Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design[†]

Andrew R. Leach,* John Bradshaw, Darren V. S. Green, and Michael M. Hann

Medicines Research Centre, Glaxo Wellcome Research & Development, Gunnels Wood Road,
Stevenage, Hertfordshire SG1 2NY, U.K.

John J. Delany III

Daylight Chemical Information Systems Inc., Santa Fe Research Office, 419 East Palace Avenue,
Santa Fe, New Mexico 87501

Received April 21, 1999

We describe an integrated suite of computational tools which are used to assist in the selection of compounds for biological assays and the design of combinatorial libraries. These functions are delivered in a platform-independent manner via a corporate intranet and are used by computational experts and nonexperts alike. While the system was primarily designed to be used prior to synthesis, it can also be used to provide structural information for library registration and for decoding beads in tagged libraries. We describe a simple statistical method for monomer selection and compare it to computationally more demanding approaches.

INTRODUCTION

Much of the published research concerning the theoretical and computational aspects of high-throughput screening and combinatorial chemical libraries has concentrated on the problem of designing “diverse” libraries or selecting diverse sets of compounds. However, it is now generally recognized that diversity alone is rarely sufficient; 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 compound selection or library design phase. The key is to achieve an appropriate degree of chemical diversity while simultaneously incorporating this target knowledge. The relationship is an approximately inverse one; the more knowledge one has, the more focused (i.e., less chemically diverse) the selection should be. For example, if we only knew the primary amino acid sequence of a biological target, then we would probably aim for a rather greater amount of diversity (however, that is measured) than would be the case if we had available a series of protein−ligand X-ray structures for the target.

In addition to the incorporation of target knowledge, it is also desirable to try to ensure that the molecules screened have “sensible”, “druglike” properties. The expectation is not only that this will increase the chances of achieving biological activity but that any hits from the assay will represent more attractive starting points for lead optimization. A number of groups have developed methods which are designed to distinguish sets of druglike molecules from sets of nondruglike compounds. In such approaches a number of properties (descriptors) are calculated for each compound; these act as the input to a computational model which predicts whether the molecule is druglike or not. The model may be extremely simple (as in the case of the so-called “rule of 5” [1]) or may be a rather detailed neural network[2] or genetic algorithm-derived equation.[3] The important feature of these methods is that they involve the calculation of relevant properties from the structure, which are then used to assess the quality of that particular molecule. However, it should always be remembered that such approaches are only very general in nature and that any specific target may require molecules that violate one or more of these more general criteria.

The methods described in this paper have evolved to meet the needs of a growing number of scientists at Glaxo Wellcome who wish to select compounds for assays and to design combinatorial libraries. The general strategy that we adopt is to provide a variety of filters that can be applied to reagents or compounds, to eliminate unwanted molecules from further consideration. The system (called ADEPT, A Daylight Enumeration and Profiling Tool) delivers these capabilities over the World Wide Web in a platform-independent manner via a company-wide intranet. Some of the elements in ADEPT are also present in other systems which have been described in the literature or reported at conferences. These include the cyclops system at Novartis[4] together with systems at Vertex,[5] Abbott,[6] and UCSF.[7]

While there are many different ways to access the methods present in ADEPT, the flowchart shown in Figure 1 identifies the key stages where it can be employed. Nonexpert users are recommended to follow this flowchart; the workflow is facilitated by the judicious use of hyperlinks. The first step involves the identification of potential compounds or reagents via some form of database search (e.g., a substructure or similarity search). This initial list is then pruned using various criteria to give an intermediate set. When a library is designed, the virtual library would be fully enumerated at

[†] Some of this work was presented at the 1998 European Daylight Users meeting in Cambridge; a semilive version of the software is available at http://www.daylight.com/meetings/emug98/Leach/index.html.

* To whom correspondence should be addressed. E-mail: arl22958@GlaxoWellcome.co.uk.

10.1021/ci9904259 CCC: $18.00
Published on Web 10/29/1999

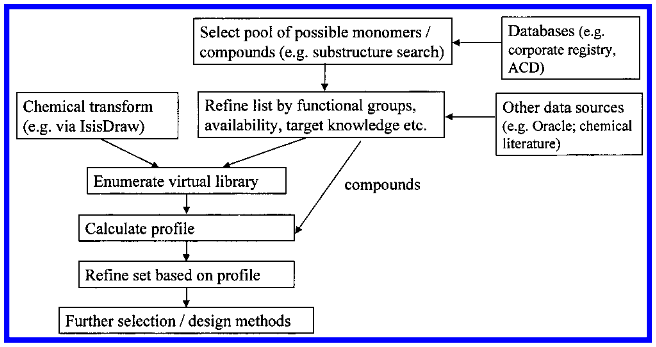

Select pool of possible monomers / compounds (e.g. substructure search)
Databases (e.g. corporate registry, ACD)
Chemical transform (e.g. via IsisDraw)
Refine list by functional groups, availability, target knowledge etc.
Other data sources (e.g. Oracle; chemical literature)
Enumerate virtual library
compounds
Calculate profile
Refine set based on profile
Further selection / design methods

Figure 1. Process flowchart recommended for use with ADEPT.

this point. Various properties are now calculated for the structures in this virtual library, and then the list is pruned using these properties to remove any molecules which are not acceptable. The subsequent steps could involve the selection of a subset of compounds using cluster analysis or a cell-based method, docking to a protein structure,[8] or the combinatorial selection of a diverse set of reagents for a library while matching to a desired profile.[9] We now describe each of the key functionalities in ADEPT with reference to their role in a strategy for compound selection and library design.

IDENTIFICATION OF THE INITIAL COMPOUND/ REAGENT POOL

The first task is to select the initial pool of compounds or reagents which will form the basis for the subsequent steps. This is most commonly achieved by a substructure search. For reagent selection the substructure query will usually be rather simple, comprising just the key reactive functionality (e.g., carboxylic acid, primary amine, α-haloketone, etc.). For compound selection the query may be more complex, often involving a "disconnected" substructure consisting of a number of features. For example, one might define an antihistamine substructure containing two aromatic rings and a tertiary amine. Within ADEPT these initial sets are usually identified by searching one or more Daylight databases using an appropriate SMARTS[10] query. The SMARTS expressions for commonly used functional groups are predefined within the system; in other cases the user can generate a SMARTS expression via a chemical drawing package such as ISISDraw as described below, or if sufficiently confident can define their own.

It may be desired to restrict the search to compounds from a set of preferred suppliers in the case of the available chemical database (ACD[11]). Supplier information is present within the Daylight database and so can be considered during the substructure search. Another very useful filter to apply at this stage for in-house molecules is an availability filter to ensure that sufficient compound is available. Availability information is stored within an Oracle database which is checked once the initial substructure search has been completed.

This initial pool of potential reagents (which in the case of a common functional group may contain a very large number of molecules) is then further refined according to additional chemical requirements. The objective here is to eliminate molecules which, by virtue of their chemical composition, can be deemed inappropriate. For example, a reductive amination reaction can be problematic should the aldehyde component also contain any ketones, alkyl bromides, or unprotected carboxylic acids. In other cases a specific functional group may be required, but not multiple occurrences (due to the possibility of multiple products being generated). The user is presented with a wide selection of the most commonly encountered chemical functional groups (Figure 2). For each functional group one of five different possibilities can be specified: (a) the group must not be present, (b) there must be exactly one of the functional group; (c) there must be at least one of the group, (d) there must be more than one of the group, or (e) it is not relevant to impose any restrictions on the group (this is the default).

