Journal of Biochemistry and Molecular Biology, Vol. 31, No. 3, May 1998, pp. 209–220

JOURNAL OF
Biochemistry
and
Molecular Biology

In vitro Constructive Approaches to the Origin of Coding Sequences

Kiyotaka Shiba
PRESTO, Japan Science and Technology Corporation (JST)
and Department of Cell Biology, Cancer Institute,
Japanese Foundation for Cancer Research, Kami-Ikebukuro, Toshima-ku, Tokyo 170-8455, Japan

Received 18 March 1998

How did nature create the first set of genes at the beginning of life on Earth? One of the goals of molecular biology is to elucidate the fundamental rules governing how genes and, therefore, proteins were created. Through experiments carried out in the emerging field of *"in vitro"* or "benchtop" evolution studies, we are gaining new insights into the origins of genes and proteins as well as the origins of their functions (e.g., catalysis). In this review, I present an overview of recent experimental approaches to the question of the origin and evolution of genes. In addition, I will introduce a novel *in vitro* protein emergence system that was recently developed in my laboratory.

Keywords: Evolutionary molecular engineering, *Ex vivo* evolution, Microgene polymerization reaction (MPR), Molecular diversity, Periodic structure.

Introduction

All the activities of human cells are maintained by the function of approximately 70,000 to 100,000 genes. The precise number of our genes and their complete classification will be known by the year 2005 when the Human Genome Project completes its DNA sequencing of the 3 billion base pairs that make up the human genome. In parallel with the Human Genome Project, genome sequencing of selected microorganisms has also been ongoing. As a result of these efforts, we presently know all of the genes in thirteen microorganisms from three main taxonomic domains: Archaea, Bacteria, and Eucarya.[1]

By revealing their genomic structures, we learned that *Saccharomyces cerevisiae* contains approximately 6500 genes (Goffeau *et al.*, 1996), but *Mycoplasma genitalium* has no more than 500 genes (Fraser *et al.*, 1995). Many of the genes found in any given organism have their orthologs in other organisms (Tatusov *et al.*, 1997), even if they belong to distinct taxonomic domains. For instance, orthologs of most aminoacyl-tRNA synthetases are found in all 3 taxonomic domains (Brown and Doolittle, 1995; Shiba *et al.*, 1997). This circumstance highlights the fact that these genes are descendants of ancient genes comprising the primordial genome of "cenancestor" [the common ancestor of Archaea, Eucarya, and Bacteria (Fitch and Upper, 1987)]. Although the number of genes in *Homo sapiens* is much greater than those of microorganisms, many human genes must be categorized as paralogs that evolved from preexisting genes through duplication. Thus, the history of many existing genes may trace back to the ancient genome of cenancestor. With this in mind, the question is: how did the first genes emerge at the very beginning of life on Earth? The first set of genes did not have master genes from which they could evolve by duplication or shuffling; therefore, they must have emerged *de novo*. Moreover, a single gene may not be sufficient to initiate life. Thus, a set of genes may have emerged almost simultaneously to originate the first living creature on Earth. The origin of the first genes may be a much more challenging issue than the evolution of genes, because the former resists application of the reductionist approach of molecular phylogenetics.

In the 1990s, a group of experimental protocols was developed which were characterized by the keywords *"in vitro* evolution" (Tuerk and Gold, 1990), *"in vitro* selection" (Robertson and Joyce, 1990), *"in vitro* genetics" (Szostak, 1992), "combinatorial libraries" (Houghten *et al.*, 1991), "molecular diversity" (Geysen *et al.*, 1995), and "evolutionary molecular engineering" (Kumar *et al.*, 1995). Although these experiments belong to different

* To whom correspondence should be addressed.
Tel: 81-3-3918-0111; Fax: 81-3-5394-3903
E-mail: kshiba@jfcr.or.jp

[1] The current status of microorganisms genome projects is shown at web site, http://www.tigr.org/tdb/mdb/mdb.html.

disciplines, their aims are identical. One goal is practical in nature: development of novel molecules or molecules with improved functionality. The other goal is to obtain basic knowledge of the origin of genes and proteins as well as their functions (e.g., catalysis). In this review, I focus on the latter aspect of experimental evolution and discuss how we can address the issue of the origin of coding sequences (proteins).

Experimental approaches to the origin of life

Early research Aside from the very early experiments of F. Redi (17th century), L. Spallanzani (18th century), and L. Pasteur (19th century), who provided evidence against the theory of the spontaneous generation of life, the first experimental approach to the question of the origin of life can be traced back to the classical experiment of S. L. Miller in 1953 (Miller, 1953). In this experiment, Miller observed the formation of amino acids (alanine, glycine, and others) in a flask containing a mixture of methane, water, ammonia, and hydrogen; a series of electrical sparks was passed through the mixture for a period of one week. Thus, inside Miller's flask was a simulation of Earth's primordial atmosphere, and in contrast to earlier ideas (17th century–19th century), the underlying concept in this experiment is that life may have spontaneously generated from simple chemicals (Oparin, 1938). Miller's "spark experiment" has been followed by many other elegant experiments investigating the process of "chemical evolution". One of the principle goals of this research is the construction of a model system that simulates conditions of the prebiotic world in order to observe whether self-replicating macromolecules, the most plausible first forms of life, will spontaneously emerge (Li and Nicolaou, 1994; Ferris *et al.*, 1996; Lee *et al.*, 1996).

Molecular biologists have also shown a great interest in constructing "*in vitro* evolution" systems. In particular, S. Spiegelman and his colleagues began a series of experiments in 1967 that they called "an extracellular Darwinian experiment" (Mills *et al.*, 1967). For this procedure, they used a very simple *in vitro* evolution system consisting of the RNA genome of bacteriophage Qβ, a replication enzyme (Qβ replicase), and triphosphates. They had already shown that Qβ RNA can be very efficiently replicated *in vitro* using Qβ replicase. Under the appropriate selective pressure, such as the presence of ethidium bromide, they observed that adapted Qβ RNA variants evolved in the population after serial transfers of the reaction mixture (Kramer *et al.*, 1974). The molecular basis of this "Darwinian evolution" is the very high error rate of Qβ replicase (10^{-3} to 10^{-4} per nucleotide incorporated); this high error rate meant that variant RNA offspring emerged at each replication step, and more rapidly replicating variants (adapted offspring) overgrew other RNA molecules. The spontaneous generation of variants in offspring, and the fact that fitted variants propagated more rapidly in the system, supported the notion that Spiegelman's group had constructed a model of Darwinian evolution. In the 1970s, this work inspired M. Eigen to conceive of "sequence space" and "quasispecies" and to propose the "hyper cycle model" of molecular evolution, which are now important concepts to design benchtop evolution systems (for details, see Eigen, 1992; Biebricher and Gardiner, 1997).

Evolution experiments in the '90s For several decades following Spiegelman's excellent early work, further development of experimental evolution was hindered, primarily, because of practical limitations: investigators lacked facile methods with which to carry out more sophisticated experiments. In particular, development of the protocols for site directed mutagenesis (Zoller and Smith, 1982), *in vitro* transcription (Milligan *et al.*, 1987), and polymerase chain reaction (PCR) (Saiki *et al.*, 1988) were critical for the construction of the modern systems introduced below. In addition, it should be pointed out that the discovery of self-splicing RNA by T. R. Cech (Kruger *et al.*, 1982) revived interest in the experimental evolution of RNA originally begun by Spiegelman. Below, I present an overview of the experimental approaches to evolution revived in the 1990s; they are classified into 4 fields: "Evolving systems based on nucleic acids", "Phage display experiments", "Catalytic antibodies", and "Combinatorial chemistry".

