(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2004-507247
(P2004-507247A)
(43)公表日　平成16年3月11日(2004.3.11)

(51)Int.Cl.[7]	F I	テーマコード（参考）
C12N 15/09	C12N 15/00 ZNAA	4B024

審査請求 未請求　予備審査請求 有　（全 161 頁）

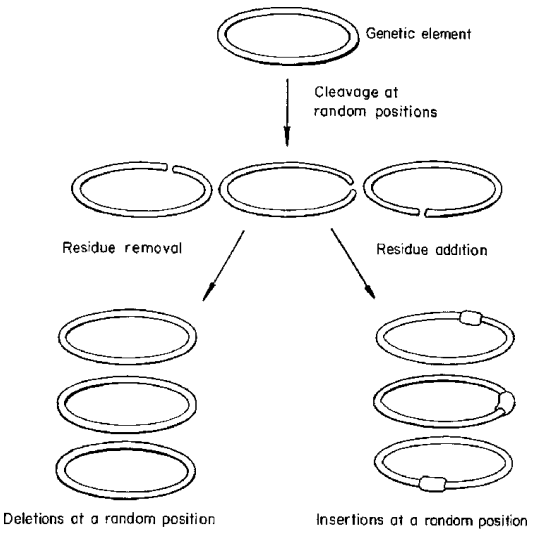
(21)出願番号	特願2002-522312(P2002-522312)
(86)(22)出願日	平成13年8月17日(2001.8.17)
(85)翻訳文提出日	平成15年2月17日(2003.2.17)
(86)国際出願番号	PCT/US2001/025788
(87)国際公開番号	WO2002/016642
(87)国際公開日	平成14年2月28日(2002.2.28)
(31)優先権主張番号	60/226,477
(32)優先日	平成12年8月18日(2000.8.18)
(33)優先権主張国	米国(US)

(71)出願人　503064637
インテグリジェン，　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４９４９，　ノバト，　ユニット　６，　ディジタル　ドライブ　４２
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100062409
弁理士　安村　高明
(74)代理人　100113413
弁理士　森下　夏樹

最終頁に続く

(54)【発明の名称】ＤＮＡ末端改変を使用する指向された分子進化のための方法および組成物

(57)【要約】
図３に示されるように、指向された進化のための方法が記載され、ここで、遺伝的要素は、ランダムに切断され、ポリヌクレオチドの欠失もしくは付加または両方が、付加または欠失を有する関連した遺伝的要素のライブラリーを生成することを可能にする。対応するライブラリー集団もまた記載される。これらのプロセスは、遺伝子の指向された進化に必要である配列空間の重要なサンプリングを可能にする。目的の遺伝的要素において非常に小さいヌクレオチド欠失を行うための方法が、さらに記載される。



【特許請求の範囲】
【請求項１】
遺伝的要素の配列中の異なる位置にヌクレオチド欠失を有するポリヌクレオチド配列のライブラリーを生成するための方法であって、該方法は、以下の工程：
（ａ）該遺伝的要素を含む複数コピーの環状ポリヌクレオチドを、ランダムな切断に供して、複数の線状ポリヌクレオチドを獲得する工程であって、該ポリヌクレオチドの各々は、少なくとも１つの３’末端および５’末端を有する、工程；および
（ｂ）工程（ａ）由来の該ポリヌクレオチドを、該ポリヌクレチドの該末端の１つから少なくとも１つのヌクレオチドを取り除くプロセスに供して、欠失ポリヌクレオチド配列のライブラリーを生成する工程であって、該ライブラリーは、異なるランダムな位置に欠失を有する複数の欠失ポリヌクレオチド配列を含む、工程、
を包含する、方法。
【請求項２】
工程（ｂ）由来の前記ポリヌクレオチドが、前記３’末端および５’末端を互いに共有結合するプロセスに供される、請求項１に記載の方法。
【請求項３】
前記ポリヌクレオチドのライブラリーが、目的の機能について選択するプロセスにさらに供される、請求項１に記載の方法。
【請求項４】
前記切断がエンドヌクレアーゼで生じる、請求項１に記載の方法。
【請求項５】
前記エンドヌクレアーゼがＳ１である、請求項４に記載の方法。
【請求項６】
前記欠失ポリヌクレオチドのライブラリーが、少なくとも５個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな欠失を有する請求項１に記載の方法。
【請求項７】
前記欠失ポリヌクレオチドのライブラリーが、少なくとも１０個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな欠失を有する請求項１に記載の方法。
【請求項８】
前記欠失ポリヌクレオチドのライブラリーが、少なくとも３０個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな欠失を有する請求項１に記載の方法。
【請求項９】
前記複数コピーの環状ポリヌクレオチドの組成物が、前記遺伝的要素に対する天然に存在するホモログを含まない、請求項１に記載の方法。
【請求項１０】
工程（ａ）および（ｂ）が繰り返される、請求項１に記載の方法。
【請求項１１】
工程（ｂ）が欠失の位置にヌクレオチドを挿入するためのプロセスをさらに含む、請求項１に記載の方法。
【請求項１２】
工程（ｂ）で１～３個のヌクレオチドが欠失されている、請求項１に記載の方法。
【請求項１３】
工程（ｂ）で５０～１００個のヌクレオチドが欠失されている、請求項１に記載の方法。
【請求項１４】
実質的に純粋な組成物であって、該組成物は、異なる３’末端および５’末端を各々有する複数の線状ポリヌクレオチドのライブラリーを含むが、該線状ポリヌクレオチドの各々は、環状にされる場合、他のポリヌクレオチドと同一である、組成物。

【請求項１５】
前記ライブラリーが、異なる３’末端および５’末端を有する少なくとも５個のポリヌクレオチドを含む、請求項１４に記載の組成物。
【請求項１６】
少なくとも２個の欠失ポリヌクレオチドのライブラリーを含む実質的に純粋な組成物であって、該ポリヌクレオチドは、異なるランダムな欠失を有することによってのみ各々他と異なる、組成物。
【請求項１７】
前記欠失ポリヌクレオチドが、欠失の位置に挿入された少なくとも１個のヌクレオチドを更に含む、請求項１６に記載の実質的に純粋な組成物。
【請求項１８】
前記ライブラリーが、少なくとも５個のポリヌクレオチドを有する、請求項１６に記載の組成物であって、該ポリヌクレオチドは、異なるランダムな欠失を有することによってのみ各々他と異なる、組成物。
【請求項１９】
遺伝的要素中のランダムな位置にヌクレオチド付加を有するポリヌクレオチド配列のライブラリーを生成するための方法であって、該方法は、以下の工程：
（ａ）該遺伝的要素を含む複数コピーの環状ポリヌクレオチドの組成物を、ランダムな切断に供して、複数の線状ポリヌクレオチドを獲得する工程であって、該ポリヌクレオチドの各々は、少なくとも１つの３’末端および５’末端を有する、工程；および
（ｂ）工程（ａ）由来の該ポリヌクレオチドを、該ポリヌクレチドの該末端の１つに少なくとも１つのヌクレオチドを付加するプロセスに供して、付加ポリヌクレオチド配列のライブラリーを生成する工程であって、該ライブラリーは、異なるランダムな位置に付加を有する複数の付加配列を含む、工程、
を包含する、方法。
【請求項２０】
工程（ｂ）由来の前記付加ポリヌクレオチドが、前記３’末端および５’末端を互いに共有結合するプロセスに供される、請求項１９に記載の方法。
【請求項２１】
前記ポリヌクレオチドのライブラリーを、目的の機能について選択するプロセスに供する工程をさらに包含する、請求項１９に記載の方法。
【請求項２２】
前記切断がエンドヌクレアーゼで生じる、請求項１９に記載の方法。
【請求項２３】
前記エンドヌクレアーゼがＳ１である、請求項２２に記載の方法。
【請求項２４】
前記付加ポリヌクレオチドのライブラリーが、少なくとも５個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな付加を有する、請求項１９に記載の方法。
【請求項２５】
前記付加ポリヌクレオチドのライブラリーが、少なくとも１０個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな付加を有する、請求項１９に記載の方法。
【請求項２６】
前記付加ポリヌクレオチドのライブラリーが、少なくとも３０個の個々のポリヌクレオチドを含み、該ポリヌクレオチドの各々は、他のポリヌクレオチドとは異なる位置にランダムな付加を有する、請求項１９に記載の方法。
【請求項２７】
前記複数コピーの環状ポリヌクレオチドの組成物が、前記遺伝的要素に対する天然に存在するホモログを含まない、請求項１９に記載の方法。

【請求項２８】
工程（ａ）および（ｂ）が繰り返される、請求項１９に記載の方法。
【請求項２９】
工程（ｂ）が付加の位置でヌクレオチドを欠失させるためのプロセスを含む、請求項１９に記載の方法。
【請求項３０】
工程（ｂ）で１～３個のヌクレオチドが付加されている、請求項１９に記載の方法。
【請求項３１】
工程（ｂ）で３～５０個のヌクレオチドが付加されている、請求項１９に記載の方法。
【請求項３２】
工程（ｂ）で５０～１００個のヌクレオチドが付加されている、請求項１９に記載の方法。
【請求項３３】
少なくとも２個の付加ポリヌクレオチドのライブラリーを含む実質的に純粋な組成物であって、該ポリヌクレオチドは、異なるランダムな付加を有することによってのみ各々他と異なる、組成物。
【請求項３４】
少なくとも５個の付加ポリヌクレオチドのライブラリーを含む実質的に純粋な組成物であって、該ポリヌクレオチドは、異なるランダムな付加を有することによってのみ各々他と異なる、組成物。
【請求項３５】
ポリヌクレオチドの末端から短い欠失を生成するための方法であって、該方法は、１０～５００ｍＭ塩の存在下で０℃～２４℃の温度にて、エキソヌクレアーゼと共に該ポリヌクレオチドの集団をインキュベートし、それにより、該ポリヌクレオチドの少なくとも１つの末端からの１～１００残基の欠失を含むポリヌクレオチドの集団を生成することによる、方法。
【請求項３６】
前記ポリヌクレオチドが二本鎖である、請求項３５に記載の方法。
【請求項３７】
前記エキソヌクレアーゼがエキソヌクレアーゼＩＩＩである、請求項３５に記載の方法。
【請求項３８】
前記二本鎖核酸が、平滑末端を生成するために一本鎖エンドヌクレアーゼと共にインキュベートされる、請求項３６に記載の方法。
【請求項３９】
請求項３５に記載の方法であって、さらに、前記末端に欠失を含む前記得られたポリヌクレオチドの集団が、少なくとも第２の末端に共有結合され、内部位置に欠失を含むポリヌクレオチドの集団を生成する、方法。
【請求項４０】
前記一本鎖エンドヌクレアーゼがＳ１ヌクレアーゼである、請求項３８に記載の方法。
【請求項４１】
共有結合から生じた前記ポリヌクレオチドが、環状ポリヌクレオチドである、請求項３９に記載の方法。
【請求項４２】
前記ポリヌクレオチドの集団が、該ポリヌクレオチドの少なくとも１つの末端からの１～５０残基の欠失を含む、請求項３５に記載の方法。
【請求項４３】
前記ポリヌクレオチドの集団が、該ポリヌクレオチドの少なくとも１つの末端からの１～３０残基の欠失を含む、請求項３５に記載の方法。
【請求項４４】
少なくとも２個のポリヌクレオチドの実質的に純粋な組成物であって、該ポリヌクレオチ

ドは、各々２つの末端を有し、そして１つの末端または両方の末端で１～１００残基の異なる欠失を有することによってのみ各々互いに異なる、組成物。
【請求項４５】
前記ポリヌクレオチドの組成物が、１つの末端または両方の末端で１～５０残基の欠失によって互いに異なる、請求項４４に記載の組成物。
【請求項４６】
前記ポリヌクレオチドの組成物が、１つの末端または両方の末端で１～３０残基の欠失によって互いに異なる、請求項４４に記載の組成物。
【請求項４７】
前記ポリヌクレオチドの組成物が、１つの末端または両方の末端で１～１０残基の欠失によって互いに異なる、請求項４４に記載の組成物。
【請求項４８】
少なくとも２つのポリヌクレオチドの実質的に純粋な組成物であって、該組成物の各々は、該ポリヌクレオチド内の特定の内部位置での１～１００残基の欠失によってのみ互いに異なる、組成物。
【請求項４９】
前記ポリヌクレオチドが、前記特定の内部位置での１～５０残基の欠失によって互いに異なる、請求項４８に記載の実質的に純粋な組成物。
【請求項５０】
前記ポリヌクレオチドが、前記特定の内部位置での１～３０残基の欠失によって互いに異なる、請求項４８に記載の実質的に純粋な組成物。
【請求項５１】
前記ポリヌクレオチドが、前記特定の内部位置での１～１０残基の欠失によって互いに異なる、請求項４８に記載の実質的に純粋な組成物。
【発明の詳細な説明】
【０００１】
（発明の分野）
本発明は、指向された進化に関し、遺伝子操作およびタンパク質操作に適用することができる方法を包含する。指向された進化を用いて、遺伝子またはタンパク質の機能を改善または変化させることを目標に、遺伝子配列を進化させる。指向された進化は、医薬品開発、バイオレメディエーション（ｂｉｏｒｅｍｅｄｉａｔｉｏｎ）、バイオリーチング（ｂｉｏｌｅａｃｈｉｎｇ）、および化学産業を含むが、これらに限定されない多くの分野に適用することができる。
【０００２】
（発明の背景）
近年、インビトロでの進化プロセスをシミュレートし、それにより特定遺伝子中で遺伝子変化を誘導して、それらの機能を変化または改善させる試みがなされてきた。遺伝子を変化させる技術がここ数年の間に知られてきたが、一般に、これらの方法を成功させるために、コードタンパク質構造および機能に関する詳細な特徴が必要とされた。ＤＮＡシャッフリング技術は、この障壁をある程度まで克服し、ここ数年でいくつかの遺伝子を首尾よく進化させるために適用されてきた［Ｍｉｎｓｈｕｌｌ＆Ｓｔｅｍｍｅｒ、米国特許第５，８３７，４５８号（１９９８年）］。
【０００３】
天然の進化は、環境中の遺伝子に関して、何百万年もかけて起こった。インビトロ進化は、数日または数週間で天然プロセスを模倣することを試みるものである。インビトロ戦略が成功するためには、進化理論のいくつかの様相を理解しなくてはならない。第１に、配列空間の概念が、既定長のタンパク質の考え得る配列の総数を規定する［Ｋａｕｆｆｍａｎ，（１９９３）］。したがって、
【０００４】
【数１】

$$S = 20^N$$

であり、式中、配列空間Ｓは、考え得る配列数であり、Ｎは、タンパク質長である。インビトロ進化実験では、最も改善または変化した活性を有するタンパク質の分画を特定するために、目的のタンパク質のＳ配列を探索することが最も望ましい。中程度の５０個のアミノ酸を有するタンパク質は、$２０^{５０}$個の考え得る異なる配列のＳを有することがすぐに理解でき、その数は、現在の分子生物学技術による分析に関して事実上無限である。第２に、ほとんどのアミノ酸変化が、タンパク質にとって有害であることが明らかである。これらの変化は、タンパク質を不活性にし得るか、適切なフォールディングの崩壊を引き起こし得るか、またはインビボでのタンパク質もしくはｍＲＮＡに対する不安定性を引き起こし得る。有害な変異に対する有利な変異の平均比率は、$１０^{５}$分の１であると推定されている［Ｒａｄｍａｎら、Ａｎｎ，Ｎ．Ｙ．Ａｃａｄ．Ｓｃｉ．８７０：１４６－５５（１９９９）］。これに関して、変異率は、それらの機能を改善するために遺伝子を変異させる場合に重要なパラメータである。変異率が高すぎると、有害な変異が、有利な変異とともにｃｉｓで生じ、その状態により、有利な変異を含有する得られたタンパク質が、付随する有害な変異に起因して不活性であるため、有利な変異を有する遺伝子を特定することが不可能となる。第３に、より高い変異率の結末を克服するために、相同組換えを利用して、二重クロスオーバー事象により有害な変異を除去し得る。第４に、任意のインビトロ進化技術は、タンパク質の機能を改善または変化させる配列を特定するために、選択スクリーニングを必要とする。

【０００５】

現在の分子進化の主な障壁は、目的のタンパク質に関する配列空間を効率的に探索することができないことである。これに関して、１つより多い残基が異なる配列を生成および特定する能力が非常に重要であり、ここで、これらの配列は、タンパク質機能に対してさらなる効果を有し得る。このさらなる効果は、アミノ酸相互依存にて記載し得る。例えば、残基ｉに単一の変異を有するタンパク質は、ｊでの付随する変異もまた存在しない限り、検出可能な機能のいかなる増大をも有さない場合がある。これに関して、進化が成功するためには、標的配列の考え得る２変異改変体すべてが、サンプリングされ、機能改善に関して試験されるべきである。一般に、長さＮのタンパク質のＲ変異体数は、以下：

【０００６】

【数２】

$$S_R = \frac{20^R N!}{(N-R)!R!}$$

によって表され、ここで、Ｒは、変異改変体の数であり、２０は、各位置での考え得るアミノ酸数を表す。したがって、長さ５０のタンパク質に関しては、４９０，０００個の異なる２変異改変体が存在する。

【０００７】

配列空間のこれらの統計学的解析において、臨界値は、タンパク質長である（すなわち、Ｒ変異改変体数は、タンパク質長に依存する）。しかし、本質的には、任意の目的のタンパク質長は、三次元空間のアミノ酸残基の整列ほど、その機能にとって重要ではない場合がある。実際に、「触媒作業空間（ｃａｔａｌｙｔｉｃ　ｔａｓｋ　ｓｐａｃｅ）の仮説的概念は、この原理を説明すると提唱されている（Ｋａｕｆｆｍａｎ，１９９３）。タンパク質長Ｎを変化させることなくアミノ酸残基を変化させることは、Ｎを増加または減少させるいくつかの方法において、タンパク質の三次元構造に影響を及ぼし得ない。あるいは、Ｎの変化は、タンパク質の生物学的機能を全く変化させ得ない。相同タンパク質の事実上任意のファミリーの分析により、メンバーは、時には実質的な挿入または欠失を伴っ

て異なる長さを有するが、区別不可能な生物学的機能を保持し得ることが明らかである。したがって、上記式は、おそらく、生物学的機能の改善または変化に関して探索する場合、スクリーニングされるべき様々なＲ変異改変体の正確な見解を提供しない。研究室では、タンパク質のＲ変異体の近隣と、ヌクレオチドがあらゆる位置で付加または欠失される多くの変異体数の全てを探索することが最適である。

【０００８】

欠失の場合には、Ｄ変異体欠失数は、以下：

【０００９】

【数３】

$$S_D = \frac{N!}{(N-D)!\,D!}$$

で表され、式中、Ｎは、タンパク質の初期長であり、Ｄは、欠失が起きる位置の数である。アミノ酸付加の場合には、考え得るすべての付加に関する同様の式が、２０個のアミノ酸のいずれかが任意の位置で付加され得るという事実を説明し：

【００１０】

【数４】

$$S_A = \frac{20^A\,N!}{(N-D)!\,A!}$$

であり、式中、Ａは、考え得る付加変異体数である。付加および欠失変異体の場合では、これらの式はともに、唯一のアミノ酸が各位置で付加または欠失されると仮定している。しかし、インビトロ分子進化に関しては、各位置で付加または欠失した１個、２個、３個、・・・Ｃ個全ての数のアミノ酸を探索することが最適である。したがって、欠失変異体に関して、各位置で欠失された可変アミノ酸を有する配列数は、以下：

【００１１】

【数５】

$$S_{CD} = \frac{C_D\,N!}{(N-D)!\,D!}$$

であり、式中C_Dは、各位置で欠失したアミノ酸数を表し、Ｄは、欠失が起きる位置の数である。付加変異体に関して、式は、以下：

【００１２】

【数６】

$$S_{CA} = \frac{C_A\,20^A\,N!}{(N-A)!\,A!}$$

となり、式中、C_Aは、各位置での付加したアミノ酸数であり、Ａは、付加が生じる位置の数である。

【００１３】

現在の分子生物学技術のみが、空間全体の分画を、目的のタンパク質に関して生成およびサンプリングすることを可能にするため、生成されるべき実験空間について記載する式を定義することができる。この式はまた、ライブラリー構築技術の改善のモニタリングを可

