JOURNAL OF combinatorial CHEMISTRY

Perspective

On the Representation of Combinatorial Libraries

Derek Maclean, and Eric J. Martin

J. Comb. Chem., **2004**, 6 (1), 1-11• DOI: 10.1021/cc0340325 • Publication Date (Web): 25 November 2003

Downloaded from http://pubs.acs.org on March 20, 2009

JOURNAL OF
combinatorial
CHEMISTRY

Volume 6, Number 1 January/February 2004

Perspective

On the Representation of Combinatorial Libraries

Derek Maclean[‡] and Eric J. Martin*[,†]

PETNET Pharmaceuticals, 6140 Bristol Parkway, Culver City, California 90230, and Chiron Corporation, 4560 Horton Street, Emeryville, California 94608

Received July 29, 2003

Introduction

Combinatorial chemistry exhibits both quantitative and qualitative differences from other fields of chemistry, posing unique challenges to data storage and communication within the field. The most striking feature is the sheer number of compounds that may be studied. It is possible to prepare more compounds in one library than exist in entire chemical registries, but the level of characterization may range from definitive analysis to computer prediction. More subtly, combinatorial synthesis demands definition of the *relationships* between compounds. There is a great difference between testing a mixture of one hundred compounds and testing the same set of compounds individually. Finally, as in any emerging field, the need for scientists to describe their work generates a new vocabulary. As terminology and structural representations develop on an ad hoc basis, the opportunities for confusion grow. As a multidisciplinary endeavor, combinatorial chemistry is particularly susceptible to this issue.

To address these issues, a working party was formed within the Medicinal Chemistry section of IUPAC. The direction of this effort has been to capture the means by which combinatorial libraries are described, and to search for common patterns of usage within the field. We have also tried to identify the benefits to the field of improved clarity in communication and data exchange. In particular, we believe that this may in part be achieved by enabling a more standardized representation for combinatorial libraries. This article describes some aspects of the resulting analysis. Topics that will be addressed include the terminology of combinatorial chemistry and structural representation of libraries, which covers analysis of generic structures, building blocks, pool notation, and generic reaction schemes.

Terminology of Combinatorial Chemistry

An attempt to capture the terminology of the field has been the most visible contribution thus far of the IUPAC Working Party. The *Glossary of Terms Used in Combinatorial Chemistry* was published in 1999[1] with definitions of almost 150 terms. An excerpt is shown in Figure 1. In IUPAC terminology this is a Technical Report; that is, it is an authoritative review of usage in the field, but it does not carry the full weight of IUPAC Recommendations.[2] Our goal is to bring the document to Recommendation status after a suitable period of consideration. It will also be desirable to provide the Glossary in a web format to improve its accessibility and visibility and to offer a more dynamic format in which entries could be added or modified as justified during the transition to Recommendation status. The challenge is to achieve the right degree of flexibility to new developments while preserving the authoritative status of the resource. The parallel IUPAC initiative *Standard XML Dictionaries for Chemistry*[3] is designed to maximize accessibility and exposure to resources such as the Glossary.

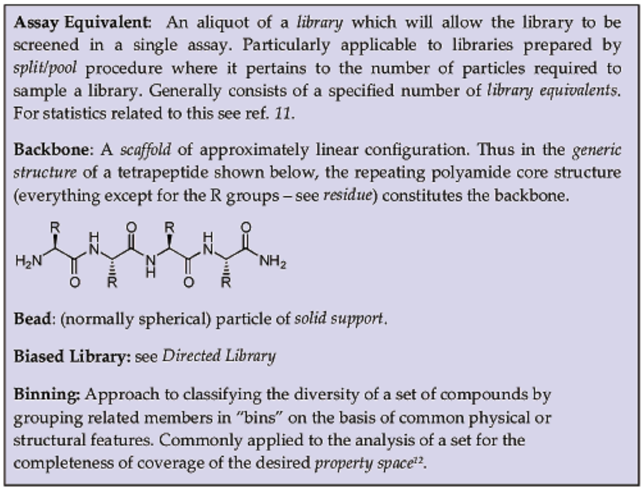
The development of documents such as the Glossary is important in facilitating communication of concepts and ideas

* E-mail: eric_martin@chiron.com.

[†] Chiron Corporation.

[‡] Current address: KAI Pharmaceuticals, 270 Littlefield Ave., South San Francisco, CA 94080. E-mail: maclean@netcom.com.

10.1021/cc0340325 CCC: $27.50 © 2004 American Chemical Society
Published on Web 11/25/2003



Assay Equivalent: An aliquot of a *library* which will allow the library to be screened in a single assay. Particularly applicable to libraries prepared by *split/pool* procedure where it pertains to the number of particles required to sample a library. Generally consists of a specified number of *library equivalents*. For statistics related to this see ref. *11*.

Backbone: A *scaffold* of approximately linear configuration. Thus in the *generic structure* of a tetrapeptide shown below, the repeating polyamide core structure (everything except for the R groups – see *residue*) constitutes the backbone.

Bead: (normally spherical) particle of *solid support*.

Biased Library: see *Directed Library*

Binning: Approach to classifying the diversity of a set of compounds by grouping related members in "bins" on the basis of common physical or structural features. Commonly applied to the analysis of a set for the completeness of coverage of the desired *property space*[12].

Figure 1. Excerpt from *Glossary of Terms Used in Combinatorial Chemistry* (ref 1).

within the field. A practical illustration of the utility of such a tool is the VCH thesaurus of chemistry by the publishers of *Angewandte Chemie*.[4] Authors must choose at least two out of five keywords from the thesaurus to help ensure a common frame of reference and increase the quality and completeness of literature searches.

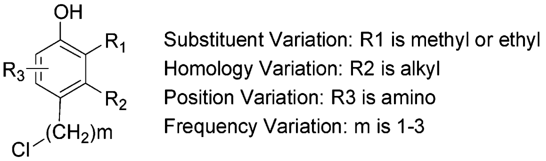
Structural Representation of Combinatorial Chemistry

The specialized structural representation of combinatorial chemistry is as important and potentially confusing as its specialized vocabulary. This article will focus on four problems of structural representation that are unusual to combinatorial chemistry: (1) representing the compounds that comprise a library, (2) describing the reactions that generate a library, (3) expressing substructural similarities among library members (such as specification of a common scaffold), and (4) specifying subsets of a library (e.g., pooled members). In each of these areas, a variety of approaches have been described in the literature, and that which works best generally depends on the author's specific purpose. In the subsequent sections, we wish to point out advantages and disadvantages of some of these representations, and where possible, to identify improved strategies to anticipate and address the multiple goals that may be desirable in representation.

Specification of Library Members

The most basic description of a combinatorial library is to specify the compounds that comprise the library. This can always be achieved by tabulating the enumerated library members. However, the scale of combinatorial libraries presents a significant limitation on presentation via the printed page, slide, computer screen, or even computer memory. Compressed library representations that maintain clarity, accuracy, and completeness are therefore important and desirable.

In this respect, combinatorial libraries have much in common with the chemical patent literature, in which the description of large numbers of related compounds is a frequent goal. In the patent field, however, completeness and accuracy are the defining principles, with clarity and conciseness distant or irrelevant goals.[5] Nevertheless, the patent literature shares with combinatorial chemistry the desire to find an appropriate method for structure representation and compression.