Evolving systems based on nucleic acids The systems termed "*in vitro* nucleic acid selection" or "*in vitro* nucleic acid evolution" are regarded as the historical descendants of Spiegelman's experiments. For Spiegelman, the phenotype to be selected was limited to the fitness of RNA molecules to serve as substrates for Qβ replicase. In contrast, the systems established in the 1990s made use of a much wider array of phenotypes including binding to specific ligands (Ellington and Szostak, 1990; Tuerk and Gold, 1990), altered substrate specificity (Robertson and Joyce, 1990), and emergence of catalytic activities (Bartel and Szostak, 1993; Breaker and Joyce, 1994).

A flow chart illustrating the steps that make up benchtop evolution experiments is shown in Fig. 1. Experiments entailing *in vitro* RNA selection usually start with pools of random nucleic acids (Ellington and Szostak, 1990; Tuerk and Gold, 1990). At the present time, populations containing 10^{13}–10^{15} different DNA molecules can be easily generated by combinatorial polymerization of four deoxyribonucleotides using a commercially available DNA synthesizer. These DNA pools can then be converted to random RNA pools by *in vitro* transcription. Alternatively, in evolution experiments, existing molecules (e.g., a ribozyme) are used as starting materials (Robertson and Joyce, 1990). In that case, variant offspring of the starting

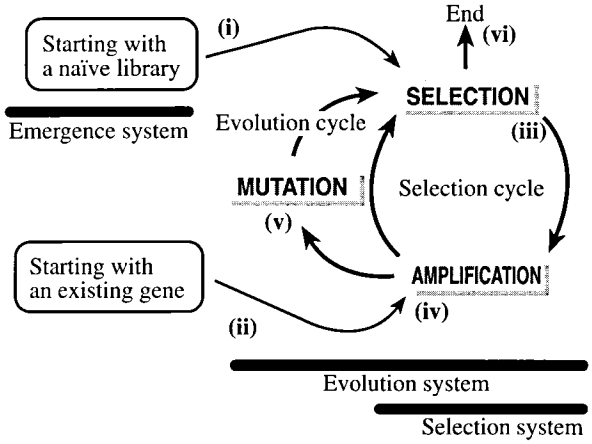

Fig. 1. Flow chart illustrating the steps that comprise benchtop evolution experiments. Experiments begin with either (i) a naïve library or (ii) an existing gene, depending upon the aim of the experiment. Naïve libraries are often generated by combinatorial polymerization of building blocks (nucleotides, amino acids, V genes for antibodies, microgenes, chemicals), and I refer to the strategies for generating naïve libraries as "emergence systems". When starting with existing genes, a pool of their variants must first be generated by (iv) "amplification" and (v) "mutation". Molecules possessing targeted phenotypes are enriched by iterative processes of (iii) selection and (iv) amplification ("selection cycle"). A mutation step can be incorporated into the cycle resulting in an "evolution cycle" consisting of (iii) selection, (vi) amplification, and (v) mutation. The simplest paradigm, which entails (i) synthesis of a naïve library followed by (iii) the selection from that library, has proved very useful in the area of combinatorial chemistry.

material must first be prepared using appropriate methods, such as error-prone PCR (Cadwell and Joyce, 1992) or targeted random mutagenesis (Matteucci and Heyneker, 1983). Whether simulating selection or evolution, it is important that the starting materials have a large molecular diversity.

Once a starting population of RNA (or DNA) is prepared, the next step is to select the molecules that possess the desired phenotype. If, for example, ligand binding is the targeted phenotype, selection is performed using affinity chromatography (Ellington and Szostak, 1990; Bock *et al.*, 1992) or other physical separation methods (Tuerk and Gold, 1990). RNA aptamer folds into a specific tertiary structure and forms a complex with its target ligand (Jiang *et al.*, 1996). To select for the catalytic activities of RNA or DNA, specific strategies must be developed for each experiment (Robertson and Joyce, 1990; Bartel and Szostak, 1993). When selecting from large numbers of starting molecules (e.g., 10^{13}), a single selection operation is not sufficient to yield the fittest molecules. Each selection cycle enriches the population of fit molecules within the pool. To obtain enough enrichment of the pool for purposes of cloning, iterative processes of selection and amplification are necessary (Joyce, 1989; Ellington and Szostak, 1990). In this protocol, the selected population of RNA is converted to DNA using reverse-transcriptase and then amplified by PCR to the original library size. The amplified pool is then further enriched by the next selection operation and so on. If necessary, mutagenesis can be incorporated into the amplification step by employing error-prone PCR (Bartel and Szostak, 1993), thereby constructing a Darwinian evolution cycle composed of amplification, mutation, and selection (Joyce, 1989).

The procedure just described is often referred to as "SELEX" (systematic evolution of ligands by exponential enrichment) (Tuerk and Gold, 1990), and the nucleic acid ligands that specifically bind to target molecules are called "aptamers" (Ellington and Szostak, 1990). Generation of aptamers by SELEX has had a great impact on areas of applied science (Gold *et al.*, 1995). The clinical use of aptamers as pharmaceutical agents is currently being tested at several biotechnology firms (Ellington and Conrad, 1995; Gold, 1995). In addition, many novel ribozymes have been created by J. W. Szostak's group and by others. These include, among others, ribozymes having ligase activity (Bartel and Szostak, 1993), kinase activity (Lorsch and Szostak, 1994), aminoacyl-tRNA synthesis activity (Illangasekare *et al.*, 1995), and alkylating activity (Wilson and Szostak, 1995).

Phage display experiments When making use of evolving systems based on nucleic acids, RNA (or DNA) serves both as a carrier of hereditary information (genotype) and as an entity to execute its functional activity (phenotype; Fig. 2). Cech's discovery of RNA molecules with catalytic activity (ribozymes) (Kruger *et al.*, 1982) made RNA the most suitable molecule for use in molecular evolution experiments because RNA could serve as both genotype and phenotype (Joyce, 1989). On the other hand, if one wants to use a protein (or polypeptide), which is the translated product of the genetic information encoded in the DNA, as the functional moiety in an evolution system, a physical link between the DNA (genotype) and the encoded protein (phenotype) must be established. In "phage display" experiments, the linkage is mediated by a bacteriophage (Figs. 2 and 3).

The first phage display experiment was reported in 1985 by G. P. Smith (Smith, 1985). Smith inserted fragments of DNA encoding for the "*Eco*RI" endonuclease/methylase into gene III of phage f1. The f1 (or its relative M13) phage is filamentous (6.5 nm in diameter and 600 nm long) and carries a single-stranded DNA genome within its particle (Fig. 3). The phage infects the F pilus of *E. coli* and produces its progeny without killing the host cell. The product of gene III is located at one end of the phage particle as a minor coat protein and is essential for phage infection to F pilus. Smith demonstrated that segments of *Eco*RI were displayed on phage particles as a fusion

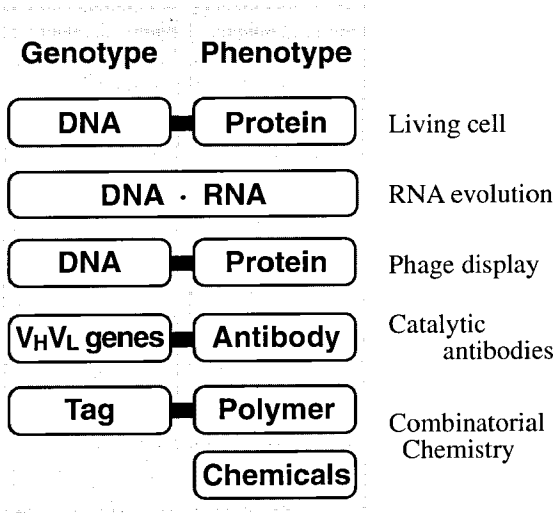

Fig. 2. Strategies for linking phenotype and genotype. In evolving systems based on nucleic acids, phage display protocols, and catalytic antibody development, the carrier of hereditary information (genotype) is always DNA (or RNA) which may be subjected to amplification and/or mutagenesis. In contrast, combinatorial chemistry experiments do not use heritable molecules as the genotype since it is merely a code for the structures of functional molecules (phenotype). Indeed, most combinatorial chemical libraries do not contain molecules for the genotype. In those experiments, generation of libraries containing large molecular diversity, as well as the selection of target molecules, are focused.