能とし、タンパク質機能に関連する空間の解析を可能にする。実験的に探索されるべき空間全体を、以下：

【００１４】

【数７】

$$S_{EX} = S_R S_{CD} S_{CA}$$

として定義することができ、ここで、アミノ酸は、様々な組合せおよび順列で、他の残基に対して変異される（S_R）か、欠失される（S_{CD}）か、または付加される（S_{CA}）。もちろん、現在の分子化学技術により、ライブラリーを創出することが可能であり、ここでＲ＝１であればS_R＝Ｎであり、Ｄ＝１であればS_{CD}＝Ｎであり、Ａ＝１であればS_{CA}＝２０Ｎである。続いて、Ｎ＝５０のタンパク質について、この仮説的ライブラリーは、$20^* N^3 = 2.5 \times 10^6$個の異なる配列を有し、ここで１つの位置での変化、欠失および付加の順列すべてが表される。

【００１５】

タンパク質進化に関する配列空間に関する上記議論は、種々の方法で、進化配列のインビトロ操作に適用してもよい。本質的に、酵素ファミリーの異なる触媒活性の進化は、大きく２つのカテゴリー：１）活性部位のアミノ酸は同一であるが、構造フォールド（ｆｏｌｄ）の差異が酵素に異なる基質特異性をもたせるもの［Ｐｅｒｏｎａ＆Ｃｒａｉｋ，Ｊ．Ｂｉｏｌ．Ｃｈｅｍ．２７２：２９９８７－９０（１９９７）］、および２）酵素構造は同一であるが、活性部位の残基の差異が、酵素に異なる反応を触媒させるもの［Ｂａｂｂｉｔｔ＆Ｇｅｒｌｔ，Ｊ．Ｂｉｏｌ．Ｃｈｅｍ．２７２：３０５９１－４（１９９７）］に類別することができる。前者の例は、セリンプロテアーゼファミリーであり、後者の例は、エノラーゼスーパーファミリーである。

【００１６】

これらのカテゴリー間の差異は、取るに足らないことのように思えるが、それらは、分子進化の方法および配列空間の概念にとって重要な意味を持つ。類似の構造フォールドを有するファミリーの酵素に関して、触媒機構のために触媒活性部位が同一残基を必要とするようであるため、タンパク質長全体にわたって配列空間をサンプリングする分子進化アプローチは、酵素の特異性を変化させる最適な戦略であるようである。しかし、第２の型の酵素に関しては、タンパク質の全長にわたって探索する配列空間を増加することはおそらく必要ではない。それよりも、重要な触媒ドメインの配列空間サンプリングを増加させることが、分子進化プロセスを最適化する。これに関して、全遺伝子配列にわたって展開される20^5個の配列空間をサンプリングするよりも、２０個のアミノ酸それぞれに対して重要な５個のアミノ酸を変化させ、このより限定された空間（20^5個）をサンプリングすることがより良好である。さらに、重要な領域における考え得る付加または欠失の変異体と同数をサンプリングすることはまた、インビトロ進化プロトコルの考え得る成功に寄与する。したがって、第２の型の酵素ファミリーに属する遺伝子の分子進化を最適化する方法は、非常に重要であり、かつ堅固な技術。

【００１７】

遺伝子ドメインの交換（ｓｗａｐｐｉｎｇ）は、生体分子の新たな機能または改善された機能を進化させるために効率的な手段である。単一ヌクレオチド残基の変化は、遺伝子およびタンパク質機能に影響を及ぼし得るものの、遺伝子中の複数残基の大量の交換が、タンパク質機能に劇的に影響を及ぼし得る。例えば、Ｅ．ｃｏｌｉおよびＳａｌｍｏｎｅｌｌａは、高度に関連した細菌種であるが、これらの遺伝的内容の差異は、単一残基の変化ではなく遺伝子交換事象にほぼ完全に起因する。さらに、大量のＤＮＡの変換が遺伝子を創出する交換事象は、凝固カスケードのような経路で数回起きたと考えられ、ならびに転位により新規転写カセットを創出することが考えられる［Ｂｅｌｌ，（１９９７）；Ｐａｔｔｈｙ，（１９９９）］。

【００１８】

天然に存在する分子進化の周知の例は、免疫系での抗体産生の基礎を成すものである。哺乳類の前リンパ球（ｐｒｅ－ｌｙｍｐｈｏｃｙｔｅ）では、天然の分子進化が日常的に首尾よく生じる。抗体は、混乱させるアレイの種々の抗原を結合することが可能であるが、類似のアミノ酸配列および二次構造を有している。抗体遺伝子は、遺伝子セグメントとして生殖系列で整列される（図２）。リンパ球成熟中、これらのセグメント（可変または「Ｖ」、多様性または「Ｄ］、および結合性または「Ｊ」と呼ぶ）は、Ｖ（Ｄ）Ｊ組換えと称したプロセスにおいて互いに並置され、機能的抗体またはＴ細胞受容体遺伝子を創出する。複数のＶセグメント、Ｄセグメント、およびＪセグメントは、相当量の多様性を可能とし、したがって、種々の抗原結合特異性が、リンパ球の最終レパートリーで創出される。この機構により創出された多様性を、コンビナトリアル多様性（ｃｏｍｂｉｎａｔｏｒｉａｌ　ｄｅｖｅｒｓｉｔｙ）と称する。別の型の多様性もまた、Ｖ（Ｄ）Ｊ組換え中に創出され、これは、コンビナトリアル多様性と同様に重要である［Ｄａｖｉｓ＆Ｂｊｏｒｋｍａｎ，Ｎａｔｕｒｅ　３３４：３９５－４０２（１９８８）］。この多様性を結合多様性と称し、それは、ヌクレオチドが遺伝子セグメントの結合部で損失または獲得される場合に創出される。重要なことに、これらの結合部は、抗原と接触する抗体分子の領域をコードし、したがって、この型の多様性は、多様性であるが機能的な免疫系を創出するために重要である。

【００１９】

免疫系によって利用される２つの型の多様性は、分子進化の実施に関して、以下の方法で特徴付けられ得る。免疫グロブリン遺伝子におけるコンビナトリアル多様性の生成は、複数の機能性Ｖ遺伝子セグメント、Ｄ遺伝子セグメントおよびＪ遺伝子セグメントを提供することによって配列空間全体のサンプリングが可能となり、その各メンバーは、配列がわずかに異なるが、依然としてセグメントのファミリーの他のメンバーに対して相同である。これに関して、Ｖ遺伝子セグメント、Ｄ遺伝子セグメントおよびＪ遺伝子セグメントの組合せ再配列は、新規抗体遺伝子を生成するために、「ドメイン交換」事象として機能する。結合多様性の生成は、連結されるべきＤＮＡの末端にランダムヌクレオチドを付加または欠失させる機構により、抗原を接触するのに重要な残基で配列空間のより大きな局所的サンプリングを可能にする。

【００２０】

遺伝子進化に関する上述の問題、すなわち莫大な配列空間の探索における困難、ランダム変異誘発の有害な変異の優勢、およびアミノ酸相互依存に起因して、研究室で機能的配列空間を探索する堅固な方法を考案することが困難であった。ライブラリー形式で変異タンパク質を創出するのに現在広範に使用される方法は、誤りがちなポリメラーゼ連鎖反応［Ｃａｌｄｗｅｌｌ＆Ｊｏｙｃｅ，（１９９２）；Ｇｒａｍら、Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　８９：３５７６－８０（１９９２）］、およびカセット変異誘発［Ａｒｋｉｎ＆Ｙｏｕｖａｎ，Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　８９：７８１１－５（１９９２）；Ｈｅｒｍｅｓら、Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　８７：６９６－７００（１９９０）；Ｏｌｉｐｈａｎｔら、Ｇｅｎｅ　４４：１７７－８３（１９８６）；Ｓｔｅｍｍｅｒ＆Ｍｏｒｒｉｓ，Ｂｉｏｔｅｃｈｎｉｑｕｅｓ　１３：２１４－２０（１９９２）］であり、これらの方法では最適化される特定領域が、合成的に変異誘発されたオリゴヌクレオチドで置き換えられる。あるいは、宿主細胞の変異誘発遺伝子（ｍｕｔａｔｏｒ）系統が、変異頻度を加算するために使用されてきた［Ｇｒｅｅｎｅｒら、Ｍｏｌ　Ｂｉｏｔｅｃｈｎｏｌ　７：１８９－９５（１９９７）］。各場合において、「変異雲（ｍｕｔａｎｔ　ｃｌｏｕｄ）」［Ｋａｕｆｆｍａｎ，（１９９３）］は、元の配列中のある種の部位付近に生成される。

【００２１】

誤りがちなＰＣＲは、長い配列にわたって低レベルの点変異をランダムに導入するために、低忠実度の重合条件を使用する。誤りがちなＰＣＲをまた使用して、未知の配列のフラグメントの混合物を変異誘発し得る。誤りがちなＰＣＲは、ｄＩＴＰの存在下でｄＮＴＰの個々の濃度を変化させることによって、遺伝子をランダムに変異させることができる［

Ｃａｌｄｗｅｌｌ＆Ｊｏｙｃｅ，（１９９２）：Ｌｅｕｎｇ＆Ｍｉｙａｍｏｔｏ，Ｎｕｃｌｅｉｃ Ａｃｉｄｓ Ｒｅｓ １７：１１７７－９５（１９８９）；Ｓｐｅｅら、Ｎｕｃｌｅｉｃ Ａｃｉｄｓ Ｒｅｓ ２１：７７７－８（１９９３）］。

【００２２】

しかし、コンピュータシミュレーションは、点変異単独では、多くの場合緩やかすぎて、連続した配列進化に必要なブロック変化を可能にし得ないことを示唆した。公開されている誤りがちなＰＣＲプロトコルは一般に、０．５～１．０ｋｂより長いＤＮＡフラグメントの信頼性高い増幅に不適切であり、それらの実際の適用は限られる。さらに、誤りがちなＰＣＲの繰り返しのサイクルは、中立変異の蓄積を引き起こし、それは例えば、タンパク質を免疫原性にし得る。

【００２３】

オリゴヌクレオチド指向性変異誘発では、短い配列が、合成的に変異誘発されたオリゴヌクレオチドで置き換えられる。このアプローチは、遠位変異の組合せを生成せず、したがって有意にコンビナトリアルではない。莫大な配列長に対して限定されたライブラリーサイズは、タンパク質最適化のために多数回の選択が回避できないことを意味する。合成オリゴヌクレオチドによる変異誘発は、各回の選択後に個々のクローンを配列決定し、続いてファミリーに類別し、単一ファミリーを任意に選択し、そのファミリーをコンセンサスモチーフへと縮小させる必要があり、そのコンセンサスモチーフは、再合成され、単一遺伝子に再挿入され、続いてさらに選択される。このプロセスは、統計学的ボトルネックを構成し、それは多数回の変異誘発について集中的に労力を有し、また実用的ではない。

【００２４】

制限部位を組込むランダムプライマーまたは部分的縮重プライマーを利用する飽和変異誘発方法についても記載されている［Ｈｉｌｌら、Ｍｅｔｈｏｄｓ Ｅｎｚｙｍｏｌ １５５：５５８－６８（１９８７）；Ｏｌｉｐｈａｎｔら、Ｇｅｎｅ ４４；１７７－８３（１９８６）；Ｒｅｉｄｈａａｒ－Ｏｌｓｏｎら、Ｍｅｔｈｏｄｓ Ｅｎｚｙｍｏｌ ２０８：５６４－８６（１９９１）］。

【００２５】

「カセット」変異誘発は、変異タンパク質のライブラリーを創出するための別の方法である［Ｂｏｃｋら、米国特許第５，８３０，７２０号（１９９５年）；Ｃｈｒｉｓｔｏｕ＆ＭｃＣａｂｅ、米国特許第５，８３０，７２８号（１９９８年）；Ｈｉｌｌら、Ｍｅｔｈｏｄｓ Ｅｎｚｙｍｏｌ １５５：５５８－６８（１９８７）；Ｍｉｌｌｅｒら、米国特許第５，８３０，７４０号（１９９８年）；Ｓｈｉｒａｉｓｈｉ＆Ｓｈｉｍｕｒａ，Ｇｅｎｅ ６４：３１３－９（１９８８）；Ｓｔｅｍｍｅｒ＆Ｃｒａｍｅｒｉ、米国特許第５，８３０，７２１号（１９９８年）］。カセット変異誘発は典型的に、部分的にランダム化された配列で、鋳型の配列ブロック長を置き換える。したがって、得られ得る最大の情報内容は、カセットのランダム化された部分中のランダム配列の数に統計学的に限定される。

【００２６】

プロトコルもまた開発されており、それにより、オリゴヌクレオチドの合成は、非天然ホスホルアミダイトで「ドープ（ｄｏｐｅ）」され、ランダム変異誘発を標的化する遺伝子セクションのランダム化を生じる［Ｗａｎｇ＆Ｈｏｏｖｅｒ，Ｊ Ｂａｃｔｅｒｉｏｌ １７９：５８１２－９（１９９７）］。この方法により、ランダム置換率を維持しながら、位置の選択の制御が可能となる。

【００２７】

ＺａｃｃｏｌｏおよびＧｈｅｒａｒｄｉ（１９９９）は、ピリミジンヌクレオチシドアナログおよびプリンヌクレオチシドアナログを利用するランダム変異誘発方法について記載している［Ｚａｃｃｏｌｏ＆Ｇｈｅｒａｒｄｉ，Ｊ Ｍｏｌ Ｂｉｏｌ ２８５：７７５－８３（１９９９）］。この方法は、セファロスポリン、セフォタキシムに対する触媒速度の増加を伴う －ラクタマーゼを示す置換変異を達成するのに成功した。Ｃｒｅａは、「ウォークスルー（ｗａｌｋ ｔｈｒｏｕｇｈ）」方法を記載し、ここでは、既定のアミ

ノ酸が、あらかじめ選択された位置で標的配列に導入される［Ｃｒｅａ，米国特許第５，７９８，２０８号（１９９８年）］。

【００２８】

挿入変異および／または欠失変異により標的遺伝子を変異させる方法が開発されている。挿入変異は、ｓｔａｐｈｙｌｏｃｏｃｃａｌヌクレアーゼの内部に蓄積され得ることが実証された［Ｋｅｅｆｅら、Ｐｒｏｔｅｉｎ Ｓｃｉ ３：３９１－４０１（１９９４）］。開発された欠失変異誘発方法の例としては、エキソヌクレアーゼ（例えば、エキソヌクレアーゼＩＩＩまたはＢａｌ３１）の利用、または点欠失を組み込むオリゴヌクレオチド指向性欠失によるものが挙げられる［Ｎｅｒら、Ｎｕｃｌｅｉｃ Ａｃｉｄｓ Ｒｅｓ １７：４０１５－２３（１９８９）］。さらに、Ｌｉｅｔｚは、ランダム配列を有するオリゴヌクレオチドが挿入および欠失を誘導するためにＰＣＲと組合わせられ得る方法について記載する。この技術による機能の強化は示されておらず、過剰変異誘発（すなわち、ポリヌクレオチド１つ当たり多すぎる挿入または欠失を作製する）の容量が、この方法では重要である［Ｌｉｅｔｚ，米国特許第６，２５１，６０４号（２００１年）］。

【００２９】

インビトロでタンパク質を進化させるのに最も頻繁に使用される技術は、「ＤＮＡシャッフリング」として知られている。この方法では、遺伝子改変のライブラリーは、遺伝子の相同配列をフラグメント化し、そのフラグメントを互いにランダムにアニーリングさせ、ポリメラーゼを用いてオーバーハングを充填することにより創出される。次に、完全長遺伝子ライブラリーは、ポリメラーゼ連鎖反応（ＰＣＲ）により再構築される。この方法の有用性は、アニーリングの工程に生じ、それにより相同配列は、互いにアニーリングし得、両方の出発配列の特性を有する配列を産生し得る。実際に、この方法は、相同性であるが、いくつかの位置で有意な差異を含む２つ以上の遺伝子間の組換えに影響を及ぼす。いくつかの相同配列を用いるライブラリーの創出により、ランダムに変異した単一出発配列を用いる場合よりも多くの配列空間をサンプリングすることが可能となることを示した［Ｃｒａｍｅｒｉら、Ｎａｔｕｒｅ ３９１；２８８－９１（１９９８）］。この効果は、進化の年月が、異なる種のホモログ間の種々の有利な変異または中立の変異に関してすでに選択されていたという事実に起因するようである。ホモログから出発すると、次いで、スクリーニングされるべきライブラリーの創出において有害な変異数をかなり限定する。ホモログの有利な位置をコンビナトリアルに再配列すると、明らかに生化学反応を触媒するのに最適な二次タンパク質構造を可能にすることができる。得られた進化タンパク質は、出発配列各々に起因する明確な特徴を含有するようであり、選択後、劇的に改善された機能を生じる。

【００３０】

ＤＮＡシャッフリング技術の変法が考案された。一プロセスは、「付着伸長（ｓｔａｇｇｅｒｅｄ ｅｘｔｅｎｓｉｏｎ）」プロセス、またはＳｔＥＰと呼ばれる。伸長プライマーにより創出されるフラグメントのプールを再構築するのではなく、完全長遺伝子を鋳型の存在下で直接アセンブリする。ＳｔＥＰは、変性、続く極端に省略されたアニーリング／伸長工程の繰り返しサイクルから構成される。各サイクルでは、伸長フラグメントを、相補性に基づいた種々の鋳型にアニーリングさせ得、さらに少し伸長して、「組換えカセット」を創出することができる。この鋳型転換により、ほとんどのポリヌクレオチドは、異なる親遺伝子由来の配列を含有する（すなわち、新規組換え体である）。このプロセスは、完全長遺伝子が形成するまで繰り返される。それに続いて、任意の遺伝子増幅工程を行うことができる［Ａｒｎｏｌｄら、米国特許第６，１７７，２６３号（２００１年）］。

【００３１】

別の技術では、初期ＤＮＡのフラグメント化は、標的遺伝子の付加物（ａｄｄｕｃｔ）形成を誘導することで、伸長反応中のポリメラーゼの早発性終結により達成することができる［Ｓｈｏｒｔ、米国特許第５，９６５，４０８号（１９９９年）］。異なる技術では、融合遺伝子のライブラリーを生成するために２つのホモログの各々に漸増性切断を誘導す

ることにより、ライブラリーが創出され、ライブラリーの各々が、各ホモログからのドメインを含有する［Ｏｓｔｅｒｍｅｉｅｒら、Ｎａｔ．Ｂｉｏｔｅｃｈｎｏｌ．１７：１２０５－９（１９９９）］。このアプローチの利点は、以前の方法のアニーリング工程が省略されるため、出発配列間の有意な相同性が必要とされないことである。しかし、選択技術をライブラリーに適用する後に、この技術の変法が実際に改善された遺伝子機能の生成を引き起こすか否かは明らかではない。

【００３２】

種々の生物体由来の遺伝子の対立遺伝子を用いる遺伝子シャッフリングのこれまで記載されてきた方法は、コンビナトリアル多様性を生じさせるが、出発配列に見出される相同性により限定される。さらに、これらの方法は、抗体遺伝子セグメントのＶ（Ｄ）Ｊ結合により形成される結合多様性を生成する機構を提供しない。本発明は、指向性様式またはランダム様式のいずれかでタンパク質配列または核酸配列由来の残基を付加および欠失させることで、結合多様性に類似した機構を利用する。本発明はまた、コンビナトリアルＶ（Ｄ）Ｊ組換えにより生成されるコンビナトリアル多様性に類似した「遺伝子交換」事象を提供する。これは、遺伝子をインビトロで進化させる手段を大いに強化する。

