PRODUCTION OF FUNCTIONAL HYBRID GENES AND PROTEINS

FIELD OF THE INVENTION

This invention relates to methods for creating novel DNA and amino acid sequences, especially the production of gene libraries encoding polypeptides or proteins, and corresponding protein libraries. Libraries can be made for individual proteins, or for hybrid proteins comprised of fragments from different proteins. The method can also be used to make random circular permutations of a protein. Random protein fragments can be made that complement one another to form a functional protein. Protein fragments that assemble to form a functional protein can be identified, for example by screening or selection. The method can also be used to create novel hybrid or chimeric proteins by assembling fragments taken from different parent proteins, or by creating circular permutations containing fragments from different parent proteins. For example, DNA libraries can be made which encode for hybrid or chimeric proteins of an N-terminal part originating from one protein fused to a C-terminal part of another protein. Screening or selection of the resulting library can be used to identify proteins with useful properties. Genes encoding any useful protein or proteins can be used as parents, starting materials, or templates for the invention. This invention also relates to a method which may be used for the selection of gene library repertoires that have a continuous reading frame.

BACKGROUND OF THE INVENTION

The publications and reference materials noted herein and listed in the appended Bibliography are each incorporated by reference in their entirety.

New and useful proteins can be obtained in many ways, often by altering known proteins to obtain new or altered properties. One strategy to generate proteins with improved properties over existing, *i.e.* wild type, proteins is called directed evolution. For this purpose, DNA recombination techniques, including techniques known as "DNA shuffling", have become powerful tools. In one technique, called bisection, it has been found that there are proteins which tolerate being cut or synthesized as separate fragments. The fragments can be reassembled *in vitro* or *in vivo* to yield functional proteins in the form of dimers. This method, called protein fragment complementation, is thought to rely on interchain packaging interactions between the protein fragments to restore biological function. (*See* Bibi and Kaback, 1990; Burbaum and Schimmel, 1991; Hall and Frieden, 1989; Hantgan and Taniuchi, 1977; Labhardt, 1982; Shiba and Schimmel, 1992; Taniuchi et al., 1977; and Yang and Schachman, 1993.)

In a variation of these methods, a protein or polypeptide can be connected via the original amino (N) and carboxy (C) terminals and bisected to yield new molecules in a process known as circular permutation. (*See* Mullins et al., 1994; Protasova et al., 1994; Vignais et al., 1995; Yang and Schachman 1993; and Zhang et al., 1993.) Circular permutation reorganizes the primary sequence of the protein so that the original amino and carboxy terminals are covalently closed, and new terminals are created at a different site within the sequence. Covalent closure of the natural terminals can involve insertion of one or more amino acids, for example if the terminals are not close enough in space to be directly linked to each other. Proteins reorganized in this way may retain some or all of their original biological function and properties, and may have new functions or properties.

Traditionally, cleavage or bisections sites, and the sites for new amino and carboxy terminals of circularly permuted proteins, are chosen based on some knowledge of the protein structure or behavior. Typical sites have included, for example, cleavage sites of limited proteolytic digestion, or regions of the protein thought to be flexible (e.g. loops).

Graf and Schachman, 1996, produced variants of aspartate transcarbamoylase (ATC) by random circular permutation, and also by constructing a gene homodimer connected by a short linker sequence. Thereafter, the gene dimer was cut at a specific site by digestion with

a restriction endonuclease that recognizes a unique site in the gene. After circularizing the obtained fragments by ligation, using the cohesive ends left by the digest, the fragments were randomly linearized by treatment with DNase I. This approach created random circular permutations or permutants of one protein, but could not be used to create libraries for protein fragment complementation.

Another method for creating a library of complementary protein fragments is suggested by Ostermeier et al., 1999. In this method, a library is made from two different parent proteins, to create heterodimers in which one fragment comes from a first parent and the other from a second parent. However, incremental gene truncation is used to create a library of fragment pairs which are largely useless or non-functional. Most fragments can not be combined to produce or sum up to a single complete protein. Also, a large fraction of the protein fragments will not be able to fold properly, or will not be able to dimerize.

Various DNA shuffling methods have been developed to produce protein libraries that are hybrids between two or more parent proteins. These include *in vitro* methods (*See* Shao et al.,1998; Stemmer 1994; and Zhao et al.,1998) and *in vivo* methods (*See* Okkels, 1997; and Volkov, 1999) These methods use protein monomers, and they require regions of high DNA sequence identity between parent proteins, generally at least 80%. Fragments obtained from more distant parents can not be recombined, for example because crossover will not occur. However, many evolutionarily related proteins have highly similar structures, and may have similar functions, but they do not share much or a high degree of DNA sequence identity. For example, it is not uncommon for proteins which have similar three-dimensional structures to have only 20-30% sequence identity.

Accordingly, there is a need for combinatorial approaches to protein design which do not require homology or a high degree of sequence identity. Methods which do not depend on detailed knowledge of parent proteins would also be useful. Given the vast numbers of proteins about which little is known, there is a need for random or unbiased methods to identify cleavage sites that yield functional proteins when one or more parent proteins are permutated, or when protein fragments are made for complementation.

Furthermore, there is a need for a method to recombine sequence elements from multiple proteins which do not require high levels of DNA sequence identity. Such a method could also be used to create a library of hybrid, or chimeric, proteins with fragments from two (or more) parent sequences. In particular, random methods which are capable of producing

relatively large numbers of hybrids relatively quickly, and from which molecules exhibiting desired characteristics can be identified by screening, would be advantageous.

In addition, the current techniques for protein fragment complementation and circular permutation tend to generate a bisection of each polypeptide chain somewhere between the original N- and C- terminals of the protein. This may impair protein folding, and also influence or impair the ability of the resulting hybrid protein to function. (*See* Graf et al., 1996; and Hennecke et al., 1999). Such a constraint may severely and unnecessarily limit the proportion of functional proteins in the library. Thus, there is a need for techniques preserving or recreating the original N– and C-terminal sequences of the parent protein(s).

SUMMARY OF THE INVENTION

The present invention provides an improved method for creating gene and protein libraries. The invention can be used to make random libraries of circularly permuted variants of genes encoding a single protein, or hybrid proteins containing fragments from two or more parent proteins. The invention can also be used to create a library for protein fragment complementation, in which fragments originate either from one protein or from different proteins, typically from two proteins. In one embodiment, hybrid proteins can be created from two parent proteins independent of sequence similarity between the parent proteins. In another embodiment, the invention may be used to create a library of genes which are mostly of a size appropriate for successful recombination into full-length proteins. This provides a significantly high likelihood that a relatively high fraction of complements in the library will be functional. In still another embodiment, the invention can be used to create a library of truncated or elongated genes from one or more parent genes.

The invention also provides a method to create libraries of hybrid proteins, especially single-chain hybrids, that may have an N-terminal part originating from one protein fused to a C-terminal part of a second protein, with both parts varying in length, while the total length is comparable to a parent protein. This technique is designed to further increase the fraction of functional proteins expressed or produced using hybrid genes in the library. Methods provided herein can also be used to create a random library of single-chain hybrid proteins that consist of fragments of several proteins. In addition, the method can be used to make libraries of proteins that have small interior sequence duplications or deletions of random length and

at random positions. Especially, the invention can produce libraries of hybrid proteins with preserved N– and C-terminal sequences.

Thus, the invention provides a method which, particularly when coupled with screening or selection, can be used to create and identify new gene and protein libraries, including new proteins with useful properties. The basic strategy for creating these libraries involves manipulation of the DNA encoding the protein or proteins, followed by expression, either *in vivo* (*e.g.* in host cells) or *in vitro*.

Random Fragments for Protein Fragment Complementation and Circular Permutation

In one embodiment, a gene dimer is constructed as a homodimer or heterodimer, *i.e.* a polynucleotide is made from two genes or portions of genes that encode for the same protein or for different proteins. Typically, each dimer comprises two complete and non-identical genes, placed in tandem on a single piece of DNA, and separated by a linker sequence. The linker sequence encodes for at least one restriction site that is unique in the dimer construct. If so desired, gene concatemers can also be made.

To construct a random circular permutation library, the linker sequence is preferably designed such that the reading frame is continuous, and the original 5' and 3' terminal ends of the structural gene are connected. Appropriate linker sequences can either insert, delete, or mutate amino acids in the protein sequence, or they can leave the protein sequence unchanged, except for covalent attachment of the N– and C-terminal amino acids.

To construct a library for protein fragment complementation, the linker sequence should encode a stop translation signal of the upstream gene fragment of the dimer and a translation initiation signal of the downstream gene fragment of the dimer.

The gene dimer can be constructed, for example, using the polymerase chain reaction and subcloned into a suitable vector for amplification. The constructed gene dimer is then excised and purified after separation from other components of the mixture. The purified gene dimer is subjected to limited fragmentation, resulting in a mixture consisting of DNA fragments varying in size. From this mixture, fragments having a predetermined size, or being within a predetermined size range, can be isolated. In one approach, DNA fragments approximately the size of a gene monomer are isolated using any one of a range of techniques, including gel electrophoresis. The resulting DNA will consist of a population of DNA

molecules approximately the size of the parent gene or genes, but with different 5' and 3' termini.

The purified DNA is then treated as necessary and ligated into a suitable expression plasmid to create a library of random circular permuted genes or proteins, or a library for protein fragment complementation. The expression plasmid can be used to transform a suitable host for expression of the proteins. The genes can also be expressed by phage display (Johansson et al., 1999) or *in vitro* transcription-translation systems. Functional circular permutants or complementary fragments that yield functional protein are identified by screening or selection. Optionally, the repertoire of hybrid variants whose parental fragments are in one continuous reading frame may be increased by ligating the N-terminals of fragments in the gene library to a gene encoding for a reporter protein. Preferably, the start codon of translation (ATG) of this reporter protein has been modified (or removed) to prevent its independent translation.

Hybrid proteins with preserved terminal sequences

The invention also provides improved methods for creating functional hybrid or chimeric proteins from two or more parent proteins, by preserving the N- and C- terminals of the original protein or proteins, or by providing terminal ends which are appropriate for, or compatible with, the proteins. This includes, for example, terminals which promote functional protein folding, and is particularly useful for proteins which are sensitive to alternations in the C- or N- terminal, or which are sensitive to folding conditions associated with one or both terminal ends. To facilitate this method, gene dimers can be made with linkers which preferably have at least two unique restriction sites.

In one embodiment, randomly generated gene monomer-length DNA-fragments are circularized by ligating the 3'-end of the truncated gene (the second gene of the dimer) to the 5'-end of the truncated gene (the first gene of the dimer). This procedure results in the fusion of the corresponding new C-terminus of the second protein with the new N-terminus of the first protein. After digestion of the circular DNA fragments with, e.g., restriction enzymes that cut within the linker sequence, amplification by PCR when appropriate, and ligation into a corresponding expression vector, the resulting hybrid proteins maintain the original N-terminus of the second protein and the C-terminus of the first protein. They also contain intervening covalent crossovers between the two proteins.

In another embodiment, gene concatemers can be constructed from the same, or several different, parent genes. After one or more additional cycles of random fragmentation, selection, and circularization, a gene or protein library can be obtained which corresponds to hybrid proteins consisting of several different fragments of the parent protein(s).

The invention thus provides for methods to modify chemical, physical and/or functional properties of a protein by creating a hybrid between the protein and another protein having different properties. For example, one property residing in the N-terminal of one protein may be combined with a property residing in the C-terminal of another protein, and a hybrid protein created which fully or partially retains desirable properties of the parent proteins.

The above features and many other advantages of the invention will become better understood by reference to the following detailed description when taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A shows a schematic description of the construction of a gene dimer. Each dimer can be a homodimer or a heterodimer.

FIG. 1B shows a strategy for constructing a library of gene fragments corresponding in size to the size of the gene or genes that encode for the original protein(s).

FIG. 2A shows restriction sites for the digestion of two isolated plasmids from active clones of Green Fluorescent Protein (GFP).

FIG. 2B shows results of inserts deduced from the double enzyme digestion of isolated plasmids from active clones of Green Fluorescent Protein. The double enzyme digestion consisted of *Bam*HI + *Eco*RI, *Bam*HI + *Xho*I, and *Bam*HI+*Sfi*I. Obtained fragments (insert types) are: (a) intact GFP gene with extra fragment upstream; (b) intact GFP with an extra fragment downstream; (c) two overlapped fragments; (d) recovered wild-type or wild-type-like genes; (e) complementary fragments; and (f) truncated genes.

FIG. 3 shows the construction of a gene heterodimer according to one embodiment of the invention.

FIG. 4A and 4B show two possible strategies for constructing a library of gene fragments, using heterodimers of FIG. 3, to obtain hybrid genes corresponding in size to the size of the genes that encode the original proteins.

FIG. 5 shows a strategy for constructing a library of hybrid proteins with one crossover between the two parent proteins, or with small interior sequence deletions or duplications. "**X**" designates the position of the crossover between the two proteins.

FIG. 6 shows one strategy for constructing a library of hybrid proteins with several crossover points between two or more parent proteins. "**X**" designates the position of the crossovers between the different proteins.

FIG. 7 shows another strategy for constructing a library of hybrid proteins with several crossover points between two or more parent proteins.

FIG. 8 shows N-terminal nucleotide and amino acid sequences for two hybrid proteins constructed according to the invention. Sequences originating from BM3 are in bold letters and sequences originating from 1A2 are in italic letters.

FIG. 9 shows the nucleotide sequence for human cytochrome P450 1A2 having a modified N-terminus (*See* Fischer et al., 1992) **[SEQ ID NO: 27]**.

FIG. 10 shows the nucleotide sequence for the heme domain of mutant P450 BM3 (*See* Schwaneberg et al., 1999) **[SEQ ID NO: 28]**.

FIG. 11 shows the nucleotide sequence for a hybrid gene of the invention (RC1) **[SEQ ID NO: 29]**.

FIG. 12 shows the nucleotide sequence for a hybrid gene of the invention (RC2) **[SEQ ID NO: 30]**.

FIG. 13 shows the nucleotide sequence for a hybrid gene of the invention (RC3) **[SEQ ID NO: 31]**.

FIG. 14 shows the nucleotide sequence for a hybrid gene of the invention (RC4) **[SEQ ID NO: 32]**.

FIG. 15 shows the nucleotide sequence for a hybrid gene of the invention (RC5) **[SEQ ID NO: 33]**.

DETAILED DESCRIPTION OF THE INVENTION

The object of this invention is to provide improved methods for creating novel protein sequences, including the production of libraries of genes encoding for polypeptides or proteins. These methods involve the creation of gene or protein libraries for single proteins, and for hybrid proteins which contain fragments from several different proteins. In a preferred embodiment, the protein library is constructed so that the N– and C-terminal ends of the protein are preserved. In particular, the method provides for the efficient creation of random or partially random libraries which can be screened for functional proteins.

Definitions

In any identified embodiments, the terms *about*, *approximately*, and variants thereof, means within 50%, preferably within 25%, and more preferably within 10% of a given value or range. Alternatively, the term "about" means that the value is within an acceptable standard error of the mean, when considered by one of ordinary skill in the art.

The term *library*, as used herein, means a collection of proteins, polypeptides or polynucleotides. A *gene* or *DNA library* is a collection of polynucleotides or DNA sequences, and generally includes polynucleotides or sequences that correspond to, are derived from, or are in some way related to one or more parent genes that can be expressed to produce one or more polypeptides or proteins. A *protein library* is a collection of polypeptides or amino acid

sequences that correspond to, are derived from, or are in some way related to one or more parent polypeptides or proteins, and may also encompass a corresponding gene library.

A protein, polypeptide, polynucleotide or gene, may be *native* or *wild-type*, meaning that it occurs in nature; or it may be a *hybrid, mutant*, *variant* or *modified*, meaning that it has been made, altered, derived, or is in some way different or changed from a native protein or gene, or from another mutant. A hybrid gene or protein can also be called a *chimeric* gene or protein.

A *crossover* is used to describe a point in a hybrid or chimeric polynucleotide or polypeptide sequence at which a section of the hybrid or chimeric polynucleotide or polypeptide sequence originating from one parent is connected to a section originating from another parent.

