Pergamon

Bioorganic & Medicinal Chemistry, Vol. 5, No. 3, pp 473–491, 1997

PII: S0968-0896(96)00277-5
0968-0896/97 $17.00 + 0.00

Review Article

Combinatorial Chemistry in Drug Research from a New Vantage Point

Hubert Maehr*
Roche Research Center, Hoffmann–La Roche Inc., Nutley, NJ 07110, U.S.A.

Contents

Combinatorial synthesis can be conducted in two distinctively different ways; the product can be generated either by parallel synthesis so that its components are available individually or it may be prepared as a mixture of compounds. Maintenance of spacial segregation of the individual reaction products provides the particular advantage of their direct comparability, but limits the size of the number of components for practical reasons. If mixtures are chosen, this number may be much larger; lead identification, however, is then no longer a direct process.

The lead compounds derived from large molecular libraries (e.g., those produced by recombinant techniques) are generally identified after receptor–affinity selection, followed by direct or indirect sequence analyses. Some libraries are constructed with concurrently introduced identifying groups such as peptides, oligonucleotides, or halogenated aromatic compounds. One of the recently explored possibilities to secure product identification involves solid-phase synthesis in microreactors containing imbedded memory devices, addressable via radiofrequency transmission to an external receiver. Alternate provisions for lead identification, the so-called deconvolution strategies, must be built into the synthetic methodology.

*Facsimile: (201) 235-2663, e-mail: hubert.maehr@roche.com

While recombinant peptide libraries are expected to continue to play a crucial role in exploiting potentially new drug targets resulting from human genomic research, there is presently a trend toward smaller libraries, produced on a milligram scale, often as mixtures, and targeting compounds with molecular weights usually not exceeding 500 dalton. The building blocks used in their synthesis deliberately exclude encoded amino acids.

Accommodating this trend, the present account reviews combinatorial technology with the aid of set theory. Consequently, concise descriptions of both synthesis and lead-identification strategies are possible. The focus centers on molecular libraries prepared as mixture whose components are devoid of solid supports or identifying tags.

1. Introduction

The methods for drug discovery have changed drastically during the past years. To appreciate and visualize this change, including the increasing popularity of combinatorial chemistry, drug research and development should be understood as a value-added process. The intrinsic premarketing drug value (IDV) of an entity can then be viewed as a function of a set of essential steps, $s = \{s_1, s_2, \ldots, s_n\}$, that reflects the drug research and development phase as shown in abbreviated form in Figure 1. For a chemical compound to acquire the status of a marketable drug, its IDV must be unity, an event that commands a successful outcome at every challenge point s_i. Since $s_i \in \{s\}$, any s_i provides an essential contribution that must be positive and is expressed as $s_i \neq 0$. If an event $s_i = 0$ occurs, the IDV reverts to zero. The research and development phase is therefore afflicted with an inherent propensity for failure. Moreover, the price of failure increases as the event $s_i = 0$ moves closer to the final event s_n.

The search for a marketable drug should therefore commence with a large number of chemical entities so that the probability to find a substance that satisfies the condition $s_1 \neq 0$ is increased. Once such lead structures have been identified, a success rate at the challenge points must be sustained and accomplished within a predetermined time frame. As a result, rapid, or even partially concurrent, onslaughts at the challenge points is essential. The intricacy of the research and development phase is underlined by the observation that of 5000 compounds that satisfy the condition $s_1 \neq 0$ only five enter clinical trials, and of those five only one will achieve a IDV = 1.[1] These statistical facts explain the frequent exploitation of reduced-risk scenarios. One example is the acquisition of an advanced drug lead from an external source. Several members of the set {s} are then already certified to conform with the condition $s_i \neq 0$ and an event characterized as $s_i = 0$ has not yet been encountered. Another one entails the development and production of the enantiopure form of an existing racemic drug, often referred to as a 'racemic switch'.[2] Most common, however, is the explosion of research activity at the periphery of a successful or promising competitor's drug in the hope to find an improved drug version and perhaps even a patent loophole.

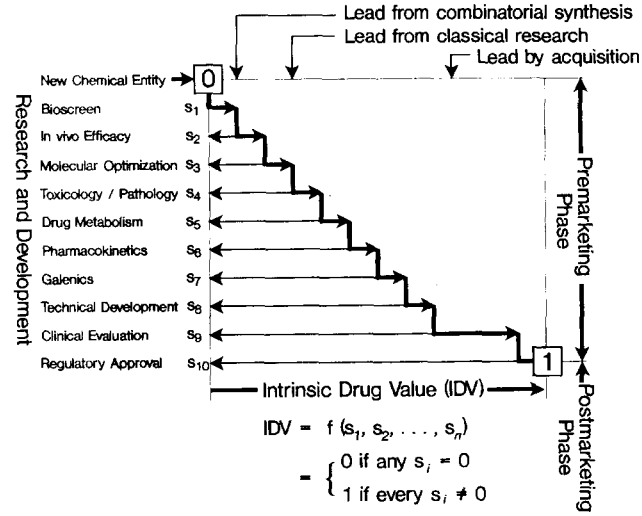

Figure 1. Premarketing drug value as a function of pharmaceutical research and development activities.

Classical drug research usually starts with novel natural products or derivatives thereof, or with new chemical entities synthesized as well-characterized compounds, often preceded by elaborate planning stages. These candidates are frequently tested directly in in vivo models. Surely, a favorable test result for such a compound places it relatively high on the IDV scale. Since the entity exhibited a favorable response in an animal model, it must have reached efficacious blood levels, it is obviously devoid of acute toxicity at the level tested, and molecular optimization may be less important in view of the individual attention the compound has received.

Modern drug research is based on rational drug design. The deep insights into biological processes underlying individual disease states that have been gained recently, enables the search for ligands that bind to a ligate such as a receptor, an antibody, enzyme, transcription factor, growth factor, or any other consequential host molecule. The ensuing test methods have evolved into automated high-throughput screens capable of analysing a huge number of compounds in a short time and generating an unprecedented number of compounds that satisfy the condition $s_1 \neq 0$. Traditionally, a medicinal chemist produced an average of ca. 30 new and fully characterized compounds per year. At such a rate of sample generation it is obviously no longer possible to provide enough testing material commensurate with the capacity that modern screening facilities provide. In direct response to this demand, large sample quantities, often mixtures of related compounds, are generated either by some biological system or by reiterative synthetic processes carried out in the chemical laboratory furnishing 'combinatorial libraries'. Regardless of their geneses, leads from these sources

will typically enter the IDV scale at a low level: The successful translation of efficacy in a high-flux bioscreen to a desired drug action in vivo is highly problematical and uncertain. Moreover, the chemical constitution of a library's member is usually unrefined.

Combinatorial technology[3–9] serves not only the discovery, but also the chemical optimization of lead substances and the corresponding efforts differ. While structural requirements are deliberately vague in the search for new bioactive compounds, encouraging the use of sizable libraries of elements with widely different properties, subsequent drug development is based on the substantial features of a lead structure that imposes constraints on the library's components that need to be accessible individually. In either case, the transition from a lead's chemical structure to that of the final drug is accelerated if the difference between them is as narrow as possible, or, expressed differently, if the need for lead optimization is minimal. The awareness that research at the s_1 level is of low risk and that the original lead structure should be closely related to that of the final drug, fosters new attitudes. Libraries of peptides, oligonucleotides, and other linear biopolymers are being de-emphasized. Rather, library construction, even for drug discovery purposes, may commence with specially selected building blocks, often not commercially available, with the goal to enhance an envisioned recognition center or to probe for pharmacophoric activity in a broader sense. The use of privileged elements, frequently reflecting structural features of natural products or established drugs, are either incorporated as preformed units by the building block concept, or generated as part of the library construction.[10,11] These experiments are combined, with increasing frequency, with forays into new bond-formation technologies in solid phase,[12–17] adaption of established tools to combinatorial chemistry in liquid phase,[18,19] or on soluble polymeric supports,[20] and deployment of highly automated and robotic synthesizers.

2. Definitions and Nomenclature

The term *combinatorial* chemistry has become popular for two reasons. First, it emphasizes the use of collections of countable numbers of discrete chemicals or 'building blocks', which are used in synthesis as sets and where the order in the set does not matter. Secondly, the term implies the *combinatorial* (or Cartesian) *product* of these sets to form a new set, commonly referred to as the combinatorial library.

The mathematical aspects of combinatorial chemistry can therefore be treated simply by the fundamental principles of counting. Assume that a molecule comprises six molecular subunits to form the sequence X1–A–X2–B–X3–C. Suppose three of these units are available in sets (e.g., $A = \{A_1, A_2, \ldots, A_a\}$, $B = \{B_1, B_2, \ldots, B_b\}$, and $C = \{C_1, C_2, \ldots, C_c\}$) so that A, B, and C are represented in a, b, and c different ways, respectively. The term (X1–aA–X3–bB–X5–cC) then portrays all molecules that can be synthesized from the available sets, and the resulting collection, the combinatorial library, contains *abc* elements.

'Library' is a finite set of different but related synthetic products called elements, members, or components.[21] 'N' is a reagent group, that is, a finite set of n different reagents or molecular building blocks but belonging to a specific reactivity class (e.g., amines, carboxylic acids, etc., so that $N = \{N_1, N_2, \ldots, N_n\}$).

Experimental chemistry requires that two types of set N are differentiated. In one, the elements are present as a mixture, in the other they are separate from each other, they are present in individual containers, as it were. Mathematically, this differentiation can be simulated by considering N as a set and as a partition of that set, respectively, where each element occupies one cell in the partition. Thus, $({}^{n}N) = \{N_1, N_2, \ldots, N_n\}$, is a finite set of n components, present as a mixture and in equimolar quantities. The internal superscript in $({}^{n}N)$ denotes that the elements are present as a mixture. Such blends are occasionally referred to as cocktails. An individual member of a set, whose identity is established, is distinguished by a subscript, thus $N_k \in ({}^{n}N)$ (i.e., N_k is a known element in the set N, and $\Phi \subset N \subset \Omega$, where Φ is the null set and Ω is the universal set that contains all possible molecular building blocks). Further, $p({}^{n}N)=p$ equal parts of the mixture $({}^{n}N)$.

We now consider the same set, where the individual components are present separate from each other, as a partition of set N containing the maximum number of cells; the elements are mutually exclusive and their union is N: ${}^{n}(N) = \{(N_1), (N_2), (N_3), \ldots, (N_n)\}$. The physical separation of the elements in the set is emphasized by the superscript outside of the parentheses; this external superscript again reveals the number of elements in the set. It is also implied that all elements are present in equimolar amounts.

ABC is a molecule where the molecular moieties A, B, and C are linked in the sequence A–B–C.

$(A_k\ {}^{b}B) = \{A_k B_1, A_k B_2, \ldots, A_k B_b\}$ (i.e., a set of molecules of the type AB, where A_k is known and invariant but $B = \{B_1, B_2, \ldots, B_b\}$). All elements are present as a mixture as suggested by the internal superscript at B. It follows that ${}^{b}(A_k B) = \{A_k B_1, A_k B_2, \ldots, A_k B_b\}$; it is a partition of the set $(A_k\ {}^{b}B)$; the number of cells in this partition is revealed by the external superscript b.

$({}^{a}A\ {}^{b}B)$ = a set (collection, library) of molecules of the type AB, where all members are present as a mixture; each A component is covalently linked to a B component, observing that $A = \{A_1, A_2, \ldots, A_a\}$ and $B = \{B_1, B_2, \ldots, B_b\}$. The superscript attached to each variable molecular moiety reveals how many variants of it are present within the set. The total number of elements in the set is the product of the internal super-

scripts. In other words, (^{a}A ^{b}B) represents the combinatorial (or Cartesian) product of the two sets A and B, which is the set of all ordered pairs of their elements, for example, (A_iB_j).

^{a}(A ^{b}B) = {(A_1B_1, A_1B_2, ... , A_1B_b), (A_2B_1, A_2B_2, ... , A_2B_b), ... , (A_aB_1, A_aB_2, ... , A_aB_b)}, that is, ^{a}(A ^{b}B) is a partition of the set (^{a}A ^{b}B) containing a cells where each cell is distinguished by a single invariant element, A_i, taken from the set ^{a}(A) so that the partition is appropriately referred to as the A-partition. This A-partition may also be expressed as ^{a}(A ^{b}B) = {(A_1 ^{b}B), (A_2 ^{b}B), ... , (A_a ^{b}B)}. Analogously, the B-partition of the same set takes the form: ^{b}(^{a}AB) = {(^{a}AB_1), (^{a}AB_2), ... , (^{a}AB_b)}; here B_j is invariant in a given cell, but A is present as a set. The B-partition contains b cells.

The definition of the combinatorial product of two sets as the collection of all ordered pairs extends trivially to ordered triples and z-tuples embracing multidimensional sample spaces. A library of the molecular type ABC ... Z, wherein each position is variable, can thus be represented by (^{a}A ^{b}B ^{c}C ... ^{z}Z). The total number of the set's members is given by the product of the superscripts associated with the individual, variable molecular moieties. The absence of an external superscript indicates that all members are present as a mixture.

Of particular interest in combinatorial chemistry is the arrangement of a set in the form of those partitions that are sorted with respect to the specific molecular moieties that were employed as reagent groups in the synthesis. For the general set (^{a}A ^{b}B ^{c}C ... ^{z}Z) these partitions are as follows:

^{a}(A ^{b}B ^{c}C ... ^{z}Z) = {(A_1 ^{b}B ^{c}C ... ^{z}Z), (A_2 ^{b}B ^{c}C ... ^{z}Z), ... , (A_a ^{b}B ^{c}C ... ^{z}Z)},

^{b}(^{a}A B ^{c}C ... ^{z}Z) = {(^{a}A B_1 ^{c}C ... ^{z}Z), (^{a}A B_2 ^{c}C ... ^{z}Z), ... , (^{a}A B_b ^{c}C ... ^{z}Z)},

^{c}(^{a}A ^{b}B C ... ^{z}Z) = {(^{a}A ^{b}B C_1 ... ^{z}Z), (^{a}A ^{b}B C_2 ... ^{z}Z), ... , (^{a}A ^{b}B C_c ... ^{z}Z)}, etc., and

^{z}(^{a}A ^{b}B ^{c}C ... Z) = {(^{a}A ^{b}B ^{c}C ... Z_1), (^{a}A ^{b}B ^{c}C ... Z_2), ... ,(^{a}A ^{b}B ^{c}C ... Z_z)}.