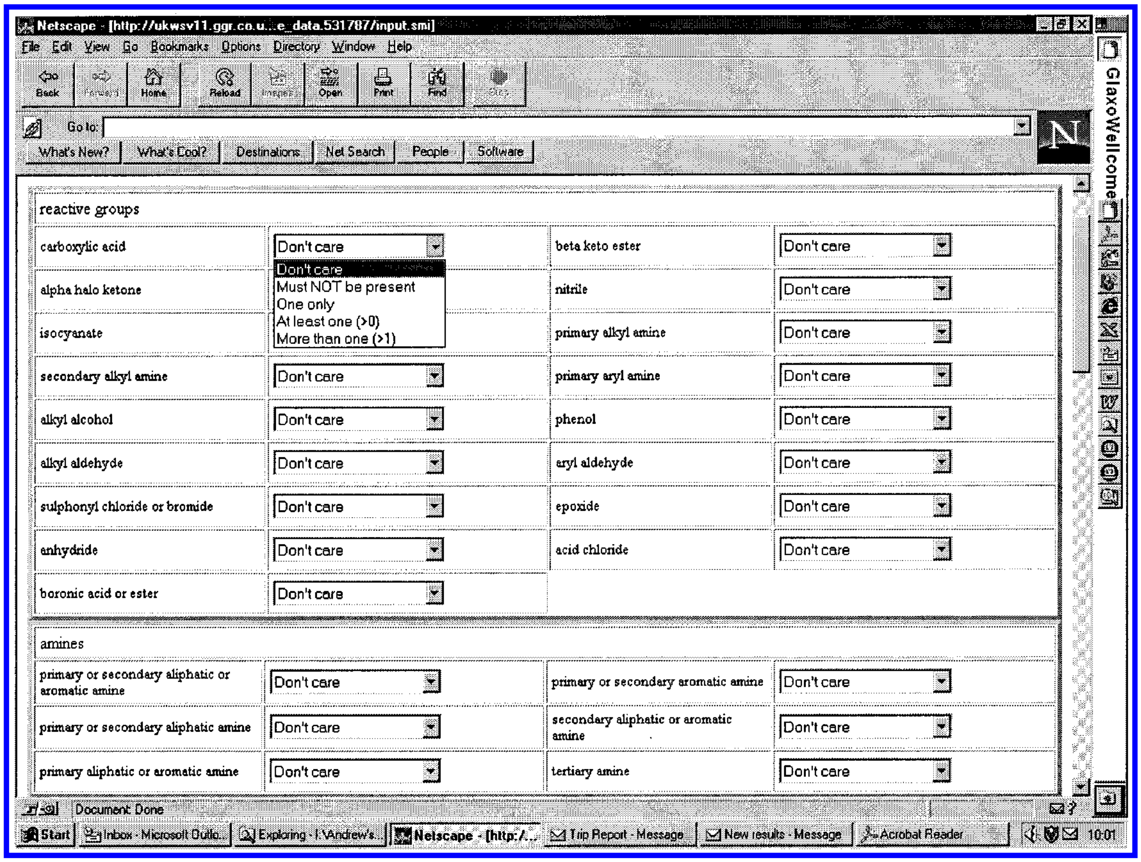
In compound selection a somewhat different scenario often arises, in which it is desired to remove from the set those compounds which contain one or more groups that often interfere with biological assays. Compounds containing groups such as acid halides, peroxides, or benzyl chlorides will frequently give rise to "false positives" in an assay. In addition there are a number of undesirable substructures such as polyfluorinated alkanes or β-lactams which can be used to identify compounds that would not generally be considered good leads. A list of such undesirable groups has been encoded as a series of SMARTS expressions which can be selected *en bloc*; any compound which matches one or more of these patterns will be eliminated from the set.

Finally, it is possible at this stage to apply two simple property filters, one based on the molecular weight and the other based on a count of the number of rotatable bonds within the molecule (we define a rotatable bond as any acyclic, unconjugated single bond which connects two atoms, each of which is in turn bonded to at least one other non-hydrogen atom). Notwithstanding our emphasis on product-based monomer selection (see below), it can often be appropriate to eliminate from consideration any potential monomers which, on account of their molecular weight or flexibility, would never be considered suitable for inclusion in a library as a monomer.

FRAGMENT-MARKING AND TRANSFORM-BASED APPROACHES TO LIBRARY ENUMERATION

A key component of ADEPT, at least as far as its use in the design of combinatorial arrays and arrays is concerned, is library enumeration. The term *enumeration* refers to the procedure by which the connection tables for the product structures in a real or virtual library are produced. It should be noted that a single compound can be considered as a library of one, and so enumeration can equally well be applied to this problem. However, whereas it might be considered appropriate for a chemist to manually draw the structure of a single compound (which may have taken days if not months or years to synthesize), it is clearly not practical to do so even for small combinatorial libraries, hence the need for automated tools to perform this procedure.

A number of software packages have been developed for tackling the problem of enumerating a combinatorial library, some of which are now commercially available. These generally fall into one of two distinct categories. The first of these is often referred to as "fragment-marking". In this method a central core template, common to all product structures, is identified. The template will contain one or more points of variation where different substituents (often


Netscape - [http://ukwsv11.ggr.co.u...e_data.531787/input.smi]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Go to:
What's New? What's Cool? Destinations Net Search People Software
N
GlaxoWellcome
Document: Done
Start Inbox - Microsoft Outlo... Exploring - I:\Andrew's... Netscape - [http:/... Trip Report - Message New results - Message Acrobat Reader 10:01

reactive groups

carboxylic acid	Don't care Don't care Must NOT be present One only At least one (>0) More than one (>1)	beta keto ester	Don't care
alpha halo ketone		nitrile	Don't care
isocyanate		primary alkyl amine	Don't care
secondary alkyl amine	Don't care	primary aryl amine	Don't care
alkyl alcohol	Don't care	phenol	Don't care
alkyl aldehyde	Don't care	aryl aldehyde	Don't care
sulphonyl chloride or bromide	Don't care	epoxide	Don't care
anhydride	Don't care	acid chloride	Don't care
boronic acid or ester	Don't care		

amines

primary or secondary aliphatic or aromatic amine	Don't care	primary or secondary aromatic amine	Don't care
primary or secondary aliphatic amine	Don't care	secondary aliphatic or aromatic amine	Don't care
primary aliphatic or aromatic amine	Don't care	tertiary amine	Don't care

Figure 2. Interface used to refine the initial list of reagents showing the five different options for each functional group.

termed R groups) can be placed. By varying the R groups at the points of substitution, different product structures can be generated. To enumerate a combinatorial library, it is first necessary to construct sets of R group substituents from the relevant monomer sets. In the simplest cases this is done by replacing the reactive functional group in the monomer with a “free valence”. By creating a bond between the template and the required R groups, the connection table for the product molecule can be generated. Enumeration of the full library corresponds to systematically generating all possible combinations of R group substituents at the different points of variation.

The alternative approach is to use the computational equivalent of a chemical reaction, or *reaction transform*. Here, one does not need to define a common template nor to generate sets of “clipped” reagents. Rather, the library can be enumerated using as input the initial reagent structures and the chemical transforms required to operate upon them. In this way it more closely replicates the stages involved in the actual synthesis, wherein reagents react together according to the rules of synthetic chemistry (at least, when the chemistry works as planned!).

The key elements of the fragment-marking and transform approach can be illustrated using as an example the synthesis of aminothiazoles from thioureas and α-haloketones[12] (Figure 3). With the reaction transform approach (Figure 3a), one would simply define an appropriate transform. The enumeration engine then applies this transform to the initial starting materials (i.e., the thiourea and the α-haloketone) to produce the aminothiazole (with water and the hydrogen halide as byproducts). Using the fragment-marking approach (Figure 3b), one would construct three sets of clipped fragments (two from each α-haloketone and one from each thiourea) which would then be grafted on to give the central thiazole core to give the appropriate products.

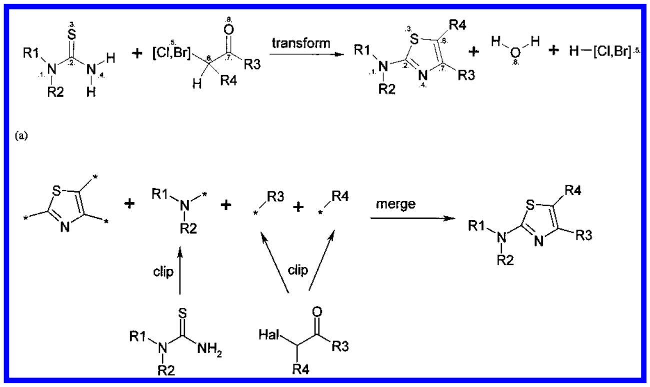


Figure 3. Schematic comparison of the transform (a) and fragment-marking (b) approaches to the enumeration of a 2-aminothiazole library. The reaction transform approach takes as input the two sets of reagents to which the transform is applied. The fragment-marking approach requires the central core template to be defined together with appropriate sets of R groups (three in this case).

Both the marking-up and reaction transform approaches have advantages and disadvantages. In favor of the marking-up approach is the fact that for some kinds of libraries (i.e., those that most obviously fit the “core plus R group” definition) it can be the fastest way to enumerate the library. This is because the fragment-marking approach only involves some rather elementary connection table operations once the R groups have been generated. Although most systems offer automated ways to generate the R groups (i.e., “clipping algorithms”), problems almost invariably arise which need to be corrected by hand. This can make the fragment-marking approach time-consuming to perform for a nonexpert unless sets of predefined R groups are already available. In addition there are certain reactions which are not properly handled by the fragment-marking approach, one well-known example being the Diels−Alder reaction where a simple fragment-marking approach would generate a number of extraneous and incorrect products.[13] Moreover, in some cases there is no clear core structure (e.g., oligomeric libraries such as peptoids). The advantages of the reaction method include the ability to enumerate directly from the reagents without having to perform any preprocessing and the ability to reuse the same transforms many times (once they have been defined). However, this method requires more computational steps and so is typically slower. Perhaps the key advantage however is that this approach models the actual chemical steps involved in the experiment, thus bringing the experimental and computational systems closer together.

IMPLEMENTATION OF LIBRARY ENUMERATION WITHIN ADEPT

Within ADEPT we use the reaction transform approach as implemented within the Daylight reaction toolkit. To enumerate a library, the user identifies the sets of reagents that are to be used in the library and chooses the chemistry that will be performed upon them. In the Daylight system reaction transforms are specified using the SMIRKS language.[14] SMIRKS can be considered an extension of the SMILES and SMARTS molecular notations. For example, the SMIRKS for the aminothiazole reaction could be written as

[N:1][C:2](=[S:3])[NH2:4]([H:99])[H:100].[Cl,Br:5]
[C:6]([H:101])[C:7]=[O:8]≫[s:3]1[c:2]([N:1])[n:4][c:7]
[c:6]1.[H:99][O:8][H:100].[Cl,Br:5][H:101]

As with SMARTS, a pair of square brackets identifies a single atom. Thus, for example, [NH2] indicates a nitrogen atom with exactly two hydrogen atoms. Generalized atom types, such as “chlorine or bromine”, can be defined using the appropriate SMARTS (i.e., [Cl,Br]). The “≫” symbol is used to separate the reagents from the products. Hydrogen atoms that are actually involved in the reaction must be specified explicitly. The integers following the “:” character are the atom maps which indicate how atoms on the reactant side of the equation correspond to the atoms in the products. In the case of the aminothiazole reaction the atom maps for non-hydrogen atoms are also indicated in Figure 3a.

