METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES AND POLYPEPTIDES HAVING DESIRED CHARACTERISTICS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation-in-part of "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES AND POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., USSN 09/416,375, filed October 12, 1999, which is a non provisional of "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES AND POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov and Stemmer, USSN 60/116,447, filed January 19, 1999 and which is also a non-provisional of "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES AND POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov and Stemmer, USSN 60/118,854, filed February 5, 1999.

This application is also a continuation-in-part of "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., Attorney Docket Number 02-296-3 US, filed herewith, which is a continuation-in-part of "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., USSN 09/408,392, filed September 28, 1999, which is a non-provisional of "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., USSN 60/118,813, filed February 5, 1999 and which is also a non-provisional of "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., USSN 60/141,049, filed June 24, 1999.

This application is also a continuation-in-part of co-filed application "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer, Attorney Docket Number 3271.002WO0 (filed by Majestic, Parsons, Siebert & Hsue) which is a continuation-in-part of "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer, USSN 09/416,837, filed October 12, 1999.

This application is also related to "USE OF CODON VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., USSN 09/408,393, filed September 28, 1999.

The present application claims priority to and benefit of each of the applications listed in this section, as provided for under 35 U.S.C. §119(e) and/or 35 U.S.C. §120, as appropriate. All of the preceding applications are incorporated herein by reference.

COPYRIGHT NOTIFICATION

Pursuant to 37 C.F.R. 1.71(e), Applicants note that a portion of this disclosure contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

FIELD OF THE INVENTION

This invention is in the field of genetic algorithms and the application of genetic algorithms to nucleic acid shuffling methods.

BACKGROUND OF THE INVENTION

Recursive nucleic acid recombination ("shuffling") provides for the rapid evolution of nucleic acids, *in vitro* and *in vivo*. This rapid evolution provides for the generation of encoded molecules (e.g., nucleic acids and proteins) with new and/or improved properties. Proteins and nucleic acids of industrial, agricultural and therapeutic importance can be created or improved through DNA shuffling procedures.

A number of publications by the inventors and their co-workers describe DNA shuffling. For example, Stemmer et al. (1994) "Rapid Evolution of a Protein" Nature 370:389-391; Stemmer (1994) "DNA Shuffling by Random Fragmentation and Reassembly: in vitro Recombination for Molecular Evolution," Proc. Natl. Acad. USA 91:10747-10751; Stemmer, U.S. Patent No. 5,603,793 "METHODS FOR IN VITRO RECOMBINATION;" Stemmer et al., U.S. Pat. No. 5,830,721, "DNA MUTAGENESIS BY RANDOM FRAGMENTATION AND REASSEMBLY;" and Stemmer et al., U.S. Pat. No. 5,811,238 "METHODS FOR GENERATING POLYNUCLEOTIDES HAVING DESIRED CHARACTERISTICS BY ITERATIVE SELECTION AND RECOMBINATION" describe, e.g., a variety of shuffling techniques.

Many applications of DNA shuffling technology have also been developed by the inventors and their co-workers. In addition to the publications noted above, Minshull et al., U.S. Pat. No. 5,837,458 METHODS AND COMPOSITIONS FOR CELLULAR AND METABOLIC ENGINEERING provides for the evolution of new metabolic pathways and the enhancement of bio-processing through recursive shuffling techniques. Crameri et al. (1996), "Construction And Evolution Of Antibody-Phage Libraries By DNA Shuffling" Nature Medicine 2(1):100-103 describe, e.g., antibody shuffling for antibody phage libraries. Additional details regarding DNA shuffling can also be found in various published applications, such as WO95/22625, WO97/ 20078, WO96/33207, WO97/33957, WO98/27230, WO97/35966, WO98/ 31837, WO98/13487, WO98/13485 and WO989/42832.

A number of the publications of the inventors and their co-workers, as well as other investigators in the art also describe techniques which facilitate DNA shuffling, e.g., by providing for reassembly of genes from small fragments of genes, or even oligonucleotides encoding gene fragments. In addition to the publications noted above, Stemmer et al. (1998) U.S. Pat. No. 5,834,252 "END COMPLEMENTARY POLYMERASE REACTION" describe processes for amplifying and detecting a target sequence (e.g., in a mixture of nucleic acids), as well as for assembling large polynucleotides from fragments.

Review of the foregoing publications reveals that DNA shuffling is an important new technique with many practical applications. Thus, new techniques which facilitate DNA shuffling are highly desirable. In particular, techniques which reduce the number of physical manipulations needed for shuffling procedures would be particularly useful. The present invention provides significant new DNA shuffling protocols, as well as other features which will be apparent upon complete review of this disclosure.

SUMMARY OF THE INVENTION

The present invention provides new "in silico" DNA shuffling techniques, in which part, or all, of a DNA shuffling procedure is performed or modeled in a computer system, avoiding (partly or entirely) the need for physical manipulation of nucleic acids. These approaches are collectively termed Genetic Algorithm Guided Gene Synthesis or "GAGGS."

In a first aspect, the invention provides methods for obtaining a "chimeric" or "recombinant" polynucleotide or polypeptide (or other bio-polymer) having a desired characteristic. In the methods, at least two parental character strings encoding sequence information for one or more polypeptides and/or for one or more single-stranded or double-stranded polynucleotides are provided. All or a part of the sequences (i.e., one or more

subsequence regions) contain areas of identity and areas of heterology. A set of character strings of a pre-defined or selected length is provided that encodes single-stranded oligonucleotide sequences which include overlapping sequence fragments of at least a part of each of the parental character strings, and/or at least a part of polynucleotide strands complementary to the parental character strings.

In one class of embodiments, the invention provides methods of generating libraries of biological polymers. The method include generating a diverse population of character strings in a computer, where the character strings are generated by alteration (recombination, mutagenesis, etc.) of pre-existing character strings. The diverse population of character strings is then synthesized to comprise the library of biological polymers (nucleic acids, polypeptides, peptide nucleic acids, etc.). Typically, the members of the library of biological polymers are selected for one or more activity. In one recursive aspect of the invention, an additional library or an additional set of character strings is filtered by subtracting the additional library or the additional set of character strings with members of the library of biological polymers which display activity below a desired threshold. In an additional or complementary recursive aspect of the invention, the additional library or additional set of character strings is filtered by biasing the additional library, or the additional set of character strings, with members of the library of biological polymers which display activity above a desired threshold.

A set of single-stranded oligonucleotides made according to the set of sequences defined in the character strings is provided. Part or all of the single stranded nucleotides produced are pooled under denaturing or annealing conditions, where at least two single-stranded oligonucleotides represent parts of two different parental sequences. The resultant population of the single-stranded oligonucleotides is incubated with a polymerase under conditions which result in annealing of the single-stranded fragments at areas of identity to form pairs of annealed fragments. These areas of identity are sufficient for one member of the pair to prime replication of the other, resulting in an increase in the length of the oligonucleotides. The resulting mixture of double- and single-stranded oligonucleotides are denatured into single-stranded fragments. These steps are repeated, such that at least a part of the resultant mixture of single-stranded chimeric and mutagenized polynucleotides are used in the steps of subsequent cycles. Recombinant polynucleotides having evolved toward a desired property are selected or screened for.

In another aspect, the invention provides for the use of genetic operators, e.g., in a computer. In these methods, sequence strings corresponding to the oligonucleotides noted above are selected by the computer from sequence strings corresponding to one or more of the following sets of single-stranded oligonucleotides:

a) oligonucleotides synthesized to contain randomly or non-randomly pre-selected mutations of the parental sequences according to modified sequences including replacement of one or more characters with another character, or deletion or insertion of one or more characters;

b) oligonucleotide sequences synthesized to contain degenerate, mixed or unnatural nucleotides, at one or more randomly or non-randomly pre-selected positions; and,

c) chimeric oligonucleotides synthesized according to artificial sequences of character substrings designed to contain joined partial sequences of at least two parental sequences.

In certain embodiments, oligonucleotides of set (c) contain one or more mutated or degenerate positions defined in sets (a) and (b). The oligonucleotides of set (c) are optionally chimeric nucleotides with crossover points selected according to a method allowing identification of a plurality of character substrings displaying pairwise identity (homology) between any or all of the string pairs comprising sequences of different parental character strings.

Crossover points for making chimeric oligonucleotide sequences are optionally selected randomly, or approximately in the middle of each or a part of the identified pairwise identity (homology) areas, or by any other set of selection criteria.

In one aspect, at least one crossover point for at least one chimeric oligonucleotide sequence is selected from those not within detected identity areas.

In one aspect, the mixtures of single stranded oligonucleotides described above are pooled at least once with an additional set of polynucleotides comprising one or more double-stranded or single-stranded polynucleotide encoded by a part and/or by an entire character string of any of the parental sequences provided, and/or by another character string(s) which contains areas of identity and areas of heterology with any of the parental character strings provided.

The polynucleotides from the additional set of polynucleotides can be obtained by oligonucleotide synthesis of oligonucleotides corresponding to any parental character string (or homolog thereof), or by random fragmentation (e.g., by enzymatic cleavage e.g., by a

DNAse, or by chemical cleavage of the polynucleotide) and/or by a restriction-enzyme fragmentation of polynucleotides encoded by character strings defined above, and/or by another character string(s) which contains areas of identity and areas of heterology with any of the parental character strings provided. That is, any nucleic acid generated by GAGGS can be further modified by any available method to produce additionally diversified nucleic acids. Furthermore, any diversified nucleic acid can serve as a substrate for further rounds of GAGGS.

The above methods are suitably adapted to a wide range of lengths for synthetic oligos (e.g., 10-20 nucleotides or more, 20-40 nucleotides or more, 40-60 nucleotides or more, 60-100 nucleotides or more, 100-150 nucleotides or more, etc.), a wide variety of types of parental sequences (e.g., for therapeutic proteins such as EPO, insulin, growth hormones, antibodies or the like; agricultural proteins such as plant hormones, disease resistance factors, herbicide resistance factors (e.g., p450s,) industrial proteins (e.g., those involved in bacterial oil desulfurization, synthesis of polymers, detoxification proteins and complexes, fermentation or the like)) and for a wide variety in the number of selection/screening cycles (e.g., 1 or more cycle, 2 or more cycle, 3-4 or more cycles, 10 or more cycles, 10-50 or more cycles, 50-100 or more cycles, or more than 100 cycles). Rounds of GAGGS evolution can be alternated with rounds of physical nucleic acid shuffling and/or selection assays under various formats (*in vivo,* or *in vitro*). Selected nucleic acids (i.e., those with desirable properties) can be deconvoluted by sequencing or other procedures such as restriction enzyme analysis, real-time PCR analysis or the like, so that the processes can be started over using the sequence information to guide gene synthesis, e.g., without any physical manipulation of DNA obtained from previous GAGGS rounds.

Typically in the methods above, synthesis of polynucleotides from single-stranded oligonucleotides is performed by assembly PCR. Other options for making nucleic acids include ligation reactions, cloning and the like.

In typical embodiments, the sets of character strings, encoding single-strand oligonucleotides comprising fragments of parental strings, including chimeric and mutated/degenerate fragments of a pre-defined length, are generated using a device comprising a processing element, such as a computer with software for sequence string manipulation.

In one aspect, the invention provides for single parent GAGGS. These methods are set forth in more detail in the examples herein.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 is a flow chart describing a portion of directed evolution by GAGGS.

Fig. 2 is a flow chart describing a portion of directed evolution by GAGGS. The flow chart of Fig. 2 is optionally contiguous from Fig. 1.

Fig. 3 is a flow chart describing a portion of directed evolution by GAGGS. The flow chart of Fig. 3 is optionally contiguous from Fig. 2.

Fig. 4 is a flow chart describing a portion of directed evolution by GAGGS. The flow chart of Fig. 4 is optionally contiguous from Fig. 3.

Fig. 5 is a chart and relational tree showing percent similarity for different subtilisins (an exemplar shuffling target).

Fig. 6 is a pairwise dot-plot alignment showing homology areas for different subtilisins.

Fig. 7 is a pairwise dot-plot alignment showing homology areas for 7 different parental subtilisins.

Fig. 8, Panels A-C are pairwise histograms showing conditions determining probability of crossover point selection can be independently controlled for any region over a selected gene length, as well as independently for the pairs of parents.

Fig. 9 is a chart showing introducing indexed crossover points marker into the sequence of each parent.

Fig 10 shows a procedure for oligonucleotide assembly to make nucleic acids.

Fig. 11 is a continuation of figure 13 showing an oligonucleotide assembly scheme.

Fig. 12 is a difference plot and relatedness tree for shuffling Naphthalene deoxygenase.

Fig. 13 is a schematic of a digital system of the invention.

Fig. 14 is a schematic showing a geometric relationship between nucleotides.

Fig. 15 is a schematic of an HMM matrix.

DETAILED DESCRIPTION

In the methods of the invention, "genetic" or "evolutionary" algorithms are used to produce sequence strings which can be converted into physical molecules, shuffled and tested for a desired property. This greatly expedites forced evolution procedures, as the ability to pre-select substrates for shuffling reduces actual physical manipulation of nucleic acids in shuffling protocols. In addition, the use of character strings as "virtual substrates" for

shuffling protocols, when coupled with gene reconstruction methods, eliminates the need to obtain parental physical molecules encoding genes.

Genetic algorithms (GAs) are used in a wide variety of fields to solve problems which are not fully characterized or which are too complex to allow for full characterization, but for which some analytical evaluation is available. That is, GAs are used to solve problems which can be evaluated by some quantifiable measure for the relative value of a solution (or at least the relative value of one potential solution in comparison to another). The basic concept of a genetic algorithm is to encode a potential solution to a problem as a series of parameters. A single set of parameter values is treated as the "genome" or genetic material of an individual solution. A large population of candidate solutions are created. These solutions can be bred with each other for one or more simulated generations under the principle of survival of the fittest, meaning the probability that an individual solution will pass on some of its parameter values to subsequent solution sets is directly related to the fitness of the individual (i.e., how good that solution is relative to the others in the population for the selected parameter). Breeding takes place through use of operators such as crossovers which simulate basic biological recombination, and mutation. The simple application of these operators with reasonable selection mechanisms has produced startlingly good results over a wide range of problems.

An introduction to genetic algorithms can be found in David E. Goldberg (1989) Genetic Algorithms in Search, Optimization and Machine Learning Addison-Wesley Pub Co; ISBN: 0201157675 and in Timothy Masters (1993) Practical Neural Network Recipes in C++ (Book&Disk edition) Academic Pr; ISBN: 0124790402. A variety of more recent references discuss the use of genetic algorithms used to solve a variety of difficult problems. *See*, e.g., http://garage.cse.msu.edu/papers/papers-index.html and the references cited therein; http://gaslab.cs.unr.edu/ and the references cited therein; http://www.aic.nrl.navy.mil/ and the references cited therein; http://www.cs.gmu.edu/research/gag/ and the references cited therein and http://www.cs.gmu.edu/research/gag/pubs.html and the references cited therein.

In the present invention, a genetic algorithm (GA) is used to provide a character string-based representation of the process of generating biopolymer diversity (computational evolution of character strings by application of one or more genetic operators to a provided population (e.g., a parent library) of character strings, e.g. gene sequences).

A representation of a GA-generated character string population (or "derivative library") is used as a sequence instruction set in a form suitable to control polynucleotide

synthesis (e.g. via non-error-prone synthesis, error-prone synthesis, parallel synthesis, pooled synthesis, chemical synthesis, chemoenzymatic synthesis, (including assembly PCR of synthetic oligonucleotides), and the like). Synthesis of polynucleotides is conducted with sequences encoded by a character string in the derivative library. This creates a physical representation (a library of polynucleotides) of the computation-generated "gene" (or any other character string) diversity.

Physical selection of the polynucleotides having desired characteristics is also optionally (and typically) conducted. Such selection is based on results of physical assays of properties of polynucleotides, or polypeptides, whether translated in-vitro, or expressed in-vivo.

Sequences of those polynucleotides found to have desired characteristics are deconvoluted (e.g., sequenced, or, when positional information is available, by noting the position of the polynucleotide). This is performed by DNA sequencing, by reading a position on an array, real time PCR (e.g., TaqMan), restriction enzyme digestion, or any other method noted herein, or currently available.

These steps are optionally repeated, e.g., for 1-4 or more cycles, each time optionally using the deconvoluted sequences as an information source to generate a new, modified set of character strings to start the procedure with. Of course, any nucleic acid which is generated in silico can be synthesized and shuffled by any known DNA shuffling method, including those taught in the references by the inventors and their co-workers cited herein. Such synthesized DNAs can also be mutagenized or otherwise modified according to existing techniques.

In summary, GAGGS is an evolutionary process which includes an information manipulation step (application of a genetic algorithm to a character string representing a biopolymer such as a nucleic acid or protein), to create a set of defined information elements (e.g., character strings) which serve as templates for synthesizing physical nucleic acids. The information elements can be placed into a database or otherwise manipulated in silico, e.g., by the recursive application of a GA to the sequences which are produced. Corresponding physical nucleic acids can be subjected to recombination/selection or other diversity generating procedures, with the nucleic acids being deconvoluted (e.g., sequenced or otherwise analyzed) and the overall process repeated, as appropriate, to achieve a desired nucleic acid.

Example Advantages of GAGGS

There are a variety of advantages to GAGGS as compared to the prior art. For example, physical access to genes/organisms is not required for GAGGS, as sequence information is used for oligo design and selection. A variety of public databases provide extensive sequence information, including, e.g., Genbank™ and those noted *supra*. Additional sequence databases are available on a contract basis from a variety of companies specializing in genomic information generation and storage.

Similarly, sequences from inaccessible, non-cultivable organisms can be used for GAGGS. For example, sequences from pathogenic organisms can be used without actual handling of the pathogens. All of the sequence types suitable for physical DNA shuffling, including damaged and incomplete genes (e.g., pseudo genes), are amenable to GAGGS.

All genetic operators, including different types of mutagenesis and crossovers can be fully and independently controlled in a reproducible fashion, removing human error and variability from physical experiments with DNA manipulations. GAGGS has applicability to the self-learning capability of artificial intelligence (optimization algorithm output parameter profiles based on feedback entry of yields, success rates and failures of physical screens, etc.).

In GAGGS procedures, sequences with frame-shift mutations (which are generally undesirable) are eliminated or fixed (discarded from the character set, or repaired, *in silico*). Similarly, entries with premature terminations are discarded or repaired and entries with loss of sequence features known to be important for display of a desired property (e.g. conservative ligands for metal binding) are discarded or repaired.

Furthermore, wild-type parents do not contaminate derivative libraries with multiple redundant parental molecules, as, in one preferred embodiment, only *a priori* modified genes are subjected to physical shuffling and/or screening (which, in some cases can be expensive, or low throughput, or otherwise less than ideal, depending on the assay available).

In addition, because no actual physical recombination is required, *protein* sequences can be shuffled in the same way *in silico* as nucleic acid sequences, and retrotranslation of the resulting shuffled sequences can be used to alleviate codon usage problems and to minimize the number of oligos needed to build one or more library of coding nucleic acids. In this regard, protein sequences can be shuffled *in silico* using genetic operators that are based on recognition of structural domains and folding motifs, rather than being bound by annealing-based homology criteria of DNA sequences, or simple homology of

AA sequences. Furthermore, rational structure-based biases are easily incorporated in library construction, when such information is available.

The only significant operational costs of running GAGGS is the cost of synthesis of large libraries of genes represented *in silico*. Synthetic assembly of genes can be done, e.g., by assembly PCR from 40-60 bp oligos, which can be synthesized inexpensively by current techniques.

DIRECTED EVOLUTION BY GAGGS:

All changes in any DNA sequence during any evolutionary process can be described by a finite number of events, each resulting from action of an elementary genetic operator. In any given parental sequence subspace, these changes can accurately be accounted for and simulated in a physical representation of an evolutionary process aimed to generate sequence diversity for subsequent physical screening for desired characteristics. Physical double stranded polynucleotides are not required for starting GAGGS processes; instead, they are generated following initial GAGGS processes with the purpose of physical screening and/or selection, and/or as a result of this screening or selection. Generating very large libraries for screening/ selection is not required.

Genetic algorithms (GA).

CHARACTER STRINGS: in general, a character string can be any representation of an array of characters (e.g., a linear array of characters provides "words" while a non-linear array can be used as a code to generate a linear array of characters). For practicing GAGGS, character strings are preferably those which encode polynucleotide or polypeptide strings, directly or indirectly, including any encrypted strings, or images, or arrangements of objects which can be transformed unambiguously to character strings representing sequences of monomers or multimers in polynucleotides, polypeptides or the like (whether made of natural or artificial monomers).

GENETIC ALGORITHM: Genetic algorithms generally are processes which mimic evolutionary processes. Genetic algorithms (GAs) are used in a wide variety of fields to solve problems which are not fully characterized or too complex to allow full characterization, but for which some analytical evaluation is available. That is, GAs are used to solve problems which can be evaluated by some quantifiable measure for the relative value of a solution (or at least the relative value of one potential solution in comparison to another). In the context of the present invention, a genetic algorithm is a process for selecting or manipulating character

strings in a computer, typically where the character string can be corresponded to one or more biological polymer (e.g., a nucleic acid, protein, PNA, or the like). A biological polymer is any polymer which shares some structural features with naturally occurring polymers such as an RNAs, DNAs and polypeptides, including, e.g., RNAs, RNA analogues, DNAs, DNA analogues, polypeptides, polypeptide analogues, peptide nucleic acids, etc.

DIRECTED EVOLUTION OF CHARACTER STRINGS OR OBJECTS:

A process of artificially changing a character string by artificial selection, i.e., which occurs in a reproductive population in which there are (1) varieties of individuals, with some varieties being (2) heritable, of which some varieties (3) differ in fitness (reproductive success determined by outcome of selection for a predetermined property (desired characteristic). The reproductive population can be, e.g., a physical population or a virtual population in a computer system.

GENETIC OPERATORS (GOs): user-defined operations, or sets of operations, each comprising a set of logical instructions for manipulations of character strings. Genetic operators are applied to cause changes in populations of individuals in order to find interesting (useful) regions of the search space (populations of individuals with predetermined desired properties) by predetermined means of selection. Predetermined (or partially predetermined) means of selection include computational tools (operators comprising logical steps guided by analysis of information describing libraries of character strings), and physical tools for analysis of physical properties of physical objects, which can be built (synthesized) from matter with the purpose of physically creating a representation of information describing libraries of character strings. In a preferred embodiment, some or all of the logical operations are performed in a computer.