A *parent* or *template* polynucleotide or gene, is any polynucleotide or gene from which any other polynucleotide or gene is derived or made, using any methods, tools or techniques, and whether or not the parent is itself a native or mutant polynucleotide or gene. Likewise, a parent or template polypeptide or protein is any polypeptide or protein from which any other polypeptide or protein is derived or made, using any methods, tools or techniques, and whether or not the parent is itself a native or mutant polypeptide or protein.

The terms *monomer*, *dimer*, or *polymer* describe a polypeptide, polynucleotide, protein, or gene, in the form of one, two, or several components, or "mers", respectively. Further, a "homodimer" may be a polypeptide, polynucleotide, protein, or gene, made from two components originating from the same parent polypeptide, polynucleotide, protein, or gene, in native or modified form. A "heterodimer" is a polypeptide, polynucleotide, protein, or gene made from two components originating from different parents, each of which encodes or corresponds to all or part of a different protein, native or modified. The term gene "concatemer" herein is a polynucleotide consisting of several genes or gene fragments, from one or more parent genes, in sequence with or without linker DNA in between each fragment.

The term *fragment* means any part of a larger whole, including any rearrangement of parts which make up the whole. This includes polypeptide sequences obtained from, or corresponding to, all or part of the amino acid sequence of a functional protein. The term fragment also includes polynucleotide sequences obtained from or corresponding to all or part of the nucleotide sequence of a gene. For example, in the molecular cloning of a gene from genomic DNA, DNA fragments are generated, some of which will encode the desired gene.

The DNA may be *restricted*, *i.e.*, cleaved or cut, into fragments at specific sites using various restriction enzymes. Any suitable restriction enzyme may be used, including, but not limited to, *Xho*I, *Eco*RI, *Pst*I, *Sac*I, *Hind* III, *Stu*I, *Xba*I, *Bam*HI, *Sal*I, and *Mfe*I, Alternatively, one may use DNase in the presence of manganese to fragment or digest the DNA, or the DNA can be physically sheared, as for example, by sonication. DNA fragments can then be separated according to size by standard techniques, including but not limited to, agarose and polyacrylamide gel electrophoresis and column chromatography.

A *limited* treatment or digestion of DNA means to treat or digest DNA under such conditions that a substantial portion of the treated or digested DNA fragments are approximately of a predetermined size, or approximately within a predetermined size range. The degree of digestion can be controlled, *e.g.*, by limiting the time of the treatment/digestion process, or by altering the treatment/digestion conditions so as to slow down or limit the DNA fragmentation. The optimal processing time and/or conditions to achieve the desired degree of DNA fragmentation are advantageously determined experimentally for each specific treatment/digestion.

A *polypeptide* (one or more peptides) or *protein* is a chain of chemical building blocks called amino acids that are linked together by chemical bonds called peptide bonds.

The *properties* of a polypeptide or protein include chemical, physical, or functional properties, which may be derived from characteristics such as amino acid composition and peptide chain folding. Chemical and physical properties are represented by, *e.g.*, charge, isoelectric point (IP), water solubility, cell membrane solubility and/or binding, hydrophobicity, hydrophilicity, lipophobicity, lipophilicity, size, and stability. Functional properties of a protein or enzyme include, but are not limited to, foldability (*i.e.*, the ability of the enzyme to fold in the desired manner), expressability (*i.e.*, the ability of the enzyme to be expressed in the desired manner and/or amount), the specific reaction catalyzed, substrate specificity, reaction product, and enzyme activity.

A *membrane-associated* protein or a polypeptide is a protein or polypeptide which can have least one part of its polypeptide chain integrated or associated with a cell membrane.

DNA (deoxyribonucleic acid) means any chain or sequence of the chemical building blocks adenine (A), guanine (G), cytosine (C) and thymine (T), called nucleotide bases, that are linked together on a deoxyribose sugar backbone. DNA can have one strand of nucleotide bases, or two complimentary strands which may form a double helix structure.

A *polynucleotide*, *nucleotide sequence* or *oligonucleotide* is a series of nucleotide bases (also called "nucleotides") in DNA, and means any chain of two or more nucleotides. A nucleotide sequence typically carries genetic information, including the information used by cellular machinery to make polypeptides, proteins and enzymes. These terms include double or single stranded genomic and cDNA, as well as any synthetic and genetically manipulated polynucleotide. The DNA and polynucleotides herein may be flanked by natural regulatory sequences, or may be associated with heterologous (non-native) sequences, including promoters, enhancers, response elements, signal sequences, polyadenylation sequences, introns, 5'- and 3'- non-coding regions, linker regions, sequences containing specific sites recognized by restriction enzymes, and the like. The nucleic acids in the present invention may also be modified by the many means known in the art.

The single- or double-stranded polynucleotide sequences described herein may be multiplied or *amplified* by any means know in the art. One preferred technique is the polymerase chain reaction, or *PCR*. Generally, PCR involves the use of (1) one or more templates, which in this context relates to DNA sequences to be amplified; and (2) *primers*, which are DNA sequences, generally of limited length, which are specific for or complementary to regions of DNA. Primers may thus be used to, *e.g.*, initiate DNA polymerization *in vitro* in the presence of DNA polymerase. When coupled to a reporter molecule such as a radionuclide or a fluorescent molecule, primers may also be used to identify whether a certain DNA segment contains a complementary sequence. If desired, *error-prone* PCR may be used to create variants or mutants of a template molecule.

The single- or double-stranded polynucleotide sequences described herein may be *ligated*, *i.e.*, joined. For example, several DNA strands can be joined to one *linear* sequence, forming e.g. a gene dimer, concatemer, or the like. Also, a *circular* or *circularized* polynucleotide can be obtained when ligating the ends of one single strand of DNA, a process which may also be referred to as *circularization*. The term *linearization* can be used to describe the formation of a linear or *linearized* sequence from a circular sequence by, *e.g.*, cutting the circular sequence with a restriction or other enzyme. Any methods known in the art may be used for DNA ligation. Ligation conditions may be designed to favor circularization over concatemerization, or the reverse, by e.g. choice of DNA concentration, or treating the ends of the DNA strands. An example of the latter is to convert *staggered* ends, having single-stranded cohesive ends, to *blunt* ends, or by treating the DNA strands with

suitable restriction enzymes. Further, the term ligation may also be used e.g. in a context describing the insertion of a gene into a vector, as described herein.

Transcriptional and *translational* control sequences are DNA regulatory sequences, such as promoters, enhancers, terminators, and the like, that provide for the expression of a coding sequence in a host cell. In eukaryotic cells, polyadenylation signals are control sequences. A *promoter* or *promoter sequence* is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. The promoter sequence is bounded at its 3' terminus by a transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site (conveniently defined for example, by mapping with nuclease S1), as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase. As described, promoter DNA is a DNA sequence which initiates, regulates, or otherwise mediates or controls the expression of the coding DNA. A promoter may be "inducible", meaning that it is influenced by the presence or amount of another compound (an "inducer"). For example, an inducible promoter includes those which initiate or increase the expression of a downstream coding sequence in the presence of a particular inducer compound. A "leaky" inducible promoter is a promoter that provides a high expression level in the presence of an inducer compound and a comparatively very low expression level, and at minimum a detectable expression level, in the absence of the inducer.

A *coding sequence* or a sequence *encoding* a polypeptide, protein or enzyme is a nucleotide sequence that, when expressed, results in the production of that polypeptide, protein or enzyme, i.e., the nucleotide sequence encodes an amino acid sequence for that polypeptide, protein or enzyme. A coding sequence is under the control of transcriptional and translational control sequences in a cell when RNA polymerase transcribes the coding sequence into mRNA, which is then trans-RNA spliced and translated into the protein encoded by the coding sequence. Preferably, the coding sequence is a double-stranded DNA sequence which is transcribed and translated into a polypeptide in a cell in vitro or in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a translation stop codon at the 3' (carboxyl) terminus. More than one stop codon can be used to terminate

the transcription of a DNA sequence. For example, to ensure termination of transcription of a DNA segment that has been truncated at the 5' and/or 3' end, stop codons can be provided in all three reading frames proximal, *i.e.* near, the 3' end. A coding sequence can include, but is not limited to, prokaryotic sequences, cDNA from eukaryotic mRNA, genomic DNA sequences from eukaryotic (e.g., mammalian) DNA, and even synthetic DNA sequences. If the coding sequence is intended for expression in a eukaryotic cell, a polyadenylation signal and transcription termination sequence will usually be located 3' to the coding sequence.

The term *gene*, also called a *structural gene* means a DNA sequence that codes for or corresponds to a particular sequence of amino acids which comprise all or part of one or more proteins or enzymes, and may or may not include regulatory DNA sequences, such as promoter sequences, which determine for example the conditions under which the gene is expressed. A gene encoding a protein of the invention for use in an expression system, whether genomic DNA or cDNA, can be isolated from any source, particularly from a human cDNA or genomic library. Methods for obtaining genes are well known in the art. (*See* e.g. Sambrooke et al., 1989) Accordingly, any animal cell potentially can serve as the nucleic acid source for the molecular cloning of the gene of interest. The DNA may be obtained by standard procedures known in the art, such as from cloned DNA (*e.g.*, a DNA "library"), from cDNA library prepared from tissues with high level expression of the protein, by chemical synthesis, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell. Clones derived from genomic DNA may contain regulatory and intron DNA regions in addition to coding regions; clones derived from cDNA will not contain intron sequences.

Proteins and enzymes are made in the host cell using instructions in DNA and RNA, according to the genetic code. Generally, a DNA sequence having instructions for a particular protein or enzyme is *transcribed* into a corresponding sequence of RNA. The RNA sequence in turn is *translated* into the sequence of amino acids which form the protein or enzyme.

The term *reporter* herein means any molecule, or a portion thereof, that is detectable, or measurable, for example, by optical detection. In addition, the reporter may associate or be associated with a molecule or a particular marker or characteristic of the molecule, or is itself detectable, to permit identification of the molecule or the presence or absence of a characteristic of the molecule. In the case of molecules such as polynucleotides such characteristics include size, molecular weight, the presence or absence of particular

constituents or moeties (such as particular nucleotide sequences or restrictions sites), and polypeptides which the reporter polynucleotide encodes. The term *label* can be used interchangeably with "reporter". The reporter is typically a dye, fluorescent, ultraviolet, or chemiluminescent agent, chromophore, or radiolabel, any of which may be detected with or without some kind of stimulatory event, e.g., fluoresce with or without a reagent. A reporter protein or polypeptide can be expressed from a reporter polynucleotide *in vitro* or in a cell, and such expression may be indicative of the presence of another protein that may or may not be coexpressed with the reporter. A reporter may also include any substance on or in a cell that causes a detectable reaction, for example by acting as a starting material, reactant or a catalyst for a reaction which produces a detectable product.

An *amino acid sequence* is any chain of two or more amino acids. Each amino acid is represented in DNA or RNA by one or more triplets of nucleotides. Each triplet forms a codon, corresponding to an amino acid. For example, the amino acid lysine (Lys) can be coded by the nucleotide triplet or codon AAA or by the codon AAG. (The genetic code has some redundancy, also called degeneracy, meaning that most amino acids have more than one corresponding codon.) Because the nucleotides in DNA and RNA sequences are read in groups of three for protein production, it is important to begin reading the sequence at the correct amino acid, so that the correct triplets are read. The way that a nucleotide sequence is grouped into codons is called the *reading frame*.

The terms *express* and *expression* mean allowing or causing the information in a gene or DNA sequence to become manifest, for example producing a protein by activating the cellular functions involved in transcription and translation of a corresponding gene or DNA sequence. A DNA sequence is expressed in or by a cell to form an "expression product" such as a protein. The expression product itself , e.g. the resulting protein, may also be said to be "expressed" by the cell. A polynucleotide or polypeptide is expressed recombinantly, for example, when it is expressed or produced in a foreign host cell under the control of a foreign or native promoter, or in a native host cell under the control of a foreign promoter.

The terms *vector*, *cloning vector* and *expression vector* mean the vehicle by which a DNA or RNA sequence (e.g. a foreign gene) can be introduced into a host cell, so as to transform the host and promote expression (e.g. transcription and translation) of the introduced sequence. Vectors typically comprise the DNA of a transmissible agent, into which foreign DNA is inserted. A common way to insert one segment of DNA into another

segment of DNA involves the use of enzymes called restriction enzymes that cleave DNA at specific sites (specific groups of nucleotides) called restriction sites. Generally, foreign DNA is inserted at one or more restriction sites of the vector DNA, and then is carried by the vector into a host cell along with the transmissible vector DNA. A segment or sequence of DNA having inserted or added DNA, such as an expression vector, can also be called a "DNA construct."

A common type of vector is a *plasmid*, which generally is a self-contained molecule of double-stranded DNA, that can readily accept additional (foreign) DNA and which can be readily introduced into a suitable host cell. A plasmid vector often contains coding DNA and promoter DNA and has one or more restriction sites suitable for inserting foreign DNA. Promoter DNA and coding DNA may be from the same gene or from different genes, and may be from the same or different organisms. A large number of vectors, including plasmid and fungal vectors, have been described for replication and/or expression in a variety of eukaryotic and prokaryotic hosts. Recombinant cloning vectors will often include one or more replication systems for cloning or expression, one or more markers for selection in the host, e.g. antibiotic resistance. In general, the choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention.

The term *host cell* means any cell of any organism that is selected, modified, transformed, grown, or used or manipulated in any way, for the production of a substance by the cell, for example the expression by the cell of a gene, a DNA or RNA sequence, a protein or an enzyme. Appropriate host cells for expressing protein include bacteria, Archaebacteria, fungi, especially yeast, and plant and animal cells, especially mammalian cells. Of particular interest are E. coli, B. subtilis, S. cerevisiae, Sf9 cells, C129 cells, 293 cells, Neurospora, and CHO cells, COS cells, HeLa cells, and immortalized mammalian myeloid and lymphoid cell lines.

The term *expression system* means a host cell and compatible vector under suitable conditions, e.g. for the expression of a protein coded for by foreign DNA carried by the vector and introduced to the host cell. Preferred expression systems include bacteria (e.g. *E. coli* and *B. subtilis*) or yeast (*e.g. S. cerevisiae*) host cells and plasmid vectors, and insect host cells and Baculovirus vectors.

Isolation or *purification* of a polynucleotide, DNA fragment, polypeptide, or protein refers to the derivation of the polypeptide by removing it from its original environment (for

example, from its natural environment if it is naturally occurring, or from the host cell if it is produced by recombinant DNA methods). Methods for polypeptide purification are well-known in the art, including, without limitation, preparative electrophoresis, isoelectric focusing, HPLC, reversed-phase HPLC, gel filtration, ion exchange and partition chromatography, and countercurrent distribution. A purified polynucleotide or polypeptide may contain less than about 50%, preferably less than about 75%, and most preferably less than about 90%, of the cellular components with which it was originally associated. A "substantially pure" enzyme indicates the highest degree of purity which can be achieved using conventional purification techniques known in the art.

The terms *sequence similarity* or *sequence identity* refers to the difference between the amino acid sequence of a modified protein and that of the parent protein or enzyme, or the nucleotide sequence of a modified polynucleotide or gene and that of the parent polynucleotide or gene. The percent sequence identity or similarity between any two protein, amino acid, polynucleotide, or gene sequences can be determined according to an alignment scheme, such as, *e.g.*, the Cluster Method, wherein similarity/identity is based on the MEGALIGN algorithm.

DNA shuffling is one approach to the creation of modified or hybrid proteins. For instance, a gene may be randomly fragmented and reassembled by error-prone PCR. After screening, the iterative process may be repeated until a protein with the desired properties is produced (*See*, *e.g.*, Stemmer, 1994). The term "*shuffling*" herein means performing DNA shuffling, and includes various shuffling strategies, such as for example those described in Ness et al., 1999; Chang et al., 1999; Minshull and Stemmer, 1999; Christians et al., 1999; Crameri et al., 1998; Crameri et al., 1997; Zhang et al., 1997; Patten et al., 1997; Crameri et al. (1), 1996; Crameri et al. (2), 1996; Stemmer (1) 1994; Stemmer (2), 1994; U.S. Patent No. 5,605,793; U.S. Patent No. 5,811,238; U.S. Patent No. 5,830,721; U.S. Patent No. 5,837,458; U.S. Patent No. 5,965,408; WO 95/22625; WO 97/20078; WO 97/35966; WO 98/31837; WO 98/27230; WO 00/00632; WO 00/09679; WO 98/42832; WO 00/18906; EP 752008; and EP 0932670.

