Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 2779–2782, April 1997
From the Academy

This paper serves as a summary of a symposium session as part of the Frontiers of Science series, held November 7–9, 1996, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Combinatorial thinking in chemistry and biology

JONATHAN ELLMAN*, BARRY STODDARD†, AND JIM WELLS‡

*Department of Chemistry, University of California, Berkeley, CA 94720; †Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104; and ‡Department of Protein Engineering, Genentech, 460 Point San Bruno Boulevard, South San Francisco, CA 94080

Every scientific and technical field has a collection of problems that are exceedingly difficult, if not impossible, to solve simply because of the sheer number of possible answers. The advent of increasingly powerful computers is now allowing some of these problems to be addressed. For example, the problem of describing and modeling turbulence in fluid dynamics (considered an intractable problem for many decades) is now increasingly amenable to computational simulation and analysis (1). Problems of similar complexity are the biochemist's attempts to computationally design molecules that bind tightly to a specific macromolecule. This goal has proven to be very difficult, because the thermodynamic and kinetic factors that determine the specificity and affinity of a binding event are extremely complex. As a result, most drug leads have been identified as a result of the random screening of biological extracts or libraries of thousands of unrelated compounds. Such methods, however, represent a relatively sparse sampling of the almost countless number of potential molecules that can be synthesized through current technologies.

Therefore, any method that accelerates the discovery of such molecules, and provides an experimental foothold for rigorous computational studies, is worthy of being described as a "Frontier of Science." The techniques described in this session, termed "combinatorial" chemistry, provide methods for the efficient synthesis and screening of libraries of related compounds with well-defined levels of diversity. These methods can be used either to generate and screen large, unbiased chemical libraries for a novel binding activity, or to create smaller, less diverse libraries of compounds that are all descended from a parental molecule with a previously determined biological activity. Combinatorial experiments are attractive to biochemists because they allow the systematic, rigorous screening of a large number of related compounds, in search of molecules that can be further optimized for specific purposes. As illustrated by the two talks in this session, combinatorial chemistry has been facilitated by the development of several technologies: (*i*) efficient methods for the parallel synthesis of many unique compounds, each produced by the coupling of individual reactants selected from large collections of related building blocks; (*ii*) DNA cloning and expression, which allows the generation of large numbers of protein or nucleic acid molecules; and (*iii*) automated hardware for the screening and analysis of the resulting libraries of compounds.

This session described the field of combinatorial thinking in two stages. First, the session built a general definition of what is meant by combinatorial synthesis, including the issues of molecular diversity and screening strategies (Stoddard). This was followed by a pair of talks describing two very different applications of combinatorial chemistry. The first (Ellman) described the synthesis and screening of small libraries of closely related organic compounds, generated from an initial parental molecule. These libraries are screened for specific enzyme inhibitors that might be candidates for new drug molecules. The second talk (Wells) built on these themes, describing the generation of large libraries of protein growth hormone mutants using genetic techniques. These libraries are screened for antagonists to signaling by the hormone receptor.

What Is Combinatorial Chemistry?

There are three common features that describe a combinatorial chemistry project (reviewed in refs. 2–9). The first is the *type* of molecules that comprise the library itself. Combinatorial libraries have been described that are composed of completely random sequences of peptides or oligonucleotides. Libraries have also been described that consist of random, site-directed mutants of a specific protein or nucleic acid oligonucleotide, and are therefore composed of many variants of an initial parental molecule. Finally, combinatorial libraries of small organic molecules can be generated by a variety of synthetic methods, leading to the synthesis and screening of a family of specific small molecules for potential utility as a drug.

In any combinatorial library, regardless of the type of molecules represented, all of the compounds are related to one another. Their structures are all built from a common set of chemical building blocks, with each molecule possessing a unique combination or sequence of these building blocks at each synthetically incorporated position. Additionally, the molecules all possess a common structural core or synthetic linkage, dictated by the type of molecules in the library and by the actual synthetic strategy employed. For example, collections of peptides or protein molecules in a combinatorial library are usually built from the 20 naturally occurring amino acids, and possess a common synthetic linkage (an amide bond) between each position in the polymeric molecule.