Genetic Operators

All changes in any population of any type of character strings (and thus in any physical properties of physical objects encoded by such strings) can be described as a result of random and/or predetermined application of a finite set of logical algebraic functions comprising various types of genetic operators.

In its mathematical nature, this statement is not a postulated abstract axiom. In fact, this statement is a derivative theorem with stringent formal proof readily derived from Wiles' proof of Fermat's last theorem. The fundamental implication of the Wiles' Proof for evolutionary molecular biology is in the proof of the central conjecture stating that all elliptic curves are in essence modular forms. Particularly, all of the diversity and evolution of living

matter in the universe (i.e., the plurality of objects whose properties can be described by a finite number of elliptical curves) can be described in the language of five basic arithmetic operations: addition, subtraction, multiplication, division and modular forms (i.e., evolution of life can be effectively described by a finite combination of simple changes of information in a finite population of character strings, e.g. all DNA in the universe). This being the case, it is possible to determine the language of nucleic acid-based forms of life, and to define all basic types of genetic operators which apply to nucleic acids under evolutionary selection.

Mathematical modeling of certain genetic operations have been proposed, e.g., in Sun (1999) "Modeling DNA Shuffling" Journal of Computational Biology 6(1):77-90; Kelly et al. (1994) "A test of the Markovian model of DNA evolution" Biometrics 50(3):653-64; Boehnke et al. (1991) "Statistical methods for multipoint radiation hybrid mapping" Am. J. Hum. Genet. 49:1174-1188; Irvine et al. (1991) "SELEXION: systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis" J. Mol. Biol. 222:739-761; Lander and Waterman (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis" Genomics 2:231-239; Lange (1997) Mathematical and Statistical Methods for Genetic Analysis Springer Verlag, NY; Sun and Waterman (1996) "A mathematical analysis of *in vitro* molecular selection-amplification" J. Mol. Biol. 258:650-660; Waterman (1995) Introduction to Computational Biology Chapman and Hall, London, UK.

The following provides a description of certain basic genetic operations applicable to the present invention.

MULTIPLICATION (including duplication and replication) is a form of reproduction of character strings, producing additional copies of character strings comprising parental population/library of strings. Multiplication operators can have many variations. They can be applied to individual strings or to groups of identical or non-identical strings. Selecting groups of strings for multiplication can be random or biased.

MUTATION: all mutation types in each member of a set of strings can be described by several simple operations which can be reduced to elements comprising replacement of one set of the characters with another set of characters. One or more characters can be mutated in a single operation. When more than one character is mutated, the set of characters may or may not be continuous over an entire string length (a feature useful to simulate closely clustered mutations by certain chemical mutagens). A Single point mutation operator replaces a single character with another single character. The nature of the new characters can vary, and they can be from the same set of characters making up parental

strings, or from different, (e.g., to represent degenerate nucleobases, unnatural nucleobases or amino acids, etc). A Deletion mutation is a more complex operator which removes one or more characters from strings. Individual single point deletions in nucleic acid-encoding strings may be not desirable for manipulating strings representing polynucleotide sequences; however, 3x clustered (continuous or dispersed) deletions may be acceptable ("triple deletion frameshifts"). Single point deletions, though, are useful and acceptable for evolutionary computation of strings encoding polypeptides. Insertion mutations are operationally similar to deletions, except that one or more new characters are inserted. The nature of the added characters optionally vary, and they can be from the same set of characters making up parental strings, or from different, (e.g., to represent degenerate nucleobases, unnatural nucleobases or amino acids, etc). Death can simply be defined as a variation of the deletion operator. It takes place when the result of an application of a genetic operator (or combinations thereof) yields a deletion of an entire individual character string, or entire (sub) population of character strings. Death can also be defined as a variation of an elitism-prone multiplication operator (multiplication of values defining abundance level of one or more strings by zero). Death can also be defined as a default non-selection action in operators, effecting selections of sub-populations of string and transfer manipulations with various sorting and indexing operations of indexed libraries of strings (all non-transferred strings can be considered as dead or non-existent for subsequent computations).

FRAGMENTATION OF STRINGS are an important class of non-elemental (complex) optional operators which can have advantages for simulating evolution of strings in various formats of DNA shuffling. Operationally, fragmentation can be described as a formal variation of a combination of a deletion operator and a multiplication operator. One of skill will appreciate, however, that there are many other simple algorithmic operations which allow any given character string be fragmented to give a progeny of shorter strings. Fragmentation operations may be random or biased. Different ranges of fragment sizes can be predetermined. String fragments may be left in the same population with parental stings, or they may be transferred to different population. Strings fragments from various population strings can be pooled to from new populations.

CROSSOVER (RECOMBINATION)- This operator formally comprises joining a continuous part of one string with a continuous part of another string in such a way that one or two hybrid strings are formed (chimeras), where each of the chimeras contain at least two connected continuous string areas each comprising partial sequence of two different

recombining strings. The area/point where sequence characters from different parental strings, is termed the crossover/recombination area/point. Crossover operations can be combined with mutation operations affecting one or more characters of the recombined strings in a proximity of the crossover area/point of joining. When applied recursively to a population of character strings, complex chimeras comprising consequently connected partial sequences of more than two parental strings can be formed.

LIGATION is a variation of an insertion mutation operator where essentially the entire content of one string is combined with the entire content of another string in a way that the last character of one string is followed by the first character of another string. Ligation operation can be combined with mutation operation affecting one or more characters of the ligated strings in a proximity of the point of joining. Ligation can also be viewed as a means of forming chimeras.

ELITISM is a concept that provides a useful form of bias which imposes discriminating criteria for use of any of the genetic operators, and various types of positive and negative biases can be designed and implemented. The rational for the design of elitist operators is based on the concept of fitness. Fitness can be determined using string analysis tools which recognize various sequence-specific features (GC- content, frameshifts, terminations, sequence length, specific substrings, homology properties, ligand-binding and folding motifs, etc) and/or indexed correlated parameters acquired from physical selection of physical representations of character strings (enzyme activity, stability, ligand binding, etc.). It is understood that different elitism criteria can be applied separately to any of the above described genetic operators, or combinations of operators. It is also possible to use elitism, in the same evolutionary computation process, with several operators of the same type, where input/output parameters of each of the similar operators can be controlled independently (or interdependently). Different elitism criteria can be used to control changes in the string character populations caused by action of each of the individual operators.

SEQUENCE HOMOLOGY or SEQUENCE SIMILARITY is an especially important form of sequence-specific elitism useful for controlling changes in populations of character strings caused by crossover/ recombination operators in those genetic algorithms used to evolve character strings encoding polynucleotide and polypeptide sequences.

Various approaches, methods and algorithms known in the art can be used to detect homology or similarity between different character strings. Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith &

Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by even by visual inspection (*see generally*, Ausubel *et al.*, *infra*).

One example algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity

provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence (and, therefore, homologous) if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, or less than about 0.01, and or even less than about 0.001.

An additional example of a useful sequence alignment algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Doolittle, *J. Mol. Evol.* 35:351-360 (1987). The method used is similar to the method described by Higgins & Sharp, *CABIOS*5:151-153 (1989). The program can align, e.g., up to 300 sequences of a maximum length of 5,000 letters. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster can then be aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program can also be used to plot a dendogram or tree representation of clustering relationships. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison.

Thus, different types of similarity of with various levels of identity and length can be detected and recognized. For example, many homology determination methods have been designed for comparative analysis of sequences of biopolymers, for spell-checking in word processing, and for data retrieval from various databases. With an understanding of double-helix pair-wise complement interactions among 4 principal nucleobases in natural polynucleotides, models that simulate annealing of complementary homologous polynucleotide strings can also be used as a foundation of sequence-specific elitism useful for controlling crossover operators.

Homology-based elitism of crossover operators can thus be used (a) to find suitable recombination pairs of strings in a population of strings, and/or (b) to find/predetermine particularly suitable/desired areas/points of recombination over lengths of character strings selected for recombination.

Setting predetermined types and stringency of similarity/homology as a condition for crossover to occur is a form of elitism for control of formation of chimeras between representative parental character strings of various degree of homology.

RECURSIVE USE OF GENETIC OPERATORS FOR EVOLUTION OF CHARACTER STRINGS. All of the described genetic operators can be applied in a recursive mode, and specific parameters for each application occurrence can remain the same or can be systematically or randomly varied.

RANDOMNESS IN THE APPLICATION OF GENETIC OPERATORS FOR EVOLUTION OF CHARACTER STRINGS. Each genetic operator can be applied to randomly selected strings and/or to randomly selected positions over one or more string's length, with occurrence frequencies randomly selected within a range.

ARRANGEMENT OF GOS IN GAS. Order determining applications of individual GOs to product derivative libraries of character strings may be different and may depend on the composition of a particular set of individual GOs selected for practicing various formats of GAGGS. The order may be linear, cyclic, parallel, or a combination of the three and can typically be represented by a graph. Many GO arrangements can be used to simulate natural sexual and mutagenic processes for generating genetic diversity, or artificial protocols, such as single-parent or family DNA shuffling. However, the purpose of GA is not in limited to simulation of some known physical DNA manipulation methods. Its main aim is in the provision of a formal and intelligent tool, based on understanding natural and artificial evolution processes, for creation and optimization of evolutionary protocols of practical utility which may provide effective advantages over currently practiced methods.

Gene Synthesis.

The physical synthesis of genes encoded by derivative libraries of character strings, obtained by operation of genetic algorithms, is the primary means to create a physical representation of matter that is amenable to a physical assay for a desired property or to produce substrates that are further evolved in physical diversity generation procedures. Thus, one aspect of the present invention relates to the synthesis of genes with sequences selected following one or more computer shuffling procedure as set forth herein.

For GAGGS to be a time and resource effective technology, gene synthesis technology is used, typically to construct libraries of genes in a consistent manner, and in close adherence to the sequence representations produced by GA manipulations. GAGGS typically uses gene synthesis methods which allow for rapid construction of libraries of 10^4-10^9 "gene"

variations. This is typically adequate for screening/selection protocols, as larger libraries are more difficult to make and maintain and sometimes cannot be as completely sampled by a physical assay or selection method. For example, existing physical assay methods in the art (including, e.g., "life-and-death" selection methods) generally allow sampling of about 10^9 variations or less by a particular screen of a particular library, and many assays are effectively limited to sampling of 10^4-10^5 members. Thus, building several smaller libraries is a preferred method, as large libraries cannot easily be completely sampled. However, larger libraries can also be made and sampled, e.g., using high-throughput screening methods.

Gene Synthesis Technologies

There are many methods which can be used to synthesize genes with well-defined sequences. Solely for the purpose of clarity of illustration, this section focuses on one of the many possible and available types of known methods for synthesis of genes and polynucleotides.

Current art in polynucleotide synthesis is best represented by well-known and mature phosphoramidite chemistry which permits effective oligo perparation. It is possible, but somewhat impractical, to use this chemistry for routine synthesis of oligos significantly longer than 100 bp, as the quality of sequence deteriorates for longer oligos, with longer synthetic oligos generally being purified before use. Oligos of a "typical" 40-80 bp size can be obtained routinely and directly with very high purity, and without substantial sequence deterioration.

For example, oligonucleotides *e.g.,* for use in *in vitro* amplification/ gene reconstruction methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers (1981), *Tetrahedron Letts.*, 22(20):1859-1862, *e.g.,* using an automated synthesizer, as described in Needham-VanDevanter *et al.* (1984) *Nucleic Acids Res.*, 12:6159-6168. Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill. There are many commercial providers of oligo synthesis services, and thus this is a broadly accessible technology. Any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (http://www.genco.com), ExpressGen Inc. (www.expressgen.com), Operon Technologies Inc. (Alameda, CA) and many others. Similarly, peptides and antibodies can be custom ordered from any of a variety of sources,

such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (http://www.htibio.com), BMA Biomedicals Ltd (U.K.), Bio·Synthesis, Inc., and many others.

A relevant demonstration of total gene synthesis from small fragments which is readily amenable to optimization, parallelism and high throughout is set forth in Dillon and Rosen (Biotechniques, 1990, 9(3)298-300). Simple and rapid PCR-based assembly process of a gene from a set of partially overlapping single-strand oligonucleotides, without use of ligase, can be performed. Several groups have also described successful applications of variations the same PCR-based gene assembly approach to the synthesis of various genes of increasing size, thus demonstrating its general applicability and its combinatorial nature for synthesis of libraries of mutated genes. Useful references include Sandhu, et al. (Biotechniques, 1992, 12(1)15-16), (220 bp gene from 3 oligos of 77-86 bp); Prodromou and Pearl (Protein Engineering, 1992, 5(8)827-829 (522 bp gene, from 10 oligos of 54-86 bp); Chen et al, 1994 (JACS, 1194(11):8799-8800), (779 bp gene); Hayashi et al, 1994 (Biotechniques, 1994, 17:310-314) and others.

More recently Stemmer et al (Gene, 1995, 164:49-53) showed that, e.g., PCR-based assembly methods are effectively useful to build larger genes of up to at least 2.7 kb from dozens or even hundreds of synthetic 40 bp oligos. These authors also demonstrated that, from four basic steps comprising PCR-based gene synthesis protocols (oligo synthesis, gene assembly, gene amplification, and, optionally, cloning) the gene amplification step can be omitted, if a 'circular' assembly PCR is used.

A number of the publications of the inventors and their co-workers, as well as other investigators in the art also describe techniques which facilitate DNA shuffling, e.g., by providing for reassembly of genes from small fragments, or even oligonucleotides. One aspect of the present invention is the ability to use family shuffling oligonucleotides and cross over oligonucleotides as recombination templates/intermediates in various DNA shuffling methods.

Indeed, a number of the publications by the inventors and their co-workers, as well as other investigators in the art also describe techniques which facilitate reassembly of genes from small fragments, including oligonucleotides. In addition to the publications noted above, Stemmer et al. (1998) U.S. Pat. No. 5,834,252 END COMPLEMENTARY POLYMERASE REACTION describe processes for amplifying and detecting a target sequence (e.g., in a mixture of nucleic acids), as well as for assembling large polynucleotides from fragments. Crameri et al. (1998) Nature 391: 288-291 provides basic methodologies for gene reassembly, as does Crameri et al. (1998) Bio techniques 18(2): 194-196.

More recently, a number of gene reassembly protocols which simultaneously recombine and reconstruct genes have been described in several applications of the inventors and their co-workers, such as "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed February 5, 1999 (USSN 60/118,813) and filed June 24, 1999 (USSN 60/141,049) and filed September 28, 1999 (USSN 09/408,392) and "USE OF CODON-BASED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., filed September 28, 1999 (USSN 09/408,393). In these embodiments, synthetic recombination methods are used, in which oligonucleotides corresponding to different homologues are synthesized and reassembled in PCR or ligation reactions which include oligonucleotides which correspond to more than one parental nucleic acid, thereby generating new recombined nucleic acids.

One advantage of oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even to recombine non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of fragmented nucleic acids is recombined, e.g., with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination. Such oligonucleotide sets are selected in silico according to the methods herein.

When recombining homologous nucleic acids, sets of overlapping family gene shuffling oligonucleotides (which are derived by comparison of homologous nucleic acids and synthesis of oligonucleotide fragment sets, which correspond to regions of similarity and regions of diversity derived from the comparison) are hybridized and elongated (e.g., by reassembly PCR), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. Typically, the set of overlapping family shuffling gene oligonucleotides include a plurality of oligonucleotide member types which have consensus region subsequences derived from a plurality of homologous target nucleic acids.

Typically, family gene shuffling oligonucleotide are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are

synthesized (serially or in parallel) which correspond to at least one region of sequence diversity. Further details regarding family shuffling is found in USSN 09/408,392, cited above.

Sets of fragments, or subsets of fragments used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (e.g., with a DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments can be used in conjunction with family gene shuffling oligonucleotides, e.g., in one or more recombination reaction to produce recombinant nucleic acids.

Gene assembly by PCR from single-strand complementary overlapping synthetic oligos is a method of choice for practicing in GAGGS. Optimization of this method can be performed e.g., including varying oligo length, the number of oligos in the recombination reaction, the degree of oligonucleotide overlap, levels and nature of sequence degeneracy, specific reaction conditions and particular polymerase enzymes used in the reassembly, and in controlling the stringency of gene assembly to decrease or increase the number of sequence deviations during gene synthesis.

The method can also be practiced in a parallel mode where each of the individual library members, including a plurality of the genes intended for subsequent physical screening, are synthesized in spatially separated vessels, or arrays of vessels, or in a poolwise fashion, where all, or part, of the desired plurality of genes are synthesized in a single vessel. Many other synthesis methods for making synthetic nucleotides are also known, and specific advantages of use of one vs. another for practicing GAGGS may be readily determined by one skilled in the art.

Sequence Deconvolution.

Sequence deconvolution is performed on those variants of polynucleotides which are found to have desired properties, in order to confirm changes in corresponding character strings (i.e., corresponding to physical sequences for biopolymers) yielding desired changes in the relevant composition of matter (e.g., a polynucleotide, polypeptide, or the like).

Sequencing and other standard recombinant techniques useful for the present invention, including for sequence deconvolution are found, e.g., in Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc.,

San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1999) ("Ausubel")). In addition to sequencing GAGGS products, unique restriction sites can also be used for detecting particular sequences. Sufficient information to guide one of skill through restriction enzyme digestion is also found in Sambrook, Berger and Ausubel, *id.*

Methods of transducing cells, including plant and animal cells, with GAGGS generated nucleic acids, e.g., for cloning and sequencing and/or for expression and selection of encoded molecules are generally available, as are methods of expressing proteins encoded by such nucleic acids. In addition to Berger, Ausubel and Sambrook, useful general references for culture of animal cells include Freshney (Culture of Animal Cells, a Manual of Basic Technique, third edition Wiley- Liss, New York (1994)) and the references cited therein, Humason (Animal Tissue Techniques, fourth edition W.H. Freeman and Company (1979)) and Ricciardelli, et al., In Vitro Cell Dev. Biol. 25:1016-1024 (1989). References for plant cell cloning, culture and regeneration include Payne et al. (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York, NY (Payne); and Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) (Gamborg). A variety of Cell culture media are described in Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Additional information for plant cell culture is found in available commercial literature such as the Life Science Research Cell Culture Catalogue (1998) from Sigma- Aldrich, Inc (St Louis, MO) (Sigma-LSRCCC) and, e.g., the Plant Culture Catalogue and supplement (1997) also from Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

In vitro amplification methods can also be used to amplify and/or sequence GAGGS generated nucleic acids, e.g., for cloning, and selection. Examples of techniques sufficient to direct persons of skill through typical *in vitro* amplification and sequencing methods, including the polymerase chain reaction (PCR) the ligase chain reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques (*e.g.*, NASBA) are found in Berger, Sambrook, and Ausubel, *id.*, as well as in Mullis *et al.*, (1987) U.S. Patent No. 4,683,202; PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic

Press Inc. San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1, 1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; Kwoh *et al.* (1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli *et al.* (1990) Proc. Natl. Acad. Sci. USA 87, 1874; Lomell *et al.* (1989) J. Clin. Chem 35, 1826; Landegren *et al.*, (1988) Science 241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene 4, 560; Barringer *et al.* (1990) Gene 89, 117, and Sooknanan and Malek (1995) Biotechnology 13: 563-564. Improved methods of cloning *in vitro* amplified nucleic acids are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng *et al.* (1994) Nature 369: 684-685 and the references therein, in which PCR amplicons of up to 40kb are generated. PCR reassembly techniques are discussed *supra.* One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. *See*, Ausbel, Sambrook and Berger, *all supra.*

If gene synthesis is essentially error-free (stringent), and, e.g., performed in a parallel mode where each of the individual members of library was originally synthesized in spatially separated areas or containers (vessels), then deconvolution is performed by referencing the positional encoding index for each intended sequence of all library members. If the synthesis is performed in a poolwise fashion (or if library members were pooled during selection), then one of the many known polynucleotide sequencing techniques is used.

Recursive GAGGS processes.

The recursive nature of directed evolution methods, aimed at stepwise/roundwise improvement of desired properties of polynucleotides and polypeptides, is well understood. In directed evolution (DE) by GAGGS, one or more deconvoluted character string(s), encoding sequences of those variants displaying certain changes in level of desired properties (where the level is arbitrarily defined by increase/decrease/ratios between measures of several properties), can be used to comprise a new library of character strings for a new round of GAGGS. Recursive GAGGS, unlike typical DNA shuffling, does not use physical manipulation of the polynucleotides in order to produce subsequent generations of gene diversity. Instead, GAGGS simply uses sequence information describing acquired beneficial changes as a foundation for generating additional changes leading to subsequent changes (improvements) in the desired properties of molecules encoded by the character strings. Recursive GAGGS can be performed until strings of characters are evolved to the point when the encoded polynucleotides and polypeptides attain arbitrarily set levels of desired

characteristics or until further changes in characteristics cannot be obtained (e.g., enzyme turnover reached theoretical diffusion rate limit under conditions of a physical assay). Genetic algorithm parameters, gene synthesis methods and schemes, as well as physical assays and sequence deconvolution methods can vary in each of the different rounds/cycles of directed evolution by recursive GAGGS.).

One particular advantage of this approach is that an initially random or pseudo random approach to library generation can become progressively more directed as information on activity levels becomes available. For example, any heuristic learning approach or neural network approach gradually becomes more efficient at selecting "correct" (active) sequences. A variety of such approaches are set forth below, including principle component analysis, use of negative data, data parameterization and the like.

Integration of GAGGS, DNA Shuffling and other Directed Evolution Technologies.

GAGGS constitutes a self-sufficient and independent technology which can be practiced regardless of DNA shuffling or any other available directed evolution methods. However, one or more rounds of GAGGS can be, and often is, practiced in combination with physical shuffling of nucleic acids, and/or in combination with site directed mutagenesis, or error-prone PCR (e.g. as alternating cycles of a directed evolution process) or other diversity generation methods. GAGGS-generated libraries of polynucleotides can be subjected to nucleic acid shuffling, and polynucleotides found to have desired characteristics following rounds of in silico and/or physical shuffling can be selected and sequenced to provide character strings to evaluate GAGGS processes or to form character strings for further GAGGS operations. Thus, GAGGS can be performed as a stand-alone technology, or can be followed by shuffling, mutagenesis, random priming PCR, etc.