In the context of this paper, we refer to these specific element distributions simply as 'partitions'. More specifically, the first line represents a partition whose cells are ordered with respect to the elements of set A and is therefore referred to as the A-partition. The subsequent expressions denote the B-, C-, and Z-partition of the set. The number of cells in each partition is given by the corresponding superscript. Mixing the cells of any partition reconstitutes the parent set (^{a}A ^{b}B ^{c}C ... ^{z}Z). Similarly, ^{2,3}(X1 X2 ^{20}X3) delineates a partition of the set (^{2}X1 ^{3}X2 ^{20}X3) as six cells, each containing 20 compounds; these are defined in terms of X1 and X2, while the multiplicity of 20 is due to the diversity in X3: ^{2,3}(X1 X2 ^{20}X3) = {($X1_1$ $X2_1$ ^{20}X3), ($X1_1$ $X2_2$ ^{20}X3), ($X1_1$ $X2_3$ ^{20}X3), ($X1_2$ $X2_1$ ^{20}X3), ($X1_2$ $X2_2$ ^{20}X3), ($X1_2$ $X2_3$ ^{20}X3)}.

The two external superscripts correspond, respectively, to the first and second variable moieties of the molecular type X1–X2–X3. The term ^{2,3}(X1 X2 ^{20}X3) is thus referred to as the X1, X2-partition of the set (^{2}X1 ^{3}X2 ^{20}X3).

Partitions designated by numerical external superscripts could occasionally be ambiguous with respect to the correlation between superscript and the corresponding element. Internal notations associated with certain elements would then clarify the situation, so that the term ^{2,3}(ABC_kD_k ^{3}E) unambiguously defines the partition since C_k and D_k refer to individual elements and are thus invariant, but the term ^{2,3}(ABCD ^{3}E) is inadequate.

Branched molecules are treated in analogy to the conventional chemical symbolism, which would denote 2-phenylbutane as $(C_6H_5)C(CH_3)(C_2H_5)$. Branches are defined by brackets, however. Thus, X[A][B] = a molecule of the type A–X–B where X is derived from a bifunctional agent that binds covalently both A and B. The expression A[B][C][D] symbolizes a branched species where A is a trifunctional agent; it is the branching scaffold from which substituents B, C, and D emanate.[22]

It has already been pointed out that the description of a combinatorial synthetic step usually implies equimolarity of the reaction partners. The conversion (^{5}A) + B → (^{5}AB), therefore, implies a reaction of an equimolar mixture of the five members of set (^{5}A) with B, employing five molecular equivalents of B with respect to A_i, in the sense of (^{5}A) + 5B → (^{5}AB). The reaction symbolism remains obviously unchanged if (^{5}A) is present in solid phase and is allowed to react with an excess of B to enforce reaction completion.

3. Synthetic Processes

The practical synthesis of organic-chemical combinatorial libraries requires the concurrent completion of many similar chemical reactions, for example, ^{a}(A) + ^{b}(B) → ^{a,b}(AB). Libraries of more complicated members may require not just one, but sets of consecutively performed condensation steps employing the products of the previous set as the educt in the next set. A typical second coupling reaction could be formulated as ^{a,b}(AB) + ^{c}(C) → ^{a,b,c}(ABC). These condensations can be carried out either in liquid or solid phase. In one method, the individual reaction sequences (e.g., A_i + B_j → A_iB_j, and A_iB_j + C_k → $A_iB_jC_k$) can be conducted in segregated spaces, such as multiple reactors,[23] pins,[24,10] beads,[25,26] microreactors,[13,27,28] tea bags,[29] and on macroscopic[23,30,31] and microscopic[32] surfaces. This technique is generally referred to as parallel synthesis. The term ^{a,b,c}(ABC) reveals the individuality of the set's members and attests to the use of this synthetic method.

Some of the above mentioned spaces in which the chemical reactions are performed, such as multiple reactors and tea bags, can also serve for the practice of the other method, where the individual *sets* of these reactions are performed each at *one* reaction site, leading to mixtures of reaction products, for example, $(^{a}A) + B_k \rightarrow (^{a}AB_k)$.

The concise descriptions of both the chemical events leading to combinatorial libraries and the subsequent lead-identification programs are of profound importance. Employing the definitions outlined in section 2, a terminology is now feasible that allows the description of all combinatorial chemical processes commensurate with the present state of the art.

To pave the way for further discussions, it is, however, necessary to categorize the synthetic operations. In a rudimentary step, pure reactant A reacts with pure B to yield AB. Since neither A nor B are sets of compounds (no mixtures are employed), the process is called an M∅ operation. To render the description more general, the term M∅ is further modified to reflect the number of components that take part in the reaction and are represented in the product. Thus, A + B → AB is a two-component condensation, abbreviated as 2C. The overall process is described as M∅-2C. If one of the reaction partners in a given synthesis is employed as a mixture, for example, $(^{a}A) + B \rightarrow (^{a}AB)$, the step is of the M1-2C category. Analogously, a conversion involving both educt and reactant as mixtures belongs to the M2-2C type.[33]

The conversion $(^{a}AB) + (^{c}C) \rightarrow (^{a}AB\ ^{c}C)$ is a special form of an M2-2C process. The reaction loci in the members of the set (^{a}AB) are identical and are only modified in their reactivity by the variant, but distant, A moieties. To highlight this special feature, we designate such a process as M2*-2C. Clearly, an M2*-2C process may be more manageable in terms of reaction uniformity than one of the M2-2C category as it resembles an M1-2C step of the form AB + (^{c}C) → (AB ^{c}C). The characterization of the steps $(^{a}AB) + {}^{c}(C) \rightarrow {}^{c}(^{a}ABC)$ as *c* M1*-2C operations is self-explanatory.

3.1. M2-2C Syntheses

Multicomponent mixtures can be prepared, in the most simple case, by the reaction of a mixture of educts with a reagent group as a mixture, such as $(^{a}A) + (^{b}B) \rightarrow (^{a}A\ ^{b}B)$. Let us assume that the synthesis is initiated by the reaction of one equivalent of a monofunctional starter unit E (e.g., a functionalized resin) with a mixture of 20 different, monoprotected amino acids, 1:20 equivalent of each, and the reaction is allowed to go to completion, $E + (^{20}X) \rightarrow E(^{20}X)$. Deprotection and repetition of the coupling step with the same mixture would now produce 400 new products, $E(^{20}X) + (^{20}X) \rightarrow E(^{20}X\ ^{20}X)$, in one step and in one-pot, and after four further repetitions a mixture of 20^6 hexapeptides, $E(^{20}X\ ^{20}X\ ^{20}X\ ^{20}X\ ^{20}X)$, would result, provided that all coupling reactions proceeded to some extent. If the educt E were bifunctional and not symmetric, allowing bidirectional and simultaneous chain extensions, a maximum of $20^3 \times 20^3 = 20^6$ peptide products could be obtained already after three steps, $E[^{20}X\ ^{20}X\ ^{20}X][^{20}X\ ^{20}X\ ^{20}X]$. A hypothetical educt with six equally reactive and non-equivalent functional groups could yield the same number of products, $E[^{20}X][^{20}X][^{20}X][^{20}X][^{20}X][^{20}X]$, after a single condensation step. Thus, the total number of products P, theoretically produced is a function of the number of sequentially performed coupling steps, s, the number of components, n, employed in each condensation step, and the number of reactive sites (nonsymmetric branches), b, in the educt. If n is constant during the entire synthetic sequence, then $P = n^{bs}$. The effect of b and s on P, where P is expressed as a multiple of n, is illustrated in Figure 2.

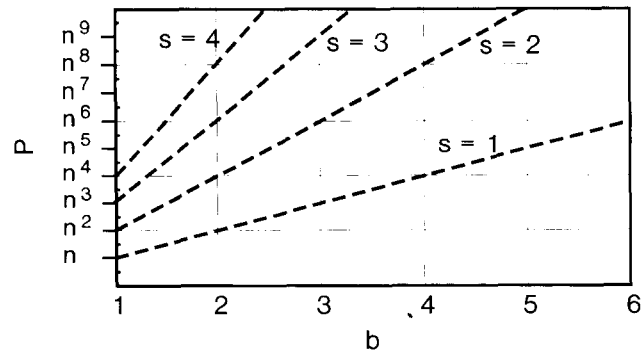

Figure 2. Theoretical number of products (P) derived from typical combinatorial processes.

Although leading to a higher degree of molecular diversity, in a single step, than an M1-2C process, the disadvantages of M2-2C syntheses in combinatorial assemblages are obvious: Each macroscopic condensation step involves many substrates and reagents; the concurrence of *ab* different rate constants for the condensation $(^{a}A) + (^{b}B) \rightarrow (^{a}A\ ^{b}B)$ underlines the expected difficulty to achieve complete reactions. Further, an M2-2C process, in its simplest form produces a mixture of the type $(^{a}A\ ^{b}B)$ without the intermediacy of the corresponding partitions $^{a}(A\ ^{b}B)$ or $^{b}(^{a}AB)$, so that identification of the active substance cannot proceed via their evaluation. Nonetheless, the 'mixed amino acid method' has been used in the synthesis of peptide libraries (section 3.2.3), where the mixture of carboxyl-activated amino acids are no longer present in equimolar ratios; rather, the concentrations are now adjusted to reflect their respective reactivities (i.e., the concentration of an educt is inversely proportional to the expected reaction rate constant).[34] M2-2C steps are readily identified by the involvement of reaction partners, or reaction sites, which are distinguished by internal superscripts, for example, $(^{a}A) + (^{b}B) \rightarrow (^{a}A\ ^{b}B)$, or $(A_k\ ^{b}B) + (^{c}C) \rightarrow (A_k\ ^{b}B\ ^{c}C)$.

3.2. M1-2C Syntheses

This synthetic category refers to the condensation of a mixture of components with one particular reactant as expressed by $({}^{a}A) + B \rightarrow ({}^{a}AB)$, or $A + ({}^{b}B) \rightarrow (A\ {}^{b}B)$. Although apparently similar, there are profound differences between these two processes both in terms of practicality and product constitution. One of these schemes may be preferred over the other. As an example, it may be easier to condense a mixture of amines with a particular activated ester than to prepare a mixture of activated esters and to allow the mixture to react with a single amine. If the element B bears a latent functional group for further reaction, then a comparison between $({}^{a}AB)$ and $(A\ {}^{b}B)$ reveals yet another important difference: $({}^{a}AB)$ may be employed in M1*-2C and M2*-2C processes, whereas $(A\ {}^{b}B)$ is restricted to reactions in the sense of M1-2C and M2-2C operations.

The successful identification of the bioactive component in an organic-synthetic combinatorial library is usually provided by a structured synthetic protocol that furnishes at least one partition of the set. The synthesis of the molecular library must also proceed with an absolute minimum of side-reaction; any unknown product that gives a positive response in the screening phase can cause serious complications. A simple, but ingenious solution to achieve the result of an M2-2C process with maximum reaction homogeneity, but employing an M1-2C protocol, has been advanced by Furka et al.[35,36] and was termed PMM (proportioning-mixing method). The method is also known as 'divide, couple, and recombine'[37] and 'split synthesis'.[38] An application of the method is shown below in section 3.2.1.1; it is applicable to solid and liquid-phase syntheses alike, but offers particular advantages for the solid-phase technique; there is the facility of easy portioning, and further, a set of reactants linked to a solid support can be deprotected as a mixture and subjected repeatedly to condensation conditions until all resin-bound reactant sites have reacted.

3.2.1. Synthesis of the last partition of a set.

3.2.1.1. Split synthesis on solid support. The synthesis of the library $({}^{5}A\ {}^{5}B\ {}^{5}C)$ serves as an example and is illustrated in Figure 3. Element assembly involves the addition of B to A and C to AB. The molecular moieties A, B, and C are each present in sets of five. Construction commences by attaching the members of the first reagent set, $^{5}(A)$, to the solid support, R. This is accomplished by dividing the solid support into five equal portions and to couple each with the individual members of the set $^{5}(A)$ until the reactions are

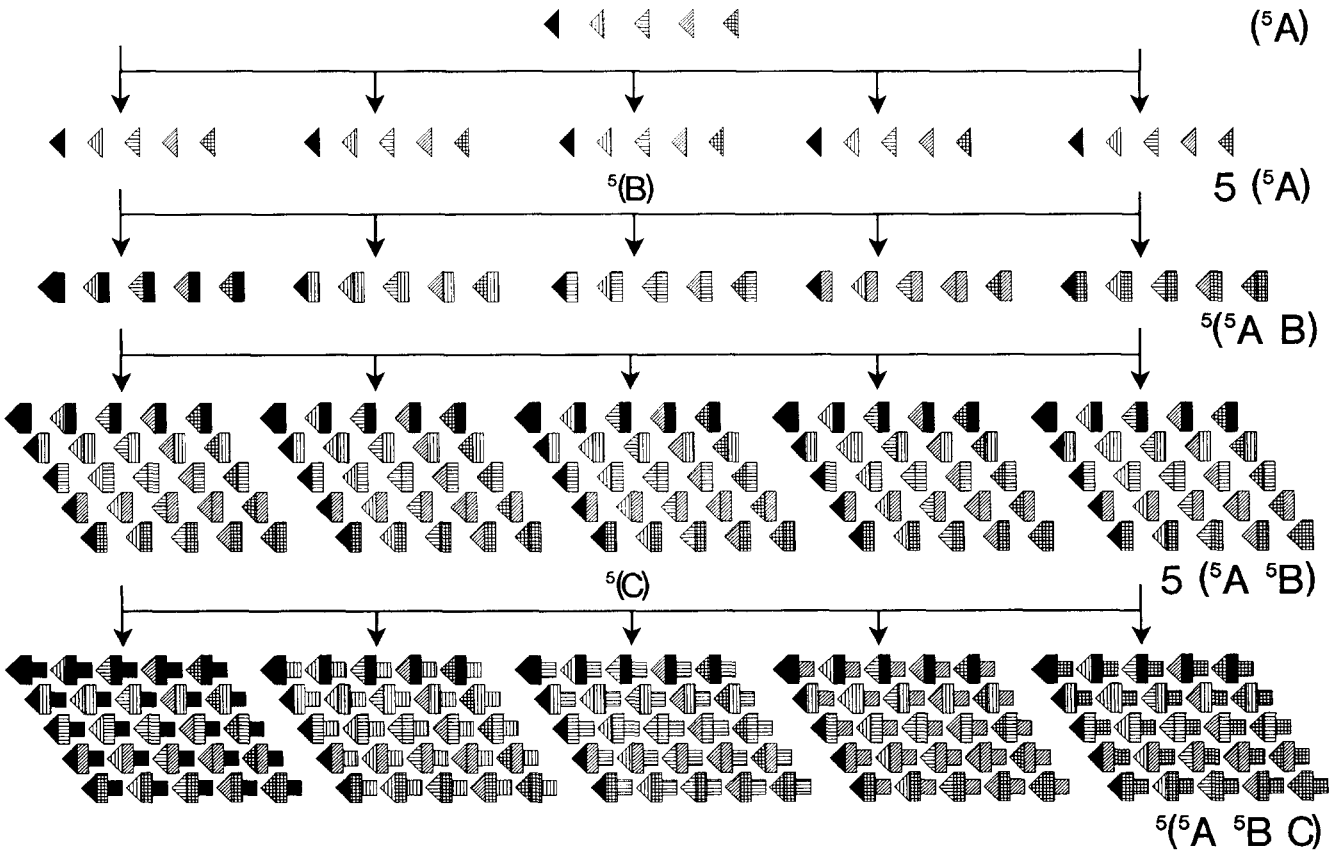

Figure 3. Synthesis diagram of the C-partition of the set $({}^{5}A\ {}^{5}B\ {}^{5}C)$. The building blocks of type A are symbolized by triangles, B by rectangles, and C by squares, the differences within each category are emphasized by different shades of the symbols. The solid support component is not shown.

complete: 5R + ^{5}(A) → R ^{5}(A). The five resin batches are combined, mixed, and redistributed into five equal parts: R ^{5}(A) → 5 R(^{5}A). The latent functionalities in the A moieties are then activated, either at this time point or prior to the redistribution, and each portion is coupled with a different member of the second set ^{5}(B): 5 R(^{5}A) + ^{5}(B) → R ^{5}(^{5}A B).