The second feature of a combinatorial experiment is the *diversity* that can be experimentally attained and exploited. Any library that can be encoded genetically is potentially capable of containing hundreds of millions of separate, related molecules. For example, the second talk of this session (Wells) described the screening of over 10^7 mutated variants of the human growth hormone (hGH), using recombinant DNA methods to screen each separate molecule on the surface of a unique viral clone. Because any one clone contains, in a single viral package, expressed copies of the actual molecule of interest *and* the genetic sequence encoding that molecule, the recovery of a single copy of a useful construct allows the determination of the precise sequence and structure of that molecule.

In contrast, combinatorial experiments that rely on the manual chemical synthesis of individual molecules face a more serious problem of attainable and useful diversity, as described by Jon Ellman. Unlike genetic combinatorial methods that

0027-8424/97/942779-4$2.00/0
PNAS is available online at **http://www.pnas.org**.

Abbreviation: hGH, human growth hormone.

specifically encode enormous numbers of molecular sequences in a retrievable format (i.e., the DNA sequence of viral or bacterial clones), a synthetic small-molecule library must either incorporate an interpretable, unique synthetic code that is physically associated with each molecule or alternatively the library must be designed in a "spatially addressable" manner, meaning that the chemical structure of each molecule may be inferred from its actual position in the library. Since such methods require that each individual molecular type be synthesized in a separate reaction "vessel," the resulting synthetic combinatorial libraries are usually limited to a diversity of thousands of compounds, reflecting the current limits of hardware and software in addressing individual compounds.

The third feature of combinatorial experimentation, after the design and synthesis of the library, is the *screening* process itself. The methods employed can be very diverse, ranging from chromatographic affinity selection for specific binding partners from a communal pool of all the members of the library to enzyme inhibition assays performed on each individual compound in a spatially addressable system. Finally, with the right screening procedure, any combinatorial strategy in which library diversity is created through recombinant DNA methods can be improved by a cyclical process of selection and optimization, in a manner that has been likened to "molecular evolution" in the test tube. In such a manner, initial hits can be transformed into tight-binding leads for further development.

As illustrated in two different examples during the session, combinatorial methods are now being used to select for a wide array of potentially useful molecules (Fig. 1 and Table 1). The

