CHIMIA 2001, 55, No. 3

Chimia 55 (2001) 114–118

ISSN 0009–4293

Towards Random Polypeptide Synthesis

Richard M. Thomas[a]*, Jan W. Vrijbloed[b], and Pier-Luigi Luisi[a]

Abstract: Modern naturally occurring proteins have been produced by a lengthy selective evolutionary process. While, in general, they are all composed of the same 20 amino acids there is a distinct bias in their average amino acid composition. This bias may have arisen due to evolutionary mechanisms, the degeneracy of the genetic code, the primordial availability of suitable monomers, their relative reactivity or a number of other, equally speculative, causes. Mathematics appears to dictate that Nature could not have sampled all possible amino acid sequences and selected the most suitable for a particular function, suggesting that the proteins observed today may have evolved from a relatively small number of precursors. If this is true it would imply that there is a vast set of possible proteins that have simply never existed and that may possess interesting or useful properties. This article investigates whether the structural space occupied by proteins that do not currently exist can be sampled. One approach suggests itself – random polypeptide synthesis in which all possible residue types are inserted at all possible positions of an amino acid sequence of a given length. It is abundantly clear that the truly random synthesis of even a small set of such protein sequences is precluded by simple mathematics. The issues that this raises are discussed and different practical approaches to the problem described.

Keywords: Origin of life · Phage display · Protein domain · Random polypeptide sequences

Introduction

The synthesis of libraries of small peptides with random amino acid sequences and the use of techniques in which pre-existing protein structures are subjected to systematic random mutation have become relatively commonplace. Methodology of this kind has been used in the search for novel pharmaceuticals, for the design of enzymes with altered catalytic and stability properties and in the production of protein libraries used for screening purposes. All of these activities can be described as 'directed' in that randomization is essentially being used for the design of novel molecules with specific properties, however diverse. The concepts to be described here are, in many ways, the antithesis of design in that the aim is to produce peptide/protein molecules in as random a way as possible with no preconception as to what their properties might be. An operational discrimination between the terms 'polypeptide' and 'protein' is of importance here: clearly all proteins are polypeptides, but what sets them apart, irrespective of their functional properties, is their ability to fold spontaneously to form a defined and specific three-dimensional structure. This is the definition that will be employed here and the potential and an exploration of the use of random synthesis to produce molecules that fold ('minimal proteins') will made with this as the only criterion for success. If specific activities can be found within the folded population, it will be seen as a bonus.

The stimulus to investigate this problem comes from two different lines of thought. On the one hand there is the field of the theories of the origin of life on earth and, more particularly, that of the origin of biological macromolecules and biopolymers. Mechanisms by which long polypeptide chains of, say, hundreds of residues could have arisen from a mixture of amino acids are unknown and cannot, therefore, be reproduced in the laboratory. On the other hand there is the simpler question that explores the realm of possible proteins that do not, however, exist within the envelope of the proteins observed in living organisms today. This discussion will concentrate on the latter aspect largely escaping many of the problems and circularities imposed when the concept is viewed from a biological standpoint.

In its simplest form, the problem can be resolved into four basic questions which will be posed and then considered separately:

- What restrictions are there on the size of possible proteins? In the current context this can be translated into 'what is the minimal size required for a polypeptide to fold?'.
- Which amino acids should be employed in the construction of random polypeptides?
- How random is 'random'?

*Correspondence: Dr. R.M. Thomas[a]
Tel.: +41 1 632 55 40
Fax: +41 1 632 10 73
E-Mail: rthomas@ifp.mat.ethz.ch
[a]Institute for Polymers
ETH-Zentrum
CH-8092 Zürich
[b]Department of Chemistry
University of Zürich
Winterthurerstr. 190
CH-8057 Zürich.

Once these questions have been addressed, our current experimental approaches for the investigation of the problem will be described.

Does Size Matter?