As apparent, the resin-bound moieties that participate in this coupling step differ from each other, as they are derived from the set (^{5}A), but only one particular component B reacts with each R(^{5}A) aliquot. The condensation with the individual components of the set ^{5}(B) can therefore be repeated, if required. The contents of the five reaction vessels are combined, mixed and redistributed into five equal parts: R ^{5}(^{5}A B) → 5 R(^{5}A ^{5}B). The deprotection and condensation step with the individual members of the new set ^{5}(C) proceeds according to the expression 5 R(^{5}A ^{5}B) + ^{5}(C) → R ^{5}(^{5}A ^{5}BC). Cleavage of the individual reaction mixtures from the solid support affords the last, or C-partition of the set, ^{5}(^{5}A ^{5}BC) and blending of the five cells yields the set (^{5}A ^{5}B ^{5}C).

In summary, the library (^{5}A ^{5}B ^{5}C) is assembled by two M1-2C protocols, (^{5}A) + ^{5}(B) → ^{5}(^{5}AB), and (^{5}A ^{5}B) + ^{5}(C) → ^{5}(^{5}A ^{5}BC), involving a total of 2 × 5 condensation reactions. Assuming ideal conditions for all reactions, the mixture (^{5}A ^{5}B ^{5}C) could be prepared by two M2-2C processes involving only two consecutive coupling steps, virtually in a one-pot process, as follows: (^{5}A) + (^{5}B) → (^{5}A ^{5}B); (^{5}A ^{5}B) + (^{5}C) → (^{5}A ^{5}B ^{5}C).

3.2.1.2. Split synthesis in liquid phase. While reaction completion can be enforced on solid supports by repetition of the individual coupling steps, for example, R(^{a}A ^{b}B) + C_j → R(^{a}A ^{b}BC_j), to arrive eventually at the conversion R(^{a}A ^{b}B) + ^{c}(C) → R ^{c}(^{a}A ^{b}BC), there is no practical equivalent for reactions in solution. M1-2C processes in liquid phase are only successful under prevalence of favorable condensation conditions for all set members that take part in a library construction and must therefore be preceded by meticulous process development.[39]

Depending upon the actual chemical steps involved, the schematic plan for the liquid-phase synthesis of the last partition may be as described in section 3.2.1.1 or may be a modified version thereof: Without the need to link one reagent group to a solid support, the construction of the set (^{5}A ^{5}B ^{5}C) can also commence by pooling the elements of the reagent set ^{5}(B) and redistribution of the mixture into five equal parts. Each part is then allowed to react with a specific member of the set ^{5}(A): ^{5}(A) + (^{5}B) → ^{5}(A ^{5}B). The five reaction mixtures are combined and their latent functional groups activated, ^{5}(A ^{5}B) → (^{5}A ^{5}B); the bulk is redistributed into five equal parts and each part is then coupled with a specific member of the set ^{5}(C): 5 (^{5}A ^{5}B) + ^{5}(C) → ^{5}(^{5}A ^{5}BC).

3.2.2. Synthesis of the penultimate partition. The penultimate partition of the example set is ^{5}(^{5}AB ^{5}C) and its synthesis is shown in Figure 4. As revealed by the notation, the synthesis requires the intermediacy of the partition ^{5}(^{5}AB) of the set (^{5}A ^{5}B). Two reasonable avenues, adaptable to both liquid- and solid-phase methods, lead to the goal. The first one proceeds by the condensation ^{5}(^{5}AB) + (^{5}C) → ^{5}(^{5}AB ^{5}C) as shown in Figure 4. It should be noted that this step is of the M2*-2C type; each cell of the set ^{5}(^{5}AB) comprises members with a unique moiety B_j that undergoes bond formation. Clearly, repetitions of the condensation step with (^{5}C) to enforce reaction completion is not acceptable. The advantage of enhanced product uniformity in the

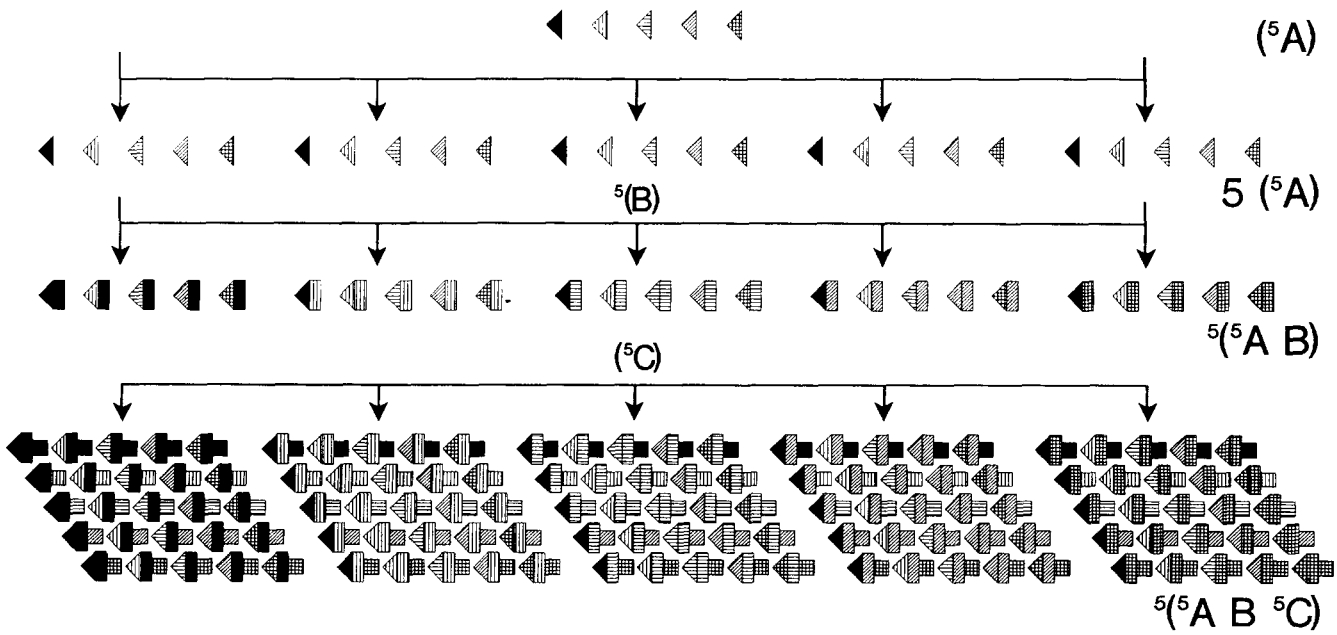

Figure 4. Synthesis diagram of the B-partition of the set (^{5}A ^{5}B ^{5}C). For symbol legend see Figure 3.

second one, preferably conducted on solid support, is compromised by the need for robotics, especially for the construction of larger libraries. The process commences again with the partition R ^{5}(^{5}AB) whose cell contents are condensed individually with the elements of the reagent group C: R ^{5}(^{5}AB) + ^{5}(C) → R ^{5,5}(^{5}ABC). The reactions involved are of the M1*-2C type and the resulting products represent the B,C-partition of the set R(^{5}A ^{5}B ^{5}C):

R ^{5,5}(^{5}ABC) =

{R(^{5}AB$_1$C$_1$), R(^{5}AB$_1$C$_2$), R(^{5}AB$_1$C$_3$), R(^{5}AB$_1$C$_4$), R(^{5}AB$_1$C$_5$);
R(^{5}AB$_2$C$_1$), R(^{5}AB$_2$C$_2$), R(^{5}AB$_2$C$_3$), R(^{5}AB$_2$C$_4$), R(^{5}AB$_2$C$_5$);
R(^{5}AB$_3$C$_1$), R(^{5}AB$_3$C$_2$), R(^{5}AB$_3$C$_3$), R(^{5}AB$_3$C$_4$), R(^{5}AB$_3$C$_5$);
R(^{5}AB$_4$C$_1$), R(^{5}AB$_4$C$_2$), R(^{5}AB$_4$C$_3$), R(^{5}AB$_4$C$_4$), R(^{5}AB$_4$C$_5$);
R(^{5}AB$_5$C$_1$), R(^{5}AB$_5$C$_2$), R(^{5}AB$_5$C$_3$), R(^{5}AB$_5$C$_4$), R(^{5}AB$_5$C$_5$)}.

The rows of this table can be summarized as ΣR(^{5}A$\bullet$B$_j$ ^{5}C) = R ^{5}(^{5}AB ^{5}C), and the columns as ΣR(^{5}A ^{5}B C$_j$) = R ^{5}(^{5}A ^{5}BC) where B$_j$ and C$_j$ represent any specific element of the sets (^{5}B) and (^{5}C), respectively. In other words, the contents of the 25 vessels can be sorted and pooled with respect to either B or C, leading to the B- or C-partition, respectively. The process of sorting and pooling with respect to B implies the combining of those containers with variable C but invariant B elements, and corresponds to pooling along *rows*: R ^{5,5}(^{5}ABC) → {R(^{5}AB$_1$ ^{5}C), R(^{5}A B$_2$ ^{5}C), R(^{5}AB$_3$ ^{5}C), R(^{5}AB$_4$ ^{5}C), R(^{5}AB$_5$ ^{5}C)} = R ^{5}(^{5}AB ^{5}C), which is the B-partition of the set. Sorting and pooling with respect to C implies the combining of those containers with variable B but invariant C elements; that is, pooling along *columns* leads to the C-partition: R ^{5,5}(^{5}ABC) → {R(^{5}A ^{5}BC$_1$), R(^{5}A ^{5}BC$_2$), R(^{5}A ^{5}BC$_3$), R(^{5}A ^{5}BC$_4$), R(^{5}A ^{5}BC$_5$)} = R ^{5}(^{5}A ^{5}BC).

3.2.3. Synthesis of the other partitions. The strategies for syntheses in solid and liquid phases are alike. The synthesis of the A-partition of the set (^{5}A ^{5}B ^{5}C), which is ^{5}(A ^{5}B ^{5}C), requires ^{5}(A ^{5}B) as educt. Coupling of each cell of this partition with each member of the set ^{5}(C) furnishes 25 reaction mixtures in the form of the A,C-partition, ^{5}(A ^{5}B) + ^{5}(C) → ^{5,5}(A ^{5}BC) and is summarized below:

^{5,5}(A ^{5}BC) =

{(A$_1$ ^{5}BC$_1$), (A$_1$ ^{5}BC$_2$), (A$_1$ ^{5}BC$_3$), (A$_1$ ^{5}BC$_4$), (A$_1$ ^{5}BC$_5$);
(A$_2$ ^{5}BC$_1$), {A$_2$ ^{5}BC$_2$), (A$_2$ ^{5}BC$_3$), (A$_2$ ^{5}BC$_4$), (A$_2$ ^{5}BC$_5$);
(A$_3$ ^{5}BC$_1$), (A$_3$ ^{5}BC$_2$), (A$_3$ ^{5}BC$_3$), (A$_3$ ^{5}BC$_4$), (A$_3$ ^{5}BC$_5$);
(A$_4$ ^{5}BC$_1$), {A$_4$ ^{5}BC$_2$), (A$_4$ ^{5}BC$_3$), (A$_4$ ^{5}BC$_4$), (A$_4$ ^{5}BC$_5$);
(A$_5$ ^{5}BC$_1$), {A$_5$ ^{5}BC$_2$), (A$_5$ ^{5}BC$_3$), (A$_5$ ^{5}BC$_4$), (A$_5$ ^{5}BC$_5$)}.

The rows of this table can be expressed as Σ(A$_j$ ^{5}B ^{5}C) = ^{5}(A ^{5}B ^{5}C), and the columns as Σ(^{5}A ^{5}BC$_j$) = ^{5}(^{5}A ^{5}BC). Subsequent sorting and pooling with respect to A (pooling along rows) leads to ^{5}(A ^{5}B ^{5}C). Figure 5 shows the product distributions of the final and intermediate set partitions.

Employing the principles of the split synthesis renders the last partition of any set directly. The synthetic complexity for the production of other partitions of the set (^{a}A ^{b}B ^{c}C ... ^{z}Z) increases with increasing distance from the last, or Z-partition. The schematic protocols for the synthesis of all partitions of the set (^{a}A ^{b}B ^{c}C ^{d}D), inclusive of M2-2C processes (shown in dotted lines), are summarized in Figure 6.