Protein fragment complementation means the mixing together of protein fragments to restore biological function. (*See e.g.* Bibi and Kaback, 1990; Burbaum and Schimmel, 1991; Hall and Frieden, 1989; Hantgan and Taniuchi, 1977; Labhardt, 1982; Shiba and Schimmel, 1992; Taniuchi et al., 1977; and Yang and Schachman, 1993.) The fragments can

be obtained by, for instance, treating native or hybrid proteins with digestive enzymes or the like, or be expressed from native or modified gene fragments. The subsequent complementation *in vitro* or *in vivo* results in the conversion of monomers to dimers, or polymers. For example, this method is useful for proteins too large to be synthesized as monomers by current biochemical techniques.

The term *circular permutation* herein means cleaving or bisecting a protein at one point and reconnecting it via the C- and N-termini to yield mutant proteins, including functional mutants of the parent protein (*See e.g.* Mullins et al., 1994; Protasova et al., 1994; Vignais et al., 1995; Yang and Schachman 1993; and Zhang et al., 1993.). In addition, circular permutation includes circularizing a gene, polynucleotide, or modified versions of the same, by connecting the 5' and 3' ends with or without a linker sequence, followed by cleavage at selected or random sites. The resulting modified gene or polynucleotide can then be used for the expression of a modified protein.

Gene Libraries

According to the invention, circular permutation and protein complementation techniques can be adapted to produce hybrid genes and functional mutant proteins. These techniques can be combined with tools adapted from DNA shuffling, directed evolution, and useful screening methods, to produce gene and protein libraries containing functional mutants. In particular, the invention provides gene dimers, comprising two gene monomers joined by a polynucleotide linker. The invention also includes adaptations to the use of gene concatemers, i.e. constructs of more than two monomers, to create gene and protein libraries by techniques outlined herein.

As outlined in **FIG 1A**, a parent gene corresponding to a parent protein or polypeptide is selected. Any source of nucleic acid, preferably in purified form can be utilized as a starting material or parent gene of the invention. Nucleic acid sequences may be any length and of various lengths, although preferably the parent comprises a structural gene for a protein of interest, and is from 50 to 50,000 base pairs. A duplicate gene is constructed by joining two genes, or monomers, to form a dimer. Each monomer may be identical to the parent gene or different from the parent gene, for example by modification of the nucleotide sequence. When both genes of the dimer are from the same parent, the resulting DNA construct can be called a homodimer.

Alternatively, two genes from two parents may be selected, each of which can be called a gene monomer. The two parent genes may encode related or unrelated proteins, including for example structurally or functionally related proteins from two different organisms. For example, a parent gene from one organism may encode a protein having a relatively high biological activity but relatively poor stability. A parent gene encoding a similar or related protein from another organism may encode a protein with less biological activity but greater stability. Also, the proteins encoded by the different parent genes may have different physical properties in terms of e.g. solubility, hydrophobicity, lipophilicity, or charge. These parent genes can be used in native or modified form. The monomers are then combined, according to the invention, to produce a gene dimer. This dimer, also called a heterodimer, can be used to generate a library of hybrids, including functional mutant proteins and chimeric proteins, some of which may combine the high biological activity and high stability of each respective parent, or display other desirable properties.

As shown in **FIG. 1A**, a first parent gene can be obtained having a structural gene flanked by an upstream primer 3 (containing a restriction site R1) region and a downstream primer 1 (containing a restriction site R3) region. A second parent gene, which can be the same as or different from the first parent, is flanked by an upstream primer 2 (containing a restriction site R3) region and a downstream primer 4 (containing a restriction site R2) region. The restriction site R3, which can be a native or an engineered site, is common to the downstream end of the first parent and the upstream end of the second parent. Thus, primers 1 and 2 can be called "linker primers" or "linking primers", via the common R3 region.

One advantageous feature of the invention is that a high degree of sequence similarity between two parent proteins is not a requirement. Accordingly, the sequence identity of two parent genes may be from 0-100%. In one embodiment, the sequence identity of two parent genes is 100%. In another embodiment, the sequence identity is less than 75%. In still another embodiment, the sequence identity is less than 50%, or even less than 30%. In a preferred embodiment, the sequence identity is between 15% and 50%, *e.g.*, as determined by BLAST analysis (*see, e.g.*, Altschul et al., 1990; Henikoff and Henikoff, 1992; or Karlin and Altschul, 1993).

A sufficient amount of genes can be obtained, for example, by amplifying DNA containing the parent genes (including the primers and restriction sites) using polymerase chain reaction (PCR) techniques, followed by purification and analysis as necessary. The

amplified DNA products are restricted with specific restriction enzymes (R1 and R3 for the first parent; R2 and R3 for the second parent). The resulting DNA fragments are ligated by joining the downstream end of the first parent to the upstream end of the second parent (at the common R3 restriction site). A linker, having at least one restriction site that is unique in the dimer construct, is interposed between the first and second parent genes, as shown. The resulting DNA construct, a gene dimer of the two parents, is ligated, or subcloned, into a vector for further amplification, *e.g.*, by transformation into host cells, or by PCR. The amplified DNA is then digested with restriction enzymes which excise the gene dimer by cutting at R1 and R2, thus provide quantities of the gene dimer. The dimer may then be purified and separated from the other components, using methods known in the art.

In one embodiment, suitable for producing a random circular permutation library, the linker sequence is preferably designed such that the reading frame is continuous and the original 5' and 3' (upstream and downstream) ends of the structural gene are connected. Appropriate linker sequences can encode to insert, delete, or mutate amino acids in the protein sequence, or they can leave the protein sequence unchanged, except that the N- and C-terminal ends of proteins encoded by the hybrid genes will be different from the parent proteins. For another embodiment, *e.g.*, to construct a library suitable for protein fragment complementation, the linker should preferably include a sequence to encode the stop translation signal of the upstream fragment (*e.g.* in linker primer 1 of the first parent), and a start of translation signal of the downstream fragment. (*e.g.* in linker primer 2 of the second parent).

As shown in **FIG. 1B**, the purified gene dimer is then cut or fragmented, for example by limited digestion with an enzyme such as, *e.g.*, a nuclease, or DNase I, or by mechanical shearing forces such as sonication. DNA fragments of various sizes are generated in this way. Even if the dimer is cut at random sites, the type or relative degree of fragmentation can often be modulated in the chosen fragmentation technique, for example by time of exposure. Appropriate conditions for each chosen application may require individual optimization, based upon knowledge in the art. *See*, step 1 of **FIG. 1B**. Using any suitable method, or combination of methods, for screening, isolating, separating, or purifying DNA, a population of DNA pieces or fragments is selected. For example, a population of DNA fragments having a predetermined size, or being within a predetermined size range, can be selected and isolated. One possible technique is gel electrophoresis. *See*, step 2 of **FIG. 1B**. Alternatively, fragments

can be made by random primer extension (*See* Shao et al., 1998). If the resulting DNA fragments are too small, they can be subjected to limited overlap extension (*See* Stemmer 1994) or StEP recombination (*See* Zhao et al., 1998) until they have reached the desired average size.

Fragments made by any of the above mentioned, or other techniques, can be further separated, for instance according to size, in order to obtain a high fraction of pieces with the desired properties. In a preferred embodiment, the isolated fragments are within a predetermined size range encompassing, comparable to, or consistent with, about the size of a parent gene (*e.g.*, the selected fragments are the same or similar to a gene monomer in size). In another preferred embodiment, most of the isolated fragments are within the following size range: at least the size of about the smaller parent gene up to, and including, about the size of the larger parent gene (*e.g.*, the size of the selected fragments is somewhere in-between the sizes of the two parent genes). Each fragment is likely to have different 5' and 3' ends, and, consequently, different intervening sequences.

The DNA fragments obtained using these techniques comprise a gene library according to the invention.

The purified DNA fragments are treated as necessary or desired, for example, with a DNA-modifying enzyme (*e.g.*, a single strand specific nuclease, or a DNA polymerase such as T4 DNA polymerase) to convert staggered ends to blunt ends. *See*, step 3 of **FIG. 1B**. The DNA is then ligated into a suitable expression vector, typically a plasmid. *See*, step 4 of **FIG. 1B**. The result is a plasmid, vector, or gene library of hybrid or permuted genes, or complementary fragments. The expression vector is designed so that gene length is controlled (stop codons are provided at all three reading frames). The presence of contaminants or undesired components, e.g. wild-type genes, in this library should be relatively low, but could be further reduced by optimizing the technique(s) used for amplifying and/or separating different components.

The expression plasmids can be used to transform suitable host cells for expressing the proteins. The genes can also be expressed using techniques such as, for example, phage display (Johansson et al., 1999) and *in vitro* transcription-translation systems. The expressed proteins and polypeptides comprise a protein library of the invention.

Briefly, the genes and proteins that are evolved using these methods can be rapidly screened. Functional hybrids, circular permutants or complementary fragments that yield

functional protein are identified by suitable screening or selection methods. When an expression system is used, functional proteins can be readily isolated and purified from the expression system or from the expression media if secreted by the host cells. For example, assays can be used to test functional activity of the particular protein in native form.

Optionally, in cases where the parental fragments are in one continuous reading frame, the number of hybrid variants can be reduced by, for example, ligating the N-terminals of the fragments in the gene library to a gene encoding for a suitable reporter protein whose start codon of translation (ATG) has been modified or removed to prevent its independent translation.

In a preferred embodiment, a gene library of the invention may be generated by a method comprising the steps of: (a) constructing a gene dimer containing a linker sequence; (b) performing limited digestion of the gene dimer to produce a population of fragments of varying sizes; (c) isolating DNA fragments of approximately the same size as a parent gene; and (d) inserting isolated DNA fragments into a suitable expression vector.

Preserved terminal sequences

When constructing a library of hybrid proteins, it may be desirable that they retain the N- and C- terminal ends of the original proteins. Conventional protein fragment complementation and circular permutation generate a bisection of the polypeptide chain somewhere between the original N- and C- terminals, and create hybrids with new N- and C-terminal ends. This tends to reduce the number or fraction of functional hybrids in the library, because new terminal ends, mismatched sequences in the hybrid, or the cleavage into two separate polypeptide fragments, can impair the ability of the resulting proteins to fold properly. This, in turn, may have an impact on protein function (See Graf and Schachman, 1993; and Hennecke et al., 1999).

This problem has been solved in one embodiment of the invention. Libraries of hybrid proteins, especially single-chain proteins, can be made that have an N-terminal part originating from one protein and a C-terminal part of a second protein, with both parts varying in length. As described above, the sequence similarity between the two parent proteins may be in the range 0-100%, since sequence similarity is not a requirement. A preferred, although not limiting, sequence identity of the genes encoding the parent proteins is in the range 15-50%.

To obtain monomeric hybrid proteins with matched terminal ends, a gene dimer, or concatemer, is made according to the strategy outlined in **FIG. 3**. This technique is similar to the one outlined in **FIG. 1A**, except that in the current example, one of the parent genes has an additional restriction site in an upstream or downstream region, *e.g.* a non-coding sequence (shown in the figure as R5 of gene 2).

With reference to **FIG. 3**, a gene construct is made, in this case a heterodimer comprising the genes of two different parent proteins. In other applications of the invention, the parent proteins may originate from the same or different organisms, and may or may not exhibit different functional or physical properties. The two genes are placed in tandem on a single piece of DNA and are separated by a linker sequence. The linker sequence contains one or more, preferably two, restriction sites (as shown) that are unique in the dimer construct. The gene dimer can be constructed and amplified, for example, using PCR and is ligated or subcloned into a suitable cloning vector. After amplification, the constructed gene dimer is excised, and purified as necessary or desired.

The gene dimer is fragmented (*e.g.*, by limited digestion with an enzyme such as, *e.g.*, a nuclease, or DNase I, by sonication, or by random primer extension). These procedures, outlined in **FIG. 4**, are similar to those described above in connection with **FIG. 1B.** A population of fragments is provided, and the resulting mixture of fragments is sorted, separated, or purified by size, for example using gel electrophoresis or other methods described herein. Preferably, the separation or sorting procedure selects a range of fragment sizes encompassing, or being at least comparable with, the size of about a parent gene monomer. If the DNA fragments are too small, they can be subjected to limited overlap extension (*See* Stemmer 1994) or StEP recombination (*See* Zhao et al. 1998) until they are on average the approximate size of a gene monomer. Each of these fragments is likely to have unique 5' and 3' termini, as well as DNA sequence.

An alternative way to produce DNA fragments with the approximate length of a gene monomer is to use Exonuclease III (*See* Henikoff 1984). When linear DNA fragments having blunt ends or 5'-protruding single-strand overhangs are treated with Exonuclease III, one nucleotide at a time is removed from the 3'-end. When a population of DNA fragments of a unique length is subjected to limited treatment with Exonuclease III, the size distribution of the obtained truncated fragments follows a Poisson distribution. This distribution has a deviation of about 20 to 25 % of the average length of the removed DNA fragments (*See*

Hoheisel 1993). When a gene monomer has *n* nucleotides, the desired deviation should be *n*/2 to obtain a library of fragments with the DNA ends covering the entire length of each gene. Therefore, the average length of the DNA to be removed from either side of the gene fragment should be around 2*n* to 3*n*.

It is therefore possible to put the gene dimer into a vector that is about twice (for both sides of the dimer) the size of two to three times the size of the gene monomer. The vector should have a unique restriction site opposite the cloning sites that were used to insert the gene dimer. This unique restriction site is used to linearize the DNA. The linear DNA is then digested with Exonuclease III, followed by a treatment with a single-strand-specific nuclease (e.g. Mung Bean Nuclease, S1-Nuclease) so that the average size of the truncated DNA fragments is the size of the gene monomer. The S1-nuclease digest results in DNA fragments that are blunt-ended which is a requirement for the ligation procedure. The DNA fragments are then separated (e.g. on an agarose gel) and fragments which are approximately the size of the gene monomer are purified.

Yet another approach to produce DNA fragments with the length of a gene monomer uses the inability of Exonuclease III to cut and remove alpha-thionucleotides (*See* Putney et al., 1981; and King and Goodbourne, 1992). When the gene dimer is amplified by PCR using dNTPs and a small amount of alpha-thio dNTPs, the alpha-thio-dNTPs are randomly incorporated over the entire length of the gene dimer. When the DNA fragments are subsequently treated with Exonuclease III, they are truncated to the first thionucleotide on each 3'-end. Therefore, the gene dimer is amplified by PCR using an amount of alpha-thionucleotides that is adjusted such that the exonuclease and subsequent single-strand-specific nuclease treatment will result in DNA fragments which are on average about the size of the gene monomer. As described above, gene fragments are then separated and purified.

The purified DNA is treated with a DNA-modifying enzyme, as needed or desired (also as described above). See **FIGs. 1B and 4**. For example, a single strand specific nuclease or DNA polymerase can be used to convert staggered ends to blunt ends to facilitate subsequent steps.

The protein or polypeptide encoded by the linear construct at this stage would have a new C-terminus in the second protein and a new N-terminus in the first protein. The linear construct may be single stranded or double stranded. In a preferred embodiment, the linear construct is double-stranded. According to the invention, the linear DNA fragments are then

circularized by e.g. intramolecular blunt-end ligation. See **FIG. 5**. The 3'-end of the truncated gene, originating from the second gene of the original dimer, is fused to the 5'-end of the truncated first gene of the dimer. Circularization results in the fusion of the DNA ends encoding for tentative new termini, corresponding to the site marked "**X**" in **FIG. 5**. The position of the ligation site in relation to the linker sequence varies between different constructs, as outlined in the figure. The circular DNA fragments are then treated with restriction enzymes that cut only within the linker sequences. This eliminates the new termini that otherwise would result, and opens the circularized construct in such a way that preferentially preserves or reintroduces one or more original termini. Shown in FIG. 5, step 5, are examples of double-stranded linearized constructs, with 4 base-pair overhangs resulting from the restriction. When necessary, the DNA fragments can be amplified by PCR using PCR-primers that recognize the two original termini.