Peptide and protein structures form a continuum in terms of sequence length or molar mass. At the low end of the scale, small peptides, which for current purposes will be considered as structurally random, give way to folded structures as the sequence length exceeds 30–40 residues. The proteins found in this range can be loosely described as 'binding proteins', that is, they either bind ligands or to a target and, in this category are to be found hormones, inhibitors and small metal-binding proteins. One of the smallest proteins to have had its crystal structure solved, the avian pancreatic hormone APP has 36 amino acid residues [1]. As the sequence length increases into the 100+ range the first small enzymes, such as RNAse and lysozyme appear. The realm of monomeric proteins now starts to overlap with that of multimeric proteins and many proteins above a mass of, say, 50 KDa are composed of multiple subunits. The scale is effectively open ended and eventually leads to the area of massive protein aggregates such as ribosomes. Although clearly a very broad generalization, folded monomeric proteins can therefore be considered to lie approximately in the 5-50 KDa mass range. For practical purposes it is the lower limit that is of interest and it would appear that, in order to be able to encode sufficient information for the formation of a defined three dimensional structure, a peptide must be of the order of 40 residues long. It is at this sequence length that the transition from 'polypeptide' to 'protein' may be considered to occur. It should be noted that this description also fits the concept of a protein domain, that is, an autonomously folding unit within a protein and the formation of domains alone, without any apparent function, will fit current requirements. Searching the available databases reveals that there are probably of the order of hundreds of proteins in the 30–50 residue class although most have never been isolated. There are thought to be somewhere in the range of 10^5–10^{10} proteins of all sizes that do exist but this depends very much on the criteria used in such predictions. For the purposes of the random synthesis of folded units it will be assumed that the sequence must be at least forty residues long.

Which Amino Acids Should Be Used?

Inquiring as to whether peptides and proteins could arise by chance polymerization in a primordial scenario requires speculation into which amino acids may have been available as well as their concentrations and reactivities. In modern terms there are 20 proteogenic amino acids, that is, amino acids that are commonly found in proteins, and they are all members of the α-amino series and are all L-configured. There are, however, many other α-L-amino acids that are found in the free state as metabolites or metabolic intermediates, a range of 'unusual' amino acids encountered in 'lower' organisms as well as amino acid derivatives that arise as a result of modification of their original structure. It is reasonable to enquire into the apparent absence of, for example, β-amino acids and members of the D-series (although some occur in bacteria) and there appears to be no reason to exclude these species from the set of primordial monomers. Indeed D-proteins have been synthesized and shown to possess activity [2]. Attempts have been made to reduce the set of possible ancestral amino acids by extrapolation of the modern set backward in time using a variety of arguments [3]. The degeneracy of the genetic code offers some support for such exercises although, depending on the criteria selected, differing sets of amino acids emerge. Classic experiments aimed at simulating conditions pertaining at primordial times suggest that only a restricted set of amino acids may have been present [4] although, again, this requires a total knowledge of initial conditions, which is unavailable. As will be seen, the use of a small set of amino acids greatly reduces the size of the set of polypeptides of a given length that may be synthesized. One way that such a restriction can be accomplished is to divide the amino acids up into classes (hydrophobic, anionic, *etc.*) and then to select a restricted set of members member to represent the class. However, this can be predicted to lead to a predetermined set of polypeptides, some of which might indeed fold, simply by an analysis of the possible patterns within the sequences that might be produced. An approach of this kind clearly suffers from a loss of randomness but has led to considerable success [5].

Selection of the set of amino acids to be used can therefore be performed on the basis of a variety of criteria. In the absence of any strong reason to prefer any one of these over the others it will be taken as an observable fact that the proteins that exist today consist of a particular set of 20 amino acids and that these will be used for experimental purposes. While this may be seen as arbitrary, such an approach presumably increases the probability of what may be termed 'biocompatability', that is the products of a random synthesis may be susceptible to detection, purification *etc.* using pre-existing biological systems, and also means that there is a heightened chance that the synthetic products might exhibit determinable activity. The distribution of proteogenic amino acids in modern proteins is noticeably biased [6] and there are several possible explanations for this which will not be considered here. As the conformational properties of the individual amino acids are well-understood, it can be argued that certain distributions of monomer types in combination are required to produce the distribution of structure observed within proteins. It is, however, hard to predict that this is the case, *a priori*, due to the bias in composition seen in proteins of known structure.

How Random is 'Random'?

The kind of calculation that follows has been performed many times and the results are usually taken, quite reasonably, to indicate that the random synthesis of proteins of any useful size is beyond contemplation. If a peptide of forty residues is taken as a starting point, there are 20^{40} (~ 10^{52}) possible different products that could appear with an average molar mass of about 5 kg. If only one molecule of each of these were to be synthesized, approximately 10^{28} moles of material with a total mass of 5.10^{28} kg would be produced. This quantity corresponds to ~10^4 times the mass of the earth. Looked at another way, if this set of peptides could be synthesized at a rate of 10^6 molecules per second, it would take ~10^{44} years to complete the synthesis. The 'real' world would probably be better modeled by assuming an average sequence closer to, say, 150 amino acid residues and there are ~10^{200} possible sequences of this length. At this point such calculations are normally abandoned. However they raise several important issues and possibly provide pointers to ways in which a random synthesis might be achieved. The numerical problems are illustrated in Fig. 1.