If the use of the set (^{d}D) as the reagent in an M2*-2C process is chemically feasible, the C-partition synthesis involves only *c* reactions in addition to those that produced the intermediate ^{c}(^{a}A ^{b}BC). Should it be required that each cell of the intermediate partition ^{c}(^{a}A ^{b}BC) reacts individually with the members of the set ^{d}(D), then *cd* additional reaction steps are necessary. The B-partition, ^{b}(^{a}AB ^{c}C), can be obtained either in *b* steps using (^{c}C) or in *bc* steps utilizing the individual

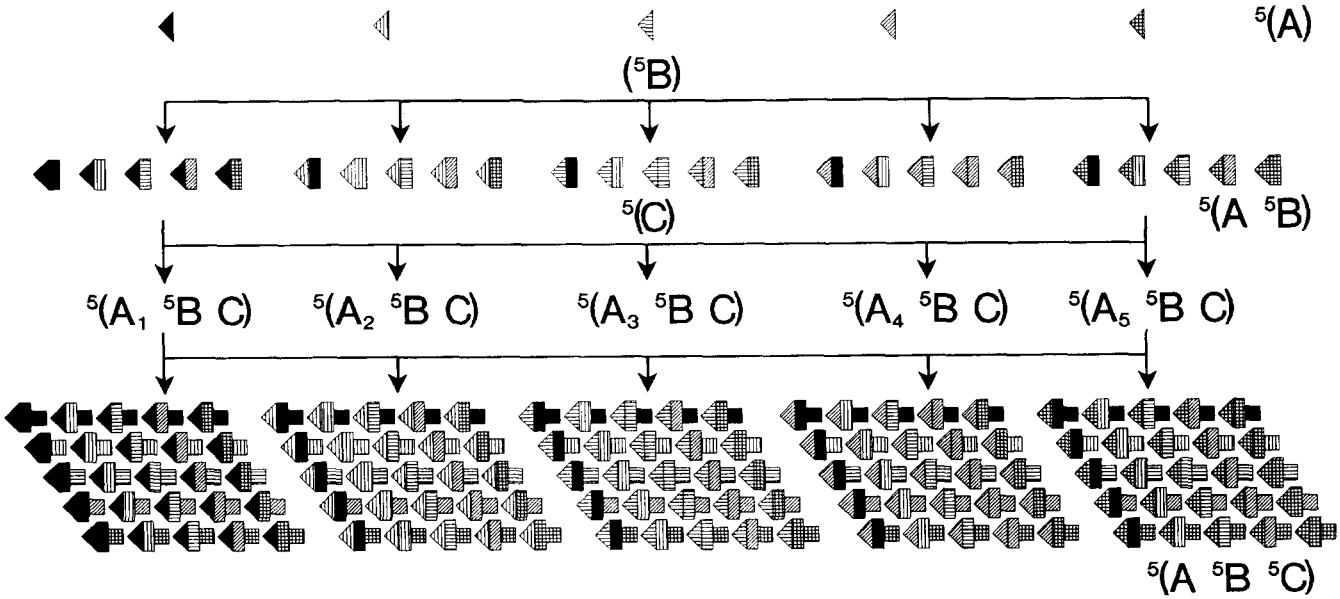

Figure 5. Synthesis diagram of the A-partition of the set (^{5}A ^{5}B ^{5}C). For symbol legend see Figure 3.

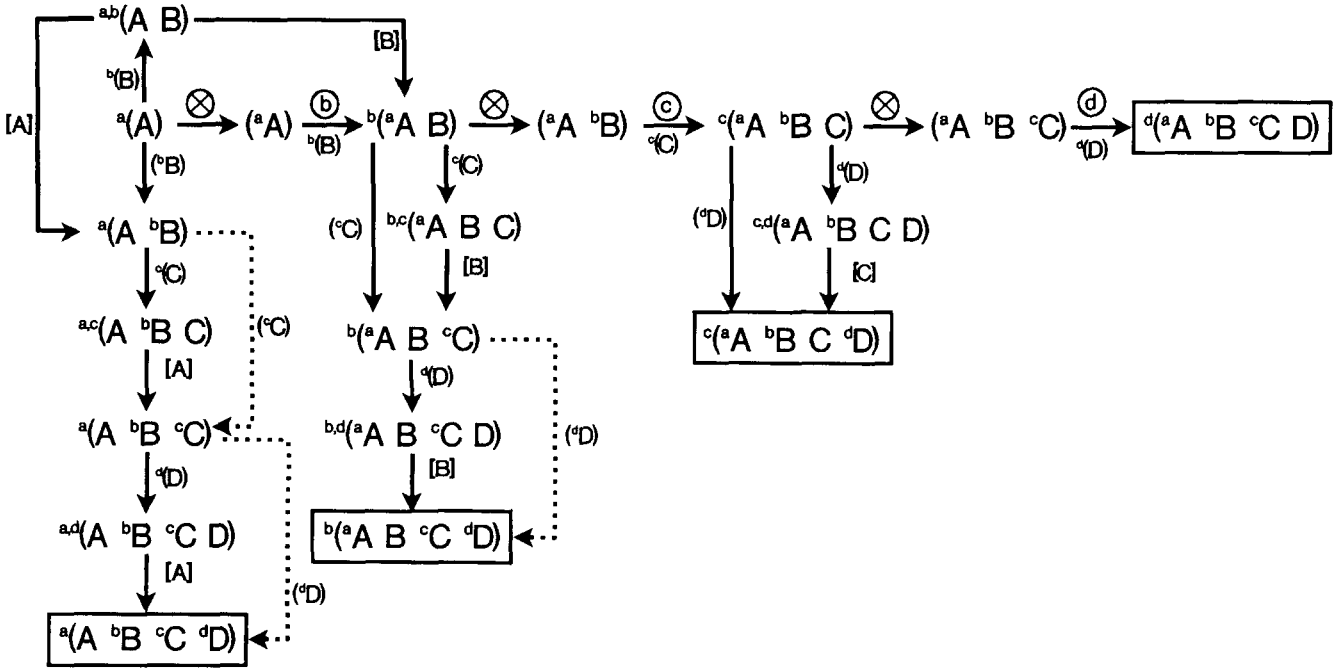

Figure 6. Synthesis diagram of the A-, B-, C- and D-partitions of the set (^{a}A ^{b}B ^{c}C ^{d}D). The synthesis of the D-partition by the split synthesis is shown horizontally, the others vertically. ⊗: mix all cells, ⓒ: redistribute into c equal parts, [A]: combine with respect to A.

members of the set ^{c}(C) with ^{b,c}(^{a}ABC) as intermediate. The synthesis of ^{b}(^{a}AB ^{c}C ^{d}D) proceeds in *bd* steps via the B,D-partiton. Similarly, the A-partition synthesis requires *a* steps toward ^{a}(A ^{b}B), *ac* steps to ^{a}(A ^{b}B ^{c}C), and *ad* steps to ^{a}(A ^{b}B ^{c}C ^{d}D).

Unless the sets involved in the combinatorial processes are small, syntheses of partitions other than the last and penultimate one, must rely on the use of robotics. It should be mentioned, however, that M2-2C processes have been employed, apparently successfully, in the constructions of all six partitions of hexapeptides[34,40–42] and orthogonal combinatorial libraries[43] as elaborated in sections 3.1 and 4.4.1, respectively.

3.3. MØ-C2 Syntheses

The rudimentary chemical reaction between two sets of compounds that are present each as single entities and applied as such in synthesis, is an important operation in combinatorial chemistry and is generally referred to as parallel synthesis. Sets of advanced lead candidates can no longer be tested as mixtures; consequently, final lead optimization is typically accomplished using MØ processes to obtain products such as ^{a}(AX$_k$) and ^{a,b}(AX$_k$B), where X$_k$ may be a pharmacophore and A and B modifying groups thereof. The intended product diversity is now more restricted and usually focused on a specific molecular feature.

To build the array, each element of the set ^{a}(A) is divided into *b* equal parts and each aliquot is placed into an individual chamber of a multiple reactor where it is allowed to react with a specific element of the set B. The corresponding conversions can be summarized as ^{a}(A) + ^{b}(B) → ^{a,b}(AB). Condensation reactions of these types are used not only for biological evaluation but also for the construction of certain set partitions as discussed in sections 3.2.2 and 3.2.3. Selective pooling of reaction products mimics M1-2C and M2-2C processes with complete reaction control.

3.4. Multicomponent condensations

Multicomponent condensations can be used to great advantage in the generation of libraries as shown by Armstrong.[14] Description of the processes can be readily accommodated with the available tools. Although multicomponent condensations are typically conducted in multiple reactor arrays using single entities each as reaction component, the use of mixtures in not inconceivable. Thus, a C3 process, such as an azomethine–ylide cycloaddition to an alkene, employing 1/2 equivalent each of two different aldehydes (A) but using the β-aminocarbonyl (B) and alkene (C) components as single entities, can be described as

$$(^{2}\text{A}) + \text{B} + \text{C} \rightarrow (^{2}\text{AB}_k\text{C}_k)$$

and the category as M1-C3. Similarly, an Ugi reaction, normally conducted as an MØ-4C step could be carried out as an M1-4C process by employing a pair of carboxylic acids while maintaining one amine, ketone and isonitrile component each. If carried out in a 96-well format, employing 192 different acids, one reaction set would furnish 192 products in pairs: ^{96}(^{2}A) + B$_k$ + C$_k$ + D$_k$ → ^{96}(^{2}AB$_k$ C$_k$ D$_k$).

4. Identification of Active Components

As one departs from the straight-forward evaluation of libraries whose components are individually addressable (e.g., those obtained by multiple MØ processes) it becomes necessary to find the active components in mixtures, such as sets of reaction products or partitions thereof. The evaluation of these mixtures requires a number of assumptions concerning both their syntheses and the biological assays employed. We have to assume equimolarity of the set's elements, absence of interactions between them, and perfect conduct at binding sites during the assay. Most notably, it is presupposed that the building blocks' contributions to the bioactivity of each member of the set are specific and additive. Ideally, the number of bioactive members of a library is relatively small and the differences in bioactivities are large.

Attempts to identify the most bioactive member in a library of compounds such as ABC ... Z focuses on the relative contributions of building blocks represented in each variable position. Bioassays of set partitions that are ordered with respect to the sets of building blocks used in the synthesis, will identity subsets of building blocks that are of relevance. Such a subset of building blocks is designated i_N, where the subscript identifies the original library partition from which the subset is derived; thus, $i_N \subset N = \{N_1, N_2, \ldots, N_n\}$. Lead identification from the combinatorial product of these subsets is based on the fundamental theorem in combinatorial analysis: given α elements in set i_A, β elements in set i_B, etc. and ω elements in set i_Z, there are $\alpha \times \beta \times \ldots \times \omega$ ordered Z-tuples ($^{\alpha}A\ ^{\beta}B \ldots\ ^{\omega}Z$) representing the pool of lead compounds. Since the bioassay of a set partition not only identifies the cells of biological significance, as specified by i_N for the N-partition, but also provides a quantitative measure, the combinatorial product of these sets can be ordered with respect to the magnitude of the predicted bioactivities.[44] Applications of this concept will be illustrated in the following sections.

4.1. Evaluation of all partitions

The simultaneous analysis of all partitions is unique in two aspects. First, the entire selection process is based on the presence of all combinations of building blocks, and secondly, the process can be completed in one step. The method was first used by Houghten in his 'positional scanning'.[40–42] Let us first use a probability argument to elaborate on the concept of lead identification by analysis of all partitions. For simplicity, we assume that only one bioactive substance is present in the entire library. The synthesis of all partitions of a set such as ($^{a}A\ ^{b}B\ ^{c}C$) provides ($a + b + c$) cells comprising a total of abc components. If one would randomly select a single member from the set ($^{a}A\ ^{b}B\ ^{c}C$), the probability to choose the 'active' one in one trial is $1/abc$. If the bioscreen highlights a particular cell of the A-partition, the probability P_A to find the active component in that cell by randomly choosing one member, is $1/bc$. Similarly, the cell of the B-partition that exhibits bioactivity allows selection of the active compound with a probability $P_B = 1/ac$, and from the analogous cell of the C-partition with the probability $P_C = 1/ab$. Using the assay results of the A- and B-partitions, limits to two the number of cells from which the active compound can be selected. The corresponding probability for the correct choice is now increased to $P_{AB} = P_A b = 1/c$. A correct selection based on two bioactive cells, one from the B- the other from the C-partition, occurs with the probability $P_{BC} = P_B c = 1/a$. If the screening results of all three partitions are available, there are three cells from which the active compound can be chosen, one cell from each partition, and the probability to correctly identify the active substance is $P_A bc = P_B ac = P_C ab = 1$.

In reality, however, an 'active' library can typically show significant biological responses in many cells of a given partition thus complicating the selection process. Returning to the example set ($^{5}A\ ^{5}B\ ^{5}C$), we first assume that the mixture elicited a positive response in a bioscreen. The A- and B-partitions (Figs 4 and 5) were then synthesized and evaluated simultaneously with the already available C-partition (Fig. 3). Figure 7(A) illustrates the corresponding hypothetical screening results.