Substituent Variation: R1 is methyl or ethyl
Homology Variation: R2 is alkyl
Position Variation: R3 is amino
Frequency Variation: m is 1-3

Figure 2. Forms of variation in generic structures (after Barnard and Downs (ref 6).

The *generic* or *Markush* structure is the primary tool used to condense the structural representation of sets of compounds. Generic structures can depict on a single page libraries that would fill a book if fully enumerated. This compression is possible as a result of the regularity of the library. Generic structures have a long history of use in patents and are ubiquitous in combinatorial chemistry publication. They consist of the common (core) structure of the library with one or more "superatoms" attached (often represented by R, for "residue" or "radical") indicating the existence of variable substituents at that location.

Barnard, Downs, and colleagues have extensively described the use and limitations of generic structures.[6] They define four forms of variation found in these structures (see Figure 2). Of these, the first (*substituent variation*) is essentially the only form of variation found in real combinatorial libraries. This implies the provision of a list of specific chemical substructures that may be interchanged in all possible combinations at the indicated positions. The other types of variation may help depict the composition of a combinatorial library, for instance, *homology* or *positional variation* may help condense the generic structure for more concise presentation.

The combination of generic structure plus substituent lists may be termed the *generic representation* of a library. For a library in which all combinations of substituent are present (*fully combinatorial*) the generic representation will be an accurate description of library composition. If only a subset of possible combinations are present, then the generic representation is simply an indication of the scope of the library.

A number of factors can complicate the use of generic representations: libraries may have multiple cores, ring forming attachments, or correlated sets of substituents. Several examples are shown in Figure 3. Barnard and Downs[6a,b] discuss many additional cases in which defining simple (intuitive) generic representations is challenging. As they conclude, it is in general possible to adequately describe combinatorial libraries with more complex, nonintuitive generic representations, which can be coded precisely for computational analysis, although they may be unwieldy for visual inspection. In these cases, it is recommended that a small sampling of enumerated whole products is displayed as examples to aid the reader.

Additional complications may arise in the case of libraries which do not belong to the typical class of small organic molecules, for instance, libraries of polymers or organome-

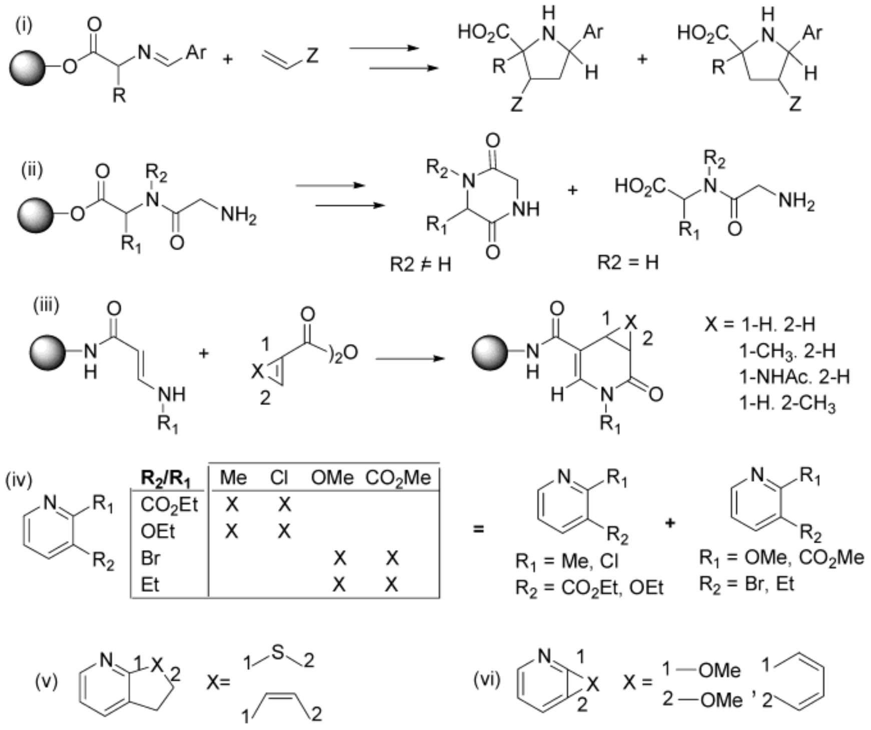


Figure 3. Complications in generic presentations. (i, ii) Multiple cores arising from regioisomers in cycloaddition reaction (i) or substituent-dependent reactivity of intermediates (ii) (dipeptide fails to cyclize if R2 = H). In each case, the resultant library (a "mixture of mixtures") can be represented as the union of two generic structures (the various possible stereochemical outcomes in (i) further complicates the issue). (iii, iv) In "correlated substituents", only certain permutations of substituents are used. For irregular combinations, that is, because two or more substituents derive from the same building block, this can be represented using disconnected bivalent substituents, that is a "pseudo ring" (iii). In more regular cases, as in a "block" combinatorial library, this can be presented using a list (iv, left) or registered as multiple generics (iv, right), analogous to (i) and (ii). (v, vi) Ring substituents may be indicated by bivalent residues, although this may be confusing when the resulting ring varies in size (v). When substituents sometimes join to form a ring (vi) they can be joined into a single, sometimes disconnected, bivalent substituent.

tallic complexes. Unfortunately, at this point in time, the systematic nomenclature of these classes of compounds lags that for more simple organic molecules, even for individual representatives of these classes.

Choice of Generic Structure. Any combinatorial library contains a large number of possible generic structures, ranging from the structural overlap between any two library members to the *maximum common substructure* (MCS) for the entire library. Which of these possible cores is most desirable depends on the context in which it will be used. Often, the most useful choice of the core derives from the synthetic route used to create the library. To an experienced chemist, this indicates the additional diversity likely to be accessible for the series. For easier visualization of the diversity of an actual synthesized library, the most useful core is more often the MCS among all the compounds, which may be much larger than the synthetic core. Comparing a peptide sequence representation to the alternative representation as a substituted polyglycine [Figure 4i] illustrates this distinction. In other instances the most desirable core structure is *larger* than the MCS, such as when the inclusion of a Markush atom in the core avoids breaking a ring [Figure 4ii].

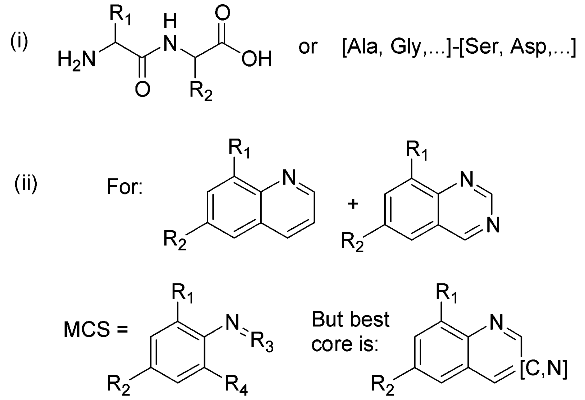


Figure 4. Maximum common substructures. (i) A combinatorial dipeptide library represented as a sequence or as substituents on a polyglycine core. (ii) The best core may be larger than the MCS.

