformation and the constraints of the binding site. Combinations of these reagents according to the combinatorial library reaction scheme give product molecules that contain the desired core template and the key functional/pharmacophoric groups, and would be able to adopt a conformation compatible with the original molecular skeleton without any unfavorable intermolecular or intramolecular interactions. We discuss how this strategy compares with and relates to alternative approaches to both structure-based library design and* de novo *design. © 2000 by Elsevier Science Inc.*

Keywords: de novo *design, structure-based design, combinatorial libraries*

INTRODUCTION

There has been much discussion concerning the most effective way to use high-throughput screening and combinatorial chemistry techniques as part of the drug-discovery process. The introduction of these two technologies led some to believe that "numbers alone" would be the way forward, and all that was required was the synthesis and testing of large numbers of compounds. In favorable circumstances, combinatorial chemistry methods do offer a way to generate a large number of molecules at a much lower unit cost than making the equivalent number of individual compounds. However, this in itself is of little value unless the molecules produced have some useful activity.

It now is generally recognized that one stands a much better chance of finding an active molecule (or series of molecules) if one can incorporate relevant information and knowledge about the biological target during the library design phase.[1] The key then is to achieve an appropriate degree of chemical diversity while simultaneously incorporating this knowledge. Here we will be concerned with the situation where one has available the structure of the biological target in the form of an X-ray or nuclear magnetic resonance (NMR) structure (or possibly an homology model). This represents a significant constraint on the types of molecules that one might wish to synthesize and, as such, in principle should enable the search space to be reduced dramatically.

One possible approach to the problem of structure-based library design would be to first enumerate all possible molecules that could be synthesized from readily available starting materials, using established libraries chemistries. Each of these suggested molecules then would be docked into the active site of the target protein, using a program such as DOCK[2] and scored using an appropriate estimation of the binding free energy (of which there are several available[3]). Libraries would be selected for synthesis according to the number of "active" molecules they contained (on the basis of the docking calcu-

Color Plates for this article are on page 526.

Corresponding author: Andrew R. Leach, Glaxo Wellcome Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom. Tel.: 44-0-1438-763383 *E-mail address*: arl22958@glaxowellcome.co.uk (A.R. Leach).

Presented in part at the Molecular Graphics and Modelling Society meeting on Combinatorial Chemistry at Sheffield, 1998. Present addresses: [1]Department of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom. [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.

Journal of Molecular Graphics and Modelling 18, 358–367, 2000
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00062-0

lations and scoring function) and on their ability to satisfy the combinatorial constraint (i.e., combinations of monomers can be found that, when combined in a combinatorial fashion, give rise to an acceptable number of active product molecules).

Such an approach has the advantage of being simple to implement, but suffers from the practical drawback that the size of the virtual library to be docked often can run to many millions (if not billions) of compounds, even if the enumerated products are subjected to a severe set of 2D filters (to remove undesirable molecules) prior to the more time-consuming docking stage. Current algorithms are not able to perform such large numbers of independent dockings within a reasonable time scale.

An alternative way to tackle the combinatorial problem is to separately dock individual monomers and then identify those combinations of monomers that are in close proximity within the binding site.[4] This requires that the docking of a product molecule can be constructed from the individual dockings of each isolated monomer, which is not guaranteed. A related approach is to dock the library "core" into the active site and then add the various R-groups independently. Having determined which R-groups can fit, only then are combinations of R-groups considered, in order to identify which combinations could give rise to acceptable dockings of the whole product molecule. We[5] and others[6–8] have described variants on this strategy. Such approaches are, of course, dependent on the initial docking of the template. In those cases where one is exploring just one R-group based on a fairly large template or where the template corresponds to some group that covalently bonds to the protein, then this is probably quite a reasonable approach. However, in other cases, the template may not make any significant interactions with the protein, and a large number of template dockings distributed throughout the binding site may be the only way to guarantee coverage.

In this article, we describe an alternative approach that tackles the time-consuming 3D docking problem through the use of *generic structures.* The use of generic structures is popular in approaches to structure-based *de novo* design. A number of comprehensive reviews of structure-based *de novo* design have been published in recent years.[9–11] Also pertinent is the article by Lewis and Leach,[12] who described two categories of *de novo* design algorithms: "inside-out" or "outside-in." The "core and R-group" approach to library design can be considered an example in the former category, where one starts from within the molecule (i.e., the core template) and works outward. In this article, we are concerned with the "outside-in" approach, where one starts with a set of disconnected fragments that interact with the protein and determines how they might be connected together.

The first step in the "outside-in" approach is to identify regions where specific functional groups or chemical entities would be expected to show a strong interaction. Many methods are available to achieve this, including GRID,[13] MCSS,[14] and LUDI.[15] The binding modes of appropriate groups also may be deduced from experimental structures of protein-ligand complexes. In addition, it may be possible to determine the locations of weakly binding small molecules using X-ray crystallography[16] or NMR.[17] One then wants to identify chemically tractable molecules that link these groups in an appropriate low-energy conformation. One approach is to search a database of existing molecules or of molecular fragments, as in programs such as CAVEAT[18] or HOOK.[19] This approach can be powerful, but the resulting suggestions can only ever reflect the contents of the database. An alternative is to attempt to design the connecting molecules from scratch. It is typical to adopt a two-stage procedure here, as described by Lewis and Dean.[20,21] The first stage involves the construction of generic "skeletons" that connect together the fragments. These skeletons are often built solely from carbon atoms. In the second stage these skeletons must then be converted into "real" molecules, typically by assigning atom types to the skeleton.