【００３３】

（発明の要旨）

本発明は、以下の工程：

（ａ）ポリヌクレオチド中にランダムにヌクレオチド残基を付加または欠失させて、付加または欠失を含有するポリヌクレオチドのライブラリーを生成する工程；および

（ｂ）必要に応じて、工程（ａ）のポリヌクレオチドのプールを、所望の機能または特徴をコードするポリヌクレオチドを特定することが可能な選択手順に供する工程、

による核酸配列の指向された分子進化を包含する。工程（ａ）および（ｂ）は、必要に応じて反復し得る。本発明の方法により生成されるライブラリーもまた、記載され、そして意図される。

【００３４】

独自に、本発明は、タンパク質二次構造に有意に影響を及ぼす配列を含む配列空間のサンプリングを可能とし、したがって進化した遺伝子中の変化した機能または改善した機能を特定する蓋然性を高めることが可能である。さらに、本発明は、他の現在の技術によりサンプリングすることができない配列空間のサンプリングを可能にする。さらに、本発明を用いて創出したポリヌクレオチドのライブラリーは、他の現在の技術を利用して得ることができない。

【００３５】

いくつかの方法および組成物が、以下に記載されそして意図される。本発明の１つの方法は、遺伝的要素の配列中の異なる位置にヌクレオチド欠失を有するポリヌクレオチド配列のライブラリーを生成するための方法であって、以下の工程：

（ａ）この該遺伝的要素を含む複数コピーの環状ポリヌクレオチドを、ランダムな切断に供して、複数の線状ポリヌクレオチドを獲得する工程であって、このポリヌクレオチドの各々は、少なくとも１つの３’末端および５’末端を有する、工程；および

（ｂ）工程（ａ）由来の上記ポリヌクレオチドを、このポリヌクレチドのＤＮＡ末端の１つから少なくとも１つのヌクレオチドを取り除くプロセスに供して、欠失ポリヌクレオチド配列のライブラリーを生成する工程であって、このライブラリーは、異なるランダムな位置に欠失を有する複数の欠失ポリヌクレオチド配列を含む、工程、を包含する。さらに所望される場合、工程（ｂ）由来のポリヌクレオチドが、上記３’末端および５’末端を互いに共有結合するプロセスに供され得、そして上記ポリヌクレオチドのライブラリーは、目的の機能について選択するプロセスにさらに供され得る。欠失ポリヌクレオチドのライブラリーは、２個より多くかまたはそれ以上、例えば、少なくとも１０個、２０個もしくは３０個またはそれ以上の欠失を含んでもよく、あるいはさらには各々が他と異なる位置にランダム欠失を有する５０～１００個の別個のヌクレオチドが得られてもよい。作製される欠失数は、出発材料および技術者の目標に依存する。いくつかの実施形態では、欠

失ポリヌクレオチドのライブラリーは、少なくとも１個、２個、３個、４個、もしくは５個またはそれ以上の別個のヌクレオチドの非常に短い欠失を含む。異なる実施形態では、ライブラリーは、５０～１００個またはそれ以上のヌクレオチドのより大きな欠失を含んでもよい。別の実施形態では、環状ポリヌクレオチドの複数コピーの組成物は、遺伝的要素に対する天然に存在するホモログを含まない。さらに、工程（ａ）および（ｂ）は、任意に反復され得る。別の任意の方法は、工程（ｂ）で欠失位置にヌクレオチドを挿入するプロセスを包含する。

【００３６】

各々異なる３’末端および５’末端を有する、複数の（好ましくは２個より多い、より好ましくは５個より多い、最も好ましくは１０個より多い）線状ポリヌクレオチドのライブラリーを含む実質的に純粋な組成物であって、環化される場合には各々の線状ポリヌクレオチドが他のものと同一である組成物について、記載されそして意図される。

【００３７】

異なるランダム欠失を有することによってのみ各々が他と異なる、少なくとも２個の（好ましくいは５個より多い、より好ましくは１０個より多い）欠失ポリヌクレオチドのライブラリーを含む実質的に純粋な組成物もまた、記載されそして意図される。必要に応じて、このような欠失ポリヌクレオチドはさらに、欠失位置に挿入された少なくとも１つのヌクレオチドを含む。

【００３８】

本発明の別の方法は、遺伝的要素中のランダムな位置にヌクレオチド付加を有するポリヌクレオチド配列のライブラリーを生成するための方法であって、以下の工程：

（ａ）この遺伝的要素を含む複数コピーの環状ポリヌクレオチドの組成物を、ランダムな切断に供して、複数の線状ポリヌクレオチドを獲得する工程であって、このポリヌクレオチドの各々は、少なくとも１つの３’末端および５’末端を有する、工程；および

（ｂ）工程（ａ）由来のポリヌクレオチドを、このポリヌクレチドの末端の１つに少なくとも１つのヌクレオチドを付加するプロセスに供して、付加ポリヌクレオチド配列のライブラリーを生成する工程であって、このライブラリーは、異なるランダムな位置に付加を有する複数の付加配列を含む、工程、

を包含する。さらに、所望される場合、工程（ｂ）由来の付加ポリヌクレオチドは、上記３’末端および５’末端を互いに共有結合するプロセスに供され得る。必要に応じて、上記ポリヌクレオチドのライブラリーは、目的の機能について選択するプロセスに供され得る。

【００３９】

本明細書に記載される方法のいずれかにおいて、切断は好ましくは、エンドヌクレアーゼ、好ましくはＳ１の使用により起きる。この方法は、付加ポリヌクレオチドのライブラリーが、各々が他と異なる位置にヌクレオチドのランダム付加を有する任意数の異なるポリヌクレオチド、例えば、少なくとも５個、１０個、２０個または３０個の別個のポリヌクレオチドを含むことを可能にする。本発明の１つの実施形態では、環状ポリヌクレオチドの複数コピーの組成物は、遺伝的要素に対する天然に存在するホモログを含まない。必要に応じて、この方法の工程（ａ）および（ｂ）は、反復され得る。別のオプションは、工程（ｂ）での付加点でヌクレオチドを欠失させるプロセスを包含する。任意数のヌクレオチドは、出発分子および技術者の目標に応じて、工程（ｂ）で付加され得、例えば１～３個、３～５０個、もしくは５０～１００個またはそれ以上のヌクレオチドが、工程（ｂ）で付加され得る。

【００４０】

異なるランダム付加を有することでのみ各々が他と異なる、少なくとも２個（好ましくは、少なくとも５個、最も好ましくは、少なくとも１０個）の付加ポリヌクレオチドのライブラリーを含む、実質的に純粋な組成物が、意図される。

【００４１】

さらに、驚くべきことに、本発明は、ポリヌクレオチドの末端で短い欠失を作製して、末

端に短い欠失（１～１００個、好ましくは１～３５個、最も好ましくは１～１０個）を有するポリヌクレオチドの集団を生産する方法を提供する。次いで、このような欠失を有するＤＮＡ末端を他のＤＮＡ末端と共有結合させることができ、特定の内部位置に欠失を含有するポリヌクレオチドのライブラリーを生成することができる。多くの場合、連結されるべき２つの末端は、得られた連結産物が環状ポリヌクレオチドを含むように、同一ＤＮＡ分子上に存在する。このような方法および組成物は、タンパク質工学および指向された進化の領域で重要である。

【００４２】

（発明の詳細な説明）

遺伝子交換（ｓｗａｐｐｉｎｇ）事象は、高分子の進化における主要な駆動力（ｄｒｉｖｅｒ）を構成する。交換事象は、ヌクレオチドの挿入、欠失、または置換を含み得る。交換事象は、相同組換えを介して発生し得るが、抗体遺伝子セグメントに用いられるＶ（Ｄ）Ｊ組換えおよびＤＮＡ末端結合機構において発生するように、非相同的手段によっても発生し得る［Ｓｍｉｄｅｒ＆Ｃｈｕ，Ｓｅｍ．Ｉｍｍｕｎ．９：１８９－９７（１９９７）］。分子進化に関する現在の技術は、遺伝子交換のために一般的に適用可能な非相同的手段を提供するものではない。

【００４３】

本発明の適用には、改善した機能または変化した機能を有する新規な遺伝的要素の産生が含まれる。これらの遺伝的要素は、かなりの商業的価値を有し得る。例えば、遺伝的要素はタンパク質製剤の産生を強化し得る。遺伝的要素は、モノクローナル抗体、または病気を処置するために用いられる酵素などのタンパク質製剤をコードし得る。さらに、遺伝的要素は、化学品製造などの工業プロセスにおいて重要な酵素をコードし得るか、あるいは洗濯用界面活性剤（すなわち、プロテアーゼ、リパーゼまたはエステラーゼ）のような製品中に用いられ得る。さらに、遺伝的要素は、病原菌耐性のための手段を提供する、または、植物種による新規な栄養素の産生を可能にするといったような、農業における重要な用途を有し得る。さらに、遺伝的要素は、新規な抗生物質、色素または他の低分子のような、ヒトへの使用のための新規な製品を生産するため、微生物において用いられ得る。このように、その機能を改善または変化させるための遺伝的要素の改変は、いくつかの異なる産業に無数の適用を有する。

【００４４】

本発明を記載する目的のために、以下の用語は有用であり、かつ、以下のような意味を有する。

【００４５】

（定義）

用語「塩基」は、アデニン、グアニン、チミン、シトシンまたはウラシルのいずれかからなる核酸成分をいう。さらに、「プリン」は、アデニンまたはグアニンのいずれかを指し、「ピリミジン」は、チミン、シトシンまたはウラシルのいずれかをいう。

【００４６】

用語「ヌクレオシド」は、ピリミジンまたはプリンのペントース環（例えば、リボースまたはデオキシリボース）との共有結合を含む分子をいう。

【００４７】

用語「ヌクレオチド」は、ヌクレオシドのリン酸エステルをいう。

【００４８】

用語「ポリヌクレオチド」は、ホスホジエステル結合のような結合を介して、少なくとも１個の他のヌクレオチドの１つの３’ヒドロキシルに共有結合したあるヌクレオチドの少なくとも１個の５’ヒドロキシルを含有する分子をいう。ポリヌクレオチドは必然的に、以下に定義されるような、「残基」を含有する「位置」からなる。

【００４９】

ポリヌクレオチド配列またはポリペプチド配列に関する場合、用語「位置」は、ポリヌクレオチドまたはポリペプチド鎖中の目的の残基の場所をいう。例えば、ポリヌクレオチド

配列中の「位置」は、少なくとも１個の他のヌクレオチドに関する、ポリヌクレオチド鎖中のヌクレオチドの場所として定義される。例えば、単純なポリヌクレオチドＴＧにおいて、Ｔは位置１（自身に関して）であり、Ｇは位置２（位置１のＴに関して）である。最も遠位の５’ヌクレオチドを基準として標識し、位置１として標識することが慣例であることが多い。ＤＮＡのような、遺伝子をコードする二本鎖のポリヌクレオチドでは、時に、遺伝子の翻訳開始部位が位置１として標識されることが多い。これはしばしば、ＡＴＧの翻訳開始配列におけるアデニンである。ＡＴＧからの５’に配置された位置は、負の位置（例えば、－１１、－３５など）が与えられ、ＡＴＧに対して３’に配置された位置は正の位置が与えられる。当業者は、用語「位置」の性質を、ポリヌクレオチドの配列における番号付けスキームに関するものとして理解する。「配列」は、各位置を占める残基の構成に起因する記号列（ｓｔｒｉｎｇ）をいう。例えば、配列ＡＴＧは、塩基アデニンがチミンの直前の位置を占め、チミンがグアニンの直前の位置を占めることを意味する。「特定の位置」は、その配列および構成が公知である少なくとも２個のヌクレオチド間のポリヌクレオチド中の位置をいう。

【００５０】

ポリヌクレオチドまたはポリペプチドに関する場合、用語「残基」は、ポリヌクレオチドについてはプリンヌクレオチドまたはピリミジンヌクレオチドをいい、ポリペプチドについてはアミノ酸をいう。

【００５１】

「遺伝的要素」は、機能をコードするポリヌクレオチドの配列を意味する。例えば、「遺伝的要素」は、ポリペプチド配列をコードし得、プロモーター機能、エンハンサー機能、翻訳開始部位または停止部位、あるいはＲＮＡスプライシング部位などをコードし得る。遺伝的要素は、他の遺伝的要素と作動可能に連結することができ、例えば、プロモーターは、タンパク質をコードする遺伝的要素と作動可能に連結して、目的の細胞型でのタンパク質の発現を可能にする。用語「遺伝子」および「目的の遺伝子」は、ポリペプチドをコードすることができるポリヌクレオチドをいう。

【００５２】

ポリヌクレオチドに関する用語「交換」または「遺伝子交換」は、以下のいずれかを意味する：１）ポリヌクレオチド中の連続した位置を占める少なくとも２個の残基の欠失の発生、または２）ポリヌクレオチドへの連続した位置を占める少なくとも２個の残基の付加の発生、または３）ポリヌクレオチド中の連続した位置を占める少なくとも２個の残基の、他の残基との置換。

【００５３】

ポリヌクレオチドに適用される場合、用語「ヌクレオチド欠失」は、ポリヌクレオチドが、得られるポリヌクレオチドを親の配列、野生型配列または他の参照配列と比較した場合、ポリヌクレオチド鎖中の１以上の位置から１個以上の特定の残基が除去されたことを意味する。

【００５４】

用語「ヌクレオチド挿入」または「ヌクレオチド付加」は、親の配列、野生型の配列、または他の参照配列と比較した場合、ポリヌクレオチドがポリヌクレオチド鎖に付加された特定の残基を有し、これにより少なくとも１個の元の残基がポリヌクレオチド中の新たな位置を現在占めていることを意味する。

【００５５】

用語「ポリヌクレオチド配列のライブラリー」は、ポリヌクレオチドの混合物をいい、ここで、この混合物中の少なくとも１つの配列が少なくとも１つの他の配列と、配列の構成または長さにおいて異なっており、例えば、２つの配列を比較した場合に少なくとも１つの位置が異なるヌクレオチドによって占められているか、または、他方の配列と比較した場合に少なくとも１つのヌクレオチド位置がもう一方の配列中に存在しない、ポリヌクレオチドの混合物をいう。

【００５６】

用語「ＤＮＡ」は、デオキシリボ核酸をいう。当業者は、ＤＮＡに関して本明細書中で記載されている操作がＲＮＡにも適用し得ることを理解する。

【００５７】

用語「ＤＮＡ末端」または末端は、ホスホジエステル結合が分解したＤＮＡ鎖中の位置をいう。一本鎖ＤＮＡの末端では、１個のヌクレオチドが他の１個のヌクレオチドと共有結合しているのみである。「二本鎖ＤＮＡ末端またはＲＮＡの末端」は、分子がもはや二本鎖でない、二本鎖ＤＮＡ分子またはＲＮＡ分子における位置をいう。一般に、ＤＮＡ末端は当業者に認識可能である。二本鎖ＤＮＡの末端は、５’オーバーハング、３’オーバーハングまたはヘアピン構造を有する平滑部分として特徴付けられる。ＤＮＡ末端は５’リン酸基を含有してもしなくてもよい。

【００５８】

本明細書中で使用される場合、用語「切断」は、ホスホジエステル結合のような２個のヌクレオチドの間の結合の開裂をいう。

【００５９】

用語「環状ポリヌクレオチド」は、二本鎖ＤＮＡ末端が全く存在しないポリヌクレオチドをいう。環状ポリヌクレオチドは一本鎖であっても二本鎖であってもよい。しかし、環状ポリヌクレオチドは一本鎖ＤＮＡ末端を含有していてもよい。環状ポリヌクレオチドは、一本鎖ＤＮＡ末端が存在するが互いにハイブリダイズした二本鎖分子の２つの鎖を水素結合が維持し、その結果二本鎖ＤＮＡ末端が互いに接近した２つの一本鎖末端の存在により形成されない場合に存在する。このような環状二本鎖ポリヌクレオチドは、「ニック（ｎｉｃｋｅｄ）」と呼ばれることが多い。

【００６０】

用語「線状ポリヌクレオチド」は、少なくとも１つの、しかしほとんどは２つのＤＮＡ末端を含有するポリヌクレオチドである。線状ポリヌクレオチドは、一本鎖であっても二本鎖であってもよい。

【００６１】

ポリヌクレオチドに適用される場合、用語「ランダム」または「ランダムな位置」は、任意の特定の残基位置が選択され得るプロセスをいう。本明細書中で用いられるランダムは、ヌクレオチドの切断点または位置のすべてが等しい頻度で選定（ｓｅｌｅｃｔ）または選択される（ｃｈｏｓｅｎ）ことを意味しない。むしろランダムは、プロセスの予測不可能な性質に関し、すなわち、ある事象がどこで発生するのか、または、任意の塩基がどの位置を有するかを作業者が演繹的に予測できない。最終的に、利用可能な位置または塩基についてランダムとなるべきプロセスに関して、すべての位置が切断のために利用可能である必要はない。例えば、長さＮのポリヌクレオチドは、操作により影響を受けるその位置（すなわち、１、２、・・・Ｎ）のいずれかまたは全てを有し得る。残基の付加（挿入）または欠失では、ポリヌクレオチドは必然的に共有結合（例えばホスホジエステル結合）は切断されねばならず、その後その残基は欠失されるかまたは付加される（すなわち、位置の総数はそれぞれ増加するかまたは減少する）。長さＮのポリヌクレオチド中の「ランダムな位置での欠失」と記載する際、任意またはすべてのＮ（環状ポリヌクレオチド中）またはＮ－１（線状ポリヌクレオチド中）のヌクレオチド間の共有結合（すなわち、ホスホジエステル結合）が分解し、末端における少なくとも１個のヌクレオチドが再連結に先立って除去されることを意味する。それゆえ、「ランダムな位置での欠失」を引き起こすプロセスでは、ポリヌクレオチドの最終的な長さ（Ｎ、すなわち位置の数）は必然的に減少する。同様に、長さＮのポリヌクレオチド中の「ランダムな位置での挿入」と記載する際、任意またはすべてのＮ（環状ポリヌクレオチド中）またはＮ－１（線状ポリヌクレオチド中）のヌクレオチド間の共有結合（すなわち、ホスホジエステル結合）は分解し、少なくとも１個の新たなヌクレオチド（すなわち、新たな位置）が再連結に先立って末端に付加されることを意味する。それゆえ、「ランダムな位置での挿入」を引き起こすプロセスでは、ポリヌクレオチドの最終的な長さ（Ｎ、すなわち位置の数）は必然的に増加する。「ランダムな位置での欠失」および「ランダムな位置での挿入」を含むプロセスの組

合せにより、ポリヌクレオチドの最終的な長さを不変に維持することが可能な場合もある（すなわち、付加が欠失を相殺し、最終的な位置の数が同じに維持されるが、その位置を占めるヌクレオチドが異なり得る）。長さＮのポリヌクレオチド中の「ランダムな切断」または「シングルランダム分解（ｓｉｎｇｌｅ　ｒａｎｄｏｍ　ｂｒｅａｋ）」と記載する場合、単一のポリヌクレオチド分子中の残基位置間のＮ（環状ポリヌクレオチド中）またはＮ－１（線状ポリヌクレオチド中）の共有結合のいずれか１つが切断されることを意味する。したがって、ポリヌクレオチドの多くのコピーを含む１つの容器中で、シングルランダム分解が種々の分子の種々の位置で発生し得る。

【００６２】

本明細書中で使用される場合、「実質的に純粋な」とは、対象種が、存在する優先種であり（すなわち、モルを基準とした場合、組成物中に他のいかなる個別の高分子種よりもより多く存在している）、好ましくは、実質的に精製された分画が、存在するすべての高分子種の少なくとも約５０％（モルを基準として）を対象種が含有する組成物であることを意味する。一般に、実質的に純粋な組成物は、組成物中に存在するすべての高分子種の約８０～９０％を超えて構成する。最も好ましくは、対象種は、基本的に均一になるまで精製され（組成物中に従来の検出方法により汚染種を検出することができない）、ここで、この組成物は、基本的に単一の高分子種からなる。溶媒種、低分子（５００ダルトン未満）および元素イオン種は、高分子種とはみなされない。