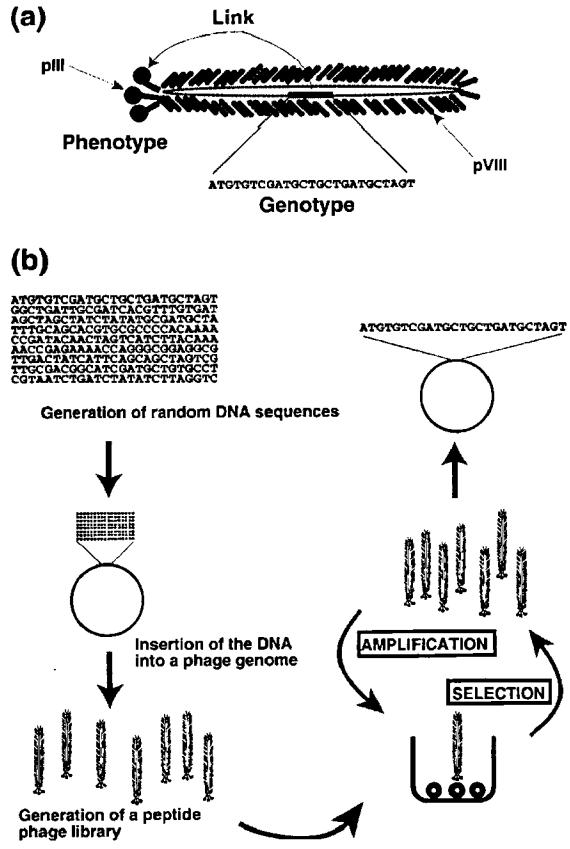

Fig. 3. Phage display system. (a) Schematic representation of a phage particle. The single stranded DNA genome is encased in coat proteins. pIII is a minor coat proteins (3–5 molecules per particle) and is located at one end of a particle. pVIII is a major coat protein. Foreign peptides or proteins are displayed on phage as fusion proteins with pIII or pVIII. (b) An example of peptide selection using phage. Random DNA sequences made by a DNA synthesizer are inserted into the gene III region and libraries of phage displaying random peptide sequences are made. Phage are then propagated in *E. coli* cells. Phage lysate is panned with immobilized target molecules, and those that bind to the target are enriched. Cycles of selection and re-amplification of bound phage in *E. coli* yield a population of binding phage. The structures of displayed peptides are determined by sequencing the DNA inserted into phage genomes.

protein with pIII. Because the phage genome is packed in the phage particle, the physical link between phenotype (a protein segment of *Eco*RI) and genotype (a DNA fragment coding for the segments of *Eco*RI) is firmly coupled in this experiment. Smith also showed that enrichment of phage displaying the enzyme fragment within a mixture of displayed and nondisplayed phages can be accomplished using an immobilized polyclonal antibody to *Eco*RI (Smith, 1985). This experiment demonstrates that selection and evolution cycles can be applied to the phage display system as well as to an RNA evolution system.

In 1990, three groups using phage display systems independently succeeded in selecting functional peptides that bind either to specific antibodies (Cwirla *et al.*, 1990; Scott and Smith, 1990) or to streptoavidin (Devlin *et al.*, 1990). They inserted DNA pools into the genes coding for coat proteins in order to make phage libraries which display random peptide sequences and, after several selection cycles, they isolated phages displaying specific binders to the targets. Since this early work, large numbers of peptides that specifically bind to target molecules have been isolated using peptide display systems.

Another important experiment that employs the phage display methodology is the "phagebody" system (Lerner *et al.*, 1992; Marks *et al.*, 1992). The first *ex vivo* generation of an immunoglobulin repertoire from the combinatorial assembly of V_H and V_L genes was reported in 1989 using λ phage (Huse *et al.*, 1989). Soon thereafter, phage display methodology was incorporated into the field, resulting in a display of Fab and scFv on phage particles. Currently, the methodology is routinely used to isolate V genes for specific antigens. The phagebody system is also used as a means to evolve high-affinity antibodies from low-affinity ones without *in vivo* maturation processes (Barbas and Burton, 1996; Hoogenboom, 1997).

Many proteins, in addition to antibodies, have been displayed on phages for purposes of directed evolution. These include alkaline phosphatase (McCafferty *et al.*, 1991), neutrophil elastase inhibitors (Roberts *et al.*, 1992), and growth hormones (Bass *et al.*, 1990) among others.

Thus, the filamentous phage display system has proven to be a powerful tool in the investigation of peptide and protein evolution. Now, in addition to the filamentous phage, display systems using $Q\beta$ phage (Kozlovska *et al.*, 1996), λ phage (Sternberg and Hoess, 1995; Mikawa *et al.*, 1996), and baculovirus (Mottershead *et al.*, 1997) have been developed. Display on cells rather than phage has also been employed using PhoE (Agterberg *et al.*, 1990), OmpA (Georgiou *et al.*, 1996), and flagellin (Lu *et al.*, 1995) of *E. coli*, and Aga2p of *S. cerevisiae* (Boder and Wittrup, 1997). Finally, another emerging field of study incorporates the cell-free "polysome display" technique (Mattheakis *et al.*, 1994; Hanes and Pluckthun, 1997; Nemoto *et al.*, 1997; Roberts and Szostak, 1997) which can accommodate much larger libraries (10^{14}–10^{15}) than the phage display system (10^{9}).

Catalytic antibodies What is the nature of biological catalysis, and what is the origin of catalytic activity? Can we create novel proteins that catalyze reactions for which no known enzymes exist? These are questions addressed by experiments in catalytic antibodies (Lerner *et al.*, 1991). The history of this field goes back to the late-1940s when L. Pauling noted that the difference between enzymes and antibodies is that the former bind transition state molecules, while the latter bind ground state molecules (Pauling, 1948). So, if an antibody binds to a haptenic group whose structure resembles the transition state of a given reaction, the antibody could act as a catalyst for that reaction (Jencks, 1969). Since the first experimental demonstrations of catalytic antibodies by R. A. Lerner, P. G. Schultz and their coworkers (Pollack *et al.*, 1986; Tramontano *et al.*, 1986), antibodies catalyzing more than 50 reactions have been developed (Benkovic, 1992). Preparation of catalytic antibodies is generally started by immunizing mice with a hapten that is a stable analog of a transition state molecule of the reaction of interest (Fig. 4). During the primary immune response, antibodies that bind to the hapten are selected from a large pool of antibodies generated by combinatorial rearrangement of V, D, and J gene segments. The antibodies are further evolved during the secondary immune response when somatic mutation provides them with higher affinity for the hapten (Fig. 4). Thus, the immune response resembles natural selection and evolution, and in this regard, generation of catalytic antibodies can be viewed as being analogous to benchtop evolution. The marriage of catalytic antibodies and the phage display methodology has given birth to a new field (Janda *et al.*, 1994) in which catalytic antibodies are generated through selection targeted to the catalytic activities of antibodies (Janda *et al.*, 1997).