A variety of skeleton-construction methods have been described. The earliest algorithms produced acyclic linkers. Lewis and Dean described the use of two-dimensional[20,21] and three-dimensional[22] regular lattices. Lewis used a ring-closure algorithm to produce a more general solution,[23] whereas Leach and Kilvington[24] used the "tweak" algorithm. An alternative is to use an irregular lattice derived from collections of molecules docked into the active site.[25] Ring-bracing algorithms enable theses acyclic chains to be converted into cyclic structures.[26] More recently, Todorov and Dean[27] described a more generic template-based approach that uses simulated annealing to identify scaffolds that can meet geometric and connectivity constraints. The SPROUT program also uses a generic template library, but here an exhaustive, systematic exploration of the space is performed.[28]

Some interesting approaches to the "atom assignment problem" have been described. The algorithm of Chan, Chau, and Goodman[29] calculates atomic partial atomic charges that optimize the electrostatic complementarity between ligand and protein. Barakat and Dean[30–34] published a series of articles that placed fragments onto a 3D molecular graph, using simulated annealing. A related branch-and-bound algorithm was described by Todorov and Dean.[35] These approaches were primarily intended to optimize either the electrostatic or the hydrophobic complementarity with the protein, rather than address the problem of ease of synthesis. There is no guarantee that appropriate reagents would be available to enable the synthesis of the target molecule to be easily addressed. In reality, synthetic tractability often is assessed via visual inspection of the program's suggestions by the computational chemist working alongside a trained synthetic or medicinal chemist. This can be a time-consuming process. Moreover, it often may require a considerable degree of compromise between the structure suggested by the computer and a molecule that could actually be synthesized.

The approach that we describe in this article is a form of *de novo* design directed toward those molecules for which a chemical synthesis should be relatively straightforward. Specifically, the program only considers molecular skeletons that can be made using combinatorial chemistry methods (and which therefore should be relatively amenable to synthesis). These partial skeletons then form the basis for substructural searches for possible starting materials from a chemical database, so establishing the link between the theoretical skeleton and the "real world" of molecules that could be made from available starting materials. A flow chart of the overall scheme is shown in Figure 1. We now describe in more detail these two key steps, fragment connection and reagent identification.

FRAGMENT CONNECTION ALGORITHM

The usual starting point for the fragment connection algorithm is a collection of functional groups distributed in space, within

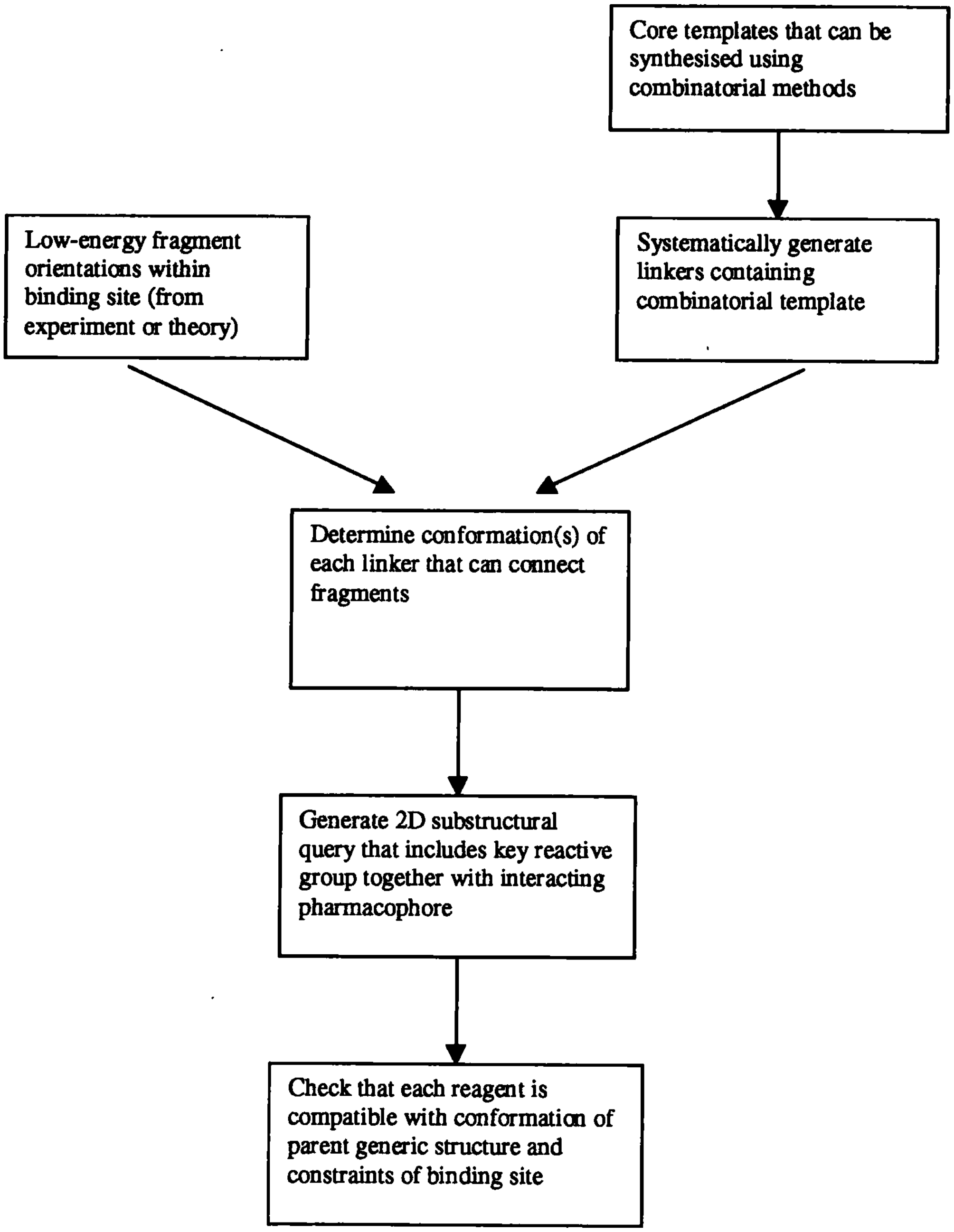
Core templates that can be synthesised using combinatorial methods
Low-energy fragment orientations within binding site (from experiment or theory)
Systematically generate linkers containing combinatorial template
Determine conformation(s) of each linker that can connect fragments
Generate 2D substructural query that includes key reactive group together with interacting pharmacophore
Check that each reagent is compatible with conformation of parent generic structure and constraints of binding site