The possible generic structures have an inherent hierarchy, since smaller core structures (more generic, less specific) are fragments of larger ones, and common features of large subsets are fragments of the core structure of small sets. Sublibraries can be arranged in a (nonunique) nested fashion, since all the members of the subset must be members of the library (see Figure 5).

In a real-world example, Linusson et al. described a library whose common synthetic core was *m*-hydroxyphenol sulfate (Figure 6).[7] This small core with three positions of variability shows the diversity which is accessible in principle by the chemistry employed (Figure 6i). However, the authors chose instead to present the MCS for the set of compounds they actually synthesized (Figure 6ii). This core is more than twice

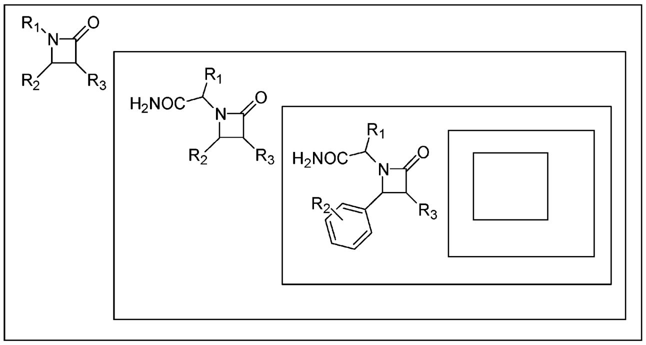


Figure 5. Hierarchy in generic structures. β-Lactams may be constructed with variable substituents at three sites (R_1-R_3; outer rectangle). A library which makes use of only α-amino acids at the R_1 position has a more narrowly defined generic structure and is a subset of the larger library. Similarly, a further structural restriction uses only aromatic building blocks at the R_2 position, and so on. Any of the three generic structures shown are valid for the library described by the innermost. However, the innermost generic may not be valid for another library described by the other generics but with different choice of substituents.

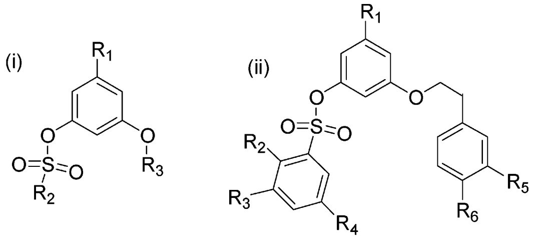


Figure 6. Choice of generic structure highlights either: (i) synthetic flexibility or (ii) medicinal chemistry SAR.

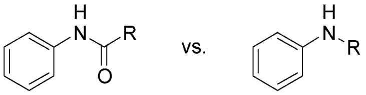


Figure 7. Choosing the smaller core with the conformational restricted amide bond attachment, by absorbing the carbonyl into each of the substituents, aids pharmacophore similarity calculations in the OSPreyS method.

as large and shows that the actual library members were quite similar, with many small substituents attached to various positions around several constant benzene rings designed to interact with various pockets in the binding site of the desired biological target.

Another series of contrasting examples can be found in the computational chemistry literature. Distributions of properties for virtual libraries can sometimes be computed as a function of the property distributions of the corresponding fragments. In this case, the core should be chosen to factor the properties as independently as possible. For example, the OSPPreyS method estimates pharmacophore similarity between library product molecules and assumes the same set of rotamers around the bond that attaches the R-group to the core.[8] Hence, choosing an attachment bond with a strong conformational preference, such as an amide bond, minimizes the influence of the core and other substituents on the conformations of the isolated substituents (Figure 7).

By contrast, the topomer databases of Cramer et al. use a fragment-based method to allow fast similarity searching of huge virtual combinatorial databases and are most powerful whenever a reaction is described as the formation of an acyclic bond between the clipped reagents, so the authors recommend that the common core of the virtual library is ideally reduced to a single amide (or other) bond.[9]

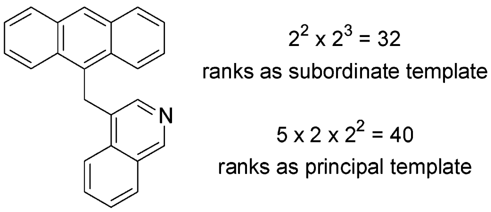


Figure 8. Designation of principal generic structure. After Katritzky et al. (ref 11). The isoquinoline and anthracene moieties are recognized and scored as potential generic structures according to simple rules, with greater (numerical) priority in this example being accorded to the former.

The closely related 3-D QSAR method of topomeric CoMFA, on the other hand, assumes that the differences in activity can originate only from the differing portions of the structures, so Cramer recommends that the largest common core should be selected.[10]

In a Perspective in the *Journal of Combinatorial Chemistry*, Katritzky et al. addressed the topic of identifying a "principal" generic structure of a library by proposing a systematic, rule-based, quantitative assessment of interlibrary diversity.[11] A set of rules were developed (reminiscent of the IUPAC rules for assigning priority in organic compound nomenclature) to analyze the alternative generic structures that may be represented within a set of compounds. An example is shown in Figure 8. The rules are based substantially on principles of medicinal chemistry. While the weighting factors are necessarily somewhat arbitrary, this effort begins to address the question of where one library may be considered to stop and another starts within a set of compounds. One significant limitation of this treatment is that the selection of principal generic is dependent on the choice of substituents which decorate it. Thus, choosing different building blocks for the same library chemistry may lead to designation of a different principal generic (compare the situation illustrated in Figure 5). A useful extension to this approach would be a more flexible and systematic strategy to identify all relevant generic structures and a means for selecting that best suited to the desired task. A computational approach would be desirable, perhaps based on the well-known clustering algorithms for grouping compounds according to some measure of structural relatedness.

Generic Reaction Schemes. Just as a generic structure is an abbreviation for all of the enumerated compounds in a library, the combination of generic structures for some or all of the library reactants, reagents, intermediates, and products comprises the *generic reaction scheme*. Unlike the library products, however, which are often enumerated, the complete set of individual reactions used to form a combinatorial library are virtually never enumerated.

Generic reaction schemes are often presented as a "reaction-based" alternative to "product-based" representations of a combinatorial library. While it is true that a synthetic chemist can often infer the contents of a library from a generic reaction scheme and a list of reactants, and vice versa,

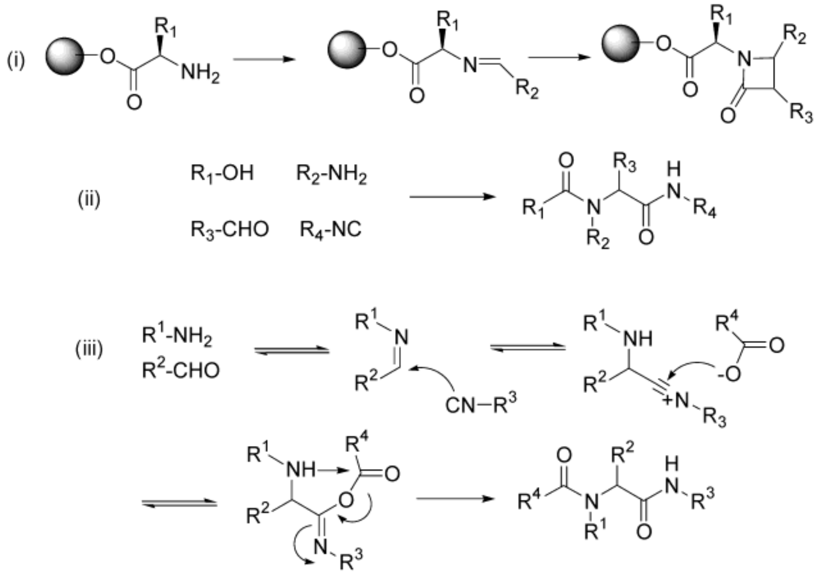


Figure 9. Temporal and positional residue labeling. (i) Temporal labeling: R_1-R_3 indicate the order of incorporation of the corresponding reagents. (ii) Positional labeling: R_1-R_4 indicate the location of the residue within the generic structure. (iii) Temporal labeling for one-step reaction. Note differences in R-numbering between (ii) and (iii).