Where the methods of the invention entail performing physical recombination ("shuffling") and screening or selection to evolve individual genes, whole plasmids, viruses, multigene clusters, or even whole genomes, the techniques of the inventors and their co-workers are particularly useful. For example, reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest which are generated by performing a GO on a character string (e.g., followed by synthesis of corresponding oligonucleotides, and gene generation/regeneration, e.g., by assembly PCR).

The following publications describe a variety of recursive recombination procedures and/or related diversity generation methods which can be practiced in conjunction

with the in silico processes of the invention: Stemmer, et al., (1999) "Molecular breeding of viruses for targeting and other clinical properties. Tumor Targeting" 4:1-4; Nesset al. (1999) "DNA Shuffling of subgenomic sequences of subtilisin" Nature Biotechnology 17:893-896; Chang et al. (1999) "Evolution of a cytokine using DNA family shuffling" Nature Biotechnology 17:793-797; Minshull and Stemmer (1999) "Protein evolution by molecular breeding" Current Opinion in Chemical Biology 3:284-290; Christians et al. (1999) "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling" Nature Biotechnology 17:259-264; Crameriet 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; Zhang et al. (1997) "Directed evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening" Proceedings of the National Academy of Sciences, U.S.A. 94:4504-4509; Patten et al. (1997) "Applications of DNA Shuffling to Pharmaceuticals and Vaccines" Current Opinion in Biotechnology 8:724-733; Crameri et al. (1996) "Construction and evolution of antibody-phage libraries by DNA shuffling" Nature Medicine 2:100-103; Crameri et al. (1996) "Improved green fluorescent protein by molecular evolution using DNA shuffling" Nature Biotechnology 14:315-319; Gates et al. (1996) "Affinity selective isolation of ligands from peptide libraries through display on a lac repressor 'headpiece dimer'" Journal of Molecular Biology 255:373-386; Stemmer (1996) "Sexual PCR and Assembly PCR" In: The Encyclopedia of Molecular Biology. VCH Publishers, New York. pp.447-457; Crameri and Stemmer (1995) "Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wildtype cassettes" BioTechniques 18:194-195; Stemmer et al., (1995) "Single-step assembly of a gene and entire plasmid form large numbers of oligodeoxyribonucleotides" Gene, 164:49-53; Stemmer (1995) "The Evolution of Molecular Computation" Science 270: 1510; Stemmer (1995) "Searching Sequence Space" Bio/Technology 13:549-553; Stemmer (1994) "Rapid evolution of a protein in vitro by DNA shuffling" Nature 370:389-391; and Stemmer (1994) "DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution." Proceedings of the National Academy of Sciences, U.S.A. 91:10747-10751.

Additional details regarding DNA shuffling methods are found in U.S. Patents by the inventors and their co-workers, including: United States Patent 5,605,793 to Stemmer (February 25, 1997), "METHODS FOR IN VITRO RECOMBINATION;" United States Patent 5,811,238 to Stemmer et al. (September 22, 1998) "METHODS FOR GENERATING

POLYNUCLEOTIDES HAVING DESIRED CHARACTERISTICS BY ITERATIVE SELECTION AND RECOMBINATION;" United States Patent 5,830,721 to Stemmer et al. (November 3, 1998), "DNA MUTAGENESIS BY RANDOM FRAGMENTATION AND REASSEMBLY;" United States Patent 5,834,252 to Stemmer, et al. (November 10, 1998) "END-COMPLEMENTARY POLYMERASE REACTION," and United States Patent 5,837,458 to Minshull, et al. (November 17, 1998), "METHODS AND COMPOSITIONS FOR CELLULAR AND METABOLIC ENGINEERING."

In addition, details and formats for nucleic acid shuffling are found in a variety of PCT and foreign patent application publications, including: Stemmer and Crameri, "DNA MUTAGENESIS BY RANDOM FRAGMENTATION AND REASEMBLY" WO 95/22625; Stemmer and Lipschutz "END COMPLEMENTARY POLYMERASE CHAIN REACTION" WO 96/33207; Stemmer and Crameri "METHODS FOR GENERATING POLYNUCLEOTIDES HAVING DESIRED CHARACTERISTICS BY ITERATIVE SELECTION AND RECOMBINATION" WO 97/0078; Minshul and Stemmer, "METHODS AND COMPOSITIONS FOR CELLULAR AND METABOLIC ENGINEERING" WO 97/35966; Punnonen et al. "TARGETING OF GENETIC VACCINE VECTORS" WO 99/41402; Punnonen et al. "ANTIGEN LIBRARY IMMUNIZATION" WO 99/41383; Punnonen et al. "GENETIC VACCINE VECTOR ENGINEERING" WO 99/41369; Punnonen et al. OPTIMIZATION OF IMMUNOMODULATORY PROPERTIES OF GENETIC VACCINES WO 9941368; Stemmer and Crameri, "DNA MUTAGENESIS BY RANDOM FRAGMENTATION AND REASSEMBLY" EP 0934999; Stemmer "EVOLVING CELLULAR DNA UPTAKE BY RECURSIVE SEQUENCE RECOMBINATION" EP 0932670; Stemmer et al., "MODIFICATION OF VIRUS TROPISM AND HOST RANGE BY VIRAL GENOME SHUFFLING" WO 9923107; Apt et al., "HUMAN PAPILLOMAVIRUS VECTORS" WO 9921979; Del Cardayre et al. "EVOLUTION OF WHOLE CELLS AND ORGANISMS BY RECURSIVE SEQUENCE RECOMBINATION" WO 9831837; Patten and Stemmer, "METHODS AND COMPOSITIONS FOR POLYPEPTIDE ENGINEERING" WO 9827230; Stemmer et al., and "METHODS FOR OPTIMIZATION OF GENE THERAPY BY RECURSIVE SEQUENCE SHUFFLING AND SELECTION" WO9813487.

Certain U.S. Applications provide additional details regarding DNA shuffling and related techniques, including "SHUFFLING OF CODON ALTERED GENES" by Patten et al. filed September 29, 1998, (USSN 60/102,362), January 29, 1999 (USSN 60/117,729), and September 28, 1999, USSN PCT/US99/22588; "EVOLUTION OF WHOLE CELLS AND

ORGANISMS BY RECURSIVE SEQUENCE RECOMBINATION", by del Cardyre et al. filed July 15, 1999 (USSN 09/354,922); "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed February 5, 1999 (USSN 60/118,813) and filed June 24, 1999 (USSN 60/141,049) and filed September 28, 1999 (USSN 09/408,392), and "USE OF CODON-BASED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., filed September 28, 1999 (USSN 09/408,393).

As review of the foregoing publications, patents, published applications and U.S. patent applications reveals, shuffling (or "recursive recombination") of nucleic acids to provide new nucleic acids with desired properties can be carried out by a number of established methods. Any of these methods are integrated with those of the present invention by incorporating nucleic acids corresponding to character strings produced by performing one or more GO on one or more selected parental character string. Any of these methods can be adapted to the present invention to evolve GAGGS produced nucleic acids as discussed herein to produce new nucleic acids with improved properties. Both the methods of making such nucleic acids and the nucleic acids produced by these methods are a feature of the invention.

In brief, at least 5 different general classes of recombination methods can be performed (separately or in combination) in accordance with the present invention. First, nucleic acids such as those produced by synthesis of sets of nucleic acids corresponding to character strings produced by GO manipulation of character strings, or available homologues of such sets, or both, can be recombined in vitro by any of a variety of techniques discussed in the references above, including e.g., DNAse digestion of nucleic acids to be recombined followed by ligation and/or PCR reassembly of the nucleic acids. Second, sets of nucleic acids corresponding to character strings produced by GO manipulation of character strings, and/or available homologues of such sets, can be recursively recombined in vivo, e.g., by allowing recombination to occur between the nucleic acids while in cells. Third, whole cell genome recombination methods can be used in which whole genomes of cells are recombined, optionally including spiking of the genomic recombination mixtures with desired library components such as with sets of nucleic acids corresponding to character strings produced by GO manipulation of character strings, or available homologues of such sets. Fourth, synthetic recombination methods can be used, in which oligonucleotides corresponding to different homologues are synthesized and reassembled in PCR or ligation reactions which include oligonucleotides which correspond to more than one parental nucleic acid, thereby generating new recombined nucleic acids. Oligonucleotides can be made by standard nucleotide addition

methods, or can be made by tri-nucleotide synthetic approaches. Fifth, purely in silico methods of recombination can be effected in which GOs are used in a computer to recombine sequence strings which correspond to nucleic acid or proteins homologues. The resulting recombined sequence strings are optionally converted into nucleic acids by synthesis of nucleic acids which correspond to the recombined sequences, e.g., in concert with oligonucleotide synthesis/ gene reassembly techniques. Any of the preceding general recombination formats, separately or together, can be practiced in a reiterative fashion to generate a diverse set of recombinant nucleic acids.

The above references in conjunction with the present disclosure provide these and other basic recombination formats as well as many modifications of these formats. Regardless of the format which is used, the nucleic acids of the invention can be recombined (with each other or with related (or even unrelated) nucleic acids to produce a diverse set of recombinant nucleic acids, including homologous nucleic acids.

Other diversity generating approaches can also be used to modify character strings or nucleic acids. Additional diversity can be introduced into input or output nucleic acids by methods which result in the alteration of individual nucleotides or groups of contiguous or non-contiguous nucleotides, i.e., mutagenesis methods. Mutagenesis methods include, for example, recombination (PCT/US98/05223; Publ. No. WO98/42727); oligonucleotide-directed mutagenesis (for review *see*, Smith, Ann. Rev.Genet. 19: 423-462 (1985); Botstein and Shortle, Science 229: 1193-1201 (1985); Carter, Biochem. J. 237: 1-7 (1986); Kunkel, "The efficiency of oligonucleotide directed mutagenesis" in Nucleic acids & Molecular Biology, Eckstein and Lilley, eds., Springer Verlag, Berlin (1987)). Included among these methods are oligonucleotide-directed mutagenesis (Zoller and Smith, Nucl. Acids Res. 10: 6487-6500 (1982), Methods in Enzymol. 100: 468-500 (1983), and Methods in Enzymol. 154: 329-350 (1987)) phosphothioate-modified DNA mutagenesis (Taylor et al., Nucl. Acids Res. 13: 8749-8764 (1985); Taylor et al., Nucl. Acids Res. 13: 8765-8787 (1985); Nakamaye and Eckstein, Nucl. Acids Res. 14: 9679-9698 (1986); Sayers et al., Nucl. Acids Res. 16:791-802 (1988); Sayers et al., Nucl. Acids Res. 16: 803-814 (1988)), mutagenesis using uracil-containing templates (Kunkel, Proc. Nat'l. Acad. Sci. USA 82: 488-492 (1985) and Kunkel et al., Methods in Enzymol. 154:367-382)); mutagenesis using gapped duplex DNA (Kramer et al., Nucl. Acids Res. 12: 9441-9456 (1984); Kramer and Fritz, Methods in Enzymol. 154:350-367 (1987); Kramer et al., Nucl. Acids Res. 16: 7207 (1988)); and Fritz et al., Nucl. Acids Res. 16: 6987-6999 (1988)). Additional suitable methods include point

mismatch repair (Kramer et al., Cell 38: 879-887 (1984)), mutagenesis using repair-deficient host strains (Carter et al., Nucl. Acids Res. 13: 4431-4443 (1985); Carter, Methods in Enzymol. 154: 382-403 (1987)), deletion mutagenesis (Eghtedarzadeh and Henikoff, Nucl. Acids Res. 14: 5115 (1986)), restriction-selection and restriction-purification (Wells et al., Phil. Trans. R. Soc. Lond. A 317: 415-423 (1986)), mutagenesis by total gene synthesis (Nambiar et al., Science 223: 1299-1301 (1984); Sakamar and Khorana, Nucl. Acids Res. 14: 6361-6372 (1988); Wells et al., Gene 34:315-323 (1985); and Grundström et al., Nucl. Acids Res. 13: 3305-3316 (1985). Kits for mutagenesis are commercially available (e.g., Bio-Rad, Amersham International, Anglian Biotechnology).

Other diversity generation procedures are proposed in U.S. Patent No. 5,756,316; U.S. Patent No. 5,965,408; Ostermeier et al. (1999) "A combinatorial approach to hybrid enzymes independent of DNA homology" Nature Biotech 17:1205; U.S. Patent No. 5,783,431; U.S. Patent No. 5,824,485; U.S. Patent 5,958,672; Jirholt et al. (1998) "Exploiting sequence space: shuffling in vivo formed complementarity determining regions into a master framework" Gene 215: 471; U.S. Patent No. 5,939,250; WO 99/10539; WO 98/58085; WO 99/10539 and others. These diversity generating methods can be combined with each other or with shuffling reactions or in silico operations, in any combination selected by the user, to produce nucleic acid diversity, which may be screened for using any available screening method.

Following recombination or other diversification reactions, any nucleic acids which are produced can be selected for a desired activity. In the context of the present invention, this can include testing for and identifying any detectable or assayable activity, by any relevant assay in the art. A variety of related (or even unrelated) properties can be assayed for, using any available assay.

Accordingly, a recombinant nucleic acid produced by recursively recombining one or more polynucleotide of the invention (produced by GAGGS methods) with one or more additional nucleic acid forms a part of the invention. The one or more additional nucleic acid may include another polynucleotide of the invention; optionally, alternatively, or in addition, the one or more additional nucleic acid can include, e.g., a nucleic acid encoding a naturally-occurring sequence or a subsequence, or any homologous sequence or subsequence.

The recombining steps can be performed in vivo, in vitro, or in silico as described in more detail in the references above and herein. Also included in the invention is a cell containing any resulting recombinant nucleic acid, nucleic acid libraries produced by

recursive recombination of the nucleic acids set forth herein, and populations of cells, vectors, viruses, plasmids or the like comprising the library or comprising any recombinant nucleic acid resulting from recombination (or recursive recombination) of a nucleic acid as set forth herein with another such nucleic acid, or an additional nucleic acid. Corresponding sequence strings in a database present in a computer system or computer readable medium are a feature of the invention.

By way of example, a typical physical recombination procedure starts with at least two substrates that generally show at least some identity to each other (*i.e.*, at least about 30%, 50%, 70%, 80% or 90% or more sequence identity), but differ from each other at certain positions (however, in purely in silico or cross-over oligonucleotide mediated formats, nucleic acids can show little or no homology). For example, two or more nucleic acids can be recombined herein. The differences between the nucleic acids can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in about 1-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials differ from each other in at least two nucleotide positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position. Of course, even if only one initial character string is provided, any GO herein can be used to modify the nucleic acid to produce a diverse array of nucleic acids that can be screened for an activity of interest.

In physical shuffling procedures, starting DNA segments can be natural variants of each other, for example, allelic or species variants. More typically, they are derived from one or more homologous nucleic acid sequence. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness. The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example, one segment can be a subsequence of another. The

segments can be present as part of larger molecules, such as vectors, or can be in isolated form. In one option, the nucleic acids of interest are derived from DE by GAGGS.

CODON-VARIED OLIGONUCLEOTIDE METHODS

Codon-varied oligonucleotides are oligonucleotides, similar in sequence but with one or more base variations, where the variations correspond to at least one encoded amino acid difference. They can be synthesized utilizing tri-nucleotide, i.e., codon-based phosphoramidite coupling chemistry, in which tri-nucleotide phosphoramidites representing codons for all 20 amino acids are used to introduce entire codons into oligonucleotide sequences synthesized by this solid-phase technique. Preferably, all of the oligonucleotides of a selected length (e.g., about 20, 30, 40, 50, 60, 70, 80, 90, or 100 or more nucleotides) which incorporate the chosen nucleic acid sequences are synthesized. In the present invention, codon-varied oligonucleotide sequences can be based upon sequences from a selected set of nucleic acids, generated by any of the approaches noted herein. Further details regarding tri-nucleotide synthesis are found in USSN 09/408,393 "USE OF CODON VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch, et al., filed 09/28/1999.

Oligonucleotides can be made by standard nucleotide addition methods, or can be made by tri-nucleotide synthetic approaches. An advantage of selecting changes which correspond to encoded amino acid differences is that the modification of triplets of codons results in fewer frameshifts (and, therefore, likely fewer inactive library members). Also, synthesis which focuses on codon modification, rather than simply on base variation, reduces the total number of oligos which are needed for a synthesis protocol.

OLIGO SETS

In general, sets of oligos can be combined for assembly in many different formats and different combinations schemes to effect correlation with genetic events and operators at the physical level.

As noted, overlapping sets of oligonucleotides can be synthesized and then hybridized and elongated to form full-length nucleic acids. A full length nucleic acid is any nucleic acid desired by an investigator which is longer than the oligos which are used in the gene reconstruction methods. This can correspond to any percentage of a naturally occurring full length sequence, e.g., 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% or more of the corresponding natural sequence.

Oligo sets often have at least about 5, sometimes about 10, often about 15, generally about 20, or more, nucleotide overlap sequences to facilitate gene reconstruction. Oligo sets are optionally simplified for gene reconstruction purposes where regions of fortuitous overlap are present, i.e., where repetitive sequence elements are present or designed into a gene sequence to be synthesized. Lengths of oligos in a set can be the same or different, as can the regions of sequence overlap. To facilitate hybridization and elongation (e.g., during cycles of PCR), overlap regions are optionally designed with similar melting temperatures.

Parental sequences can be gridded (conceptually or physically) and the common sequences used to select common sequence oligos, thereby combining oligo members into one or more sets to reduce the number of oligos required for making full-length nucleic acids. Similarly, oligonucleotides with some sequence similarity can be generated by pooled and/or split synthesis where pools of oligos under synthesis are split into different pools during the addition of heterologous bases, optionally followed by rejoined synthesis steps (pooling) at subsequent stages where the same additions to the oligos are required. In Oligo shuffling formats, heterologous oligos corresponding to many different parents can be split and rejoined during synthesis. In simple degenerate synthetic approaches, more than one nucleobase can be added during single synthetic steps to produce two or more variations in sequence in two or more resulting oligonucleotides. The relative percentage of nucleobase addition can be controlled to bias synthesis towards one or more parental sequence. Similarly, partial generacy can be practiced to prevent the insertion of stop codons during degenerated oligonucleotide synthesis.

Oligos which correspond to similar subsequences from different parents can be the same length or different, depending on the subsequences. Thus, in split and pooled formats, some oligos are optionally not elongated during every synthetic step (to avoid frame-shifting, some oligos are not elongated for the steps corresponding to one or more codon).

When constructing oligos, crossover oligos can be constructed at one or more point of difference between two or more parental sequences (a base change or other difference is a genetic locus which can be treated as a point for a crossover event). The crossover oligos have a region of sequence identity to a first parental sequence, followed by a region of identity to a second parental sequence, with crossover point occurring at the locus. For example, every natural mutation can be a cross over point.

Another way of biasing sequence recombination is to spike a mixture of oligonucleotides with fragments of one or more parental nucleic acid (if more than one parental

nucleic acid is fragmented, the resulting segments can be spiked into a recombination mixture at different frequencies to bias recombination outcomes towards one or more parent). Recombination events can also be engineered simply by omitting one or more oligonucleotide corresponding to one or more parent from a recombination mixture.

In addition to the use of families of related oligonucleotides, diversity can be modulated by the addition of selected, pseudo-random or random oligos to elongation mixture, which can be used to bias the resulting full-length sequences. Similarly, mutagenic or non-mutagenic conditions can be selected for PCR elongation, resulting in more or less diverse libraries of full length nucleic acids.

In addition to mixing oligo sets which correspond to different parents in the elongation mixture, oligo sets which correspond to just one parent can be elongated to reconstruct that parent. In either case, any resulting full-length sequence can be fragmented and recombined, as in the DNA shuffling methods noted in the references cited herein.

Many other oligonucleotide sets and synthetic variations which can be correlated to genetic events and operators at the physical level are found in "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed February 5, 1999 (USSN 60/118,813) and filed June 24, 1999 (USSN 60/141,049) and filed September 28, 1999 (USSN 09/408,392) and "USE OF CODON-BASED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., filed September 28, 1999 (USSN 09/408,393).

TARGETS FOR CODON MODIFICATION AND SHUFFLING

Essentially any nucleic acid can be shuffled using the GAGGS methods herein. No attempt is made herein to identify the hundreds of thousands of known nucleic acids. Common sequence repositories for known proteins include GenBank EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

One class of preferred targets for GAGGS methods includes nucleic acids encoding therapeutic proteins such as erythropoietin (EPO), insulin, peptide hormones such as human growth hormone; growth factors and cytokines such as epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1δ, MCP-1, epidermal growth factor, fibroblast growth factor, hepatocyte growth factor, insulin-like growth factor, the interferons, the interleukins, keratinocyte growth factor, leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF etc. Many of these

proteins and their corresponding coding nucleic acids are commercially available (*See*, e.g., the Sigma BioSciences 1997 catalogue and price list), and, in any case, the corresponding genes are well-known.

Another class of preferred targets for GAGGS are transcription and expression activators. Example transcriptional and expression activators include genes and proteins that modulate cell growth, differentiation, regulation, or the like. Expression and transcriptional activators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, and animals, including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, e.g., by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and degrading RNA. Expression activators include cytokines, inflammatory molecules, growth factors, their receptors, and oncogene products, e.g., interleukins (e.g., IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, e.g., Mos, Ras, Raf, and Met; and transcriptional activators and suppressors, e.g., p53, Tat, Fos, Myc, Jun, Myb, Rel, and steroid hormone receptors such as those for estrogen, progesterone, testosterone, aldosterone, the LDL receptor ligand and corticosterone.