Figure 1. Flow chart of the overall scheme.

a binding site of known three-dimensional structure. These functional groups may be obtained from X-ray or nuclear magnetic resonance (NMR) experiments or from calculations of the type described earlier. We also use an in-house Monte Carlo minimization search procedure similar in spirit to MCSS, which can suggest low-energy locations of functional groups in protein binding sites. The key attribute of our approach is the use of precomputed grids of the electrostatic and van der Waals potential, from which it is possible to rapidly calculate the interaction energy of a fragment within a binding site. Such potential grids have been used elsewhere for scoring orientations generated by the DOCK program[36] and in a molecular dynamics docking procedure.[37] Our implementation is based on the OPLS force field, in part because its van der Waals term is more suited to the use of precalculated potential fields due to its use of the geometric mean, $\surd(\sigma_{AA}\sigma_{BB})$ for the Lennard-Jones σ parameter for the interaction between two dissimilar atoms A and B rather than the arithmetic mean, $1/2(\sigma_{½AA}+\sigma_{BB})$.

Given a set of functional group orientations (together with their specified connection atoms), the algorithm then attempts to identify candidate structures that can position the functional groups in the desired relative orientation, in a low-energy conformation, and without experiencing any unfavorable interactions with the protein (it should be noted that the methods here also can be used starting from a 3D pharmacophore). The algorithm is represented schematically in Figure 2. Crucial to the approach is a library of "core" templates. These represent

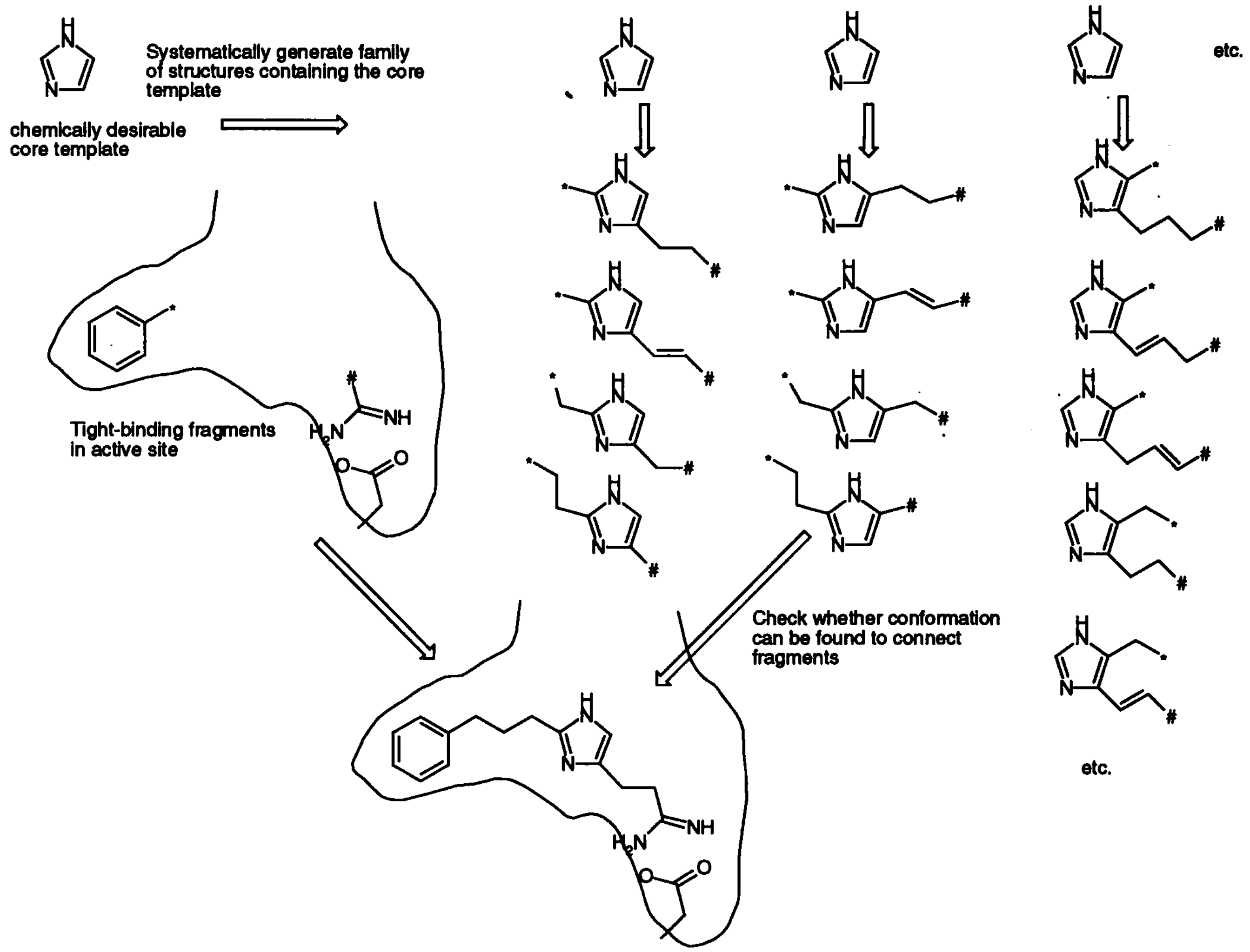
Systematically generate family of structures containing the core template
chemically desirable core template
Tight-binding fragments in active site
Check whether conformation can be found to connect fragments
etc.
etc.