Combinatorial chemistry In all of the experiments described so far, biological materials such as filamentous phage, *E. coli*, and mice were used. Even in the case of *in*

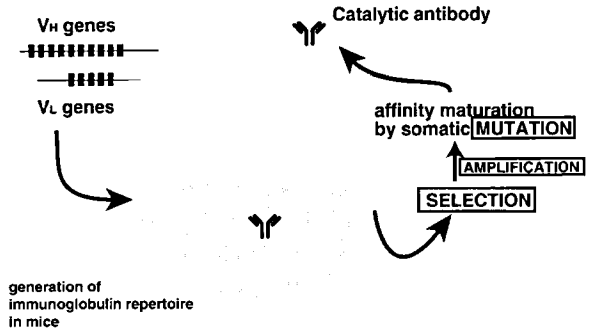

Fig. 4. Selection and evolution of catalytic antibodies *in vivo*. Generation of catalytic antibodies follows a protocol of experimental evolution. Combinatorial assembly of blocks of V, D, and J genes generates an immunoglobulin repertoire, which in animals, consists of approximately 10^{7} different combinations. The immunization of a hapten (an analog of a transition state molecule for a given reaction) selects antibodies from the repertoire that recognize the hapten. Expansion (amplification) of B cells and incorporation of somatic mutations into antibody genes improves (evolves) the affinity of the antibodies. The process of antibody development provides an *in vivo* example of directed evolution.

vitro RNA evolution, a DNA polymerase isolated from bacteria, an RNA polymerase from a bacteriophage, and a viral reverse transcriptase are essential components for the experiments. Thus, as long as iterative strategies of selection, mutation, and amplification are employed (Fig. 1), protocols in which "genotype" = "DNA" and "phenotype" = "RNA or protein" are the best (Fig. 2), and biological materials that process the genetic information must be used. If, on the other hand, the amplification and mutation steps are not focused, the genotype can be thought of as just a molecular "tag" recording the structure of the phenotype. In that case, the choice of molecules that can serve as genotype and phenotype are dramatically extended (Brenner and Lerner, 1992; Needels *et al.*, 1993; Ohlmeyer *et al.*, 1993). Furthermore, the emerging field of combinatorial chemistry, whose historical work also dates from the early '90s (Houghten *et al.*, 1991; Lam *et al.*, 1991), can be thought of as an *"in vitro* selection" system that does not have "phenotype" (Fig. 1). With combinatorial chemistry, however, the generation of a large pool of molecular diversity and the methods for selection are primarily focused (see reviews, Gallop *et al.*, 1994; Gordon *et al.*, 1994; Ecker and Crooke, 1995; Schultz and Schultz, 1996; Myers, 1997).

Evolution from existing genes and selection from naïve libraries

The benchtop evolution experiments introduced above can be classified into two categories based on their starting materials. One group starts with an existing gene and the

other group starts with a naïve library (Fig. 1). Examples of the former include the isolation of ethidium bromide-resistant Qβ RNA (Kramer *et al.*, 1974), the isolation of a ribozyme having altered substrate specificity (Robertson and Joyce, 1990), the affinity maturation of an antibody on a phage (Hawkins *et al.*, 1992), and the alternation of the target specificity of a DNA binding protein by phage display (Rebar and Pabo, 1994) among others. In each of these experiments, an existing gene was used as parental material, and its variants pool was first generated by mutagenesis. Starting from this variants pool, the fittest offspring were chosen by a series of selection or evolution cycles. Thus, these are typical Darwinian evolution protocols consisting of amplification, mutation, and selection.

An alternative approach has been to make use of naïve libraries as starting material. This method has enabled investigators to isolate aptamers (Ellington and Szostak, 1990; Tuerk and Gold, 1990), to select new ribozymes (Bartel and Szostak, 1993), and to generate catalytic antibodies. In addition, naïve libraries are used in most experiments in the area of combinatorial chemistry. As an example, Szostak's group was able to select RNA aptamers that bind to organic dyes from 10^{13} random 100-nt RNA sequences (Ellington and Szostak, 1990). The random RNA pool was transcribed from random DNA sequences that were synthesized by combinatorial polymerization of four nucleotide blocks (A, T, G, and C) and do not result from mutagenesis of an existing gene. They also succeeded in selecting novel ribozymes having weak ligase activity from 10^{15} random 220-nt RNA sequences. They then used an "evolution cycle" to improve the ligase activity of the ribozyme (Bartel and Szostak, 1993; Ekland *et al.*, 1995). Catalytic antibodies have also been selected from a mouse immunoglobulin repertoire that was formed from combinatorial assemblages of V, D, and J gene blocks. There is no reason to believe that the immunoglobulin repertoire contained the parental catalyst for any particular reaction which suggests that new catalysts were selected from a naïve library.

In vitro protein evolution experiments

Investigations into directed protein evolution often begin with existing genes. Examples include maturation of antibody affinity (Hawkins *et al.*, 1992), alteration of the efficacy of bovine pancreatic trypsin inhibitor (Roberts *et al.*, 1992), alternation of the target specificity of a DNA binding protein (Rebar and Pabo, 1994), and alternation of the substrate specificity of a galactosidase (Zhang *et al.*, 1997). The methodologies used in experimental protein evolution include phage display as well as "sexual PCR", which was developed by W. P. C. Stemmer (Stemmer, 1994a; 1994b) and has proven to be a powerful tool for making "sparse libraries" of a starting gene (Crameri *et al.*, 1998). Sexual PCR entails fragmentation of related genes (or of a single gene) and reassembly by primerless PCR in order to generate a large library of chimeric genes. Such chimeric structures result in "sparse sampling of sequence space" (or mutation into very distant sequences); pools of variants like these are difficult to obtain using standard error-prone PCR (Fig. 5). Moreover, the sparse sampling of libraries made by sexual PCR is believed to accelerate directed evolution of proteins (Crameri *et al.*, 1998).

The "binary code strategy" is another method used to generate sparse libraries of a given protein (Kamtekar *et al.*, 1993). With this method, the locations of polar and nonpolar residues within the protein of interest are specified, and their identities are relaxed by using degenerate codons (NAN for polar residues and NTN for

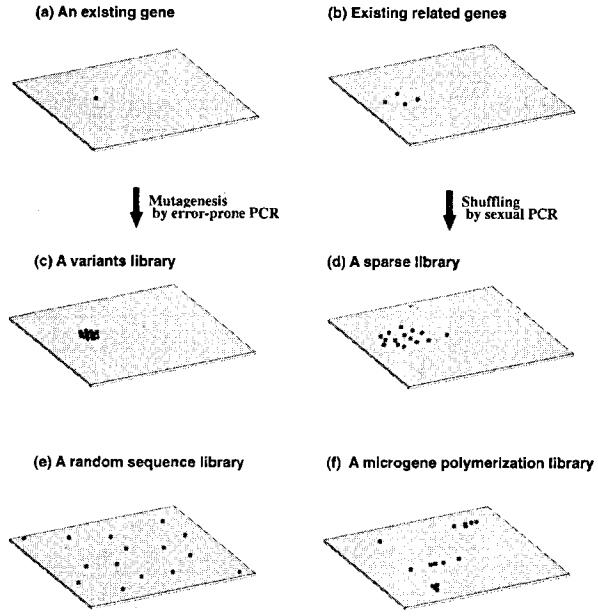

Fig. 5. Sampling spaces of starting libraries in various evolution experiments. Sequence space is the theoretical representation of all possible variants of a sequence (Eigen, 1992). For example, the sequence space of a 100 amino acids protein is composed of 20^{100} (1.3×10^{130}) variants; our universe has only 10^{80} molecules. Here, the sequence space of theoretical polymers (DNA, RNA, or proteins) of given length are schematically represented as square planes; the locations of sequences included in the starting libraries are shown as dots in the planes. (a) A single gene is located at a single position within the sequence space. (c) Amplification with mutagenesis of the single gene creates a variants library; the sampling space of such a library is clustered around the parental sequence. (d) Sexual PCR can shuffle the segments of related genes (b) yielding a sparse sampling of sequence spaces. (e) Random sequence libraries that are generated by combinatorial polymerization of monomer blocks (nucleotide or amino acids) exhibit very naïve sampling of sequence space. (f) Microgene-based libraries can also sample large portions of sequence space, but the sampling profile is constrained by the method used to generate the microgene polymers.

nonpolar locations, where N is A, T, G, or C). Using the binary code strategy, novel heme proteins were successfully selected from pools of sequences that were targeted to fold into four-helix bundles (Rojas *et al.*, 1997).