Similarly, proteins from infectious organisms for possible vaccine applications, described in more detail below, including infectious fungi, e.g., *Aspergillus*, *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (e.g., *aureus*), *Streptococci* (e.g., *pneumoniae*), *Clostridia* (e.g., *perfringens*), *Neisseria* (e.g., *gonorrhoea*), *Enterobacteriaceae* (e.g., *coli*), *Helicobacter* (e.g., *pylori*), *Vibrio* (e.g., *cholerae*), *Capylobacter* (e.g., *jejuni*), *Pseudomonas* (e.g., *aeruginosa*), *Haemophilus* (e.g., *influenzae*), *Bordetella* (e.g., *pertussis*), *Mycoplasma* (e.g., *pneumoniae*), *Ureaplasma* (e.g., *urealyticum*), *Legionella* (e.g., *pneumophila*), *Spirochetes* (e.g., *Treponema*, *Leptospira*, and *Borrelia*), *Mycobacteria* (e.g., *tuberculosis*, *smegmatis*), *Actinomyces* (e.g., *israelii*), *Nocardia* (e.g., *asteroides*), *Chlamydia* (e.g., *trachomatis*), *Rickettsia*, *Coxiella*, *Ehrilichia*, *Rochalimaea*, *Brucella*, *Yersinia*, *Fracisella*, and *Pasteurella*; protozoa such as sporozoa (e.g., *Plasmodia*), *rhizopods* (e.g., *Entamoeba*) and flagellates (*Trypanosoma*, *Leishmania*, *Trichomonas*, *Giardia*, etc.); viruses

such as (+) RNA viruses (examples include Poxviruses e.g., *vaccinia*; Picornaviruses, e.g. *polio*; Togaviruses, e.g., *rubella*; Flaviviruses, e.g., HCV; and Coronaviruses), (-) RNA viruses (examples include Rhabdoviruses, e.g., VSV; Paramyxovimses, e.g., RSV; Orthomyxovimses, e.g., influenza; Bunyaviruses; and Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, i.e., Retroviruses, e.g., especially HIV and HTLV, and certain DNA to RNA viruses such as Hepatitis B virus.

Other nucleic acids encoding proteins relevant to non-medical uses, such as inhibitors of transcription or toxins of crop pests e.g., insects, fungi, weed plants, and the like, are also preferred targets for GAGGS. Industrially important enzymes such as monooxygenases, proteases, nucleases, and lipases are also preferred targets. As an example, subtilisin can be evolved by shuffling selected forms of the gene for subtilisin (von der Osten et al., J. Biotechnol. 28:55-68 (1993) provide a subtilisin coding nucleic acid). Proteins which aid in folding such as the chaperonins are also preferred.

Preferred known genes suitable for codon alteration and shuffling also include the following: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, Apolipoprotein, Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Atrial peptides, C-X-C chemokines (e.g., T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4, SDF-1, PF4, MIG), Calcitonin, CC chemokines (e.g., Monocyte chemoattractant protein-1, Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309, R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, Collagen, Colony stimulating factor (CSF), Complement factor 5a, Complement inhibitor, Complement receptor 1, Factor IX, Factor VII, Factor VIII, Factor X, Fibrinogen, Fibronectin, Glucocerebrosidase, Gonadotropin, Hedgehog proteins (e.g., Sonic, Indian, Desert), Hemoglobin (for blood substitute; for radiosensitization), Hirudin, Human serum albumin, Lactoferrin, Luciferase, Neurturin, Neutrophil inhibitory factor (NIF), Osteogenic protein, Parathyroid hormone, Protein A, Protein G, Relaxin, Renin, Salmon calcitonin, Salmon growth hormone, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, i.e., Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Toxic shock syndrome toxin (TSST-1), Exfoliating toxins A and B, Pyrogenic exotoxins A, B, and C, and M. arthritides mitogen, Superoxide dismutase, Thymosin

alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha) and Urokinase.

Other preferred genes for shuffling include p450s (these enzymes represent a very diverse set of natural diversity and catalyze many important reactions); *see, e.g.,* Ortiz de Montellano (ed.) (1995) Cytochrome P450 Structure Mechanism and Biochemistry, Second Edition Plenum Press (New York and London) and the references cited therein For an introduction to cytochrome P450. Other monooxygenases, as well as dioxygenases, acyl transferases (cis-diol), halogenated hydrocarbon dehalogenases, methyl transferases, terpene synthetases, and the like, can be shuffled.

THE USES OF CONSENSUS GENES IN DIRECTED EVOLUTION, INCLUDING "DIPLOMACY"

One of the factors involved in standard family shuffling of parental genes is the extent of identity of the genes to be physically recombined. Genes of limited identity are difficult to recombine without cross-over oligonucleotides, and often result in shuffled libraries having unacceptable knockout rates, or no chimera formation, no activity, no functional library, etc. In one aspect, the present invention overcomes this difficulty by providing for in silico design of a "diplomat" sequence which has an intermediate level of homology to each of the sequences to be recombined, thereby facilitating cross-over events between the sequences and facilitating chimera formation. This diplomat sequence can be a character string produced by any of a variety of GO to establish intermediate sequence similarity in the diplomat sequence as compared to the sequences to be recombined, including by alignment of the sequences to select a consensus sequence, codon modification to optimize similarity between diverse nucleic acids, or the like.

As noted, one way in which to design a diplomat sequence is simply to select a consensus sequence, e.g., using any of the approaches herein. The consensus sequence is generated by comparison and lining-up/pile-up of a family of genes (DNA consensus), or of amino acid sequence line-up/pile-up (aa consensus). In the latter case, the amino acid consensus sequence are optionally back-translated using a desirable codon bias to further enhance homology, or to enhance host organism for expression, or to select for alternate codon usages in order to enable access to alternative sets of amino acid codons. One can also use different subsets of gene families to generate consensus sequences.

Furthermore, the consensus sequence itself may encode an improved enzyme. This has been observed elsewhere (e.g. presentation at International Conference "Enzyme

Opportunities on the Next Millenium", Chicago, IL, May 5-7, by Dr. Luis Pasamontes, Roche Vitamins, Inc., on "Development of Heat Stable Phytase"- a consensus phytase had an increase of 16 degrees C in thermostability). Another example of a consensus protein having improved properties is consensus Interferon (IFN-con1).

Accordingly, separately or in conjunction with any of the techniques herein, diplomat sequences can be designed using selected GO criteria and, optionally, physically synthesized and shuffled using any of the techniques herein.

EXAMPLE PROCESSING STEPS FOR REVERSE TRANSLATION AND OLIGO DESIGN

Automatic processing steps (e.g., performed in a digital system as described herein) that perform the following functions facilitate selection of oligonucleotides in synthetic shuffling techniques herein.

For example, the system can include an instruction set which permits inputting of amino acid sequences of a family of proteins of interest of interest.

These sequences are back-translated with any desired codon usage parameters, e.g., optimal usage parameters for one or more organism to be used for expression, or to optimize sequence alignments to facilitate recombination, or both. For example, codon usage can be selected for multiple expression hosts, e.g. *E. coli* and *S. cerevisiae*. In some cases, simply optimizing codon usage for expression in a host cell will result in making homologous sequences more similar, as they will lose their natural species codon bias.

Sequences are aligned, and a consensus sequence is produced, optionally showing degenerate codons.

Oligonucleotides are designed for synthetic construction of one or more corresponding synthetic nucleic acid for shuffling. Input parameters on oligonucleotide design include minimum and maximum lengths, minimum length of identical sequence at the ends, maximum degeneracy per oligonucleotide, length of oligo overlap, etc.

As noted, an alternative to back-translation to achieve optimal codon usage for expression in a particular organism is to back-translate sequences to optimize nucleotide homology between family members. For example, amino acid sequences are aligned. All possible codons for each amino acid are determined and codons that minimize differences between the family of aligned sequences are chosen at each position.

USE OF FAMILY SHUFFLING TO IDENTIFY STRUCTURAL MOTIFS CONFERING SPECIFIC PROTEIN PROPERTIES

It is often of interest to identify regions of a protein that are responsible for specific properties, to facilitate functional manipulation and design of related proteins. This identification is traditionally done using structural information, usually obtained by biophysical techniques such as X-ray crystallography. The present invention provides an alternative method in which variants are obtained and analyzed for specific properties which are then correlated with sequence motifs.

The sequences of naturally occurring enzymes that catalyze similar or even identical reactions can vary widely: sequences may be only 50% identical or less. While a family of such enzymes can each catalyze an essentially identical reaction, other properties of these enzymes may differ significantly. These include physical properties such as stability to temperature and organic solvents, pH optima, solubility, the ability to retain activity when immobilized, ease of expression in different host systems, etc. They also include catalytic properties including activity (k_{cat} and K_m), the range of substrates accepted and even the chemistries performed.

The method described here can also be applied to non-catalytic proteins (i.e. ligands such as cytokines) and even nucleic acid sequences (such as promoters that may be inducible by a number of different ligands), wherever multiple functional dimensions are encoded by a family of homologous sequences.

Because of the divergence between enzymes with similar catalytic functions, it is not usually possible to correlate specific properties with individual amino acids at certain positions, as there are simply too many amino acid differences. However, libraries of variants can be prepared from a family of homologous natural sequences by DNA family shuffling. These libraries contain the diversity of the original set of sequences, in a large number of different combinations. If individuals from the library are then tested under a specific set of conditions for a particular property, the optimal combinations of sequences from the parental set for those conditions can be determined.

If the assay conditions are then altered in only one parameter, different individuals from the library will be identified as the best performers. Because the screening conditions are very similar, most amino acids are conserved between the two sets of best performers. Comparisons of the sequences (e.g., in silico) of the best enzymes under the two different conditions identifies the sequence differences responsible for the differences in

performance. Principal component analysis is a powerful tool to use for identifying sequences conferring a particular property. For example, Partek Incorporated (St. Peters, Missouri; www.partek.com) provides software for pattern recognition (e.g., which provide Partek Pro 2000 Pattern Recognition Software) which can be applied to genetic algorithms for multivariate data analysis, interactive visualization, variable selection, neural & statistical modeling. Relationships can be analyzed, e.g., by Principal Components Analysis (PCA) mapped scatterplots and biplots, Multi-Dimensional Scaling (MDS) mapped scatterplots, Star plots, etc.

Once sequence motifs have been identified, proteins are manipulated in, e.g., any of a number of ways. For example, identified changes are optionally deliberately introduced into other sequence backrounds. Sequences conferring different specific properties can be combined. Identified sequence regions of importance for a specific function can be targeted for more thorough investigation, for example by complete randomization using degenerate oligonucleotides, e.g., selected by in silico processes.

IDENTIFICATION OF PARENTAL CONTRIBUTORS TO CHIMERAS PRODUCED BY FAMILY SHUFFLING

This example provides a method for the identification of parental contributors to chimeras produced by family shuffling.

The method takes as an input the sequences of parental genes, and the sequences of chimeras, and compares each chimera with each parent. It then builds sequence and graphical maps of each chimera, indicating the parental source of each chimeric fragment. Correlation of this with functional data permits identification of parents that contribute to specific properties and thereby facilitates the selection of parents for new more focused libraries which can be made by any of the methods noted herein, and screened for any desired functional property.

In one simple example, family 3 and 4 genes contribute to an activity at, e.g., pH 5.5, while family 1 and 2 genes are better at pH 10. Thus, for an application at low pH, the parental composition to create a library would be biased towards 3 and 4, while for high pH a predominantly 1 and 2-based library would be appropriate. Thus, a GO can be implemented which selects oligonucleotides for gene reconstruction predominantly from families 3 and 4. Additional details regarding gene blending methods utilized in oligonucleotide shuffling are found in "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by

Crameri et al., filed February 5, 1999 (USSN 60/118,813) and filed June 24, 1999 (USSN 60/141,049) and filed September 28, 1999 (USSN 09/408,392).

Gene blending is similar to principal component analysis (PCA) for identification of specific sequence motifs. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called "principal components." The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Traditionally, principal component analysis is performed on a square symmetric SSCP (pure sums of squares and cross products) matrix, covariance (scaled sums of squares and cross products) matrix, or correlation (sums of squares and cross products from standardized data) matrix. The analysis results for objects of type SSCP and covariance are similar. A correlation object is used when the variances of individual variates differ substantially or the units of measurement of the individual variates differ. Objectives of principal component analysis include, e.g., to discover or to reduce the dimensionality of a data set, to identify new meaningful underlying variables, and the like.

The main difference is that the present operation gives information about which parents to use in a mix (and so directs construction of new natural gene-based libraries), while PCA identifies specific motifs and so is well suited for more general synthetic shuffling, with identified discrete regions being altered in either a directed or randomized way.

Partek (PCA) software discussed above has an "experimental design" component, that identifies variables that appear to have an effect on a specific function. As applied to the present example, this is useful in an iterative process in which a family library is constructed and screened and the resulting chimeras analyzed for functional correlations with sequence variations. This is used to predict sequence regions for a particular function, and a library is selected in silico by any desired GO directed changes of the region which correlates to functional activity. A focused library which has diversity in those regions is constructed, e.g., by oligonucleotide synthetic methods as described herein. The resulting library members (chimeras) are analyzed for functional correlations with sequence variations. This approach focuses the search of variation in sequence space on the most relevant regions of a protein or other relevant molecule.

After sequences which are active are deconvoluted, the resulting sequence information is used to refine further predictions for in silico operations, e.g., in a neural net training approach.

For example, neural net approaches can be coupled to genetic algorithm-type programming. for example, NNUGA (Neural Network Using Genetic Algorithms) is an available program (http://www.cs.bgu.ac.il/~omri/NNUGA/) which couples neural networks and genetic algorithms. An introduction to neural networks can be found, e.g., in Kevin Gurney (1999) An Introduction to Neural Networks, UCL Press, 1 Gunpowder Square, London EC4A 3DE, UK. and at http://www.shef.ac.uk/psychology/gurney/notes/index.html. Additional useful neural network references include those noted above in regard to genetic algorithms and, e.g., Christopher M. Bishop (1995) Neural Networks for Pattern Recognition Oxford Univ Press; ISBN: 0198538642; Brian D. Ripley, N. L. Hjort (Contributor) (1995) Pattern Recognition and Neural Networks Cambridge Univ Pr (Short); ISBN: 0521460867.

COUPLING OF RATIONAL PROTEIN DESIGN AND SHUFFLING

A protein design cycle, involving cycling between theory and experiment, has led to recent advances in rational protein design. A reductionist approach, in which protein positions are classified by their local environments, has aided development of appropriate energy expressions. Protein design programs can be used to build or modify proteins with any selected set of design criteria. *See*, e.g., http://www.mayo.caltech.edu/; Gordon and Mayo (1999) "Branch-and-Terminate: A Combinatorial Optimization Algorithm for Protein Design" Structure with Folding and Design 7(9):1089-1098; Street and Mayo (1999) "Intrinsic ß-sheet Propensities Result from van der Waals Interactions Between Side Chains and the Local Backbone" Proc. Natl. Acad. Sci. USA, 96, 9074-9076; Gordon et al. (1999) "Energy Functions for Protein Design" Current Opinion in Structural Biology 9(4):509-513 Street and Mayo (1999) "Computational Protein Design" Structure with Folding and Design 7(5):R105-R109; Strop and Mayo (1999) "Rubredoxin Variant Folds Without Iron" J. Am. Chem. Soc. 121(11):2341-2345; Gordon and Mayo (1998) "Radical Performance Enhancements for Combinatorial Optimization Algorithms based on the Dead-End Elimination Theorem" J. Comp. Chem 19:1505-1514; Malakauskas and Mayo (1998) "Design, Structure, and Stability of a Hyperthermophilic Protein Variant" Nature Struct. Biol. 5:470. Street and Mayo (1998) "Pairwise Calculation of Protein Solvent-Accessible Surface Areas" Folding & Design 3: 253-258. Dahiyat and Mayo (1997) "De Novo Protein Design: Fully Automated Sequence

Selection" Science 278:82-87; Dahiyat and Mayo (1997) "Probing the Role of Packing Specificity in Protein Design" Proc. Natl. Acad. Sci. USA 94:10172-10177; Dahiyat et al. (1997) "Automated Design of the Surface Positions of Protein Helices" Prot. Sci. 6:1333-1337; Dahiyat et al. (1997) "De Novo Protein Design: Towards Fully Automated Sequence Selection" J. Mol. Biol. 273:789-796; and Haney et al. (1997) "Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus*" Proteins 28(1):117-30. These design methods rely generally on energy expressions to evaluate the quality of different amino acid sequences for target protein structures. In any case, designed or modified proteins or character strings corresponding to proteins can be directly shuffled in silico, or reverse translated and shuffled in silico and/or by physical shuffling. Thus, one aspect of the invention is the coupling of high-throughput rational design and in silico or physical shuffling and screening of genes to produce activities of interest.

Similarly, molecular dynamic simulations such as those above and, e.g., Ornstein et al. (http://www.emsl.pnl.gov:2080/homes/tms/bms.html; Curr Opin Struct Biol (1999) 9(4):509-13) provide for "rational" enzyme redesign by biomolecular modeling & simulation to find new enzymatic forms that would otherwise have a low probability of evolving biologically. For example, rational redesign of p450 cytochromes and alkane dehalogenase enzymes are a target of current rational design efforts. Any rationally designed protein (e.g., new p450 homologues or new alkaline dehydrogenase proteins) can be evolved by reverse translation and shuffling against either other designed proteins or against related natural homologous enzymes. Details on p450s can be found in Ortiz de Montellano (ed.) 1995, Cytochrome P450 Structure and Mechanism and Biochemistry, Second Edition Plenum Press (New York and London).

Comparison of protein crystal structures to predict crossover points based on structural rather than sequence homology considerations can be conducted and crossover can be effected by oligos to direct chimerization as discussed herein.

MULTIVARIATE SEQUENCE-ACTIVITY MODELING OF PROTEINS; OPTIMIZATION OF ENZYMATIC ACTIVITY BY RATIONAL STATISTICS.

This section describes how to analyze a large number of related protein sequences using modern statistical tools and how to derive novel protein sequences with desirable features using rational statistics and multivariate analysis.

Background Multivariate Data Analysis

Multivariate data analysis and experimental design is widely applied in industry, government and research centers. It is typically used for things like formulating gasoline, or optimizing a chemical process. In the classic example of gasoline formulation, there may be more than 25 different additives that can be added in different amounts and in different combinations. The output of the final product is also multifactorial (energy level, degree of pollution, stability etc. etc.). By using experimental design, a limited number of test formulations can be made where the presence and amounts of all additives are altered in a non-random fashion in order to maximally explore the relevant "formulation space." The appropriate measurements of the different formulations are subsequently analyzed. By plotting the datapoints in a multivariate (multidimensional) fashion, the formulation space can be graphically envisioned and the ideal combination of additives can be extracted. One of the most commonly used statistical tools for this type of analysis is Principal Component Analysis (PCA).

In this example, this type of matrix is used to correlate each multidimensional datapoint with a specific output vector in order to identify the relationship between a matrix of dependant variables Y and a matrix of predictor variables X. A common analytical tool for this type of analysis is Partial Least Square Projections to Latent Structures (PLS). This, for instance, is often used in investment banker's analysis of fluctuating stock prices, or in material science predictions of properties of novel compounds. Each datapoint can consist of hundreds of different parameters that are plotted against each other in an n-dimensional hyperspace (one dimension for each parameter). Manipulations are done in a computer system, which adds whatever number of dimensions are needed to be able to handle the input data. There are previously mentioned methods (PCA, PLS and others) that can assist in finding projections and planes so that hyperspace can be properly analyzed.

Background Sequence Analysis

Nucleotide or amino acid sequence analysis has traditionally been concentrated on qualitative pattern recognition (e.g., sequence classification). This mainly involves identifying sequences based on similarity. This works well for predictions or identifications of classification, but does not always correlate with quantitative values. For instance, a consensus transcriptional promoter may not be a good promoter in a particular application, but is, instead, the average promoter among an aligned group of related sequences.

To access the quantitative features of related biological sequences (DNA/RNA or amino acids) one can analyze the systematic variations (i.e. systematic absence of similarity) among aligned sequences with related biological activity. By applying different multivariate analysis tools (such as PLS) to protein sequences, it is possible to predict a sequence that generates better catalytic activity than the best one present in the analyzed set. Experimental data showing the success of the general method is described below.

Background promoter activity multivariate analysis

One of the few references where multivariate data analysis has been applied to biological sequences was focused on analyzing a set of defined transcriptional promoters in order to see if one could predict a stronger promoter than any of those found in the training set (Jonssonet al. (1993) Nucleic Acids Res. 21:733-739).

In this example, promoter sequences were parametrizised. For simplicity, the physical-chemical differences between the respective nucleotides (A,C,G,T) were selected as equal, i.e. no nucleotide was considered more closely related to any other nucleotide. They were represented as four diametrically opposed corners of a cube forming a perfect tetrahedron. By assigning an origin to the center, each corner can be represented by a numerical coordinate, reflecting the numeric representation. Since the descriptors solely represents equal distribution of properties, any nucleoside can be positioned at any corner. *See,* Figure 17.

Previous studies (Brunner and Bujard (1987) EMBO J. 6;3139-3144; Knaus and Bujard (1988) EMBO J. 7:2919-2923; Lanzer and Bujard (1988) Proc. Natl. Acad. Sci. 85:8973-8977) analyzed a set of 28 promoters that had been studied in detail in the identical context. The promoters included *E. coli* promoters, T5 phage promoters and a number of chimeric and synthetic promoters. They were all cloned as 68 bases (-49 to +19) inserted in front of a DHFR coding region. Relative transcriptional level was measured by dot-blot using the vector derived ß-lactamase gene as internal standard.

In this example, when the 28 promoters, each with 68 nucleotides, were parameterized with the three descriptors as defined in Fig. 14, the result is a 28x204 matrix (28 promoters x 68 nucleotides x 3 parameters (steric bulk, hydrophobicity and polarizability)). The unique sequence of each promoter can be represented as a single point in a 204-dimensional hyperspace. This compilation of 28 promoters thus formed a cluster of 28 points in this space. The experimental data from previous studies noted above was repeated and the generated transcription levels plotted against the cluster of 28 points generated in the 204

dimensional hyperspace using PLS. Half of the promoters (14) were subsequently used to build a statistical model and the other half to test it. This was showed to generate a good correlation for calculated vs. observed promoter strength.

Two new promoters were constructed based on extrapolations of the generated model, and in both cases were shown to be significantly better than the best promoter present among the 28 initial promoters.