Cells with low bioactivity are typically ignored but for argument's sake we shall consider first all, and then parts of the available data. The search for the lead candidate of the library ($^{5}A\ ^{5}B\ ^{5}C$) is then equivalent to the task of identifying and finding one particular cm^3 of space, representing the most bioactive molecule $A_iB_jC_k$, in a cube with a volume of 125 cm^3 which defines the sample space. In addressing such a situation we have to

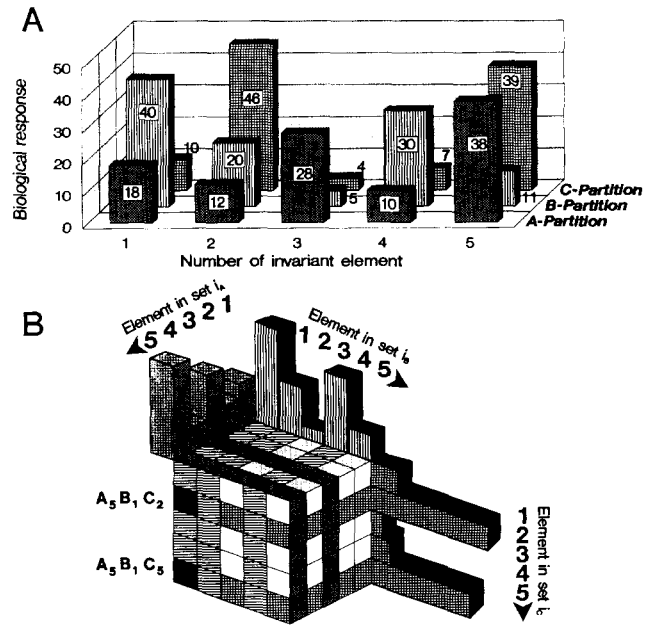

Figure 7. (A): Assay results of the A-, B- and C-partition of the set ($^{5}A\ ^{5}B\ ^{5}C$); (B): Lead identification via the combinatorial product $i_A \times i_B \times i_C$. The sets i, shown as coordinates, are derived from the assay data of the corresponding partitions.

build the combinatorial product of the entire sample space, $i_A \times i_B \times i_C$, wherein each set contains the same number of elements as the cells of original set partition used for the assay. This combinatorial product, accommodating 5^3 ordered triples, each representing an element of the library, is illustrated in Figure 7(B). The A-, B-, and C-partitions provide the three coordinates, and hence the exact location, of each individual cm^3 of space, and the assay results, in the form of the three sets i_A, i_B, and i_C, quantify the entries on each coordinate. For example, compound $A_1B_1C_1$, expected to derive its bioactivity from contributions of its building blocks in the three increments proportional to 18, 40, and 10, should exhibit the theoretical value of 68:(3 × 25) = 0.9. The compound represented by the neighboring cube, $A_1B_2C_1$, would receive the value (18 + 20 + 10):(3 × 25) = 0.6, and so on. The predicted sequence of the elements of the set (^{5}A ^{5}B ^{5}C) ranges from the most active component $A_5B_1C_2$ with a value of 1.3 to the least active $A_4B_3C_3$ with a value of 0.3. The compilation is most easily accomplished via a tree diagram in a spread-sheet program such as Excel or Lotus. Allowing a short notation for specific elements, for example $A_4B_3C_1$ = 431, the 12 leading components in the example set, by considering all available assay values, are 512 > 515 > 312 > 542 > 315 > 545 > 112 > 342 > 522 > 212 > 115 > 345. Upon reduction of the sets of the type i_N, which are employed in the construction of this virtual pool of bioactive substances, fewer members are contained therein, and, as anticipated, the predicted sequence within the selection exhibits some alterations. With a threshold value of 20, for example, 12 lead candidates can be predicted on the basis of $i_A = \{A_5, A_3\}$, $i_B = \{B_1, B_4, B_2\}$, and $i_C = \{C_2, C_5\}$, and their order is 512 > 515 > 542 > 312 > 545 > 315 > 522 > 342 > 525 > 345 > 322 > 325. The sequence reveals the unchanged dominance of the previously identified two leads. Restricting the subsets further and allowing only A_5, B_1, and $i_C = \{C_2, C_5\}$ to enter the combinatorial product, limits the prediction to the two candidates, 512 > 515, which are accentuated in Figure 7(B).

In view of the experimental effort required for the synthesis of the partitions other than those of the Z- and (Z-1)-type, this method of lead identification is obviously attractive if the library under investigation has only two variable elements. The hypothetical assay results for the A- and B-partitions of the example library (^{5}A ^{5}B) are shown in Figure 8(A). Lead identification by construction of the product set $i_A \times i_B$ as depicted in Figure 8(B) is now particularly simple in view of its two-dimensional nature. By considering a threshold value of 14, two sets are obtained, $i_A = \{A_2, A_3\}$ and $i_B = \{B_2, B_4\}$. The line intensities shown in Figure 8(B) are directly proportional to the bioactivities of the cells of the corresponding partitions, so that intersections of bolder lines reveal elements of higher bioactivity as emphasized by larger circles, suggesting the sequence $A_3B_4 > A_2B_4 > A_3B_2 > A_2B_2$.[45] The power of this method has been recognized and demonstrated. In one example an amide/ester library of the

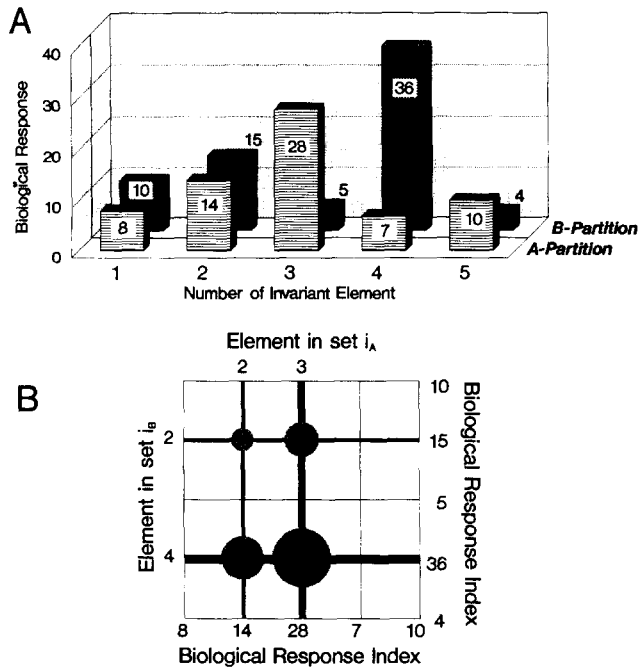

Figure 8. (A): Assay results for A- and B-partitions; (B): Identification of bioactive components via the combinatorial product of sets $i_A = \{A_2, A_3\}$ and $i_B = \{B_2, A_4\}$.

type (^{40}A ^{40}B),[46] in another, a carbamate library (^{9}A ^{6}B) was evaluated.[47]

4.2. Evaluation of the last partition followed by set reductions[48]

Since the last partition of a set is most readily obtained synthetically, it is advantageous to build the identification protocol on the corresponding intermediate sets. To illustrate the concept, the synthetic example described previously would typically be studied according to the diagram shown in Figure 9(A), where the actual synthetic steps employed in the construction of the general library (^{a}A ^{b}B ^{c}C) are delineated by solid lines. The first step in the identification process invariably involves the analysis of the C-partition (e.g., Fig. 3). The test results will then establish the C-moiety, C_k, or an entire subset i_C whose members contribute significantly to the bioactivity. The next step commences with the A-partition of the set (^{a}A ^{b}B). Each cell of this partition is allowed to react with the selected member C_k of the subset i_C to lead to the A-partition of the first reduced set, ^{a}(A ^{b}BC_k). The assay results of this partition will then identify the set i_A. To identify i_B, the second reduced set in the form of the partition ^{b}(A_kBC_k) is required. The syntheses commence with the most prominent member A_k, which is allowed to react with ^{b}(B), to give the partition ^{b}(A_kB), where $A_k \in i_A$. Further couplings with C_k complete the partition of the second reduced set ^{b}(A_kBC_k). Its evaluation identifies i_B and hence predicts a lead candidate $A_kB_kC_k = A_1B_5C_2$ as seen in Figure 9(B) for the example set (^{5}A ^{5}B ^{5}C).

In section 4.1 we pointed to the advantage of the simultaneous availability of the entire three-dimensional sample space containing *abc* elements. Inspection of Figure 9(B) illustrates how specific, consecutive selections, one from each of the sets i_C and i_B, gradually reduce the sample spaces that are available for the selection process. Once a single unit from the set i_C has been chosen (e.g., C_2) the remaining sample space contains only *ab* elements, and after a further, single selection from the set i_B (e.g., B_5) only *a* elements remain as defined by the partition of the second reduced set.[49] As a result, the set i_A is biased by the previous selection C_k from i_C, and i_B is prejudiced by the selection of A_k from i_A. One could therefore argue that the generation of a *series* of lead candidates, by the inclusion of these sets i_A and i_B in the combinatorial product $i_A \times i_B \times i_C$, may be flawed as it does involve the recreation of a three-dimensional sample space containing two subjective coordinates. Under the assumption of building block additivity, however, the simulation of the three-dimensional sample space embracing all *abc* elements is permissible. Clearly, the identification of a lead candidate is of primary concern in a deconvolution strategy. Predictions of the order of the library's elements in terms of their bioactivities, however, can be a powerful tool for the discovery of trends within the set. The succeeding paper provides an illustrative application of this method.[50]

With the aid of the tools described, the lead identification program of a more complex library can be described analogously. Consider, for example, a library of 8000 hexapeptides containing 20 different amino acids. Let us assume that the members of this set are defined in terms of three amino acids (B_1, B_2, and B_3) but contain the preceding three (X1, X2, and X3) in all possible arrangements. The construction of this set (^{20}X1 ^{20}X2 ^{20}X3 $B_1B_2B_3$) can be accomplished by application of the split synthesis leading first to the X3-partition, $^{20}(^{20}$X1 ^{20}X2 X3 $B_1B_2B_3$). The assay results of this partition will point to the set i_{X3} from which $X3_k$ can be selected. A return to the X2-partition of the set (^{20}X1 ^{20}X2), available from the synthesis of the original set, will permit the reactions $^{20}(^{20}$X1 X2) + $X3_k \rightarrow$ $^{20}(^{20}$X1 X2 $X3_k$) and consecutive condensations with B_1, B_2, and B_3 furnish the partition $^{20}(^{20}$X1 X2 $X3_k$ $B_1B_2B_3$) leading, based on the bioassay, to the set i_{X2} from which $X2_k$ can be chosen. The next step commences with the set ^{20}(X1) whose members are consecutively coupled to $X2_k$, $X3_k$, B_1, B_2, and B_3, furnishing i_{X3} and $X3_k$ via ^{20}(X1 $X2_k$ $X3_k$ $B_1B_2B_3$).

Reduced sets as a tool in lead identification were first used by Geysen as part of the mimotope strategy,[51,52] also referred to as 'iterative deconvolution'.[53]

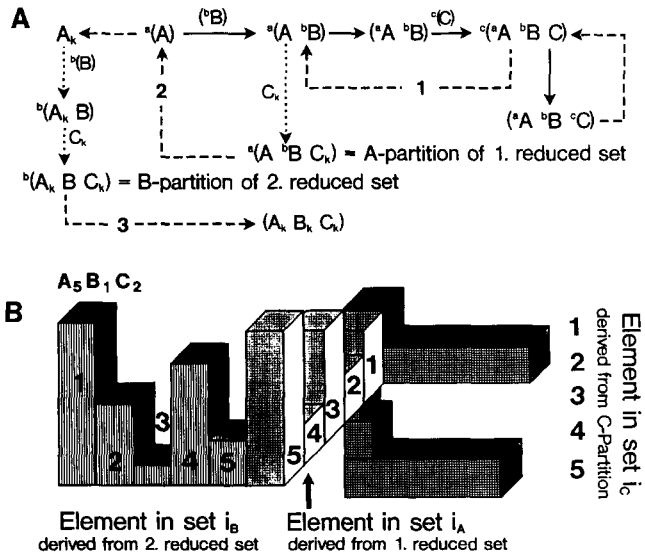

Figure 9. (A): Identification of bioactive components in the set (^{a}A ^{b}B ^{c}C) by evaluation of the C-partition, followed by partitions of two reduced sets. Solid lines show the synthesis of the original library and its C-partition, dotted lines indicate the generation of the reduced sets, and dashed lines portray the action sequence after the library was shown to be bioactive; **1**: test C-partition to get C_k, **2**: test partition of first reduced set to get A_k, **3**: test partition of second reduced set to get B_k. (B): Example of lead identification in set (^{5}A ^{5}B ^{5}C) commencing with the assay data of the C-partition ($i_C = C_2$), followed by the first ($i_A = A_5$) and the second ($i_B = B_1$) reduced set revealing the lead $A_5B_1C_2$. The bioassay data for the two reduced sets are assumed to be identical with those of the A- and B-partitions shown in Figure 7.

4.3. Evaluation of the last and penultimate partitions followed by set reductions

As discussed previously, lead identification is accelerated by the simultaneous analysis of all partitions. One of the disadvantages associated with this approach lies in the relative inaccessibility of partitions that are remote from the last one as illustrated in Figure 6. The penultimate partition of any set of the type ^{y}(... ^{x}X Y ^{z}Z), however, may be accessible either by the reaction ^{y}(... ^{x}X Y) + (^{z}Z) → ^{y}(... ^{x}X Y ^{z}Z), or ^{y}(... ^{x}X Y) + ^{z}(Z) → ^{y,z}(... ^{x}X Y Z) → ^{y}(... ^{x}X Y ^{z}Z) involving M2*-2C or M1*-2C processes, respectively. Two partitions of the set can then be evaluated simultaneously, so that the identity of i_Y and i_Z become known at the outset. Building on the resulting optimized terminal sequence Y_k Z_k, the assay results of the first reduced set, prepared according to ^{x}(X) + Y_k → ^{x}(XY_k), and ^{x}(XY_k) + Z_k → ^{x}(XY_k Z_k), furnishes the set i_X from which X_k is selected. If a library has additional variable positions, the protocol of reduced-set synthesis is self-explanatory, for example, ^{w}(W) + X_k → ^{w}(W X_k), ^{w}(W X_k) + Y_k → ^{w}(W X_k Y_k), and ^{w}(W X_k Y_k) + Z_k → ^{w}(W X_k Y_k Z_k), leading to i_W and W_k.

The diagram of the synthetic processes and the contribution of the products toward the identification of the bioactive member is illustrated as a simple example in Figure 10(A). Returning again to the example library (^{5}A ^{5}B ^{5}C) and using the assay data shown in Figure 7, the B- and C-partitions furnish, considering a threshold value of 35, $i_B = \{B_1, B_4, B_2\}$ and $i_C = \{C_2, C_5\}$. Their combinatorial product identifies six ordered pairs, shown as $B_1C_2 > B_1C_5 > B_4C_2 > B_4C_5 > B_2C_2 > B_2C_5$ (Fig. 10B).