The synthesis schemes used in libraries vary greatly in the number of chemical steps involved and the complexity of those steps. At one end of the spectrum would be a simple one-step bimolecular reaction. At the other extreme could be a multistep solid-phase synthesis involving in addition to the key reactions coupling/decoupling to a solid support, and the selective removal of protecting groups. Other reactions used in combinatorial chemistry involve three or four components.

Within ADEPT all types of libraries, from the simplest two-component reaction to the most complex solid-phase scheme, are enumerated using a single enumeration engine. However, we have developed a variety of Web interfaces to reflect our diverse user communities. Thus, in the simplest case the user just needs to define the two sets of reagents (for example, as a file of registry numbers or via a chemical drawing) and to select the appropriate reaction transform from a predefined list (Figure 4). Intermediate in complexity would be a bimolecular reaction where some chemical transformations (e.g., coupling to a resin and removal of protecting groups) are performed on one or more of the starting materials and on the products (Figure 5). The most generic interface is shown in Figure 6 where one is able to define a scheme involving many components, each of which, along with the intermediate and final products, may be subjected to coupling/protection/deprotection steps. We also use this “flow scheme” for multicomponent reactions such as the Ugi reaction. In these situations no reaction is selected until all of the components have been defined.

SPECIFICATION OF REACTIONS USING THE MTZ LANGUAGE

The first step in the actual enumeration is the specification of the synthesis using an extended version of the “molecules transform zap” (MTZ) language[15] in which the molecules are specified as SMILES strings, and the reaction transforms as SMIRKS. The “zap” facility enables byproducts to be eliminated by specifying an appropriate SMARTS expression which can remove matching components. We then use a reaction toolkit program which interprets the MTZ statements and performs the actual enumeration. A simple illustration of the MTZ language is the following amide-formation reaction between acetic acid and methylamine with elimination of the water molecule:

MOLECULES: CC(=O)O

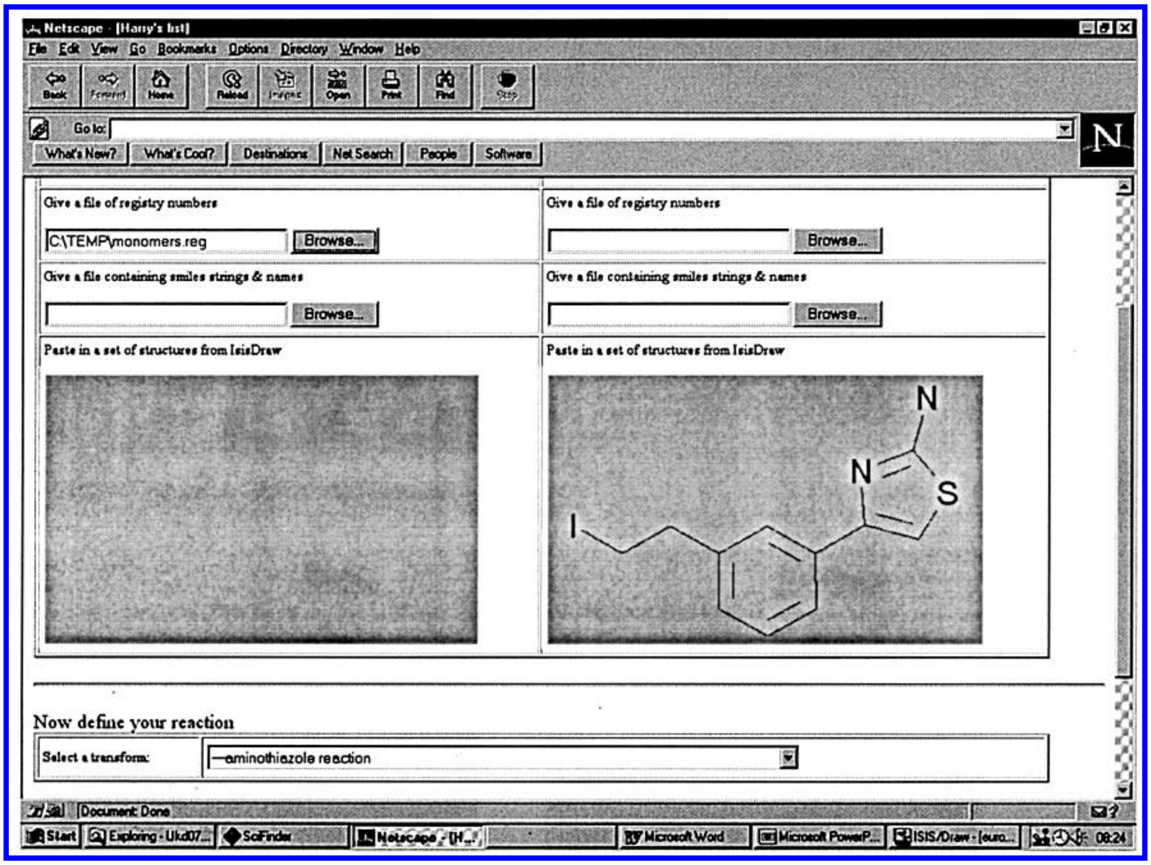
MOLECULES: CN

TRANSFORMATION: [O:0]=[C:1][O:2][#1:99].
[$([NH2][CX4]):3][#1:100]≫[O:0]=[C:1][N:3].
[#1:99][O:2][#1:100]

ZAPPAT: [OH2]

Here we have written the SMIRKS to match any primary amine (but not secondary amines or ammonia). More recent versions of the reaction toolkit enable the “transformation” and “zap” functions to be encapsulated within the same SMIRKS expression (atoms which do not appear on the right-hand side of the SMIRKS are eliminated). For a variety of reasons (including the fact that the first versions of ADEPT were based on earlier releases of the reaction toolkit, which required a properly balanced equation), we currently prefer to separate the actual chemical transformation from the byproduct elimination process.

The definition of a reaction for use within ADEPT typically requires one or more SMIRKS transformations to be specified together with some “ZAPPAT” statements which enable trivial byproducts to be eliminated (e.g., the water in a condensation reaction, the hydrogen halide in a nucleophilic displacement at an alkyl halide, etc.). In some cases we also include within the reaction definition one or more reagent molecules (as SMILES) that must be present for the reaction to proceed but which the user would expect not to have to define as a separate monomer. Two examples of this would be the water molecule required for ester hydrolysis or the hydrogen molecules required in a hydrogenation.


Netscape - [Hany's list]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Go to:
N
What's New? What's Cool? Destinations Net Search People Software
Browse...
N
N
S
I
Start
Exploring - Ukd07...
SciFinder
Netscape - [H...
Microsoft Word
Microsoft PowerP...
ISIS/Draw - [euro...
08:24