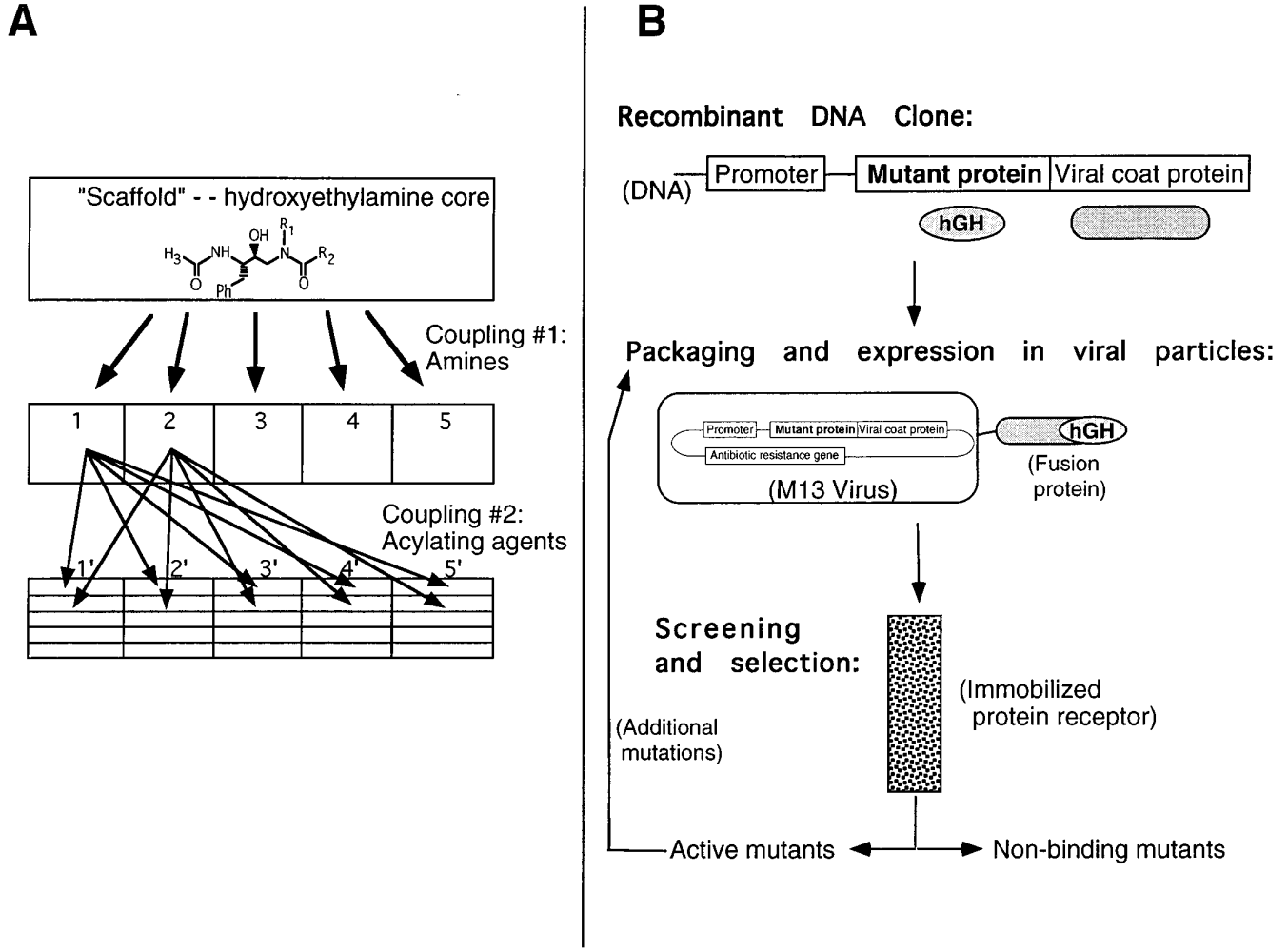

FIG. 1. Schematic illustrations of two strategies of combinatorial synthesis and screening summarized in this session. (*A*) Small molecule libraries synthesized using separate types of reactants for each variable position on the final molecule. The theoretical diversity of the library of compounds is limited only by the number of available reagents for each position on the growing molecule. In practice, however, the compounds in the library are not "encoded" with decipherable information that describes their chemical composition. Therefore, most small molecule libraries are synthesized so that each compound is "spatially addressable." For example, compounds may be synthesized while immobilized on a resin so that aliquots of the growing sets of molecules may be split at each step, resulting in a single molecular species at each position in the library. Alternatively, a unique combination of reagents may be coupled in each of many different reaction "vessels," such as wells in microtiter plates. In either case, the identity of any compound in the library is defined by its physical position in the synthetic matrix. The limiting factor is generally the ability of the experimenter and the hardware to cope with a large number of separate reaction products and the difficulty of adapting synthetic reactions to proceed to completion in a spatially addressable format. (*B*) The combinatorial synthesis or mutational variation of proteins or oligonucleotides can often be used to create a library of many millions of separate DNA, RNA, or protein molecules. Such molecules may be encoded by recombinant DNA and expressed with high fidelity (synthesized) by bacterial or viral clones containing that DNA. Therefore, any single clone possesses both the actual molecule being screened and the genetic information encoding that molecule. Enormous numbers of clones may be screened en masse; a single positive clone may be amplified by passaging that virus to regenerate all the unique information relating to that molecule. For example, in the work described by Wells, variants of the hGH protein were all produced and displayed by virus clones on their external surface, using a method called "phage (viral) display." The recovery of a single viral copy with the desired binding affinity of the mutated hormone for its receptor allows subsequent amplification and sequencing of the DNA information for that particular molecule.