Multivariate Analysis And Protein Sequence

The same analytical methods can be applied to protein sequences. Signal peptides have been characterized using multivariate analysis showing good correlation between location in hyperspace and final physical localization (Sjöström et al. (1987) EMBO J. 6:823-831). The main difference between nucleotides and amino acids being that instead of qualitative descriptors of the nucleotides (*see* Fig. 14), quantitative descriptors have to be used to parameterize amino acids. The relevant features of the amino acids (steric bulk, hydrophobicity and polarizability) have been determined and can be extracted from the literature (Hellberg et al. (1986) Acta Chem. Scand. B40:135-140; Jonsson et al., (1989) Quant. Struct. –Act. Relat. 8:204-209).

Following shuffling and characterization of shuffled proteins, the mutated proteins are analyzed, both those that are "better" and those that are "worse" than the initial sequences. Statistical tools (such as PLS) can be used to extrapolate novel sequences that are likely to be better than the best sequence present in the analyzed set.

MODELLING PROTEIN SEQUENCE SPACE

As described above, any protein encoded by a DNA sequence can be plotted as a distinct point in multidimensional space using statistical tools. A "normal" 1kb gene can constitute, e.g., about 330 amino acids. Each amino acid can be described by, e.g., three major physico-chemical quantitative descriptors (steric bulk, hydrophobicity and polarizability) for each amino acid (other descriptors for proteins are largely dependent on these three major descriptors). *See also*, Jonsson et al. (1989) Quant. Struct.- Act. Relat. 8:204-209. Thus, a 1 kb gene is modeled in 330 (number of amino acids) X 20 (possible amino acids at each position) X 3 (the three main descriptors noted above, for each amino acid) = 19,800 dimensions. Because of the extended nature of sequence space, a number of shuffled sequences are used to validate sequence activity related predictions. The closer the surrounding sequences are in space (percent similarity), the higher the likelihood that

predictive value can be extracted. Alternatively, the more sequence space which is analyzed, the more accurate predictions become. This modeling strategy can be applied to any available sequence.

The prognostic value of plotting a large number of shuffled sequences is described above. Two additional approaches can also be used. First, plotting as many chimeric progeny in a library as possible vs. an enzymatic activity using, for example, PLS (partial least square projections to latent structures) can be performed. If enough data is available, the sequence-activity plot forms a function that can be extrapolated outside of the experimental data to produce an in-silico sequence corresponding to an activity higher than the best training set. Second, all related sequences can be plotted and certain sequences grouped with given related activities. Using this matrix, subsequent genes can immediately be grouped with appropriate activity or new related activities can be directly screened for by a subset of the sequenced clones.

An overall issue for the above strategy is the availability of enough related sequences generated through shuffling to provide useful information. An alternative to shuffling sequences is to apply the modeling tools to all available sequences, e.g., the GenBank database and other public sources. Although this entails massive computational power, current technologies make the approach feasible. Mapping all available sequences provides an indication of sequence space regions of interest. In addition, the information can be used as a filter which is applied to in silico shuffling events to determine which virtual progeny are preferred candidates for physical implementation (e.g., synthesis and/or recombination as noted herein).

SHUFFLING OF CLADISTIC INTERMEDIATES

The present invention provides for the shuffling of "evolutionary intermediates." In the context of the present invention, evolutionary intermediates are artificial constructs which are intermediate in character between two or more homologous sequences, e.g., when the sequences are grouped into an evolutionary dendogram.

Nucleic acids are often classified into evolutionary dendograms (or "trees") showing evolutionary branch points and, optionally, relatedness. For example, cladistic analysis is a classification method in which organisms or traits (including nucleic acid or polypeptide sequences) are ordered and ranked on a basis that reflects origin from a postulated common ancestor (an intermediate form of the divergent traits or organisms). Cladistic

analysis is primarily concerned with the branching of relatedness trees (or "dendograms") which shows relatedness, although the degree of difference can also be assessed (a distinction is sometimes made between evolutionary taxomomists who consider degrees of difference and those who simply determine branch points in an evolutionary dendogram (classical cladistic analysis); for purposes of the present invention, however, relatedness trees produced by either method can produce evolutionary intermediates).

Cladistic or other evolutionary intermediates can be determined by selecting nucleic acids which are intermediate in sequence between two or more extant nucleic acids. Although the sequence may not exist in nature, it still represents a sequence which is similar to a sequence in nature which had been selected for, i.e., an intermediate of two or more sequences represents a sequence similar to the postulated common ancestor of the two or more extant nucleic acids. Thus, evolutionary intermediates are one preferred shuffling substrate, as they represent "pseudo selected" sequences, which are more likely than randomly selected sequences to have activity.

One benefit of using evolutionary intermediates as substrates for shuffling (or of using oligonucleotides which correspond to such sequences) is that considerable sequence diversity can be represented in fewer starting substrates (i.e., if starting with parents A and B, a single intermediate "C" has at least a partial representation of both A and B). This simplifies the oligonucleotide synthesis scheme for gene reconstruction/ recombination methods, improving the efficiency of the procedure. Further, searching sequence databases with evolutionary intermediates increases the chances of identifying related nucleic acids using standard search programs such as BLAST.

Intermediate sequences can also be selected between two or more synthetic sequences which are not represented in nature, simply by starting from two synthetic sequences. Such synthetic sequences can include evolutionary intermediates, proposed gene sequences, or other sequences of interest that are related by sequence. These "artificial intermediates" are also useful in reducing the complexity of gene reconstruction methods and for improving the ability to search evolutionary databases.

Accordingly, in one significant embodiment of the invention, character strings representing evolutionary or artificial intermediates are first determined using alignment and sequence relationship software and then synthesized using oligonucleotide reconstruction methods. Alternately, the intermediates can form the basis for selection of oligonucleotides used in the gene reconstruction methods herein.

Several of the following sections describe implementations of this approach using hidden *Markov* model threading and other approaches.

In Silico Shuffling Using Hidden *Markov* Model Threading

A concern with synthetic shuffling is the assumption that each amino acid present among the parents is an independent entity and add (or does not add) function in a given functional dimension by itself. When shuffling using DNAse I based methods, this problem is avoided because recombination during assembly occurs as 20-200 bp fragments and, thus, each amino acid exists in its evolutionary context among other amino acids that have co-evolved in a given direction due to selective pressure of the functional unit (gene or promoter or other biological entity). By capturing the co-variance normally existing within a family of genes, either wild type or generated through regular shuffling, a significant number of biologically inactive progeny are avoided, improving the quality of the generated library. Artificially generated progeny may be inactive due to structural, modular or other subtle inconsistencies between the active parents and the progeny.

One way of weeding out unwanted non co-variance progeny is to apply a statistical profile such as Hidden Markov Model (HMM) on the parental sequences. An HMM matrix generated (e.g., as in Fig. 15) can capture the complete variation among the family as probabilities between all possible states (i.e. all possible combinations of amino acids, deletions and insertion). The matrix resulting from the analyzed family is used to search assorted databases for additional members of the family that are not similar enough to be identified by standard BLAST algorithms of any one particular sequence, but which are similar enough to be identified when probing using a probabilistic distribution pattern based on the original family.

The HMM matrix shown in Fig. 15 exemplifies a family of 8 amino acid peptides. In each position, the peptide can be a specific amino acid (one of the 20 present in the boxes), an insertion (diamonds), or a deletion (circles). The probability for each to occur is dependent on how often it occurs among the compiled parents. Any given parent can subsequently be 'threaded' through the profile in such way that all allowed paths are given a probability factor.

HMM can be used in other ways as well. Instead of applying the generated profile to identify previously unidentified family members, the HMM profile can be used as a template to generate de novo family members (e.g., intermediate members of a cladistic tree of nucleic acids). For example, the program, HMMER is available (http://hmmer.wustl.edu/).

This program builds a HMM profile on a defined set of family members. A sub-program, HMMEMIT, reads the profile and constructs de novo sequences based on that. The original purpose of HMMEMIT is to generate positive controls for the search pattern, but the program can be adapted to the present invention by using the output as in silico generated progeny of a HMM profile defined shuffling. According to the present invention, oligonucleotides corresponding to these nucleic acids are generated for recombination, gene reconstruction and screening.

As the sequence context of each position is accounted for in a probabilistic fashion, the number of non-active progeny is significantly lower than a shuffling reaction that simply randomly selects such progeny. Crossover between genetic modules (structural, functional, or non-defined) occur where they occur in nature (i.e. among the parents) and co-evolution of point mutations or structural elements is retained throughout the shuffling process.

Example Alogrithm for Generating Sequence Intermediates from Sequence Alignments

The following is an outline for a program for generating sequence intermediates from alignments of related parental nucleic acids.

Given an alignment of sequences which code for the parents, and an alignment which codes for the children

for each child sequence

    for each parent sequence

        for each window

            if the parent sequence and child sequence match for this window

                If this window not already covered by sequences in segment list

                    Try to expand window, 5' and 3' until too many mismatches

                    Add final expanded segment to segment list for this sequence

for each child sequence

    set position to start of sequence

    do the following until the end of the child sequence is reached

        Search through segments finding a segment that extends the longest from a point before position (this is most like the parent segment)

        If one is found add to optimal path list and set position to end of found segment

if one is not found increment current position

display segments from optimal path list

NORMALIZATION OF LIBRARIES—USE OF POSITIVE OR NEGATIVE ACTIVITY DATA

One aspect of the present invention is to use positive or negative data in sequence design and selection methods, either in silico, or in physical processing steps, or both. The use of positive or negative data can be in the context of a learning heuristic, a neural network or by simply using positive or negative data to provide logical or physical filters in design and library synthesis processes. Learning networks are described *supra.* and provide one convenient way of using positive or negative data to increase the chances that additional sequences which are subsequently generated will have a desired activity. The ability to use negative data to reduce the size of libraries to be screened provides a considerable advantage, as screening is often a limiting step in generating improved genes and proteins by forced evolution methods. Similarly, the use of positive data to bias libraries towards sequences of interest is another way of focusing libraries.

For example, as noted, in addition to a neural net learning approach, positive or negative data can be used to provide a physical or logical "filter" for any system of interest. That is, sequences which are shown to be inactive provide useful information about the likelihood that closely related sequences will also prove to be inactive, particularly where active sequences are also identified. Similarly, sequences which are active provide useful information about the likelihood that closely related sequences will also prove to be active, particularly where inactive sequences are also identified. These active or inactive sequences can be used to provide a virtual or physical filter to bias libraries (physical or virtual) toward production of more active members.

For example, when using negative data, physical subtraction methods use hybridization to inactive members under selected stringency conditions (often high stringency, as many libraries produced by the methods herein comprise homologous members) to remove similar nucleic acids from libraries which are generated. Similarly, hybridization rules or other parameters can be used to select against members that are likely to be similar to inactive sequences. For example, oligonucleotides used in gene reconstruction methods can be biased against sequences which have been shown to be inactive. Thus, in certain methods, libraries or character strings are filtered by subtracting the library or set of character strings with

members of an initial library of biological polymers which display activity below a desired threshold.

When using positive data, physical enrichment methods use hybridization to active members under selected stringency conditions (often high stringency, as many libraries produced by the methods herein comprise homologous members) to isolate similar nucleic acids which comprise the members of libraries to be produced. Similarly, hybridization rules or other parameters can be used to select for members that are likely to be similar to active sequences. For example, oligonucleotides used in gene reconstruction methods can be biased towards sequences which have been shown to be active. Thus, in certain methods, libraries or character strings are filtered by biasing the library or set of character strings with members of an initial library of biological polymers which display activity above a desired threshold.

Similarly, in silico approaches can be used to produce libraries of inactive sequences, rather than active sequences. That is, inactive sequences can be shuffled in silico to produce libraries of clones that are less likely to be active. These inactive sequences can be physically generated and used to subtract libraries (typically through hybridization to library members) generated by other methods. This subtraction reduces the size of the library to be screened, primarily through the elimination of members that are likely to be inactive.

Example--Motif Filtering

Selections or screens often yield too many "positive" clones to cost-effectively sequence all of the positive or negative clones. However, if sequence motifs are identified that are enriched or depopulated in either the positive or the negative clones, then this bias is used in the construction of synthetic libraries that are biased toward "good" motifs and biased away from "bad" motifs.

If each contiguous selected region or motif (e.g., the selected window can be, e.g., a 20 base region) is thought of as a separate gene or gene element, one can measure the change in gene frequencies before and after a selection or screen. Motifs that increase in frequency in the positive clones are characterized as "good" and motifs whose frequency is reduced in the positive clones are characterized as "bad." Second generation libraries are synthesized in which the library is selected to be enriched for good motifs and depopulated with bad motifs, using any filtering or learning process as set forth herein.

A variety of methods for measuring frequencies of motifs in populations of genes are available. For example, one can hybridize analyte sequences to a gene chip or other array of nucleic acids with the motifs of interest encoded in a spatially addressable fashion,

e.g., using gene chips as provided by Affymetrix (Santa Clara, CA), or other gene chip manufacturers. Similarly, hybridization to membranes containing spatially addressable motifs and measuring relative signal intensities for probe before and after selection can also be performed in an essentially similar fashion, e.g., using standard Southern or northern blot methods. Relative ratios of identified desirable/undesirable features on the chips also provides an indication of overall library quality. Similarly, phage display or other expression libraries can be used to assess library features, i.e., by evaluating expression products.

Alternatively, real time quantitative PCR (e.g., TaqMan) can be performed where PCR oligos are highly discriminating for the feature of interest. This can be done by, for example, having a polymorphism unique to the motif present at or near the 3' end of an oligo such that it will only prime the PCR efficiently if there is a perfect match. Real time PCR product analysis by, e.g., FRET or TaqMan (and related real time reverse-transcription PCR) is a family of known techniques for real time PCR monitoring that has been used in a variety of contexts (*see*, Laurendeau et al. (1999) "TaqMan PCR-based gene dosage assay for predictive testing in individuals from a cancer family with INK4 locus haploinsufficiency" Clin Chem 45(7):982-6; Laurendeau et al. (1999) "Quantitation of MYC gene expression in sporadic breast tumors with a real-time reverse transcription-PCR assay" Clin Chem 59(12):2759-65; and Kreuzer et al. (1999) "LightCycler technology for the quantitation of bcr/ab1 fusion transcripts" Cancer Research 59(13):3171-4.

If the gene family of interest is highly similar to begin with (for example, over 90% sequence identity), then one can simply sequence the population of genes before and after selection. If several sequencing primers up and down the gene are used, then one can look at the sequences in parallel on a sequencing gel. The sequence polymorphisms near the primer can be read out to see the relative ratio of bases at any given site. For example, if the population starts out with 50%T and 50% C at a given position, but 90% T and 10% after selection, one could easily quantitate this base ratio from a sequencing run that originates near the polymorphism. This method is limited, because as one gets further from the primer and reads through more polymorphisms, the mobilities of the various sequences gets increasingly variable and the traces begin to run together. However, as the cost of sequencing continues to decline and the cost of oligos continues to decrease, one solution is simply to sequence with many different oligos up and down the gene.

Example: Fractional Distillation of Sequence Space

Typical sequence spaces are very large compared to the number of sequences that can be physically cloned and characterized. Computational tools exist with which to describe subsets of sequence space which are predicted to be enriched for clones with properties of interest (*see, supra*). However, there are assumptions and computational limitations inherent to these models. Methods for fractionating sequence space such that it is enriched for molecules which are predicted by a given model to have greater or lesser fitness with respect to a phenotype of interest would be useful for testing such predictive models.

A simple example of how this would work is as follows. There are about 10^{27} possible shuffled IFNs (a fairly typical protein in terms of size) based on the naturally occurring human IFN gene family. This is larger than the number that can be easily screened. If one's goal is to evolve shuffled human IFNs that are active, e.g., on mouse cells, then one could use the information in the literature that shows that residues 121 and 125 from human IFN alpha 1 confer improved activity when transplanted onto other human IFNs such as IFN-α2a. If one assumes that this motif confers improved activity in many different contexts, then one can create a large pool of shuffled IFN genes (typically on the order of 10^9 - 10^{12}), convert them to ssDNA, pass them over an affinity column consisting of an oligonucleotide complementary human IFN alpha1 over these residues, wash under appropriate stringency, elute the bound molecules, PCR amplify the eluted genes, clone the material, and perform functional tests on the expressed clones. This protocol allows one to physically bias a library of shuffled genes strongly in favor of containing this motif which is predicted by this very simple model to confer an improvement in the desired activity.

Ideally, one takes populations that are enriched for the motif and populations that are depopulated with the motif. Both populations are analyzed (say 1000 clones from each population). The hypothesis is then "tested" asking whether this fractionation of sequence space biased the average fitness in the predicted way. If it did, then one could "accept" the hypothesis and scale up the screen of that library. One could also test a number of design algorithms by affinity based fractionation, accept the ones that are supported by the results of the experiment, and then perform the affinity selections in series so that one enriches for clones that meet the design criteria of multiple algorithms.

In this model, shuffling, such as family shuffling, is used as the first order design algorithm. However, additional design algorithms are integrated downstream of the shuffling to further fractionate a sequence space based on simple design heuristics. The method

can be performed at the nucleic acid level with any design algorithm that can be translated into a nucleic acid selection scheme.

A number of variations on this example are useful for reducing the size of libraries that are produced by physical or virtual filtering processes.

For example, affinity electing oligonucleotides that encode motifs of interest prior to gene recombination/resynthesis (either physically or in silico) reduces the diversity of populations of nucleic acids that are produced in gene recombination/ resynthesis methods as noted herein.

Similarly, oligonucleotides encoding motifs can be selected by enzymatically degrading molecules that are not perfectly matched with the oligos, e.g., again prior to gene recombination/ resynthesis methods. Alternately, genes that match imperfectly with the oligonucleotides can be selected for, e.g., by binding to mutS or other DNA mismatch repair proteins.

Polymerization events during recombination/gene synthesis protocols can be primed using one or more oligos encoding the motif(s) of interest. That is, mismatches at or near the 3' end of hybridized nucleic acids reduce or block elongation. In this variation, only newly polymerized molecules are allowed to survive (used in subsequent library construction/ selection steps). This can be done, for example, by priming reverse transcription of RNA and then degrading the RNA.

Another approach is to make the template specifically degradeable. For example, DNA with a high frequency of uracil incorporation can be synthesized. Polymerase-based synthesis is primed with oligos and extended with dNTPs containing no uracil. The resulting products are treated with uracil glycosylase and a nuclease that cleaves at apurinic sites, and the degraded template removed. Similarly, RNA nucleotides can be incorporated into DNA chains (synthetically or via enzymatic incorporation); these nucleotides then serve as targets for cleavage via RNA endonucleases. A variety of other cleavable residues are known, including certain residues which are targets for enzymes or other residues and which serve as cleaveage points in response to light, heat or the like. Where polymerases are currently not available with activity permitting incorporation of a desired cleavage target, such polymerases can be produced using shuffling methods to modify the activity of existing polymerases, or to acquire new polymerase activities.

Localized motifs can easily be translated into affinity selection procedures. However, one sometimes wants to impose a rule that molecules have multiple sequence

features that are separated in space in the gene (e.g., 2, 3, 4, 5, 6, etc. sequence features). This can be timed into a selection by making a nucleic acid template that contains all motifs of interest separated by a flexible linker. The T_m for molecules having all motifs is greater than for molecules having only one or two of the motifs. It is, therefore, possible to enrich for molecules having all motifs by selecting for molecules with high T_ms for the selecting oligo(s).

A "gene" of many such motifs strung together can be synthesized separated by flexible linkers or by bases such as inosine that can base pair promiscuously. One would then select for genes with high T_ms for the selecting nucleic acid. Careful design of the selecting nucleic acid template allows one to enrich for genes having a large number of sequence motifs that are predicted to bias genes containing them toward having a phenotype of interest.

If there is little information about whether any given motif is predicted to favorably bias the library, the technique can still be used. A set of motifs is defined, e.g., based upon sequence conservation between different homologs, or the motifs can even be randomly selected motifs. As long as the sequence space is not isotropic (equally dense with good members in all directions), then one can simply fractionate the sequence space based on a designed or on a random set of motifs, measure the average fitness of clones in the region of sequence space of interest, and then prospect more heavily in the regions that give the highest fitness.

In addition to simple sequence alignment methods, there are more sophisticated approaches available for identifying regions of interest such as macromolecule binding sites. For example, United States Patent 5,867,402 to Schneider, et al (1999) "COMPUTATIONAL ANALYSIS OF NUCLEIC ACID INFORMATION DEFINES BINDING SITES" proposes methods in which binding sites are defined based upon the individual information content of a particular site of interest. Substitutions within the binding site sequences can be analyzed to determine whether the substitution causes a deleterious mutation or a benign polymorphism. Methods of identifying new binding sites using individual information content are also proposed. This approach can be used in the context of the present invention as one way of identifying sequences of interest for in silico manipulation of the sequences.

MOTIF BREEDING

Rational design can be used to produce desired motifs in sequences or sequence spaces of interest. However, it is often difficult to predict whether a given designed motif will be expressed in a functional form, or whether its presence will affect another property of

interest. An example of this is the process of designing glycosylation sites into proteins such that they are accessible to cellular glycosylation machinery and such that they do not negatively affect other properties of the protein such as blocking binding to another protein by virtue of steric hindrance by the attached polysaccaride groups.

One way of addressing these issues is to design motifs or multiple variations of motifs into multiple candidate sites within the target gene. The sequence space is then screened or selected for the phenotype(s) of interest. Molecules that meet the specified design criterion threshold are shuffled together, recursively, to optimize properties of interest.

Motifs can be built into any gene. Exemplary protein motifs include: N-linked glycosylation sites (i.e. Asn-X-Ser), 0-linked glycosylation sites (i.e. Ser or Thr), protease sensitive sites (i.e. cleavage by collagenase after X in P-X-G-P) Rho-dependent transcriptional termination sites for bacteria, RNA secondary structure elements that affect the efficiency of translation, transcriptional enhancer elements, transcriptional promoter elements, transcriptional silencing motifs, etc.

HIGH THROUGHPUT RATIONAL DESIGN

In addition to, or in conjunction with the rational design approaches described *supra*, high throughput rational design methods are also useful. In particular, high throughput rational design methods can be used to modify any given sequence in silico, e.g., before recombination/ synthesis. For example, Protein Design Automation (PDA) is one computationally driven system for the design and optimization of proteins and peptides, as well as for the design of proteins and peptides.