In contrast to directed evolution started from existing proteins, experiments using naïve libraries are very limited. Random DNA sequences that were successfully used in RNA/DNA evolution can not simply be applied to protein evolution experiments. The reading frames of random DNA sequences contain many translation termination codons (TAA, TAG, and TGA) which prevent such libraries from coding for larger proteins. Repeats of NNK (where K is G or T) eliminate two thirds of the termination codons in one reading frame and are frequently used as a library source for peptide selection experiments (Cwirla *et al.*, 1990; Scott and Smith, 1990). Codon-based polymerization is another solution for preparation of a stop codon-free single frame (Glaser *et al.*, 1992; Davidson and Sauer, 1994; Auld and Schimmel, 1995). In addition, a degenerated 16-mer oligonucleotide has been designed, which can be translated into highly randomized protein sequences and is devoid of any termination codons in all reading frames (Prijambada *et al.*, 1996). All of these methods generate libraries for proteins having nearly-random sequences. Nevertheless, some of these random sequence proteins exhibit properties similar to those of native proteins (Davidson *et al.*, 1995; Yamauchi *et al.*, 1998), supporting the hypothesis that proteins emerged from random sequences (White and Jacobs, 1993).

Sequence space and sampling space

Sequence space is the theoretical representation of all possible variants of a sequence of equal length (Eigen, 1992). For instance, the sequence space of a 5 amino acids oligopeptide is $5^{20} = 10^{14}$. Related sequences are clustered in sequence space by its definition. Sequence space is an important concept to consider when conducting *in vitro* evolution experiments (Kauffman, 1993; Stemmer, 1995). This is because the sequence spaces of large polymers are huge, and we can search only limited portions of the space (which is called sampling space). For example, the sequence space of 100-nt random RNA sequences contains 1.6×10^{60} possible variants which means that, using standard methodologies, we can search only $6 \times 10^{-46}\%$ of that space by synthesizing 10^{13} RNA sequences. Thus, we must be careful about the results of the evolution experiments. Although data from both experimental and theoretical approaches to the problems of sampling space have been accumulating (Aita and Husimi, 1996; Joyce, 1997; Crameri *et al.*, 1998), we are still far from completely understanding the problems. In Fig. 5, I schematically show conceptual views of the sampling spaces under various evolution experiments. The sampling space of the experiment starting with an existing gene is clustered around the parental sequence (c). This type of sampling profile would be fit for optimizing a parental activity. The sampling space is rather scattered using powerful mutagenesis methods such as sexual PCR or binary code strategy (d). Sparse sampling may avoid a trap in local minimum in optimization. Random sequence libraries exhibit very naïve sampling of sequence space (e). Experiments from Szostak have showed that many functional RNAs whose sequences were not related to each other were selected from this type of naïve sampling, indicating that sequences possessing a given function scatter here and there in sequence space (Ellington and Szostak, 1990; Bartel and Szostak, 1993).

Emergence of coding sequences from polymerization of microgenes.

Experiments to attempt to generate proteins from random sequences were introduced above. However, as mentioned, random DNA sequences are not the best source for protein emergence systems, because random nucleotide sequences contain many termination codons. Did nature create genes from random sequences of nucleic acid? If there were termination codons in a primordial genetic code system, it might be impossible that larger genes have emerged from random sequences. Here, I introduce alternative views to the hypothesis of the random origin of genes.

The "exon theory of genes" was proposed by W. Gilbert who postulated that novel genes could emerge from the assembly of exons (microgenes) (Gilbert, 1987). The fact that an existing gene can be dissected into multiple minigene units without loss of its activity (Shiba and Schimmel, 1992a; 1992b) supports the idea that microgenes serve as building blocks for generating larger genes (Seidel *et al.*, 1992; Shiba, 1995). Several *in vitro* protein emergence systems which mimic this "exon shuffling" have been developed (Nord *et al.*, 1995; 1997; Fisch *et al.*, 1996; Mikheeva and Jarrell, 1996; Shiba *et al.*, 1996). With these methods, short stretches of DNA (RNA), but not nucleotide blocks, are polymerized in a combinatorial manner. However, while the involvement of exon shuffling in the generation of new genes late in eukaryotic evolution is evident, its role in the generation of primordial genes is questionable (Palmer and Logsdon, 1991).

In the 1980s, S. Ohno published a series of papers that pointed out the repetitious nature of coding sequences (Ohno, 1981; Ohno and Matsunaga, 1982). From these observations, Ohno suggested that coding frames had emerged from repeats of short oligonucleotides (Ohno and Epplen, 1983; Ohno, 1987). This hypothesis proposes that nucleotide oligomers, which arose in the prebiotic world and were internal doubles, have progressively elongated as a result of unequally primed replication processes (Ohno, 1987). These oligomeric repeats may have served as

templates for the emergence of the first set of coding frames on Earth. Primordial coding sequences may have been filled with a multitude of repeats, and present day coding sequences would therefore be in the process of periodic-to-chaotic transition (Ohno, 1989). One advantage of this hypothesis is that, if a starter sequence is devoid of termination codons, oligomeric repeats of the sequence could contain rather large open reading frames and would be relatively tolerant to insertions and deletions; a second advantage is that translated proteins from such repeated sequences have periodic amino acid sequences and, consequently, might be expected to have higher propensities to form secondary structures (Ohno and Epplen, 1983). Since Ohno's proposal, numerous studies have confirmed the repetitious nature of not only coding sequences (Tsonis *et al.*, 1991; Korotkov and Korotkova, 1995; Tsonis and Tsonis, 1997) but also the genomic structure (Wolfe and Shields, 1997) and tertiary structures of many proteins (Yura *et al.*, 1993; Kobe, 1996).

We recently established a novel *in vitro* protein emergence system that mimics Ohno's scenario for the birth of coding sequences (Shiba *et al.*, 1997). In this system, a short stretch of DNA (microgene) is tandemly polymerized using a newly developed microgene polymerization reaction (MPR) (Shiba *et al.*, 1997). At the junctions where microgenes were joined, nucleotide insertions and deletions occurred randomly. Consequently, the generated microgene polymers were able to serve as combinatorial libraries of 2×3 reading frames from a single microgene (Fig. 6a). As long as the starting microgene is devoid of stop codons, the sequences could have long open reading frames whose products had a repetitious nature (Fig. 6b). The MPR-based polymerization of a microgene should serve as a new protein emergence system for protein evolution experiments. Also, repetitive polypeptides have been receiving attention in materials science (Tirrell, 1991; Ball, 1994; Brown, 1997). The MPR system would be used as a new methodology in this field. Attempts to select functional proteins from the microgene polymers are currently in progress in our laboratory.