the generic reaction scheme both carries additional information and also puts limitations on the description of the products. Notably, many of the choices of what to consider the scaffold or how to represent a mixture of mixtures are already determined by the context of the reaction scheme. Many database registration systems allow inputting the reaction to simplify library registration, as the reactant structures are already available and this obviates the steps of drawing a scaffold and inputting or clipping the substituents. If the database stores only the enumerated products, and not the generic substructures of the products, this is fine. However, if the database stores an abbreviated Markush structure for brevity and clarity, this approach often dictates a nonoptimal choice for the core (see above) or a reaction scheme that is even more complicated than clipping the reagents. The cycloaddition reactions presented above (Figure 3i,iii) that formed the heterocyclic core during the reaction, and the SAR example (Figure 6) provide examples of this dilemma.

The R designation with an ordinal suffix (e.g., R_1, R_2, etc) often indicates the order in which the corresponding building block was introduced in the generic reaction scheme. This *temporal labeling* may be contrasted with *positional labeling* (see Figure 9), in which the residue number simply indicates the (arbitrary) location of the substituent. Temporal labeling of residues has greater information content than positional labeling, since the order of incorporation may be inferred even from an isolated generic structure numbered in this fashion. Thus, the numbering scheme in the final generic structure in Figure 9ii provides synthetic and mechanistic insight even in the absence of the reaction scheme. In the context of a multicomponent reaction, although there may be no specific order of reagent (and R group) addition (e.g., the Ugi reaction in Figure 9ii), there may perhaps be value in numbering R groups to indicate the order of incorporation in the proposed mechanism (iii).

Positional encoding may have utility in particular cases, for instance, in discussion of structure–activity relationships within sets of compounds for which it may be useful to have a common orientation of substituents independent of the reaction scheme.

Treatment of Building Blocks: Residues and Superatoms

We described above how the composition of a library can be defined by a generic representation which combines a generic structure (which defines that portion common to all library members) and one or more lists of substituents (which define the diversity of the library). These lists tend to be larger and more difficult to compress than the generic core. For a generic reaction scheme, the list will comprise the full structure of the reagents that were used to prepare the library. For a generic representation of the library, only the residual portion of each building block that becomes incorporated in the final library products is listed. The process of stripping off the extraneous parts of the reagent to the residue or radical is known as *clipping* (see below). For libraries of even modest size these lists may be lengthy, and visual inspection may be of limited utility. Of course, the building blocks or clipped residues may be represented by their own generic representations. In general, the depiction of lengthy lists of building blocks may be best left to the supporting information section of publications, especially if a suitable electronic form of these lists were to be made available (see below).

Figure 10 illustrates strategies that have been used to deal with clipping and some of the issues that may arise. It is important to ensure clarity and accuracy in representating the link between core and residue. This is especially true for bifunctional residues which have an inherent ambiguity in their attachment.

As described above, the use of the superatom R is widespread; however, many other superatoms are represented in the literature. In a general sense, it is desirable that the superatom designation be as informative as possible. Thus, the use of Ar, alkyl, PG, LG, etc. to indicate that only certain types of substituent have been used may be encouraged, as

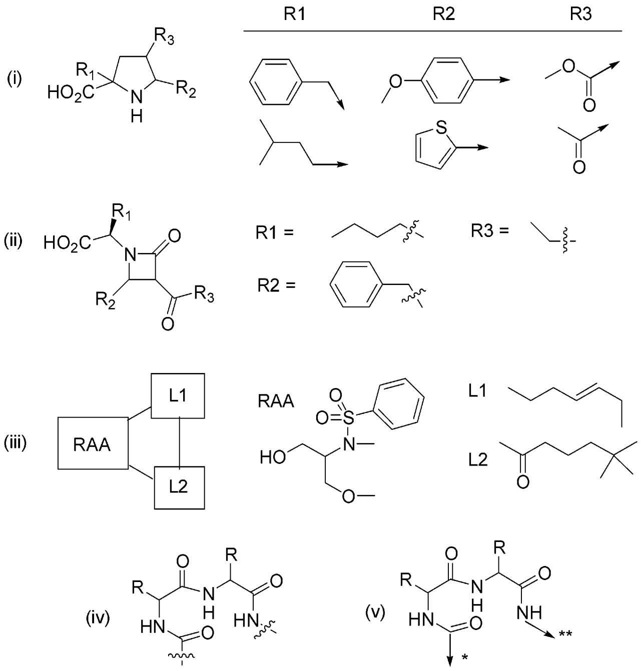


Figure 10. Representation of clipped residues. Examples (i), (ii): simple residues where the bond joining the residue to the core structure is indicated by an arrow or wavy bond. Examples (iii)−(v) show residues attached to the core through two bonds, where designation of orientation of attachment is useful. Thus, in (v) the single or double star would correspond to a similar mark on the generic core. In contrast, (iii) and (iv) have equivalent substitution positions and offer some potential for confusion.

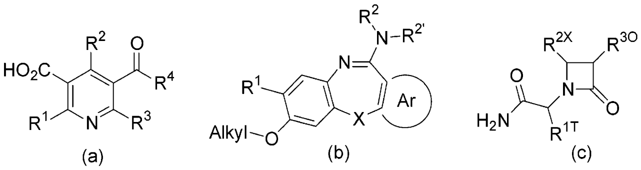


Figure 11. Use of superatoms. (a) Straightforward use of R1−4. (b) R2, R2′: indication that substituents derive from the same reagent and are not independently varied; Alkyl, more descriptive than R; Ar with ring, fused aromatic ring; X, bifunctional residue. (c) Additional qualifiers indicate relationship between library members using the pool notations described in the text (see Figure 16). Thus, the indicated library has R3 separate pools, each containing R2 × R1 members, with the R1 position being encoded.

more specific information is conveyed than the more generic R designation.

The use of additional modifiers may be useful, such as R_1^T to indicate that a variable position is associated with tags in an encoded library. Figure 11 illustrates the use of alternative superatoms. Similarly, modifiers X and O may be used to convey information pertaining to pool notation in a structural context (see the section on pool notation, below). The risk with these additional layers of complexity is that the generic structure may become cluttered or confusing. Authors should assess each situation on its merits.