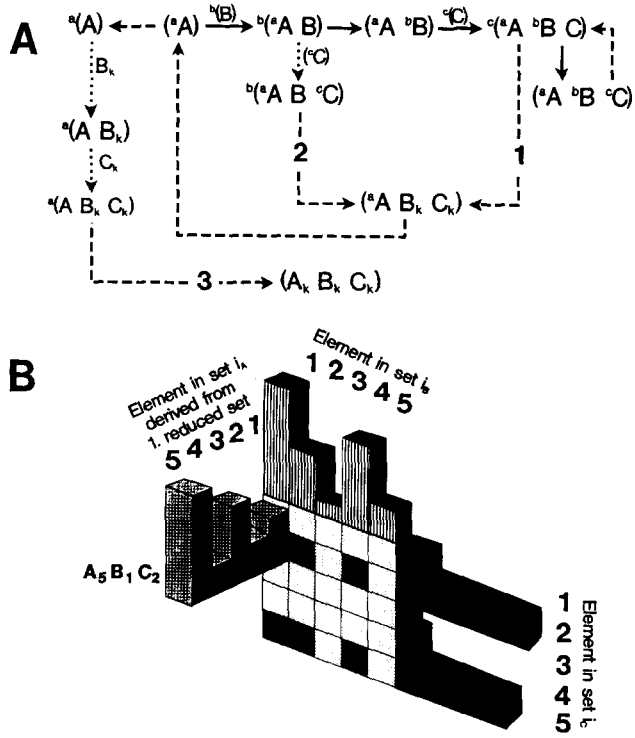

Figure 10. (A): Identification of bioactive components in the set (^aA ^bB ^cC) by evaluations of the C- and B-partitions, followed by the partition $^a(AB_kC_k)$ of the reduced set. For line legend see Figure 9; **1**: test C-partition to get C_k, **2**: test B-partition to get B_k, **3**: test partition of first reduced set to get A_k. (B): example of lead identification.

Each of these could be used as educt for the syntheses of a reduced set of the type $^5(AB_iC_j)$ furnishing, upon bioassay, a set i_A. Restricting the choice to the most prominent pair, namely B_1C_2, suggests synthesis of a reduced set in the form of the partition $^5(AB_1C_2)$. Biological evaluation identifies i_A = {A_5, A_3} and hence the lead candidate $A_5B_1C_2$.[54] As illustrated in Figure 10(B), this method of analysis again reconstitutes a three-dimensional sample space generated by the product of $i_{B,C}$ and the i_A-axis.[55] Of these multiplicands, the i_B, i_C-plane is generated in the presence of all elements; the axis represented by the set i_A, however, is based on a specific selection from the set $i_{B,C}$ and is therefore biased by this choice. Thus, the resulting three-dimensional sample space contains only one subjective coordinate. Discounting the effort required for the synthesis of the last and penultimate partitons, this method of prediction of the leads' order should be an attractive tool.

4.4. Evaluation of orthogonal partitions

4.4.1. Orthogonal partitions of the entire sample space. As outlined in section 3.1, M2-2C synthetic processes can generate large libraries entailing a minimum of experimental effort. Since the mixtures are generated in one step, lead identification would pose a particularly complex situation, especially if large reagent sets are involved. Addressing this problem, Tartar et al.[43] proposed a theoretically interesting application of M2-2C protocols. Consider the synthesis of the library (^{25}N ^{25}N ^{25}N) by the two steps (^{25}N) + (^{25}N) → (^{25}N ^{25}N) and (^{25}N ^{25}N) + (^{25}N) → (^{25}N ^{25}N ^{25}N). Assuming ideal reaction conditions, the resulting product would constitute a mixture with 15,625 components but lacking a rational avenue for lead identification. Thus, the required reagent set ^{25}(N) = {a, b, c, d, ... , y} is partitioned into five cells according to the diagram shown in Figure 11, so that ^{25}(N) = ^{5}(^{5}N), where ^{5}(^{5}N) = ^{5}(A) = {A1, A2, A3, A4, A5}, and the subsets A1 = {a, f, k, p, u}, A2 = {b, g, l, q, v}, etc. We refer to this partition of the set ^{25}(N) as 'type A' to accommodate the letter used by the authors. The partition of set ^{25}N is then repeated, 'orthogonally' to the one of the A-type, as revealed in Figure 11: ^{25}(N) = ^{5}(^{5}N), where ^{5}(^{5}N) = ^{5}(B) = {B1, B2, B3, B4, B5}, and the subsets B1 = {a, b, c, d, e}, B2 = {f, g, h, i, j}, etc. We refer to this partition of the set ^{25}(N) as 'type B'.

As a consequence of the orthogonal relationship between the A- and B-type partitions, each cell of the type-A shares exactly one element with any cell of the B-type partition. The cell A1 = {a, f, k, p, u}, for example, shares only the element *f* with the cell B2 = {f, g, h, i, j}. Each of the five cells of the A-type partition is then condensed, by application M2-2C protocols, with each of the cells of the A-type partition to furnish 25 mixtures of dimers embracing all possible arrangements of the A-cells, ranging from (A1 A1) to (A5 A5): A1 + A1 → (A1 A1), A1 + A2 → (A1 A2), etc. Let A represent any column, and B any row in the orthogonal partition of the set N (Fig. 11), so that ^{5}(A) = ΣA and ^{5}(B) = ΣB, then the synthetic array of dimers is identical with the Cartesian product ^{5}(A) × ^{5}(A) = ^{25}(AA); since ^{5}(A) is actually a partition of N containing five cells, there are 25 mixtures of dimers and since each of these cells is represented by five building blocks, there are 25 × 5 × 5 = 625 dimers of type N_iN_j contained therein. In other words, the product array is a partition of the set (^{25}N ^{25}N). Further condensations of each of these 25 mixtures with all five cells of the A-type partition, individually, then yields the product ^{25}(A) × ^{5}(A) → ^{125}(AAA) in the form of 125 mixtures of trimers

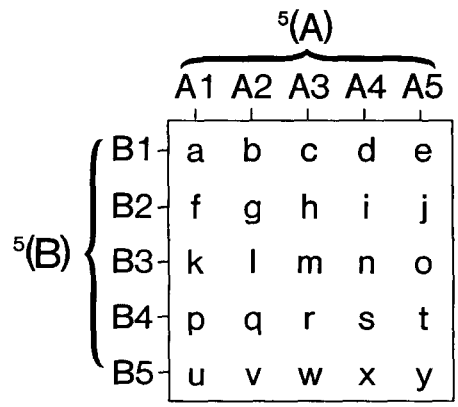

Figure 11. Orthogonal partition of the set ^{25}(N) = {a, b, c, d, ... , y}.

as shown below. Each of these 125 mixtures contains components with three variable positions, represented five times each, so that there are a total of 125 × 5 × 5 × 5 = 15,625 individual compounds present.

(A1 A1 A1) (A1 A2 A1) (A1 A3 A1) (A1 A4 A1) (A1 A5 A1)
(A1 A1 A2) (A1 A2 A2) (A1 A3 A2) (A1 A4 A2) (A1 A5 A2)
(A1 A1 A3) (A1 A2 A3) (A1 A3 A3) (A1 A4 A3) (A1 A5 A3)
(A1 A1 A4) (A1 A2 A4) (A1 A3 A4) (A1 A4 A4) (A1 A5 A4)
(A1 A1 A5) (A1 A2 A5) (A1 A3 A5) (A1 A4 A5) (A1 A5 A5)

(A2 A1 A1) (A2 A2 A1) (A2 A3 A1) (A2 A4 A1) (A2 A5 A1)
(A2 A1 A2) (A2 A2 A2) (A2 A3 A2) (A2 A4 A2) (A2 A5 A2)
(A2 A1 A3) (A2 A2 A3) (A2 A3 A3) (A2 A4 A3) (A2 A5 A3)
(A2 A1 A4) (A2 A2 A4) (A2 A3 A4) (A2 A4 A4) (A2 A5 A4)
(A2 A1 A5) (A2 A2 A5) (A2 A3 A5) (A2 A4 A5) (A2 A5 A5)

(A3 A1 A1) (A3 A2 A1) (A3 A3 A1) (A3 A4 A1) (A5 A5 A1)
(A3 A1 A2) (A3 A2 A2) (A3 A3 A2) (A3 A4 A2) (A5 A5 A2)
(A3 A1 A3) (A3 A2 A3) (A3 A3 A3) (A3 A4 A3) (A5 A5 A3)
(A3 A1 A4) (A3 A2 A4) (A3 A3 A4) (A3 A4 A4) (A5 A5 A4)
(A3 A1 A5) (A3 A2 A5) (A3 A3 A5) (A3 A4 A5) (A5 A5 A5)

(A4 A1 A1) (A4 A2 A1) (A4 A3 A1) (A4 A4 A1) (A4 A5 A1)
(A4 A1 A2) (A4 A2 A2) (A4 A3 A2) (A4 A4 A2) (A4 A5 A2)
(A4 A1 A3) (A4 A2 A3) (A4 A3 A3) (A4 A4 A3) (A4 A5 A3)
(A4 A1 A4) (A4 A2 A4) (A4 A3 A4) (A4 A4 A4) (A4 A5 A4)
(A4 A1 A5) (A4 A2 A5) (A4 A3 A5) (A4 A4 A5) (A4 A5 A5)

(A5 A1 A1) (A5 A2 A1) (A5 A3 A1) (A5 A4 A1) (A5 A5 A1)
(A5 A1 A2) (A5 A2 A2) (A5 A3 A2) (A5 A4 A2) (A5 A5 A2)
(A5 A1 A3) (A5 A2 A3) (A5 A3 A3) (A5 A4 A3) (A5 A5 A3)
(A5 A1 A4) (A5 A2 A4) (A5 A3 A4) (A5 A4 A4) (A5 A5 A4)
(A5 A1 A5) (A5 A2 A5) (A5 A3 A5) (A5 A4 A5) (A5 A5 A5)

The synthesis of 125 mixtures is then repeated employing the five cells of the original set's B-type partition leading to ^{125}(BBB). Due to the orthogonal relationship between the sets of cells, any one of the 125 product mixtures derived from the partition of the A-type will only share one of its 125 elements (trimers) with any of the 125 elements taken from any of the 125 products derived from the synthesis employing the B-type cells. As an example, the product mixture (A3 A2 A4) which is composed of the cells A3 = {c, h, m, ***r***, w}, A2 = {b, g, ***l***, q, v} and A4 = {***d***, i, n, s, x} shares only one element with, say, (B4 B3 B1). The latter product mixture is constructed from the cells B4 = {p, q, ***r***, s, t}, B3 = {k, ***l***, m, n, o} and B1 = {a, b, c, ***d***, e}. Consequently, the only common element is (***r l d***).

In reality, one would expect that both the A- and B-type partitions provide a plethora of bioactive cells each, so that two *sets* of mixtures of the type (AAA) and (BBB) are identified. Allowing a mixture such as (A1 A2 A3) be represented by the three-digit number 123, then the significant members of these sets i_A and i_B, which serve for lead prediction, consist of members, each represented by three digits, and each digit reflecting a set of five building blocks according to Figure 11. The combinatorial product of these two sets of three-digit numbers, $i_A \times i_B$, yields ordered pairs of three-digit numbers that are related to each other orthogonally. Each pair can therefore identify a lead candidate. Moreover, since each set member can be associated with an assay value, each ordered pair that is part of the combinatorial product can be assigned the sum of two assay values. Consequently, ranking of the candidates in terms of their bioactivity is again theoretically feasible.

To briefly illustrate the concept, three cells in both the A- and B-type partitions are assumed to exhibit particularly high bioactivities and are selected for the identification process, namely (A2 A3 A1) [10], (A4 A4 A1) [6], (A3 A1 A2) [2], (B5 B4 B2) [8], (B2 B1 B4) [5], and (B3 B1 B3) [3], where the number in brackets represents an arbitrary assay value. The combinatorial product consists of nine ordered pairs of three-digit numbers, each pair identifying a three-letter word, deciphered according to the code contained in Figure 11, and hence describing a distinct chemical entity. Nine members are thus identified and listed in decreasing order of activity: vrf, gcp, xsf, lck, idp, wpg, ndk, haq, mal.

In spite of its theoretical appeal, one disadvantage associated with such an orthogonal combinatorial library is the need for sequential, M2-2C synthetic steps. Further, there is a limitation as to element selection for the variable position in the core molecule: All building blocks are derived from a common set N = $\{N_1, N_2, \ldots, N_n\}$, leading to libraries of the type (^{n}N ^{n}N ^{n}N) where $\sqrt{n}$ is an integer x. Library types (^{n}N ^{p}P ^{q}Q) and (^{n}N ^{p}N ^{q}N), where n, p, and q are arbitrary integers, are thus excluded. Another problem will surface whenever many components with significant bioactivities are present in a given mixture as the protocol can only predict one candidate from each intersection between any (AAA) and (BBB). Thus, one candidate must be selected from a relatively large sample space. Let ^{x}(A) = ΣA and ^{x}(B) = ΣB, and restrict the consideration to libraries of trimers, then the sample space which is subject to the selection process is $x^3 = n\sqrt{n}$, so that $x^3 \subset n^3$, where n^3 is the entire sample space.

4.4.2. Evaluation of the last partition in conjunction with orthogonal partitions. The weakness of M2-2C syntheses with regard to product homogeneity has been emphasized in section 3.1. This problem is obviously compounded if two M2-2C processes are conducted in sequence as required in the strategy described in section 4.4.1. To avoid two consecutive M2-2C steps, the following alternative is feasible. Aiming for the same combinatorial space as described in section 4.4.1 but compressing the space from which elements are selected by the orthogonal strategy, the first step consists of the synthesis of the last partition of the set, in the form of ^{25}(^{25}N ^{25}N N), by the split synthesis: (^{25}N)

$+ {}^{25}(N) \rightarrow {}^{25}({}^{25}N\ N)$, ${}^{25}({}^{25}N\ N) \rightarrow ({}^{25}N\ {}^{25}N)$, $({}^{25}N\ {}^{25}N) + {}^{25}(N) \rightarrow {}^{25}({}^{25}N\ {}^{25}N\ N)$. Analysis of the 25 cells of this partition will furnish the biologically relevant element set i_N from which N_k is selected to identify the subset $({}^{25}N\ {}^{25}N\ N_k)$. Proceeding with the orthogonal partitions of the set $({}^{25}N)$ and the syntheses of the dimer mixtures ${}^{25}(AA)$ and ${}^{25}(BB)$ as described in the previous section, the necessity to conduct a second M2-2C step no longer exists. The two product groups ${}^{25}(AA)$ and ${}^{25}(BB)$ are coupled to N_k to furnish 25 new products each, which serve for lead identification: ${}^{25}(AA) + N_k \rightarrow {}^{25}(AAN_k)$ ${}^{25}(BB) + N_k \rightarrow {}^{25}(BBN_k)$.

The composition of ${}^{25}(AAN_k)$ is shown below.