Figure 2. Schematic illustration of the structure-generation phase. A "core template" (chosen to be an imidazole ring for the purposes of illustration) that can be readily synthesised by combinatorial means is incorporated into a series of acyclic chains. The position of the fragment within the chain, the length of the chain, and the bond orders in the chain are all explored systematically (note that only three of the possible substitution patterns are shown together with just a few of the linker sequences for each case). The conformational space of each chain is checked to determine whether it can link together the fragments in the binding site.

key common substructures that can be synthesized by combinatorial methods. An increasing number of such core structures now are available and have been described in the literature; these range from the ubiquitous amide group through aromatic systems such as imidazoles[38] to much larger cores such as the benzdiazepine template.[39]

Each of the core templates may have an associated list of permitted substitution patterns, reflecting the synthetically feasible alternatives. If no such constraints are imposed, the algorithm automatically generates all symmetrically unique substitution patterns. For example, there are three ways to substitute two groups in a benzene ring (1,2; 1,3; 1,4) and 10 ways to substitute three groups (1,2,3; 1,2,4; 1,2,5; 1,2,6; 1,3,2; 1,3,4; 1,3,5; 1,3,6; 1,4,2; 1,4,3). The imidazole synthesis of Bilodeau and Cunningham[38] can give rise to substitution on the three carbon atoms in the ring, and so there would be six different ways to attach two substituents (2,4; 2,5; 4,5; 4,2; 5,2; 5,4). For this template, there also are six unique ways to attach three different substituents. For each substitution pattern, a series of trial skeletons is constructed first by appending acyclic bonds with normal bond lengths and angles and then finally joining the functional groups. The total number of bonds inserted varies between a minimum and maximum value specified by the user. For a given number of bonds, the template is inserted into all possible places along the chain, as shown in Figure 2. By default, the linker chains are constructed using just single bonds, but, if required, additional linkers may be constructed by systematically inserting double bonds (and triple bonds if so desired) into the chain, provided they give a chemically sensible result (i.e. no more than one multiple bond per atom). The linkers are constructed from carbon atoms that are assumed to be appropriately representative of other heteroatoms, at least from the first row of the periodic table. In this way, the algorithm systematically generates all possible linker "molecules" containing the template, as shown in Figure 2.

Each trial structure initially is generated in an arbitrary conformation by the model-building component of our COBRA program.[40] The next stage involves an exploration of the conformational space of the molecule to try and determine whether or not a reasonable conformation can be found that enables the linker to connect the functional groups in their desired orientations within the binding site. To facilitate this, a series of distance constraints is established between pairs of atoms in the fragments to be joined. The atoms that are used to derive these distance constraints may be specified by the user but usually are determined automatically by the program. Before performing any conformational manipulation, however, the algorithm first checks that the skeleton is consistent with the geometrical constraints by performing triangle smoothing to generate upper and lower bounds for the interatomic distances. If the upper bound of any one pair of distances is less than the desired distance in the active site, then clearly that particular skeleton cannot fulfill the distance constraints and so the algorithm moves on to the next linker.

It is possible to use one of two different algorithms for searching the conformational space in order to connect the fragments with the trial structure. One approach uses the directed tweak method of Hurst.[41] The directed tweak algorithm provides a rapid way to modify the conformation of an acyclic chain in order to satisfy a set of distance constraints. The directed tweak algorithm is closely related to the tweak algorithm[42] that we used in some of our earlier approaches to *de novo* design.[24] These algorithms calculate the derivative of each distance with respect to the torsion angles of the intervening rotatable bonds, from which it is possible to calculate by how much each torsion angle must change to satisfy the constraints. A crucial feature of the tweak algorithms is that their complexity varies with the number of constraints rather than the number of rotatable bonds. The success of the procedure depends on the starting conformation, and so a number of random conformations are generated and "tweaked" in order to cover the conformational space. This also enables the algorithm to search for alternative solutions that can satisfy the constraints.

The second conformational searching method employs the fragment-building method implemented in COBRA, an approach that has been described in a number of publications.[40,43,44] In the current context, we use the COBRA algorithm to generate low-energy conformations of the structure. Analyses of protein-ligand complexes indicate that the conformations of small-molecule ligands are generally quite close to the local energy minima that COBRA is designed to generate.[45,46] However, some deviation from the "ideal" torsion angles in individual bonds inevitably is present. This could lead to a high number of failures if only the "near minimum-energy" conformations generated by COBRA were considered. Should a COBRA conformation fail to satisfy the constraints, it is subjected to a maximum of two iterations of the tweak method (with each rotatable bond able to change by a maximum of 10° per iteration) to determine whether this slight modification does enable it to satisfy the the imposed distance constraints.

Having successfully generated a conformation of the linker molecule that satisfies the distance constraints, it is positioned in the site so that the fragment groups are overlaid on their initial orientations, using a least-squares fitting procedure. If the root mean square (RMS) fit is below a specified threshold (typically 0.5–0.75Å), the structure is checked further to ensure that it does not represent a high-energy internal conformation nor that it interacts unfavorably with the receptor. Simple distance-based "bump checks" are used to achieve this. It also can be useful to calculate the buried surface area of the structures and to use this to filter out structures that are particularly exposed to solvent. Conformations that meet all these criteria are output for the next part of the procedure.

Very similar conformations may be produced by these methods (especially the directed tweak approach). The conformations for each skeleton are subjected to a cluster analysis in order to identify a suitably representative set of structures. The representative for each cluster is chosen as the molecule with the lowest internal energy.

To illustrate the fragment connection algorithm, we consider the much examined but familiar case of methotrexate bound in the active site of dihydrofolate reductase. We used as our two fragments the pterin ring system and the carboxyl group that interacts with Arg57, taken from the X-ray structure[47] (PDB file 4dfr). The objective then was to determine how these fragments might be connected via a benzene template (assumed for the purposes of this illustration to correspond to the key "combinatorial" functionality) within the confines of the active site. All three substitution patterns around the ring were considered (i.e., ortho, meta, and para) and the path between the two fixed fragments was allowed to vary between 7 and 9 atoms (methotrexate itself has a path of 9 atoms). Of a total 476 trial molecules, 208 had an inappropriate bond order sequence (both single and double bonds were considered). Of the remaining 268, one or more acceptable conformations could be found for 47 structures (a total of 94 different conformations), using the COBRA-based search. An overlay of some of the these generic structures is shown in Color Plate 1. Also shown is the conformation that most closely corresponds to the actual binding mode of methotrexate; a double bond is used as a geometrical isostere for the amide bond in methotrexate.