Table 1. Summary of combinatorial experiments described in this session

	Cathepsin D inhibitors	Human growth hormone
Library type	Small organic molecules	Protein site-directed mutants
Synthesis method	Solid-state, spatially addressable	Recombinant DNA and viral display
Theoretical diversity	Greater than 10^9 (based on available synthetic reagents)	Greater than 20^{185} (total randomization)
Actual library size	Two libraries, 1,000 members each	Greater than 10^7
Target molecule	Cathepsin D (protease enzyme)	Growth hormone receptor
Desired effect	Inhibition of enzyme	Binding to receptor and antagonism of signalling
Screening method	Individual enzyme inhibition assays	Collective screening of recombinant virus
Optimization?	Use of directed library starting with crystal structure of enzyme complex	Selection and amplification of tight-binding viral clones

first talk described the synthesis and screening of synthetic libraries of small organic compounds, all related to a parent molecule that is capable of inhibiting a specific class of enzymes. The second talk described the generation and screening of mutant variants of a cytokine growth factor, in search of a protein molecule that acts as an antagonist of the normal hormone receptor. These two talks provided a direct contrast between synthetic and genetic combinatorial techniques. The diversity of the libraries and the methods of screening and identifying useful compounds are quite different; however, a number of unifying features are present in both projects (summarized in Table 1):

(*i*) Both projects generate and screen a huge number of molecules that might bind to a specific macromolecule target (an enzyme or receptor) with the goal of discovering novel molecules that elicit a desired therapeutic effect in the organism.

(*ii*) Both projects are designed to generate a drug lead through combinatorial diversification of an initial molecule. An important issue addressed by both investigators is whether to bias the design of the library toward a specific "privileged" structure or conduct a more random combinatorial synthesis.

(*iii*) Both projects describe a library in which the *theoretical* diversity (defined by the number of synthetically available sites and the number of building blocks to be selected from at each site) is much greater than the *actual* diversity of the library attained during screening. In both cases, a successful screen is conducted for useful molecules, indicating that sparse sampling of all possible variants using combinatorial screening is an effective strategy.

Combinatorial Synthesis and Screening of a Small Molecule Library of Cathepsin D Inhibitors (Ellman)

The identification of a number of nonpeptide inhibitors of cathepsin D, a proteolytic enzyme implicated in a number of inflammation processes, provided a good example of the issues to be considered when using combinatorial synthesis and evaluation approaches. Cathepsin D is a proteolytic enzyme (it cleaves other proteins) that induces localized increases in vascular permeability, fluid accumulation, and inflammation. Any specific inhibitor of cathepsin D could be an effective anti-inflammatory agent, as well as a potentially useful drug for several other pathogenic conditions. A number of relatively nonspecific inhibitors of the enzyme have been characterized. The experimental problem is to alter any of these compounds to make it bind more tightly, while also increasing its specificity. Stated differently, the goal is to produce an inhibitor that only binds to cathepsin D at low concentrations, while avoiding the hundreds of similar enzymes in the body. The one-by-one synthesis of individual variants of these inhibitors would be a slow and costly endeavor. However, methods have been developed by the Ellman Lab (ref. 10 and E. K. Kick, D. C. Roe, A. G. Skillman, G. Liu, T. J. A. Ewing, Y. Sun, I. D. Kuntz, and J.A.E., unpublished work) that describe how to create a "scaffold" or precursor of an inhibitor with several positions capable of coupling to many different chemical groups of similar reactivity, but different structure (Fig. 1*A*). This is an excellent problem for combinatorial methods, because the investigators need to sample a large number of variations of a specific molecule through many different combinations of chemical groups.