Another ubiquitous feature of combinatorial reactions is the use of solid supports, linkers, and supported reagents. In many generic reaction schemes, these are handled as a special type of superatom with pictorial notation (see Figure 12). This is an area of particular importance to the field of combinatorial chemistry and which continues to develop at a rapid pace. The past few years have brought to prominence fluorous supports,[12] dendrimers,[13] and other additions to the more traditional solid or soluble supports. A variety of creative representations have been used to illustrate this diversity of techniques. Some of these are shown in Figure 12a and b. The nature of the support is important information to describe in primary publications and record in secondary sources such as reaction databases; however, this is often not captured effectively in such resources. The IUPAC Glossary included definition of some of the more commonly used supports and linkers (TentaGel, ArgoGel, Rink, Wang), but a more general approach was not attempted because of the complexity of the area. The standardization of simple notations may be useful, such as the use of a filled symbol (circle, square, or whatever) to indicate a primarily insoluble support, or an open symbol for a soluble support. The use of the circle to indicate a beaded support is already ubiquitous. Some degree of standardization may improve the elegance, precision, and clarity of communication, but it is more important that novelty and creativity are not hampered by such an effort.[14]

Interestingly, a significant factor in the selection of representations for solid supports is the choice available in chemical drawing software packages. In fact, the s orbital has become the de facto standard for depicting a resin bead (perhaps much more widely used than for the designed purpose?), and other drawing tools are commonly used for other types of support. It is surprising that after many years of practice of solid-phase chemistry and the widespread use of solid-supported combinatorial methodologies, that software vendors have not yet provided a special set of pictograms to represent beads, surfaces, etc. A gaping deficit in the use of existing tools, such as the s orbital, is that no point of attachment is available on these devices. Every practicing chemist in the field will be familiar with the frustration of keeping a molecule "attached" to the "resin bead" while constructing chemical schemes. A "smart bead" which will accept molecular linkages has been suggested by these authors to a number of software providers, but without effect (or acknowledgment!) to date.

Recently, a standardized terminology for linkers was proposed by Comely and Gibson.[15] Their PACT notation (Point of Attachment Converted To; see Figure 13) focuses on the residue left on the compound of interest after cleavage from the linker. This notation will be useful for categorizing linkers and certainly could form a field in computer registry for combinatorial syntheses. A useful addition to this scheme would be notation describing the cleavage conditions.

Specification of Noteworthy Subsets; Pool Notations

It is often important to specify particular combinatorial subsets of a larger library: when split−pool solid-phase library synthesis result in pools of compounds sharing one or more common reagent, when individual compounds can be intentionally pooled to facilitate screening (e.g., as orthogonal mixtures), and when successively smaller subsets are synthesized during deconvolution. To define an appropriate notation for pools or mixtures of compounds, we may first identify that the synthesis route can be described in linear terms even for nonlinear products (see Figure 14). Thus,

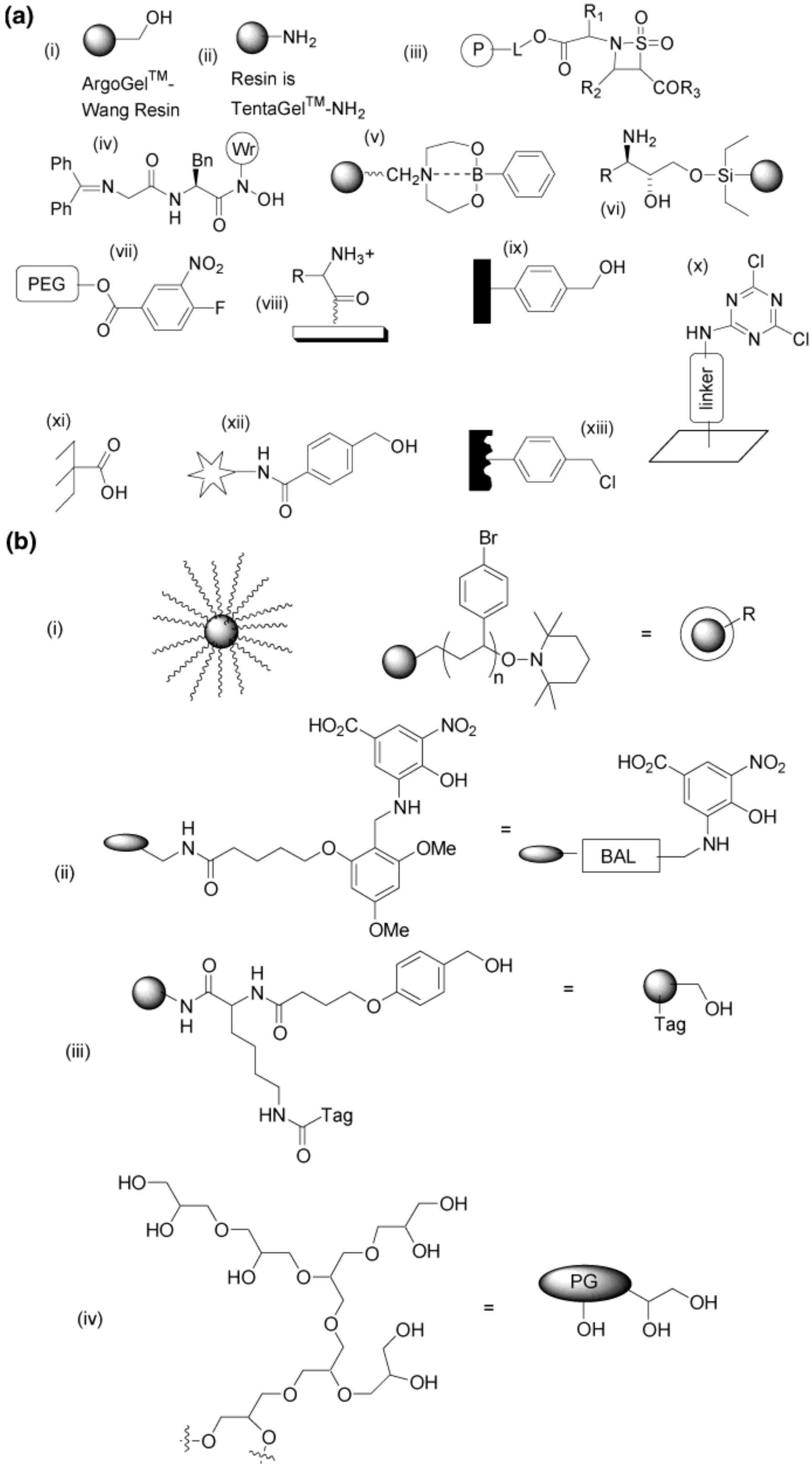


Figure 12. (a) Representation of supports and linkers for library synthesis. Examples (i) and (ii) show "standard" functionalized beaded supports with explicit annotation of the nature of support and linker. In (iii), the nature of the support and linker would be expected to be defined elsewhere, like (iv) where the unusual "Wr" symbol on the bead indicates a "nonstandard" beaded support and directs the reader to the text for complete description. Examples (v) and (vi) suggest standard beads with unusual linkers; (vii) indicates the use of a soluble support; (viii)−(xiii) indicate various unusual solid supports, respectively (as defined in the corresponding texts) grafted MicroTube, glass slide, polypropylene membrane, polymer Crown, monolithic polymer disk. (b). Representation of supports and linkers for library synthesis. In each example, a shorthand depiction of the support is defined by initial depiction of a more complete entity. This allows clear and concise illustration of complex supports, such as the tagged resin in (iii) with two types of functionality or the dendritic polyglycerol in (iv) with three different classes of hydroxyl residue. It is useful to distinguish cartoons, which are purely illustrative (such as the "hairy bead" in (i)), from depictions with real chemical meaning.