REAGENT SEARCHING ALGORITHM

The linker-generation phase typically gives rise to a range of structures, in each of which the functional groups are linked together via the combinatorial template. It is not necessary for this core template to make any interactions itself with the receptor, although if this is possible it may be useful to consider it when comparing different libraries. The next stage of the procedure involves the identification of potential reagents from which each of these structures could be synthesized. The structural input to this second phase need not necessarily be derived from our fragment connection algorithm, but could come from an existing lead compound or series of lead compounds. The reagent searching algorithm is illustrated schematically in Figure 3.

The starting point for the reagent searching algorithm is the 3D conformation (or conformations) of a specific structure within the binding site. From the associated 2D structure, we first retrosynthetically break the molecule into its constituent pieces according to the combinatorial chemistry scheme. We use the Daylight reaction toolkit[48] to achieve this. Where the template is obtained from a bimolecular reaction then this would give two structures; where the template is synthesized from a trimolecular reaction then three structures would result. For example, if the linking template is an amide, the resulting monomers would comprise a primary or secondary amine and

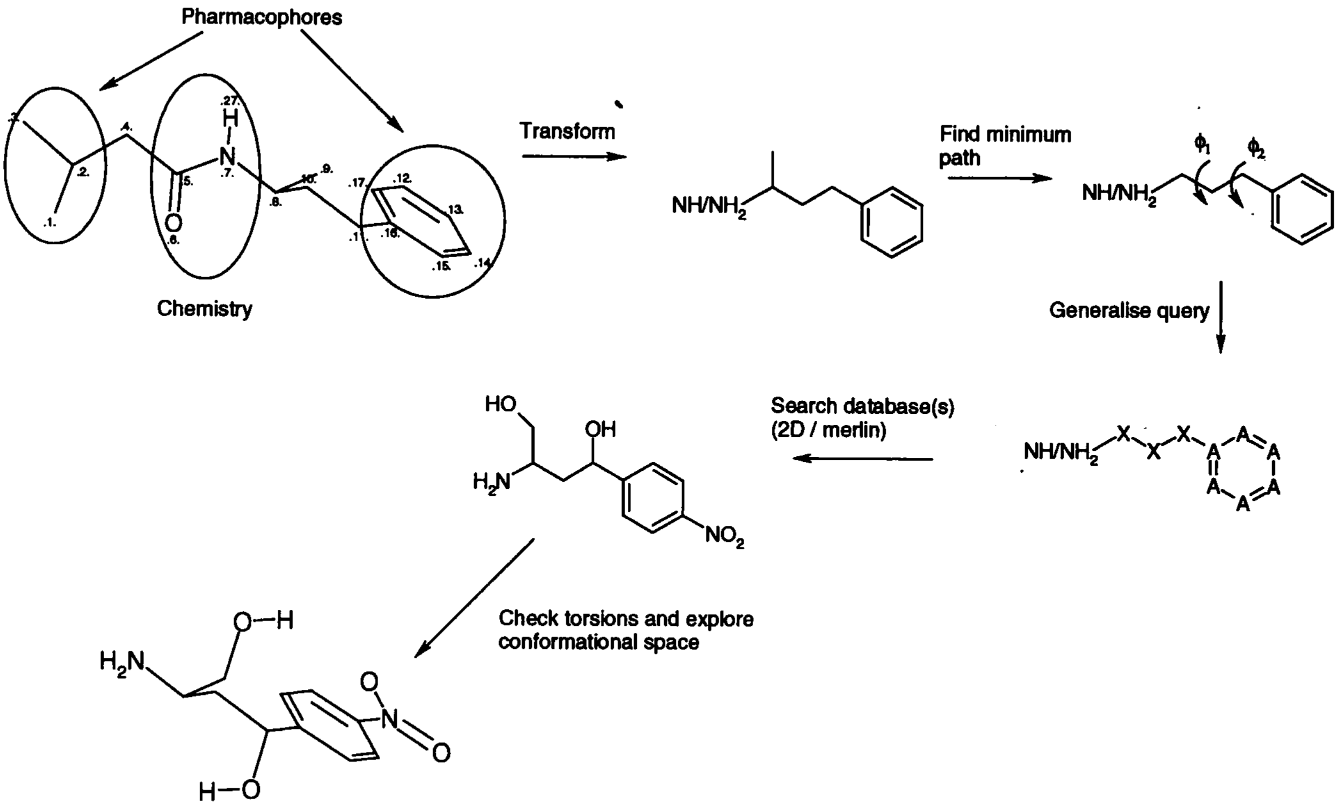

Figure 3. Reagent searching algorithm, illustrated using an amide "core template," which is decomposed into a carboxylic acid and an amine. The figure illustrates generation of a database query and examination of any hits found against the original conformation for just the amine component.

a carboxylic acid, as in Figure 3. Each of the "products" from this retrosynthesis step then is considered in turn.

The first step is to determine the minimum path between the reacting group generated by the retrosynthetic transformation and the interacting pharmacophoric group, deleting any nonessential pendant groups along this path. This reduced representation of the reaction product then can be mapped back onto the original monomer fragment found in the parent compound, calculating coordinate and dihedral information along the minimum path. This minimal fragment forms the basis of a 2D substructure search of an appropriate database, using the Daylight SMARTS query language.[49] A set of generalizing rules can be used to produce the search query as the path is retraced from the pharmacophoric group back to the reacting group. For example, aromatic carbons of a phenyl ring can be designated to match any aromatic atom. Similarly, aliphatic carbons can be permitted to match any saturated atom type. In general, it is the geometry of the atom that matters rather than its chemical nature. This is consistent with the use within COBRA of generalized 3D templates for those portions of the molecule for which the specific template is not present in the template library.[50] For example, a double bond can be considered the geometrical equivalent of an amide.[51] The database search will provide possible reagents that match the SMARTS query and so will contain the key features of the generic structure. It also is possible to apply additional constraints on the chemical functionality that the reagent may (or may not) include as part of the reaction definition, together with limits on the size or flexibility of the molecules using similar procedures to those employed within our ADEPT system for compound selection, library enumeration and profiling.[52]