【００６３】

用語「相同（ｈｏｍｏｌｏｇｏｕｓ）」または「相同（ｈｏｍｅｏｌｏｇｏｕｓ）」は、１つの一本鎖核酸配列が相補的な一本鎖核酸配列にハイブリダイズし得ることを意味する。ハイブリダイゼーションの程度は、配列間の同一性の量、および後述するように温度や塩濃度といったハイブリダイゼーションの条件を含む多くの要因に依存し得る。好ましくは、同一性の領域は約５ｂｐを超え、より好ましくは同一性の領域は１０ｂｐを超える。このように、「ホモログ」は、同一でないが、生理学的条件下で互いにハイブリダイズし得る核酸分子である。二本鎖ホモログは、変性させた後に互いにハイブリダイズし得る。

【００６４】

用語「異種」は、１つの一本鎖核酸配列が、別の一本鎖核酸配列またはその相補体にハイブリダイズすることができないことを意味する。したがって、異種の範囲とは、核酸フラグメントまたはポリヌクレオチドが、配列中に別の核酸またはポリヌクレオチドにハイブリダイズすることができない範囲または領域を有することを意味する。このような領域または範囲は、例えば、変異領域である。

【００６５】

用語「同一」または「同一性」は、２つの核酸配列が同一の配列または相補的配列を有することを意味する。したがって、「同一性領域」とは、核酸フラグメントまたはポリヌクレオチドの領域または範囲が、別のポリヌクレオチドまたは核酸フラグメントと同一であるかまたはそれに相補的であることを意味する。

【００６６】

用語「増幅」は、核酸フラグメントのコピー数が増大することを意味する。

【００６７】

用語「野生型」は、核酸フラグメントがいかなる変異をも含まないことを意味する。「野生型」タンパク質とは、このタンパク質が天然に見られる活性に匹敵するレベルで活性であることを意味し、天然に見られるアミノ酸配列を典型的に含む。本発明の１つの局面では、用語「野生型」または「親配列」は、配列の操作に先立つ、開始配列または参照配列を示す。

【００６８】

用語「関連ポリヌクレオチド」は、ポリヌクレオチドの領域または範囲が同一であり、ポリヌクレオチドの領域または範囲が異種であることを意味する。

【００６９】

用語「キメラポリヌクレオチド」は、野生型であるヌクレオチド領域および変異した領域

をこのポリヌクレオチドが含むことを意味する。また、この用語は、このポリヌクレオチドが、あるポリヌクレオチド由来の野生型領域および別の関連ポリヌクレオチド由来の野生型領域を含むことを意味する。

【００７０】

本明細書中で使用される場合、用語「集団」は、ポリヌクレオチド、核酸フラグメントまたはタンパク質のような構成成分を集めたものを意味する。「混合集団」とは、核酸またはタンパク質と同一のファミリーに属する（すなわち、関連する）が、配列が異なり（すなわち、同一でない）、したがって、その生物活性が異なる構成成分を集めたものを意味する。「ライブラリー」は、少なくとも２つの構成成分がいくつかの点（化学組成、長さ等）で異なっている集団を必然的に意味する。

【００７１】

用語「特定の核酸フラグメント」は、特定の末端地点を有しかつ特定の核酸配列を有する、核酸フラグメントを意味する。１つの核酸フラグメントが第２の核酸フラグメントの一部と同一の配列を有するが異なる末端を有する、２つの核酸フラグメントは、２つの異なる特異的核酸フラグメントを含む。同一の配列を有するが異なる５’または３’末端を有する２つの核酸フラグメントは、２つの異なる特異的核酸フラグメントを含む。

【００７２】

用語「変異」は、野生型の核酸配列の配列変化またはペプチドの配列変化を意味する。このような変異は、転位または転換などの点変異であり得る。変異は、欠失、挿入または複製であってもよい。

【００７３】

本明細書中で使用されるポリペプチドの表記では、標準的な用法および慣例に従い、左向きの方向はアミノ末端方向であり、右向きの方向はカルボキシ末端方向である。同様に、特記しない限り、一本鎖ポリヌクレオチド配列の左側末端は５’末端であり、二本鎖ポリヌクレオチド配列の左向き方向は５’方向をいう。新生ＲＮＡ転写物の５’から３’付加の方向は、転写方向をいい、このＲＮＡと同一の配列を有し、かつＲＮＡ転写物のこの５’末端に対して５’であるＤＮＡ鎖上の配列領域を、「上流配列」と呼び、このＲＮＡと同一の配列を有し、かつコードＲＮＡ転写物の３’末端に対して３’であるＤＮＡ鎖上の配列領域を、「下流配列」という。

【００７４】

本明細書中である対象に対して適用される場合、用語「天然に存在する」は、対象が天然に見出しされ得るという事実をいう。例えば、天然の供給源から単離され得る生物（ウイルスを含む）中に存在し、研究室で人間の手により意図的に改変されていないポリペプチド配列またはポリヌクレオチド配列は、天然に存在する。一般に、用語天然に存在するは、非病理的（病気になっていない）個体中に存在するような対象、例えば、その種に典型的なものをいう。

【００７５】

本明細書中で使用される場合、用語「生理学的条件」は、温度、ｐＨ、イオン強度、粘性などをいい、これは、生存可能な微生物に適合し、および／または生存可能な培養酵母細胞もしくは哺乳動物細胞中の細胞内に典型的に存在する生化学的パラメータをいう。例えば、典型的な研究培養条件下で増殖させた酵母細胞中の細胞内条件が、生理学的条件である。インビトロ転写カクテルに適したインビトロ反応条件が、通常、生理学的条件である。一般に、インビトロの生理学的条件は、５０～２００ｍＭのＮａＣｌまたはＫＣｌ、ｐＨ６．５～８．５、２０～４５℃、および０．００１～１０ｍＭの２価カチオン（例えば、Mg^{++}、Ca^{++}）、好ましくは、約１５０ｍＭ　ＮａＣｌまたはＫＣｌ、ｐＨ７．２～７．６、５ｍＭ　２価カチオンを含み、０．０１～１．０％の非特異的タンパク質（例えば、ＢＳＡ）を含む場合が多い。非イオン性界面活性剤（Ｔｗｅｅｎ、ＮＰ－４０、Ｔｒｉｔｏｎ　Ｘ－１００）が、通常約０．００１～２％で、典型的には０．０５～０．２％（ｖ／ｖ）で、しばしば存在し得る。特定の水性条件が、従来の方法に従い専門家によって選択され得る。一般的な手引きとしては、以下の緩衝化水性条件が適用可能であり

得る：１０～２５０ｍＭ　ＮａＣｌ、５～５０ｍＭのＴｒｉｓ　ＨＣｌ、ｐＨ５～８、２価カチオン（複数可）および／または金属キレート剤および／または非イオン性界面活性剤および／または膜画分および／または消泡剤および／または閃光剤（ｓｃｉｎｔｉｌｌａｎｔ）を任意に添加する。

【００７６】

本明細書中で使用される場合、「リンカー」または「スペーサー」とは、ＤＮＡ結合タンパク質およびランダムペプチドのような２個の分子を接続し、この２個の分子を好ましい配置に位置させるように機能し、例えば、その結果、ランダムペプチドがＤＮＡ結合タンパク質から最少の立体障害でレセプターに結合することができる分子または分子団をいう。

【００７７】

本明細書中で使用される場合、用語「作動可能に連結された」は、機能的関連におけるポリヌクレオチド要素の連結をいう。核酸は、別の核酸配列との機能的関連に配置された場合、「作動可能に連結され」る。例えば、プロモーターまたはエンハンサーは、コード配列の転写を影響する場合、このコード配列に作動可能に連結される。作動可能に連結されるとは、連結されるＤＮＡ配列が典型的に連続し、２つのタンパク質コード領域を結合する必要がある場合に、連続し、リーディングフレーム中にある。

【００７８】

（進化ランダム分子のライブラリー生産）

本発明は、ランダムな位置でのヌクレオチドの欠失、挿入または欠失および挿入の組合せのいずれかを含有するポリヌクレオチドのライブラリーを作製する方法を提供する。実際は、本発明は、相同性または増幅技術を必要とすることなく遺伝的要素を「交換」する手段を提供する。遺伝的要素の交換は高分子、細胞および生物の進化の推進力であることが知られている［Ｏｓｔｅｒｍｅｉｅｒ＆Ｂｅｎｋｏｖｉｃ，Ａｄｖ　Ｐｒｏｔｅｉｎ　Ｃｈｅｍ　５５：２９－７７（２０００）］。ＰＣＲに基づく遺伝子シャッフリングのような現在の技術では、相同性と独立した遺伝的要素を有意に交換することができない。

【００７９】

（欠失）

１つの実施形態では、本発明は、集団のメンバーが単一のランダムな位置での欠失の存在により互いに異なる、ヌクレオチド集団を作製する方法を提供する。本発明の一方法は、例えば、以下の工程：

（ａ）２つの末端を作製するために、複数コピーのポリヌクレオチドの組成物をランダムな位置で切断する工程；

（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの末端のうちの一末端から、少なくとも１個のヌクレオチドを除去するプロセスに供する工程；および

（ｃ）工程（ｂ）からの上記ポリヌクレオチドを、必要に応じて、上記末端を互いに共有結合させ、１つの位置での欠失により他のものと異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスに供する工程、

を包含する。

【００８０】

さらに、本発明は、ポリヌクレオチドの集団を提供し、この集団のメンバーは、単一のランダムな位置での欠失の存在により互いに異なる。欠失が、遺伝的要素の有害または不要な機能の除去を可能にすることが意図される。これらの機能は、プロテアーゼ部位、イオン結合ドメイン、阻害的転写因子に関するＤＮＡ結合配列、タンパク質の免疫原性ドメイン等を含み得る。

【００８１】

さらなる実施形態では、本発明は、例えば、１つより多い位置で欠失を含有するポリヌクレオチドを生成する方法を提供する。１つの方法は、以下の工程：

（ａ）２つの末端を作製するために、ランダムな位置で複数コピーのポリヌクレオチドの組成物を切断する工程；

（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの上記末端から少なくとも１個のヌクレオチドを除去するプロセスに供する工程；および
（ｃ）任意に、工程（ｂ）からの上記ポリヌクレオチドを、上記末端を互いに共有結合させ、１つの位置での欠失により他のものと異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスに供する工程、
を包含する。次いで、所望されるならば、目的の機能が、選択されてもよい（工程（ｄ））。さらに、所望されるならば、工程（ａ）～（ｃ）または工程（ａ）～（ｄ）が、１～５０回以上反復され得る。

【００８２】

さらに、本発明は、１つより多い位置での欠失を含有するポリヌクレオチドの集団を提供する。複数の位置での欠失が、遺伝的要素の多くの有害または不要な機能の除去を可能にすることが意図される。これらの機能は、当業者には十分理解されるように、プロテアーゼ部位、イオン結合ドメイン、阻害的転写因子に関するＤＮＡ結合配列、タンパク質の免疫原性ドメインまたは目的の他の機能の任意の組合せを含み得る。

【００８３】

（挿入）
１つの実施形態では、本発明は、ポリヌクレオチドの集団を作製する方法を提供し、この集団のメンバーは、単一のランダムな位置での挿入の存在により互いに異なる。１つの方法は、以下の工程：
（ａ）２つの末端を作製するために、ランダムな位置でポリヌクレオチドの複数コピーの組成物を切断する工程；
（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの少なくとも１つの末端に、少なくとも１個のヌクレオチドを挿入するプロセスに供する工程；および
（ｃ）任意に、工程（ｂ）からの上記ポリヌクレオチドを、上記末端を互いに共有結合させ、１つの位置での挿入により他のものと異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスを付す工程、
を包含する。

【００８４】

さらに、本発明は、ポリヌクレオチドの集団を提供し、この集団のメンバーは、単一のランダムな位置での挿入の存在により互いに異なる。本発明のこの実施形態は、遺伝的要素に新たな融合を発生させ得る。例えば、毒素は、標的化される分子（例えば、抗体）に融合され得、重要な代謝経路の酵素モジュール（例えば、ポリケチドシンダーゼ（ｓｙｎｔｈｅｔａｓｅ））は、新たな方法で融合され得、または結合ドメイン（すなわち、核酸結合ドメイン、低分子結合ドメインまたはイオン結合ドメイン、プロテアーゼ部位、または他の翻訳後修飾モジュール）のような新たな機能は、、既存の遺伝的要素中に組み込まれ得る。

【００８５】

同様に、別の実施形態では、本発明は、１つより多い位置での挿入を含有するポリヌクレオチドを生成する方法を提供する。１つの方法は、以下の工程：
（ａ）ランダムな位置で複数コピーのポリヌクレオチドの組成物を切断する工程；
（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ヌクレオチドの少なくとも１つの上記末端に、少なくとも１個のヌクレオチドを挿入するプロセスに供する工程；および
（ｃ）任意に、工程（ｂ）からの上記ポリヌクレオチドを、上記ＤＮＡ末端を互いに共有結合させ、１つの位置での挿入により他のものと異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスに供する工程；および
（ｄ）任意に、目的の機能を選択する工程、
を包含する。工程（ａ）～（ｂ）、（ａ）～（ｃ）または（ａ）～（ｄ）は、１～５０回以上繰り返されてもよい。

【００８６】

さらに、本発明は、１つより多い位置での挿入を含有するポリペプチドの集団を提供する

。本発明のこの実施形態は、遺伝的要素に複数の新たな融合を発生させ得る。例えば、以下のものは、目的の遺伝子にコンビナトリアル様式で融合され得る；毒素は、標的化される分子（例えば、抗体）に融合され得、重要な代謝経路の酵素モジュール（例えば、ポリケチドシンダーゼ）は、新たな方法で融合され得、または複数の結合ドメイン（すなわち、核酸結合部位、イオン結合ドメイン、プロテアーゼ部位、または他の翻訳後修飾モジュール）のような新たな機能は、既存の遺伝的要素中に組み込まれ得る。

【００８７】

（挿入および欠失の組合せ）

１つの実施形態では、本発明は、ポリヌクレオチドの集団を作製する方法を提供し、この集団のメンバーが単一のランダムな位置での欠失および挿入の存在により互いに異なる。この方法は、以下の工程：

（ａ）２つの末端を作製するために、ランダムな位置で複数コピーのポリヌクレオチドの組成物を切断する工程；

（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの上記末端のうちの１つの末端から、少なくとも１個のヌクレオチドを除去するプロセスに供する工程；

（ｃ）工程（ｂ）からの上記ポリヌクレオチドを、工程（ｂ）からの上記ポリヌクレオチドの上記ＤＮＡ末端のうちの少なくとも１つの末端に、少なくとも１個のヌクレオチドを挿入するプロセスに供する工程；

（ｄ）任意に、工程（ｃ）からの上記ポリヌクレオチドを、上記ＤＮＡ末端を互いに共有結合させ、１つの位置での欠失または挿入により他のものと異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスに供する工程、

を包含する。

【００８８】

さらに、本発明は、ポリペプチドの集団を提供し、そのメンバーが単一のランダムな位置での欠失および挿入の組合せにより互いに異なる。本実施形態は、新たな異種ドメインが目的の遺伝子中のドメインと置き換わることを可能にすることを意図する。これに関して、新たな機能、例えば、リガンド結合または酵素触媒が、遺伝的要素に付与され得る。同様に、ネイティブな機能は、本実施形態を利用して強化され得る。

【００８９】

別の実施形態では、本発明は、１つより多い位置での挿入および欠失を含有するポリヌクレオチドを生成する方法を提供する。この点で、欠失は、挿入とは異なる位置で発生してもよく、または、欠失と挿入は、同じ位置で発生してもよい。さらに、欠失および／または挿入は、複数の位置で発生し得る。本方法は、以下の工程：

（ａ）２つの末端を作製するため、ランダムな位置で複数コピーのポリヌクレオチドの組成物を切断する工程；

（ｂ）工程（ａ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの上記末端ののうちの１つの末端から、少なくとも１個のヌクレオチドを除去するプロセスに供する工程；

（ｃ）任意に、工程（ｂ）からの上記ポリヌクレオチドを、上記ポリヌクレオチドの上記末端のうちの少なくとも１つの末端に、少なくとも１個のヌクレオチドを挿入するプロセスに供する工程；

（ｄ）任意に、工程（ｃ）からの上記ポリヌクレオチドを、上記末端を互いに共有結合させ、１つの位置での欠失および挿入により他とのもの異なる少なくとも１個のポリヌクレオチドを含有するポリヌクレオチドのライブラリーを生産するプロセスに供する工程；

（ｅ）任意に、目的の機能について選択する工程；および、任意に工程（ａ）～（ｄ）のいずれかを、１～５０回以上繰り返す工程、

を含む。

【００９０】

さらに、本発明は、１つより多い位置での挿入および欠失を含有するポリヌクレオチドの

集団を提供する。本発明のこの実施形態は、古典的な指向された進化を可能にすることが意図され、この進化では、複数回のランダムな位置での挿入、ランダムな位置での欠失、ならびに挿入および欠失の組合せが生じ、各回の間で遺伝的要素は、必要に応じて選択に供される。本実施形態は、遺伝的要素の機能の改善または改変を可能にする。

【００９１】

（出発材料）

本発明は、研究者にとって目的の任意のポリペプチドに適用可能である。ポリペプチドは、核酸、すなわち、ＲＮＡまたはＤＮＡであってもよい。ポリヌクレオチドは、しばしば、遺伝的要素または１つ以上の目的の遺伝子からなるＤＮＡである。出発材料は、天然の供給源より得ることができるか、あるいは、研究室で合成（例えば、遺伝子合成）されたポリヌクレオチドであっても、研究室で操作された天然の供給源由来のポリヌクレオチドであってもよい。ポリヌクレオチドのいくつかの供給源は、公共のデータバンク、例えば、Ｇｅｎｂａｎｋ（ｈｔｔｐ：／／ｗｗｗ．ｎｃｂｉ．ｎｌｍ．ｎｉｈ．ｇｏｖ：８０／Ｇｅｎｂａｎｋ／ｉｎｄｅｘ．ｈｔｍｌ）から入手できるか、または、市販されている（Ｃｅｌｅｒａ，Ｒｏｃｋｖｉｌｌｅ，ＭＤ；Ｉｎｃｙｔｅ，Ｐａｌｏ　Ａｌｔｏ，ＣＡ；Ｃｌｏｎｔｅｃｈ，Ｐａｌｏ　Ａｌｔｏ，ＣＡ；Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）。

【００９２】

核酸は、任意の供給源から、例えば、ｐＢＲ３２２のようなプラスミドから、クローニングされたＤＮＡまたはＲＮＡから、あるいは、細菌、酵母、ウイルスおよび植物または動物のようなより高等生物を含む任意の供給源からの天然のＤＮＡまたはＲＮＡから得られ得る。ＤＮＡまたはＲＮＡは、血液または組織材料から抽出してもよい。鋳型ポリヌクレオチドは、ポリヌクレオチド連鎖反応（ＰＣＲ）を用いた増幅により得てもよい［Ｍｕｌｌｉｓ、米国特許第４，６８３，２０２号（１９８７年）；Ｍｕｌｌｉｓら、米国特許第４，６８３，１９５号（１９８７年）］。あるいは、ポリヌクレオチドは、細胞中に存在するベクター中に存在してもよく、当該分野において公知の方法で細胞を培養し、細胞から核酸を抽出することにより、十分な核酸を得ることができる。

【００９３】

ベクターの選択は、ポリヌクレオチド配列のサイズ、および、本発明の方法で用いられる宿主細胞に依存する。鋳型は、プラスミド、ファージ、コスミド、ファージミド、ウイルス（例えば、レトロウイルス、パラインフルエンザウイルス、ヘルペスウイルス、レオウイルス、パラミクソウイルスなど）、またはそれらの選択された部分（例えば、コートタンパク質、スパイク糖タンパク質、キャプシドタンパク質）であってもよい。例えば、コスミド、ファージミド、ＹＡＣおよびＢＡＣが好ましく、ここで、これらのベクターは大きな核酸フラグメントを安定に増殖させることができるため、変異される特定の核酸配列はより大きい。