Give a file of registry numbers

C:\TEMP\monomers.reg

Give a file containing smiles strings & names

Paste in a set of structures from IsisDraw

Give a file of registry numbers

Give a file containing smiles strings & names

Paste in a set of structures from IsisDraw

Now define your reaction

Select a transform: —aminothiazole reaction

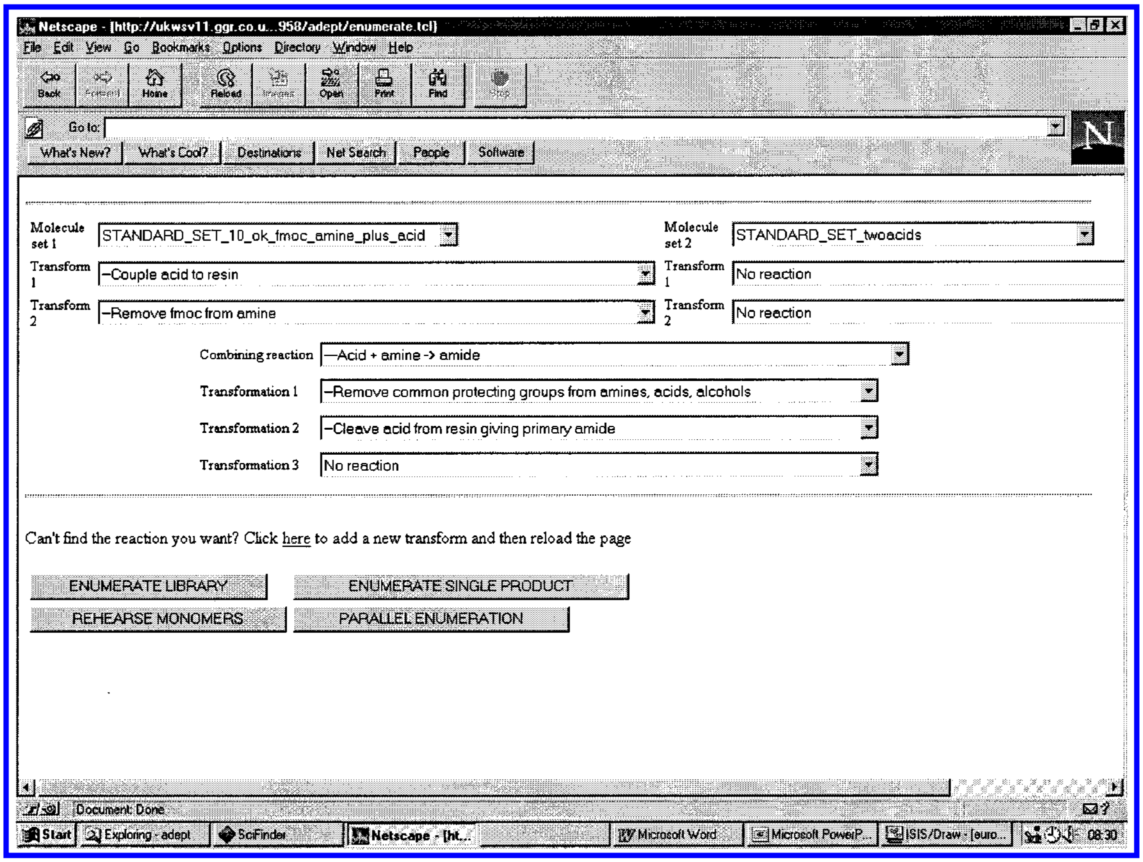
Document: Done

Figure 4. Basic interface used to enumerate the products of a simple bimolecular reaction where one monomer set is specified as a file of registry numbers and the other is input as a chemical drawing.

Inevitably, some reactions do not proceed as the user expected, and so it is necessary to be able to inform the user when problems are detected. The two most common enumeration problems are due to reagents which do not react according to the transform definition, or due to the generation of multiple products. A simple example of the former would be the failure of a secondary amine to couple with a carboxylic acid should the transform definition be defined specifically for the reaction of primary amines. An example of the latter would be the reaction between 2-methyl-1,4-diaminobutane and acetic acid where the two nonequivalent amino groups give rise to two distinct monoamides. Of course in some cases we would want to obtain mixtures of products (e.g., certain types of pericyclic reaction which can give rise to more than one product). Such potential failures are due to the presence of inappropriate reagents in the monomer sets, to the selection of an inappropriate transform, or to the transform definition being incorrect. When this does occur, an appropriate warning message is produced and a series of hyperlinks are provided to enable the user to view the problematic structures.

We usually recommend that each user defines his or her reaction scheme so that it matches the actual stages followed in the synthesis. This means that a number of distinct enumerations may be required in the case of a multicomponent, multistep solid-phase synthesis involving a detailed protection/deprotection strategy. In such cases an alternative approach would be to perform a single multicomponent "reaction" that would have no real-world experimental counterpart but which would enable the computer to enumerate the library in one step. Clearly this second approach has the advantage of speed, but it does suffer from two drawbacks. The first is that a potentially large number of rather specific reaction transforms would have to be defined, whereas the alternative approach enables a relatively small number of transforms (corresponding to the commonly encountered library chemistries) to be defined and reused in many projects. The other advantage to the multistep enumeration approach is that when problems arise, it is much easier to determine the cause if we have used the stepwise approach.

The enumeration of large libraries with many different steps can take a significant amount of computational time, and so it is desirable to ensure (as far as possible) that one has chosen an appropriate set of monomers and the correct reaction transforms. As an illustration, a recent three-component library containing 640 000 enumerated products required approximately 18 h (Silicon Graphics R10K processor) to enumerate (10 structures/s). However, the reaction scheme for this library actually involved six distinct steps for a total of 1.9 million individual computational reactions (30 reactions/s). To assist in this process, we have developed the computational equivalent of the experimental monomer rehearsal. This allows us to evaluate each monomer to check that it will not give rise to any problems without first having to generate the full set of combinatorial products. To achieve this, each of the monomer sets is taken in turn. If a monomer set contains N_i reagents, then we would enumerate the N_i products that correspond to reacting these N_i reagents with one reagent selected from each of the other monomer sets. The computational rehearsal thus involves the enumeration of $N_1 + N_2 + N_3 + ...$ product structures rather than the N_1


Netscape - [http://ukwsv11.ggr.co.u...958/adept/enumerate.tcl]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Go to:
What's New? What's Cool? Destinations Net Search People Software
ENUMERATE LIBRARY
ENUMERATE SINGLE PRODUCT
REHEARSE MONOMERS
PARALLEL ENUMERATION
Document: Done
Start Exploring - adept SciFinder Netscape - [ht... Microsoft Word Microsoft PowerP... ISIS/Draw - [euro... 08:30

Molecule set 1: STANDARD_SET_10_ok_fmoc_amine_plus_acid

Transform 1: --Couple acid to resin

Transform 2: --Remove fmoc from amine

Molecule set 2: STANDARD_SET_twoacids

Transform 1: No reaction

Transform 2: No reaction

Combining reaction: --Acid + amine -> amide

Transformation 1: --Remove common protecting groups from amines, acids, alcohols

Transformation 2: --Cleave acid from resin giving primary amide

Transformation 3: No reaction

Can't find the reaction you want? Click here to add a new transform and then reload the page

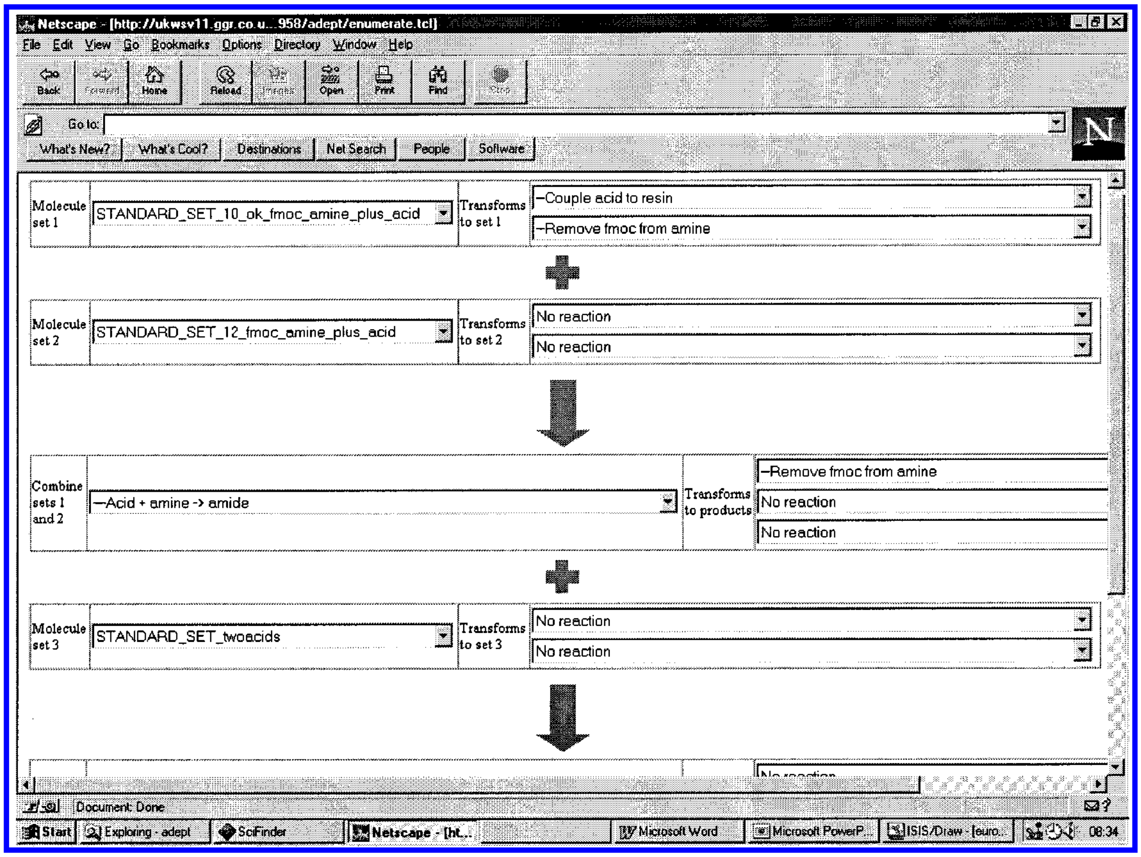
Figure 5. Web interface used to enumerate a two-component library involving first the coupling of a set of fmoc-protected amino acids to a resin followed by removal of the protecting group, reaction with a set of acids, then removal from the resin, and removal of any remaining protecting groups.

$\times$ N_2 $\times$ N_3 $\times$ ... products in the combinatorial library. If this rehearsal is successful (i.e., no errors are produced), then we can be confident that the full combinatorial enumeration should proceed without errors. Another (less commonly used) enumeration strategy is the parallel enumeration. Here we react the first monomer from each of the sets together to give the first product, the second monomers from each of the sets to give the second product, and so on. Each of the monomer sets is thus required to contain the same number of molecules. This parallel enumeration method is rarely used but is important for encoded libraries (see below).