Each of the hits from the database search then must be considered in 3D, in the context of the parent conformation and the receptor site. Each atom in the potential reagent molecule that matches an atom in the SMARTS query is positioned as in the initial skeleton(s) and an initial check performed on the conformation of the atoms in the minimum path. In particular, matches with minimum paths that involve ring compounds are rejected unless the torsion angles fall within some prescribed tolerance (e.g., 10°) of the parent structure. There may be other atoms in the reagent (due, for example, to substitution along the chain). This is shown for the example in Figure 3 where the database hit contains pendant CH_2OH, OH, and nitro groups not present in the original query. The conformational space of these extra parts of the molecule is now explored using the fragment-building approach embodied in COBRA to determine whether one or more low-energy conformations can be found that are compatible with the 3D structure of the parent structure and with the binding site. Herein lies the crux of the reagent selection algorithm, for although it might not be possible to identify the one single reagent that corresponds exactly to the generic all-carbon minimum-path chain (indeed, this reagent might not be synthetically tractable or could even be unstable), potential reagents that contain the key minimum path and are available in the database *are* identified. As these potential reagents may have functionality additional to the minimum path, we need to check that they are compatible (in a 3D sense) with the rest of the ligand and with the binding site.

For each skeleton, this second phase thus generates lists of potential reagents, which when combined synthetically should provide molecules that can adopt reasonable low-energy conformations consistent with the skeleton and the constraints imposed by the surrounding binding site. To illustrate this second phase, we consider a published series of 4-amino and 4-guanidino-4H-pyran-2-carboxylic acid 6-carboxamide inhibitors of the enzyme neuraminidase (Figure 4) [53]. We considered the "hit" molecule to be the R(cis) = $-H$; R(trans) = $-(CH_2)_2Ph$ compound, that the phenyl ring was one of the key pharmacophore groups, and that the amide bond constitutes the library template. The retrosynthetic reaction transform breaks this molecule into a carboxylic acid and an amine. Here we just concentrate on the amine-containing portion. The generic search query produced comprised a primary or secondary amine connected by two saturated sp^3 atoms to a six-membered aromatic ring. An initial test of the procedure was performed by constructing a small database containing the inhibitors described in Smith et al.[53] to confirm that the algorithm was able to identify the compounds actually synthesized during the project. The conformational analysis was performed based on the binding mode of the parent compound in the active site.[54] One of the more potent of the derivatives described has R(cis) = $-(CH_2)_2CH_3$; R(trans) = $-(CH_2)_2Ph$. The predicted COBRA conformations for this molecule are shown in Color Plate 2 superimposed on the experimental X-ray structure. Color Plate 3 shows a predicted structure for the biphenyl compound. Note in both these cases that the derived compounds have additional functionality beyond that contained within the minimum path (i.e., both contain a propyl group on the amine nitrogen and the second compound has an extra phenyl ring). To illustrate the database searching capabilities, the Available Chemicals Database[55] (99.1 version) was searched, with the additional constraints that the potential reagent should contain just one amine function and no carboxylic acids, and that the molecular weight of the product (when combined with the 4-amino-4H-pyran-2-carboxylic acid 6-carboxamide) should not exceed 600. This gave just over 850 monomers, of which 109 were found to have at least one acceptable conformation. The subsequent steps for a library would involve identifying those combinations of monomers that should give sensibly docked product molecules, and then dealing with the combinatorial subset selection problem.[56]

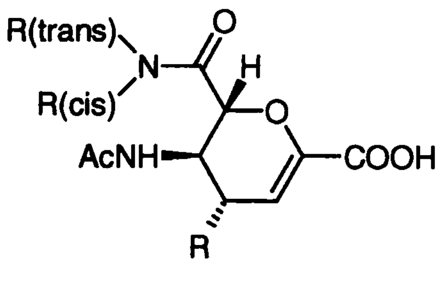

Figure 4. Generic structure of the of 4-amino and 4-guanidino-4H-pyran-2-carboxlic acid 6-carboxamide inhibitors series of neuraminidase inhibitors.

DISCUSSION

In previous articles, we described a stepwise approach to structure-based *de novo* design using a procedure that first involves the identification of favorable binding groups, the linking together of such groups using acyclic chains, and then the rigidification of the chains using ring templates. The resulting skeletons then would need to be examined visually to assess their suitability for synthesis. The approach we consider here provides a mechanism to address this important "synthesizability" problem by requiring each skeleton to contain a key core template, the synthesis of which has been well established through combinatorial chemistry methods. Potential starting materials consistent with the skeleton then are identified through a combination of database searching and conformational analysis.

It should be noted that the two stages can be used separately; any skeletons generated after the first phase can still be considered synthetic targets in their own right. If there were a significant number of core templates to consider, then it might be more appropriate to use a set of generic structures for these as well (e.g., benzene could act as a surrogate for pyridine, pyrimidine etc.). It then would be necessary to define several retrosynthetic reaction transforms that would convert the generic core template to the precursors for each chemical scheme in turn. The approach presented here is more effective than our previous ring-bracing procedure, because the molecule is preformed to contain the ring in those cases where the combinatorial template itself contains a ring or a ring system.

A major practical limitation to structure-based library design is that, at some stage, one needs to consider the 3D conformation of the ligand within the binding site. In general, 3D methods are significantly slower than those that solely use the 2D structure of the molecule. If each potential product molecule from a virtual combinatorial library has to be docked independently into the active site, then this may require a prohibitive amount of time. One of the ways in which this combinatorial problem has been tackled is to assume that the docking of a product molecule can be constructed from the independent dockings of the monomers, or can be based on a docking of the central core alone. These methods certainly can be appropriate in certain circumstances, but would not be expected to be universally applicable. Here we adopt a somewhat different strategy, in which a single generic structure is used as a surrogate for many potential product molecules during the time-consuming site-based conformational analysis, thus saving computational effort. The number of such generic structures usually is relatively limited (certainly compared to the number of combinations of possible reagents in most cases), making it possible to explore the conformational space relatively quickly even for a number of possible library templates and for several functional groups distributed within the binding site. We then search for reagents that could adopt a conformation consistent with those generic structures that were identified by the procedure.

