Combinatorial chemistry: 20 years on...

'The mass of a protein library...would exceed that of the universe by more than two hundred orders of magnitude.'

Árpád Furka, Dept of Organic Chemistry, Eötvös Loránd University, Hungary

In traditional pharmaceutical research, drug candidates were prepared and tested one by one. The philosophy behind the combinatorial approach brought about a revolutionary change in this methodology and can be best expressed by the question: Why deal with a single compound when you could prepare and test thousands or millions of compounds? However, new synthetic and screening methods were needed to apply this philosophy. The development of one of the most efficient combinatorial synthetic methods began 20 years ago in our laboratory at the Eötvös Loránd University, Budapest.

The roots

During 1964–1965, I was a postdoctoral fellow at the University of Alberta, Canada, and participated in the determination of the amino acid sequence of a pro-enzyme, chymotrypsinogen-B. After returning to Budapest, I wondered from how many possible structures was the correct sequence chosen? Given that the protein contained 245 amino acid residues and that each position could be one of the 20 natural amino acids, the number of possible sequences was 20^{245} ($= 5.65 \times 10^{318}$). This certainly seemed to be a very large number, but to really realize its magnitude we had to compare it with an entity that was also very large. Finally, I found an estimate of the mass of the universe based on an Einstein formula. The comparison showed that the mass of a protein library, in which each sequence is represented by a single molecule, would exceed that of the universe by more than two hundred orders of magnitude. Initially, this immense diversity of molecules shocked me; however, it later induced a speculation about the possibilities of preparing such peptide libraries.

Although the number of components was far from being as huge as the number of protein sequences, it was still too large to consider synthesizing such libraries using techniques available at that time. By 1980, we were generally still preparing one peptide at a time using solution-phase or solid-phase methods with an elongation rate of one amino acid per day. Based on this rate, the synthesis of all the possible 160,000 tetrapeptides, for example, would have taken around 1800 years of uninterrupted work. This time could have been reduced to a still-unrealistic 460 years by optimization of synthesis by splitting the products of each coupling step according to the branches of a combinatorial tree.

The split–mix method

The first idea that was considered for synthesizing peptide libraries was the use of amino acid mixtures in couplings. It was clear, however, that as a result of such couplings the products would form in unequal molar quantities as a consequence of the differences in the reactivities of each amino acid. The differences in molarities would be amplified in each successive coupling step, leading to a mixture with uncertain composition. For this reason, the idea of amino acid mixtures was abandoned. However, the problem was finally solved by splitting the resin into portions and performing the couplings with single amino acids. The split–mix method [1,2] enabled the production of peptide mixtures that contained millions of components. Such mixtures, however, appeared unacceptable in the conventional drug discovery practice where single compounds were used in pure form. Hence, there was an urgent need to also present an efficient strategy for identification of the bioactive substance that might be present in the complex synthetic mixture. This task, however, was comparable to finding the proverbial needle in a (huge) haystack. Nevertheless, a theoretical solution was developed in a very short time. I called it 'synthetic back-searching strategy', which later proved to be, in principle, identical to the 'iteration strategy' published by others [3].

The first description of the method

I was fully aware of the important role that the combinatorial approach might play in pharmaceutical research, but one of the leading Hungarian pharmaceutical companies I contacted showed no interest at all. In addition, the split–mix method was considered by patent attorneys only

1359-6446/02/$ – see front matter PII: S1359-6446(02)02085-2

as a potential research tool and, for this reason, was judged not to be patentable. They suggested, however, that I should describe the method in a document and, in the event of future priority disputes, to notarize it [4]. I did this, and the document (written in Hungarian), in which the principles of combinatorial chemistry, including both synthesis and screening were first clearly explained, was notarized in May 1982. The photograph of the first and last pages of this document is shown in Fig. 1, and an abbreviated form of the English translation can be found in Box 1.

The use of combinatorial methods

The use of combinatorial chemistry techniques began to spread in the early nineties and the field has continued to explode. New powerful combinatorial synthetic and deconvolution methods have been published, such as the production of phage-displayed peptide libraries [5] and the light-directed synthesis of libraries on a glass surface [6]. The earlier-developed parallel synthesis [7], although not as powerful as the real combinatorial methods, became popular also, particularly when different automatic machines were developed and became commercially available.