POTENTIAL PROBLEMS WITH LIBRARY ENUMERATION AND MTZ EXTENSIONS

The majority of users can be accommodated using straightforward reaction transforms, where a single SMIRKS expression is used to operate on a set of molecules to generate the products. However, there are two particular situations which have required extensions to this. The first arises for sets of reagents that contain molecules with more than one of a particular type of group, such as a monomer that contains both a primary and a secondary alkylamine. Should such a monomer be subjected to a generic amide coupling transform (i.e., one defined for both primary and secondary amines), then two products would result. This may be counter to the expected experimental scenario whereby the more reactive primary amine would preferentially react to the exclusion of the secondary amine to give just the single monoamide product. However, the set may also contain reagents that have no primary amines, just secondary amines. For these monomers we *do* want the secondary amine to react. We deal with such circumstances by specifying within the reaction definition file not one but multiple reaction transforms that are treated as a single group. The first transform is applied. If products are successfully generated, then we stop. However, should no products be generated with this first transform, then the next transform is applied, and so on until either we get a successful reaction or there are no more transforms left. Thus, in our simple amine example the first transform in the block would be restricted to just primary amines; the second transform would include secondary amines. We identify such a group of transforms using a "CASCADE_TRANSFORMATION" label within our extended MTZ notation.

The second type of situation arises when the sets of monomers are deliberately designed to contain reagents with multiple reactive groups and which the user expects to react multiple times. For example, we may have a set of reagents, each of which contains two amine groups. As indicated above, under normal circumstances each reagent would give rise to two products, corresponding to the reaction of the constitutionally different amine groups. However, if the acid component were present in excess, then the correct product would be the diamide. Within the enumeration program the diamide can be generated by applying the reaction transform a second time to a mixture of the monoamide intermediates and the set of initial reagents. Should we have a triamine, then it would be necessary to repeat the cycle another time to obtain the triamide product. We identify such a situation as an "EXHAUSTIVE_TRANSFORMATION" or as a


Netscape - [http://ukwsv11.ggr.co.u...958/adept/enumerate.tcl]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Go to:
What's New? What's Cool? Destinations Net Search People Software
Start Exploring - adept SciFinder Netscape - [ht... Microsoft Word Microsoft PowerP... ISIS/Draw - [euro... 08:34

Molecule set 1: STANDARD_SET_10_ok_fmoc_amine_plus_acid
Transforms to set 1: --Couple acid to resin; --Remove fmoc from amine

Molecule set 2: STANDARD_SET_12_fmoc_amine_plus_acid
Transforms to set 2: No reaction; No reaction

Combine sets 1 and 2: --Acid + amine -> amide
Transforms to products: --Remove fmoc from amine; No reaction; No reaction

Molecule set 3: STANDARD_SET_twoacids
Transforms to set 3: No reaction; No reaction

No reaction

Document: Done

Figure 6. Generic interface to enumeration engine which can be used to enumerate libraries involving many sets of reagents, multicomponent reactions, and protection/deprotection strategies.

"COMPLETE_TRANSFORMATION" within MTZ; this causes the enumeration engine to repeatedly apply the transformation until no new structures are generated. A limit is imposed on the number of iterations that are possible; otherwise a diamine and a diacid would continue to react indefinitely to give a polymer. If we only require the single "final" product then we would use the COMPLETE_TRANSFORMATION tag; if we also required all of the intermediate structures, than we would use the EXHAUSTIVE_TRANSFORMATION option. In the diamine case this latter option would provide two monoamides plus the diamide; for the triamine example this would give three monoamides, three diamides, and the single triamide.

There are two further tags within our extended MTZ language. First is the ability to associate an identifier with each input molecule via a NAMES field. An identifier is generated for each product structure on the basis of the identifiers of the constituent monomers. The second option is primarily used for the removal of protecting groups. Immediately prior to this final stage there may be one or more of a large number of protecting groups present within a molecule. It would be possible to remove these using the reaction toolkit (via the EXHAUSTIVE_ENUMERATION option), but we find that it is more efficient to specify these as SMARTS expressions (identified as DELETE_GROUP within our MTZ). Removal of most common protecting groups can be considered as replacement of a specific group of atoms with a hydrogen atom (e.g., the hydrolysis of a *tert*-butyl ester to give the free acid corresponds to replacing the *tert*-butyl group with a hydrogen atom). When a DELETE_GROUP tag is noted, any atom within the molecule that is matched by an atom in the SMARTS pattern is deleted from the molecule, with the atom that matches the first atom in the SMARTS pattern being replaced by a hydrogen atom.

INCORPORATING USER-DEFINED REACTIONS

ADEPT contains a few tens of transforms which are available for general use. We find that this is sufficient to cover the chemistries commonly employed in combinatorial or array chemistry. However, it was recognized from the start that there would be a need for users to be able to define their own reactions within the system. This is particularly the case in lead optimization programs where a wider range of chemistry is employed. Similarly, it is desirable for users to be able to define their own substructure queries for database searching. Of course the most obvious way to define new chemical transforms within ADEPT is as a SMIRKS expression, and substructures as SMARTS. The complexity of SMARTS and SMIRKS expressions (apparent even in the relatively straightforward examples above) does however mean that is not an approach that would be considered acceptable by the vast majority of bench scientists! However, most chemists are familiar with the capabilities of tools such as ISISDraw[16] or ChemDraw.[17] We have therefore provided the option to paste structures or reactions from ISISDraw into a Chime[18] window within the Web browser. The drawing is transformed into a mol or rxn file[19] as appropriate on the Web server and is then converted into a SMARTS or SMIRKS expression as appropriate. Some of the features of the way in which this conversion is performed are worthy of mention.[20] First, our program makes no assumptions about

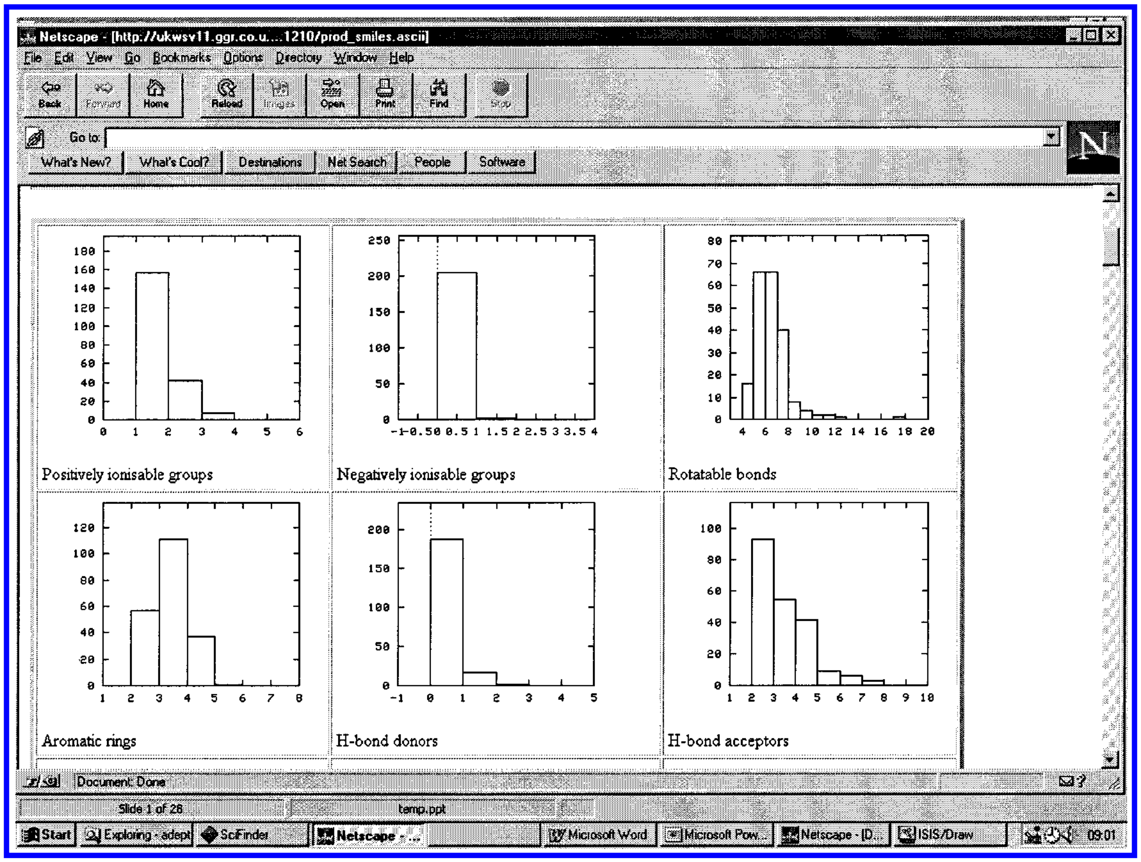

Netscape - [http://ukwsv11.ggr.co.u....1210/prod_smiles.ascii]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Go to:
What's New? What's Cool? Destinations Net Search People Software
180 160 140 120 100 80 60 40 20 0
0 1 2 3 4 5 6
Positively ionisable groups
250 200 150 100 50 0
-1-0.50 0.5 1 1.5 2 2.5 3 3.5 4
Negatively ionisable groups
80 70 60 50 40 30 20 10 0
4 6 8 10 12 14 16 18 20
Rotatable bonds
120 100 80 60 40 20 0
1 2 3 4 5 6 7 8
Aromatic rings
200 150 100 50 0
-1 0 1 2 3 4 5
H-bond donors
100 80 60 40 20 0
1 2 3 4 5 6 7 8 9 10
H-bond acceptors
Document: Done
Slide 1 of 28
temp.ppt
Start Exploring - adept SciFinder Netscape - ... Microsoft Word Microsoft Pow... Netscape - [D... ISIS/Draw 09:01