Typically, PDA starts with a protein backbone structure and designs the amino acid sequence to modify the protein's properties, while maintaining it's three dimensional folding properties. Large numbers of sequences can be manipulated using PDA, allowing for the design of protein structures (sequences, subsequences, etc.). PDA is described in a number of publications, including, e.g., Malakauskas and Mayo (1998) "Design, Structure and Stability of a Hyperthermophilic Protein Variant" Nature Struc. Biol. 5:470; Dahiyat and Mayo (1997) "De Novo Protein Design: Fully Automated Sequence Selection" Science, 278, 82-87. DeGrado, (1997) "Proteins from Scratch" Science, 278:80-81; Dahiyat, Sarisky and Mayo (1997) "De Novo Protein Design: Towards Fully Automated Sequence Selection" J. Mol. Biol. 273:789-796; Dahiyat and Mayo (1997) "Probing the Role of Packing Specificity in Protein Design" Proc. Natl. Acad. Sci. USA, 94:10172-10177; Hellinga (1997) "Rational Protein

Design – Combining Theory and Experiment" Proc. Natl. Acad. Sci. USA, 94:10015-10017; Su and Mayo (1997)" Coupling Backbone Flexibility and Amino Acid Sequence Selection in Protein Design" Prot. Sci. 6:1701-1707; Dahiyat, Gordon and Mayo (1997) "Automated Design of the Surface Positions of Protein Helices" Prot. Sci., 6:1333-1337; Dahiyat and Mayo (1996) "Protein Design Automation" Prot. Sci., 5:895-903. Additional details regarding PDA are available, e.g., at http://www.xencor.com/.

In the context of the present invention, PDA and other design methods can be used to modify sequences in silico, which can be synthesized/ recombined in shuffling protocols as set forth herein. Similarly, PDA and other design methods can be used to manipulate nucleic acid sequences derived following selection methods. Thus, design methods can be used recursively in recursive shuffling processes.

NON-OLIGONUCLEOTIDE DEPENDENT IN SILICO SHUFFLING

As discussed herein, many of the methods of the invention involve generating diversity in sequence strings in silico, followed by oligonucleotide gene recombination/ synthesis methods. However, non-oligonucleotide based recombination methods are also appropriate. For example, instead of generating oligonucleotides, entire genes can be made which correspond to any diversity created in silico, without the use of oligonucleotide intermediates. This is particularly feasible when genes are suffciently short that direct synthesis is possible.

In addition, it is possible to generate peptide sequences directly from diverse character string populations, rather than going through oligonucleotide intermediates. For example, solid phase polypeptide synthesis can be performed. For example, solid phase peptide arrays can be constructed by standard solid phase peptide synthesis methods, with the members of the arrays being selected to correspond to the in silico generated sequence strings.

In this regard, solid phase synthesis of biological polymers, including peptides has been performed at least since the early "Merrifield" solid phase peptide synthesis methods, described, e.g., in Merrifield (1963) J. Am. Chem. Soc. 85:2149-2154 (1963). Solid-phase synthesis techniques are available for the synthesis of several peptide sequences on, for example, a number of "pins." *See* e.g., Geysen et al. (1987) J. Immun. Meth. 102:259-274, incorporated herein by reference for all purposes. Other solid-phase techniques involve, for example, synthesis of various peptide sequences on different cellulose disks supported in a column. *See,* Frank and Doring (1988) Tetrahedron 44:6031-6040. Still other solid-phase

techniques are described in U.S. Patent No. 4,728,502 issued to Hamill and WO 90/00626. Methods of forming large arrays of peptides are also available. For example, Pirrung et al., U.S. Patent No. 5,143,854 and Fodor et al., PCT Publication No. WO 92/10092, disclose methods of forming arrays of peptides and other polymer sequences using, for example, light-directed synthesis techniques. *See also*, Stewart and Young, Solid Phase Peptide Synthesis, 2d. ed., Pierce Chemical Co. (1984); Atherton et al. (1989) Solid Phase Peptide Synthesis, IRL Press, Greene, et al. (1991) Protective Groups In Organic Chemistry, 2nd Ed., John Wiley & Sons, New York, NY and Bodanzszyky (1993) Principles of Peptide Synthesis second edition Springer Verlag, Inc. NY. Other useful information regarding proteins is found in R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982); Deutscher, Methods in Enzymology Vol. 182: Guide to Protein Purification, Academic Press, Inc. N.Y. (1990); Sandana (1997) Bioseparation of Proteins, Academic Press, Inc.; Bollag et al. (1996) Protein Methods, 2nd Edition Wiley-Liss, NY; Walker (1996) The Protein Protocols Handbook Humana Press, NJ, Harris and Angal (1990) Protein Purification Applications: A Practical Approach IRL Press at Oxford, Oxford, England; Harris and Angal Protein Purification Methods: A Practical Approach IRL Press at Oxford, Oxford, England; Scopes (1993) Protein Purification: Principles and Practice 3rd Edition Springer Verlag, NY; Janson and Ryden (1998) Protein Purification: Principles, High Resolution Methods and Applications, Second Edition Wiley-VCH, NY; and Walker (1998) Protein Protocols on CD-ROM Humana Press, NJ; and the references cited therein.

In addition to proteins and nucleic acids, it should be appreciated that character string diversity generated in silico can be corresponded to other biopolymers. For example, the character strings can be corresponded to peptide nucleic acids (PNAs) which can be synthesized according to available techniques and screened for activity in any appropriate assay. *See*, e.g., Peter E. Nielsen and Michael Egholm (eds) (1999) Peptide Nucleic Acids: Protocols and Applications ISBN 1-898486-16-6 Horizon Scientific Press, Wymondham, Norfolk, U.K for an introduction to PNA synthesis and activity screening.

ASSAYS--PHYSICAL SELECTION.

Directed Evolution by GAGGS, as in DNA shuffling, or classical strain improvement, or any functional genomics technology, can use any physical assays known in the art for detecting polynucleotides encoding desired phenotypes.

Synthetic genes are amenable to conventional cloning and expression approaches; thus, properties of the genes and proteins they encode can readily be examined after their expression in a host cell. Synthetic genes can also be used to generate polypeptide products by in-vitro (cell-free) transcription and translation. Polynucleotides and polypeptides can thus be examined for their ability to bind a variety of predetermined ligands, small molecules and ions, or polymeric and heteropolymeric substances, including other proteins and polypeptide epitopes, as well as microbial cell walls, viral particles, surfaces and membranes.

For example, many physical methods can be used for detecting polynucleotides encoding phenotypes associated with catalysis of chemical reactions by either polynucleotides directly, or by encoded polypeptides. Solely for the purpose of illustration, and depending on specifics of particular pre-determined chemical reactions of interest, these methods may include a multitude of techniques well known in the art which account for a physical difference between substrate(s) and product(s), or for changes in the reaction media associated with chemical reaction (e.g. changes in electromagnetic emissions, adsorption, dissipation, and fluorescence, whether UV, visible or infrared (heat). These methods also can be selected from any combination of the following: mass-spectrometry; nuclear magnetic resonance; isotopically labeled materials, partitioning and spectral methods accounting for isotope distribution or labeled product formation; spectral and chemical methods to detect accompanying changes in ion or elemental compositions of reaction product(s) (including changes in pH, inorganic and organic ions and the like). Other methods of physical assays, suitable for use in GAGGS, can be based on the use of biosensors specific for reaction product(s), including those comprising antibodies with reporter properties, or those based on in vivo affinity recognition coupled with expression and activity of a reporter gene. Enzyme-coupled assays for reaction product detection and cell life-death-growth selections in vivo can also be used where appropriate. Regardless of the specific nature of the physical assays, they all are used to select a desired property, or combination of desired properties, encoded by the GAGGS-generated polynucleotides. Polynucleotides found to have desired properties are thus selected from the library.

The methods of the invention optionally include selection and/or screening steps to select nucleic acids having desirable characteristics. The relevant assay used for the selection will depend on the application. Many assays for proteins, receptors, ligands and the like are known. Formats include binding to immobilized components, cell or organismal viability, production of reporter compositions, and the like.

In high throughput assays, it is possible to screen up to several thousand different shuffled variants in a single day. For example, each well of a microtiter plate can be used to run a separate assay, or, if concentration or incubation time effects are to be observed, every 5-10 wells can test a single variant (e.g., at different concentrations). Thus, a single standard microtiter plate can assay about 100 (e.g., 96) reactions. If 1536 well plates are used, then a single plate can easily assay from about 100- about 1500 different reactions. It is possible to assay several different plates per day; assay screens for up to about 6,000-20,000 different assays (i.e., involving different nucleic acids, encoded proteins, concentrations, etc.) is possible using the integrated systems of the invention. More recently, microfluidic approaches to reagent manipulation have been developed, e.g., by Caliper Technologies (Mountain View, CA) which can provide very high throughput microfluidic assay methods.

In one aspect, cells, viral plaques, spores or the like, comprising GAGGS shuffled nucleic acids, are separated on solid media to produce individual colonies (or plaques). Using an automated colony picker (e.g., the Q-bot, Genetix, U.K.), colonies or plaques are identified, picked, and up to 10,000 different mutants inoculated into 96 well microtiter dishes containing two 3 mm glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or viruses in plaque applications). The time the pin is in the colony, the number of dips to inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each parameter can be controlled and optimized.

The uniform process of automated colony picking such as the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are optionally shaken in a temperature and humidity controlled incubator. Optional glass balls in the microtiter plates act to promote uniform aeration of cells and the dispersal of cellular (e.g., mycelial) fragments similar to the blades of a fermenter. Clones from cultures of interest can be isolated by limiting dilution. As also described supra, plaques or cells constituting libraries can also be screened directly for the production of proteins, either by detecting hybridization, protein activity, protein binding to antibodies, or the like. To increase the chances of identifying a pool of sufficient size, a prescreen that increases the number of mutants processed by 10-fold can be used. The goal of the primary screen is to quickly identify mutants having equal or better product titers than the parent strain(s) and to move only these mutants forward to liquid cell culture for subsequent analysis.

One approach to screening diverse libraries is to use a massively parallel solid-phase procedure to screen cells expressing shuffled nucleic acids, e.g., which encode enzymes for enhanced activity. Massively parallel solid-phase screening apparatus using absorption, fluorescence, or FRET are available. *See*, e.g., United States Patent 5,914,245 to Bylina, et al. (1999); *see also*, http://www.kairos-scientific.com/; Youvan et al. (1999) "Fluorescence Imaging Micro-Spectrophotometer (FIMS)" Biotechnology et alia<www.et-al.com> 1:1-16; Yang et al. (1998) "High Resolution Imaging Microscope (HIRIM)" Biotechnology et alia, <www.et-al.com> 4:1-20; and Youvan et al. (1999) "Calibration of Fluorescence Resonance Energy Transfer in Microscopy Using Genetically Engineered GFP Derivatives on Nickel Chelating Beads" posted at www.kairos-scientific.com. Following screening by these techniques, sequences of interest are typically isolated, optionally sequenced and the sequences used as set forth herein to design new sequences for in silico or other shuffling methods.

Similarly, a number of well known robotic systems have also been developed for solution phase chemistries useful in assay systems. These systems include automated workstations like the automated synthesis apparatus developed by Takeda Chemical Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Beckman Coulter, Inc. (Fullerton, CA)) which mimic the manual synthetic operations performed by a scientist. Any of the above devices are suitable for use with the present invention, e.g., for high-throughput screening of molecules encoded by codon-altered nucleic acids. The nature and implementation of modifications to these devices (if any) so that they can operate as discussed herein will be apparent to persons skilled in the relevant art.

High throughput screening systems are commercially available (*see, e.g.*, Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures including all sample and reagent pipetting, liquid dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization.

The manufacturers of such systems provide detailed protocols the various high throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like.

A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical or other assay images, *e.g.*, using PC (Intel x86 or pentium chip- compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (*e.g.*, SUN™ work station) computers.

Integrated systems for analysis typically include a digital computer with GO software for GAGGS, and, optionally, high-throughput liquid control software, image analysis software, data interpretation software, a robotic liquid control armature for transferring solutions from a source to a destination operably linked to the digital computer, an input device (e.g., a computer keyboard) for entering data to the digital computer to control GAGGS operations or high throughput liquid transfer by the robotic liquid control armature and, optionally, an image scanner for digitizing label signals from labeled assay components. The image scanner can interface with image analysis software to provide a measurement of probe label intensity. Typically, the probe label intensity measurement is interpreted by the data interpretation software to show whether the labeled probe hybridizes to the DNA on the solid support.

Current art computational hardware resources are fully adequate for practical use in GAGGS (any mid-range priced Unix system (e.g., for Sun Microsystems) or even higher end Macintosh or PCs will suffice). Current art in software technology is adequate (i.e., there are a multitude of mature programming languages and source code suppliers) for design of an upgradable open-architecture object-oriented genetic algorithm package, specialized for GAGGS users with a biological background.

A DIGITAL APPARATUS FOR GOs

Various methods and genetic algorithms (GOs) can be used to perform desirable functions as noted herein. In addition, digital or analog systems such as digital or analog computer systems can control a variety of other functions such as the display and/or control of output files.

For example, standard desktop applications such as word processing software (e.g., Microsoft Word™ or Corel WordPerfect™) and database software (e.g., spreadsheet software such as Microsoft Excel™, Corel Quattro Pro™, or database programs such as Microsoft Access™ or Paradox™) can be adapted to the present invention by inputting one or more character string into the software which is loaded into the memory of a digital system,

and performing a GO as noted herein on the character string. For example, systems can include the foregoing software having the appropriate character string information, e.g., used in conjunction with a user interface (e.g., a GUI in a standard operating system such as a Windows, Macintosh or LINUX system) to manipulate strings of characters, with GOs being programmed into the applications, or with the GOs being performed manually by the user (or both). As noted, specialized alignment programs such as PILEUP and BLAST can also be incorporated into the systems of the invention, e.g., for alignment of nucleic acids or proteins (or corresponding character strings) as a preparatory step to performing an additional GO on the resulting aligned sequences. Software for performing PCA can also be included in the digital system.

Systems for GO manipulation typically include, e.g., a digital computer with GO software for aligning and manipulating sequences according to the GOs noted herein, or for performing PCA, or the like, as well as data sets entered into the software system comprising sequences to be manipulated. The computer can be, *e.g.*, a PC (Intel x86 or Pentium chip- compatible DOS,™ OS2,™ WINDOWS,™ WINDOWS NT,™ WINDOWS95,™ WINDOWS98,™ LINUX, Apple-compatible, MACINTOSH™ compatible, Power PC compatible, or a UNIX compatible (*e.g.*, SUN™ work station) machine) or other commercially common computer which is known to one of skill. Software for aligning or otherwise manipulating sequences can be constructed by one of skill using a standard programming language such as Visualbasic, Fortran, Basic, Java, or the like, according to the methods herein.

Any controller or computer optionally includes a monitor which can include, e.g., a cathode ray tube ("CRT") display, a flat panel display (*e.g.*, active matrix liquid crystal display, liquid crystal display), or others. Computer circuitry is often placed in a box which includes numerous integrated circuit chips, such as a microprocessor, memory, interface circuits, and others. The box also optionally includes a hard disk drive, a floppy disk drive, a high capacity removable drive such as a writeable CD-ROM, and other common peripheral elements. Inputting devices such as a keyboard or mouse optionally provide for input from a user and for user selection of sequences to be compared or otherwise manipulated in the relevant computer system.

The computer typically includes appropriate software for receiving user instructions, either in the form of user input into a set parameter fields, *e.g.*, in a GUI, or in the form of preprogrammed instructions, *e.g.*, preprogrammed for a variety of different specific

operations. The software then converts these instructions to appropriate language for instructing the system to carry out any desired operation. For example, in addition to performing GO manipulation of character strings, a digital system can instruct an oligonucleotide synthesizer to synthesize oligonucleotides for gene reconstruction, or even to order oligonucleotides from commercial sources (e.g., by printing appropriate order forms or by linking to an order form on the internet).

The digital system can also include output elements for controlling nucleic acid synthesis (*e.g.*, based upon a sequence or an alignment of a sequences herein), i.e., an integrated system of the invention optionally includes an oligonucleotide synthesizer or an oligonucleotide synthesis controller. The system can include other operations which occur downstream from an alignment or other operation performed using a character string corresponding to a sequence herein, e.g., as noted above with reference to assays.

In one example, GOs of the invention are embodied in a fixed media or transmissible program component containing logic instructions and/or data that when loaded into an appropriately configured computing device causes the device to perform a GO on one or more character string. Figure 13 shows example digital device 700 that should be understood to be a logical apparatus that can read instructions from media 717, network port 719, user input keyboard 709, user input 711 or other inputting means. Apparatus 700 can thereafter use those instructions to direct GO modification of one or more character string, e.g., to construct one or more data set (e.g., comprising a plurality of GO modified sequences corresponding to nucleic acids or proteins). One type of logical apparatus that can embody the invention is a computer system as in computer system 700 comprising CPU 707, optional user input devices keyboard 709, and GUI pointing device 711, as well as peripheral components such as disk drives 715 and monitor 705 (which displays GO modified character strings and provides for simplified selection of subsets of such character strings by a user. Fixed media 717 is optionally used to program the overall system and can include, e.g., a disk-type optical or magnetic media or other electronic memory storage element. Communication port 719 can be used to program the system and can represent any type of communication connection.

The invention can also be embodied within the circuitry of an application specific integrated circuit (ASIC) or programmable logic device (PLD). In such a case, the invention is embodied in a computer readable descriptor language that can be used to create an ASIC or PLD. The invention can also be embodied within the circuitry or logic processors of a

variety of other digital apparatus, such as PDAs, laptop computer systems, displays, image editing equipment, etc.

In one preferred aspect, the digital system comprises a learning component where the outcomes of physical oligonucleotide assembly schemes (compositions, abundance of products, different processes) are monitored in conjunction with physical assays, and correlations are established. Successful and unsuccessful combinations are documented in a database to provide justification/preferences for user-base or digital system based selection of sets of parameters for subsequent GAGGS processes involving the same set of parental character strings/nucleic acids/proteins (or even unrelated sequences, where the information provides process improvement information). The correlations are used to modify subsequent GAGGS processes to optimize the process. This cycle of physical synthesis, selection and correlation is optionally repeated to optimize the system. For example, a learning neural network can be used to optimize outcomes.

EMBODIMENT IN A WEB SITE.

The methods of this invention can be implemented in a localized or distributed computing environment. In a distributed environment, the methods may implemented on a single computer comprising multiple processors or on a multiplicity of computers. The computers can be linked, e.g. through a common bus, but more preferably the computer(s) are nodes on a network. The network can be a generalized or a dedicated local or wide-area network and, in certain preferred embodiments, the computers may be components of an intra-net or an internet.

In one internet embodiment, a client system typically executes a Web browser and is coupled to a server computer executing a Web server. The Web browser is typically a program such as IBM's Web Explorer, Internet explorer, NetScape or Mosaic. The Web server is typically, but not necessarily, a program such as IBM's HTTP Daemon or other WWW daemon (e.g., LINUX-based forms of the program). The client computer is bi-directionally coupled with the server computer over a line or via a wireless system. In turn, the server computer is bi-directionally coupled with a website (server hosting the website) providing access to software implementing the methods of this invention.

A user of a client connected to the Intranet or Internet may cause the client to request resources that are part of the web site(s) hosting the application(s) providing an implementation of the methods of this invention. Server program(s) then process the request to

return the specified resources (assuming they are currently available). A standard naming convention has been adopted, known as a Uniform Resource Locator ("URL"). This convention encompasses several types of location names, presently including subclasses such as Hypertext Transport Protocol ("http"), File Transport Protocol ("ftp"), gopher, and Wide Area Information Service ("WAIS"). When a resource is downloaded, it may include the URLs of additional resources. Thus, the user of the client can easily learn of the existence of new resources that he or she had not specifically requested.

The software implementing the method(s) of this invention can run locally on the server hosting the website in a true client-server architecture. Thus, the client computer posts requests to the host server which runs the requested process(es) locally and then downloads the results back to the client. Alternatively, the methods of this invention can be implemented in a "multi-tier" format wherein a component of the method(s) are performed locally by the client. This can be implemented by software downloaded from the server on request by the client (e.g. a Java application) or it can be implemented by software "permanently" installed on the client.

In one embodiment the application(s) implementing the methods of this invention are divided into frames. In this paradigm, it is helpful to view an application not so much as a collection of features or functionality but, instead, as a collection of discrete frames or views. A typical application, for instance, generally includes a set of menu items, each of with invokes a particular frame--that is, a form which manifest certain functionality of the application. With this perspective, an application is viewed not as a monolithic body of code but as a collection of applets, or bundles of functionality. In this manner from within a browser, a user would select a Web page link which would, in turn, invoke a particular frame of the application (i.e., subapplication). Thus, for example, one or more frames may provide functionality for inputing and/or encoding biological molecule(s) into one or more character strings, while another frame provides tools for generating and/or increasing diversity of the encoded character string(s).

In particular preferred embodiments, the methods of this invention are implemented as one or more frames providing, e.g., the following functionalit(ies). Function(s) to encode two or more biological molecules into character strings to provide a collection of two or more different initial character strings wherein each of said biological molecules comprises a selected set of subunits; functions to select at least two substrings from the character strings; functions to concatenate the substrings to form one or more product

strings about the same length as one or more of the initial character strings; functions to add (place) the product strings to a collection of strings, and functions to implement any feature of GAGGS or any GO or GA as set forth herein.

The functions to encode two or more biological molecules can provide one or more windows wherein the user can insert representation(s) of biological molecules. In addition, the encoding function also, optionally, provides access to private and/or public databases accessible through a local network and/or the intranet whereby one or more sequences contained in the databases can be input into the methods of this invention. Thus, for example, in one embodiment, where the end user inputs a nucleic acid sequenced into the encoding function, the user can, optionally, have the ability to request a search of GenBank and input one or more of the sequences returned by such a search into the encoding and/or diversity generating function.