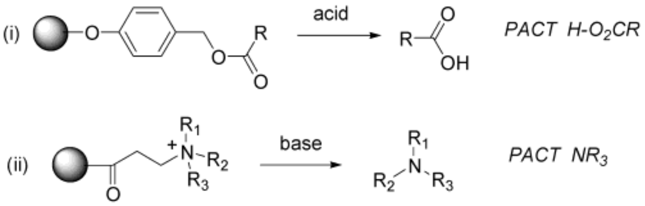


Figure 13. Comely−Gibson linker notation. Linkers are defined by the residual functionality on the released compound. Thus, the Wang linker (i) is defined as a PACT H−O$_2$CR linker, where the point of attachment is converted to the H−O residue. The hyphen indicates the new bond that is formed on cleavage. Cleavage of the REM linker (ii) is PACT NR3. Note no hyphen because no new bond is formed on cleavage.

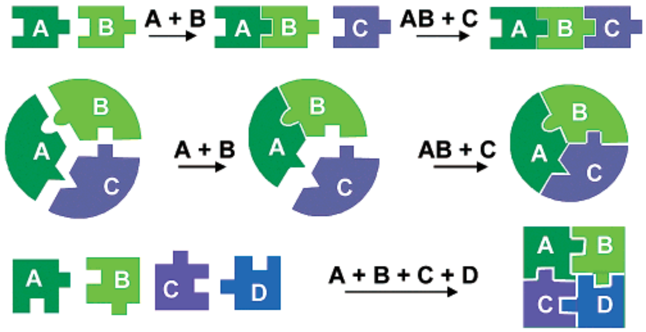


Figure 14. Linear notation (ABC...) describes distinct product types. Linear notation can describe both linear and nonlinear products, arising from a series of stepwise reactions or single-step, multicomponent condensation.

while oligomeric compounds such as peptides are naturally described by a string of letters, the structure of any member of a combinatorial library may be defined by equivalent shorthand. By simple extension, a nomenclature for the relationship of all compounds in a library may be defined.

Given the well-defined assignment of residues, the nomenclature of any given library member becomes clear. The *Journal of Combinatorial Chemistry* has defined notation for library members, whereby **3**{*8,4,1*} represents a compound of generic structure **3** with R$_1$ = 8, R$_2$ = 4, and R$_3$ = 1, indicating the particular residue from a separately defined list.

This scheme is extended to the *ChemSet* notation to describe pools of compounds. Thus **4**{*1−35*; *1−12*; *1−20*} describes a set of compounds of generic structure **4** having R$_1$ = all of residues 1−35, R$_2$ = residues 1−12, and R$_3$ = 1−20. This is a concise description of 8400 compounds. However it is not clear from this notation what the *relationship* between the compounds may be. Is it one pool of 8400 compounds, 8400 separate samples, or something else?

An alternative treatment is that of Houghten, originally used for peptide libraries.[16] In addition to the usual single letter codes for amino acids, the use of the letters X and O describe, respectively, “all possible residues, mixed together”, and “all possible residues, kept separate.” Figure 15 shows some examples of the use of X and O. These may be useful additions to the ChemSet nomenclature, since the information conveyed by O is hard to represent in ChemSet, and X is a more concise version of {*1−35*}. X is also more precise, since it may be unclear if {*1−35*} is all or only a subset of possible residues at the indicated position. Thus, these additions may more precisely define the relationships between the indicated compounds. An illustration of two possible consensus notations using the best features of both systems is shown in Figure 16. Note in particular the fourth row, which effectively describes collections of single compounds from parallel, split−sort, or split−only synthesis and has no comparable notation in the existing ChemSet scheme.

It should be noted that description of a collection of a “cherry picked” set of discrete compounds prepared by parallel synthesis cannot be readily condensed by these methods. In this case, the compounds may either be represented by a list of the residue combinations used, or be fully enumerated.

Enhanced Data Exchange for Combinatorial Chemistry: Electronic Databases

The quantity of data associated with combinatorial library synthesis, analysis, and testing is a serious obstacle to the efficient utilization of literature data in this field. In addition, the printed page is a less than adequate medium as a primary repository of much of the information associated with such studies. Building on reported data is the foundation of all scientific endeavor, and opportunities are undoubtedly being missed to capitalize on the rich resource of reported studies in combinatorial chemistry, owing to the difficulty in accurately accessing primary information.

It is not difficult to imagine how enhanced data access could impact the field. For instance, the development and critical comparison of tools for library analysis, such as diversity metrics, would be facilitated. The use of a particular type of building block or the synthesis of related compounds

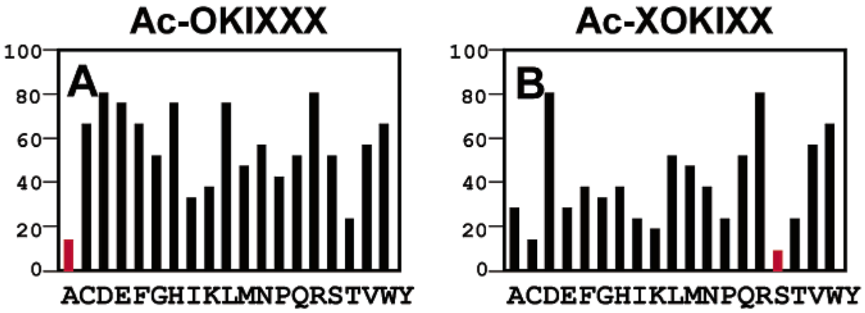


Figure 15. Positional scanning of peptide libraries. After Houghten (ref 16). Illustrating the use of X and O for describing the relationship between pools. Each graph shows the biological activity for 20 pools of peptides. Pools are defined by the generic sequence above each graph. Each pool contains a single, defined, amino acid residue at the position indicated by O, and a mixture of all 20 possible residues at the positions indicated by X.

JCC Chemset	Consensus1	Consensus2
8{*1-10; 1-20; 5*}	**8**{*1-10, X, 5*}	**8**{*O 1-10; X 1-20; 5*}
6{*8; 4; 1-10*}	**6**{*8, 4, O*}	**6**{*8; 4; O 1-10*}
2{*1-35; 1-35; 1-35*}	**2**{*X, X, O*}	**2**{*X 1-35;X 1-35;O 1-35*}
5(parallel synthesis)	**5**{*O, O, O*}	**5**{*O 1-20, O 1-20, O 1-20*}

Figure 16. Comparison of pool nomenclature. Consensus 1 combines Houghten and ChemSet nomenclature. Consensus 2 adds the additional specification of the number of building blocks in each reagent set.

may be more readily tracked across different studies. Development of pharmacophore models could be achieved by associating structure with biological activity between publications.