If desired, the DNA fragments can also be analyzed by PCR. A PCR reaction that uses a primer pair of which one primer is specific for one gene while the other one is specific for the other gene, a product will only be obtained when there is a crossover in the region of the two genes that is flanked by the two primers. Alternatively, when using two primers that are both specific for one gene, the lack of a product indicates a crossover region. The presence or absence of PCR products, therefore, reveals whether the crossover has taken place in a specific region or not.

The linear DNA fragments obtained using these techniques comprise a gene library according to the invention.

The fragments can thereafter be ligated into a suitable expression vector. The vector is pretreated in such a way that DNA ends are compatible for ligation with the DNA fragments, and enable correct transcription of the inserted genes as well as the correct initiation and termination of its translation. The expression vector might also contain a sequence encoding a propeptide or a prepropeptide (e.g. signal sequence) that is necessary for the correct localization and/or folding of the protein. Optionally, the expression vector might also contain the sequence of a reporter gene, which 5' end has been fused in the same reading frame to the ligated genes of the hybrid protein variants. Preferably, the intrinsic start codon of the reporter gene has been removed, to promote a selection for those gene variants that encode the hybrid proteins in one continuous reading frame. The result is a vector gene library of hybrid or permuted genes, or fragments encoding for complementary protein fragments.

In another embodiment, hybrid polynucleotides in a gene library according to the invention, or in a gene library produced according to the invention, can be further mutated by any suitable methods known in the art. For example, the entire library, a selected group of hybrid genes from the library, or polynucleotides selected from the library can be subjected to error-prone PCR, methods for introducing point mutations, and/or various DNA shuffling techniques known in the art (*see*, for example Stemmer, 1994 and Zhao et al., 1998).

The expression vector, *e.g.* a plasmid, can be used to transform a suitable host for expression of the proteins, creating a protein library according to the present invention. Alternatively, the genes can be expressed *in vitro*, *e.g.* using an *in vitro* transcription--translation system. The resulting hybrid proteins maintain the original N-terminus of the second protein and the C-terminus of the first protein -- while containing single crossovers between the two proteins in between. No artificial linker has to be used to fuse the original termini (circular permutation), and the method can therefore be applied to proteins which have buried termini. It can also be applied to proteins which have no independently folding domains, since full-length polypeptide chains are produced. Functional hybrid proteins, circular permutants, or complementary fragments, that yield functional protein are preferably identified by screening or selection.

In yet another embodiment of the invention, a modification of the techniques described above can be applied to obtain a library of hybrid proteins that have more than one crossover at structurally related sites. In this procedure, a unique site for a DNA-cleaving enzyme that leaves nonidentical ss-DNA protruding ends, for example type II restriction enzymes, is introduced beforehand in the linker sequence shown in **FIG. 3**. After limited DNA digestion, isolation of selected gene fragments and construction of circular fragments, constructed as outlined in **FIGs. 4** and **5**, the circularized DNA fragments are cut with these specific enzymes. *See* **FIG. 6** (section 5). The obtained linear fragments are then ligated to each other under conditions that favor intermolecular ligation over intramolecular ligation, in order to obtain long concatemers of gene fragments. All genes will be in the right orientation (5'-3':5'-3'...) in the concatemers, because the protruding ends on each side are not identical. The concatemers are then subjected to another cycle of fragmentation and separation, similar to that described in **FIG 4**, to obtain fragments that are approximately the length of a parent gene monomer. After creating blunt ends, these new fragments are circularized according to **FIG. 5** . The circular DNA-fragments can be used for more rounds of shuffling of different parents

and fragments according to **FIG. 6**. After the chosen number of shuffling cycles, the circularized fragment is cut, for example with restriction enzymes R4 and R5 in **FIG. 5**, to generate a library of linear gene constructs to be fused into expression vectors.

The corresponding protein library consists of hybrid proteins made of multiple fragments from the proteins encoded by the original gene dimer. In addition, if this procedure is applied to a mixture of heterodimers; or to concatemers of various combinations of the genes corresponding to several proteins, for example produced by ligating a mixture of all the proteins which have the linker with the appropriate type IIs restriction site introduced already, gene libraries encoding for hybrid proteins, and corresponding protein libraries, consisting of fragments of multiple parent proteins can be produced.

In still another embodiment of the invention, the techniques described above can be extended to produce hybrid proteins with more than one crossover site, as shown in **FIG. 7**. In this procedure, a second library of single crossover hybrids is obtained similarly as described above, with the exception that the two parent proteins are exchanged ("mirror" library). Thus, the gene that is on the 5' end in one library is on the 3' end in this second library. Both hybrid gene libraries can be mixed and used in a conventional DNA-shuffling experiment. (*See*, for example Stemmer, 1994 and Zhao et al., 1998). In the members of the shuffled library many crossover sites may be recombined and complete multiple shuffling is achieved.

Examples of practicing the invention are provided, and are understood to be exemplary only, and do not limit the scope of the invention or the appended claims. A person of ordinary skill in the art will appreciate that the invention can be practiced in many forms according to the claims and disclosures here. All polynucleotide and polypeptide sequences referred to in the Examples are listed in Table 1 and 2, respectively, together with sequence identification numbers (SEQ ID NOS).

TABLE 1 - Nucleotide Sequences

Name	SEQ ID NO	Nucleotide Sequence [1]
P1	1	TAA GGG GAA CTC GAG ATG AGT AAA GGA GAA GAA CTT

Name	SEQ ID NO	Nucleotide Sequence [1]
P2	2	CAT CTC GAG TTC CCC TTA TTT GTA TAG TTC ATC CAT
L1&2	3	TAA GGG GAA CTC GAG ATG
P3	4	ACC ATG ATT ACG CCA AGC TTG
P4	5	GGG CCC GTA CGG CCG ACT AGT
Pstp	6	p-CGA TTA TTT TTC AGC TAG GCC TAG TTG GTA ATG GTA GCG AC [2]
P1a2u	7	GCA GGG CCC CAG AGC TCA TGG CTC TGT TAT TAG CAG TTT TTC
P1a2d	8	CGC TCT AGA GGT ACC CCA ATT GAT GGA GAA GCG CCG C
Pbm3u	9	CGA CGG ATC CGG AAG GAA GGG CCC ATA TGA CAA TTA AAG AAA TGC CTC AG
Pmund	10	GCA AAG ACC AAT CGT ATC AAG CG
Pmunu	11	GCT TGA TAC GAT TGG TCT TTG C
Pnded	12	CGT TTA GCA TGT GCG TTA ATA AAT C
Pndeu	13	GAT TTA TTA ACG CAC ATG CTA AAC G
Pbm3d	14	CGT CGG TAC CCT CGA GTG AAG TGC TAG GTG AAG GAA TAC C
pB-1B linker sequence	15	TTG GGG TAC CTC TAG AAC TAG TGG ATC CGG AAG GAA GGG CCC AT
P1a2i2r	16	CAG CTG GGG TCT GTC AGA GAG C
Pcatc	17	CGA CGA TCT AGA TTA CGC CCC GCC CTG CCA CTC
pCW1A2cat linker sequence	18	TGG CCT GGG TCC CCT GCT
Pscctu	19	CCT GGG TCC CCT GCT AGC GAG AAA AAA ATC ACT GGA TAT ACC

Name	SEQ ID NO	Nucleotide Sequence [1]
Psectd	20	GGT ATA TCC AGT GAT TTT TTT CTC GCT AGC AGG GGA CCC AGG
Pbm3bam	21	CCA GGA TCC ATC GAT GCT TAG GAG GTC ATA TGA CAA TTA AAG AAA TGC CTC
RC1 N-terminal sequence	22	ATG ACA ATT AAA GAA ATG CCT CAG CCA AAA ACG TTT GGA GAG GTG CTC AAG GGT TTG AGG
RC2 N-terminal sequence	23	ATG ACA ATT AAA GAA ATG CCT CAG CCA AAA CGT TTG GAG AGC CCC AAA GGC CTG AAA AGT
RC3 crossover region	24	AGC ATG CGT TTA AAC CGT TTG GAC CCC TCT GAG TTC CGG C
RC4 crossover region	25	CAG TCA GCA GGC AAT GAA AGG CCC GCC GGC GCC TGG CCC A
RC5 crossover region	26	TAT CGC CGT GCA GCT TGT TCC AAG GGC CGG CCT GAC CTC T
modified P450 1A2	27	*See* **FIG. 9**
heme domain of mutant P450 BM3	28	*See* **FIG. 10**
RC1 complete sequence	29	*See* **FIG. 11**
RC2 complete sequence	30	*See* **FIG. 12**
RC3 complete sequence	31	*See* **FIG. 13**
RC4 complete sequence	32	*See* **FIG. 14**

Name	SEQ ID NO	Nucleotide Sequence [1]
RC5 complete sequence	33	*See* **FIG. 15**
[1] Sequences are listed in 5'→3' direction if not otherwise indicated. [2] "p" indicates that the 5' end is phosphorylated.		

TABLE 2 - Peptide sequences

Name	SEQ ID NO	Peptide Sequence
Peptide expressed from pCW1A2cat linker sequence	34	WPGSPA
RC1 N-terminal sequence	35	MTIKEMPQPKTFGEVLKGLR
RC2 N-terminal sequence	36	MTIKEMPQPKRLESPKGLKS

EXAMPLE 1

Random complementary fragment library of green fluorescent protein

GFP (Green Fluorescent Protein) is a protein produced by the jellyfish *Aequorea victoria* which fluoresces in the lower green portion of the visible spectrum. This Example describes the production of a GFP library suitable for protein fragment complementation. A gene homodimer consisting of two GFP monomers connected by a linker sequence, was constructed. After a limited digestion of the gene dimer, fragments approximating a gene monomer in size were retrieved and inserted into an expression vector. The plasmid library was thereafter screened to identify functional GFP variants.

Plasmid pGFP containing the complete GFP coding sequence under the lac promoter (Clontech Laboratories) was used either intact or modified. This plasmid was transformed into *E. coli* strain XL1-Blue for amplification of the plasmid and for GFP expression. The GFP gene monomer consists of 714 bp.

a) Construction of gene homodimer with linker sequence

A GFP gene dimer in which a linker sequence was inserted between the two copies of the GFP gene was constructed by PCR. Two linking primers were used; P1 (forward) **[SEQ ID NO: 1]** and P2 (reverse) **[SEQ ID NO: 2]**. Each of these linking primers contains the same (forward) linker sequence L1&2 **[SEQ ID NO: 3]**. Another two primers flanking the GFP gene, and used for PCR, were P3 (forward) **[SEQ ID NO: 4]**, and P4 (reverse) **[SEQ ID NO: 5]**.

Two PCR reactions were carried out. The first reaction used primers P1 and P4, and the second used P2 and P3. In both cases, the template was pGFP. The PCR reactions were carried out in 100 μL volume, with 25 cycles of 94°C for 1 minute, 52°C for 40 seconds, 72°C for 1 minute, with an increment of 1 second each cycle. The PCR products were checked by loading 3 μL of the reaction mixture onto a 1% agarose gel for electrophoresis. In both cases, the expected DNA fragment (about 900 bp in size) was found to be the sole product. The PCR products were purified using a Qiagen PCR purification kit. The purified product from the first reaction was restricted by restriction enzymes XhoI and EcoRI. The product from the second PCR reaction was restricted by XhoI and PstI. The resulting DNA was used in the following three-piece ligation reaction.

2-3 μg of plasmid pGFP was restricted with PstI and EcoRI, and the about 2.6 kbp band was purified from an agarose electrophoresis gel using the Qiagen extraction kit. This fragment was ligated with the above purified PCR reaction products in a 3-piece ligation reaction by standard cohesive-end ligation. The ligation mixture was transformed in XL1-Blue competent cells by the heat shock method. Transformed cells were plated out onto LB/ampicillin plates. Ten colonies were picked at random and grown up 2-3 mL cultures in order to purify the plasmid DNA (by mini-prep). All ten colonies picked contained the duplicated GFP gene. This new plasmid was designated GFP2x.

b) Generation of GFP gene fragments from gene dimer

The GFP gene homodimer was subjected to limited digestion using DNase I. About 80 μg of the GFP gene dimer was obtained by restriction of about 200 ug of pGFP2x with BamHI and EcoRI, and the about 1.5 kbp DNA band was purified from a 1% agarose electrophoresis gel. The dimer gene DNA was digested by adding an appropriate amount of DNase I (about 30 μL of 0.0015 U/μL) in 100 μL reaction mixture in 50 mM Tris-HCl, pH 7.5/1 mM $MnCl_2$ and 50 μg/mL BSA for 20-60 min at room temperature. The progress of the reaction was checked by agarose gel electrophoresis every 5 minutes. The reaction was

stopped when the digestion products gives an even smear on the gel from about 1.5 kbp down to about 50 bp. Fragments of about 500 bp to about 850 bp in size were purified from a 1% agarose electrophoresis gel using the Qiagen DNA extraction kit. The purified DNA (2-4 μg) was treated with 1 U of T4 DNA polymerase in the presence of 0.2 mM of each of the four dNTPs in T4 polymerase reaction buffer (New England Biolabs). The reaction (25 μL total volume) was allowed to proceed for 15 min at 16°C. The reaction was stopped by addition of one volume of 15 mM Tris, pH 7.5 and two volumes of phenol:chloroform (1:1). After two more extractions with phenol-chloroform, the DNA was precipitated by ethanol, washed once with 70% ethanol, dried and dissolved in 20 μL water.

c) Construction of expression plasmid

A modified pGFP plasmid designated pGFP-stp, to be used as an expression vector for the fragmented GFP gene dimers, was prepared. The pGFP-stp plasmid was constructed so that stop codons were introduced in all three reading frames following the GFP-coding sequence in pGFP. PCR was used to introduce the stop codons and associated sequence alterations. Primer Pstp (reverse) was designed to introduce a stop codon in each reading frame and a new StuI site. Primer Pstp is 5-prime phosphorylated, and the sequence of Pstp is listed in Table 1 **[SEQ ID NO: 6]**.

The PCR used P3 (forward) and Pstp (reverse) with pGFP as the template. Conditions were the same as those described above. The PCR product (about 850 bp) was restricted with Hind III and purified using the Qiagen kit. The restricted PCR product was ligated with the about 2.6 kbp fragment isolated from the digestion of pGFP with Hind III and StuI. The ligation mixture was used to transform XL1 -Blue competent cells by the heat shock method. Since the majority of the colonies contained the pGFP-stp plasmid (as shown by the control ligation experiment), and some minor fraction of wild-type pGFP at this stage does not affect the final result, a pool of 60 colonies was used to grow cells for pGFP-stp plasmid preparation.

d) Construction of a plasmid library containing the GFP gene fragments.

After restriction, the pGFP-stp plasmid was dephosphorylated and ligated with the GFP gene dimer fragments to form a plasmid library. The pGFP-stp plasmid (2-5 μg) was restricted with SmaI and StuI at 22°C for 12 hrs and then at 37°C for another hour. The about 2.6 kbp fragment was purified from a 0.8% agarose electrophoresis gel using the Qiagen DNA extraction kit. The purified DNA (in 20 μL dephosphorylation reaction buffer) was treated

with 0.3 U of shrimp alkaline phosphatase (US Biochemical) at 37°C for 30 min. Fresh enzyme (0.3 U) was added every 30 min. This 5'-dephosphorylated plasmid vector fragment was ligated with the blunt-ended GFP DNA insert (500-850 bp, prepared as described above) using a standard blunt-end ligation protocol. The ligation mixture (about 40 μL) was transformed into XL1-Blue competent cells. The transformed cells were plated out onto LB plates supplemented with ampicillin and IPTG.

e) Screening for active GFP

The library produced above was subjected to two rounds of screening to identify functional GFP fragments. The first was based on fluorescence of the functional protein, and the second was based on the restriction digestion pattern of the plasmids.

Fluorescence.

Two batches of plates prepared from two separate ligation reactions were screened. A total of about 11,400 colonies was screened visually by shining 366 nm UV light briefly on each plate with a hand UV lamp (UVP, Model UVGL-58) in a dark room. The colonies that emitted green light were marked on the bottom of the plate. A total of 184 clones emitting green light upon UV illumination were obtained and used for the next round of screening.