The most striking conclusion is that Nature has had neither the time or requisite amount of material to produce all

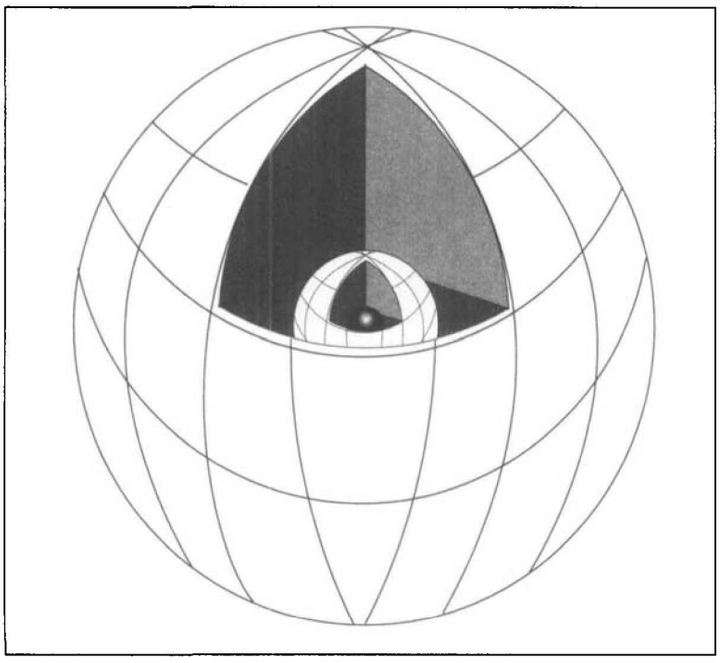

Fig. 1. A diagrammatic representation of the numerical problem in random polypeptide synthesis. The large outer sphere represents the space occupied by all possible 40 residue amino acid sequences (~10^{52}). Going towards the center, the next sphere is the space of all possible 40-mers constructed using the natural occurring frequencies of the 20 amino acids (~10^{40}). The innermost sphere represents the current upper limit on estimates of the number of proteins of all sizes that actually exist (~10^{10}). The radii of the three spheres are drawn on a logarithmic scale; were this not the case, the two inner spheres would simply vanish.

possible proteins and select from them those that are best suited for particular tasks. This in turn strongly suggests that all proteins observed today have arisen from a small set of precursors. It is also clearly possible that the precursors themselves may have arisen by the fusion of smaller, independently folding entities. If we take the estimate of 10^{10} as being the number of proteins that do exist, they represent less than noise within the space of all possible proteins (Fig. 1). Furthermore, the biased composition and relatively narrow size distribution (at least of monomeric forms) of current proteins suggests that they are sequestered within a small region of this space.

In order to try to see how a random synthesis could be accomplished, a couple of points should be made. Firstly, it is obviously not necessary or possible to synthesize all the members of a random permutation – any single sequence of the set is as random as any other. This leads naturally on to the question of statistical analysis. Here, a complete synthesis of all 'randomers', although impossible, would lead to a full statistical answer to questions of the type 'what proportion of all possible 40-mers are folded' and so forth. In practice, what is required is some estimate of the number of sequences that would be have to be synthesized in order to provide a reasonable statistical description. Unfortunately, as there is no method which will predict whether a protein is folded or not (other than, perhaps by analogy with known proteins of known structure) it is impossible to provide reasonable *a priori* answers to this point. Indeed, it would seem that the only way to arrive at some conclusion on points of this nature is to attack the problem empirically. Faced with this situation, methodology needs to be found that can not only produce as many sequences as possible but that can do so in such a way that the products are amenable to structural analysis.

Experimental Approaches

Three potential strategies for the exploration of the possibility of forming folded polypeptides on a random basis can be identified:

- Classical peptide synthesis
- Molecular biological techniques
- Computational methods

The third approach, the most attractive, can be ruled out at present as it suffers from the insurmountable drawback that it is currently impossible to predict the folded structure of a protein given its primary sequence alone.