Methods of implementing Intranet and/or Intranet embodiments of computational and/or data access processes are well known to those of skill in the art and are documented in great detail (see, e.g., Cluer et al. (1992) A General Framework for the Optimization of Object-Oriented Queries, Proc SIGMOD International Conference on Management of Data, San Diego, California, Jun. 2-5, 1992, SIGMOD Record, vol. 21, Issue 2, Jun., 1992; Stonebraker, M., Editor; ACM Press, pp. 383-392; ISO-ANSI, Working Draft, "Information Technology-Database Language SQL", Jim Melton, Editor, International Organization for Standardization and American National Standards Institute, Jul. 1992; Microsoft Corporation, "ODBC 2.0 Programmer's Reference and SDK Guide. The Microsoft Open Database Standard for Microsoft Windows.TM. and Windows NT.TM., Microsoft Open Database Connectivity.TM. Software Development Kit", 1992, 1993, 1994 Microsoft Press, pp. 3-30 and 41-56; ISO Working Draft, "Database Language SQL-Part 2:Foundation (SQL/Foundation)", CD9075-2:199.chi.SQL, Sep. 11, 1997, and the like). Additional relevant details regarding web based applications are found in "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer, Attorney Docket Number 3271.002WO0.

EXAMPLES

The following examples are intended to further illustrate the present invention and should not be considered to be limiting. One of skill will immediately recognize a variety of parameters which can be changed to achieve essentially similar results.

EXAMPLE 1: DECISION TREE FOR EXAMPLE GAGGS PROCESS

A set of flow schematics which provide a general representation of an exemplary process of Directed Evolution (DE) by GAGGS are enclosed (Figs. 1-4). Fig. 1 provides an example decision making process from an idea of a desired property to selection of a genetic algorithm. Figure 2 provides a directed evolution decision tree from selection of the genetic algorithm to a refined library of parental character strings. Figure 3 provides example processing steps from the refined parental library to a raw derivative library of character strings. Figure 4 processes the raw character strings to strings with a desired property.

Generally the charts are schematics of arrangements for components, and of process decision tree structures. It is apparent that many modifications of this particular arrangement for DEGAGGS, e.g., as set forth herein, can be developed and practiced. Certain quality control modules and links, as well as most of the generic artificial neural network learning components are omitted for clarity, but will be apparent to one of skill. The charts are in a continuous arrangement, each connectable head-to tail. Additional material and implementation of individual GO modules, and many arrangements of GOs in working sequences and trees, as used in GAGGS, are available in various software packages. Suitable references describing exemplar existing software are found, e.g., at http://www.aic.nrl.navy.mil/galist/ and at http://www.cs.purdue.edu/coast/archive/clife/FAQ/www/Q20_2.htm. It will be apparent that many of the decision steps represented in Figs. 1-4 are performed most easily with the assistance of a computer, using one or more software program to facilitate selection/ decision processes.

EXAMPLE 2: MODELING COST ESTIMATES

Use of degenerate synthetic oligos with very limited degree/low level of positional degeneracy (under 0.01-5% per position) can offer a very substantial cost saving in building those libraries which incorporate substantial mutagenicity. For PCR assembly gene synthesis, however, representation of all of the crossover events between parental entries uses synthesis of two dedicated oligos per simulated crossover event.

However, as will be apparent from the examples below and from the combinatorial nature of nucleic acid evolution algorithms, even building very large (10^9-10^{10}) gene libraries for physical screening uses less than 10^3 individual 40-mer oligos for evolution of a family of typical genes of ~1.6 kb size.

Several typical examples below provide examples of costs of gene synthesis components in GAGGS, where the cost calculation is based arbitrarily at $0.7 per base (for a 40-50 nmol quantity, which is adequate for gene reassembly procedures) for exemplary purposes. Larger volume demand in oligo synthesis service leads to substantially lower unit cost (e.g., to a decrease of as much as 10 fold) and the general costs of oligo synthesis are in decline. Oligo synthesis is an inheritantly parallel and routine process easily amenable to automation and thus to increases in throughput. Currently, non-chip parallel devices for oligo synthesis provide an effective capacity to complete simultaneous (single-load) synthesis of 196 (2x96) individual 60-mer oligos in less than 5 hours, with the cost of hardware under $100K, and the cost of reagents under $0.07 per base. Therefore, with an understanding of these costs, the cost estimates made in the examples below can be reduced by at least 8 fold.

EXAMPLE 3: GAGGS OF A SINGLE PARENT LOW MUTAGENICITY LIBRARY.

This example describes GAGGS of a single parent low mutagenicity library derived from an average gene (~1.6 kb), given the sequence information of a single 1.6 kb gene (encoding 500 aa + "convenience" start/end oligos). The goal is to build a library of gene variants with all possible single amino acid changes, one aa change per each gene copy in the library.

Relevant parameters include the number of oligos and cost to build 1 parental 1.6 kb gene, e.g., from 40 mer oligos, with complete 20+20 base overlaps e.g.,, by non-error prone assembly PCR, the number of all possible single aa replacement mutations, the number of distinct non-degenerate 40-mer oligos used to "build-in" all possible single aa mutations, the minimal number of all distinct fixed-position single-codon-degenerate oligos used to incorporate all possible single aa mutations, but not terminations, and the minimal number of all distinct fixed-position single-codon-fully degenerate oligos used to incorporate all possible single aa mutations.

For a 1.6 kb gene, 1 ∞ 1,600 : 40 ∞ 2= 80 oligos; $0.7 ∞ 40 ∞ 80 = $2,240. N=500 ∞ 19= 9,500 9,500 x 2 = 19,000; $532,000 @ $0.7/base $56 per gene, 1 per pool 500 x 2 x 3 = 3,000; $84,000 @ 0.7/base, $8.85 per gene, 20 phenotypes per pool, normalized abundance (e.g. by using only three variable codons, two of which are degenerate: NNT, VAA, TGG) 500 x 2 = 1,000; $28,000 @0.7/base $2.94 per gene, 20 phenotypes per pool, skewed abundance (this results in the presence of significant numbers of truncated genes in the synthesized library).

The same physical oligo inventory used for the first round GAGGS is used in the second round of GAGGS to synthesize a library which contains ~95% of all possible combinations of any of two single aa changes. To have 100% coverage (to include for combinations of mutations within +/- 20 bp proximity, additional oligos are used. Where at least one mutation from the previous round has been identified as beneficial, coverage of all combinations of new mutations within +/- 20 bp of the beneficial mutations uses synthesis of no more than 42 new oligos). The cost of subsequent rounds of GAGGS grow only marginally, and linearly, while diversity sampled in a recursive mode grows exponentially.

EXAMPLE 4:GAGGS OF RECOMBINOGENIC (NON-MUTAGENIC) LIBRARY PARENTED BY A FAMILY OF GENES (GAGGS EQUIVALENT OF SINGLE ROUND OF FAMILY DNA SHUFFLING).

Given sequence information for six fairly average (1.6 kb) size genes, each having six areas of homology with each of the other parental genes (six "heads" and "tails" for chimerizing each area of homology).

Relevant parameters include: the number of oligos and resulting cost to build 6 parental 1.6 kb genes (from 40 mer oligos, complete 20+20 overlaps, by non-error-prone assembly PCR), the number of distinct pairwise crossovers between all matching homology areas, assuming 1 crossover event per pairwise homology region), the number of all possible chimeras using the crossovers, the theoretical library size, and the number of distinct oligos and cost to build all possible chimeras. As above, $6 \infty 1{,}600 : 40 \infty 2 = 480$ oligos, $\$0.7 \infty 40 \infty 480 = \$13{,}440$, i.e., $2,240 per gene built. N=180, calculated according to the formula $N=k \infty m \infty (m-1)$, where m=6 number of parents, and k=6 is the number of pairwise homology areas satisfying crossover conditions. $X=\sim 5.315 \infty 10^9$, calculated according to the formula:

$$X = \sum_{n=1}^{k} \left\{ C_{k}^{n} \times m \times [m \times (m-1)]^{n} \right\}$$

where X is the theoretical library size and n is number of crossovers in each library entry (integer from 1 to k) $2 \infty 180 + 480 = 840$ oligos; $\$0.7 \infty 40 \infty 840 = \$23{,}520$; $0.000048 per gene built. If only 10^6 are screened, then the cost of oligos is $0.024 per gene built; if only 10^5, then the cost of oligos is $0.24 per gene built, if 10^4 are screened then the cost of oligos is $2.35 per gene built.

The cost of running multiple rounds of GAGGS is not additive, as most of the excess oligos from previous rounds can be reused in synthesis of the later generation libraries. Even if only a small fraction of all genes built is actually screened (e.g. 10^4, with cost of oligos $2.35 per gene built), the oligo expenses are comparable with cost of assays on per gene-assay basis. In addition, industry wide oligo synthesis costs are declining.

EXAMPLE 5: STEPWISE GAGGS

This example provides a GAGGS family model stepwise protocol.

A family of genes/proteins (DNA or AA sequence) is selected. All possible pairwise alignments are made to identify pairwise homology regions satisfying crossover operator conditions (length, % identity, stringency). Crossover points are selected, one per each of the pairwise homology substrings, in the middle of each substring, or randomly, or according to an annealing-based probability model built on histograms of crossover probability ranks for every pair of parents. Oligos are selected for assembly PCR and synthesized. Genes/libraries are assembled from synthesized oligos. The libraries are screened/ selected as set forth above.

EXAMPLE 6: SUBTILISIN FAMILY MODEL

Amino acid sequences were aligned (Codon usage can be optimized on retrotranslation for a preferred expression system, and number of oligos for synthesis can be minimized). A Dot plot pairwise alignment of all possible pairs of 7 parents was made (Figs. 5, 6, 7). Figure 5 is a percent similarity alignment for 7 parents. Amino acid sequences are aligned, with the leader peptide excluded. Figure 6 is a dot-plot alignment of the sequences to identify regions of similarity. Figure 7 is a dot plot showing pairwise crossover points in the alignment.

Pair 6 and 7 show 95% percent identity per each window of $\geq$7aa, while all other pairs show 80% percent identity per each window of $\geq$7aa. Note that the stringency of alignment (and subsequent representation of crossover between parents) can be manipulated individually for each pair, so that low homology crossovers can be represented at the expense of highly homologous parents. No structural biases or active site biases were incorporated in this model.

As an example GAGGS calculation for the subtilisin family model, assuming 7 parents, of about 400 amino acids, and 1200 bp each (including the leader) or about 275 amino acids and 825 bp for mature protein, 7 x 825 x 2 + about 500 = 12 kb of total sequence to be

generated by gene synthesis. From 40 mers, with full overlap assembly (20 +20 bp overlaps), about 300 oligos are used.

For pairwise crossover oligos to build chimeras, based on alignment results, with one crossover per each homologous substring, there are about 180 homologous substrings, with 170 in the coding region and 10 in the leader region. With 2 60 mers per each crossover point, and 2 head-tail sets for each pair of parents, about 360 additional oligos dedicated to build crossovers can be used. The total number of oligos is about 660 (300 40 mers and 360 60 mers). At a total cost of oligos of $0.70 per base, the oligos would be about $23,520. The cost of reagents would run about $0.07 per base, for a total cost of about $2,252 dollars.

EXAMPLE 7: NAPTHALENE DIOXYGENASE

Napthalene dioxygenase is a non-heme reductive dioxygenase. There are at least three closely related but catalytically distinct types of Napthalene dioxygenases. Figure 12 provides a schematic of a percent similarity plot for the three different Napthalene dioxygenase types, with the amino acid sequence for the ISP large subunit (which is responsible for substrate specificity) being provided.

At a size of about 1,400 amino acids, there are 3 X 1,400 total base pairs = 260 40 mer oligos for 20+20 overlap gene synthesis. A plot of the sequence alignment reveals that there are 14 +19 +23 = 112 60 mer high stringency oligos used in the recombination. The cost of oligos at $0.70 per base would yield a cost of about $12,000 for synthesis, using about 9 hours of synthesizer time to make the oligos. The estimated library size would be about 9.4 X 10^9 chimeras.

EXAMPLE 8: SINGLE PARENT GAGGS CALCULATION

As noted above, one aspect of the invention provides for single parent GAGGS. In these methods, polynucleotides having desired characteristics are provided. This is accomplished by: (a) providing a parental sequence character string encoding a polynucleotide or polypeptide; (b) providing a set of character strings of a pre-defined length that encode single-stranded oligonucleotide sequences comprising overlapping sequence fragments of an entire parental character string, and an entire polynucleotide strand complementary to the parental character string (splitting the sequence of a parent into oligos suitable for assembly PCR); (c) creating a set of derivatives of parental sequence comprising variants with all possible single point mutations, with, e.g., one mutation per variant string (defining all possible single point mutations); (d) providing a set of overlapping character strings of a pre-defined

length that encode both strands of the parental oligonucleotide sequence, and a set of overlapping character strings of a pre-defined length that encode sequence areas including the mutations (oligos incorporating single point mutations, suitable for the same assembly PCR scheme); (e) synthesizing sets of single-stranded oligonucleotides according to the step (c) (e.g., to build or rebuild the parental sequence or a variant thereof e.g., incorporating single point mutations during gene assembly); (f) assembling a library of mutated genes in assembly PCR from the single-stranded oligonucleotides (pooling, partial pooling, or one per container).

For one gene per container approaches (or other approaches involving physically separating library components, e.g., in arrays), wild type oligos are excluded at mutations; and (g) selecting or screening for recombinant polynucleotides having evolved toward a desired property. In an additional optional step (h), the method includes deconvoluting sequence of the mutated polynuceotides (i.e., determining which library member has a sequence of interest, and what that sequence is) having evolved toward a desired property to determine beneficial mutations (when assembly PCR is one per container format, this is done by positional sequence deconvolution, rather than actual sequencing, i.e., the physical location of the components are adequate to provide knowledge of the sequence). In an optional additional step (i), the method includes assembling a library of recombinant variants which combine some or all possible beneficial mutations in some or all possible combinations, from single-stranded oligos by assembly PCR. This is performed from the same set of oligos; if some of the mutations are positionally close (within any one oligo), then additional single strand oligos are made which incorporate combinations of mutations. An optional step (j) includes selecting or screening for recombinant polynucleotides having evolved further toward a desired property.

An example single parent GAGGS calculation, per 1kb of sequence follows.

Genome length: 1000 bp.

First round mutation rate: 1 amino acid/ gene.

Number of oligonucleotides to build wild-type gene: 52 (40 mers, 20+20 overlap synthesis scheme).

Number of oligonucleotide to provide for all possible single point non-terminating mutations (333 total possible): non degenerate oligos: 13320. Partially degenerate 40 mers, one pg position per oligo: 1920. Fully degenerate oligos, 40 mers, one fg position per oligo: 666.

Error prone PCR assembly will also work, but sequence deconvolution is performed, e.g., by sequencing before subsequent rounds.

The number of additional oligos to allow for construction of all possible recombinations with beneficial mutations would use about 10% of the preceding number of oligos. However, about 95% of all possible recombinants having beneficial mutations can be made from the initial set produced above.

EXAMPLE 9: A PROCESS FOR DESIGN OF CROSSOVER OLIGONUCLEOTIDES FOR SYNTHESIS OF CHIMERICAL POLYNUCLEOTIDES:

First, substrings are identified and selected in parental strings for applying a crossover operator to from chimeric junctions. This is performed by: a) identifying all or part of the pairwise homology regions between all parental character strings, b) selecting all or part of the identified pairwise homology regions for indexing at least one crossover point within each of the selected pairwise homology regions, c) selecting one or more of the pairwise non-homology regions for indexing at least one crossover point within each of the selected pairwise non-homology regions ("c" is an optional step which can be omitted, and is also a step where structure-activity based elitism can be applied), thereby providing a description of a set of positionally and parent-indexed regions/areas (substrings) of parental character strings suitable for further selection of crossover points.

Secondly, further selection of crossover points within each of the substrings of the set of the substrings selected in step 1 above is performed. The steps include: a) randomly selecting at least one of the crossover points in each of the selected substrings, and/or b) selecting at least one of the crossover points in each of the selected substrings, using one or more of annealing-simulation-based models for determining probability of the crossover point selection within each of the selected substrings and/or c) selecting one crossover point approximately in the middle of each of the selected substrings, thereby creating a set of pairwise crossover points, where each point is indexed to corresponding character positions in each of the parental strings desired to from a chimeric junction at that point.

Thirdly, optional codon usage adjustments are performed. Depending on methods used to determine homology (strings encoding DNA or AA), the process can be varied. For example, if a DNA sequences was used: a) adjustment of codons for the selected expression system is performed for every parental string, and b) adjustment of codons among parents can be performed to standardize codon usage for every given aa at every corresponding

position. This process can significantly decrease total number of distinct oligos for gene library synthesis, and may be particularly beneficial for cases where AA homology is higher than DNA homology, or with families of highly homologous genes (e.g. 80%+ identical).

This option has to be exercised with caution, as it is in essence an expression of an elitism mutation operator. Thus, one considers the benefits of cutting the number and resulting costs of oligos vs. introduction of this bias, which can have undesirable consequences. Most typically, one uses codons which encode AA at a given position in a majority of parents.

If AA sequences are used: a) retrotranslate sequence to degenerate DNA; b) define degenerate nucleotides using position-by-position referencing to codon usage in original DNA (of majority of parents or of corresponding parent), and/or - exercise codon adjustments suitable for the selected expression system where a physical assay will be performed.

This step can also be used to introduce any restriction sites within coding parts of the genes, if any, for subsequent identification/QA/deconvolution/manipulations of library entries. All crossover points identified in step 2 above (indexed to pairs of parents) are correspondingly indexed to the adjusted DNA sequences.

Fourth, oligo arrangements are selected for a gene assembly scheme. This step includes several decision steps:

Uniform 40-60 mer oligos are typically used (using longer oligos will result in decrease of # of oligos to build parents, but uses additional dedicated oligos for providing representation of closely positioned crossovers/mutations.

Select whether Shorter/Longer Oligos are allowed (i.e., a Yes/No decision). A "Yes" decision cuts the total number of oligos for high homology genes of different lengths with gaps (deletion/insertion), esp for 1-2aa.

Select the overlap length (typically 15-20 bases, which can be symmetrical or asymmetrical).

Select whether degenerate oligos are allowed (Yes/No). This is another potent cost cutting feature and also a powerful means to obtain additional sequence diversity. Partial degeneracy schemes and minimized degeneracy schemes are especially beneficial in building mutagenic libraries.

If software tools are used for these operations, several variations of the parameters are run to select maximum library complexity and minimal cost. Exercising complex assembly schemes using oligos of various length significantly complicates indexing

processes and, subsequently, assembly of the library in positionally encoded parallel or partial pooling formats. If this is done without sophisticated software, a simple and uniform scheme (e.g. all oligos 40 bases long with 20 bases overlap) can be used.

Fifth, "convenience sequences" are designed in front and in the back of the parent strings. Ideally, it is the same set which will be built in every library entry at the end. These include any restriction sites, primer sequences for assembled product identifications, RBS, leader peptides and other special or desirable features. In principle, the convenience sequences can be defined at a later stage, and at this stage, a "dummy" set of appropriate length can be used, e.g. a substring from an easily recognizable forbidden letters.

Sixth, an indexed matrix of oligo strings for building every parent is created, according to the selected scheme. An index of every oligo includes: a parent identifier (parentID), indication of coding or complementary chain, and position numbers. Crossover-points are determined for indexed coding string of every parent with head and tail convenience substrings. A complementary chain of every string is generated. Every coding string is selected according to the selected assembly PCR scheme in step 4 above (e.g. in increments of 40 bp). Every complement string is split according to the same scheme (e.g. 40 bp with 20 bp shift)

Seventh, an indexed matrix of oligos is created for every pairwise crossover operation. First, all oligos which have pairwise crossover markers are determined. Second, all sets of all oligos which have the same position and same pair of parents crossover markers (4 per crossover point) are determined. Third, every set of 4 oligo strings are taken which have been labeled with the same crossover marker, and another derivative set of 4 chimeric oligo strings comprising of characters encoding 2 coding and 2 complement chains (e.g. with 20 bp shift in 40=20+20 scheme) are made. Two coding strings are possible, having a forward end sequence substring of one parent followed by the backward end of the second parent after crossover point. Complement strings are also designed in the same fashion, thereby obtaining an indexed complete inventory of strings encoding oligos suitable for gene library assembly by PCR.

This inventory can further be optionally refined by detecting all redundant oligos, counting them and deleting from inventory, accompanied by the introduction of the count value to an "abundance=amount" field in the index of each oligo string. This may be a very beneficial step for reducing total number of oligos for library synthesis, particularly in the cases if parental sequences are highly homologous.

EXAMPLE 10: PROGRAM ALGORITHM FOR DESIGNING OLIGONUCLEOTIDES FOR SYNTHESIS

The following is a program outline for designing oligonucleotides for use in synthetic/ recombination protocols.

Given an alignment of proteins and a codon bias table:
For each position in an alignment of proteins
find a set of minimally degenerate codons that code for Amino Acids at this position using codon bias table
For each sequence in alignment
Add three letter codon (DNA) that codes for the amino acid at this position in this sequence to DNA version of sequence
!Gaps are represent by a special codon ~~~
for each sequence of DNA created by above
!Note gaps are ignored in this step
For each window=rough oligo size
check end degeneracy
Try to increase and decrease window length to minimize end degeneracy while staying within length bounds
add oligos given window bounds and all sequences
add oligos given reverse window bounds and all sequences
for each position in dnaseqs
for each sequence
If the current position is the start of a gap
Add oligos given bounds which contains a minimum amount of sequence from current sequence 5' of the gap and a minimum amount of sequence from current sequence 3' of the gap
Repeat add oligos for reverse bounds
add Oligos: given a list of sequences (DNA) and bounds
for each position in the bounds
get all unique bases at this position from DNA sequences in the list
generate a base (or degenerate base symbol) for this position
if total degenerate positions is greater than user defined number

split sequence list in two, add Oligos given sequence list one, add Oligos given sequence list two (recursive)

else add this oligo (set of bases for each position) to oligo list

display all oligo in oligo list

EXAMPLE 11: CROSSOVER POINT SELECTION

Figs. 8-11 are schematics of various processes and process criteria for oligonucleotide selection for recombination between parental nucleic acids. Fig. 8, panel A shows a typical dot-plot alignment of two parents and the increase in crossover probability that results in regions of similarity. Panel B shows that crossovers can be selected based upon a simple logical/physical filter, i.e., the physical or virtual annealing temperature of oligonucleotides, e.g., using a linear annealing temperature. Panel C shows various more complex filters which vary annealing temperature to achieve specific crossovers, i.e., by appropriately controlling the physical or virtual annealing temperature.