Developments in the synthetic side were paralleled by the appearance of the high-throughput screening methods and automation of analytical characterization methods. Handling and storage of huge numbers of compounds, along with their analytical and screening data, also required the development of computerized data management methods. Software was developed to aid library design, including databases of available starting materials and generation of 'virtual libraries' and filters to select promising structures and reduce the number of compounds to be

Box 1. English translation of original notarized document (abbreviated version)

Study on the possibilities of systematic searching for pharmaceutically useful peptides

The proposal outlined here constitutes a research project that makes possible a search for biologically active peptides with much greater chance than before. As I write this project I am fully aware of its potential importance in industry.

The essence of the proposal is that instead of synthesizing peptides seperately, peptide mixtures are prepared that contain several hundred or several thousand peptides in approximately 1:1 molar ratio, and these peptide mixtures are then submitted to screening tests. It will be shown that significant time and labour can be saved both in the synthetic work and in the screening experiments.

Method for synthesis of peptide mixtures

As mixtures of peptides rather than single peptides are synthesized, post-synthetic purification and removal of by-products is not possible. The solid-phase method must therefore be used for the synthesis of peptide mixtures.

Realization of the synthesis

The resin is divided into k_1 equal portions (i.e. into the same number of portions as the intended number of C-terminal variations). Each portion of resin is then coupled with one of the k_1 types of amino acids followed by removal of the amino-protecting group from each sample. A small quantity is removed from each sample and put aside for later use, and then the remaining samples are thoroughly mixed. The mixture of aminoacyl resins is then divided into k_2 equal portions and each one is coupled with one of the k_2 types of protected amino acids followed by removal of the amino-protecting groups from each sample. Before mixing, small samples are removed and taken aside as before. The mixture of dipeptides is cleaved from a small portion of the mixed resin for use in biological tests. The rest of the mixed resin is divided into k_3 equal parts and the amino acids intended to occupy the third position are coupled to them. The synthesis is likewise continued until the mixture of *n*-residue peptides is reached.

There are several points that should be noted. The amount of labour required could be significantly reduced by using mixtures of properly protected amino acids in acylation reactions. This, however, does not seem to be an acceptable solution because of the differences in reactivity of the activated amino acids, which would lead to the formation of peptides in significantly different concentrations, thereby causing problems in the screening experiments. Formation of peptides in equal concentrations can only be assured by mechanical mixing of samples followed by their division into equal portions. This makes possible a complete conversion for every amino acid component.

The number of peptides (N_n) formed in the synthesis (that is, the number of components in the peptide mixtures) can generally be calculated by the following formula:

$$N_n = k_1 k_2 .. k_{n-1} k_n \quad (1)$$

where k_1, k_2,... are the numbers of amino acids used in couplings.

The number of synthetic steps (S_n) in the synthesis of a peptide mixture containing N_n peptides (considering the attachment of the first amino acid to the resin as a separate step) is:

$$S_n = k_1 + k_2 + \ldots + k_{n-1} + k_n \quad (2)$$

The formulae show the advantage of synthesizing peptide mixtures: the number of the synthetic steps can be calculated by the sum of the numbers of the varied amino acids, while the number of peptides is given by the product of the numbers of the varied amino acids.

As an example, synthesis of a mixture of tetrapeptides prepared by variation of 20 different amino acids, needs

synthesized to a reasonable level, a particular requirement of the low-efficiency parallel synthetic methods. The problem with software currently used in library design is that the rules of molecular interactions are far from being understood. It is well known, for example, that the shapes of molecules are of primary importance. Every molecule can have a multitude of shapes and the software cannot take all of them into account, particularly if they are deduced from 2D structural formulae. We can hope, however, that the huge amount of data accumulating from the testing of combinatorial libraries will eventually help to better our understanding of how the molecular interactions depend on structure. Indeed, it would be very useful if all data produced by the various companies could be collated to form a large database that was then made available for scientific analysis. However, it would of course be equally important to preserve the interests of the companies who supply the data.

In most cases, libraries are designed for a single target. In the near future, the success of the Human Genome Project will make available thousands of new targets. For this situation, different approaches would also be worth considering, such as exploiting the power of combinatorial synthetic methods [8,9] for the synthesis of large libraries, and testing all compounds against all potential targets. To achieve this, faster and cheaper screening methods need to be developed.

In contrast to libraries applied in drug discovery, where the components are prepared by a combination of molecular building blocks, there is a large class of important materials where the properties depend on the proportion of mixed ingredients and on the conditions of preparation.

only 80 synthetic steps! It should be noted that in the same run, all possible shorter peptides – that is, the 400 dipeptides and the 8000 tripeptides – are also formed. By contrast, the traditional synthesis of these peptides would require 168,400 synthetic steps.