It is perfectly feasible to construct peptides with randomly selected sequences using available chemical synthesis methodology. Here, however, the problem is of numbers and, as already discussed, it is not possible to predict the size of sample of random peptides that would be needed to be produced in order to have a statistically reasonable likelihood of finding folded molecules. Modern equipment permits the synthesis of approximately 100 peptides at a time in separate reaction vessels, and while this could be repeated many times, it would take considerable investment to even begin to approach the synthesis of the numbers of molecules that can be predicted to be required. Additionally, screening of the products for 'foldedness' would become a formidable task. One-pot synthesis, in which amino acids are added at random to a growing chain in a single reaction vessel, has some attraction, especially if a method could be found in which folded chains could be specifically extracted from the reaction mixture. It should be borne in mind that, because of the synthetic strategy, many of the products will be very similar to others in the mixture, further complicating identification of individual molecules. In any case, this approach rapidly runs into what may be termed the Avogadro Number barrier. This is reached at a chain length of approximately 18 residues, if all 20 amino acids are used, as, at this point 1 mole of synthetic material (~ 2 kg) will contain just one molecule of each of the possible sequences. This chain length is almost certainly far too short to produce a stable, folded structure but such an approach might well help in aiding in answering precisely the question as to what the minimum chain length for folding actually is. It is clear that substantial concessions to mathematics, with concomitant loss in randomness would have to be made for there to be any reasonable chance for such methodology to work. For example, the set of amino acids could be considerably reduced, although the synthesis of a 40-mer using only four amino acids would already approach the Avogadro Number barrier. An alternative method in which preconstructed peptide blocks are added together in a random fragment condensation strategy offers some numerical hope but at a cost to the underlying random principles, as it creates the problem of selecting the sequences for the initial blocks. However, this approach has the advantage that suitable chemistry

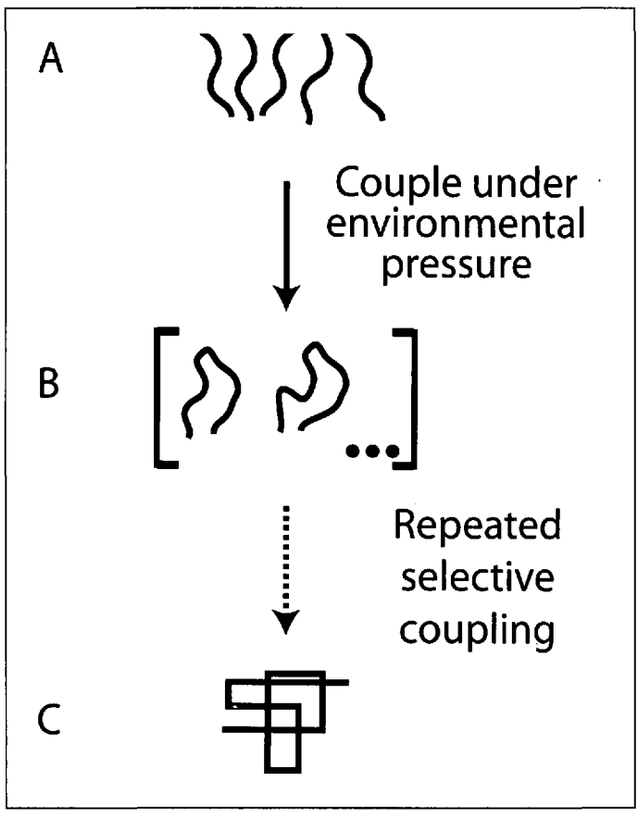

Fig. 2. Fragment condensation under environmental pressure. A set of short oligopeptide monomers, **A**, are condensed together in a random manner to form dimers. Due to the environmental stress imposed only some of the possible products will be formed, representing only a small set of all the possible dimers, **B**. Repeated application of this procedure may lead to folded, and possibly functional species, **C**.

has been established [7]. Fragment condensation of short peptide blocks of, say, ten residues could be used in a three-step process to arrive at the critical length of 40 residues (Fig. 2). If 100 blocks were to be used in the initial step there would be 100^2 possible 20 residue products. If, now, a form of environmental selectivity were to be employed during the condensation that favored certain combinations and disfavored others, it might be possible to reduce the number of products substantially. The type of environmental pressure to be employed would include such factors as, for example, solubility at different pH values, salt concentrations and temperatures. Repeating the procedure with the 20 residue products under similar conditions could then be use to construct a relatively restricted set of 40-mers. Whilst it may appear that the selection procedures would be purely arbitrary, the aim would be simply to show that it is possible to produce a limited number of folded proteins from among the enormous number of theoretical possibilities. A second, somewhat fanciful, selective method would rely on the properties of the randomly synthesized peptides themselves. Imagine that a large library of polypeptides has been synthesized (neglecting the polymerization methods used). It will be assumed that within this population certain members are capable of folding and that of those there are some that possess biological activity. Furthermore it will be supposed that of the biologically active members a small number are proteases. As the majority of the polypeptides present will be structurally random and, hence, highly susceptible to proteolytic degradation the end result should be a small number of folded proteases. In reality, it is asking too much to expect that a protease might arise by chance, however the general concept can be used as a method for discriminating between folded and unfolded chains, as will be discussed below.