The sequence space that can be sampled by the microgene based libraries described above is biased by the choice of microgene sequences, and has an alternative sampling profile (Fig. 5f), which may not be obtained from libraries of existing genes or from libraries of random sequences. The unique profile would modify the outcome of evolution experiments. Furthermore, we should bear in mind that nature did not search all sequence spaces during the emergence of genes. All possible sequence spaces of 100 base nucleotide polymers would include 1.6×10^{60} species and would weight 8×10^{37} kg by synthesizing all variants for one molecule each; in contrast, the weight of the Earth is only 6×10^{24} kg. Thus, nature did not search all sequence space for creating genes. The creation of genes must have proceeded under certain constraints such as the elongation of oligomeric repeats (Ohno, 1987).

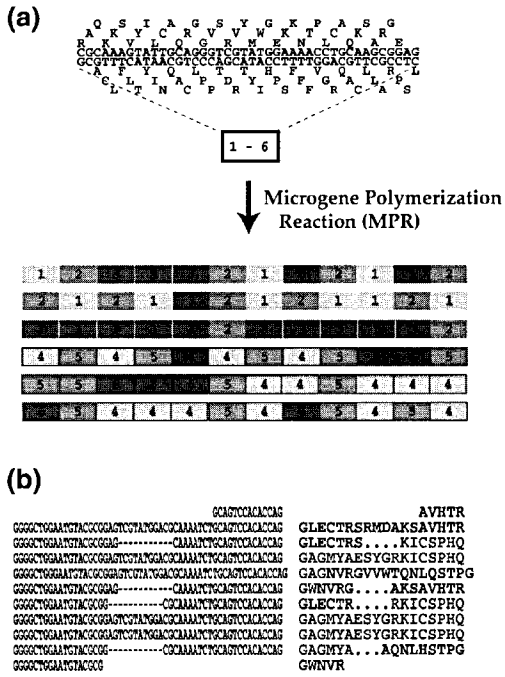

Fig. 6. Construction of a protein library from polymerization of a microgene. (a) A single short stretch of DNA provides 6 reading frames. The MPR method makes tandem polymers of the microgene. Random insertions or deletions at microgene junctions and within the microgene sequence mean that the resultant polymers are comprised of combinatorial libraries of reading frames. Polypeptides produced from such libraries have periodic amino acid sequences with a higher propensity to form secondary structures. (b) An example of a microgene polymer. The left-hand panel shows the DNA sequence, and the right-hand panel shows the translated sequence of 1 reading frame.

Perspective

Miller's experiment (Miller, 1953) was a constructive approach to the origin of life and contrasts with the reductionist approach of molecular biology. Although benchtop evolution protocols emerged from the field of molecular biology, these experiments can, nevertheless, be thought of as constructive approaches to the origin of genes. Rapid progress in the genome projects and in structural biology should help elucidate the fundamental grammar governing the structure of genes and proteins (Ohno, 1989; 1992; 1994; Solovyev, 1993; Tsonis, 1997). The construction of *in vitro* gene or protein emergence systems based on these grammars would then provide additional insight into the origins and evolution of genes.

Acknowledgment The first part of this article is based on a review written in Japanese and appeared in KAGAKU (Shiba, 1997).

References

Agterberg, M., Adriaanse, H., van Bruggen, A., Karperien, M., and Tommassen, J. (1990) Outer-membrane PhoE protein of *Escherichia coli* K-12 as an exposure vector: possibilities and limitations. *Gene* **88**, 37–45.

Aita, T. and Husimi, Y. (1996) Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. *J. Theor. Biol.* **182**, 469–485

Auld, D. S. and Schimmel, P. (1995) Switching recognition of two tRNA synthetases with an amino acid swap in a designed peptide. *Science* **267**, 1994–1996.

Ball, P. (1994) Polymers made to measure. *Nature* **367**, 323–324.

Barbas, C. F. and Burton, D. R. (1996) Selection and evolution of high-affinity human anti-viral antibodies. *Trends Biotech.* **14**, 230–234.

Bartel, D. P. and Szostak, J. W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science* **261**, 1411–1418.

Bass, S., Greene, R., and Wells, J. A. (1990) Hormone phage — an enrichment method for variant proteins with altered binding properties. *Proteins Struct. Funct. Genet.* **8**, 309–314.

Benkovic, S. J. (1992) Catalytic antibodies. *Annu. Rev. Biochem.* **61**, 29–54.

Biebricher, C. K. and Gardiner, W. C. (1997) Molecular evolution of RNA *in vitro*. *Biophys. Chem.* **66**, 179–192.

Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H., and Toole, J. J. (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* **355**, 564–566.

Boder, E. T. and Wittrup, K. D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotech.* **15**, 553–557.

Breaker, R. R. and Joyce, G. F. (1994) A DNA enzyme that cleaves RNA. *Chem. Biol.* **1**, 223–229.

Brenner, S. and Lerner, R. A. (1992) Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. USA* **89**, 5381–5383.

Brown, S. (1997) Metal-recognition by repeating polypeptides. *Nature Biotech.* **15**, 269–272.

Brown, J. R. and Doolittle, W. F. (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**, 2441–2445.

Cadwell, R. C. and Joyce, G. F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Applic.* **2**, 28–33.

Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. C. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291.

Cwirla, S. E., Peters, E. A., Barrett, R. W., and Dower, W. J. (1990) Peptides on phage — a vast library of peptides for identifying ligands. *Proc. Natl. Acad. Sci. USA* **87**, 6378–6382.

Davidson, A. R. and Sauer, R. T. (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **91**, 2146–2150.

Davidson, A. R., Lumb, K. J., and Sauer, R. T. (1995) Cooperatively folded proteins in random sequence libraries. *Nature Struct. Biol.* **2**, 856–864.

Devlin, J. J., Panganiban, L. C., and Devlin, P. E. (1990) Random peptide libraries — a source of specific protein binding molecules. *Science* **249**, 404–406.

Ecker, D. J. and Crooke, S. T. (1995) Combinatorial drug discovery: which methods will produce the greatest value? *Bio/Technology* **13**, 351–360.

Eigen, M. (1992) *Steps Towards Life: A Perspective on Evolution*, Oxford University Press, Oxford.

Ekland, E. H., Szostak, J. W., and Bartel, D. P. (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364–370.

Ellington, A. D. and Szostak, J. W. (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822.

Ellington, A. D. and Conrad, R. (1995) Aptamers as potential nucleic acid pharmaceuticals. *Biotech. Annu. Rev.* **1**, 185–214.

Ferris, J. P., Hill, A. R., Liu, R. H., and Orgel, L. E. (1996) Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* **381**, 59–61.

Fisch, I., Kontermann, R. E., Finnern, R., Hartley, O., Solergonzalez, A. S., Griffiths, A. D., and Winter, G. (1996) A strategy of exon shuffling for making large peptide repertoires displayed on filamentous bacteriophage. *Proc. Natl. Acad. Sci. USA* **93**, 7761–7766.

Fitch, W. M. and Upper, K. (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 759–737.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., and Venter, J. C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. A., and Gordon, E. M. (1994) Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **37**, 1233–1251.

Georgiou, G., Stephens, D. L., Stathopoulos, C., Poetschke, H. L., Mendenhall, J., and Earhart, C. F. (1996) Display of β-lactamase on the *Escherichia coli* surface: Outer membrane phenotypes conferred by Lpp′-OmpA′-β-lactamase fusions. *Protein Engng.* **9**, 239–247.

Geysen, H. M., Houghten, R. A., Kauffman, S., Lebl, M., Moos, M. H., Pavia, M. R., and Szostak, J. W. (1995) Molecular diversity comes of age! *Mol. Divers.* **1**, 1–3.

Gilbert, W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901–905.

Glaser, S. M., Yelton, D. E., and Huse, W. D. (1992) Antibody engineering by codon-based mutagenesis in a filamentous phage vector system. *J. Immunol.* **149**, 3903–3913.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996) Life with 6000 genes. *Science* **274**, 546.