Screening of peptide mixtures: 'backsearching' for the active peptide

If the peptide mixture is found to contain an active component (that is, if the mixture shows a new biological effect), the next step is the isolation and structure determination of the active peptide followed by its synthesis.

Backsearching step No. 1

The experiment begins by selecting the k_n samples that were produced in the synthesis of the *n*-residue peptides before the final mixing step. The mixtures of *n*-residue peptides are then cleaved from each resin sample. The mixtures of peptides differ from each other in the *n*-th (that is, the N-terminal) residue of their component peptides. Each peptide mixture is submitted to a quantitative activity determination, which shows how activity varies depending on the terminal amino acid residue. Hence, this way we can determine the N-terminal residue of the active peptide and, in addition, show the effect of its replacement by other amino acid residues. Let's suppose, for example, that the N-terminal residue in the sample showing the highest activity (as well as in the active peptide) is Phe. It should be noted that if there are several samples showing equally high activity, it is practical to choose the cheapest or the synthetically less-problematic amino acid as the N-terminal residue of the active peptide. This point also applies to the subsequent backsearching steps.

Backsearching step No. 2

The k_{n-1} samples produced in the synthetic stage of the (n–1)-residue peptides are then selected. The amino acid determined before (Phe in our example), is coupled to each sample. Cleavage of the peptides from the support gives k_{n-1} different peptide mixtures. Their common feature is that every peptide has Phe in the N-terminal position. By submitting the peptide mixtures to quantitative screening experiments, one can determine the amino acid residue occupying position n–1 (that is, the pre-N-terminal position) in the active peptide. This experiment also shows the effect that a single amino acid has on the overall activity. For example, let's suppose that the pre-amino-terminal amino acid is Arg. It should be noted that in this backsearching step, the Phe is coupled to k_{n-1} samples and the same number (k_{n-1}) of screening experiments need to be performed. Consequently, the number of synthetic steps and the number of screening experiments are the same, k_{n-1}. It should also be noted that in the previous backsearching step, only screening tests are done (their number is k_n) and no synthetic steps are needed.

Backsearching step No. 3

We then take the k_{n-2} samples created in the synthesis of (n–2)-residue peptides. Each sample is coupled first with protected Arg and then with protected Phe. After cleaving the peptides from the support, each of the k_{n-2} peptide mixtures is submitted to activity tests to determine the amino acid residue occupying the third position counting from the N-terminal end. The number of screening tests that need to be executed is k_{n-2}. The number of the required synthetic steps is $2k_{n-2}$.

Extension of the method to other types of compounds

Applicability of the method outlined here is not restricted only to systematic searching for active peptides. The same principle applies to all other sequential types of compounds, that is, when the compounds differ from each other only in the nature or sequence of the building blocks. Among them might occur natural compounds such as oligosaccharides or oligonucleotides, as well as synthetic products, such as sequential copolymers or sequential polycondensates.