Rather than conducting a "random" combinatorial synthesis, diverse chemical groups were incorporated at two specific positions on a hydroxyethylamine molecule (Fig. 1), which is a stable analog of a reaction intermediate formed by cathepsin D. Amines were used to introduce diversity at one site on the scaffold, and acylating agents, such as carboxylic acids and sulfonyl chlorides served to introduce various groups at an additional site. Exhaustive combination of all commercially available amines and acylating agents would provide a library of over 10 billion compounds. While a library of this size could theoretically be prepared, it would require that thousands of building blocks be introduced at each position resulting in considerable expense and effort in both synthesis and evaluation.

The Ellman group, in collaboration with Irwin Kuntz and coworkers at University of California, San Francisco, chose to design two smaller libraries. Two alternative computational methods were used to select the chemical groups to be combined and displayed. A computational program designed to maximize diversity was used to select the building blocks for one library (the diverse library). Structure-based design, using the crystal structure of cathepsin D, was used to select the building blocks for the second library (the directed library).

The chemical reactions were optimized for solid-phase synthesis and both libraries, each containing 1000 separate compounds, were prepared by parallel synthesis (Fig. 1*A*). The key feature of this strategy was the development of useful, highly reactive leaving groups, incorporated at specific positions on the scaffold of the core molecule. These groups ensure uniform, high yields of coupling with many different compounds, so that the final product at each position in the library is of sufficient amount and purity to be directly assayed without further purification. The ability to successfully optimize and exploit this strategy has been the critical step in creating and screening combinatorial libraries of small molecules. Spatial addressing was achieved by synthesizing the compounds on plastic pins in different reaction vessels. Each compound in the libraries was then screened for inhibitory activity against cathepsin D with the assay performed simultaneously in the reaction vessels. The directed library yielded a "hit rate" of 6–7% at inhibitor concentrations of 1×10^{-6} M, with the most potent compound having a relatively tight inhibitor dissociation constant of 78×10^{-9} M. The diverse library provided

inhibitors >4-fold less potent, indicating the superiority of the directed library. A number of more potent inhibitors of cathepsin D were then rapidly identified by synthesizing and screening a small second library that explored variants of the most active compounds. It is anticipated that these general methods will be extended to many other enzyme targets.

Combinatorial Mutagenesis of hGH (Wells)

The second talk of this session provided a contrast to the synthesis of small organic molecule libraries by illustrating the combinatorial production and screening of several million variants of the hGH protein. The purpose of this project is to isolate a mutated version of hGH that can bind tightly to the hormone's receptor and block its signaling activity, thus acting as a therapeutic agent for certain types of hypergrowth disorders. A powerful method for isolating peptides or proteins with improved or novel binding properties for a target receptor or protein is called phage display (for reviews see refs. 11 and 12). In this method, large numbers of mutated proteins or peptides (exceeding 10^7 variants) were displayed on the surface of a virus called M13 (Fig. 1*B*). Each virus in the library also contains the gene that encodes protein variant linked to an otherwise normal protein of the virus coat. By successive rounds of affinity chromatography against the target receptor, any virus that displays a fusion protein having improved binding affinity to the hormone receptor can be sorted from those that encode weaker binders (Fig. 1*B*). After several rounds of sorting, viruses that display fusion proteins with improved binding properties were cloned and their corresponding genes sequenced to identify the fusion protein sequence.

Even with the ability to screen 10^7 variants the entire surface of hGH could not be randomized, because greater than 20^{185} variants would have to be screened for exhaustive analysis. To limit the library to functionally important variants, high-resolution mutational and structural studies of hGH were use to identify regions of the hormone that are important for binding (13). These studies indicate that there are two binding sites on hGH (called Sites 1 and 2) that sequentially associate with two receptors to form an active receptor complex. Analogs of hGH that are potent antagonists have been produced by mutating residues in Site 2 so that the hormone can only form a 1:1 complex with the receptor, blocking association with a second receptor molecule and therefore blocking signaling. The Wells group reasoned that by improving the binding affinity at Site 1 they could further improve these analogs as antagonists.