Figure 9 schematically represents the introduction of indexed crossover points into the sequence of each of the aligned parents. In brief, sequences are aligned and the positional index of each crossover point (marker field) is represented schematically by a vertical identifier mark. The crossover point, for parents m and n, as represented in Figure 8, is represented by an identifier, a position number for parent m (a head) and a position number for parent n (a tail). This process is repeated for every parent in a data set, applying the oligonucleotide gridding operator (grid of positional indexes which indicate the start and end of every oligonucleotide in a PCR assembly operation) to each of the parents. Figure 10 schematically represents the complete inventory of oligonucleotide sequences to assemble all of the parents. The data set is simplified by identifying all pairs of oligonucleotide sequences with matching pairwise crossover indexes, providing a sub inventory of oligos with crossover markers. Figure 11 provides a schematic for obtaining an inventory of sequences for chimeric oligonucleotides for each of the selected crossover points. In brief, two pairs of oligo sequences with matching pairwise crossover indexes are selected (down arrow 1). Sequences of chimeric oligonucleotides are generated around the pairwise crossover point (head-tail, tail-head, s or a chain) (down arrow 2). In a "40=20+20" assembly scheme (where a 40mer has 20 residues from each parent), only one oligo longer than 60 bp. is used for each chimerization (two for each crossover point, regardless of relative positions within each oligo). This empirical finding can be described by variations of cut and join operations at the depicted S or

A chains (e.g., as in Fig. 8), with the rule being reduced to a guidance table. In the chimeric oligos, the A or S chain and sequence head and tail subfragments are defined using a selected subset of rules from the guidance table. Selection or rules from the guidance table can be automated based upon comparison of relative positions of crossover points in the oligos (boolean operations) (down arrow 3). This process is repeated for every set of oligos with identical crossover indexes to obtain an inventory of sequences for chimeric oligos for each of the selected crossover points.

Modifications can be made to the methods and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to generate shuffled nucleic acids and/or to test shuffled nucleic acids, including in an iterative process.

An assay, kit or system utilizing a use of any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits will optionally additionally comprise instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally comprise one or more of the following: (1) a shuffled component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein; (3) one or more assay component; (4) a container for holding nucleic acids or enzymes, other nucleic acids, transgneic plants, animals, cells, or the like, (5) packaging materials, and (6) software for performing any of the decision steps noted herein related to GAGGS.

In a further aspect, the present invention provides for the use of any component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

The previous examples are illustrative and not limiting. One of skill will recognize a variety of non-critical parameters which may be altered to achieve essentially similar results. *All patents, applications and publications cited herein are incorporated by reference in their entirety for all purposes.*

* C L M S *

WHAT IS CLAIMED IS:

1. A method of making a recombinant nucleic acid, the method comprising:

providing a plurality of parental character strings corresponding to a plurality of nucleic acids, which character strings, when aligned for maximum identity, comprise at least one region of heterology;

aligning the character strings;

defining a set of character string subsequences, which set of subsequences comprises subsequences of at least two of the plurality of parental character strings;

providing a set of oligonucleotides corresponding to the set of character string subsequences;

annealing the set of oligonucleotides; and,

elongating one or more members of the set of oligonucleotides with a polymerase, or ligating at least two members of the set of oligonucleotides with a ligase, thereby producing one or more recombinant nucleic acid.

2. The method of claim 1, wherein the character strings, when aligned for maximum identity comprise at least one region of similarity.

3. The method of claim 1, wherein at least one of the parental character strings is an evolutionary or artificial intermediate.

4. The method of claim 1, wherein at least one of the parental character strings corresponds to a designed nucleic acid.

5. The method of claim 4, wherein the designed nucleic acid represents an energy minimized design for an encoded polypeptide.

6. The method of claim 1, further comprising applying one or more genetic operator to one or more of the parental character strings, or to one or more of the character string subsequences, wherein the genetic operator is selected from: a mutation of the one or more parental character strings or one or more character string subsequences, a multiplication of the one or more parental character strings or one or more character string subsequences, a fragmentation of the one or more parental character strings or one or more character string

This Sheet Has Been Replaced

subsequences, a crossover between any of the one or more parental character strings or one or more character string subsequences or an additional character string, a ligation of the one or more parental character strings or one or more character string subsequences, an elitism calculation, a calculation of sequence homology or sequence similarity of aligned strings, a recursive use of one or more genetic operator for evolution of character strings, application of a randomness operator to the one or more parental character strings or the one or more character string subsequences, a deletion mutation of the one or more parental character strings or one or more character string subsequences, an insertion mutation into the one or more parental character strings or one or more of character string subsequences, subtraction of the of the one or more parental character strings or one or more character string subsequences with an inactive sequence, selection of the of the one or more parental character strings or one or more character string subsequences with an active sequence, and death of the one or more parental character strings or one or more of character string subsequences.

7. The method of claim 1, further comprising selecting a diplomat sequence, which diplomat sequence comprises an intermediate level of sequence similarity between two or more of the plurality of character strings.

8. The method of claim 1, wherein the set of oligonucleotides comprise a plurality of overlapping oligonucleotides.

9. The method of claim 1, wherein the set of character string subsequences is defined by selecting a length for the character string and subdividing at least two of the plurality of parental character strings into segments of the selected length.

10. The method of claim 1, wherein aligning the character strings is performed in a digital computer or in a web-based system.

11. The method of claim 1, further comprising synthesizing a set of single-stranded oligonucleotides which correspond to the set of character string subsequences, thereby providing the set of oligonucleotides.



12. The method of claim 1, further comprising:

pooling all or part of the set of oligonucleotides;

hybridizing the resulting pooled oligonucleotides; and,

extending a plurality of the resulting hybridized oligonucleotides, wherein at least one of the resulting extended double stranded nucleic acids comprises sequences from at least two of the plurality of parental character strings.

13. The method of claim 11, further comprising denaturing the double stranded nucleic acids, thereby producing a heterogeneous mixture of single-stranded nucleic acids.

14. The method of claim 11, further comprising:

(i) denaturing the double stranded nucleic acids, thereby producing a heterogeneous mixture of single-stranded nucleic acids;

(ii) re-hybridizing the heterogeneous mixture of single-stranded nucleic acids; and

(iii) extending the resulting rehybridized double stranded nucleic acids with a polymerase.

15. The method of claim 13, further comprising repeating steps (i) (ii) and (iii) at least twice.

16. The method of claim 1, further comprising selecting the one or more recombinant nucleic acid for a desired property.

17. The method of claim 1 wherein the set of oligonucleotides is provided by synthesizing the oligonucleotides to comprise one or more modified parental character string subsequence, which subsequence comprises one or more of:

a parental character string subsequence modified by one or more replacement of one or more character of the parental character string subsequence with one or more different character;

a parental character string subsequence modified by one or more deletion or insertion of one or more characters of the parental character string subsequence;

a parental character string subsequence modified by inclusion of a degenerate sequence character at one or more randomly or non-randomly selected positions;

a parental character string subsequence modified by inclusion of a character string from a different character string from a second parental character string subsequence at one or more position;

This Sheet Has Been Replaced

a parental character string subsequence which is biased based upon its frequency in a selected library of nucleic acids; and,

a parental character string subsequence which comprises one or more sequence motif, which sequence motif is artificially included in the subsequence.

18. The method of claim 17, wherein the sequence motif comprises an N-linked glycosylation sequence, an O-linked glycosylation sequence, a protease sensitive sequence, a collagenase sensitive sequence, a Rho-dependent transcriptional termination sequence, an RNA secondary structure sequence that affects the efficiency of transcription, an RNA secondary structure sequence that affects the efficiency of translation, a transcriptional enhancer sequence, a transcriptional promoter sequence, or a transcriptional silencing sequence.

19. The method of claim 1, wherein the oligonucleotide set contains one or more altered or degenerate positions as compared to the corresponding subsequence of one or more parental character string.

20. The method of claim 1, further comprising selecting the one or more recombinant nucleic acid based upon its hybridization to a selected nucleic acid or to a set of selected nucleic acids.

21. The method of claim 1, wherein the one or more parental character string comprises at least two parental character strings, wherein the oligonucleotide set comprises at least one oligonucleotide member comprising a chimeric nucleic acid sequence, the at least one oligonucleotide member comprising at least two oligonucleotide member subsequences, wherein the at least two oligonucleotide member subsequences correspond to at least two subsequences from the at least two parental character strings, the at least two oligonucleotide member subsequences being separated by a crossover point.

22. The method of claim 21, wherein the crossover point is selected by identifying a plurality of parental character substrings from a plurality of the at least two parental character strings, aligning the substrings to display pairwise identity between the substrings, and selecting a point within the aligned sequence as the crossover point.

23. The method of claim 21, wherein the crossover point is selected randomly.

This Sheet Has Been Replaced

24. The method of claim 21, wherein the crossover point is selected non randomly.

25. The method of claim 21, wherein the crossover point is selected non randomly by selecting a crossover point approximately in the middle of one or more identified pairwise identity region.

26. The method of claim 21, wherein at least one crossover point for at least one oligonucleotide member is selected from a region outside of an identified pairwise homology region.

27. The method of claim 1, further comprising adding one or more oligonucleotide member of the set of oligonucleotides at a concentration which is higher than at least one or more additional oligonucleotide member of the set of oligonucleotides.

28. The method of claim 1, further comprising incubating one or more member of the oligonucleotide set with the recombinant nucleic acid and a polymerase.

29. The method of claim 1, further comprising denaturing the recombinant nucleic acid, and contacting the recombinant nucleic acid with at least one additional nucleic acid from the oligonucleotide set.

30. The method of claim 1, further comprising denaturing the recombinant nucleic acid, and contacting the recombinant nucleic acid with at least one additional nucleic acid produced by cleavage of a parental nucleic acid encoded by the at least one parental character string.

31. The method of claim 1, further comprising denaturing the recombinant nucleic acid, and contacting the recombinant nucleic acid with at least one additional nucleic acid produced by cleavage of a parental nucleic acid encoded by the at least one parental character string, which parental nucleic acid is cleaved by one or more of: chemical cleavage, cleavage with a DNAse and cleavage with a restriction endonuclease.

32. The method of claim 1, wherein the parental character string encodes one or more nucleic acid corresponding to one or more or protein or gene selected from: EPO, insulin, a peptide hormone, a cytokine, epidermal growth factor, fibroblast growth factor,

This Sheet Has Been Replaced

hepatocyte growth factor, insulin-like growth factor, an interferon, an interleukins, a keratinocyte growth factor, a leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF, an oncogene, a tumor suppressor, a steroid hormone receptor, a plant hormone, a disease resistance gene, an herbicide resistance gene, a bacterial gene, a monooxygenases, a protease, a nuclease, and a lipase.

33. The method of claim 1, wherein the set of oligonucleotides comprises one or more oligonucleotide member between about 20 and about 60 nucleotides in length.

34. The method of claim 1, further comprising selecting the recombinant nucleic acid for a desired trait or property, thereby providing a selected recombinant nucleic acid.

35. The method of claim 34, further comprising recombining the selected recombinant nucleic acid with one or more of: a homolgous nucleic acid, and an oligonucleotide member from the set of oligonucleotides.

36. The method of claim 1, further comprising selecting the recombinant nucleic acid for a desired trait or property, thereby providing a selected recombinant nucleic acid, wherein the desired trait or property is selected in an in vivo selection assay or a parallel solid phase assay.

37. The method of claim 1, further comprising selecting the recombinant nucleic acid for a desired trait or property, thereby providing a selected recombinant nucleic acid, wherein the desired trait or property is selected in an in vitro selection assay.

38. The method of claim 1, further comprising deconvolution of the recombinant nucleic acid.

39. The method of claim 1, further comprising sequencing or cloning the recombinant nucleic acid.

40. The method of claim 1, wherein the recombinant nucleic acid is synthesized in vitro by assembly PCR.

41. The method of claim 1, wherein the recombinant nucleic acid is synthesized in vitro by error-prone assembly PCR.

This Sheet Has Been Replaced

42. The method of claim 1, wherein the parental character strings, or oligonucleotide sets are selected in a computer.

43. A method of making character strings, the method comprising:

a) providing a parental character string encoding a polynucleotide or polypeptide;

b) providing a set of oligonucleotide character strings of a pre-selected length that encode a plurality of single-stranded oligonucleotide sequences comprising sequence fragments of the parental character string, and complement thereof;

c) creating a set of derivatives of the parental sequence comprising sequence variant strings, the set comprising a plurality of mutations, having one mutation per variant string.

44. The method of claim 43, wherein a plurality of the plurality of single-stranded oligonucleotide sequences are overlapping in sequence.

45. The method of claim 43, further comprising applying one or more genetic operator to the parental character string, or to one or more of the oligonucleotide character strings, wherein the genetic operator is selected from: a mutation of the parental character string, or one or more of the oligonucleotide character strings, a multiplication of the parental character string, or one or more of the oligonucleotide character strings, a fragmentation of the parental character string, or one or more of the oligonucleotide character strings, a crossover between any of the parental character string or one or more of the oligonucleotide character strings, or an additional character string, a ligation of the of the parental character string, or one or more of the oligonucleotide character strings, an elitism calculation, a calculation of sequence homology or sequence similarity of an alignment comprising the parental character string, or one or more of the oligonucleotide character strings, a recursive use of one or more genetic operator for evolution of character strings, application of a randomness operator to the parental character string, or to one or more of the oligonucleotide character strings, a deletion mutation of the parental character string, or one or more of the oligonucleotide character strings, an insertion mutation into the parental character string, or one or more of the oligonucleotide character strings, subtraction of the of the parental character string, or one or more of the oligonucleotide character strings with an inactive sequence, selection of the of the parental character string, or one or more of the oligonucleotide character strings with an active sequence, and death of the parental character string, or one or more of the oligonucleotide character strings.

This Sheet Has Been Replaced

46. The method of claim 43, further comprising:

d) providing a set of overlapping character strings of a pre-defined length that encode both strands of the parental character string sequence; and,

e) synthesizing sets of single-stranded oligonucleotides according to the step (c) and (d).

47. The method of claim 46, further comprising:

f) assembling a library of recombinant nucleic acids by assembly PCR from the single-stranded oligonucleotides.

48. A library made by the method of claim 47.

49. The method of claim 47, further comprising:

g) selecting or screening the library for one or more recombinant polynucleotide having a desired property.

50. The method of claim 48, further comprising:

h) deconvoluting the sequence of the one or more selected polynucleotide.

51. The method of claim 46, wherein the sequence of the one or more selected polynucleotide is deconvoluted by sequencing the selected polynucleotide, or by digesting the one or more selected polynucleotide.

52. The method of claim 46, wherein the sequence is deconvoluted by positional deconvolution of the one or more selected polynucleotide.

53. The method of claim 46, further comprising reiterative shuffling or selection of the library of recombinant nucleic acids.

54. A method of facilitating recombination between two or more divergent nucleic acids, the method comprising:

aligning parental character strings corresponding to the divergent nucleic acids, thereby identifying regions of sequence identity and regions of sequence diversity;

defining a diplomat character string which is intermediate in sequence between the parental character strings;

This Sheet Has Been Replaced

synthesizing at least a portion of the diplomat sequence to produce a diplomat nucleic acid; and,

recombining a mixture of selected nucleic acids comprising the parental nucleic acids, or fragments thereof, and the diplomat nucleic acid.

55. The method of claim 54, wherein the diplomat nucleic acid is synthesized by synthesizing a plurality of overlapping oligonucleotides corresponding in sequence to the diplomat sequence, hybridizing the overlapping oligonucleotides, and incubating the overlapping oligonucleotides with a polymerase.

56. The method of claim 54, further comprising synthesizing a pool of oligonucleotides corresponding to one or more of the parental character strings, which pool of oligonucleotides is present in the mixture of selected nucleic acids.

57. The mixture of selected nucleic acids produced by the method of claim 56.

58. A method of generating and recombining nucleic acids, the method comprising:

inputting a plurality of amino acid sequence character strings into a digital system;

reverse translating the amino acid character strings in the digital system into a plurality of nucleic acid character strings, wherein reverse translated nucleic acid sequences are selected for one or more of: species codon bias in a selected expression host, and optimized sequence similarity between the plurality of nucleic acid character strings; and,

synthesizing one or more sets of oligonucleotides corresponding to one or more reverse translated nucleic acid sequences.

59. The method of claim 58, further comprising hybridizing members of the one or more oligonucleotide sets to each other, or to a set of fragmented nucleic acids which encodes one or more amino acid polymer corresponding to one or more of the amino acid sequence character strings.

60. The method of claim 59, further comprising elongating one or more resulting hybridized nucleic acids with a polymerase.

61. The method of claim 58, further comprising fragmenting one or more resulting elongated nucleic acids and hybridizing the resulting secondary fragmented nucleic

This Sheet Has Been Replaced

acids with each other or with members of the one or more oligonucleotide sets, or with a set of primary fragmented nucleic acids which encodes one or more amino acid polymer corresponding to one or more of the amino acid sequence character strings.

62. A method of optimizing activity of a nucleic acid, the method comprising:

parameterizing a set of nucleic acids or proteins to provide a set of multidimensional datapoints;

extrapolating one or more postulated multidimensional datapoint from the set of multidimensional datapoints; and,

converting the postulated multidimensional datapoint to a new character string corresponding to a postulated nucleic acid nucleic acid or protein.

63. The method of claim 62, comprising synthesizing the postulated nucleic acid or protein.

64. The method of claim 62, further comprising principle component analysis of the set of multidimensional datapoints.

65. The method of claim 62, comprising shuffling the postulated nucleic acid, or a subsequence thereof, with an additional nucleic acid.

66. The method of claim 62, wherein the set of nucleic acids or proteins is parameterized by correlating each residue of the nucleic acid or protein to a matrix of numeric indicators.

67. The method of claim 66, wherein the matrix is graphically represented as a tetrahedron, having an assigned origin at the center of the tetrahedron, with each corner represented as a numeric representation, with each residue of a nucleic acid being positioned at a different corner, thereby producing the matrix of numeric indicators.

68. The method of claim 62, comprising correlating each multidimensional datapoint with an output vector to identify a relationship between a matrix of dependent Y variables and a matrix of predictor X variables.

69. The method of claim 68, wherein the correlation is performed by partial least square projections to latent structures analysis.

This Sheet Has Been Replaced

70. The method of claim 62, wherein each multidimensional datapoint comprises more than one different parameter, wherein the parameters are plotted against each other in n dimensional hyperspace, said n dimensional hyperspace comprising at least one dimension for each parameter.

71. A method of providing a library of recombinant nucleic acids which is enriched for a sequence of interest and selecting the library, the method comprising:

producing an initial library of at least about 10^6 recombinant nucleic acids, which initial library of recombinant nucleic acids comprises at least about 10^5 different member types, which 10^5 different member types are non-identical;

hybridizing the library to one or more population of nucleic acids, which one or more population of nucleic acids correspond to one or more subsequences in the different library members;

isolating members of the library which hybridize to the one or more populations of nucleic acids, thereby enriching the library of nucleic acid for members which hybridized to the one or more population of nucleic acids; and,

selecting members of the resulting enriched library for one or more property of interest.

72. The method of claim 71, wherein the initial library has between about 10^9 and 10^{12} members.

73. The method of claim 71, wherein the one or more population of nucleic acids is fixed to a solid substrate.

74. The method of claim 73, wherein the solid substrate comprises one or more of: a column matrix material and a nucleic acid chip.

75. The method of claim 71, wherein the initial library is produced by recombining one or more homologous nucleic acids.

76. The enriched library produced by the method of claim 71.

77. The method of claim 71, wherein the initial library is produced by:

providing a plurality of parental character strings corresponding to a plurality of nucleic acids, which character strings, when aligned for maximum identity, comprise at least one region of similarity and at least one region of heterology;

This Sheet Has Been Replaced

aligning the character strings;

defining a set of character string subsequences, which set of subsequences comprises subsequences of at least two of the plurality of parental character strings;

providing a set of oligonucleotides corresponding to the set of character string subsequences;

annealing the set of oligonucleotides; and,

elongating one or more member of the set of oligonucleotides with a polymerase, thereby producing the initial library of nucleic acids.

78. A method of generating a library of biological polymers, the method comprising:

generating a diverse population of character strings in a computer, which character strings are generated by alteration of pre-existing character strings; and,

synthesizing the diverse population of character strings, which diverse population comprises the library of biological polymers.

79. The method of claim 78, wherein the alteration comprises recombination of the pre-existing character strings.

80. The method of claim 78, wherein the biological polymers are selected from nucleic acids, polypeptides and peptide nucleic acids.

81. The method of claim 78, further comprising selecting members of the library of biological polymers for one or more activity.

82. The method of claim 81, further comprising filtering an additional library or an additional set of character strings by subtracting the additional library or the additional set of character strings with members of the library of biological polymers which display activity below a desired threshold.

83. The method of claim 81, further comprising filtering an additional library or an additional set of character strings by biasing the additional library or the additional set of character strings with members of the library of biological polymers which display activity above a desired threshold.

This Sheet Has Been Replaced

84. An integrated system comprising a computer having a first data set comprising a first character string, a second data set comprising a second character string, software for aligning the first and second character strings, software for performing a genetic operation on the first or second character string, an output file comprising a third data set comprising a third character string, the third character string comprising character string subsequences from the first and second character strings, and an oligonucleotide sequence output file comprising a plurality of overlapping oligonucleotide sequences corresponding to the third character string.

85. The integrated system of claim 84, the system further comprising an oligonucleotide synthesis machine for synthesizing the plurality of overlapping oligonucleotides.

86. The integrated system of claim 84, further comprising a plurality of oligonucleotides encoded by the plurality of overlapping oligonucleotide sequences, which oligonucleotides, when incubated in one or more cycles of chain extension, produce a third nucleic acid encoded by the third character string.

87. The integrated system of claim 84, wherein the system further comprises a program with an instruction set for applying one or more genetic operator to the first or second character string, or to any other character string.

88. The integrated system of claim 84, wherein the system further comprises a program with an instruction set for applying one or more genetic operator to the first or second character string, or to any other character string, wherein the genetic operator is selected from: a mutation, a multiplication, a fragmentation of the string or strings, a crossover between one or more strings, a ligation of strings, an elitism calculation, an alignment, a calculation of sequence homology or sequence similarity, a recursive use of one or more genetic operator for evolution of character strings, randomness, a deletion mutation, an insertion mutation, and death.