{A1 A1 N_k}, {A1 A2 N_k}, {A1 A3 N_k}, {A1 A4 N_k}, {A1 A5 N_k},

{A2 A1 N_k}, {A2 A2 N_k}, {A2 A3 N_k}, {A2 A4 N_k}, {A2 A5 N_k},

{A3 A1 N_k}, {A3 A2 N_k}, {A3 A3 N_k}, {A3 A4 N_k}, {A3 A5 N_k},

{A4 A1 N_k}, {A4 A2 N_k}, {A4 A3 N_k}, {A4 A4 N_k}, {A4 A5 N_k},

{A5 A1 N_k}, {A5 A2 N_k}, {A5 A3 N_k}, {A5 A4 N_k}, {A5 A5 N_k}.

In contrast to the 2×125 analyses required for the orthogonal combinatorial libraries described in section 4.4.1 that contain 125 components each, this alternate method presented here requires the analyses of the 25 cells of the partition ${}^{25}({}^{25}N\ {}^{25}N\ N)$ followed by only 2×25 additional samples. While the 25 samples of the partition ${}^{25}({}^{25}N\ {}^{25}NN)$ contain 625 compounds each, the subsets ${}^{25}(AAN_k)$ and ${}^{25}(BBN_k)$, resulting from the orthogonal partition, comprise only 25 species each. It should be noted that this protocol is not restricted to libraries of the type $({}^{n}N\ {}^{n}N\ {}^{n}N)$, but can be applied to mixtures of the type $({}^{n}N\ {}^{n}N\ {}^{p}P)$ where n differs from p and where p is any integer.

4.4.3. Evaluation of the last partition in conjunction with a reduced set and orthogonal partition. Driven by the goal to eliminate M2-2C processes and to extend the combinatorial space from $({}^{n}N\ {}^{n}N\ {}^{n}N)$ to $({}^{n}N\ {}^{p}P\ {}^{q}Q)$, a further option is realized by restricting the orthogonal strategy to the N-position. Thus, N, P, and Q may be dissimilar, and p and q may be any integer, while the restriction that $\sqrt{n}$ = an integer persists. Synthesis commences with the last partition, $({}^{25}N) + {}^{p}(P) \rightarrow {}^{p}({}^{25}N\ P) \rightarrow ({}^{25}N\ {}^{p}P)$, $({}^{25}N\ {}^{p}P) + {}^{q}(Q) \rightarrow {}^{q}({}^{25}N\ {}^{p}P\ Q)$. Bioassays will furnish i_Q and hence Q_k, permitting the synthesis of a reduced set, ${}^{p}({}^{25}N\ P) + Q_k \rightarrow {}^{p}({}^{25}N\ P\ Q_k)$ that leads to i_P and hence to P_k. With the most promising terminus P_kQ_k at hand, ${}^{25}(N)$ is then partitioned orthogonally as described in section 4.4.1 and Figure 11. Subsequent condensations with P_k and Q_k according to the scheme ${}^{5}(A) + P_k \rightarrow {}^{5}(A\ P_k)$, and ${}^{5}(A\ P_k) + Q_k \rightarrow {}^{5}(A\ P_k\ Q_k)$ are repeated with the cells of the B-type partition, without the need for any M2-2C protocols, to furnish the sublibraries detailed below:

{A1 P_k Q_k}, {B1 P_k Q_k},
{A2 P_k Q_k}, {B2 P_k Q_k},
{A3 P_k Q_k}, {B3 P_k Q_k},
{A4 P_k Q_k}, {B4 P_k Q_k},
{A5 P_k Q_k}, {B5 P_k Q_k}.

A specific example is shown in Figure 12. The bars represent the hypothetical bioscreening results of the cells of the A- and B-type partitions and provide the sets i_A and i_B, respectively. Identification of N_k for the final candidate N_k P_k Q_k, from the set i_N, proceeds again by generating the combinatorial product $i_A \times i_B$ of the assay data as shown in Figure 12. All 25 members of the resulting set i_N are ranked without the assumption of a threshold value. The order of activity is shown in numerical sequence ranging from one to 25, within the central square. The most active member is nP_k Q_k due to the prominence of (A4 P_k Q_k) and (B3 P_k Q_k), followed by kP_k Q_k. The assay data of these A- and B-type sublibraries lead to product sets expected to identify the lead compounds more reliably than the orthogonal strategies discussed previously.

4.4.4. Evaluation of last and penultimate partitions in conjunction with orthogonal partition. In another variation of the protocol elaborated above, libraries of the type $({}^{n}N\ {}^{p}P\ {}^{q}Q)$ can be evaluated again by applying the orthogonal strategy exclusively to position N. The last and penultimate partitions of the set $({}^{25}N\ {}^{p}P\ {}^{q}Q)$, are assembled in the conventional fashion: $({}^{25}N) + {}^{p}(P) \rightarrow {}^{p}({}^{25}N\ P) \rightarrow ({}^{25}N\ {}^{p}P)$, $({}^{25}N\ {}^{p}P) + {}^{q}(Q) \rightarrow {}^{q}({}^{25}N\ {}^{p}P\ Q)$, and ${}^{p}({}^{25}N\ P) + ({}^{q}Q) \rightarrow {}^{p}({}^{25}N\ P\ {}^{q}Q)$. Assays of these two

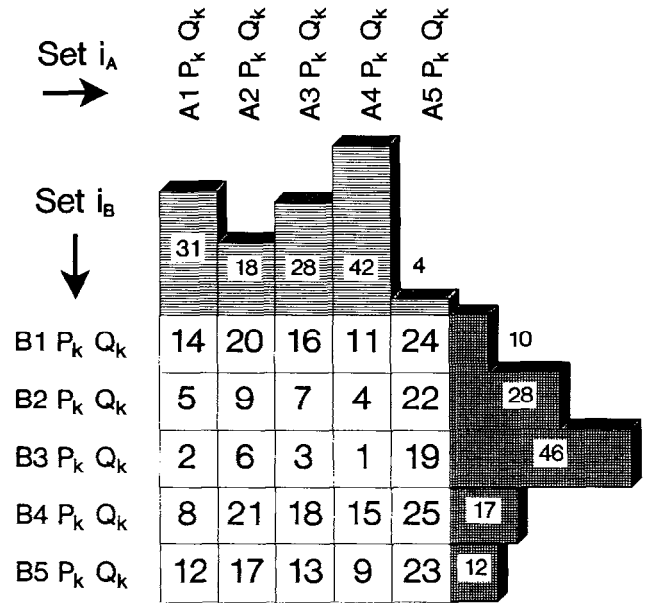

Figure 12. Assay data for A- and B-type partitions of the set $({}^{25}NP_k\ Q_k)$ are shown by bars. Crosspoints identify the elements and their ranking in terms of bioactivity is indicated by the sequence 1–25.

partitions furnish i_P and i_Q so that the subsequent orthogonal partition process for the evaluation of the last position is the same as discussed above in section 4.4.3. The considerable advantage of the simplified orthogonal synthetic strategy is partly offset by the need of products available by application of the split synthesis. These requirements include $^{p}(^{n}N\ ^{n}N\ P)$ in section 4.4.2, $^{q}(^{n}N\ ^{p}P\ Q)$ in 4.4.3, and $^{q}(^{n}N\ ^{p}P\ Q)$ and $^{p}(^{n}N\ P\ ^{q}Q)$ in 4.4.4.

4.4.5. Comparison of orthogonal strategies. In conclusion, the assay data treated in section 4.4.1 were based on x^3 mixtures of the type (AAA) and (BBB) taken from the collections ^{x,x,x}(AAA) and ^{x,x,x}(BBB), where N = {a, b, ... , n}, where $\sqrt{n}$ is an integer x, and ^{x}(A) = {A1, A2, ... , Ax}, and ^{x}(B) = {B1, B2, ... , Bx}. In section 4.4.2 the sample complexity that is subject to the orthogonal strategy was reduced to the form $(^{n}N\ ^{n}N\ ^{p}P)$ where samples of the type (AAN_k), contained in x^2 mixtures, were generated. An alternate protocol, detailed in section 4.4.3, produces mixtures of the type $^{x}(AP_kQ_k)$ containing only x cells, i.e., there are x individual mixtures AP_kQ_k, each containing only one variable moiety while the molecular portion (P_kQ_k) is invariant throughout all cells of the A- and B-type partitions. The collections $^{x}(AP_kQ_k)$ and $^{x}(BP_kQ_k)$ that lead to the identification of the last unsolved molecular position are again prevalent in section 4.4.4, but the prelude to that stage is base on the simultaneous evaluation of two partitions generated by conventional methods.

If we continue to limit our considerations to trimeric elements, the entire sample space is n^3. The sample space that is subject to lead selection by the orthogonal strategy was a fraction thereof in section 4.4.1, namely $n\sqrt{n} = x^3$. In section 4.4.2 this fraction was reduced to n, and in sections 4.4.3 and 4.4.4, the space was further reduced to $\sqrt{n}$. It is anticipated that accuracy of lead prediction by the orthogonal strategy will increase with decreasing sample space.

5. Summary and Outlook

Combinatorial organic-synthetic processes and the steps required toward the identification of the bioactive products are discussed with the aid of set theory. This treatment leads to short and accurate descriptions of all operations. A library, for example, $(^{a}A\ ^{b}B\ ^{c}C\ ...\ ^{z}Z)$, containing compounds with Z variable molecular moieties and introduced by consecutive coupling reactions, can be regarded as a set, presentable in the form of Z partitions whose cells are ordered with respect to the variable moieties. The first solution to the identification problem of the bioactive component in the set consists of the synthesis and simultaneous analysis of all Z partitions. Assuming that the moiety Z is the last of the building blocks added in the course of the library construction, then a second solution is found in the synthesis and analysis of the last or Z-partition, followed by reiterative set reductions in conjunction with the analysis of the resulting partitions. A third solution commences with the synthesis and evaluation of the last (Z) and penultimate (Z-1) partitions of the set. Simultaneous analyses of these partitions will identify the active moieties in positions Z and Z-1 of the molecule ABC ... Z, so that lead identification in a set composed of members with only two variable positions can then be realized from the product set derived from the significant assay data. Sets composed of members featuring additional variable positions can be solved in terms of their remaining degenerate positions after Z-2 consecutive set reductions. The most recent proposal for lead identification is based on orthogonal partitions of element sets and leads to two sets of libraries that complement each other in such a manner that the identification is conceivable from one collection of assay results. To cover the sample space $(^{n}N\ ^{n}N\ ^{n}N)$ requires two consecutive M2-2C synthetic operations and the total number of steps is $2n(1 + \sqrt{n})$.[56] Restricting the orthogonal methodology to two positions addresses the combinatorial space $(^{n}N\ ^{n}N\ ^{p}P)$ and reduces the complexity of the complimentary orthogonal partitions on which lead identification is based. The required number of M2-2C operations is reduced to one and the total number of steps to $5n + p$.[57] Restricting the orthogonal partition to N in the molecular motif $(^{n}N\ ^{p}P\ ^{q}Q)$, while utilizing the last partition and one reduced set for bioevaluation, eliminates the need for M2-2C syntheses entirely. As a result, the sample complexity that is subject to analysis is further reduced and the assembly requires only $2p + q + 4\sqrt{n}$ reaction steps[58] for its construction. Assuming $n = p = i$, and comparing the combinatorial spaces[59] in terms of the required number of synthetic operations under the conditions elaborated above, reveals that this number decreases in the sequence $(^{n}N\ ^{n}N\ ^{n}N) > (^{n}N\ ^{n}N\ ^{p}P) > (^{n}N\ ^{p}P\ ^{q}Q)$. These differences diminish, however, with decreasing n.

Considerable effort has been expended to determine which of the lead-identification method should give the most reliable prediction. In one study it was concluded that the evaluation, based on the last partition in conjunction with reduced sets, was more reliable than the analysis via all partitions.[60] Another one suggested that the two methods were comparable in their efficacy.[61] It is our opinion that a judgment as to the reliability of a certain method may not be made a priori. If the chemical synthesis and the bioassay would proceed under ideal conditions, and if all building blocks used in the construction of a library such as $(^{a}A\ ^{b}B\ ^{c}C)$ would indeed exhibit additive properties related to the expression of bioactivity, all methods of lead identification discussed here should be of comparable accuracy. Certain libraries, in conjunction with a given bioassay, exhibit biopotencies reflecting remarkable building block additivity and apparently ideal conduct at the binding sites in the assay system, while others do not. Lead prediction in the former case can be expected to be much more successful that in the latter regardless of the protocol used. The synthetic reliability to produce materials of undistorted product distribution is also

critical; products prepared by M2-2C processes are usually suspect in this regard.

Smaller libraries will obviously be more amenable to a deconvolution program than large ones. The spread of the actual assay values will also be consequential: Screening results of a certain partition of a set, as well as the bioactivities of individual compounds within a cell of such a partition, can be regarded as probability distributions. It is anticipated that more meaningful lead candidates will emanate from libraries where the variances of these collections are relatively small.

Pharmaceutical companies compete on the basis of product innovations and their developments. It has now been recognized that superior process development capabilities, occasionally discounted as a research and development determinant of secondary importance, can directly and positively impact product innovation.[62] The evolution of combinatorial technology supports this conviction. Combinatorial chemistry has its roots in process research where its major components, such as robotics, high-yielding bond-formation protocols, solid-phase reaction technology, etc., were originally developed for other purposes. As an established tool, however, combinatorial chemistry advances product innovation by providing novel molecular entities (s_1 level) and enhancing product development (s_3 level). Indeed, new process and product innovations may mutually benefit each other. Process innovations in nucleic acid synthesis enabeled antisense technology and created the potential for antisense drugs; at the same time, antisense technology is used in the analysis of gene function and gene expression with the potential for contributions to genomics and gene therapy, thus aiding in the discovery of drugs of a different dimension. Similarly, a new synthetic route for a certain drug candidate may furnish novel building blocks or privileged structural elements for the construction of new chemical compound libraries via combinatorial chemistry which, in turn, may lead to new development candidates. Further, combinatorial technology may serve as a tool in process research, as in the catalyst selection process, for example.

As pointed out in section 1, the chemical structures, as provided by combinatorial processes, are usually unrefined so that the time lines in the development phase come under pressure. Rapid process development and process innovation capabilities can compensate for these delays. As the competitive position is based on high levels of both product and process innovations, industrial enterprises can be weighed against each other by their relative position on the surface plot shown in Figure 13. Accordingly, the greatest competitive advantages for the future will be enjoyed by a company that assumes the highest position on this surface: it has many new molecular entities in various phases of development and leads in the area of process innovations. These capabilities support rapid process development, whenever that need arises.