【００９４】

特定の核酸配列がベクターにクローニングされる場合、各ベクターを宿主細胞に挿入し、宿主細胞にこのベクターを増幅させることによりクローン増幅され得る。核酸配列の絶対数が増大する一方、変異体の数は増大しないため、これはクローン増幅（ｃｌｏｎａｌ　ａｍｐｌｉｆｉｃａｔｉｏｎ）といわれる。

【００９５】

出発材料は、実質的に純粋な形態であるべきである。ポリヌクレオチドは、二本鎖でも一本鎖でもよいが、より好ましいのは二本鎖である。さらに、ポリヌクレオチドは、線状でも環状でもよいが、好ましい実施形態では、ポリヌクレオチドは環状である。環状形態のポリヌクレオチドは、当業者に周知の技術により、細菌、酵母、植物、または哺乳動物の細胞ののような生物由来のプラスミドＤＮＡの調製により調製されてもよい［Ｍａｎｉａｔｉｓら、（１９８９）］。反応容器内の種々の特定の核酸フラグメントの数は、少なくとも約１００個、好ましくは少なくとも約５００個、より好ましくは少なくとも約１０００個である。

【００９６】
出発材料（すなわち、ポリヌクレオチド）は、実質的に純粋な形態であるものの、ホモログまたは関連配列無しで同様に存在し得る。換言すると、最初の容器中のポリヌクレオチドはすべて同一であり得るが、これらはまた関連しているか、関連していない、すなわち異種であり得る。実際、本発明の実施は、出発材料の配列により影響されることはない。さらに、出発材料の配列は、既知であっても既知でなくてもよい。指向された進化の目的のため、必要とされるものは、ポリペプチドの機能を検出する方法（例えば、スクリーニングアッセイ）だけである。
【００９７】
（ランダムな位置でのポリヌクレオチドの切断）
概して、核酸フラグメントは、多くの異なる方法で切断されてもよい。核酸フラグメントは、容易に入手可能なＤＮＡｓｅＩ、Ｓ１ヌクレアーゼ、Ｐ１または大豆ヌクレアーゼ、あるいはＲＮＡｓｅのようなヌクレアーゼで消化され得る。他の酵素、例えば、ＲＡＧ１およびＲＡＧ２、トポイソメラーゼおよびインテグラーゼは、ポリヌクレオチドを切断することができる。核酸は、超音波処理法により、または小型のオリフィス（ｏｒｉｆｉｃｅ）を有するチューブに通すことにより、ランダムに剪断され得る。照射（例えば、γ線照射または紫外線照射）の使用もまた、ポリヌクレオチドを切断することができる。化学薬剤（例えば、ブレオマイシンまたはメタンスルホン酸メチル（ＭＭＳ））もまた、ポリヌクレオチドを切断できる。
【００９８】
挿入または欠失を含有する、機能的に変異した遺伝子の生成について実質的な重要なことは、ポリヌクレオチドを少ない回数、通常は、１回と１０回との間、好ましくは１回と５回との間、最も好ましくは１回切断することである。本発明は、反応容器中で１個のポリヌクレオチド当り１つの位置でのみ切断が起こるようにポリヌクレオチドを切断するための手段を提供する。重要なことは、本発明が、ポリヌクレオチドのほぼランダムな切断（すなわち、異なる分子中のいくつかの異なる位置での切断）のための手段を提供することである。切断は、二本鎖または一本鎖で起こり得る（すなわち、一本鎖末端または二本鎖末端を生じる）。ポリヌクレオチドを切断できる酵素の例は、ＤＮａｓｅＩ、Ｓ１ヌクレアーゼ、Ｐ１ヌクレアーゼだけでなく、トポイソメラーゼ、トランスポゾン、およびインテグラーゼを含む。切断は、トポイソメラーゼ、トランスポゾンおよびインテグラーゼなどの酵素を用いて、一過的に発生し得る。これらの酵素はポリヌクレオチドを一回または２回以上切断する。Ｓ１ヌクレアーゼは、通常ランダムな様式で二本鎖または一本鎖ポリヌクレオチドを切断するのに使用され得る。好ましい実施形態では、環状の二本鎖ＤＮＡについて、Ｓ１ヌクレアーゼがポリヌクレオチドを１回だけ切断し、２つのＤＮＡ末端を生じる（図４）。
【００９９】
高い頻度で（すなわち、ポリヌクレオチド内のいくつかの位置で）ＤＮＡを切断する１つ以上の制限酵素を用いて、特定のポリヌクレオチドが１回だけ切断され、得られる集団が、１回だけ切断されたが、異なる位置で切断された異なるポリヌクレオチドを有するポリヌクレオチドを含有するように、核酸も部分的に消化され得ることもまた意図される。制限酵素を用いた切断は、完全にランダムではないかもしれないが、目的の遺伝的要素が異なった位置に十分に特異的な制限部位を有さない場合には、切断パターンは、実質的な多様性を生成するのに十分有用であり得る。
【０１００】
ポリヌクレオチドの一回の切断は、通常ポリヌクレオチドを数回切断する他の代替的機構を介して達成され得ることが意図される。ポリヌクレオチドは、超音波処理法により、または、小型のオリフィスを有するチューブに通すことにより、ランダムに剪断され得る。照射（例えば、γ線照射または紫外線照射）の使用もまた、ポリヌクレオチドを切断可能である。これらの様式のいずれかが注意深く滴定され、精製手段が使用される場合、一回切断された分子を、実質的に純粋な形態で得ることができる（すなわち、一回切断された

分子は、切断されていないかまたは複数に切断された分子から分離させて精製することができる）。

【０１０１】

さらに、ＤＮＡを切断しかつ再結合させるよう作用する酵素（例えば、トポイソメラーゼ、トランスポゾンおよびインテグラーゼ）は、ポリヌクレオチドを効果的に切断するために用いられ得る［Ｓｉｎｇｈら、Ｐｒｏｃ Ｎａｔｌ Ａｃａｄ Ｓｃｉ ９４：１３０４－９（１９９７）］。これらの場合、切断工程および再結合工程はあわせられ得る。好ましくは、ＤＮＡ末端が連結されるか、または切断の後、互いに物理的に接近している。これは、欠失事象または挿入事象の後に、間違った末端が互いに再連結することを防ぐためである。結合された末端を維持する１つの機構は、出発材料としての環状ポリヌクレオチドの使用によるものである。この場合、末端は、介在するポリヌクレオチド鎖により連結される。このため、再連結は、分子間に対して、分子内事象であり、そしてより高い効率で進行する。末端を接近状態に保つ他の機構は、タンパク質の架橋（例えば、クロマチン（すなわち、ヒストンまたは他のＤＮＡ結合タンパク質）を介するか、または再連結と切断とを一緒にする酵素（例えば、トランスポゾン、インテグラーゼまたはトポイソメラーゼ）を介する）である。あるいは、末端は、固体支持体に反対の末端（非切断末端）の連結を介して、互いに接近状態のままであると考えられ得る。

【０１０２】

スーパーコイルプラスミドＤＮＡから構成される環状ポリペプチドの切断は、０．１～１００μｇ、好ましくは１～１０μｇで、Ｓ１ヌクレアーゼのようなヌクレアーゼとともにインキュベートすることにより達成され得る。ヌクレアーゼは、１０μｌの反応物中で、０．１～１０００ユニット、好ましくは１～１００ユニットの量で存在し得る。反応温度は、０℃と１００℃との間、好ましくは４℃と５０℃との間であってもよい。反応時間は、３０秒～１時間で変化させることができるが、好ましくは約１分と３０分との間である。線状化の程度は、図４のような、アガロースゲル上でプラスミドＤＮＡを分析することにより測定することができる。線状ＤＮＡは、好ましくは、当業者に周知の多数の方法のいずれかにより、切断されていないＤＮＡから精製されるべきである。このような方法は、アガロースゲル精製キット（Ｑｉａｇｅｎ，Ｖａｌｅｎｃｉａ，ＣＡ）、ＨＰＬＣ、カラムクロマトグラフィーなどの利用を含む。

【０１０３】

（ヌクレオチドの欠失）

ヌクレオチドの欠失は、種々の方法により、ＤＮＡ末端で生成され得る。例えば、エキソヌクレアーゼ、例えば、エキソヌクレアーゼＩＩＩは、ＤＮＡ末端から３’から５’方向でヌクレオチドを除去するために用いられ得る。次いで、得られるＤＮＡ末端は、一本鎖エンドヌクレアーゼ、例えば、Ｐ１ヌクレアーゼ、Ｓ１ヌクレアーゼまたは大豆ヌクレアーゼを用いたＤＮＡの消化により除去され得る、５’オーバーハングを含有する。Ｂａｌ３１ヌクレアーゼは、５’から３’だけでなく３’から５’の核分解（ｎｕｃｌｅｏｌｙｔｉｃ）活性を有する酵素であり、ＤＮＡ末端からヌクレオチドを欠失させるために用いられ得る。さらに、いくつかのヌクレアーゼ、例えば、Ｅ．ｃｏｌｉ由来のＤＮＡポリメラーゼ、ＫｌｅｎｏｗフラグメントおよびＴａｑポリメラーゼは、エキソヌクレアーゼ活性を有し、ＤＮＡ末端からの欠失の作製に用いられ得ると考えられ得る。すべての生物からの細胞抽出物は、ヌクレオチドを欠失させるよう作用することができるＤＮＡ修復酵素を含有し、したがって、純粋ではない細胞抽出物が、エキソヌクレアーゼ活性の供給源として用いられ得ると考えられ得る。特定の条件下ではエキソヌクレアーゼ活性を有し得ない他のヌクレアーゼは、他の条件下ではＤＮＡ末端での欠失を生じ得る。例えば、Ｓ１ヌクレアーゼは、高い酵素濃度で用いられると、短い欠失を生じることができる。さらに、ＤＮＡ末端が「解ける（ｆｒａｙｅｄ）」ような、ＤＮＡ分子の穏やかな変性は、一本鎖ヌクレアーゼ、例えば、Ｓ１ヌクレアーゼ、Ｐ１ヌクレアーゼまたは大豆ヌクレアーゼの適用の際に欠失を発生させることが意図される。

【０１０４】

好ましい実施形態では、欠失反応の条件は、各ＤＮＡ末端で発生する個々の欠失の数が十分制御され得るように設定される。例えば、塩濃度を変え、ｐＨを変え、温度を変え、または反応の他の任意の生化学的パラメータを変えることにより、多少の欠失が研究者の意図に従って発生するように、ヌクレアーゼ酵素の活性を変化させることができる（例えば、温度を低下させるか、または塩を増大させることにより、エキソヌクレアーゼの処理能力（ｐｒｏｃｅｓｓｉｖｉｔｙ）を低下させ、より少ない欠失を発生させることができる）。図５は、異なる数の欠失をＤＮＡ末端に発生させることが可能な、条件の変更を示す。いくつかの場合、大きな欠失（すなわち、遺伝的要素中の大きなドメインが完全に除去される）が保証され得、他の場合、小さな欠失（すなわち、単一のアミノ酸、またはプロテアーゼ部位を含むもののような数個のアミノ酸が除去される）が好ましい。概して、欠失は、１～１０００個の数で得ることができ、より好ましくはこれらは１～１００個である。特定の場合には、記載されるように、欠失は、１～１０個の数であってもよい。

【０１０５】

ポリヌクレオチド中のランダムな位置での切断のために、得られるポリヌクレオチド中の欠失の位置はまた、ランダムな位置に配置される。また、残基が分子の一方の末端から欠失されるので、欠失の総数は、５’末端および３’末端で発生する欠失の合計に等しい。

【０１０６】

（ヌクレオチドの付加）

ランダムな位置のポリヌクレオチドに付加を作製するため、ポリヌクレオチドは、上述のように、必然的にランダムな位置で切断される。挿入の前に、ヌクレオチドが切断事象の間に生産されるＤＮＡ末端から欠失されてもよい。また、切断反応により形成されるＤＮＡ末端は、新たなヌクレオチドまたはポリヌクレオチドが付加される基質として用いられててもよい。

【０１０７】

いくつかの異なる機構が、ポリヌクレオチドの末端にヌクレオチドを付加するために存在する。例えば、ヌクレオチドは、化学的カップリングにより付加され得る。ポリメラーゼ、例えば、末端デオキシヌクレオチジルトランスフェラーゼは、ＤＮＡ末端にヌクレオチドをセミランダムな様式で付加するために用いられ得る［Ｇａｕｓｓ＆Ｌｉｅｂｅｒ，Ｍｏｌ　Ｃｅｌｌ　Ｂｉｏｌ　１９９６　１６：２５８－６９（１９９６）］。また、切断工程は、トランスポゾンまたはインテグラーゼを挿入事象に用いる場合があり得るように、挿入事象と一緒にされ得る。

【０１０８】

Ｅ．ｃｏｌｉリガーゼまたはファージＴ４リガーゼのようなリガーゼが、新たなポリヌクレオチドを親ポリヌクレオチドに共有結合させるために用いられ得る。好ましい実施形態では、ポリヌクレオチドは、遺伝的要素または遺伝的要素のフラグメントである。遺伝的要素は本質的にいくつかの方法で機能的であるので、遺伝的要素によって、得られるポリヌクレオチドが機能を有しやすくなる。遺伝的要素は、遺伝子、遺伝子の調節要素、または有用なドメインをコードする遺伝的要素であり得る。遺伝的要素は、ｃＤＮＡライブラリーまたはゲノムＤＮＡライブラリーなどの遺伝的要素のライブラリーであり得る。遺伝的要素のフラグメントは、ポリヌクレオチドをヌクレアーゼ、例えば、ＤＮＡｓｅＩ、Ｓ１ヌクレアーゼ、Ｐ１または大豆ヌクレアーゼ、あるいはＲＮＡｓｅを用いて消化することにより生産され得る。他の酵素、例えば、制限酵素およびトポイソメラーゼもまた、ポリヌクレオチドをフラグメントに切断し得る。ポリヌクレオチドは、超音波処理法により、または小型のオリフィスを有するチューブに通すことにより、ランダムに剪断され得る。照射（例えば、γ線照射または紫外線照射）の使用もまた、ポリヌクレオチドをフラグメントに切断し得る。化学薬剤（例えば、ブレオマイシンまたはＭＭＳ）もまた、ポリヌクレオチドをフラグメントに切断し得る。

【０１０９】

遺伝的要素の集団または遺伝的要素のフラグメントを有するランダムな位置で切断された親ポリヌクレオチドと、Ｔ４　ＤＮＡリガーゼのようなリガーゼとを、適当な塩、緩衝液

および温度条件下で混合することにより、遺伝的要素が親ポリヌクレオチドと元の切断事象の位置で共有結合することができることが意図される。したがって、親ポリヌクレオチド内のランダムな位置での挿入を含むポリヌクレオチドの混合物が生産される。各挿入の内容（すなわち、配列）は、遺伝的要素または遺伝的要素のフラグメントが同一である場合には同一であり得、遺伝的要素のフラグメントが非同一である場合には異なる。

【０１１０】

（ＤＮＡ末端の再結合）

ＤＮＡ末端は、ＤＮＡ末端を、ＤＮＡ末端でヌクレオチド間にホスホジエステル結合を形成させる、ＤＮＡリガーゼのような酵素とともにインキュベートすることにより、共有結合的に再結合され得る。リガーゼの例には、Ｅ．ｃｏｌｉ　ＤＮＡリガーゼ、ファージＴ４　ＤＮＡリガーゼまたはヒトＤＮＡリガーゼが含まれる。これらの酵素は、当業者に周知の条件下で、ＤＮＡの連結に用いられ得る。他の酵素もまた、ＤＮＡ末端でヌクレオチド間に共有結合（ホスホジエステル結合のような）を作製することができる。このような酵素は、トポイソメラーゼ、トランスポゾン、インテグラーゼ、および他の再結合酵素である。他の機構が、ＤＮＡ末端の結合に用いられ得、例えば、いずれの末端（すなわち、５’および３’末端の両方）の配列にハイブリダイズして、これらの末端を水素結合で「架橋」し得る配列のオリゴヌクレオチドの利用が用いられ得る。反対側の鎖上の介在配列は、ポリメラーゼ、例えば、Ｅ．ｃｏｌｉポリメラーゼ、Ｋｌｅｎｏｗフラグメント、ファージＴ４ポリメラーゼまたはＴａｑポリメラーゼで満たされ得る。次いで、ニックは、上述にように、ＤＮＡリガーゼにより修復され得る。細胞抽出物もまた、リガーゼ活性を有し、細胞または核抽出物はＤＮＡ末端の再結合に用いられ得る。あるいは、ＤＮＡ分子は、無傷細胞中に導入され得、細胞の機構がＤＮＡ末端を相同的または非相同的に手段により再結合させ得る。

【０１１１】

（ライブラリー組成物）

本発明は、以下の組成物を例とする新規なライブラリーを提供する。

【０１１２】

（欠失）

本発明は、ポリヌクレオチドの集団を提供し、この集団のメンバーは、単一のランダムな位置での欠失の存在により互いに異なる。このような単一の欠失のライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含み得る。欠失ライブラリーは、１つのランダムな位置での少なくとも１個のヌクレオチドの欠失により少なくとも１個の他の分子と異なる、少なくとも１個の分子を含むべきである。各位置での欠失の数は、１～１０００個であり得るが、少なくとも１個であるべきである。欠失が、遺伝的要素の有害または不要な機能の除去を可能にすることが意図される。これらの機能には、プロテアーゼ部位、イオン結合ドメイン、阻害的転写因子のためのＤＮＡ結合配列、タンパク質の免疫原性ドメインなどが含まれ得る。

【０１１３】

さらに、本発明は、１つより多い位置での欠失を含有するポリヌクレオチドの集団を提供する。このようなライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含むべきである。これらの複数の欠失ライブラリーは、１つより多いランダムな位置での少なくとも１個のヌクレオチドの欠失により少なくとも１個の他の分子と異なる、少なくとも１個の分子を含むべきである。複数の位置での欠失が、遺伝的要素の複数の有害または不要な機能の除去を可能にすることが意図される。これらの機能には、複数のプロテアーゼ部位、イオン結合ドメイン、阻害的転写因子のためのＤＮＡ結合配列、タンパク質の免疫原性ドメインなどの任意の組合せが含まれ得る。

【０１１４】

（挿入）

本発明は、ポリヌクレオチドの集団を提供し、この集団のメンバーは、単一のランダムな

位置での挿入の存在により互いに異なる。挿入ライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含み得る。挿入ライブラリーは、１つのランダムな位置での少なくとも１個のヌクレオチドの挿入により少なくとも１個の他の分子と異なる、少なくとも１個の分子を含むべきである。各位置での挿入の数は、１～１０，０００個であってもよいが、好ましくは少なくとも１個である。例えば、毒素は、標的化された分子（例えば、抗体）に融合され得、重要な代謝経路の酵素モジュール（例えば、ポリケチドシンダーゼ）は、新たな方法で融合され得、または、結合ドメイン（すなわち、核酸結合ドメイン、イオン結合ドメイン、プロテアーゼ部位または他の翻訳後修飾モジュール）のような新たな機能は、既存の遺伝的要素中に組み込まれ得る。

【０１１５】

さらに、本発明は、１つより多い位置での挿入を含有するポリヌクレオチドの集団を提供する。このようなライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含むべきである。これらの複数の挿入ライブラリーは、１つより多いランダムな位置での少なくとも１個のヌクレオチドの挿入により少なくとも他の１個の分子と異なる少なくとも１個の分子を含むべきである。本発明のこの実施形態により、遺伝的要素の新規な融合が発生可能となることが意図される。本発明のこの実施形態により、遺伝的要素の複数の新たな融合が発生可能となることが意図される。例えば、以下のものは、コンビナトリアル様式で、目的の遺伝子に融合され得る：毒素は、標的化された分子（例えば、抗体）に融合され得、重要な代謝経路の酵素モジュール（例えば、ポリケチドシンダーゼ）は、新たな方法で融合され得、または、結合ドメイン（すなわち、核酸結合ドメイン、イオン結合ドメイン、プロテアーゼ部位または他の翻訳後修飾モジュール）は、既存の遺伝的要素中に組み込まれ得る。