Figure 7. Typical profile output showing graphical distribution of properties as histogram plots.

the environment or nature of an atom or bond beyond that which can be unambiguously inferred from the input. Thus, the user is required to explicitly include hydrogen atoms where they are needed. This is achieved within the SMARTS/SMIRKS language through the use of "negative logic". For example, if the user draws a carbon atom double bonded to an oxygen and also singly bonded to a second oxygen, then that would be converted to the expression O═CO which would match both carboxylic acids and esters. If the user wishes only to match carboxylic acids, then a hydrogen atom would need to be explicitly bonded to the appropriate oxygen; the resulting SMARTS would be O═C[O&!H0] (i.e. "an oxygen with not zero hydrogens"). Similarly, if a nitrogen atom is drawn with one attached hydrogen, then this would be expressed as [$(N&!H0)], able to match both primary and secondary amines but not a tertiary amine. Another feature of the conversion is that singly connected non-hydrogen atoms are permitted to match either aliphatic or aromatic atoms, and so we simply specify the atomic number of such atoms using the "#" nomenclature of SMARTS (as in [#7] for nitrogen or [#6] for carbon). A limited range of mol file functionality is recognized and converted to the appropriate SMARTS. Key among these is the ability to specify that an atom is able to match a list of atoms (e.g., either chlorine or bromine) or to specify that an atom is not allowed to match a particular set of atoms.

PROFILE CALCULATION AND MONOMER SELECTION

Having enumerated the structures of the virtual library or identified an initial set of compounds which do not contain any undesirable features, the next step in the process (Figure 1) involves the calculation of a variety of properties, which we collectively refer to as the *profile*. The list of properties included in the profile is currently restricted for reasons of speed to ones which can be determined from the 2D connection table. The ADEPT profile should usually be regarded as just the starting point for a more extensive descriptor calculation or modeling exercise. The current list of properties includes counts of hydrogen bond donor and acceptor atoms, positively and negatively ionizable groups, the number of rotatable bonds, the molecular weight, ClogP and CMR, the "maximal binding energy",[21] the number of potential chiral centers, and a simple measure of structural complexity determined by counting the number of bits present in the Daylight fingerprint. Some of these properties are calculated using external sofware (as in the case of ClogP and CMR[22]); others are determined using in-house programs.

In general the properties have been chosen to be ones which have a chemical or physical meaning that are familiar to and comprehensible by a bench scientist. The distribution of the properties over the set is determined and reported to the user both in the form of a histogram (Figure 7) and also via summary statistics for each property (mean, standard deviation). The user is then able to further refine the set of compounds by restricting the minimum and/or maximum value of one or more of these properties to a particular value. Those molecules which violate one or more of these requirements are eliminated from the set, the summary statistics are recalculated, and the graphical distributions are displayed to the user. We also provide hyperlinks to precalculated distributions from "standard" sets of compounds, such as the World Drug Index.[23] These distributions are intended to be used as a guide only; they are not intended

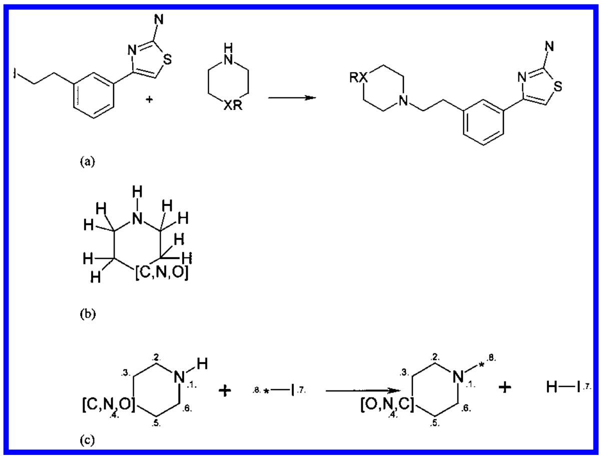


Figure 8. Generic reaction leading to formation of the aminothiazole array (a). The substructure used to identify the initial set of reagents is shown (b) together with the reaction used to enumerate the library (c).

to supplant any more detailed target-specific requirements that the user may wish to apply.

As outlined above, application of the profile constraints is often just the first phase in the library design/monomer selection process, identifying as it does those products in the virtual library which meet the requirements for that limited set of properties. The next step(s) depends on the additional factors that one wishes to take into account and upon the methods to be used for monomer selection. Perhaps the simplest situation arises when the only factors that need to be taken into account are present within the profile and where there is no combinatorial subset selection problem because there is only one point of variation (i.e., we have a 1×N array). Under such circumstances it may be appropriate simply to perform some form of cluster analysis on the remaining products to identify a structurally diverse set of monomers for the experiment. A simple hierarchical clustering method using Ward's algorithm with the intermolecular distances being determined from the Daylight fingerprints is provided within ADEPT for this purpose.

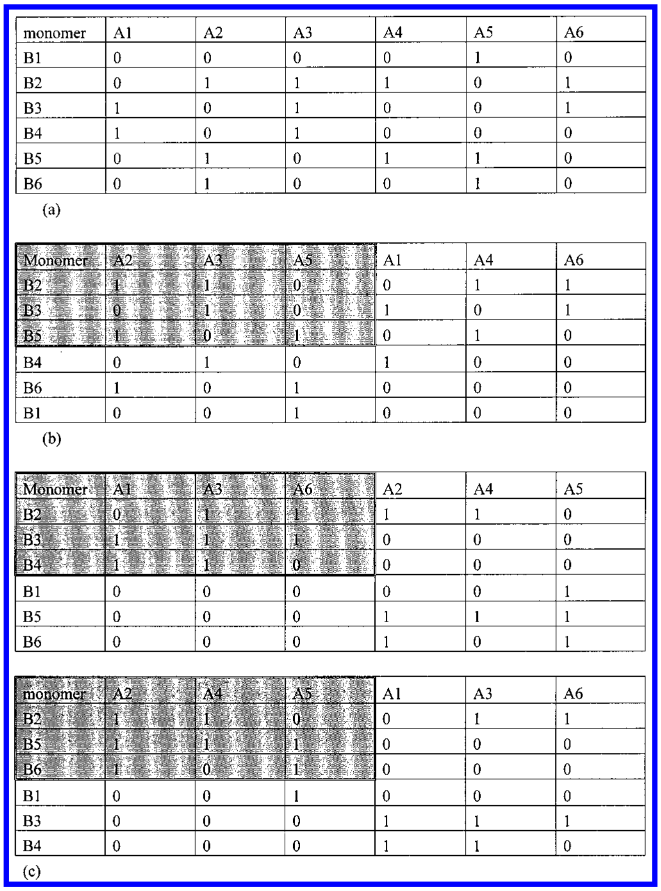
To illustrate how ADEPT would be used to follow the process outlined in Figure 1, we shall consider the synthesis of an array of aminothiazoles as described by Selway and Terrett.[24] The reaction scheme is shown in Figure 8a. Suppose we wish to identify a set of monomers from the available chemical database such that the final products must satisfy the rule of 5[1] (number of hydrogen bond donors $\leq$5; number of hydrogen bond acceptors $\leq$10; molecular weight $\leq$500; ClogP $\leq$5), have no more than eight rotatable bonds, and not contain any chiral centers. First we identify the initial reagent pool which in this case is achieved using the substructure query shown in Figure 8b. This query provides 278 hits from the available chemical database (version 97.2). This initial list was then filtered to remove molecules which did not contain just one secondary amine or which contained one of the following groups deemed inappropriate: carboxylic acid, β-ketoester, aldehyde, epoxide, isocyanate, primary amine, hydroxyl, phenol, β-lactam, hydrazine, azide, thiol, or primary halide. These filters reduced the list to 206. Enumeration of this virtual library was achieved using the reaction transform also shown in Figure 8c. Calculation of the profiles and application of the various filters reduced the list to 97, a 3-fold reduction over the initial list.

A SIMPLE STATISTICAL-BASED COMBINATORIAL SUBSET SELECTION SCHEME

Where there is more than one point of variation in a combinatorial library, the combinatorial subset selection problem must be considered. This issue was initially discussed in the literature in the context of diversity-based libraries where it was shown by Gillet et al. that more diverse libraries were obtained when the monomers were selected using a product-based selection method rather than by considering the monomers alone.[25] Such combinatorial subset selection is typically tackled using a sophisticated optimization procedure such as a genetic algorithm or simulated annealing. We have therefore investigated some simpler alternative approaches (which still make use of product properties) to determine their suitability for the more interactive Web environment. One such algorithm is as follows. For every monomer in each of the reagent sets it is possible to determine how many of the structures in the initial virtual library were considered acceptable after application of the profile constraints. Thus, each monomer has a score between 0 (i.e., none of the virtual products with that particular monomer were acceptable) and 1 (i.e., all of the products were acceptable). The monomers chosen for each position in the experimental library are then those which have the highest values.