It is interesting to consider how the generic 3D structures act as a form of filter. The use of filters is widely recognized as an effective way to reduce the enormous size of the chemical search space when designing a library or when selecting compounds for purchase or for a focused screening campaign.[52,57] It is typical to use filters based on the 2D structure of the molecules prior to filters that consider 3D properties due to the more time-consuming nature of work in 3D (especially when one has to take a protein binding site into account). However, it could be considered that, in our approach, we use two 3D-based filters prior to a 2D filter. The first 3D filter involves the use of the protein structure to screen for tight-binding molecular fragments. The subsequent "outside-in" strategy is based on the assumption that only reagents containing at least one of the key pharmacophores represented in such fragments will be considered in subsequent steps. The second 3D filter arises in the form of the docked generic skeletons. For a given combination of pharmacophoric groups in spatially different parts of the binding site, the fragment connection algorithm aims to systematically and exhaustively explore both configurational and conformational space, subject to the need to incorporate the combinatorial core template. The resulting family of 3D structures should represent all possible (generic) molecules that are compatible with the bound pharmacophores, the core template, and the protein binding site. By considering generic structures that represent "whole" molecules (in the sense that all of the basic parts of the final molecules are present), one should avoid some of the problems associated with the methods that dock independent monomers or which are based on the orientation of the core template. The subsequent 2D search then aims to identify only those available reagents that are compatible with the synthetic scheme and which would be expected to conform to the conformational requirements, although this needs to be checked due to the presence of groups additional to the basic skeleton.

In some cases, there may be relatively few acceptable and available monomers for a particular synthetic scheme so making the virtual library relatively small. It may be feasible then simply to dock the entire library to identify those virtual products that can dock to the protein. However, in such cases, the fact that the virtual library is of a limited size could be taken as an indication that some form of *de novo* design might be an appropriate strategy to expand the pool of possible molecules, even if it requires the synthesis of additional monomers.

ACKNOWLEDGMENTS

R.A.B. and A.H. both thank the EPSRC and Glaxo Wellcome for studentships. We would like to thank Dr. M.M. Hann for his comments on the manuscript.

REFERENCES

1 Hann, M.M., and Green, R.H. Chemoinformatics—A new name for an old problem? *Curr. Opin. Chem. Biol.* 1999, **3,** 379–383

2 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* 1982, **161,** 269–288

3 Ajay, and Murcko, M.A. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* 1995, **38,** 4953–4967

4 Jones, G., Willett, P., Glen, R.C., Leach, A.R., and Taylor, R. Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. *ACS Symp. Ser.* 1999, **719,** 271–291

5 Leach, A.R. Structure-based selection of building blocks for array synthesis via the World-Wide Web. *J. Mol. Graphics* 1997, **15,** 158–160

6 Sun, Y., Ewing, T.J.A., Skillman, A.G., and Kuntz, I.D. CombiDOCK: Structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Design* 1998, **12,** 597–604

7 Makino, S., Ewing, T.J.A., and Kuntz, I.D. DREAM++: Flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Design* 1999, **13,** 513–532

8 Murray, C.W., Clark, D.E., Auton, T.R., Firth, M.A., Li, J., Sykes, R.A., Waszkowycz, B., Westhead, D.R., and Young, S.C. PRO−SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J. Comput.-Aided Mol. Design* 1997, **11,** 193–207

9 Murcko, M.A. Recent advances in ligand design methods. In: *Reviews in computational chemistry, Volume 11.* Wiley-VCH, New York, 1997, pp. 1–66

10 Clark, D.E., Murray, C.W., and Li, J. Current issues in *de novo* molecular design. In: *Reviews in computational chemistry, Volume 11.* Wiley-VCH, New York, 1997, pp. 67–126

11 Gillet, V.J., and Johnson, A.P. Structure generation for *de novo* design. *Des. Bioact. Mol.* 1998, 149–174

12 Lewis, R.A., and Leach, A.R. Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Design* 1994, **8,** 467–475

13 Goodford, P.J. A Computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med. Chem.* 1985, **28,** 849–857

14 Miranker, A., and Karplus, M. Functionality maps of binding sites—A multiple copy simultaneous search method. *Prot. Struct. Funct. Genet.* 1991, **11,** 29–34

15 Böhm, H.J. LUDI—Rule-based automatic design of new substituents for enzyme-inhibitor leads. *J. Comput.-Aided Mol. Design* 1992, **6,** 593–606

16 Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A., and Ringe, D. An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* 1996, **100,** 2605–2611

17 Shuker, S.B., Hajduk, P.J., Meadows, R.P., and Fesik, S.W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 1996, **274,** 1531–1534

18 Lauri, G., and Bartlett, P.A. CAVEAT—A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Design* 1994, **8,** 51–66

19 Eisen, M.B., Wiley, D.C., Karplus, M., and Hubbard, R.E. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Prot. Struct. Funct. Genet.* 1994, **19,** 199–221

20 Lewis, R.A., and Dean, P.M. Automated site-directed drug design: The concept of spacer skeletons for primary structure generation. *Proc. R. Soc. Lond. B* 1989, **236,** 125–140

21 Lewis, R.A., and Dean, P.M. Automated site-directed drug design: The formation of molecular templates in primary structure generation. *Proc. R. Soc. Lond. B* 1989, **236,** 141–162

22 Lewis, R.A. Automated site-directed drug design: Approaches to the formation of 3D molecular graphs. *J. Comput.-Aided Mol. Design* 1990, **4,** 205–210

23 Lewis, R.A. Automated site-directed drug design: A method for the generation of general three-dimensional molecular graphs. *J. Mol. Graphics* 1992, **10,** 131–143

24 Leach, A.R., and Kilvington, S.R. Automated molecular design: A new fragment-joining algorithm. *J. Comput.-Aided Mol. Design* 1994, **8,** 283–298

25 Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R., Kuntz, and I.D. Automated site-directed drug design using molecular lattices. *J. Mol. Graphics* 1992, **10,** 66–78