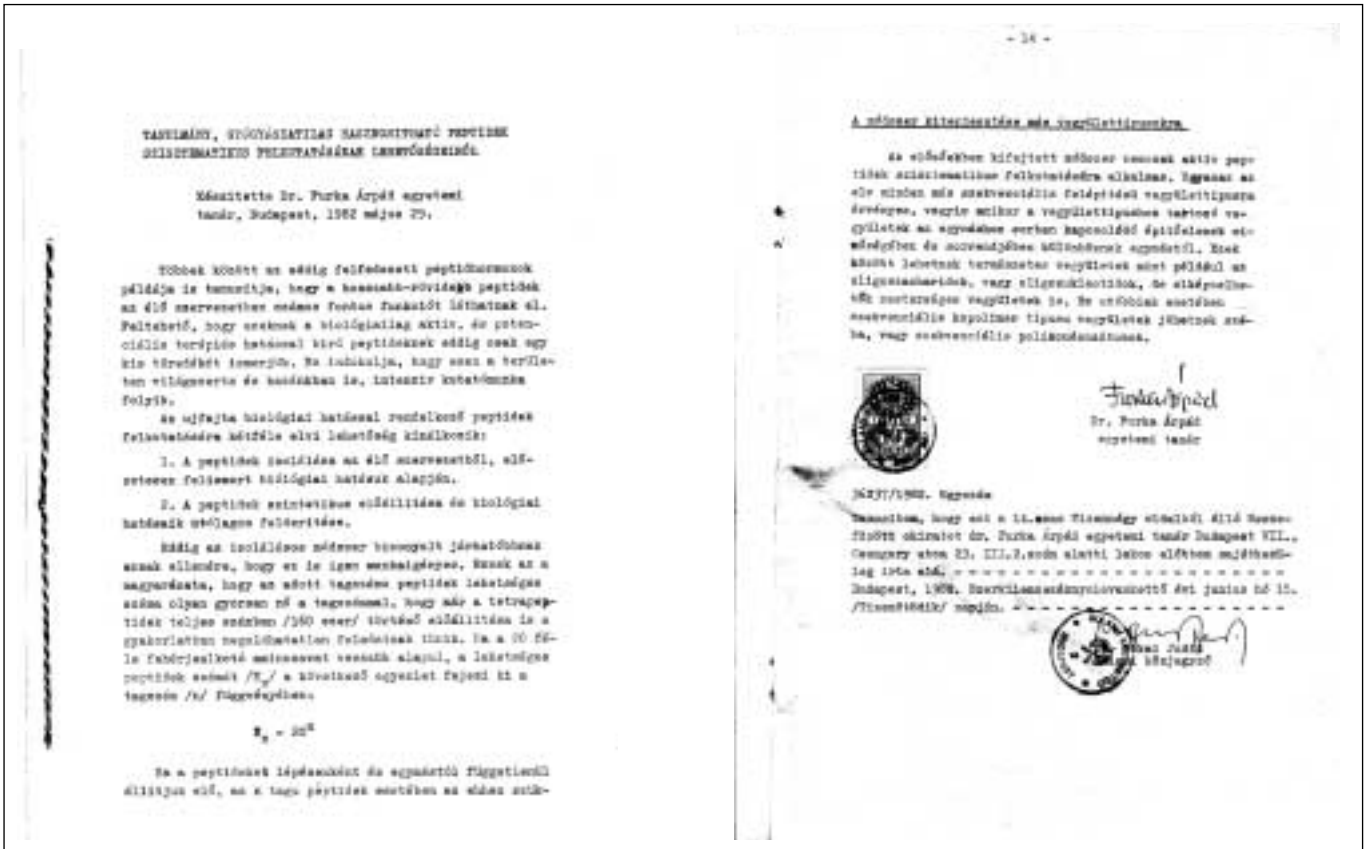

Figure 1. Photograph of the first and last pages of the original document [4] notarized in 1982 describing the principles of combinatorial chemistry and written in Hungarian. For a complete version in Hungarian and translated English, please see http://archive.bmn.com/supp/ddt/CCHUN.pdf and http://archive.bmn.com/supp/ddt/CCENG.pdf.

The principle of the systematic experiments that are expressed in the title of the original 1982 document have been similarly applied in this area with some success [10]. Extension of this principle to some areas of biology is already fact and further extension of combinatorial methods to other fields can be expected in the future.

References

1 Furka, Á. *et al.* (1988) Cornucopia of peptides by synthesis. In Highlights of Modern Biochemistry, *Proceedings of the 14th International Congress of Biochemistry,* VSP (Vol. 5), pp. 47, Utrecht, The Netherlands

2 Furka, Á. *et al.* (1991) General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Peptide Protein Res.* 37, 487–493

3 Houghten, R.A. *et al.* (1991) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* 354, 84–86

4 Furka, Á. (2001) Visszapillantás a kombinatorikus kémia korai időszakára. *Magyar Kémikusok Lapja* 56, 250–254

5 Scott, J.K. and Smith, G.P. (1990) Searching for peptide ligands with an epitope library. *Science* 249, 404–406

6 Fodor, S.P.A. *et al.* (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773

7 Geysen, H.M. *et al.* (1984) Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc. Natl. Acad. Sci. U. S. A.* 81, 3998–4002

8 Nicolaou, K.C. *et al.* (1995) Radiofrequency encoded combinatorial chemistry. *Angew. Chem. Int. Ed.* 34, 2289–2291

9 Furka, Á. *et al.* (2000) The string synthesis. A spatially addressable split procedure. *J. Comb. Chem.* 2, 220–223

10 Schultz, P.G. (1995) A combinatorial approach to materials discovery. *Science* 268, 1738–1740

Árpád Furka
Dept of Organic Chemistry
Eötvös Loránd University
H-1518 Budapest 112, Hungary
tel: +36 1209 0555
fax: +36 1372 2620
e-mail: afurka@szerves.chem.elte.huI