The obvious drawback of this simple statistical approach is that the monomer scores are based upon the performance of each monomer for the entire virtual library and not upon its performance within an appropriately sized sublibrary. The more complex optimization procedures typically involve the generation of very many possible solutions, each of which is a sublibrary that must be individually assessed. The two approaches can be contrasted using the simple 6 × 6 library which is represented in tabular form in Figure 9. In these tables the presence of a "1" in the matrix indicates an acceptable product structure, whereas a "0" indicates a structure that is unacceptable. Suppose we wish to select 3 monomers from each of the two reagent pools to give a 3 × 3 library for synthesis. This is equivalent to interchanging rows and columns in the matrix to maximize the number of 1's in the 3 × 3 sub-matrix in the top-left corner. In this contrived example the statistical method would suggest a library that contains five acceptable structures, whereas it is possible to identify libraries that contain seven acceptable structures.

We desired to quantify the performance of the simple approach relative to more exhaustive optimization methods. To do this, we first constructed a "diverse" set of 100 carboxylic acids and 100 amines by searching the "medchem" database (as supplied by Daylight) to identify those molecules with just one of the appropriate functionality. Each of the two sets of compounds was then clustered to give 100 clusters with the molecule closest to the hypothetical centroid being chosen as the representative structure. The 100 × 100 virtual library was then enumerated and a profile calculated. We then attempted to select a series of 10 acids and 10 amines to give a series of smaller 100 compound libraries. Note that our use of cluster analysis in the selection of the monomer sets is purely to try to obtain a manageable set of monomers which in some way is representative of the entire set, and so has a spread of properties. As will be seen,



monomer	A1	A2	A3	A4	A5	A6
B1	0	0	0	0	1	0
B2	0	1	1	1	0	1
B3	1	0	1	0	0	1
B4	1	0	1	0	0	0
B5	0	1	0	1	1	0
B6	0	1	0	0	1	0

(a)

Monomer	A2	A3	A5	A1	A4	A6
B2	1	1	0	0	1	1
B3	0	1	0	1	0	1
B5	1	0	1	0	1	0
B4	0	1	0	1	0	0
B6	1	0	1	0	0	0
B1	0	0	1	0	0	0

(b)

Monomer	A1	A3	A6	A2	A4	A5
B2	0	1	1	1	1	0
B3	1	1	1	0	0	0
B4	1	1	0	0	0	0
B1	0	0	0	0	0	1
B5	0	0	0	1	1	1
B6	0	0	0	1	0	1

monomer	A2	A4	A5	A1	A3	A6
B2	1	1	0	0	1	1
B5	1	1	1	0	0	0
B6	1	0	1	0	0	0
B1	0	0	1	0	0	0
B3	0	0	0	1	1	1
B4	0	0	0	1	1	0

(c)

Figure 9. Comparison of statistical and sublibrary-based selection methods. In this simple example there are six monomers each of A and B, and the objective is to select three of each. The initial data are shown in (a) where a 1 is an acceptable structure and a 0 is an unacceptable one. The statistical approach produces the library in (b) which contains five acceptable structures. However, it is possible to identify at least two libraries which contain seven acceptable structures as in (c).

our objective in this exercise was to try to quantify the performance of various subset selection methods in designing libraries that meet certain criteria. We are not trying to identify the most diverse 10 × 10 library, nor are we trying to identify the "best" library that could be designed from the entire medchem database. Rather, we desired a relatively small number of monomers (100) such that the entire virtual library could be enumerated and thus the entire search space established.

The criteria used to design these libraries are provided in Table 1. We used five properties: the number of rotatable bonds, molecular weight, the maximal binding energy, the "complexity" (the number of bits in the Daylight fingerprint), and the calculated molar refractivity (CMR). We used the mean of these properties for the set of compounds from the WDI and took as the acceptable range approximately ±1 standard deviation, ±0.75 standard deviation, and ±0.5 standard deviation to give three sets of constraints of varying severity. The actual ranges of values used are shown in the table. We also selected a subset based on the rule of 5.

In each case we used the simple statistical method to identify the top 10 scoring acids and top 10 amines on the basis of their performance over the set as a whole. We also used a genetic algorithm[26] to select a library of the same size, optimized to maximize the number of "acceptable" product structures in the 10 × 10 library. The table shows the performance of these two methods as a proportion of the structures in each 10 × 10 library which match these criteria. As can be seen the simple statistical approach performs reasonably well if the proportion of acceptable molecules is not too low, but when this ratio falls below approximately 10%, the statistical solution is significantly worse than the fully optimized solution. There are two main advantages to the simple approach. First, it is trivial to compute and can often provide a reasonable solution very quickly. Second, as implemented within the Web environment, the user has full control over the selection process, with the statistical weights being provided as a guide only; he or she is at liberty to include his or her own "preferred" monomers if he or she so chooses. Nevertheless, we also encourage the use of more rigorous methods, especially as these are now able to take many different factors into consideration including the ability to design diverse libraries which fit a particular property distribution.[9]

ENUMERATION OF TAGGED LIBRARIES

A key development in combinatorial chemistry has been the ability to incorporate some form of molecular code or tag onto the solid support (i.e., a resin bead) along with the actual synthetic product. Tagging technology is important because it eliminates a fundamental problem with traditional "split mix" synthesis wherein the identity of only the final monomer to be added is known. When a tagged library is synthesized, a unique tag (or set of tags) that identifies each monomer is added to the bead before the beads are recombined, mixed up, and redistributed for the next phase. The advantages of tagging become apparent when it becomes necessary to determine the identity of the product molecule on a particular bead (for example, because the bead has been shown to contain an active component in an assay). The identity of the product can be determined directly by cleaving the tag from the bead in question and "reading" the tag.

Tagging technology raises a number of interesting computing and chemical information issues. One of these is concerned with the determination of the enumerated product structures present on a set of beads (e.g., because they have been identified as active in an assay). One of the tagging approaches in use at Glaxo Wellcome is based upon the secondary amine procedure reported by Ni et al.[27] Here a series of secondary amine tags are incorporated into a polyamide backbone on a differentially functionalized solid-phase support. A binary encoding scheme is employed which means that n tags can code for $2^n - 1$ reactions. For example, if one wishes to make a three-component library, then just 18 tags will encode for $(2^6 - 1)^3$ compounds.[27] The tags are typically identified using reversed-phase HPLC, and the tags are chosen to have distinctive retention times.

Information about the encoding scheme is typically stored as a table, an example of which is shown in Table 2. This also serves to illustrate that the "absence" of a tag is just as informative as the presence of a tag (i.e., some monomers may be coded using just a single tag, whereas other monomers would be coded using two or three tags) and that some of the tag combinations are subsets of others. We do not use the "null" tag. Suppose we have a series of beads to analyze. The input data would consist of information (derived

Table 1. Comparison of the Performance of a Simple Statistical Approach to Monomer Selection with a Genetic Algorithm[a]

experiment	constraints applied	total number of acceptable molecules in entire 10 000 virtual library	statistical method score (%)	genetic algorithm score (%)
1	3 ≤ rotatable bonds ≤ 7, 280 ≤ molecular weight ≤ 400, 6.5 ≤ binding energy ≤ 17.7, 150 ≤ complexity ≤ 250, 7.5 ≤ CMR ≤ 11	409	39	69
2	2 ≤ rotatable bonds ≤ 8, 275 ≤ molecular weight ≤ 425, 4.0 ≤ binding energy ≤ 20.0, 125 ≤ complexity ≤ 275, 6.75 ≤ CMR ≤ 11.25	1536	78	100
3	1 ≤ rotatable bonds ≤ 9, 250 ≤ molecular weight ≤ 450, 1.0 ≤ binding energy ≤ 23.0, 100 ≤ complexity ≤ 300, 6 ≤ CMR ≤ 12	3445	99	100
4	acceptors ≤ 10, donors ≤ 5, molecular weight ≤ 500, ClogP ≤ 5 (rule of 5)	6209	100	100

[a] In each case a 10 × 10 library was selected from a set of 100 acids and 100 amines. The table indicates the constraints applied in each experiment, the total number of molecules in the virtual library which met these criteria, and the proportion of molecules in each of the two 100-member libraries that were able to meet these criteria.

Table 2. Binary Encoding Scheme Used to Associate Monomers with Secondary Amine Tags[a]

monomer	tag 1	tag 2	tag 3
A	1		
B		1	
C			1
D	1	1	
E	1		1
F		1	1
G	1	1	1

[a] A total of seven monomers can be coded using three tags ($2^3 - 1$).