26 Leach, A.R., and Lewis, R.A. A ring-bracing approach to computer-assisted ligand design. *J. Comput. Chem.* 1994, **15,** 233–240

27 Todorov, N.P., and Dean, P.M. Evaluation of a method for controlling molecular scaffold diversity in *de novo* ligand design. *J. Comput.-Aided Mol. Design* 1997, **11,** 175–192

28 Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z., Johnson, and A.P. SPROUT: Recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 207–217

29 Chan, S.L., Chau, P.L., and Goodman, J.M. Ligand atom partial charges assignment for complementary electrostatic potentials. *J. Comput.-Aided Mol. Design* 1992, **6,** 461–474

30 Barakat, M.T., and Dean, P.M. The atom assignment problem in automated *de novo* drug design. 1. Transferability of molecular fragment properties. *J. Comput.-Aided Mol. Design* 1995, **9,** 341–350

31 Barakat, M.T., and Dean, P.M. The atom assignment problem in automated *de novo* drug design. 2. A method for molecular graph and fragment perception. *J. Comput.-Aided Mol. Design* 1995, **9,** 359–372

32 Barakat, M.T., Dean, P.M. The atom assignment problem in automated *de novo* drug design. 3. Algorithms for optimization of fragment placement onto 3D molecular graphs. *J. Comput.-Aided Mol. Design* 1995, **9,** 341–350

33 Barakat, M.T., and Dean, P.M. The atom assignment problem in automated *de novo* drug design. 4. Tests for site-directed fragment placement based on molecular complementary. *J. Comput.-Aided Mol. Design* 1995, **9,** 448–456

34 Barakat, M.T., and Dean, P.M. The atom assignment problem in automated *de novo* drug design. 5. Tests for envelope-directed fragment placement based on molecular similarity. *J. Comput.-Aided Mol. Design* 1995, **9,** 457–462

35 Todorov, N.P., and Dean, P.M. A branch-and-bound method for optimal atom-type assignment in *de novo* ligand design. *J. Comput.-Aided Mol. Design* 1998, **12,** 335–349

36 Meng, E.C., Shoichet, B.K., and Kuntz, I.D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* 1992, **13,** 505–524

37 Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M., and McCammon, J.A. A molecular mechanics/grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.* 1995, **16,** 454–464

38 Bilodeau, M.T., and Cunningham, A.M. Solid-supported synthesis of imidazoles: A strategy for direct resin-attachment to the imidazole core. *J. Org. Chem.* 1998, **63,** 2800–2801

39 Bunin, B.A., Plunkett, M.J., and Ellman, J.A. The combinatorial synthesis and chemical and biological evaluation of a 1,4-benzodiazepine library. *Proc. Natl. Acad. Sci. U. S. A.* 1994, **91,** 4708–4712

40 Leach, A.R., and Prout, K. Automated conformational analysis: Directed conformational search using the A* algorithm. *J. Comput. Chem.* 1990, **11,** 1193–1205

41 Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 190–196

42 Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., and Levinthal, C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* 1987, **26,** 2053–2085

43 Leach, A.R., Prout, K., and Dolata, D.P. An investigation into the construction of molecular models by the template joining method. *J. Comput.-Aided Mol. Design* 1988, **2,** 107–123

44 Leach, A.R., Prout, K., and Dolata, D.P. Automated conformational analysis: Algorithms for the efficient construction of low-energy conformations. *J. Comput.-Aided Mol. Design* 1990, **4,** 271–282

45 Bostrom, J., Norrby, P.-O., and Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Design* 1998, **12,** 383–396

46 Leach, A.R. Unpublished results

47 Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., and Kraut, J. Crystal structures of Escherichia Coli and Lactobacillus Casei Dihydrofolate Reductase refined at 1.7Angstroms resolution. 1. General features and binding of methotrexate. *J. Biol. Chem.* 1982, **257,** 13650–13662

48 Daylight theory manual chapter 7. Daylight Chemical Information Systems, Santa Fe, and *http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html*

49 Daylight theory manual chapter 4. Daylight Chemical Information Systems, Santa Fe, and *http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html*

50 Leach, A.R., Dolata, D.P., and Prout, K. Automated conformational analysis and structure generation: Algorithms for molecular perception. *J. Chem. Inf. Comput. Sci.* 1990, **30,** 316–324

51 For example, through the use of the following SMARTS: [$(C = C),$(C(= O)N),$(NC = O)]-,= [$(C = C),$(NC = O),$(C(= O)N)]

52 Leach, A.R., Bradshaw, J., Green, D.V.S., Hann, M.M., and Delany, J.J. III. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 1161–1172

53 Smith, P.W., Sollis, S.L., Howes, P.D., Cherry, P.C., Cobley, K.N., Taylor, H., Whittington, A.R., Scicinski, J., Bethell, R.C., Taylor, N., Skarzynski, T., Cleasby, A., Singh, O., Wonacott, A., Varghese, J., and Colman, P. Novel inhibitors of influenza sialidases related to GG167. Structure-activity, crystallographic and molecular dynamics studies with 4H-pyran-2-carboxylic acid 6-carboxamides. *Bioorg. Med. Chem. Lett.* 1996, **6,** 2931–2936

54 Taylor, N.R., Cleasby, A., Singh, O., Skarzynski, T., Wonacott, A.J., Smith,P.W., Sollis, S.L., Howes, P.D., Cherry, P.C., Bethell, R., Colman, P., and Varghese, J. Dihy-

dropyrancarboxamides related to Zanamivir: anew series of inhibitors of influenza virus sialidases. 2. Crystallographic and modeling study of complexes of 4-amino-4h-pyran-6-carboxamides and sialidase from 6 influenza virus types A and B J. *Med. Chem.* 1998, **41,** 798–807

55 The Available Chemicals Database is from MDL Information Systems, Inc., San Leandro, CA

56 Gillet, V.J., Willett, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 731–740

57 Walters, W.P., Ajay, and Murcko, M.A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 1999, **3,** 384–387