【０１１６】

（挿入および欠失の組合せ）

本発明は、ポリヌクレオチドの集団を提供し、そのメンバーは、単一のランダムな位置での欠失および挿入の組合せによって互いに異なる。このようなライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含むべきである。これらの組合せライブラリーは、１つのランダムな位置での１個のヌクレオチドの挿入および少なくとも１個のヌクレオチドの欠失によって少なくとも１個の他の分子と異なる、少なくとも１個の分子を含むべきである。本実施形態により、異種ドメインを、目的の遺伝子中のドメインと置き換えることが可能となることが意図される。これに関して、新たな機能（例えば、リガンド結合または酵素触媒）が遺伝的要素に付与され得る。また、ネイティブな機能が、本実施形態を利用して強化され得る。

【０１１７】

さらに、本発明は、１つより多い位置での挿入および欠失を含有するポリヌクレオチドの集団を提供する。このようなライブラリーは、少なくとも２個の分子、好ましくは１００個の分子、最も好ましくは少なくとも約１０００個の分子を含むべきである。これらの組合せライブラリーは、１つのランダムな位置での少なくとも１個のヌクレオチドの挿入、および、１つのランダムな位置での少なくとも１個のヌクレオチドの欠失により、少なくとも１個の他の分子と異なる少なくとも１個の分子を含むべきである。本発明のこの実施形態により、古典的な指向された進化が可能となり、この進化では、ランダムな位置での複数回の挿入、ランダムな位置での欠失、および挿入と欠失の組合せからが生じ、各回の間で目的の遺伝子は、必要に応じて選択に供される。本実施形態により、遺伝的要素の機能の改善または変更が可能となる。

【０１１８】

（組成物の分析）

このようなライブラリーの組成物は、当業者に周知の機構により決定され得る。ライブラリーが挿入または欠失を含むか否かを判定するため、このライブラリーは、アガロースまたはアクリルアミドゲルの電気泳動によって分析されることが可能であり、サイズは、親

配列と比較可能である。他の方法、例えば、ＨＰＬＣ、質量分析、カラムクロマトグラフィーは、ポリヌクレオチド間のサイズの差を同定するために使用され得る。本発明は、ランダムな位置の挿入または欠失に関するので、ライブラリーの組成物を決定するための最も明確な方法は、組成物内の典型的なポリヌクレオチドを配列決定に供することであり、この方法は当業者に周知である。典型的なクローンの配列の比較により、欠失または挿入が、ライブラリーの異なった分子中のランダムな位置で発生したかどうかを判定することができる。

【０１１９】

得られるライブラリーは、ライブラリー内に含まれる、得られた変異体を発現するためのビヒクルとして用いるための発現ベクター内に連結することができる。発現ベクターの性質は、「スクリーニング」の項において後述する。

【０１２０】

（目的の機能に関するスクリーニング）

目的の機能に関するポリヌクレオチドのライブラリーの試験では、ライブラリーは、適当な発現ベクター中に挿入されるべきである。あるいは、ライブラリーは、発現ベクター中に構築され得る（すなわち、ライブラリーが発現ベクターを含む）。クローニングに用いられるベクターは、所望のサイズのＤＮＡフラグメントを受容する限り、重要ではない。ＤＮＡフラグメントの発現が所望される場合には、クローニングビヒクルは、宿主細胞中のＤＮＡフラグメントの発現を可能にするために、ＤＮＡフラグメントの挿入部位の隣に転写シグナルおよび翻訳シグナルをさらに含むべきである。細菌細胞でのスクリーニングのために好ましいベクターには、プラスミドのｐＵＣシリーズおよびｐＢＲシリーズが含まれる。

【０１２１】

得られる細菌集団は、ランダム変異を有する、多くの組換えＤＮＡフラグメントを含む。この混合された集団を試験して、所望の組換え核酸フラグメントを同定し得る。選択方法は、所望のＤＮＡフラグメントに依存する。

【０１２２】

ベクターの選択は、本発明の方法に用いられるべきポリヌクレオチド配列のサイズおよび宿主細胞に依存する。鋳型は、プラスミド、ファージ、コスミド、ファージミド、ウイルス（例えば、レトロウイルス、パラインフルエンザウイルス、ヘルペスウイルス、レオウイルス、パラミクソウイルスなど）またはこれらの選択された部分（例えば、コートタンパク質、スパイク糖タンパク質、キャプシドタンパク質）であってもよい。特定の核酸配列が比較的大きい場合、これらのベクターは、より大きな核酸フラグメントを安定に遺伝させるので、例えば、コスミド、ファージミド、ＹＡＣおよびＢＡＣが好ましい。

【０１２３】

リガンドに対して増加された結合効率を有するタンパク質をコードするＤＮＡフラグメントが所望される場合、集団またはライブラリー中のＤＮＡフラグメントの各々により発現されるタンパク質を、当該分野において公知の方法（すなわち、パンニング、アフィニティクロマトグラフィー）により、リガンドに対する結合能力に関して試験することができる。増加された薬剤耐性を有するタンパク質をコードするＤＮＡフラグメントが所望される場合、集団またはライブラリー中のＤＮＡフラグメントの各々により発現されるタンパク質を、宿主生物に薬剤耐性を付与する能力に関して試験することができる。所望のタンパク質についての知識が与えられた当業者は、容易に集団を試験し、タンパク質に所望の特性を有するＤＮＡフラグメントを同定できる。

【０１２４】

本発明の状況において、用語「陽性のポリペプチド改変体」は、対応する導入（ｉｎｐｕｔ）ＤＮＡ配列から生産可能なポリペプチドと比較して改善された機能特性を有する、得られたポリペプチド改変体を意味する。このような改善された特性の例は、例えば、強められるか、または弱められた生物学的活性、増大された洗浄性能、熱安定性、酸化安定性、基質特異性、抗生物質耐性あるいは目的であり得る他の特性と同様に異なってもよい。

【０１２５】
したがって、陽性改変体の同定のために用いられるべきスクリーニング方法は、変化が望まれる問題のポリペプチドの特性、および、その変化が望まれる方向にあるポリペプチドの特性に依存する。
【０１２６】
所望の生物学的活性について選択するための、多くの適したスクリーニング系または選択系が、当該分野において記載されている。例えば、Ｓｔｒａｕｓｂｅｒｇら［Ｓｔｒａｕｓｂｅｒｇら、Ｂｉｏｔｅｃｈｎｏｌｏｇｙ（ＮＹ）１３：６６９－７３（１９９５）］は、カルシウム依存的安定性を有するサブチリシン改変体に対するスクリーニング系を記載している。Ｂｒｙａｎら［Ｂｒｙａｎら、Ｐｒｏｔｅｉｎｓ　１：３２６－３４（１９８６）］は、強められた熱安定性を有するプロテアーゼに関するスクリーニングアッセイを記載している。
【０１２７】
当業者であれば、タンパク質のフラグメントがファージ表面上に融合タンパク質として発現される、ファージディスプレイ系（Ｐｈａｒｍａｃｉａ，Ｍｉｌｗａｕｋｅｅ，Ｗｉｓ．）を使用できることが意図される。組換えＤＮＡ分子は、その一部が組換えＤＮＡ分子によりコードされる、融合タンパク質の転写をもたらす部位でファージＤＮＡにクローニングされる。組換え核酸分子を含有するファージは、細胞での複製および転写を受ける。融合タンパク質のリーダー配列は、融合タンパク質の、ファージ粒子の先端への輸送を指向する。このため、組換えＤＮＡ分子により部分的にコードされる融合タンパク質は、ファージ粒子上に表示され、上述の方法により検出および選択される。
【０１２８】
（核酸中の標的化された短い欠失に影響する方法）
ポリヌクレオチドに短い欠失を作る能力は、通常、ＤＮＡ末端で作用するエキソヌクレアーゼの高い活性および処理能力により阻害される。ＤＮＡ末端における大きな（すなわち、１００塩基より多い）欠失を作るためのいくつかの方法が存在する［Ｓａｍｂｒｏｏｋら、（１９８９）］。しかし、１～１００塩基などの短い欠失、または、１～１０塩基のような非常に短い欠失を制御された様式で作る方法は、可能ではない。特定の部位にこのような欠失を作る能力は、タンパク質工学の分野において重要であり［Ａｌｔａｍｉｒａｎｏら、Ｎａｔｕｒｅ　４０３：６１７－２２（２０００）］、Ｖ（Ｄ）Ｊ組換え法の末端結合機構において強調されており、この方法は、抗体遺伝子に実質的多様性をもたらす［Ｓｍｉｄｅｒ＆Ｃｈｕ，Ｓｅｍ．Ｉｍｍｕｎ．９：１８９－９７（１９９７）］。
【０１２９】
（出発材料）
この欠失生成機構は、研究者にとって目的の、任意のポリペプチドに適用され得る。ポリヌクレオチドは、核酸、すなわち、ＲＮＡまたはＤＮＡであってもよい。ポリヌクレオチドは、遺伝的要素、あるいは１つまたは複数の目的の遺伝子からなるＤＮＡである場合が多い。出発材料は、天然の供給源から得てもよく、または、研究室で合成（例えば、遺伝子合成）されたポリヌクレオチドであっても、研究室で操作された天然の供給源由来のポリヌクレオチドであってもよい。ポリヌクレオチドのいくつかの供給源は、Ｇｅｎｂａｎｋ（ｈｔｔｐ：／／ｗｗｗ．ｎｃｂｉ．ｎｌｍ．ｎｉｈ．ｇｏｖ：８０／Ｇｅｎｂａｎｋ／ｉｎｄｅｘ．ｈｔｍｌ）などの公共のデータバンクを介して入手可能であるか、または、市販されている（Ｃｅｌｅｒａ，Ｒｏｃｋｖｉｌｌｅ，ＭＤ；Ｉｎｃｙｔｅ，Ｐａｌｏ　Ａｌｔｏ，ＣＡ；Ｃｌｏｎｔｅｃｈ，Ｐａｌｏ　Ａｌｔｏ，ＣＡ；Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）。
【０１３０】
核酸は、任意の供給源から、例えば、ｐＢＲ３２２などのプラスミドから、クローニングしたＤＮＡまたはＲＮＡから、あるいは、細菌、酵母、ウイルスおよび植物または動物などの高等生物を含む任意の供給源由来の天然のＤＮＡまたはＲＮＡから得ることができる。ＤＮＡまたはＲＮＡは、血液材料または組織材料から抽出され得る。鋳型ポリヌクレオ

チドは、ポリヌクレオチド連鎖反応（ＰＣＲ）を用いた増幅によって得られ得る［Ｍｕｌｌｉｓ、米国特許第４，６８３，２０２号（１９８７年）；Ｍｕｌｌｉｓら、米国特許第４，６８３，１９５号（１９８７年）］。また、ポリヌクレオチドは、細胞中に存在するベクター中に存在してもよく、この細胞を培養し、当該分野において公知の方法によって、細胞から核酸を抽出することにより、十分な核酸が得られ得る。

【０１３１】

（ヌクレオチドの欠失）

ヌクレオチドの欠失は、種々の手段によりＤＮＡ末端において生成され得る。例えば、エキソヌクレアーゼＩＩＩなどのエキソヌクレアーゼを用いて、ＤＮＡ末端から３’から５’方向にヌクレオチドを除去することができる。得られるＤＮＡ末端は、Ｐ１ヌクレアーゼ、Ｓ１ヌクレアーゼ、または大豆ヌクレアーゼなどの一本鎖エンドヌクレアーゼを用いるＤＮＡの消化により除去され得る５’オーバーハングを含有する場合が多い。他のエキソヌクレアーゼもまた、本発明において用いられ得る。Ｂａｌ３１ヌクレアーゼは、５’から３’の核分解活性ならびに３’から５’の核分解活性を有する酵素であり、ＤＮＡ末端からヌクレオチドを欠失させるために用いられ得る。エキソヌクレアーゼＴは、３’から５’の方向にヌクレオチドを除去し得る。エキソヌクレアーゼ７は、５’から３’方向のヌクレオチドを除去し得、ニックまたはギャップのように一本鎖末端で作用し得る。エキソヌクレアーゼＩは、３’から５’方向での一本鎖ＤＮＡからのヌクレオチドの除去を触媒する。λエキソヌクレアーゼは、５’から３’方向に作用し、二重鎖ＤＮＡからの５’モノヌクレオチドの除去を触媒する、非常に進行性の酵素である。ＲｅｃＪは、５’から３’方向にＤＮＡからのデオキシヌクレオチドモノリン酸の除去を触媒する、一本鎖ＤＮＡ特異的エキソヌクレアーゼである。さらに、いくつかのポリメラーゼ、例えば、Ｅ．ｃｏｌｉ由来のＤＮＡポリメラーゼＩ、Ｋｌｅｎｏｗフラグメント、およびＴａｑポリメラーゼは、エキソヌクレアーゼ活性を有し、ＤＮＡ末端から欠失を作るのに使用可能であると考えられ得る。すべての生物由来の細胞抽出物は、ヌクレオチドを欠失させるように作用し得るＤＮＡ修復酵素を含有し、したがって、不純な細胞抽出物が、エキソヌクレアーゼ活性の源として使用可能であると考えられ得る。特定の条件下ではエキソヌクレアーゼ活性を有し得る他のヌクレアーゼは、他の条件下ではＤＮＡ末端での欠失を生産し得る。例えば、Ｓ１ヌクレアーゼは、高い酵素濃度で使用されると、短い欠失を生産し得る。さらに、ＤＮＡ末端が「解ける」ような、ＤＮＡ分子の穏やかな変性により、一本鎖エンドヌクレアーゼ、例えば、Ｓ１ヌクレアーゼ、Ｐ１ヌクレアーゼまたは大豆ヌクレアーゼの適用の際に、欠失の発生が可能になる。

【０１３２】

好ましい実施形態では、欠失反応の条件は、各ＤＮＡ末端で発生する個々の欠失の数が、十分制御され得るように設定される。例えば、塩濃度および温度を変え、ｐＨを変え、または、反応の他の任意の生物学的パラメータを変えることにより、多少の欠失が研究者の目的に依存して発生するように、ヌクレアーゼの酵素活性を変えることができる。最も際立って、かつ驚くべきことに、本発明者らは、温度を低下させること、および／または、塩を増大させることにより、エキソヌクレアーゼの処理能力を低下させ、より制御された小さい欠失が得られることを見出した。反応に用いられる塩は、いかなる塩であってもよい。塩の例としては、塩化ナトリウム、酢酸ナトリウム、塩化カリウムまたは酢酸カリウムが挙げられる。好ましくは、塩は、塩化ナトリウムまたは塩化カリウムのいずれかである。塩濃度は、１０ｍＭ～１．０Ｍの範囲であり得るが、好ましくは５０ｍＭと５００ｍＭとの間である。反応温度も、本発明では、変化し得る。温度は、０℃～３０℃の範囲であり得るが、好ましくは０℃と２４℃との間である。図５は、異なった数の欠失をＤＮＡ末端上に発生させ得る、条件の変更を示す。いくつかの場合、大きな欠失（すなわち、遺伝的要素中の大きなドメインを完全に除去すること）が保証され得、他の場合、小さな欠失（すなわち、単一のアミノ酸、または、プロテアーゼ部位を含むものなどの数個のアミノ酸を除去すること）が好ましくあり得る。得られるポリヌクレオチドの集団は、開始配列の末端に、種々の量の欠失を含む。概して、欠失は、１～１０００個の数で得ることが

でき、より好ましくはその数は１～１００個である。好ましい実施形態では、欠失は、１～３０個、または、ひいては１～１０個の数であり得る。

【０１３３】

（ＤＮＡ末端の再結合）

いくつかの場合、欠失を含有する分子のＤＮＡ末端を、第２のＤＮＡ末端と結合させて、これにより、今度は欠失が内部位置で発生させることが有用であり得る。連結されるべき２つの末端が同一のＤＮＡ分子に存在することが多く、その結果、得られる接続産物は環状ポリヌクレオチドである。ＤＮＡ末端でヌクレオチド間にホスホジエステル結合を形成させるＤＮＡリガーゼのような酵素とともにＤＮＡ末端をインキュベートすることにより、ＤＮＡ末端は、再結合され得る。リガーゼの例としては、Ｅ．ｃｏｌｉ　ＤＮＡリガーゼ、ファージＴ４　ＤＮＡリガーゼ、またはヒトＤＮＡリガーゼが挙げられる。これらの酵素は、ＤＮＡを連結するための当該分野において周知の条件下で用いられ得る。他の酵素もまた、ＤＮＡ末端でヌクレオチド間に共有結合（ホスホジエステル結合のような）を作ることができる。このような酵素は、トポイソメラーゼ、トランスポゾン、インテグラーゼ、および他の組換え酵素である。他の機構（例えば、その配列がいずれかの末端（すなわち、５’および３’末端の両方）の配列とハイブリダイズして、末端同士を水素結合で「架橋」することのできるオリゴヌクレオチドの使用）が、ＤＮＡ末端の結合に使用可能である。対向する鎖上の介在配列は、ポリメラーゼ、例えば、Ｅ．ｃｏｌｉポリメラーゼ、Ｋｌｅｎｏｗフラグメント、ファージＴ４ポリメラーゼ、またはＴａｑポリメラーゼで満たされ得る。次いで、ニックは、上記のようにＤＮＡリガーゼにより修復され得る。細胞抽出物も、リガーゼ活性を含有し、細胞または核の抽出物は、ＤＮＡ末端の再結合に用いられ得る。あるいは、ＤＮＡ分子は、インタクトな細胞内に導入することができ、細胞の機構部分が相同的または非相同的手段によりＤＮＡを再結合させ得る。

【０１３４】

（欠失組成物）

１つの実施形態では、本発明は、ポリヌクレオチドの組成物を提供し、この集団のメンバーがポリヌクレオチドの一方または両方の末端での欠失の存在により互いに異なる。欠失の数は、各末端で１～１００個の範囲であるが、より好ましくは１～３０個である。

【０１３５】

さらに、本発明は、特定の内部位置（すなわち、末端ではない）での短い欠失により互いに異なるポリヌクレオチドの組成物を提供する。この組成物は、欠失を有するポリヌクレオチドの組成物を、その末端において他のＤＮＡ末端と結合させることにより得られ、その結果、今度は欠失が内部に発生する。連結されるべき２つの末端が同一のＤＮＡ分子にしばしば存在し、その結果、得られる連結産物は、環状ポリヌクレオチドである。欠失の数は、各末端で１～１００個の範囲であり得るが、より好ましくは１～３０個である。

【０１３６】

本明細書中において参照される全ての参考文献および特許公開は、本明細書中に参考として援用される。

【０１３７】

上記開示から理解されるように、本発明は、広範な種々の用途を有する。したがって、以下の実施例は、例示を目的として呈示されるものであり、決して本発明についての限定として解釈されることを意図するものではない。

（実施例）

（実施例１：プラスミドのランダムな切断）

挿入または欠失を利用する分子進化技術では、遺伝子が、少なくとも一時的に、少数回切断される必要がある。必要に応じて、混合物内の各分子は、異なるランダムな位置で一回切断される。一回切断されたＤＮＡを調製することはかなり困難であり、切断はランダムな位置で発生する。Ｂｉｏｎｄｉらは、ＤＮａｓｅＩおよびＤＮＡポリメラーゼを用いてニックを誘発し、次いでこれらのニックをさらに切断して、二本鎖の分解を生成する、面倒な方法を記載している［Ｂｉｏｎｄｉら、Ｎｕｃｌｅｉｃ　Ａｃｉｄｓ　Ｒｅｓ　２６