The best opportunity currently available for overcoming the numerical problems would appear to be the use of molecular biological techniques. Here this will be viewed simply as a synthetic method for the simultaneous production of a large a number of peptides as possible but, as will be seen, it also offers the advantage of the use of biological specificity for the analysis of the products.

In molecular biological peptide synthesis randomness is introduced at the DNA level. Randomly constructed DNA sequences are inserted into suitable expression vectors and the peptides produced isolated and characterized. It is clear that this has one immediate drawback. As the genetic code is degenerate; there are 64 possible codons of which 61 code for amino acids. There are, for example, six codons for leucine, four for threonine but only one each for tryptophan and methionine. Using randomly synthesized DNA then yields a population of products with a strongly biased composition. A trick can be employed here, however. If the first two positions of a codon are filled by random selection from among the four bases but the third is occupied by only two, T or G, the number of codons is halved to 32 of which 31 code for amino acids and each of the 20 amino acids is represented at least once. This reduces the relative frequency of amino acids that are otherwise over-represented. The synthesis of DNA fragments of up to 100 bases is routine and such fragments can be simply ligated so that the production of a peptide of, say, 40 residues is relatively facile. The next criterion to be satisfied is that of numbers – how many sequences can be produced simultaneously using such methodology? Current techniques can produce up to 10^8 different clones within a population of suitable bacteria. For the sake of simplicity, the ability of bacteria to 'edit' or alter the peptides produced will be ignored and, in any case, is poorly understood. 10^8 starts to sound like a reasonable sample when it is recalled that there are probably less than 10^{10} proteins of all sizes that actually exist. The techniques to be described here offer the opportunity of producing 10^8 peptides of just one sequence length (*e.g.* 10^8 40-mers) and, in principle, any sequence length could be sampled by such an approach although there are currently upper limits to size dictated by the length of random DNA that can be synthesized.

A practical implementation of these principles can be realized that utilizes the technique of phage display, which has entered routine use over the last few years. The underlying principle is as follows. Phage are viruses that infect bacteria and are dependent on the bacterial infrastructure for their own replication. The