Gold, L. (1995) Oligonucleotides as research, diagnostic, and therapeutic agents. *J. Biol. Chem.* **270**, 13581–13584.

Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. (1995) Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763–797.

Gordon, E. M., Barrett, R. W., Dower, W. J., Fodor, S. P. A., and Gallop, M. A. (1994) Applications of combinatorial

technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.* **37**, 1385–1401.

Hanes, J. and Pluckthun, A. (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc Natl. Acad. Sci. USA* **94**, 4937–4942.

Hawkins, R. E., Russell, S. J., and Winter, G. (1992) Selection of phage antibodies by binding affinity — mimicking affinity maturation. *J. Mol. Biol.* **226**, 889–896.

Hoogenboom, H. R. (1997) Designing and optimizing library selection strategies for generating high-affinity antibodies. *Trends Biotech.* **15**, 62–70.

Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C. T., and Cuervo, J. H. (1991) Generating and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **354**, 84–86.

Huse, W. D., Sastry, L., Iverson, S. A., Kang, A. S., Alting-Mees, M., Burton, D. R., Benkovic, S. J., and Lerner, R. A. (1989) Generation of a large combinatorial library of the immunoglobulin repertoire in phage λ. *Science* **246**, 1275–1281.

Illangasekare, M., Sanchez, G., Nickles, T., and Yarus, M. (1995) Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* **267**, 643–647.

Janda, K. D., Lo, C.-H. L., Li, T., Barbas III, C. F., Wirshing, P., and Lerner, R. A. (1994) Direct selection for a catalytic mechanism from combinatorial antibody libraries. *Proc. Natl. Acad. Sci. USA* **91**, 2532–2536.

Janda, K. D., Lo, L. C., Lo, C. H. L., Sim, M. M., Wang, R., Wong, C. H., and Lerner, R. A. (1997) Chemical selection for catalysis in combinatorial antibody libraries. *Science* **275**, 945–948.

Jencks, W. P. (1969) *Catalysis in Chemistry and Enzymology*, McGraw-Hill, New York.

Jiang, F., Kumar, R.A., Jones, R.A., and Patel, D.J. (1996) Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature* **382**, 183–186.

Joyce, G. F. (1989) Amplification, mutation and selection of catalytic RNA. *Gene* **82**, 83–87.

Joyce, G. F. (1997) Evolutionary chemistry: getting there from here. *Science* **276**, 1658–1659.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680–1685.

Kauffman, S. (1993) *The Origins of Order*. Oxford University Press, Oxford.

Kobe, B. (1996) Leucines on a roll. *Nature Struct. Biol.* **3**, 977–980.

Korotkov, E. V. and Korotkova, M. A. (1995) Latent periodicity of DNA sequences from some human gene regions. *DNA Seq.* **5**, 353–358.

Kozlovska, T. M., Cielens, I., Vasiljeva, I., Strelnikova, A., Kazaks, A., Dislers, A., Dreilina, D., Ose, V., Gusars, I., and Pumpens, P. (1996) RNA phage Qβ coat protein as a carrier for foreign epitopes. *Intervirology* **39**, 9–15.

Kramer, F. R., Mills, D. R., Cole, P. E., Nishihara, T., and Spiegelman, S. (1974) Evolution *in vitro*: sequence and phenotype of a mutant RNA resistant to ethidium bromide. *J. Mol. Biol.* **89**, 719–736.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–157.

Kumar, P. K. R., Nishikawa, S., and Nakamura, Y. (1995) Evolutionary molecular engineering in Japan. *Mol. Diversity* **1**, 135–137.

Lam, K. S., Salmon, S. E., Hersh, E. M., Hruby, V. J., Kazmierski, W. M., and Knapp, R. J. (1991) A new type of synthetic peptide library for identifying ligand-binding activity. *Nature* **354**, 82–84.

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., and Ghadiri, M. R. (1996) A self-replicating peptide. *Nature* **382**, 525–528.

Lerner, R. A., Benkovic, S. J., and Schultz, P. G. (1991) At the crossroads of chemistry and immunology: catalytic antibodies. *Science* **252**, 659–667.

Lerner, R. A., Kang, A. S., Bain, J. D., Burton, D. R., and Barbas III, C. F. (1992) Antibodies without immunization. *Science* **258**, 1313–1314.

Li, T. and Nicolaou, K. C. (1994) Chemical self-replication of palindromic duplex DNA. *Nature* **369**, 218–221.

Lorsch, J. R. and Szostak, J. W. (1994) *In vitro* evolution of new ribozymes with polynucleotide kinase activity. *Nature* **371**, 31–36.

Lu, Z., Murray, K. S., Cleave, V. V., LaVallie, E. R., Stahl, M. L., and McCoy, J. M. (1995) Expression of thioredoxin random peptide libraries on the *Escherichia coli* cell surface as functional fusions to flagellin: a system designed for exploring protein-protein interactions. *Bio/Technology* **13**, 366–372.

Marks, J. D., Hoogenboom, H. R., Griffiths, A. D., and Winter, G. (1992) Molecular evolution of proteins on filamentous phage — mimicking the strategy of the immune system. *J. Biol. Chem.* **267**, 16007–16010.

Matteucci, M. D. and Heyneker, H. L. (1983) Targeted random mutagenesis: the use of ambiguously synthesized oligonucleotides to mutagenize sequences immediately 5′ of an ATG initiation codon. *Nucleic Acids Res.* **11**, 3113–3121.

Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994) An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl. Acad. Sci. USA* **91**, 9022–9026.

McCafferty, J., Jackson, R. H., and Chiswell, D. J. (1991) Phage-enzymes: expression and affinity chromatography of functional alkaline phosphatase on the surface of bacteriophage. *Protein Engng.* **4**, 955–961.

Mikawa, Y. G., Maruyama, I. N., and Brenner, S. (1996) Surface display of proteins on bacteriophage lambda heads. *J. Mol. Biol.* **262**, 21–30.

Mikheeva, S. and Jarrell, K. A. (1996) Use of engineered ribozymes to catalyze chimeric gene assembly. *Proc. Natl. Acad. Sci. USA* **93**, 7486–7490.

Miller, S. L. (1953) A production of amino acids under possible primitive Earth conditions. *Science* **117**, 528–530.

Milligan, J. F., Groebe, D. R., Witherell, G. W., and Uhlenbeck, O. C. (1987) Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res.* **15**, 8783–8798.

Mills, D. R., Peterson, R. L., and Spiegelman, S. (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA* **58**, 217–224.

Mottershead, D., van der Linden, I., von Bonsdorff, C. H., Keinanen, K., and OkerBlom, C. (1997) Baculoviral display of the green fluorescent protein and rubella virus envelope proteins. *Biochem. Biophys. Res. Commun.* **238**, 717–722.

Myers, P. L. (1997) Will combinatorial chemistry deliver real medicines? *Curr. Opin. Biotechnol.* **8**, 701–707.

Needels, M. C., Jones, D. G., Tate, E. H., Heinkel, G. L., Kochersperger, L. M., Dower, W. J., Barrett, R. W., and Gallop, M. A. (1993) Generation and screening of an oligonucleotide-encoded synthetic peptide library. *Proc. Natl. Acad. Sci. USA* **90**, 10700–10704.

Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and Yanagawa, H. (1997) *In vitro* virus: bonding of mRNA bearing puromycin at the 3′-terminal end to the C-terminal end of its encoded protein on the ribosome *in vitro*. *FEBS Lett.* **414**, 405–408.