：４９４６－５２（１９９８）］。このプロセスは、冗長かつ時間がかかる塩化セシウムの勾配精製およびリンカーの連結工程を必要とし、分子進化のような、ハイスループット分子生物学的技術に一般的に適用できない。

【０１３８】

一本鎖エンドヌクレアーゼを用いてＤＮＡのランダムな位置での二本鎖の分解を誘発する戦略は、これまで用いられてはいなかった。これは、Ｓ１、Ｐ１または大豆ヌクレアーゼのような一本鎖ヌクレアーゼが、緊密なスーパーコイルＤＮＡの一本鎖領域を特異的に切断し、これによりニックを生成することを理由としていた。ニックは、これらの酵素についての自然の基質であり、次いで、二本鎖の分解を作り出す切断は、同じ反応で発生し得る。切断に続いて、プラスミドはもはやスーパーコイル状態ではないので、一本鎖領域はもはや存在せず、このためＤＮＡは、もはや酵素にとっての基質ではない。それゆえ、切断は生じ、一度のみしか発生しない。本実施例は、この仮説の利用性を例示するものである。

【０１３９】

ポリヌクレオチドをランダムな位置で切断できる機構を例示するため、プラスミドｐＬａｃＺｉ（Ｃｌｏｎｅｔｅｃｈ，Ｐａｌｏ Ａｌｔｏ，ＣＡ）を用いた。このプラスミドを、ＤＨ１０Ｂ Ｅ．ｃｏｌｉ細胞（Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）中で増殖させ、プラスミドを、Ｑｉａｇｅｎマキシプレップカラム（Ｑｉａｇｅｎ ｍａｘｉｐｒｅｐ ｃｏｌｕｍｎ）（Ｑｉａｇｅｎ，Ｖａｌｅｎｃｉａ，ＣＡ）により調製した。２００ｎｇ／μｌのプラスミドＤＮＡを、０．４、２．０、１０または５０ユニットのＳ１ヌクレアーゼ（Ｐｒｏｍｅｇａ，Ｍａｄｉｓｏｎ，ＷＩ）とともに、１×Ｓ１緩衝液（５０ｍＭ酢酸ナトリウム、ｐＨ４．５、２８０ｍＭ ＮａＣｌ、４．５ｍＭ $ZnSO_4$）中で１０分間、室温でインキュベートした。ＥＤＴＡを０．０２５Ｍまで添加して、７０℃で１０分間加熱することにより、反応を停止させた。タンパク質を、等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で２度、等量のエーテルで一度抽出して除去し、酢酸ナトリウムで沈殿させ、水に再懸濁させた。

【０１４０】

切断したｐＬａｃＺｉを、１．５％アガロースゲル電気泳動により分析した（図４、パネルＡ）。Ｓ１ヌクレアーゼで切断されたプラスミドをｐＬａｃＺｉを一度切断するＣｌａＩで切断されたｐＬａｃＺｉとともに移動することを、観察した。これにより、Ｓ１ヌクレアーゼは環状ＤＮＡ分子を線状化させることができる。Ｓ１ヌクレアーゼが配列特異的な様式でＤＮＡを切断することは知られていないが、Ｓ１によるプラスミドの切断は部位特異的でないと決定することは重要であった。このために、Ｓ１切断により生成された線状プラスミドをゲル精製するか（図４、パネルＢ、レーン５）、または精製してさらにＣｌａＩで切断した（レーン６）。コントロールは、スーパーコイルプラスミド（レーン２）、ＣｌａＩで線状化されたプラスミド（レーン３）またはＳ１ヌクレアーゼで線状化したが精製していないプラスミド（レーン４）を含んでいた。Ｓ１／ＣｌａＩ切断プラスミドを、１つのスメアとして観察し、これはＳ１がプラスミドのいくつかの異なる位置で切断していることを示す。Ｓ１が１つの位置しか切断しない場合、Ｓ１／ＣｌａＩ切断プラスミドは、２本のバンドとして移動する；Ｓ１が２つの位置で切断する場合、Ｓ１／ＣｌａＩプラスミドは３本のバンドなどとして移動する。本実施例の重要性は、ポリヌクレオチドが一度（すなわち、環の線状化）、しかも一度だけ、異なった複数の位置で切断されることが可能である。

【０１４１】

（実施例２：ＬａｃＺ中の部位での欠失）

ヌクレオチド欠失が、遺伝子の構造分析を目的として、かつ、ヌクレオチド配列分析を目的としてなされている。一般にこれらの欠失は、１００個のヌクレオチドをはるかに超える範囲の大きさである。通常の条件下では、例えば、エキソヌクレアーゼＩＩＩは、１分間当り１００個より多い塩基を除去する［Ｓａｍｂｒｏｏｋら、（１９８９）］。しかし、小型の欠失を作製する能力は、タンパク質中の小型のドメインを変化させるか、または

、有害な機能を除去するのに有用である。ポリヌクレオチドの末端に小型の欠失を作るため、エキソヌクレアーゼＩＩＩを、種々の塩（図５）および温度の条件下で用いた。ｐＬａｃＺｉ由来の、蛍光標識した２３２塩基対のＰＣＲ産物を、１００ｍＭ、１５０ｍＭおよび２００ｍＭのＮａＣｌに、１０μｌの６６ｍＭ Ｔｒｉｓ－Ｃｌ（ｐＨ７．４）、０．６６ｍＭの$MgCl_2$中の１０Ｕのエキソヌクレアーゼ ＩＩＩ（Ｎｅｗ Ｅｎｇｌａｎｄ Ｂｉｏｌａｂｓ，Ｂｅｖｅｒｌｙ，ＭＡ）の存在下で、１５℃で５分間の反応中に暴露した。ＥＤＴＡを０．０２５Ｍまで添加することにより反応を停止させ、等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で一度、等量のエーテルで一度抽出し、酢酸ナトリウムで沈殿させた。ＤＮＡを、２０μｌの脱イオン化したホルムアミドに再び懸濁させ、０．５μｌを、製造者の推薦に従い遺伝子スキャン（ｇｅｎｅｓｃａｎ）の設定にセットしたＡＢＩ ３７３シーケンサー（Ｐｅｒｋｉｎ－Ｅｌｍｅｒ，Ｆｏｓｔｅｒ Ｃｉｔｙ，ＣＡ）中の６％ポリアクリルアミド変性ゲルに流した。

【０１４２】

約２５個のヌクレオチドを、１００ｍＭ ＮａＣｌの条件下（図５、第２パネル）で、１５個までのヌクレオチドを１５０ｍＭ ＮａＣｌで、数個のヌクレオチドを２００ｍＭ ＮａＣｌ（下のパネル）で除去することができた。

【０１４３】

ｐＬａｃＺｉ中のＣｌａＩ部位はＬａｃＺ遺伝子のコード領域に存在している。この部位は、遺伝子自体内に短い欠失を作るのに用いられ、次いでさらにＰＣＲにより分析して、欠失が作られた程度を決定した。さらに、欠失を含むプラスミドを、４０μｇ／ｍｌのＸ－Ｇａｌを含むＬＢ寒天プレート上で選択し、ＬａｃＺ遺伝子の機能性を決定した。ｐＬａｃＺｉプラスミド（１０μｇ）を、２００μｌのＣｌａＩ中で線状化し、次いで、２０ＵのＳ１ヌクレアーゼ４００μｌ中でインキュベートして、２ｂｐの５’オーバーハングを除去した。さらに、線状化したプラスミドを濃縮し、ウルトラフリーＭＣ膜（ｕｌｔｒａｆｒｅｅ ＭＣ ｍｅｍｂｒａｎｅ）（３０ｋＤ除去用、Ｍｉｌｌｉｐｏｒｅ，Ｂｅｄｆｏｒｄ，ＭＡ）を通して濾過し、１００Ｕの仔ウシ腸ホスファターゼ（Ｎｅｗ Ｅｎｇｌａｎｄ Ｂｉｏｌａｂｓ，Ｂｅｖｅｒｌｙ，ＭＡ）を含む容量４００μｌの１×仔ウシ腸ホスファターゼ緩衝液中に入れ、室温で４５分間インキュベートした。プラスミドを、等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で、等量のエーテルで一度抽出し、酢酸ナトリウムで沈殿させ、水に再懸濁させた。次いで、このプラスミドを、実施例１に記載したのと同様に、エキソヌクレアーゼＩＩＩとともに、１００ｍＭ、１５０ｍＭまたは２００ｍＭのＮａＣｌの存在下で１５℃で５分間、１０μｌ反応物でインキュベートした。対照アームでは、プラスミドは、エキソヌクレアーゼＩＩＩとともにインキュベートせず、欠失の無い状態での脱リン酸化プラスミドの再連結の発生頻度について調べた。エキソヌクレアーゼＩＩＩ反応の５分後、Ｓ１ヌクレアーゼ５０Ｕを１×Ｓ１緩衝液中に含む混合物を添加した。この混合物を、室温で１５分間さらにインキュベートした。ＥＤＴＡを０．０２５Ｍまで添加し、７０℃で１０分間加熱することにより、この反応を停止させた。次いで、ＤＮＡを等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で一度、等量のエーテルで一度抽出し、酢酸ナトリウムで沈殿させ、１．０ＵのＴ４ ＤＮＡリガーゼ（Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）を含む１０μｌの１×リガーゼ緩衝液中に再び懸濁させた。連結反応物は１５℃で１２時間インキュベーションした。Ｅ．ｃｏｌｉ株ＤＨ１０Ｂ（Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）の電気泳動を、１．０μｌの連結混合物とともに行った。細胞を、４０μｇ／ｍｌのＸ－Ｇａｌおよび１００μｇ／ｍｌのアンピシリンを含むＬＢ寒天プレート上で平板培養し、３０℃で一晩インキュベートした。表１に平板培養実験の結果を示す。

（表１．部位特異的欠失後のコロニー特性）

【０１４４】

【表１】

	青色コロニー	白色コロニー	青／白
Exo III なし	0	0	-
Exo III, 100 mM NaCl	177	66	0.37
Exo III, 150 mM NaCl	340	140	0.41
Exo III, 200 mM NaCl	77	34	0.44

脱リン酸化プラスミドがエキソヌクレアーゼＩＩＩに暴露されない場合（第１行、表１）、バックグラウンドが見られないことに、注意すべきである。いくつかの青色および白色のコロニーが、種々の塩濃度でのエキソヌクレアーゼＩＩＩ処理で見出されている。興味深いことに、少なくとも２／３の再連結がフレーム外（ｏｕｔ　ｏｆ　ｆｒａｍｅ）にあるべきであるので、青色／白色の比率の理論的最大値は０．３３である。しかし、本実験での青色／白色比率は０．３３をわずかに超えており、塩濃度が増大するにつれ増大するように思われる。この傾向は、一端からの１塩基対の欠失がインフレームでの再連結の発生を可能にし、塩が増大するにつれ欠失がより容易でなくなる事実に起因し得る。この結果の統計的有意性は分析されておらず、本当の発生頻度は０．３３により近いものであるかもしれない。

【０１４５】

６つのコロニーを、ＣｌａＩ部位に隣接するプライマーを用いてＰＣＲにより分析した。図６にこれらの結果を示す。上側のパネルに、ｐＬａｃＺｉ由来の野生型の３１２塩基対のフラグメントを示す。クローン１は、２９１塩基のインフレーム欠失（２９１塩基のＰＣＲ産物）を含有し、青色の表現型を維持する。クローン２は、４塩基対のフレーム外の欠失（３０８塩基のＰＣＲ産物）を含有し、白色の表現型を有する。クローン３は、９塩基対のインフレーム欠失（３０３塩基のＰＣＲ産物）を含有し、白色の表現型を有する。クローン４は、６塩基対のインフレーム欠失（３０６塩基のＰＣＲ産物）を含有し、白色の表現型有する。クローン５は、７塩基対のフレーム外欠失（３０５塩基のＰＣＲ産物）を含有し、白色の表現型を有する。クローン６は、３塩基対の欠失（３０９塩基のＰＣＲ産物）を含有し、青色の表現型を有する。より短い欠失はより厳密性の低い表現型をもたらすと思われ得るが、本実験はこのことが必ずしも当てはまらないことを示す。クローン１は、７個のアミノ酸を含む欠失を含有するが機能を維持する。一方で、クローン３および４は、インフレームでのより短い欠失を含有するが機能を維持しない。さらに、本実施例は、機能的配列空間を探索するための欠失技術（ｄｅｌｅｔｉｏｎａｌ　ｔｅｃｈｎｏｌｏｇｙ）の能力を示す。

【０１４６】

（実施例３：ＬａｃＺへの挿入）

ＬａｃＺ遺伝子へのランダムＤＮＡの挿入は、ＤＮａｓｅＩを用いてＣＨＯ細胞由来ｃＤＮＡをフラグメント化し、続いてこれらのフラグメントを線状化ｐＬａｃＺｉ中に連結することにより達成された。ｃＤＮＡは自明ながら機能的であるので、ｃＤＮＡの使用が機能性タンパク質の獲得可能性を最適化するであろうことが意図される。ＣＨＯ細胞のｃＤＮＡ（５μｇ）を、０．００１ユニットのＤＮａｓｅＩを用いて、４０ｍＭ　Ｔｒｉｓ－Ｃｌ（ｐＨ７．４）および１０．０ｍＭ　$MgCl_2$ を含む緩衝液中で、室温で５分間フラグメント化させた。ＥＤＴＡを０．０２５Ｍまで添加し、１０μｇのプロテアーゼＫの存在下で７０℃まで加熱することにより、反応を停止させた。ＤＮＡを、等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で、等量のエーテルで一度抽出し、酢酸ナトリウムで沈殿させた。ＣｌａＩおよびＳ１ヌクレアーゼで線状化したプ

ラスミドを上記のように脱リン酸化し、次いで、再び等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で、等量のエーテルで一度抽出し、酢酸ナトリウムで沈殿させた。ランダムｃＤＮＡフラグメントをプラスミドＤＮＡに挿入するために、線状化し脱リン酸化したプラスミド０．２ｍｇをｃＤＮＡフラグメント１ｎｇとともに、Ｔ４ＤＮＡリガーゼ（１．０Ｕ）の存在下で、１０ｍｌの反応量で１５℃１２時間インキュベートした。コントロールとして、線状化プラスミドを、リガーゼとともにｃＤＮＡフラグメントの非存在下でインキュベートし、ｃＤＮＡフラグメントを、リガーゼとともに線状化ベクター非存在下でインキュベートした。次いで、ＤＨ１０Ｂ　Ｅ．ｃｏｌｉを、１０μｌの各連結反応混合物とともに電気泳動させた。

【０１４７】

いくつかのＥ．ｃｏｌｉコロニーを、Ｘ－Ｇａｌプレート上で白色、中間または青色の表現型のいずれかを示す実験のベクター＋挿入物アームにおいて同定した。ｃＤＮＡフラグメントに連結されたＣｌａＩで線状化されたベクターから発生したコロニーのＣｌａＩ部位についてのＰＣＲから、１００～３００塩基対のサイズの挿入を含有するいくつかのクローンが明らかとなった。これらのうちの３つを図７に示す。このように、ｃＤＮＡフラグメントの遺伝的要素への挿入が、本発明で達成可能である。

【０１４８】

（実施例４：ランダムな位置での機能的変化）

ｌａｃオペロンは、遺伝的要素が容易に研究されるモデル系である。酵素β－ガラクトシダーゼは、ＬａｃＺ遺伝子によりコードされるが、通常、環境中にラクトースが存在している場合にのみ産生される。酵素レベルの制御は、転写レベルで達成される。ｌａｃリプレッサータンパク質は、ＬａｃＺのＡＴＧ開始部位の上流のオペレーター配列に結合し、ＲＮＡポリメラーゼによる転写を阻害する。しかし、ラクトースの存在下では、リプレッサーはオペレーターから除去され、転写が進行し得る。プロモーター活性化の機構は、インデューサーであるラクトースのｌａｃリプレッサーへの結合、および、そのオペレーターに対する親和性を劇的に低下させるアロステリック変化の発生によるものである。研究室での設定では、Ｅ．ｃｏｌｉを比色分析用基質Ｘ－Ｇａｌ上で平板培養することにより、ＬａｃＺ転写を評価することができ、この基質は、β－ガラクトシダーゼにより加水分解された場合にはコロニーを青変させる。オペレーターは、ラクトースのアナログであるＩＰＴＧを用いて抑制解除可能であり、このＩＰＴＧは加水分解されず、かつ、ｌａｃリプレッサーに結合することによりＬａｃＺの転写を強力に誘導する。

【０１４９】

ランダム欠失が遺伝子機能に影響を与える能力を調べるため、ｐＢｌｕｅｓｃｒｉｐｔＩＩ　ＫＳ＋プラスミドを、実施例１および２に記載したように、Ｓ１ヌクレアーゼで線状化し、ゲル精製し、脱リン酸化し、エキソヌクレアーゼＩＩＩで消化させた。２０ｎｇ／μｌの線状化プラスミドを、１０Ｕのエキソヌクレアーゼ ＩＩＩとともに、６６ｍＭ　Ｔｒｉｓ－Ｃｌ（ｐＨ７．４）、０．６６ｍＭ　$MgCl_2$ 緩衝液中で、１５℃にて５分間インキュベートし、次いで、５０ｍＭ酢酸ナトリウム（ｐＨ４．５）、２８０ｍＭ　ＮａＣｌ、４．５ｍＭ　$ZnSO_4$ および１０ＵのＳ１ヌクレアーゼを含む１×Ｓ１溶液を添加し、室温で１５分間インキュベートした。ＥＤＴＡを０．０２５Ｍまで添加して、等量のフェノール：クロロホルム：イソアミルアルコール（２５：２４：１）で、等量のエーテルで一度抽出することにより、反応を停止させ、酢酸ナトリウムで沈殿させた。ＤＮＡを、１．０ＵのＴ４ＤＮＡリガーゼを含む１×Ｔ４ＤＮＡリガーゼ緩衝液中に再懸濁させ、１５℃で１２時間インキュベートした。次いで、連結反応物（１μｌ）を用いて、ｌａｃリプレッサータンパク質を産生するＥ．ｃｏｌｉ株ＴＯＰ１０Ｆ’（Ｉｎｖｉｔｒｏｇｅｎ，Ｃａｒｌｓｂａｄ，ＣＡ）をエレクトロポレーションした。Ｅ．ｃｏｌｉを、インデューサーとしてのＩＰＴＧを含むかまたは含まずに、かつＸ－Ｇａｌが存在するＬＢプレート上でインキュベートし、β－ガラクトシダーゼ活性を測定した。さらに、ｐＢｌｕｅｓｃｒｉｐｔプラスミドを、ＩＰＴＧの存在下または非存在下で、Ｘ－Ｇａｌを含むプレート上で平板培養した。表２に実験の結果を示す。

（表２．β－ガラクトシダーゼの転写における機能性変化）
【０１５０】
【表２】

	+IPTG		-IPTG	
	青	白	青	白
pBluescript	100%	0	0	100%
pBluescript/欠失	66%	34%	2%	98%

いくつかのコロニーは、欠失がランダムな位置で作られた実験アームにおける、インデューサーであるＩＰＴＧの非存在下で、ＬａｃＺを転写する能力を獲得した。さらに、いくつかのコロニーは、ＩＰＴＧの存在下で機能性β－ガラクトシダーゼを産生する能力を失った。ｐＢｌｕｅｓｃｒｉｐｔ／欠失アームからのＩＰＴＧの存在下における１つの白色のコロニーについて配列決定し、翻訳開始部位に８個の塩基対の欠失を有することがわかった。この配列を以下に示し、ここで、メチオニンのコドンをコードする翻訳開始部位（ＡＴＧ）に下線を付している。
【０１５１】
【化１】

CACACAGGAAA--------ACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA
CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

（それぞれ配列番号１および配列番号２）
このように、プラスミドのランダムな切断、その後のエキソヌクレアーゼＩＩＩにより作製される短い欠失により、遺伝的要素の調節領域およびタンパク質コード領域における機能性変化を引き起こすことができる。次いで、これらの変化は、その後機能性アッセイで検出可能である。
【０１５２】
【表３】