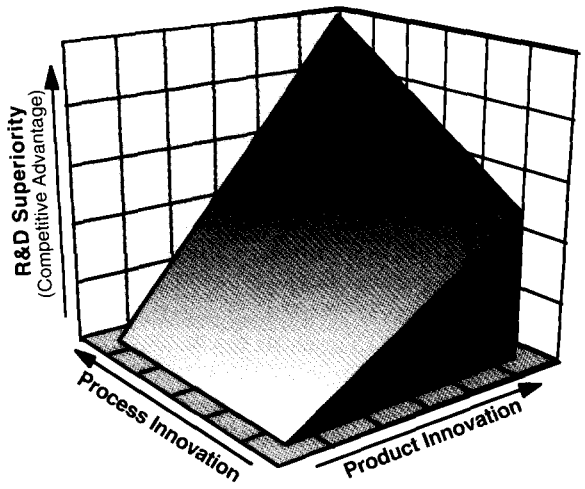

Figure 13. Integral role of process and product innovations in R&D performance.

Combinatorial chemistry may serve as a vehicle for upward advancement on this surface.

Historically, drug discovery was based on the evaluation of a single chemical entity versus a single disease target at a time. The advent of combinatorial technology changed this ratio as it is now customary to test a library of compounds against a single target. Building on the established concepts, there are opportunities for further innovations in combinatorial chemistry, probably centering first on miniturizations, advanced library design,[63] and development of new chemical technologies in solid phase. A new milestone, however, will be set on the biological side when a chemical compound library can be evaluated simultaneously and routinely against a multitude of disease targets. The terms, notations and deliberations outlined in this paper should further combinatorial processes, their cause, development, description, registry, and choice of deployment.

References and Notes

1. Beary, III, J. F. *New Medicines in Development for Aids*; Pharmaceutical Research and Manufacturers of America, 1100 15th Street, NW, Washington, DC 20005, Nov. 1995; p 14.

2. Stinson, S. C. *C & E News* 1995, 9 Oct., p 44.

3. Combinatorial Chemistry. In *Methods in Enzymology*; Abelson, J. N., Ed.; Academic: San Diego, 1996; Vol. 267.

4. Lowe, G. *Chem. Soc. Rev.* **1995**, 309.

5. Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. *Tetrahedron* **1995**, *51*, 8135.

6. Nielson, J. *Chem. & Ind.* **1994**, 901.

7. Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233.

8. Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. *J. Med. Chem.* **1994**, *37*, 1385.

9. Pavia, M. R.; Sawyer, T. K.; Moos, W. H. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 387.

10. Ellman, J. A. *Acc. Chem. Res.* **1996**, *29*, 132.

11. The advantage of commencing a drug discovery progam with a natural product, such as a secondary microbial metabolite, has ample precedent. In a further development, *combinatorial biology* holds the potential to produce novel secondary metabolites by the combination of pathways, induced by genetic manipulation of microorganisms of both marine and terrestrial origins. The resulting lead candidates would be expected to enter the IDV scale at a relatively high level. For a synopsis of combinatorial biology see: *C&E News*, 2 Oct. 1995, p 9.

12. Früchtel, J.; Jung, G. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 19.

13. Hobbs DeWitt, S.; Czarnik, A. W. *Acc. Chem. Res.* **1996**, 114.

14. Armstrong, R. W.; Combs A. P.; Tempest, P. A.; Brown S. D.; Keating T. A. *Acc. Chem. Res.* **1996**, 123.

15. Gordon, E. M.; Gallop, M. A.; Patel, D. V. *Acc. Chem. Res.* **1996**, 1144.

16. Willoughby, C. A.; Chapman, K. T. *Tetrahedron Lett.* **1996**, *37*, 7181.

17. Yun, W.; Mohan, R. *Tetrahedron Lett.* **1996**, *37*, 7189.

18. An, H.; Dan Cook, P. *Tetrahedron Lett.* **1996**, *37*, 7233.

19. Kaldor, S. W.; Siegel, M. G.; Fritz, J. E.; Dressman, B. A.; Hahn, P. J. *Tetrahedron Lett.* **1996**, *37*, 7197.

20. Jung, K. W.; Zhao, X.-Y.; Janda, K. D. *Tetrahedron Lett.* **1996**, *37*, 6491.

21. This term is also used to address ill-defined sample spaces, for example, those generated by recombinant systems with bacteriophages as vehicles.

22. It follows that A[^{20}(B)][(^{19}C)][(^{10}D)] represents a trivial partition of the set A[^{20}B][^{19}C][^{10}D] with respect to B; the expression portrays 20 mixtures, each is defined in terms of a component taken from the set ^{20}(B). Moieties C and D, however, are present in all possible combinations of 19 and 10 different building blocks, respectively, while A is invariant. The entire library contains therefore 3800 species, each of the 20 mixtures has 190 members. As a further example, the terms X[^{3}(A)][B][C] and X ^{3}(A)BC and X ^{3}(A)BC describe equally complex mixtures, but the first is based on a branched, the second on a linear and the third on a cyclic structure.

23. Jung, G.; Beck-Sickinger, A. G. *Angew. Chem., Int. Ed. Engl.* **1992**, *31*, 367.

24. Geysen, H. M.; Meloen, R. H.; Barteling, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 3998.

25. Chen, C. L.; Strop, P.; Lebl, M.; Lam, K. S. Combinatorial Chemistry. In *Methods in Enzymology*; Abelson, J. N., Ed.; Academic: San Diego, 1996; Vol. 267, pp 211–219.

26. Salomon, S. F.; Lam, K. S.; Lebl, M.; Kandola, A.; Khattri, P. S.; Wade, S.; Pátek, M.; Kocis, P.; Krchnák, V.; Thorpe, D.; Fleder, S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11708.

27. Nicolaou, K. C.; Xiao, X.-Y.; Parandoosh, Z.; Senyei, A.; Nova, M. P. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2289.

28. Morgan, E. J.; Sarshar, S.; Cargill, J. F.; Shahbaz, M. M.; Lio, A.; Mjalli, A. M. M.; Armstrong, R. *J. Am. Chem. Soc.* **1995**, *117*, 10787.

29. Houghton, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 5131.

30. Frank, R. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 425.

31. Schmidt, M.; Eichler, J.; Odarjuk, J.; Krause, E.; Beyermann, M.; Bienert, M. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 441.

32. Fodor, S. P. A.; Read, J. L.; Pirrung, M. C.; Stryer, L.; Liu, A. T.; Solas, D. *Science* **1991**, *251*, 767.

33. The descriptor part -*r*C, where *r* describes the number of components that take part in the reaction and are represented in the product, has been adopted from Armstrong's terminology which refers to an *r*-component condensation as *r*CC; see ref 14. To be included in the descriptor r, the reaction component must be presentable as a set; thus, the Strecker synthesis, although formally a 3C-condensation of amine, aldehyde and cyanide, is considered a 2C-condensation since cyanide is a unique entity.

34. Ivanetich, K. M.; Santi, D. V. Combinatorial Chemistry. In *Methods in Enzymology*; Abelson, J. N., Ed.; Academic: San Diego, 1996; Vol. 267, pp 247–260.

35. Furka, A.; Sebestyén, F.; Asgedom, M.; Dibó, G. *Int. J. Peptide Prot. Res.* **1991**, *31*, 487, and references cited therein.

36. Sebestyén, F.; Dibó, G.; Kovacs, A.; Furka, A. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 413.

37. Houghton, R. A.; Pinilla, C.; Blondelle, S. E.; Appel, J. R.; Dooley, C. T.; Cuervo, J. H. *Nature (London)* **1991**, *354*, 84.

38. Lam, K. S.; Salomon, S. E.; Hersh, E. M.; Hruby, V. J.; Kazmierski, W. M.; Knapp, R. J.; *Nature (London)* **1991**, *354*, 82.

39. Amide-bond formation is a favorite in combinatorial chemistry due to the highly developed technology of peptide synthesis. For a recent example of process development geared toward library construction employing a Wittig–Horner reaction, see Maehr, H.; Yang, R. *Tetrahedron Lett.* **1996**, *31*, 5445.

40. Pinilla, C.; Appel, J. R.; Blanc, P.; Houghton, R. A. *BioTechniques* **1992**, *13*, 901.

41. Houghton, R. A.; Appel, J. R.; Blondelle, S. E.; Cuervo, J. H.; Dooley, C. T.; Pinilla, C. *BioTechniques* **1992**, *13*, 412.

42. Dooley, C. T.; Houghton, R. A. *Life Sci.* **1993**, *52*, 1509.

43. Déprez, B.; Williard, X.; Bourel, L.; Coste, H.; Hyafil, F.; Tartar, A. *J. Am. Chem. Soc.* **1995**, *117*, 5405.

44. Reality may deviate from the underlying assumptions so that the members identified by the combinatorial product and their computed order may not be congruent with the actual ones. Practice has shown, however, that identification of the lead compounds by the methods described herein is remarkably successful.

45. With α building blocks of the type A and β building blocks of the type B one can form αβ pairs (A_jB_k) containing one building block from each set. Thus, arranging the pairs in a rectangular array with α rows and β columns, so that A_jB_k occupies the intersection point at the *j*th row and the *k*th column displays each pair only once.

46. Smith, P. W.; Lai, J. Y. O.; Whittington, A. R.; Cox, B.; Houston, J. G.; Stylli, C. H.; Banks, M. N.; Tiller, P. R. *Bioorg. Med. Chem. Lett.* **1994**, *4*, 2821. In this example, 40 acid chlorides, used in the form of ^{40}(A) and (^{40}A), are coupled in M1-2C processes, to 40 nucleophiles employed as ^{40}(B) and (^{40}B). The two required partitions are built as follows: ^{40}(A) + (^{40}B) → ^{40}(A ^{40}B) and (^{40}A) + ^{40}(B) → ^{40}(^{40}AB).

47. Pirrung, M. C.; Chen, J. *J. Am. Chem Soc.* **1995**, *117*, 1240. In this study nine alcohols (A) and six isonitriles (B) were allowed to react according to the same protocol described in ref 41 to furnish the two required partitions: (^{9}A) + ^{6}(B) → ^{6}(^{9}A B) and ^{9}(A) + (^{6}B) → ^{9}(A ^{6}B).

48. If one of the variable positions in a set, such as (^{a}A ^{b}B ^{c}C) is replaced by a constant value (e.g., C_k) we refer to the resulting mixture (^{a}A ^{b}B C_k) as a reduced set.

49. While the identification of 12 lead candidates based on the A-, B-, and C-partitions of the previous example (section 4.1, Figure 7) may appear trivial, the protocol required for the identification of several significant set members is now more involved. To illustrate this point, we return to the library (^{a}A ^{b}B ^{c}C) and assume the selection of the two most prominent cells at each selection point. This process would lead to eight candidates via two A- and four B-partitions, shown within brackets: ^{c}(^{a}A ^{b}B C) → [^{a}(A ^{b}B C_1), ^{a}(A ^{b}B C_2)] → [^{b}(A$_1$ B C_1), ^{b}(A$_{1'}$ B C_1), ^{b}(A$_2$ B C_2), ^{b}(A$_{2'}$ B C_2)] → (A$_1$ B$_1$ C_1), (A$_1$ B$_{1'}$ C_1), (A$_{1'}$ B$_2$ C_1), (A$_{1'}$ B$_{2'}$, C_1), (A$_2$ B$_3$ C_2), (A$_2$ B$_{3'}$, C_2), (A$_{2'}$ B$_4$ C_2), (A$_{2'}$ B$_{4'}$, C_2).

50. Maehr, H.; Yang, R. *Bioorg. Med. Chem.*, following article. The paper provides an example of library construction in liquid phase and illustrates lead identification and arrangement of the elements in the order of their bioactivities as suggested by the combinatorial product of a set obtained by bioassays of the last partition and two sets based on data of partitions of two reduced sets.

51. Geysen, H. M.; Rodda, S. J.; Mason, T. J.; Tribbick, G.; Schoofs. P. G. *J. Immunol. Methods* **1987**, *102*, 259.

52. Geysen, H. M.; Mason, T. J. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 397.

53. Freier et al. (ref 60) refer to the identification process via reduced sets as *synthetic unrandomization of randomized fragments* ('SURF').

54. Bioassay results of the second reduced set are assumed to be the same as those of the A-partition specified in Figure 7(A).

55. In contrast, simultaneous evaluation of all partitions (section 4.1) considered the three-dimensional sample space as the product of three coordinates.

56. If $\sqrt{n} = x$, the synthesis of (^{x}A ^{x}A) requires x^2 steps and (^{x}A ^{x}A^{x}A) requires nx additional steps. To include the B-type partition, a total of $2(n + n\sqrt{n})$ synthetic operations are needed.

57. If $\sqrt{n} = x$, the synthesis of (^{x}A ^{x}A) requires x^2 operations and (^{x}A ^{x}A P) requires n additional steps. After inclusion of the corresponding B-type partion, $2 \times 2n$ steps are needed. To obtain P_k requires the synthesis of ^{p}(^{n}N ^{n}N P) which involves n steps to reach ^{n}(^{n}N N) followed by additional p operations, so that a total of $5n + p$ steps are required.

58. The synthesis of ^{p}(^{n}N P) requires p steps, and q additional ones to reach ^{q}(^{n}N ^{p}P Q). ^{p}(^{n}N P Q_k) requires p additional steps. If $\sqrt{n} = x$, the synthesis of (^{x}A P_k Q_k) requires $x + x$ operations, and (^{x}B P_k Q_k) requires the same number of operations, so that the total number of synthetic steps is $2p + q + 4\sqrt{n}$.

59. Burgess, K.; Liaw, A. I.; Wang, N. *J. Med. Chem.* **1994**, *37*, 2985.

60. Freier, S. M.; Konings, D. A. M.; Wyatt, J. R.; Ecker, D. J. *J. Med. Chem.* **1995**, *38*, 344.

61. Wilson-Lingardo, L.; Davis, P. W.; Ecker, D. J.; Hébert, N.; Acevedo, O.; Sprankle, K.; Brennan, T.; Schwarcz, L.; Freier, S. M.; Wyatt, J. R. *J. Med. Chem.*, **1996**, *39*, 2720.

62. Pasano, G. P. *The Development Factory*; Harvard Business School: Boston, 1997.

63. Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431.

(*Received in U.S.A.* 15 *April* 1996; *accepted* 14 *November* 1996)