Digestion Pattern.

150 well-isolated colonies from the 184 green-light emitting colonies were picked and used to inoculate 2 mL LB/ampicillin cultures for plasmid preparation. Plasmid DNA mini-preparations were carried out for each culture. The purified plasmid DNA was subjected to different restriction enzyme digestions. *See,* **FIG. 2A.** First, a double digestion with BamHI and EcoRI was used to estimate the wild-type GFP background present among the active clones. The about 100 active colonies from the batch of plates made first contained a high wild-type GFP content (about 80%). This was consistent with the control experiment from this batch, in which the about 2.6 kbp plasmid fragment alone gave rise to a considerable background of active GFP colonies growing on the plates. In contrast, the about 50 active GFP clones from the plates in the second batch had a very low wild-type GFP background. Furthermore, about 70% contained the unique XhoI site, which does not exist in the wild-type GFP plasmid.

A total of 43 colonies containing plasmid with the XhoI site were identified. The length of the insert was estimated by double digestion with BamHI and SfiI. The digestion patterns for all 150 plasmids were analyzed, and the results of the 50 plasmids from the second batch of plates are summarized in **FIG. 2B**. A large portion of the active GFP-containing plasmids had a whole insert length greater than the GFP gene size and was found to contain the intact GFP gene, either in front of or following the linker sequence. Recovered wildtype or wild-type-like GFP genes from the insert library also occured frequently. A few inserts were found to be slightly shorter than the intact GFP gene (i.e. lacking both EcoRI and XhoI sites).

None of the screened active GFP plasmids contained a split gene with two complementary fragments of the whole gene, nor a considerably truncated gene. In contrast, of two inactive GFP plasmids chosen at random from the library and subjected to the same restriction treatment, one had the XhoI site almost in the middle of the gene. The second was found to be a truncated gene. The linker sequence inserted between the genes in the gene dimer was found in different positions in the final library.

It is estimated that about one-third of the genes in the blunt-end ligation product (the gene fragment or permutation library) would have the correct reading frame when the method of this example is used. For a protein of 300 amino acids, one will need to screen in the order of 5×10^4 colonies in order to cover all the diversity of positions for fragmentation at a single site.

The wild type background may be due to the presence of the wild type gene after purification of the expression plasmid vector. The presence of wild type protein can be eliminated or greatly reduced by either of the following approaches. First, during purification of the expression plasmid DNA fragment (the about 2.6 kbp fragment), a longer path agarose gel for electrophoresis can be used to better resolve the desired fragment from the partially digested plasmid that still contains the wild type gene. A second and more reliable approach is to use a plasmid vector that does not contain this gene in the first place.

Although this example showed construction of a homodimer, a heterodimer can be made from two different parent genes using substantially the same techniques.

EXAMPLE 2:

Hybrid protein library with preserved terminal sequences

This Example describes the creation of a library of protein hybrids containing sequences from two parent proteins; human cytochrome P450 1A2 and bacterial cytochrome P450 BM3. The resulting proteins consist of a single polypeptide chain that have the N-terminus of the bacterial enzyme and the C-terminus of the mammalian P450. The human P450 is membrane-associated, while the bacterial enzyme is soluble. The human P450 is active towards a range of aromatic substrates, while the bacterial enzyme prefers long-chain fatty acids. The library of hybrid proteins was therefore expected to contain P450s that are soluble like the bacterial enzyme, and exhibit the substrate specificity of the human enzyme.

Human cytochrome P450 1A2 having a modified N-terminus (*See* Fischer et al., 1992) was used **[SEQ ID NO: 27]**. The heme domain of *Bacillus Megaterium* P450 BM3 containing the mutation F87A (*See* Schwaneberg et al., 1999), resulting from four replaced nucleotides at position 261 (ATTT to GGCC), was further modified by removing two restriction sites (A to G at position 459; T to C at position 711) **[SEQ ID NO: 28]**. The modified P450 1A2 gene inserted into the expression vector pCWori (Barnes, 1996) was provided by Prof. F.P. Guengerich, and the F87A mutant of P450 BM3 inserted into the cloning vector pUC19, was provided by Dr. U Schwaneberg. The size of the P450 1A2 gene monomer used was 1,515 bp, whereas the size of the heme domain of P450 BM3 used was 1,392 bp.

a) Construction of gene heterodimer with linker sequence

In this Example, the following restriction sites were used: *Sac*I, *Xho*I, *Xba*I, *Mfe*I and *Nde*I, corresponding to R1-R5 in **FIG. 3**, respectively. A gene heterodimer consisting of mammalian and bacterial P450 connected by a linker sequence was constructed. The gene of P450 1A2 was amplified by PCR (referred to as PCR#1) from the vector pCW1A2bc using the following combination of primers, P1a2u **[SEQ ID NO: 7]** and P1a2d **[SEQ ID NO: 8]**.

Fragments of the gene of the heme domain of P450 BM3 were amplified by PCR (referred to as PCR#2-4) from the vector pcmdheme, using the following combinations of primer sequences: PCR#2: Pbm3u **[SEQ ID NO: 9]** plus Pmund **[SEQ ID NO: 10]**; PCR#3: Pmunu **[SEQ ID NO: 11]** plus Pnded **[SEQ ID NO: 12]**; PCR#4: Pndeu **[SEQ ID NO: 13]** plus Pbm3d **[SEQ ID NO: 14]**.

The fragments from PCR#2-4 were purified after separation on an agarose gel using the *Qiaexll* purification kit, combined and used as a template for a PCR (referred to as

PCR#5) with the primer pair bm3u **[SEQ ID NO: 9]** and bm3d **[SEQ ID NO: 14]**. Primer Pndeu **[SEQ ID NO: 13]** and Pnded **[SEQ ID NO: 12]** contain a mismatch, which removes an internal *Nde*I site in the gene of BM3. Primers Pmunu **[SEQ ID NO: 11]** and Pmund **[SEQ ID NO: 10]** contain a mismatch, which removes an internal *Mfe*I site. The product of PCR#5 therefore encodes the gene of BM3 with two silent mutations that remove restriction sites for *Nde*I and *Mfe*I.

The PCR reactions were carried out in 50 μl volume, with 30 cycles of 94°C for 45 sec, 52°C for 45 sec, 72°C for 2 min (PCR#1 and 5), 1 min (PCR#2, 3, 4) using Vent-Polymerase (New England Biolabs). The PCR products were separated on a 1% agarose gel and purified using a *QiaexII* purification kit. The purified PCR#1 product was then restricted by restriction enzymes *Sac*I and *Xba*I. The PCR#5 product was restricted by *Xho*I and *Bam*HI. Both were ligated into pBluescript II SK (+) (Stratagene) which had been restricted by the same pairs of restriction enzymes to yield pB-1A2 and pB-BM3, respectively, (using standard ligation procedure). The ligation mixtures were used to transform XL1-Blue cells by electroporation. Transformed cells were plated out onto LB/ampicillin/Xgal/IPTG plates. Ten white colonies were picked from each plate at random, grown up in 5 ml cultures, and plasmid DNA was prepared. The DNA sequence of the inserted genes of both pB-1A2 and pB-BM3 was checked by sequencing and found to be correct. Both vectors were then restricted by *Xho*I and *Bam*HI and the reaction products separated on an 1% agarose gel. The linearized vector of pB-1A2 and the BM3 fragment of pB-BM3 were purified from the gel and ligated to obtain vector pB-1B. Vector pB-1B contains the gene heterodimer with the gene of 1A2 on the 5' end and the gene of BM3 on the 3'-end, separated by a linker sequence listed in Table 1 **[SEQ ID NO: 15]**. *E. coli* XL1-Blue cells were transformed with pB-1B, cells were grown and plasmid DNA was prepared.

b) Generation of P450 gene fragments by DNase I digestion

About 10 μg of pB-1B was treated with the restriction enzymes *Xho*I, *Ssp*I, *Sca*I and *Asn*I. The linearized DNA was then subjected to a limited digestion by adding 1.2 μl DNase I (500 mu/μl) in 120 μl reaction in 33 mM TrisHCl, pH 7.5/10 mM $MnCl_2$ and 50 μg/ml BSA for 60 min at room temperature. The reaction was stopped by addition of 13 μl 0.5 M EDTA and cooling on ice. These conditions had been found to give an even smear on an agarose gel from 3 kbp down to about100 bp. Fragments of about 1400 bp to about 1600 bp in size were purified from a 1% agarose electrophoresis gel using the *QiaexII* DNA extraction kit.

The purified DNA (100 ng) was treated with 1 u of T4 DNA polymerase in the presence of 0.2 mM of each of the four dNTPs in T4 polymerase reaction buffer (New England Biolabs). The reaction (33 μl total volume) was allowed to proceed for 15 min at 16°C. The reaction was stopped by heating to 65°C for 10 min. The solution was debuffered using a centriprep spin column.

c) Generation of P450 fragments by Exonuclease III digestion

An alternative strategy to fragment the gene dimers is to use Exonuclease III. (*See* Detailed Description). One μg of vector pB-1B was digested with 20 u *Ssp*I, and 500 ng of vector pACYC184 (New England Biolabs) was digested with 10 u AsnI, followed by treatment with 1 u T4 DNA polymerase in the presence of 0.2 mM of each of the four dNTPs in T4 polymerase reaction buffer (New England Biolabs). The reactions were stopped and the DNA was concentrated using the *QuiaexII* Kit. Both vectors were ligated together and used to transform *E. coli* XL1-Blue cells by electroporation. Six colonies were picked, cells were grown, and DNA prepared and analyzed by restriction digestion with *Nco*I. Two clones showed the correct direction of the two ligated fragments to each other and were named pB-exo+. pB-exo+ is about 9500 bases long and has a singular restriction site of *Eag*I roughly opposite to the linker sequence which connects the genes of 1A2 and BM3.

20 μg of pB-exo+ were linearized by digestion with 50 u *Eag*I. After inactivation of the enzyme by heating to 65°C for 10 min, the DNA was precipitated with EtOH/NaAc and redissolved in 20 μl TE. In a total volume of 200 μl, the DNA was digested with 2000 units Exonuclease III at 37°C. After 11 min the reaction was stopped and the 5'-3' single strand overhangs were removed by adding 750 μl S1-Solution (Exo/S1 Kit, MBI Fermentas). These conditions had been determined to give a smear on an agarose gel from 1800 bp down to 1100 bp. Fragments of about 1400 bp to about 1600 bp in size were purified from a 1% agarose electrophoresis gel using the *QiaexII* DNA extraction kit.

d) Circularization of the gene fragments to obtain full length genes

The gene fragments obtained by the methods described in b and c were circularized by treatment with 3 Weiss units of T4-DNA Ligase for 20 h at 25°C in 30 μl of 100 mM Tris/HCl, pH 7.5, 3 mM DTT, 50 μM rATP, 10 mM $MgCl_2$.

e) Construction of a plasmid library containing the cytochrome P450 hybrids

An expression vector can be constructed, containing all necessary features for the expression of a gene incorporated at the two restriction sites, which are identical to the ones in the linker sequence. The start codon can be within the gene fragment (BM3), as well as the stop codon (1A2). Two additional stop codons can be incorporated into the expression vector (in the 3' direction of the two restriction sites) to avoid unnecessary run-off during the translation of fragments that are not in the correct reading frame. The vector can be cut with the same two restriction enzymes and ligated with the DNA fragments from above, using standard procedures for sticky end ligation (*See* Sambrook et al., 1989). The ligation mixture can be used to transform a suitable host for protein expression.

f) Analysis of library

About 100 transformants can be randomly picked and used as a template for colony PCR using the primer pair BM3 forward (Pbm3u **[SEQ ID NO: 9]**) and 1A2 reverse (P1a2d **[SEQ ID NO: 8]**). The products of the reactions can be analyzed by agarose gel electrophoresis. Only clones that have a full-length hybrid gene can show a band with the size of about 1.5 kbp. In a second colony PCR two primers that bind to the expression vector but flank the inserted gene fragments can be used. Again, the products of these reactions can be analyzed by agarose gel electrophoresis. These two experiments can provide the percentage of clones which contain a P450 fragment insert and the percentage of those which contain a hybrid of BM3 and 1A2. By restriction digestion of the PCR products of selected clones containing hybrid genes, the position of the crossover point is narrowed down. For this experiment, restriction enzymes that cause different restriction patterns in the genes of BM3 and 1A2 but that do not cut more than 3-5 times can be used. Alternatively, tile crossover point can be narrowed down by nested PCR using internal primers.

g) Screening for active P450

The library can be analyzed for active P450 variants by coexpressing a P450 reductase using a standard protocol (*See* Chang and Waxman, 1998). One third of the library is expected to contain genes that are in the correct reading frame over their entire length.

EXAMPLE 3

Hybrid protein library with preserved terminal sequences

Using techniques similar to those described in Example 2, a library of protein hybrids was created, containing sequences from the same two parent proteins, modified human cytochrome P450 1A2 **[SEQ ID NO: 27]** and an F87A mutant of *B. Megaterium* cytochrome P450 BM3 **[SEQ ID NO: 28]**. The resulting proteins consist of a single polypeptide chain and have the N-terminus of the bacterial enzyme and the C-terminus of the human (mammalian) P450. The mammalian P450 is membrane associated, while the bacterial enzyme is soluble. The mammalian P450 is active towards a range of aromatic substrates, while the bacterial enzyme prefers long-chain fatty acids. The library of hybrid proteins was screened for P450 proteins that are soluble like the bacterial enzyme, and exhibit the substrate specificity of the mammalian enzyme.

a) Construction of gene heterodimer with linker sequence

Gene heterodimers were constructed as described in Example 2, section ***a)***.

b) Generation of P450 gene fragments by DNase I digestion and S1-nuclease treatment

About 100 µg of pB-1B (*see* Example 2) was digested with the restriction enzymes *Xho*I, *Ssp*I, *Sac*I and *Asn*I and subsequently desalted. About 25 µg of that DNA was then digested by adding 2.5 µl DNase I (500 mu/µl) in 300 µl reaction in 33 mM Tris/HCl, pH 7.5/10 mM $MnCl_2$ and 50 µg/ml BSA for 15 min at 26°C. The reaction was stopped by addition of 13 µl 0.5 M EDTA and cooling on ice. After purification of the DNA (to remove the DNase) using the QiaexII DNA extraction kit, the DNA was further digested with 35 u S1-nuclease in 35 µl 25 mM potassium acetate buffer (pH 4.6) supplemented with 200 mM NaCl, 0.9 mM $ZnSO_4$, and 4 % glycerol) for 40 min at 22°C, to make the DNA fragments blunt-ended. The resulting fragments were separated on an agarose gel. These conditions had been found to give an even smear on an agarose gel from 3 kb down to about 100 bp. Fragments of about 1400 to about 1600 bp in size were purified from a 1% agarose electrophoresis gel using the *QiaexII* DNA extraction kit.

c) Generation of P450 fragments by Exonuclease III digestion

P450 fragments were also generated using an alternative strategy, Exonuclease III digestion, as described in Example 2, section ***c)***. Fragments of about 1450 bp to about 1550 bp in size were purified from a 1% agarose electrophoresis gel using the *QiaexII* DNA extraction kit.

d) Circularization of the gene fragments to obtain full-length genes

The gene fragments obtained by the digestion methods described in *b)* and *c)* were circularized by treatment with 3 Weiss units of T4-DNA Ligase for 20 h at 25 °C in 30 μl of 100 mM Tris/HCl, pH 7.5, 3 mM DTT, 50 μM rATP, 10 mM $MgCl_2$.

e) Analysis of gene libraries

Both libraries of circularized gene fragments, *i.e.* the library obtained by DNase I/S1-nuclease digestion (the "DNase I library") and the one obtained by Exonuclease III digestion (the "Exo III library"), were restricted with *Xba*I and used as PCR templates with the primers Pbm3u **[SEQ ID NO: 9]** and P1a2d **[SEQ ID NO: 8]**. The PCR reactions were carried out in 50 µl volume, with 30 cycles of 94 °C for 45 sec, 52 °C for 45 sec, 72 °C for 1 min 30sec using Vent-Polymerase (New England Biolabs). The purified PCR products were then restricted by restriction enzymes *Bam*HI and *Xba*I, separated on a 1% agarose gel, and purified using a *QiaexII* purification kit. Both were ligated into pBluescript II SK (+) which had been restricted by the same pairs of restriction enzymes (using standard ligation procedure). The ligation mixtures were used to transform XL1-Blue cells by electroporation. Transformed cells were plated out onto LB/ampicillin/ Xgal/IPTG plates. About 50 white colonies of each sample were randomly picked and used as a template for colony PCR with the primers Pbm3u and P1a2d.