Twenty of the residues in Site 1 were randomly mutated, in groups of 4 residues at a time (5 separate libraries of viruses displaying a total of 4^{20} different hGH molecules each), so as to search exhaustively for the best binding solution (13). From each library, hormone mutants were isolated that were improved by 2- to 8-fold in their binding affinities for the extracellular domain of the hGH receptor. By simply combining these mutants, the affinity was further improved in a nearly additive fashion by up to 400-fold. Virtually all of the enhanced affinity was the result of a decrease in off-rate from the receptor. Thus, hormone variants can be produced that bind with higher affinity to their native receptors. When this mutant was combined with one that cannot bind a second receptor, the resultant was a much better antagonist of the hGH receptor. This or similar antagonists may be useful in the treatment of diseases involving hGH excess, such as acromegly.

Conclusions

The clearest message from this session was that the field of combinatorial chemistry is benefitting from rapid advances in efficiently synthesizing large numbers of related compounds that differ in the order of combination of specific chemical building blocks. The field has diverged into two challenging areas, using either synthetic organic techniques to generate biased libraries of compounds related to known therapeutic leads or using recombinant DNA to generate similar, but much larger, collections of biologically active polymers such as peptides, proteins, or strands of DNA or RNA. In either case, the critical technical advance is the ability to depart from the synthesis and analysis of individually prepared molecules (for example, a single site-directed protein mutant or the multi-step synthesis of a single organic molecule), and instead develop methods to generate many compounds in parallel or in a single batch. If a screen is also developed that can give an accurate positive signal in the presence of a very small amount of an active compound, and if the chemical structure of that compound can be either decoded or deduced from the design of the library, then such experiments can systematically search for a specific chemical activity among collections of compounds that might otherwise never have been created by human hand.

It might be expected that further developments in this field will be driven by technologies that allow investigators to address increasingly large libraries, possibly by the presentation of libraries in spatial arrays using novel surface chemistries. For example, with current lithographic methods, an investigator might be able to synthesize greater than 2^{16} different compounds in an immobilized 1-cm^2 array, allowing direct screening of the library in a single step and in a very miniaturized format (14). Such technologies may eventually allow large combinatorial libraries to be distributed on chips for chemical screens tailored to specific needs. As with the other sessions on the *Frontiers of Science*, what recently seemed impossible is now conceivable, and may soon be routine.

We gratefully acknowledge the comments and revisions provided by Tom Alber and Nicholas Cozzarelli during the preparation of this summary.

1. Moin, P. & Kim, J. (1997) *Sci. Am.* **276** (1), 62–68.
2. Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. A. & Gordon, E. M. (1994) *J. Med. Chem.* **37,** 1233–1251.
3. Gordon, E. M., Barrett, R. W., Dower, W. J., Fodor, S. P. A. & Gallop, M. A. (1994) *J. Med. Chem.* **37,** 1386–1401.
4. Scott, J. K. & Craig, L. (1994) *Curr. Opin. Biotechnol.* **5,** 40–48.
5. Houghten, R. A. (1994) *Curr. Biol.* **4,** 564–567.
6. Martin, E. J. (1995) *J. Med. Chem.* **38,** 1431–1437.
7. Uphoff, K. W., Bell, S. D. & Ellington, A. D. (1996) *Curr. Opin. Struct. Biol.* **6,** 281–288.
8. Service, R. F. (1996) *Science* **272,** 1266–1267.
9. Thompson, L. A. & Ellman, J. A. (1996) *Chem. Rev.* **96,** 555–600.
10. Kick, E. K. & Ellman, J. A. (1995) *J. Med. Chem.* **38,** 1427–1430.
11. Smith, G. P. (1991) *Curr. Opin. Struct. Biol.* **1,** 668–673.
12. Clackson, T. & Wells, J. A. (1994) *Trends Biotechnol.* **12,** 173–184.
13. Wells, J. A. & de Vos, A. M. (1996) *Annu. Rev. Biochem.* **65,** 609–634.
14. Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. A. & Gordon, E. M. (1994) *J. Med. Chem.* **37,** 1233–1251.