Nord, K., Nilsson, J., Nilsson, B., Uhlén, M., and Nygren, P. A. (1995) A combinatorial library of an alpha-helical bacterial receptor domain. *Protein Engng.* **8**, 601–608.

Nord, K., Gunneriusson, E., Ringdahl, J., Stahl, S., Uhlén, M., and Nygren, P.A. (1997) Binding proteins selected from combinatorial libraries of an α helical bacterial receptor domain. *Nature Biotech.* **15**, 772–777.

Ohlmeyer, M. H., Swanson, R. N., Dillard, L. W., Reader, J. C., Asouline, G., Kobayashi, R., Wigler, M., and Still, W. C. (1993) Complex synthetic chemical libraries indexed with molecular tags. *Proc. Natl. Acad. Sci. USA* **90**, 10922–10926.

Ohno, S. (1981) Original domain for the serum albumin family arose from repeated sequences. *Proc. Natl. Acad. Sci. USA* **78**, 7657–7661.

Ohno, S. (1987) Evolution from primordial oligomeric repeats to modern coding sequences. *J. Mol. Evol.* **25**, 325–329.

Ohno, S. (1989) Modern coding sequences are in the periodic-to-chaotic transition. *Hamatol. Bluttransfus.* **32**, 512–519.

Ohno, S. (1992) Of palindromes and peptides. *Hum. Genet.* **90**, 342–345.

Ohno, S. (1994) The cardinal principle of like attracting like generates many ubiquitous oligopeptides shared by divergent proteins. *Animal Genet.* **25s**, 5–11.

Ohno, S. and Matsunaga, T. (1982) The 48-base-long primordial building block of immunoglobulin light-chain variable regions is complementary to the primordial building block of heavy-chain variable regions. *Proc. Natl. Acad. Sci. USA* **79**, 2338–2341.

Ohno, S. and Epplen, J. T. (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. *Proc. Natl. Acad. Sci. USA* **80**, 3391–3395.

Oparin, A. I. (1938) *The Origin of Life*, MacMillan, New York.

Palmer, J. D. and Logsdon, J. J. M. (1991) The recent origins of introns. *Curr. Opin. Genet. Develop.* **1**, 470–477.

Pauling, L. (1948) Chemical achievement and hope for the future. *Am. Sci.* **36**, 51.

Pollack, S. J., Jacobs, J. W., and Schultz, P. G. (1986) Selective chemical catalysis by an antibody. *Science* **234**, 1570–1573.

Prijambada, I. D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S., and Urabe, I. (1996) Solubility of artificial proteins with random sequences. *FEBS Lett.* **382**, 21–25.

Rebar, E. J. and Pabo, C. O. (1994) Zinc finger phage — affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671–673.

Roberts, B. L., Markland, W., Ley, A. C., Kent, R. B., White, D. W., Guterman, S. K., and Ladner, R. C. (1992) Directed evolution of a protein — selection of potent neutrophil elastase inhibitors displayed on M13 fusion phage. *Proc. Natl. Acad. Sci. USA* **89**, 2429–2433.

Roberts, R. W. and Szostak, J. W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl. Acad. Sci. USA* **94**, 12297–12302.

Robertson, D. L. and Joyce, G. F. (1990) Selection *in vitro* of an RNA enzyme that specifically cleaves single stranded DNA. *Nature* **344**, 467–468.

Rojas, N. R. L., Kamtekar, S., Simons, C. T., McLean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S., and Hecht, M. H. (1997) *De novo* heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512–2524.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **29**, 487–491.

Schultz, J. S. and Schultz, J. S. (1996) The combinatorial library: a multifunctional resource. *Biotechnol. Progr.* **12**, 729–743.

Scott, J. K. and Smith, G. P. (1990) Searching for peptide ligands with an epitope library. *Science* **249**, 386–390.

Seidel, H. M., Pompliano, D. L., and Knowles, J. R. (1992) Exons as microgenes? *Science* **257**, 1489–1490

Shiba, K. (1995) Dissection of an enzyme into two fragments at intron-exon boundaries; in *Tracing Biological Evolution in Protein and Gene Structures*, Go, M. and Schimmel, P. (eds.), pp. 11–21, Elsevier Science Publishers, Amsterdam.

Shiba, K. (1997) Creation of genes (Japanese). *KAGAKU* **67**, 938–947.

Shiba, K. and Schimmel, P. (1992a) Functional assembly of a randomly cleaved protein. *Proc. Natl. Acad. Sci. USA* **89**, 1880–1884.

Shiba, K. and Schimmel, P. (1992b) Tripartite functional assembly of a large class I aminoacyl tRNA synthetase. *J. Biol. Chem.* **267**, 22703–22706.

Shiba, K., Hatada, T., and Noda, T. (1996) Combinatorial assembly of microgenes. *Protein Engng.* **9**, 813–814.

Shiba, K., Motegi, H., and Schimmel, P. (1997) Maintaining genetic code through adaptations of tRNA synthetases to taxonomic domains. *Trends Biochem. Sci.* **22**, 453–457.

Shiba, K., Takahashi, T., and Noda, T. (1997) Creation of libraries with long open reading frames by polymerization of a microgene. *Proc. Natl. Acad. Sci. USA* **94**, 3805–3810.

Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317.

Solovyev, V. V. (1993) Fractal graphical representation and analysis of DNA and protein sequences. *BioSystems* **30**, 137–160.

Stemmer, W. P. C. (1994a) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391.

Stemmer, W. P. C. (1994b) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* **91**, 10747–10751.

Stemmer, W. P. C. (1995) Searching sequence space — using recombination to search more efficiently and thoroughly instead of making bigger combinatorial libraries. *Bio/*

Technology **13**, 549–553.

Sternberg, N. and Hoess, R. H. (1995) Display of peptides and proteins on the surface of bacteriophage λ. *Proc. Natl. Acad. Sci. USA* **92**, 1609–1613.

Szostak, J. W. (1992) *In vitro* genetics. *Trends Biochem. Sci.* **17**, 89–93.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**, 631–637.

Tramontano, A., Janda, K. D., and Lerner, R. D. (1986) Catalytic antibodies. *Science* **234**, 1566–1570.

Tirrell, D. A., Fournier, M. J., and Mason, T. L. (1991) Protein engineering for materials applications. *Curr. Opin. Struct. Biol.* **1**, 638–641.

Tsonis, A. A. and Tsonis, P. A. (1997). Simplicity and complexity in gene evolution *Complexity* **2(5)**, 23–30.

Tsonis, A. A., Elsner, J. B., and Tsonis, P. (1991) Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol.* **151**, 323–331.

Tuerk, C. and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment — RNA ligands to bacteriophage-T4 DNA polymerase. *Science* **249**, 505–510.

White, S. H. and Jacobs, R. E. (1993) The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.* **36**, 79–95.

Wilson, C. and Szostak, J. W. (1995) *In vitro* evolution of a self-alkylating ribozyme. *Nature* **374**, 777–782.

Wolfe, K. H. and Shields, D. C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713.

Yamauchi, A., Yomo, T., Tanaka, F., Prijambada, I. D., Ohhashi, S., Yamamoto, K., Shima, Y., Ogasahara, K., Yutani, K., Kataoka, M., and Urabe, I. (1998) Characterization of soluble artificial proteins with random sequences. *FEBS Lett.* **421**, 147–151.

Yura, K., Tomoda, S., and Go, M. (1993) Repeat of α helix-turn-helix module in DNA-binding proteins. *Protein Engng.* **6**, 621–628.

Zhang, J. H., Dawes, G., and Stemmer, W. P. C. (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. USA* **94**, 4504–4509.

Zoller, M. J. and Smith, M. (1982) Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any fragment of DNA. *Nucleic Acids Res.* **10**, 6487–6500.