The products of these PCR reactions were analyzed by gel electrophoresis and their approximate length determined by comparison with a DNA-standard (1kb-ladder, Fermentas). The two libraries each showed an average length of 1395 ±100bp (DNaseI) and 1430 ±110bp (ExoIII). Subsequently, the same colonies were used as templates for a PCR analysis to reveal approximate positions of the crossover points within the genes. PCR reactions were performed with the primer pairs Pmunu **[SEQ ID NO: 11]** and P1A2d **[SEQ ID NO: 8]**; PNdeu **[SEQ ID NO: 13]** and P1a2d **[SEQ ID NO: 8]**; and P1a2i2r **[SEQ ID NO: 16]** and P1a2d **[SEQ ID NO: 8]**. The PCR products were analyzed by agarose gel electrophoresis.

The fragments of the library created via DNaseI digestion contained the crossovers evenly distributed over the whole gene (40 % of the cross overs were found in the first third of the gene, 20 % in the following sixth, 10 % in the following sixth and 30 % in the last third). Sequencing of one randomly chosen variant from the DNase I library revealed a hybrid denoted RC3 **[SEQ ID NO: 31]**, containing the first 1182 nucleotides from BM3, followed by nucleotides 1233 to 1512 from 1A2. The crossover section is shown in the Sequence Listings **[SEQ ID NO: 24]**.

The fragments of the ExoIII library had the crossover in an area of around nucleotide No. 500 ± 300. The sequencing of two randomly chosen variants revealed one hybrid denoted RC4, with nucleotides 1-343 from BM3 and nucleotides 370-1512 from 1A2 ([**SEQ ID NO: 32**]; crossover section in [**SEQ ID NO: 25**]), and one hybrid denoted RC5, with nucleotides 1-385 from BM3 and nucleotides 282-1512 from 1A2 ([**SEQ ID NO: 33**]; crossover section in [**SEQ ID NO: 26**]).

In this experiment, the method for producing gene fragments as described under section b (see above) was therefore more suitable to produce hybrid proteins with crossovers distributed along the entire gene. The method described under c (see above) had a more limited range of the crossover distribution. Thus, the Exo III method may be the digestion method of choice if, for example, it is desirable to conserve a larger portion of the N- and/or C terminal region of the parent protein(s) due to a particular function of that region. Another potential reason could be to target the crossover to a specific region that has been identified by other methods (*e.g.* computational methods) as promising to obtain functional hybrids.

f) Construction of a vector for the expression of the cytochrome P450 hybrids

The gene for chloramphenicol acetyl transferase was amplified by PCR from the vector pACYC184 and using a combination of the primers Pcatc [**SEQ ID NO:17**] and Pcatn [**SEQ ID NO: 7**] under the following conditions: 50 µl, 1 min 95 °C, 25 cycles of 45 seconds at 94 °C, 45 seconds at 52 °C, 1 min 8 seconds at 72 °C. The PCR product was digested with *Mfe*I and *Xba*I and ligated into the accordingly digested vector pCW1A2. (*See* Barnes, 1996). The ligation mixture was used for the transformation of XL1-Blue cells. The resulting plasmid (pCW1A2cat) contains the gene for chloramphenicol acetyl transferase immediately following the gene of 1A2 which itself has lost its stop codon. Translation of the gene of 1A2 produces a fusion protein between 1A2 and cat with the linker sequence of WPGSPA [**SEQ ID NO: 34**], encoded in-between by the nucleotide sequence listed in Table 1 [**SEQ ID NO: 18**].

PCW1A2sccat was digested with *Sal*I, treated with Vent polymerase to create blunt ends, and was re-ligated to obtain pCW1A2rfcat. This vector is identical to pCW1A2cat except for a shift in the reading frame at amino acid 478 of 1A2. Using a combination of the primers Pscctu [**SEQ ID NO: 19**] and Pscctd [**SEQ ID NO: 20**], and the Quickchange mutagenesis kit (Stratagene), the intrinsic start codon of the cat gene in pCW1A2cat and pCW1A2rfcat were changed to a codon for serine (ATG → AGC) to produce pCW1A2sccat and pCW1A2rfsccat.

g) Construction of expression libraries and preselection

The DNase I and Exo III gene libraries, obtained as described in sections b-d (see above), were amplified by PCR using the Primers P1a2d **[SEQ ID NO: 8]** and Pbm3bam **[SEQ ID NO: 21]**.

Thereafter, both gene libraries were restricted with *Mfe*I and *Bam*HI, purified by gel electrophoresis, and ligated into the vector pCW1A2rfsccat. Prior to ligation, the pCW1A2rfsccat vector had been treated with *Mfe*I and *Bam*HI to remove the insert (1A2rf) and purified by gel electrophoresis. XL1-Blue cells were transformed with the ligation mixtures and plated on LB-Amp agar. About 250,000 clones were obtained for the DNase I- and about 60,000 clones were obtained for the Exo III-libraries. Cells were scraped from the agar and resuspended in LB-amp medium. Serial dilutions of cells were plated on agar consisting of expression medium (TB-medium plus 1 mM IPTG, 0,5 mM δ-Aminolevulenic acid, 1 mM Thiamine, trace elements) including 40 µg/ml chloramphenicol.

h) Screening

About 2000 colonies were picked from expression libraries D and E on the TB-selection agar and used to inoculate 25 µl TB+ medium (TB including 1mM Thiamine and trace elements) in 96-well fluorescence microtiter plates. Another 5,000 to 10,000 colonies were picked in pools of ten per well. The plates were incubated for 20 hours at 30 °C, 270 rpm. Then, 100 µl of TB++ (TB+ incl. 1mM IPTG, 0.5 mM δ-Aminolevulenic Acid) were added and the plates were incubated for another 20 h at 30 °C, 270 rpm.

To analyze for activity of the variants, 125 µl of 25 mM Tris/HCl, pH 7.4, 10 mM $MgCl_2$, 100 mM KCl, 5 µM 7-Ethoxyresorufin were added to each well. (*See* Chang and Waxman, 1998). Fluorescence at 595 nm ± 20nm after excitation at 550 nm ±10 nm was measured immediately, and after a 3 hour incubation at 37°C. By subtracting both measurements, variants were identified that showed an increase in fluorescence due to the de-ethylation of 7-ethoxyresorufin, a typical reaction for 1A2 P450.

i) Characterization of variants

Two variants, RC1 **[SEQ ID NO: 29]** and RC2 **[SEQ ID NO: 30]** were found in the library that were active in the de-ethylation of 7-ethoxyresorufin. RC1 was sequenced and revealed the N-terminal nucleotide sequence **[SEQ ID NO: 22]** and corresponding amino acid sequence **[SEQ ID NO: 35]** listed in FIG. 8 and Tables 1 and 2. RC2 was sequenced and revealed the N-terminal nucleotide sequence **[SEQ ID NO: 23]** and corresponding amino acid

sequence **[SEQ ID NO: 36]** listed in FIG. 8 and Tables 1 and 2. In FIG. 8, sequences originating from BM3 are in bold type and sequences originating from 1A2 are in italic type.

In RC1, the first 15 amino acids of 1A2 have been replaced by the 14 N-terminal amino acids of BM3. RC1, therefore, is almost a full length 1A2 with a more hydrophilic N-terminus. RC2 contains the first 44 nucleotides from BM3 but with a deletion of one of the A's in the A-quintuplet at nucleotide residues 27-31. This results in the shift in the reading frame at amino acid 11 of BM3. The crossover 12 nucleotide residues further downstream restores the correct reading frame at amino acid 25 of 1A2. RC2, therefore, also consists mainly of 1A2.

j) Analysis of hybrid variants

RC1 and RC2 both were subcloned to remove the cat-fusion from the C-termini. After preparing plasmids, the genes of both variants were cut out using *Bam*HI and *Mfe*l, gel purified, and ligated into a pCWori derivative that reintroduced the native stop codon for 1A2. XL1-Blue cells were transformed by the ligation mix and plasmids were purified from the transformants, verified by restriction analysis, and used to transform DH5α cells. Together with 1A2 wildtype, both variants were then overexpressed in this strain using volumes of 250 ml TB++ medium.

Cellular localization. To analyze the solubility of the variants, the localization of the proteins within the DH5α cells was determined. Equal amounts of cells transformed with each variant were lysed by ultrasonication and centrifuged at 100,000 g for 2 h. The upper two thirds of the supernatant were removed and re-centrifuged under the same conditions. Again, the upper two thirds were removed and saved as a membrane-free cytosolic fraction. The pellet of the first centrifugation was resuspended and saved as the membrane fraction. The rest was discarded. Both fractions were analyzed for the content of P450 enzymes using the P450 peak and also for 1A2 activity (de-ethylation of 7-ethoxyresorufin) using an NADPH regeneration system (Shimada, 1998) and P450 oxidoreductase from rat in microsomes.

While basically no wild-type 1A2 could be found in the cytosolic fraction (less than 10 nM), RC2 was detected at a concentration of about 120 nM. In addition, the cytosolic fraction with RC2 showed a strong activity, while that of 1A2 was at the detection limit. From the concentrations in the different samples, a partition of about 14 % RC2 in the cytosol could be estimated, compared to less than 2 % of wildtype 1A2. Even though some RC1 could be detected in the cytoplasm (about 5%), the majority of the protein was still bound to the

membrane. In addition, in Western blot analysis, the immunoblot analysis of the cytosolic fractions gave a very strong signal for RC2, a less strong signal for RC1, and a barely visible signal for 1A2. Thus, RC1 was less soluble than RC2 but more soluble than wild-type 1A2.

Membrane solubilization. In a second experiment, different amounts of detergents (0.5 % sodium cholate plus 0, 0.01, 0.05 or 0.2 % Triton X-100) were used to extract the P450 enzymes from the membranes. After centrifugation for 1 h (100,000 g) the supernatant as well as the pellet were analyzed for activity. None of the samples showed activity in the supernatant. The re-suspended pellet of 1A2 had activity up to 0.05 % Triton X-100, while that of RC1 and RC2 only had activity up to 0.01 % Triton X-100. Western blot analysis of the samples showed that after treatment with 0.5% sodium cholate and 0.05% Triton X-100, RC1 and RC2 were almost completely solubilized, while the vast majority of 1A2 was still membrane-bound.

Enzyme activity. The activity of the P450 enzymes was investigated by measuring the deethylation of 2.5 μM 7-ethoxyresorufin in an NADPH regeneration system (5 mM glucose-6-phosphate, 2 mM NADP+, and 0.6 u/L glucose-6 phosphatase dehydrogenase). The specific activity of both RC1 and RC2 was approximately 50%±10% of the specific activity of wild-type P450 1A2. However, due to the higher solubility of the chimeras, the total activity of RC1 and RC2 in the cytosolic fractions was approximately 1.5 and 7.5 times that of wild-type P450 1A2, respectively.

This example thus demonstrates a successful application of the invention. From a library of 2000 variants of hybrid proteins constructed from the parents BM3 and 1A2, two variants were found that have (i) an N-terminal portion of BM3 and a C-terminal portion of 1A2, (ii) P450 acticity; and (iii) improved solubility compared to the parent 1A2.

BIBLIOGRAPHY

Altschul et al. (1990), J Mol Biol 215:403-410.

Barnes, H. J. (1996), Meth Enzymol., 272 3-14.

Bibi, E. and Kaback, H. R. (1990) PNAS USA, 87, 4325-4329.

Burbaum, J. J. and Schimmel, P. (199 1) Biochemistry 30, 319-324.

Chang et al., (1999) "Evolution of a cytokine using DNA family shuffling," Nature Biotechnology, 17:793-797

Chang, T.K.H. and Waxman, D.J., (1998) Methods in Molecular Biology Vol. 107: Cytochrome P450 Protocols, Human Press Inc. Totowa, NJ.

Christians et al., (1999) "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling," Nature Biotechnology, 17:259-264

Crameri et al., (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution," Nature, 391:288-291

Crameri et al., (1997) "Molecular evolution of an arsenate detoxification pathway by DNA shuffling," Nature Biotechnology, 15:436-438

Crameri et al., (1996:1) "Construction and evolution of antibody-phage libraries by DNA shuffling," Nature Medicine, 2:100-103

Crameri et al., (1996:2) "Improved green fluorescent protein by molecular evolution using DNA shuffling," Nature Biotechnology, 14:315-319

Fischer, C.V., et al., (1992). FASEB J 6(2):759-764.

Graf, R and Schachman H.K. (1996) Proc Natl Acad Sci U S A. 93(21):11591-6.

Hall, J. G. and Frieden, C. (1989) PNAS USA 86, 3060-3064.

Hantgan, R. R. and Taniuchi, H (1977) J. Biol. Chem. 252, 1367-1374.

Henikoff, S. (1984) "Unidirectional digestion with Exonuclease-III creates targeted breakpoints for DNA sequencing", Gene 28: (3) 351-359.

Henikoff S. and Henikoff JG. (1992), PNAS USA 89:10915-9.

Hennecke, J. Sebbel, S. and Glockshuber, R. (1999) J. Mol. Biol. 286(4):1197-215.

Hoheisel, JD (1993), "On the activities of Escherichia coil Exonuclease-III", Anal. Biochem. 209: (2) 238-246.

Horton, R. M. and L. R. Pease (1991). Recombination and mutagenesis of DNA sequences using PCR. In "Directed Mutagenesis - A Practical Approach". McPherson M.J., Oxford, IRL Press: 217-247.

Johansson, K. and Ge L (1999). Phage display of combinatorial peptide and protein libraries and their applications in biology and chemistry. Curr Top Microbiol Immunol 243:87-105.

Karlin and Altschul (1993), PNAS USA 90:5873-5877.

King, P and Goodbourne S (1992) "A method for sequence-specific deletion mutagenesis" NucleiC Acid Research 20 (5) 1039-1044.

Labhardt, A. M.. (1982) J. Mol. Biol. 157, 331-355.

Minshull, J. and Stemmer, W.P.C. (1999) Curr. Opin. Chem. Biol. 3: (3) 284-290.

Mullins, L. S. et al. (1994) J. Am. Chem. Soc. 116, 5529-5533.

Ness et al., (1999) "DNA Shuffling of subgenomic sequences of subtilisin," Nature Biotechnology, 17:893-896

Okkels, J. S. PCT application WO 97/07205 (1997).

Ostermeier, M., A. E. Nixon, J. H. Shim and S. J. Benkovic (1999) PNAS USA 96, 3562-3567.

Ostermeier, M., Nixon, A.E., and Benkovic, S.J. (1999) Bioorganic & Medicinal Chem 7; 2139-2144.

Ostermeier, M., Shim, J.H., and Benkovic, S.J. (1999) Nature Biotechnol 17; 1205-1160.

Patten et al., (1997) "Applications of DNA Shuffling to Pharmaceuticals and Vaccines," Current Opinion in Biotechnology, 8:724-733

Protasova, N. Y. et al. (1994) Prot. Eng. 7, 1373-1377.

Putney, S.D. et al., (1981) Proc. Natl. Acad. Sci. USA 78, No 12,7350.

Sambrook, J., E. F. Fritsch, et al. (1989). Molecular Cloning - A Laboratory Manual, Cold Spring Harbor Laboratory press.

Schwaneberg, U., et al. (1999). Anal. Biochem., 269, 359-366.

Shao, Z., H. Zhao, L. Giver, and F.H. Arnold (1998) Nucleic Acids Res., 26, 681 *et seq.*

Shimada, T and Yanazaki, H. (1998) "Cytochrome P450 Reconstitution Systems" Methods in Molecular Biology Vol. 107: Cytochrome P450 Protocols, Humana Press Inc., Totowa, N.J.