There are several instances in which a large data set generated by one group has been reanalyzed by a second group, resulting in additional insight into library design or biological activity. Zuckermann et al. reported the identification of α1-adrenergic antagonists from a library of 5,000 N-substituted glycine peptoids.[17] Bradley et al. subsequently reanalyzed the data using an "ensemble hypothesis" method,[18] suggesting a strategy whereby the active compounds could have been identified more efficiently by synthesizing only a portion of the library.

In another example, Geysen et al.[19] prepared 512 peptides representing all permutations of D and L variants of nine amino acids in substance P, resulting in the identification of key potency-determining residues. Young and Hawkins[20] subsequently reanalyzed these data using recursive partitioning methods to more systematically explore and understand the SAR. This study was greatly facilitated because the authors were part of the same industrial organization as the Geysen group and, therefore, had access to the same computerized resources.

Data-mining exercises such as this go on constantly within organizations. It is tantalizing to consider the greater number of advances that may take place if more data were more freely accessible, even for the limited number of cases which are released to the public domain.

The 1991 announcement of the Crystallographic Information File by the International Union of Crystallography[21] resonates with the current situation in combinatorial chemistry. Below is an excerpt from the introduction to that document.

"There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of local, national, and international networks, have fuelled the need for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited their effective use. This is true even in fields where the basic data requirements are well defined. Problems of data exchange are exacerbated if the number and nature of data types change rapidly and continuously. Under these conditions specialized and local file formats have proliferated. ... A general, flexible, rapidly extensible, and universal file format protocol is now essential. It must be machine-independent and portable so that accessibility to data items is independent of their point of origin. It must allow new data items to be incorporated without the need to modify existing files. In addition to archiving data, the use of a universal file would facilitate data exchange between software within a laboratory; between different laboratories; between authors and journals, providing electronic input to the publication process; and between researchers or journals and computerized databases."

Even earlier, the Joint Committee on Atomic and Molecular Physical Data (JCAMP) initiated the development of standard exchange formats for spectroscopic data, resulting in the publication of the JCAMP-DX file form for infrared spectra in 1988[22] and subsequent extension to a variety of other spectroscopic data.[23] All major instrument manufacturers now provide the ability to export or import files in this standard, machine-independent format. Other recent examples of initiatives to facilitate data exchange can be found in the areas of microarray data[24] and handling of biomedical images.[25]

The availability of standard data exchange formats is a necessary first step in the formation and utilization of generally accessible databases. Voluntary adoption of standards has often been achieved by requiring database submission prior to publication in relevant journals. This has proven very successful in the archiving of crystal structures, gene and protein sequences, and spectral data.[26]

Might we envisage a similar scenario for combinatorial libraries, with a common reporting standard and central archive for published data? There are a number of reasons why this may not happen. It may be argued, for instance, that combinatorial synthesis has not reached a level of maturity and consistency to allow the definition of standards able to stand the test of time. If so, we at least wish to instill a sense that there is a goal to be achieved. It is, however, our belief that de facto standards are emerging which encompass the majority of reported libraries, and capitalizing on this may not be such a daunting task. Second, the dominance of industrial organizations in the production of combinatorial libraries may tend to disfavor the free archiving of library structures. Already the contents of the large majority of libraries produced in the private sector are not disclosed for reasons of secrecy or perceived lack of scientific novelty. For those that are published, submission in a standardized format should not be burdensome from an intellectual property perspective. In addition, an increasing fraction of publications describing combinatorial libraries is appearing from the academic sector.[27] Finally, there are a number of practical and logistical concerns. Where would such a database reside? Who would accept responsibility for maintenance? How would it be funded?

Computer representation of chemical structure is an area of ongoing effort.[28] There are a number of ongoing challenges: development of comprehensive, robust, and unambiguous coding schemes relating name and structure;[29] development of systematic registries for the archival of chemical information;[30] and the development of methods for translating between chemical coding schemes. Each of these areas has impact on the future of data exchange in combinatorial chemistry. Clearly, a list of enumerated products may simply be specified in the appropriate, standard data format. For large libraries, particularly virtual libraries, a Markush-type storage strategy will be desirable. The very general Extensible Markup Language (XML) has emerged as a widely used standard for electronic information interchange that separates the content, syntax, semantics, and presentation of data. XML uses "ontologies" to define a machine-readable taxonomy of classes and relationships that form the subject-specific vocabulary of each technical discipline. CambridgeSoft has published CDXML, chemistry ontology with support for many of the unique challenges of combinatorial chemistry. CDXML is an ASCII, XML equivalent to their binary CDX (ChemDraw exchange) format.[31] CDXML includes the "named alternative group" container object which holds fragments that represent alternative substituents for a query. It also includes reaction objects and objects to lay out depictions on a grid, all of which are useful for combinatorial chemistry. Support for combinatorial chemistry is also under development for several other XML applications. Chemical Markup Language (CML), and its associated JUMBO Java classes, is a widely known and freely available chemistry XML application.[32] CMLQuery, a superset of CML, is being developed to support generic representations. Queries and Markush structures are similar, as they represent a set of compounds rather than a single one.[33] MDL is developing XML wrappers for the SDfile and RDfile that will accommodate all of MDL's structure representations, including generic representations.[34] Robin Hewitt, at Dupont Pharmaceuticals Research Labs, is writing a modular "fourth generation" programming language for combinatorial chemistry computation that uses an XML representation of molecule lists for the data stream.[35]

While these current systems support some of the challenges of combinatorial chemistry, none supports them all. A first layer is just to enable the specification of the library members, including designations for superatom identity, attachment points for building blocks or residues (or clipping protocols for reaction-based strategies). Computer languages, such as CHUCKLES and CHORTLES, are already available for specifying libraries from building blocks.[36,37] A second layer will allow the designation of the relationship between compounds, perhaps by encoding the pool notation described earlier. A third layer will provide mechanisms to rapidly search virtual libraries without requiring full enumeration. Several Markush search engines are available for searching the patent literature: Merged Markush for Derwent and Questel DARC,[38] MARPAT from CAS,[39] and the Derwent Fragmentation Index,[40] which has some similarity to Markush searching. The former two use input similar to typical literature substructure search. The latter uses a chemical coding scheme which can be looked up in a chart or generated with a program that has basic structure drawing capability. These, programs, however, do not return the actual assembled molecules that would match the query. More seriously, all of these tools give false positives, returning hits where the fragments are connected in the wrong topology. Significant additional work would be required to adapt these patent searching tools to accurate virtual library searching.

A further issue is the selection of the most appropriate Markush structure. As described earlier, for any given library, this is somewhat arbitrary, but for purposes of library comparison, it is important that methods exist to identify overlap and similarity between libraries. This has been addressed by Barnard and Downs[6c] and should be taken into account in the final designation of the representation.