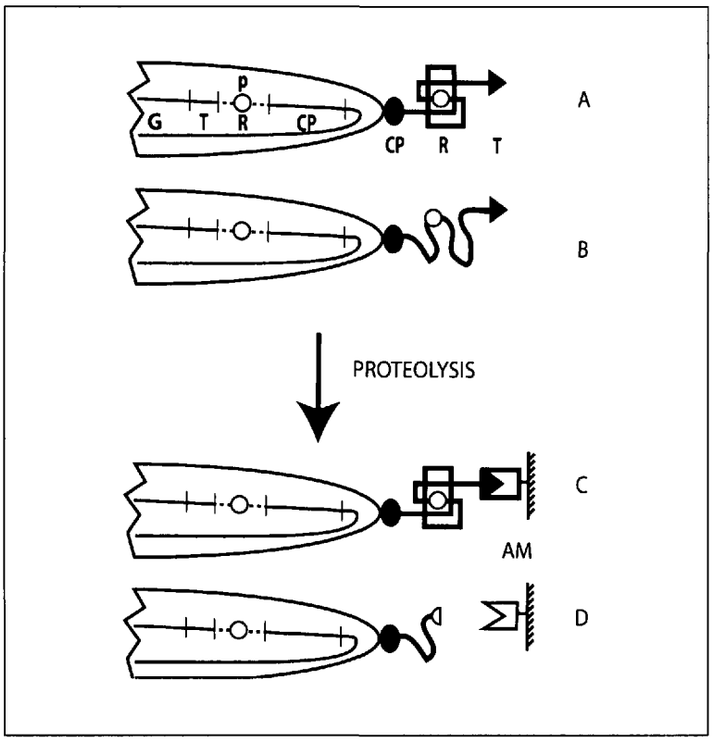

Fig. 3. Principle of the phage display method. The tip of a filamentous phage is shown including a portion of the phage genome, **G**. The construct for the production of random polypeptides encodes a phage coat protein, **CP**, a randomly synthesized DNA segment, **R**, and a tag, **T**. A protease cleavage site, **p**, is incorporated in the centre of the random sequence and the product, a fusion protein, is expressed on the outer surface of the phage. When the phage population is subjected to controlled proteolysis, those individuals in which the random portion of the fusion protein is folded, burying the protease cleavage site, **A**, are protected against digestion, while those in which the region is unfolded, **B**, are susceptible to cleavage, thereby losing the tag. When the entire population is passed over an affinity matrix, **AM**, those individuals in which the tag is intact are retained, **C**, while the remainder, **D**, are washed away. In this way, phage expressing a folded protein can be isolated and reserved for further study.

phage can be fooled into incorporating pieces of bacterial DNA offering the opportunity of manipulating a bacterial population so that designer phage can be produced. If a construct is produced at the bacterial DNA level in which a novel protein (a random sequence, for example) is fused to a protein that is normally expressed as part of the outer surface of a phage, following phage infection of the bacterial population, the phage express the desired novel protein on their outer surface. Furthermore, as the protein is the functional expression of the information coded by the genetic material contained within the phage, the properties of the protein can be used to isolate specific phage and, hence, the phage DNA encoding the property. The DNA can then be isolated, analyzed and amplified. Each phage population will express one, and only one, of the random proteins that have been introduced, and, with careful manipulation, 10^8 such populations can be created in one experiment. A scheme whereby these techniques can be exploited within the current context is shown in Fig. 3. Central to the approach is the principle that an unfolded protein is more susceptible to proteolysis than a folded one and that this property can then be used to identify folded proteins. The basic idea is that a random sequence is fused *via* its C-terminus to a phage coat protein, in the normal way for phage display, while its N-terminus is extended by a peptide that is an antigen ('tag') for a specific antibody. Within the otherwise random central sequence a specific cleavage site for a particular protease is incorporated. This concession is required for the method to function and it should be noted that while such a cleavage site can be formed by only one amino acid residue, more may be needed to create protease specificity. Up to 10^8 different phage are now produced, each population of which expresses a different random polypeptide all of which, however, contain the specific protease cleavage site. The entire phage population is now subjected to controlled proteolysis and all members of the random protein population that are unfolded are cleaved and thereby lose their N-terminal tag. The culture is now subjected to affinity selection by passing it over a surface on which the antibody for the tag is immobilized. Only those representatives which still have the tag bind to the antibody and the remainder are washed away. The phage isolated in this way are those that are expressing a folded protein and they can then be analyzed in more detail in either statistical terms or in more detail. Ultimately folded proteins detected by this method could be produced in sufficient quantity for full physicochemical and structural analysis.

Received: January 15, 2001

[1] T.L. Blundell, J.E. Pitts, I.J. Tickle, S.P. Wood, C-W. Wu, *Proc. Natl. Acad. Sci.* **1981**, *78*, 4175.
[2] a) R.C.D. Milton, S.C.F. Milton, S.B.H. Kent, *Science* **1992**, *256*, 1445; b) S. Vunnam, P. Juvvadi, K.S. Rotondi, R.B. Merrifield, *J. Pept. Res.* **1998**, *51*, 38.
[3] B.K. Davis, *Prog. Biophys. Mol. Biol.* **1999**, *72*, 157.
[4] S.L. Miller, in 'The molecular origins of life: Assembling pieces of the puzzle', Ed. A. Brack, Cambridge University Press, **1998**, p. 59.
[5] a) S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M.H. Hecht, *Science* **1993**, *262*, 160; b) S. Roy, G. Ratnaswamy, J.A. Boice, R. Fairman, M.H. Hecht, *J. Am. Chem. Soc.* **1997**, *119*, 5302.
[6] SWISS-PROT: Annotated protein sequence data-base, Release 39.10, *http://www.expasy.ch/sprot/*
[7] L.E. Canne, P. Botti, R.J. Simon, Y. Chen, E.A. Dennie, S.B.H. Kent, *J. Am. Chem. Soc.* **1999**, *121*, 8720.
[8] 'Phage Display of peptides and Proteins: a laboratory manual', Eds. B.K. Kay, J. Winter, J. McCafferty, Academic Press, San Diego, **1996**.