Shiba, K. and Schimmel, P. (1992) PNAS USA 89, 1880-1884.

Stemmer, W. P. C. (1994:1) Nature, 370, 389 *et. seq.*

Stemmer, W. P. C. (1994:2) "DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution," Proc. Natl. Acad. Sci. USA, 91:10747-10751

Taniuchi, H. et al. (1977) J. Biol. Chem. 252, 125-140.

Vignais, M.-L. et al. (1995) Protein Sci. 4, 994-1000.

Volkov, A. A., Shao Z., and Arnold F. H. (1999), Nucleic Acids Res. 27:e18.

Yang, R. Y. and Schachman, H. K. (1) (1993) Protein Sci. 2, 1013-1023.

Yang, R. Y. and Schachman, H. K. (2) (1993) PNAS USA 90, 11980-11984.

Zhang, T. et al., (1997) "Directed evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening," Proc. Natl. Acad. Sci. USA, 94:4504-4509

Zhang, T. et al. (1993) Biochemistry 32, 12311-12318.

Zhao, H., L. Giver, Z. Shao, J.A. Affholter, and F.H. Arnold (1998) Nat. Biotechnol., 16, 258 *et. seq.*.

Patent Literature

EP 932670 by Stemmer, "Evolving Cellular DNA Uptake by Recursive Sequence Recombination

EP 752008 by Stemmer and Crameri, "DNA Mutagenesis by Random Fragmentation and Reassembly

U.S. Patent No. 5,965,408 to Short "Method of DNA reassembly by interrupting synthesis"

U.S. Patent No. 5,837,458 to Minshull et al. (November 17, 1998), "Methods and Compositions for Cellular and Metabolic Engineering"

U.S. Patent No. 5,830,721 to Stemmer et al. (November 3, 1998), "DNA Mutagenesis by Random Fragmentation and Reassembly"

U.S. Patent No. 5,811,238 to Stemmer et al. (September 22, 1998), "Method for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination"

U.S. Patent No. 5,605,793 to Stemmer (February 25, 1997), "Method for In Vitro Recombination"

WO 00/18906, "Shuffling of Codon altered Genes," by Patten et al.

WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences," (Proteus)

WO 00/04190, "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," by del Cardayre et al.

WO 00/00632, "Methods for Generating Highly Diverse Libraries" by Wagner et al.

WO 98/42832, "Recombination of Polynucleotide Sequences Using Random or Defined Primers," by Arnold et al.

WO 98/31837, "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," by del Cardayre et al.

WO 98/27230, "Methods and Compositions for Polypeptide Engineering," by Patten and Stemmer

WO 97/35966, "Methods and Compositions for Cellular and Metabolic Engineering," Minshull and Stemmer

WO 97/20078, "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination," by Stemmer and Crameri

WO 95/22625, "Mutagenesis by Random Fragmentation and Reassembly," Stemmer and Crameri

CLAIMS

1. A method of screening a gene library, comprising the steps of:

(a) digesting a polynucleotide construct to form fragments, the polynucleotide construct comprising a first polynucleotide encoding a first polypeptide and a second polynucleotide encoding a second polypeptide, and the first and second polynucleotides connected by a linker sequence comprising a restriction site;

(b) selecting fragments which approximate a predetermined size;

(c) circularizing the selected fragments;

(d) linearizing the circularized fragments by treatment with a restriction enzyme specific for the restriction site to form a gene library; and

(e) screening the gene library for a linearized fragment encoding a hybrid protein having an improved property over at least one of the first and second polypeptides.

2. The method of claim 1, wherein the digesting comprises digestion of both the 3' and the 5' ends of the polynucleotide construct, and wherein the restriction site is unique in the construct.

3. The method of claim 1, wherein the digesting step comprises digestion with a DNase or a nuclease.

4. The method of claim 3, wherein the DNase is DNase I.

5. The method of claim 3, wherein the nuclease is Exonuclease III.

6. The method of claim 1, wherein the predetermined size is a size in the range of the size of the first polynucleotide to the size of the second polynucleotide.

7. The method of claim 1, wherein the selecting fragments step comprises gel electrophoresis of the fragments.

8. The method of claim 1, comprising converting the ends of each selected fragment from staggered to blunt ends prior to circularizing.

9. The method of claim 1, comprising inserting each linearized fragment into an expression vector prior to screening.

10. The method of claim 9, comprising subjecting each linearized fragment to PCR amplification prior to inserting.

11. The method of claim 10, comprising using a PCR primer specific a sequence selected from an original terminus of the first polynucleotide, an original terminus of the second polynucleotide, and a sequence located within a linker sequence.

12. The method of claim 1, wherein the first and second polypeptides are different.

13. The method of claim 12, wherein the first and second polypeptides are different in one of the following:

(a) the first polypeptide has an N-terminal sequence providing an improved property over the N-terminal of the second polypeptide;

(b) the second polypeptide has an N-terminal sequence providing an improved property over the N-terminal of the first polypeptide;

(c) the first polypeptide has a C-terminal sequence providing an improved property over the C-terminal of the second polypeptide; and

(d) the second polypeptide has a C-terminal sequence providing an improved property over the C-terminal of the first polypeptide.

14. The method of claim 13, wherein the property is selected from hydrophilicity, hydrophobicity, foldability, expressability, substrate specificity, reaction product, and enzyme activity.

15. The method of claim 1, wherein the hybrid protein has an N-terminal sequence providing an improved property over either the N-terminal region of the first polypeptide or the N-terminal sequence of the second polypeptide.

16. The method of claim 1, wherein the hybrid protein has a C-terminal sequence providing an improved property over either the C-terminal region of the first polypeptide or the C-terminal sequence of the second polypeptide.

17. The method of claim 15, wherein the property is selected from the group consisting of hydrophilicity, hydrophobicity, foldability, expressability, substrate specificity, reaction product, and enzyme activity.

18. The method of claim 1, wherein the sequence identity between the first and second polypeptides is at least 50%.

19. The method of claim 1, wherein the first and second polypeptides are identical.

20. A method of preparing a hybrid protein, comprising the steps of:

(a) digesting a polynucleotide construct to form fragments, the polynucleotide construct comprising a first polynucleotide encoding a first polypeptide and a second polynucleotide encoding a second polypeptide, and the first and second polynucleotides connected by a linker sequence comprising a restriction site;

(b) selecting fragments which approximate a predetermined size;

(c) circularizing the selected fragments;

(d) linearizing the circularized fragments by treatment with a restriction enzyme specific for the restriction site; and

(e) selecting any linearized fragment encoding a hybrid protein having an improved property over at least one of the first and second polypeptides.

SEQUENCE LISTING

<110> SIEBER, Volker
ZHANG, Ji-Hu
ARNOLD, Frances H.

<120> PRODUCTION OF FUNCTIONAL HYBRID GENES AND PROTEINS

<130> 3369/2F911-WO2

<140> PCT/US00/29717
<141> 2000-10-27

<150> U.S. 09/672,629
<151> 2000-09-28

<140> U.S. 09/672,859
<141> 2000-09-28

<150> U.S. 60/161,852
<151> 1999-10-27

<150> U.S. 60/161,850
<151> 2000-10-27

<160> 36

<170> FastSEQ for Windows Version 3.0

<210> 1
<211> 36
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 1
taaggggaac tcgagatgag taaaggagaa gaactt 36

<210> 2
<211> 36
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 2
catctcgagt tcccccttatt tgtatagttc atccat 36

<210> 3

<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 3
taaggggaac tcgagatg 18

<210> 4
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 4
accatgatta cgccaagctt g 21

<210> 5
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 5
gggcccgtac ggccgactag t 21

<210> 6
<211> 41
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<221> modified_base
<222> (1)...(1)
<223> 5'-phosphorylated

<400> 6
cgattatttt tcagctaggc ctagttggta atggtagcga c 41

<210> 7
<211> 42
<212> DNA
<213> Artificial Sequence

<220>

<223> Primer

<400> 7
gcagggcccc agagctcatg gctctgttat tagcagtttt tc 42

<210> 8
<211> 37
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 8
cgctctagag gtaccccaat tgatggagaa gcgccgc 37

<210> 9
<211> 50
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 9
cgacggatcc ggaaggaagg gcccatatga caattaaaga aatgcctcag 50

<210> 10
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 10
gcaaagacca atcgtatcaa gcg 23

<210> 11
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 11
gcttgatacg attggtcttt gc 22

<210> 12
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 12
cgtttagcat gtgcgttaat aaatc 25

<210> 13
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 13
gatttattaa cgcacatgct aaacg 25

<210> 14
<211> 40
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 14
cgtcggtacc ctcgagtgaa gtgctaggtg aaggaatacc 40

<210> 15
<211> 44
<212> DNA
<213> Artificial Sequence

<220>
<223> linker sequence

<400> 15
ttggggtacc tctagaacta gtggatccgg aaggaagggc ccat 44

<210> 16
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 16
cagctggggt ctgtcagaga gc 22

<210> 17
<211> 33

<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 17
cgacgatcta gattacgccc cgccctgcca ctc 33

<210> 18
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Linker sequence

<400> 18
tggcctgggt cccctgct 18

<210> 19
<211> 42
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 19
cctgggtccc ctgctagcga gaaaaaaatc actggatata cc 42

<210> 20
<211> 42
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 20
ggtatatcca gtgatttttt tctcgctagc aggggaccca gg 42

<210> 21
<211> 51
<212> DNA
<213> Artificial Sequence

<220>
<223> Primer

<400> 21
ccaggatcca tcgatgctta ggaggtcata tgacaattaa agaaatgcct c 51

<210> 22
<211> 60
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(42)
<223> P450-BM3

<221> source
<222> (43)...(60)
<223> P450-1A2

<400> 22
atgacaatta aagaaatgcc tcagccaaaa acgtttggag aggtgctcaa gggtttgagg 60

<210> 23
<211> 60
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(43)
<223> P450-BM3

<221> source
<222> (44)...(60)
<223> P450-1A2

<400> 23
atgacaatta aagaaatgcc tcagccaaaa cgtttggaga gccccaaagg cctgaaaagt 60

<210> 24
<211> 40
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(20)
<223> P450-BM3

<221> source
<222> (21)...(40)
<223> P450-1A2

<400> 24
agcatgcgtt taaaccgttt ggacccctct gagttccggc 40

<210> 25
<211> 40

<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(20)
<223> P450-BM3

<221> source
<222> (21)...(40)
<223> P450-1A2

<400> 25
cagtcagcag gcaatgaaag gcccgccggc gcctggccca 40

<210> 26
<211> 40
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(20)
<223> P450-BM3

<221> source
<222> (21)...(40)
<223> P450-1A2

<400> 26
tatcgccgtg cagcttgttc caagggccgg cctgacctct 40

<210> 27
<211> 1515
<212> DNA
<213> Artificial Sequence

<220>
<221> mutation
<222> (1)...(27)
<223> Modified as described in Fisher et al., 1992.

<400> 27
atggctctgt tattagcagt ttttctgttc tgcctggtat tctgggtgct caagggtttg 60
aggcctcggg tccccaaagg cctgaaaagt ccaccagagc catggggctg gcccttgctc 120
gggcatgtgc tgaccctggg gaagaacccg cacctggcac tgtcaaggat gagccagcgc 180
tacggggacg tcctgcagat ccgcattggc tccacgcccg tgctggtgct gagccgcctg 240
gacaccatcc ggcaggccct ggtgcggcag ggcgacgatt tcaagggccg gcctgacctc 300
tacacctcca ccctcatcac tgatggccag agcttgacct tcagcacaga ctctggaccg 360
gtgtgggctg cccgccggcg cctggcccag aatgccctca acaccttctc catcgcctct 420
gacccagctt cctcatcctc ctgctacctg gaggagcatg tgagcaagga ggctaaggcc 480
ctgatcagca ggttgcagga gctgatggca gggcctgggc acttcgaccc ttacaatcag 540
gtggtggtgt cagtggccaa cgtcattggt gccatgtgct tcggacagca cttccctgag 600

agtagcgatg agatgctcag cctcgtgaag aacactcatg agttcgtgga gactgcctcc 660
tccgggaacc ccctggactt cttccccatc cttcgctacc tgcctaaccc tgccctgcag 720
aggttcaagg ccttcaacca gaggttcctg tggttcctgc agaaaacagt ccaggagcac 780
tatcaggact ttgacaagaa cagtgtccgg gacatcacgg gtgccctgtt caagcacagc 840
aagaaggggc ctagagccag cggcaacctc atcccacagg agaagattgt caaccttgtc 900
aatgacatct ttggagcagg atttgacaca gtcaccacag ccatctcctg gagcctcatg 960
taccttgtga ccaagcctga gatacagagg aagatccaga aggagctgga cactgtgatt 1020
ggcagggagc ggcggccccg gctctctgac agaccccagc tgccctactt ggaggccttc 1080
atcctggaga ccttccgaca ctcctccttc ttgcccttca ccatccccca cagcacaaca 1140
agggacacaa cgctgaatgg cttctacatc cccaagaaat gctgtgtctt cgtaaaccag 1200
tggcaggtca accatgaccc agagctgtgg gaggaccccc ctgagttccg gcctgagcgg 1260
ttcctcaccg ccgatggcac tgccattaac aagcccttga gtgagaagat gatgctgttt 1320
ggcatgggca agcgccggtg tatcggggaa gtcctggcca agtgggagat cttcctcttc 1380
ctggccatcc tgctacagca actggagttc agcgtgccgc cgggcgtgaa agtcgacctg 1440
acccccatct acgggctgac catgaagcac gcccgctgtg aacatgtcca ggcgcggctg 1500
cgcttctcca tcaat 1515

<210> 28
<211> 1392
<212> DNA
<213> Artificial Sequence

<220>
<221> mutation
<222> (261)...(264)
<223> From ATTT to GGCC

<221> mutation
<222> (459)...(459)
<223> From A to G

<221> mutation
<222> (711)...(711)
<223> From T to C

<400> 28
atgacaatta aagaaatgcc tcagccaaaa acgtttggag agcttaaaaa tttaccgtta 60
ttaaacacag ataaaccggt tcaagctttg atgaaaattg cggatgaatt aggagaaatc 120
tttaaattcg aggcgcctgg tcgtgtaacg cgctacttat caagtcagcg tctaattaaa 180
gaagcatgcg atgaatcacg ctttgataaa aacttaagtc aagcgcttaa atttgtacgt 240
gattttgcag gagacgggtt ggccacaagc tggacgcatg aaaaaaattg gaaaaaagcg 300
cataatatct tacttccaag cttcagtcag caggcaatga aaggctatca tgcgatgatg 360
gtcgatatcg ccgtgcagct tgttcaaaag tgggagcgtc taaatgcaga tgagcatatt 420
gaagtaccgg aagacatgac acgtttaacg cttgatacga ttggtctttg cggctttaac 480
tatcgcttta acagctttta ccgagatcag cctcatccat ttattacaag tatggtccgt 540
gcactggatg aagcaatgaa caagctgcag cgagcaaatc cagacgaccc agcttatgat 600
gaaaacaagc gccagtttca agaagatatc aaggtgatga acgacctagt agataaaatt 660
attgcagatc gcaaagcaag cggtgaacaa agcgatgatt tattaacgca catgctaaac 720
ggaaaagatc cagaaacggg tgagccgctt gatgacgaga acattcgcta tcaaattatt 780
acattcttaa ttgcgggaca cgaaacaaca agtggtcttt tatcatttgc gctgtatttc 840
ttagtgaaaa atccacatgt attacaaaaa gcagcagaag aagcagcacg agttctagta 900
gatcctgttc caagctacaa acaagtcaaa cagcttaaat atgtcggcat ggtcttaaac 960
gaagcgctgc gcttatggcc aactgctcct gcgttttccc tatatgcaaa agaagatacg 1020

gtgcttggag gagaatatcc tttagaaaaa ggcgacgaac taatggttct gattcctcag 1080
cttcaccgtg ataaaacaat ttggggagac gatgtggaag agttccgtcc agagcgtttt 1140
gaaaatccaa gtgcgattcc gcagcatgcg tttaaaccgt ttggaaacgg tcagcgtgcg 1200
tgtatcggtc agcagttcgc tcttcatgaa gcaacgctgg tacttggtat gatgctaaaa 1260
cactttgact ttgaagatca tacaaactac gagctggata ttaaagaaac tttaacgtta 1320
aaacctgaag gctttgtggt aaaagcaaaa tcgaaaaaaa ttccgcttgg cggtattcct 1380
tcacctagca ct 1392