It is important to note that the eventual solutions to these requirements will most likely be based on existing work which may already be in widespread use. Fostering of existing methods which may represent prototypic standards will be significantly favored over the invention of totally new systems.[41]

A significant new initiative within IUPAC is the Chemical Identifier project, which strives for the first time to allocate a single, unique (i.e., *canonical*) linear code to each compound: the IChI string (for "IUPAC chemical identifier").[30] The IChI project is being developed using a "layered" approach to allow enhancements to a basic format over time. Initially, simple chemical connectivity will be handled; extensions for stereochemistry, isotopes, and tautomers are planned for the near future. Adding functionality to IChI which will allow integration with combinatorial libraries would be desirable. Beyond simple enumeration of library members, further challenges will be to add the relevant layers of specification to more completely define libraries. Attention to other issues addressed in this document may facilitate this effort. Clearly, intelligent selection of generic structure, proper treatment of residues, and accurate notation of the relationships between compounds will facilitate data storage and integration with the chemical literature. In addition, inclusion of associated properties, such as analytical data and biological assay information, will be highly desirable where available. For large libraries, particularly virtual libraries, a Markush-type storage strategy will be desirable.[42] Equipping the IChI system to handle generic structures is already being considered for subsequent phases of that project.

Conclusion

The majority of new compounds are already being prepared by combinatorial methods. Despite the huge numbers of compounds which they may contain, combinatorial collections offer the opportunity for greatly compressed representation, analysis, and manipulation as a result of regularity in their composition. Study of such libraries offers the opportunity for systematic analysis of the properties of compounds on a significantly larger scale than has heretofore been possible. However, accessing the structure and composition of combinatorial libraries and associated data is at present extremely difficult, in large part because of the lack of standardized means of reporting and archiving this information. Many other fields faced with similar challenges have overcome them with a community effort to facilitate communication through standardization. Computational methods to this end are emerging in combinatorial chemistry and may build on current efforts to build a "universal" registry of compounds. It is to be hoped that our community will be able take advantage of these initiatives to accelerate progress in the field, allowing future generations of combinatorial scientists to more easily stand on the shoulders of, and thereby see further than, those who have gone before.

References and Notes

(1) *Glossary of Terms Used in Combinatorial Chemistry* (Technical Report); *Pure Appl. Chem.* **1999**, *71* (12), 2349−2365. Reprinted *J. Comb. Chem.* **2000**, *2* (6), 562−578. German translation *Angew. Chemie* **2002**, *114*, 893−906.

(2) IUPAC Recommendations are collected in electronic form at http://www.chem.qwm.ac.uk/iupac.

(3) http://www.iupac.org/projects/2002/2002-022-1-024.html.

(4) Hindson, K. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 663. http:///www.vchgroup.de/home/angewandte.

(5) From "*Fast Alert*" March 1998: "In any normal week, a specification of over 600 pages, such as SB's WO-09806734, would merit comment as an exceptionally long document. However, this is dwarfed by WO-09806720 from Eisai, where parts of the 1342-page specification take on the appearance of abstract wallpaper design, as the applicant recites in diagrammatic form all the fused, bridged, and spiro ring systems which may be used as substituents."

(6) Barnard, J. M.; Downs, G. M.; von Scholley-Pfab; Brown, R. D. *J. Mol. Graph. Model.* **2000**, *18*, 452−463. (b) Daylight MedChem User Group meeting 1997. See: http://www.daylight.com/meetings/mug97/Barnard/970227JB.html. (c) Barnard, J. M.; Downs, G. M. *Perspectives in Drug Discovery and Design* **1997**, 7/8, 18−30. (d) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 233−253.

(7) Linusson, A.; Gottfries, J.; Olsson, T.; Örnskov, E.; Folestad, S.; Norden, B.; Wold, S. *J. Med. Chem.* **2001**, *44*, 3424−3439.

(8) Martin, E. J.; Hoeffel, T. J. *J. Mol. Graph. Model*. **2000**, *18*, 383−403.

(9) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.

(10) Cramer, R. D. *J. Med. Chem.* **2003**, *46*, 374−388.

(11) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. *J. Comb. Chem.* **2000**, *2*, 2−5.

(12) Luo, Z.; Zhang, Q.; Oderaotoshi, Y.; Curran, D. P. *Science* **2001**, *291*, 1766−1769.

(13) Haag, R.; Sunder, A.; Hebel, A.; Roller, S. *J. Comb. Chem.* **2002**, *4*, 112−119.

(14) "A foolish consistency is the hobgoblin of little minds" Ralph Waldo Emerson

(15) Comely, A. C.; Gibson, S. E. *Angew. Chem., Int. Ed.* **2001**, *40*, 1012−1032.

(16) Pinilla, C.; Appel, J.; Blondelle, S. E.; Dooley, C. T.; Dorner, B.; Eichler, J.; Ostresh, J. M.; Houghten, R. A. *Biopolymers* **1995**, *37*, 221−240.

(17) Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; et al. *J. Med. Chem*. **1994**, *37*, 2678−2685.

(18) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. *J. Med. Chem*. **2000**, *43*, 2770−2774.

(19) Wang, J. X.; DiPasquale, A. J.; Bray, A. M.; Maeji, N. J.; Geysen, H. M. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 451−456.

(20) Young, S. S. Hawkins, D. M. *J. Med. Chem.* **1995**, *38*, 2784−2788.

(21) Hall, S. R.; Allen, F. H.; Brown, I. D. *Acta Crystallogr.* **1991**, *A47*, 655−685.

(22) McDonald, R. S.; Wilks, P. A. Jr. *Appl. Spectrosc.* **1988**, *41*, 151−162.

(23) http://www.jcamp.org

(24) Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C. P.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. − *Nat. Genet.* **2001**, *29*, 365−371; *Nature* **2002**, *419*, 323.

(25) Beautiful Bioimages for the Eyes of Many Beholders. *Science* **2002**, *297*, 39.

(26) For example, see the NIST Chemistry Webbook at: http://webbook.nist.gov/chemistry/

(27) Dolle, R. E. Comprehensive Survey of Combinatorial Library Synthesis: 1999. *J. Comb. Chem.* **2000**, *2*, *383* −433.

(28) For example, see the collected reports of the 4th International Chemical Structures Conference in *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 7. See also the report of the IUPAC Strategy Round Table: Representation of Molecular Structure − Nomenclature and Its Alternative; Washington, DC, 2000. http://www.iupac.org/news/archives/2000/NRT_Report.html

(29) Brecher, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 943−950.

(30) IUPAC Chemical Identifier effort http://www.iupac.org/projects/2000/2000-025-1-800.html

(31) http://sdk.cambridgesoft.com/chemdraw/cdx/index.html

(32) http://www.xml-cml.org/

(33) Murray-Rust, P. Coauthor of CML; personal communication.

(34) Tayler, K. Personal communication.

(35) http://www.daylight.com/meetings/mug01/Hewitt/

(36) Siani, M. A.; Weininger, D.; Blaney, J. M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588−93.

(37) Siani, M. A.; Weininger, D.; James, C. A.; Blaney, J. M. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1026−33.

(38) Derwent Information, London, U.K.: http://www.derwent.com/markush/; and Institute National de la Propriete Industrielle, Paris, France: http://www.inpi.fr/inpi/mms/

(39) MARPAT: Chemical Abstracts Service, Columbus, Ohio: http://www.cas.org/ONLINE/DBSS/marpat.pdf

(40) Derwent Information, London, U.K.: http://www.derwent-.com/chemistry/indexing.html

(41) A description of the most commonly used computer representations of combinatorial libraries is provided by Barnard and Downs in ref 6c.

(42) It should be noted that the registration of virtual compounds is controversial. See: Erhardt, P. *Chem. Int.* **2002**, *24*, 16.

CC0340325