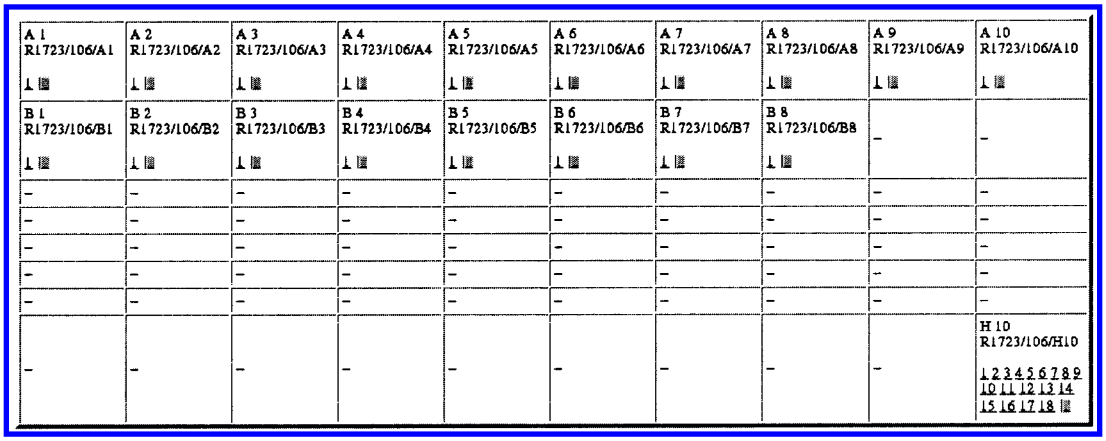
from an analysis of the RP-HPLC experiment) about which tags are present. The first step then is to determine which monomers these tags correspond to. The binary encoding scheme means that there is not a one-to-one correspondence between the presence of a given tag and a monomer; rather, it is necessary to consider the various combinations of tags. It is thus first necessary to check for the presence of monomers which are encoded by the largest number of tags and then check for monomers with fewer tags. For example, it would be necessary to check for the presence of monomer G in Table 2 (which is encoded by three tags) before the presence of monomer B (which is encoded by just one tag). Having identified the monomers present on each bead, we then create appropriate monomer lists and perform a parallel enumeration according to the reaction scheme so that we only produce the structures required. From these structures additional properties can be calculated to enable comparison with other physical analysis techniques (such as mass spectrometry).

REGISTRATION AND PLATE-MAPPING

ADEPT was primarily designed to assist chemists in designing their combinatorial libraries or arrays. However, it is also necessary to enumerate those libraries which are actually synthesized for registration purposes. Libraries or arrays containing more than a few compounds are most commonly synthesized on 96-well plates. Information about the distribution of monomers on these plates is contained within an in-house oracle-based system called GLIB. The registration of the monomers used at each well position are first registered into GLIB when the library is synthesized. In a separate step, ADEPT extracts the connection tables of the monomers from GLIB, and the user specifies the reaction scheme. The library is then enumerated. The resulting product structures are then associated with the relevant physical locations on the 96-well plates—the so-called *plate map*. This is achieved by cross-referencing the information within GLIB (which specifies which monomers were present in which wells) with that within ADEPT (which knows which product structures were generated from which combinations of monomers). The results are displayed as a series of "plates" (Figure 10). A hyperlink is provided to each structure, which provides a single mechanism to deal with both pooled libraries, and those made with one compound per well. Should several wells be of interest, then it is possible to select these wells and display the molecules they contain.

IMPLEMENTATION NOTES

ADEPT uses standard cgi form technology. A combination of tcl and perl scripts together with C and Fortran programs



A 1 R1723/106/A1 1 ▦	A 2 R1723/106/A2 1 ▦	A 3 R1723/106/A3 1 ▦	A 4 R1723/106/A4 1 ▦	A 5 R1723/106/A5 1 ▦	A 6 R1723/106/A6 1 ▦	A 7 R1723/106/A7 1 ▦	A 8 R1723/106/A8 1 ▦	A 9 R1723/106/A9 1 ▦	A 10 R1723/106/A10 1 ▦
B 1 R1723/106/B1 1 ▦	B 2 R1723/106/B2 1 ▦	B 3 R1723/106/B3 1 ▦	B 4 R1723/106/B4 1 ▦	B 5 R1723/106/B5 1 ▦	B 6 R1723/106/B6 1 ▦	B 7 R1723/106/B7 1 ▦	B 8 R1723/106/B8 1 ▦	–	–
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	H 10 R1723/106/H10 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 ▦

Figure 10. Plate-map display with wells containing both pooled and discrete samples. Each number is a hyperlink to the structure. In addition a number of wells can be selected and their contents displayed as a group.

(some from the Daylight “contrib” directory) are used on the server to process user requests and perform calculations. Html pages giving the results are produced dynamically from these scripts. Chemical structures are displayed in the browser using the Daylight daycgi software. Links to Oracle databases are achieved using the oraperl package. The ChimePro plug-in is used to enable structures to be copied to the Web browser from ISISDraw which can then be converted to the appropriate mol or rxn file using a shared object distributed with the chemscape Web server.[28]

CONCLUSIONS

ADEPT has now been used by a significant number of bench scientists as well as more expert computational chemists. We find that it enables us to provide bench chemists who have relatively little computational expertise access to some quite sophisticated software and is a good platform with which to introduce newcomers to some of the key ideas in compound selection and library design. Users can be trained to use the basic features of the system in about 2 h and then start to apply the methods in their own work. Its wide acceptance has also helped to remove some of the burden from the computational chemistry experts, thereby enabling them to concentrate on more complex problems.

ACKNOWLEDGMENT

Many individuals have contributed expertise and programs which have greatly facilitated the development of ADEPT. Darko Butina wrote early versions of the profiling and SMARTS filtering software. Xiao-Qing Lewell developed the complexity measure based on the Daylight fingerprints. Nick Hickman and Dan Ferko helped with the links to GLIB and Joe Lavelle with the availability data. Alfonso Pozzan developed the genetic algorithm used to evaluate the monomer selection method and also provided programming advice. Adrian Pipe was closely involved in the development of the tagging software. In addition to these named individuals, we thank the many other colleagues who have provided valuable feedback on the system, which has significantly improved its utility and applicability.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1−3), 3−25.

(2) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41* (18), 3325−3329.

(3) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification Of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 165−179.

(4) Gobbi, A.; Poppinger, D.; Rohde, B. Developing an inhouse system to support combinatorial chemistry. *Perspect. Drug Discov. Des.* **1997**, *7/8* (Computational Methods for the Analysis of Molecular Diversity), 131−158.

(5) Walters, P., http://www.daylight.com/meetings/mug99/Walters/index.html.

(6) Liu, M., http://www.daylight.com/meetings/mug99/Liu/index.html.

(7) Skillman, A. G., http://www.daylight.com/meetings/mug98/Skillman/recursivesmarts.html.

(8) Leach, A. R. Structure-based selection of building blocks for array synthesis via the World-Wide Web. *J. Mol. Graph.* **1997**, *15* (3), 158−160.

(9) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 169−177.

(10) Daylight theory manual, Chapter 4. Daylight Chemical Information Systems, Santa Fe, and http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

(11) The Available Chemicals Database is from Molecular Design Ltd., San Leandro, CA.

(12) Bailey, N.; Dean, A. W.; Judd, D. B.; Middlemiss, D.; Storer, R.; Watson, S. P. A convenient procedure for the solution phase preparation of 2-aminothiazole combinatorial libraries. *Bioorg. Med. Chem. Lett.* **1996**, *6* (12), 1409−1414.

(13) Chapman, D. Reaction-centered informatics for combinatorial chemistry. *Book of Abstracts*, 213th ACS National Meeting, San Francisco, April 13−17, American Chemical Society: Washington, DC, 1997; CINF-064.

(14) Daylight theory manual, Chapter 7. Daylight Chemical Information Systems, Santa Fe, and http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html.

(15) Delany, J., http://www.daylight.com/meetings/mug97/Delany/rxntk/projects0.html.

(16) ISISDraw is available from Molecular Design Ltd., San Leandro, CA.

(17) ChemDraw is available from CambridgeSoft Corp., Cambridge, MA.

(18) Chime is available from Molecular Design Ltd. and incorporates graphics code written by Roger Sayle.

(19) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244−55.

(20) A similar program has been described by Barnard, J., http://www.daylight.com/meetings/mug99/Barnard/index.html.

(21) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional group contributions to drug-receptor interactions. *J. Med. Chem.* **1984**, *27* (12), 1648−57. This is a generic binding affinity calculated using a group-additive scheme, with the group contributions being derived from a regression analysis of known drug−receptor affinities.

(22) Leo, A. J. Calculating log P_{oct} from structures. *Chem. Rev.* **1993**, *93* (4), 1281−306.

(23) The World Drug Index is available from Derwent Information, 14 Great Queen St., London WC2B 5DF, U.K.

(24) Selway, C. N.; Terrett, N. K. Parallel-compound synthesis: methodology for accelerating drug discovery. *Bioorg. Med. Chem.* **1996**, *4* (5), 645−654.

(25) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (4), 731−740.

(26) A. Pozzan, unpublished program.

(27) Ni, Z.-J, Maclean, D.; Holmes, C.; Murphy, M. M.; Ruhland, B.; Jacobs, J. W.; Gordon, E. M.; Gallop, M. A. Versatile Approach to Encoding Combinatorial Organic Syntheses Using Chemically Robust Secondary Amine Tags. *J. Med. Chem.* **1996**, *39*, 1601−1608.

(28) Chemspace is available from Molecular Design Ltd., San Leandro, CA.

CI9904259