<210> 29
<211> 1512
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(42)
<223> P450-BM3

<221> source
<222> (43)...(1512)
<223> P450-1A2

<400> 29
atgacaatta aagaaatgcc tcagccaaaa acgtttggag aggtgctcaa gggtttgagg 60
cctcgggtcc ccaaaggcct gaaaagtcca ccagagccat ggggctggcc cttgctcggg 120
catgtgctga ccctggggaa gaacccgcac ctggcactgt caaggatgag ccagcgctac 180
ggggacgtcc tgcagatccg cattggctcc acgcccgtgc tggtgctgag ccgcctggac 240
accatccggc aggccctggt gcggcagggc gacgatttca agggccggcc tgacctctac 300
acctccaccc tcatcactga tggccagagc ttgaccttca gcacagactc tggaccggtg 360
tgggctgccc gccggcgcct ggcccagaat gccctcaaca ccttctccat cgcctctgac 420
ccagcttcct catcctcctg ctacctggag gagcatgtga gcaaggaggc taaggccctg 480
atcagcaggt tgcaggagct gatggcaggg cctgggcact tcgacccctta caatcaggtg 540
gtggtgtcag tggccaacgt cattggtgcc atgtgcttcg gacagcactt ccctgagagt 600
agcgatgaga tgctcagcct cgtgaagaac actcatgagt tcgtggagac tgcctcctcc 660
gggaaccccc tggacttctt ccccatcctt cgctacctgc ctaaccctgc cctgcagagg 720
ttcaaggcct tcaaccagag gttcctgtgg ttcctgcaga aaacagtcca ggagcactat 780
caggactttg acaagaacag tgtccgggac atcacgggtg ccctgttcaa gcacagcaag 840
aaggggccta gagccagcgg caacctcatc ccacaggaga agattgtcaa ccttgtcaat 900
gacatctttg gagcaggatt tgacacagtc accacagcca tctcctggag cctcatgtac 960
cttgtgacca agcctgagat acagaggaag atccagaagg agctggacac tgtgattggc 1020
agggagcggc ggccccggct ctctgacaga ccccagctgc cctacttgga ggccttcatc 1080
ctggagacct tccgacactc ctccttcttg cccttcacca tcccccacag cacaacaagg 1140
gacacaacgc tgaatggctt ctacatcccc aagaaatgct gtgtcttcgt aaaccagtgg 1200
caggtcaacc atgacccaga gctgtgggag gacccctctg agttccggcc tgagcggttc 1260
ctcaccgccg atggcactgc cattaacaag cccttgagtg agaagatgat gctgtttggc 1320
atgggcaagc gccggtgtat cggggaagtc ctggccaagt gggagatctt cctcttcctg 1380
gccatcctgc tacagcaact ggagttcagc gtgccgccgg gcgtgaaagt cgacctgacc 1440
cccatctacg ggctgaccat gaagcacgcc cgctgtgaac atgtccaggc gcggctgcgc 1500
ttctccatca at 1512

<210> 30
<211> 1485
<212> DNA

<213> Artificial Sequence

<220>
<221> source
<222> (1)...(42)
<223> P450-BM3

<221> source
<222> (43)...(1485)
<223> P450-1A2

<400> 30

atgacaatta aagaaatgcc tcagccaaaa cgtttggaga gccccaaagg cctgaaaagt 60
ccaccagagc catggggctg gcccttgctc gggcatgtgc tgaccctggg gaagaacccg 120
cacctggcac tgtcaaggat gagccagcgc tacggggacg tcctgcagat ccgcattggc 180
tccacgcccg tgctggtgct gagccgcctg gacaccatcc ggcaggccct ggtgcggcag 240
ggcgacgatt tcaagggccg gcctgacctc tacacctcca ccctcatcac tgatggccag 300
agcttgacct tcagcacaga ctctggaccg gtgtgggctg cccgccggcg cctggcccag 360
aatgccctca acaccttctc catcgcctct gacccagctt cctcatcctc ctgctacctg 420
gaggagcatg tgagcaagga ggctaaggcc ctgatcagca ggttgcagga gctgatggca 480
gggcctgggc acttcgaccc ttacaatcag gtggtggtgt cagtggccaa cgtcattggt 540
gccatgtgct tcggacagca cttccctgag agtagcgatg agatgctcag cctcgtgaag 600
aacactcatg agttcgtgga gactgcctcc tccgggaacc ccctggactt cttccccatc 660
cttcgctacc tgcctaaccc tgccctgcag aggttcaagg ccttcaacca gaggttcctg 720
tggttcctgc agaaaacagt ccaggagcac tatcaggact ttgacaagaa cagtgtccgg 780
gacatcacgg gtgccctgtt caagcacagc aagaaggggc ctagagccag cggcaacctc 840
atcccacagg agaagattgt caaccttgtc aatgacatct ttggagcagg atttgacaca 900
gtcaccacag ccatctcctg gagcctcatg taccttgtga ccaagcctga gatacagagg 960
aagatccaga aggagctgga cactgtgatt ggcagggagc ggcggccccg gctctctgac 1020
agaccccagc tgccctactt ggaggccttc atcctggaga ccttccgaca ctcctccttc 1080
ttgcccttca ccatccccca cagcacaaca agggacacaa cgctgaatgg cttctacatc 1140
cccaagaaat gctgtgtctt cgtaaaccag tggcaggtca accatgaccc agagctgtgg 1200
gaggaccccc ctgagttccg gcctgagcgg ttcctcaccg ccgatggcac tgccattaac 1260
aagcccttga gtgagaagat gatgctgttt ggcatgggca agcgccggtg tatcggggaa 1320
gtcctggcca agtgggagat cttcctcttc ctggccatcc tgctacagca actggagttc 1380
agcgtgccgc cgggcgtgaa agtcgacctg acccccatct acgggctgac catgaagcac 1440
gcccgctgtg aacatgtcca ggcgcggctg cgcttctcca tcaat 1485

<210> 31
<211> 1465
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(1182)
<223> P450-BM3

<221> source
<222> (1183)...(1465)
<223> P450-1A2

<400> 31

atgacaatta aagaaatgcc tcagccaaaa acgtttggag agcttaaaaa tttaccgtta 60
ttaaacacag ataaaccggt tcaagctttg atgaaaattg cggatgaatt aggagaaatc 120
tttaaattcg aggcgcctgg tcgtgtaacg cgctacttat caagtcagcg tctaattaaa 180
gaagcatgcg atgaatcacg ctttgataaa aacttaagtc aagcgcttaa atttgtacgt 240
gattttgcag gagacgggtt ggccacaagc tggacgcatg aaaaaaattg gaaaaaagcg 300
cataatatct tacttccaag cttcagtcag caggcaatga aaggctatca tgcgatgatg 360
gtcgatatcg ccgtgcagct tgttcaaaag tgggagcgtc taaatgcaga tgagcatatt 420
gaagtaccgg aagacatgac acgtttaacg cttgatacga ttggtctttg cggctttaac 480
tatcgcttta acagctttta ccgagatcag cctcatccat ttattacaag tatggtccgt 540
gcactggatg aagcaatgaa caagctgcag cgagcaaatc cagacgaccc agcttatgat 600
gaaaacaagc gccagtttca agaagatatc aaggtgatga acgacctagt agataaaatt 660
attgcagatc gcaaagcaag cggtgaacaa agcgatgatt tattaacgca catgctaaac 720
ggaaaagatc cagaaacggg tgagccgctt gatgacgaga acattcgcta tcaaattatt 780
acattcttaa ttgcgggaca cgaaacaaca agtggtcttt tatcatttgc gctgtatttc 840
ttagtgaaaa atccacatgt attacaaaaa gcagcagaag aagcagcacg agttctagta 900
gatcctgttc caagctacaa acaagtcaaa cagcttaaat atgtcggcat ggtcttaaac 960
gaagcgctgc gcttatggcc aactgctcct gcgttttccc tatatgcaaa agaagatacg 1020
gtgcttggag gagaatatcc tttagaaaaa ggcgacgaac taatggttct gattcctcag 1080
cttcaccgtg ataaaacaat ttggggagac gatgtggaag agttccgtcc agagcgtttt 1140
gaaaatccaa gtgcgattcc gcagcatgcg tttaaaccgt ttggacccct ctgagttccg 1200
gcctgagcgg ttcctcaccg ccgatggcac tgccattaac aagcccttga gtgagaagat 1260
gatgctgttt ggcatgggca agcgccggtg tatcggggaa gtcctggcca agtgggagat 1320
cttcctcttc ctggccatcc tgctacagca actggagttc agcgtgccgc cgggcgtgaa 1380
agtcgacctg acccccatct acgggctgac catgaagcac gcccgctgtg aacatgtcca 1440
ggcgcggctg cgcttctcca tcaat 1465

<210> 32
<211> 1489
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(343)
<223> P450-BM3

<221> source
<222> (344)...(1489)
<223> P450-1A2

<400> 32
atgacaatta aagaaatgcc tcagccaaaa acgtttggag agcttaaaaa tttaccgtta 60
ttaaacacag ataaaccggt tcaagctttg atgaaaattg cggatgaatt aggagaaatc 120
tttaaattcg aggcgcctgg tcgtgtaacg cgctacttat caagtcagcg tctaattaaa 180
gaagcatgcg atgaatcacg ctttgataaa aacttaagtc aagcgcttaa atttgtacgt 240
gattttgcag gagacgggtt ggccacaagc tggacgcatg aaaaaaattg gaaaaaagcg 300
cataatatct tacttccaag cttcagtcag caggcaatga aaggcccgcc ggcgcctggc 360
ccagaatgcc ctcaacacct tctccatcgc ctctgaccca gcttcctcat cctcctgcta 420
cctggaggag catgtgagca aggaggctaa ggccctgatc agcaggttgc aggagctgat 480
ggcagggcct gggcacttcg acccttacaa tcaggtggtg gtgtcagtgg ccaacgtcat 540
tggtgccatg tgcttcggac agcacttccc tgagagtagc gatgagatgc tcagcctcgt 600
gaagaacact catgagttcg tggagactgc ctcctccggg aacccccctgg acttcttccc 660
catccttcgc tacctgccta accctgccct gcagaggttc aaggccttca accagaggtt 720

cctgtggttc ctgcagaaaa cagtccagga gcactatcag gactttgaca agaacagtgt 780
ccgggacatc acgggtgccc tgttcaagca cagcaagaag gggcctagag ccagcggcaa 840
cctcatccca caggagaaga ttgtcaacct tgtcaatgac atctttggag caggatttga 900
cacagtcacc acagccatct cctggagcct catgtacctt gtgaccaagc ctgagataca 960
gaggaagatc cagaaggagc tggacactgt gattggcagg gagcggcggc cccggctctc 1020
tgacagaccc cagctgccct acttggaggc cttcatcctg gagaccttcc gacactcctc 1080
cttcttgccc ttcaccatcc cccacagcac aacaagggac acaacgctga atggcttcta 1140
catccccaag aaatgctgtg tcttcgtaaa ccagtggcag gtcaaccatg acccagagct 1200
gtgggaggac ccctctgagt tccggcctga gcggttcctc accgccgatg gcactgccat 1260
taacaagccc ttgagtgaga agatgatgct gtttggcatg ggcaagcgcc ggtgtatcgg 1320
ggaagtcctg gccaagtggg agatcttcct cttcctggcc atcctgctac agcaactgga 1380
gttcagcgtg ccgccgggcg tgaaagtcga cctgaccccc atctacgggc tgaccatgaa 1440
gcacgcccgc tgtgaacatg tccaggcgcg gctgcgcttc tccatcaat 1489

<210> 33
<211> 1619
<212> DNA
<213> Artificial Sequence

<220>
<221> source
<222> (1)...(385)
<223> P450-BM3

<221> source
<222> (386)...(1619)
<223> P450-1A2

<400> 33
atgacaatta aagaaatgcc tcagccaaaa acgtttggag agcttaaaaa tttaccgtta 60
ttaaacacag ataaaccggt tcaagctttg atgaaaattg cggatgaatt aggagaaatc 120
tttaaattcg aggcgcctgg tcgtgtaacg cgctacttat caagtcagcg tctaattaaa 180
gaagcatgcg atgaatcacg ctttgataaa aacttaagtc aagcgcttaa atttgtacgt 240
gattttgcag gagacgggtt ggccacaagc tggacgcatg aaaaaaattg gaaaaaagcg 300
cataatatct tacttccaag cttcagtcag caggcaatga aaggctatca tgcgatgatg 360
gtcgatatcg ccgtgcagct tgttccaagg gccggcctga cctctacacc tccaccctca 420
tcactgatgg ccagagcttg accttcagca cagactctgg accggtgtgg gctgcccgcc 480
ggcgcctggc ccagaatgcc ctcaacacct tctccatcgc ctctgaccca gcttcctcat 540
cctcctgcta cctggaggag catgtgagca aggaggctaa ggccctgatc agcaggttgc 600
aggagctgat ggcagggcct gggcacttcg acccttacaa tcaggtggtg gtgtcagtgg 660
ccaacgtcat tggtgccatg tgcttcggac agcacttccc tgagagtagc gatgagatgc 720
tcagcctcgt gaagaacact catgagttcg tggagactgc ctcctccggg aacccccctgg 780
acttcttccc catccttcgc tacctgccta accctgccct gcagaggttc aaggccttca 840
accagaggtt cctgtggttc ctgcagaaaa cagtccagga gcactatcag gactttgaca 900
agaacagtgt ccgggacatc acgggtgccc tgttcaagca cagcaagaag gggcctagag 960
ccagcggcaa cctcatccca caggagaaga ttgtcaacct tgtcaatgac atctttggag 1020
caggatttga cacagtcacc acagccatct cctggagcct catgtacctt gtgaccaagc 1080
ctgagataca gaggaagatc cagaaggagc tggacactgt gattggcagg gagcggcggc 1140
cccggctctc tgacagaccc cagctgccct acttggaggc cttcatcctg gagaccttcc 1200
gacactcctc cttcttgccc ttcaccatcc cccacagcac aacaagggac acaacgctga 1260
atggcttcta catccccaag aaatgctgtg tcttcgtaaa ccagtggcag gtcaaccatg 1320
acccagagct gtgggaggac ccctctgagt tccggcctga gcggttcctc accgccgatg 1380
gcactgccat taacaagccc ttgagtgaga agatgatgct gtttggcatg ggcaagcgcc 1440

ggtgtatcgg ggaagtcctg gccaagtggg agatcttcct cttcctggcc atcctgctac 1500
agcaactgga gttcagcgtg ccgccgggcg tgaaagtcga cctgaccccc atctacgggc 1560
tgaccatgaa gcacgcccgc tgtgaacatg tccaggcgcg gctgcgcttc tccatcaat 1619

<210> 34
<211> 6
<212> PRT
<213> Artificial Sequence

<220>
<223> Linker sequence

<400> 34
Trp Pro Gly Ser Pro Ala
1 5

<210> 35
<211> 20
<212> PRT
<213> Artificial Sequence

<220>
<221> UNSURE
<222> (1)...(14)
<223> Sequence from P450-BM3

<221> UNSURE
<222> (15)...(20)
<223> Sequence from P450-1A2

<400> 35
Met Thr Ile Lys Glu Met Pro Gln Pro Lys Thr Phe Gly Glu Val Leu
1 5 10 15
Lys Gly Leu Arg
20

<210> 36
<211> 20
<212> PRT
<213> Artificial Sequence

<220>
<221> UNSURE
<222> (1)...(10)
<223> Sequence from P450-BM3

<221> UNSURE
<222> (15)...(20)
<223> Sequence from P450-1A2

<400> 36
Met Thr Ile Lys Glu Met Pro Gln Pro Lys Arg Leu Glu Ser Pro Lys
1 5 10 15

22 MAR 2004

Gly Leu Lys Ser
20