配列表

配列番号１

β－ガラクトシダーゼをコードする遺伝子の５’末端での変異

CACACAGGAAAACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

配列番号２

β－ガラクトシダーゼをコードする野生型遺伝子の５’末端

CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGA
ACAA

【図面の簡単な説明】
【図１】図１は、分子進化を選抜するための従来法である、ＤＮＡシャッフリングプロセスの図である。目的の遺伝子のホモログをフラグメント化して、このホモログからの一本鎖フラグメントが伸長反応において互いにプライム（ｐｒｉｍｅ）することができるように、変性および再アニーリングに供する。次いで、完全長遺伝子の増幅により、ハイブリッド遺伝子ライブラリーが生成される。次いで、遺伝子スクリーニングを適用して、変化した遺伝子または改善した遺伝子を選択する。
【図２】図２は、コンビナトリアル多様性を生成するＶ（Ｄ）Ｊ組換え、および結合部多様性を生成するＤＮＡ末端結合のプロセスを示す、免疫グロブリン重鎖遺伝子座の図である。
【図３】図３は、ポリヌクレオチドにおいてランダムな位置でヌクレオチド欠失およびヌクレオチド挿入を生じる方法の例を示す図である。標的遺伝子を切断して、各々がこの遺伝子中のランダムな位置でフラグメント化される遺伝子プールを生産する。残基は、ＤＮ

Ａ末端で欠失される（左）か、または挿入されて（右）、ランダムな位置で欠失、挿入、またはその両方を含むライブラリーを生産し得る。

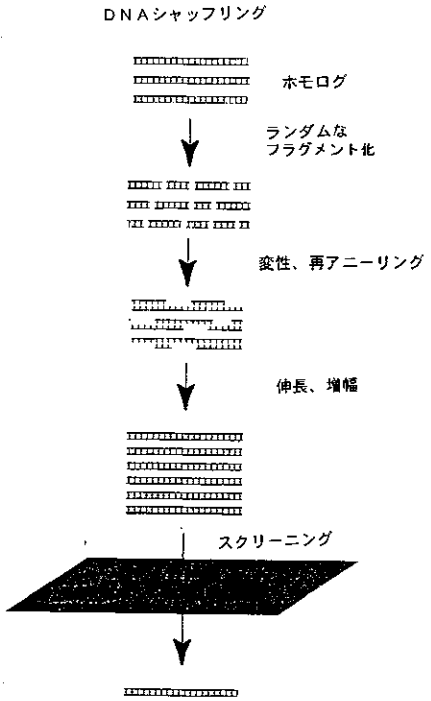
【図４】図４は、ポリヌクレオチドのランダムな切断を示す図である。パネルＡ（図４Ａ）では、ＤＮＡプラスミドｐＬａｃＺｉ（Ｃｌｏｎｔｅｃｈ，Ｐａｌｏ Ａｌｔｏ，ＣＡ）を、切断しなかった（レーン２）か、単一切断制限酵素ＣｌａＩで切断した（レーン３）か、または漸増濃度のＳ１ヌクレアーゼで切断した（レーン４～７）。レーン１および８は、λ／ＨｉｎｄＩＩＩ ＤＮＡマーカーである。パネルＢ（図４Ｂ）では、ｐＬａｃＺｉプラスミドを、切断しなかった（レーン２）か、ＣｌａＩで切断した（レーン３）か、またはＳ１ヌクレアーゼで切断した（レーン４）。Ｓ１で切断したｐＬａｃＺｉ試料をゲル精製し、レーン５に泳動したか、またはＣｌａＩで切断してレーン６で泳動した。等量のＤＮＡを、レーン２～４（１μｇ）、およびレーン５～６（１００ｎｇ）で泳動した。レーン６のスメアは、Ｓ１による切断が、部位特異的ではないことを示す。レーン１および７は、λ／ＨｉｎｄＩＩＩ ＤＮＡマーカーを含有する。

【図５】図５は、ＤＮＡ末端に短ヌクレオチド欠失を生じる方法の例を示す図である。エキソヌクレアーゼＩＩＩは、塩依存性反応で、蛍光標識した２３２ｂｐのＤＮＡフラグメントの末端からヌクレオチドを欠失させる。塩が増加するにつれ、欠失数が減少する。

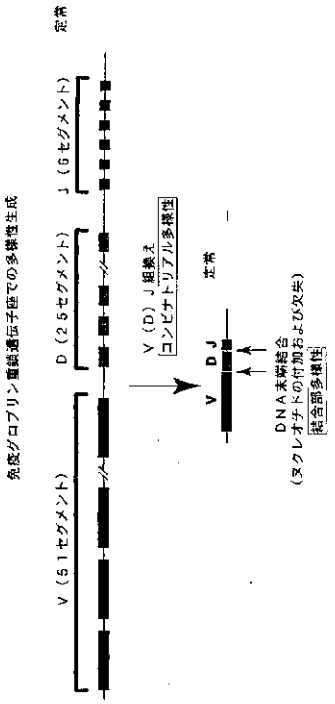
【図６】図６は、ＬａｃＺ遺伝子中のヌクレオチドの欠失を示す図である。プラスミドｐＬａｃＺｉをＣｌａＩで切断し、図５に記載するようにエキソヌクレアーゼＩＩＩで処理して、再連結させて、Ｅ．ｃｏｌｉにエレクトロポレーションして、比色ラクトースアナログＸ－Ｇａｌを含有するプレート上で平板培養した。青色または白色を有するクローンを取り出し、ＬＢ中で増殖させて、ＤＮＡを調製した。ＣｌａＩ部位に隣接するプライマーを用いて、プラスミドをＰＣＲに供した。ここで１つのプライマーを蛍光標識した。ＰＣＲ産物を、ＡＢＩ ３７３ＤＮＡシーケンサーで、６％変性アクリルアミドゲル上で泳動し、Ｇａｎｅｓｃａｎソフトウェア（Ｐｅｒｋｉｎ Ｅｌｍｅｒ，Ｆｏｓｔｅｒ Ｃｉｔｙ，ＣＡ）で分析した。最上部パネルは、３１２ｂｐフラグメントを生産する野生型ＬａｃＺ遺伝子を用いたＰＣＲを示す。クローン１～６には、多様な短い欠失が存在していた。クローン１および６は、青色表現型を有し、２～５は、白色表現型を有した。

【図７】図７は、ｐＬａｃＺｉ中に挿入を含有する３個のクローンを示す１．５％アガロースゲルである。ＣＨＯ細胞のｃＤＮＡを、ＤＮａｓｅＩでフラグメント化して、ｐＬａｃＺｉのＣｌａＩ部位に連結させて、Ｅ．ｃｏｌｉにエレクトロポレーションし、Ｘ－Ｇａｌプレート上で平板培養した。ＣｌａＩ部位に隣接するプライマーを用いたプラスミドＤＮＡのＰＣＲにより、３個のクローンを分析した。１～３と表示したレーンは、異なるサイズの挿入を含有するクローンであり、レーン４は、ｐＬａｃＺｉである。最初のレーンおよび最後のレーン中のＤＮＡは、ΦＸ１７４／Ｈａｅ ＩＩＩ ＤＮＡマーカーであり、右側に示した塩基対のサイズを有する。

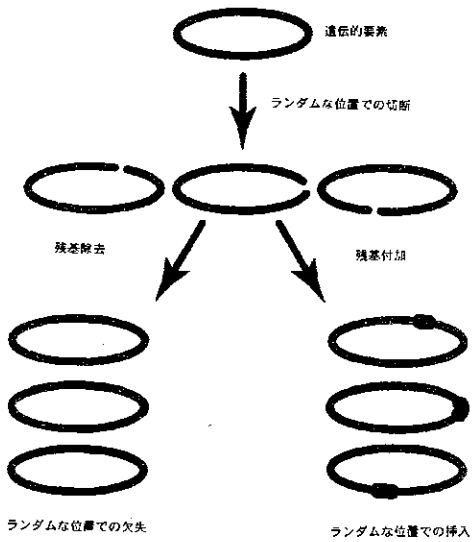
【図１】


DNAシャッフリング
ホモログ
ランダムな
フラグメント化
変性、再アニーリング
伸長、増幅
スクリーニング

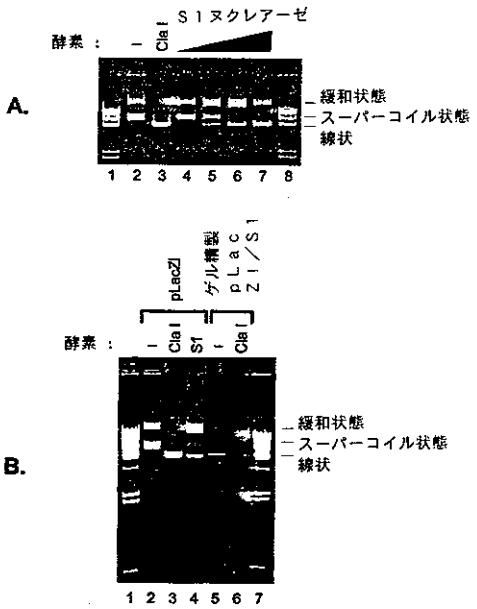
【図２】


生殖
J（6セグメント）
D（25セグメント）
V（51セグメント）
免疫グロブリン重鎖遺伝子座での多様性生成
V（D）J組換え
コンビナトリアル多様性
V
D J
正常
DNA末端結合
（ヌクレオチドの付加および欠失）
結合部多様性

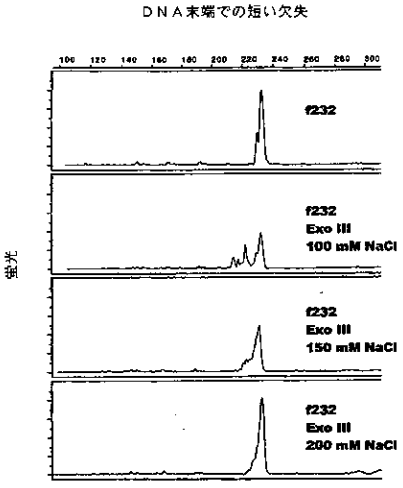
【図３】


遺伝的要素
ランダムな位置での切断
残基除去
残基付加
ランダムな位置での欠失
ランダムな位置での挿入

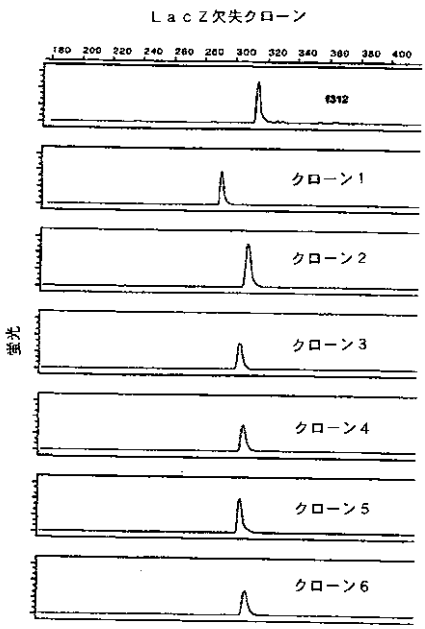
【図４】


酵素：
－
Cla I
S１ヌクレアーゼ
A.
緩和状態
スーパーコイル状態
線状
1 2 3 4 5 6 7 8
pLacZJ
ゲル精製
pLacZ1／S1
酵素：
－
Cla I
S1
－
Cla I
B.
緩和状態
スーパーコイル状態
線状
1 2 3 4 5 6 7

【図５】


ＤＮＡ末端での短い欠失
蛍光
f232
f232
Exo III
100 mM NaCl
f232
Exo III
150 mM NaCl
f232
Exo III
200 mM NaCl

【図６】


ＬａｃＺ欠失クローン
蛍光
f312
クローン１
クローン２
クローン３
クローン４
クローン５
クローン６

【国際公開パンフレット】

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 February 2002 (28.02.2002)

PCT

(10) International Publication Number
WO 02/16642 A1

(51) International Patent Classification[7]: C12Q 1/68, C12P 19/34, C12N 15/64, C07H 21/02

(21) International Application Number: PCT/US01/25788

(22) International Filing Date: 17 August 2001 (17.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/226,177 18 August 2000 (18.08.2000) US

(71) Applicant (*for all designated States except US*): INTEGRIGEN, INC. [US/US]; 42 Digital Drive, Unit 6, Novato, CA 94949 (US).

(72) Inventor; and

(75) Inventor/Applicant (*for US only*): SMIDER, Vaughn [US/US]; 1823 A Pearl Street, Alameda, CA 94501 (US).

(74) Agents: WEBER, Kenneth, A. et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th floor, San Francisco, CA 94111-3834 (US).

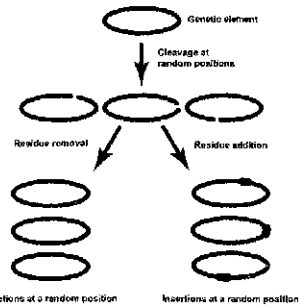
(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(Continued on next page)

(54) Title: METHODS AND COMPOSITIONS FOR DIRECTED MOLECULAR EVOLUTION USING DNA-END MODIFICATION



(57) Abstract: Methods as depicted in figure 3, for directed evolution are described where genetic elements are randomly cleaved to permit the deletion or addition of polynucleotides or both to create a library of related genetic elements with additions or deletions. Corresponding library populations are also described. These processes allow a significant sampling of sequence space which is necessary for directed evolution of genes. Further described are methods for effecting very small nucleotide deletions in genetic elements of interest.

WO 02/16642 A1

WO 02/16642 A1

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS AND COMPOSITIONS FOR DIRECTED MOLECULAR EVOLUTION USING DNA-END MODIFICATION

FIELD OF THE INVENTION

[01] The invention relates to directed evolution, which encompasses methods that can be applied to genetic engineering and protein engineering. Directed evolution is used to evolve gene sequences with the goal of improving or altering gene or protein function. Directed evolution can be applied to many areas including, but not limited to, pharmaceutical development, bioremediation, bioleaching, and the chemical industry.

BACKGROUND OF THE INVENTION

[02] Recently attempts have been made to simulate the process of evolution *in vitro*, thereby inducing genetic changes in specific genes to alter or improve their functions. Although techniques that alter genes have been known for several years, generally detailed features about the encoded protein's structure and function were required for these methods to be successful. The technique of DNA shuffling overcame this barrier to a certain extent, and has been applied to evolve several genes successfully in the past few years [Minshull & Stammer, U.S. Patent # 5,837,458 (1998)].

[03] Natural evolution occurred over millions of years for genes in the environment. *In vitro* evolution attempts to mimic the natural process in days or weeks. In order for an *in vitro* strategy to succeed, several facets of evolutionary theory must be understood. First, the concept of sequence space defines the total number of possible sequences of a protein of a given length [Kauffman, (1993)]. Thus,

$$S = 20^N$$

where S, the sequence space, is the number of possible sequences, and N is the length of the protein. In *in vitro* evolution experiments, it would be optimal to search S sequences of a given protein in order to identify the fraction of those with the most improved or altered activity. It can quickly be seen that a protein with a modest 50 amino acids has an S of 20^{50} possible different sequences, a number which is virtually infinite in terms of analysis with current molecular biology techniques. Second, it is clear that most amino acid changes are deleterious to proteins. These changes may render the protein inactive, cause disruptions in proper folding, or cause instability to the protein or mRNA *in vivo*. It has been estimated that

the ratio of advantageous to deleterious mutations on average is 1 in 10^5 [Radman et al., *Ann. N.Y. Acad. Sci.* 870: 146-55 (1999)]. In this regard, mutation rate is an important parameter when mutating genes to improve their function. If the mutation rate is too high, deleterious mutations will occur in *cis* with the advantageous mutations, a condition which makes the genes with advantageous mutations impossible to identify since the resulting proteins that contain them will be inactive due to concomitant deleterious mutations. Third, in order to overcome the consequences of higher mutation rates, homologous recombination may be utilized to remove deleterious mutation through double crossover events. Fourth, any *in vitro* evolution technique requires a selection screen in order to identify those sequences which improve or alter the function of the protein.

[04] A major barrier to current molecular evolution is the inability to efficiently search sequence space for a given protein. In this regard, of critical importance is the ability to generate and identify sequences that differ at more than one residue, wherein those differences may have an additive effect upon protein function. This additive effect may be described at amino acid interdependence. For example, a protein with a single mutation at residue *i* may not have any detectable increase in function unless a concomitant mutation at *j* is also present. In this case, in order for evolution to be successful, all possible two-mutant variants of the target sequence should be sampled and tested for improved function. In general, the number of R-mutant variants of a protein of length N is given by:

$$S_R = \frac{20^R\ N!}{(N\text{-}R)!R!}$$

where R is the number of mutant variants, and 20 denotes the number of amino acids possible at each position. Thus, for a protein of length 50, there are 490,000 different two mutant variants.

[05] In these statistical analyses of sequence space, a critical value is the length of the protein (i.e. the number of R-mutant variants depends on the length of the protein). However, in nature the length of any given protein may not be as critical for its function as the arrangement of amino acid residues in three dimensional space. Indeed, a hypothetical concept of "catalytic task space" has been proposed to account for this principle (Kauffman, 1993). Alterations in amino acid residues without altering protein length, N, may not affect the three dimensional structure of the protein in the same ways that increasing or decreasing N would. Alternatively, changing N may not change the biological function of the protein at all. Analysis of virtually any family of homologous proteins reveals that members

have different lengths, sometimes with substantial insertions or deletions, but may retain indistinguishable biological function. Thus, the above formula probably does not give an accurate view of the various R-mutant variants that should be screened when searching for improved or altered biological function. In the laboratory it would be optimum to search all of the R-mutant neighbors of a protein plus a number of variants with nucleotides either added or deleted at every position.

[06] In the case of deletions, the number of D-mutant deletions would be given by,

$$S_D = \frac{N!}{(N-D)!\,D!}$$

where N is the initial length of the protein and D is the number of positions where a deletion occurs. In the case of amino acid additions, a similar formula for all possible additions accounts for the fact that any of the 20 amino acids may be added at any position:

$$S_A = \frac{20^A\ N!}{(N-D)!\,A!}$$

where A is the number of possible addition mutants. In the cases of the addition and deletion mutants, these formula both assume that only one amino acid is added or deleted at each position. In terms of *in vitro* molecular evolution, however, it would be optimal to search all 1, 2, 3,...C number of amino acids added or deleted at each position. Thus, for deletion mutants, the number of sequences with variable amino acids deleted at each position,

$$S_{CD} = \frac{C_D\ N!}{(N-D)!\,D!}$$

where C_D denotes the number of amino acids deleted at each position and D is the number of positions where deletions occur. For addition mutants, the formula becomes:

$$S_{CA} = \frac{C_A\ 20^A\ N!}{(N-A)!\,A!}$$

where C_A is the number of amino acids added at each position, and A is the number of positions where additions occur.

[07] Since current molecular biology techniques only allow a fraction of the total space to be generated and sampled for a given protein, a formula which describes an experimental space to be generated can be defined. This expression will also allow the monitoring of improvements in library construction techniques and allow analysis of the

space which is relevant to protein function. A total space to be searched experimentally can be defined as,

$$S_{EX} = S_R S_{CD} S_{CA}$$

where amino acids are mutated to other residues (S_R), are deleted (S_{CD}), or added (S_{CA}) in various combinations and permutations. Certainly current molecular biology techniques would allow a library to be created where R = 1 so $S_R = N$, D = 1 so $S_{CD} = N$, and A = 1 so $S_{CA} = 20\ N$. For a protein of N = 50 then, this hypothetical library would contain $20 * N^3 = 2.5 \times 10^6$ different sequences where all permutations of one position changes, deletions, and additions are represented.

[08] The above discussion on the sequence space relevant to protein evolution may be applied in different ways to the *in vitro* engineering of evolved sequences. In nature, the evolution of different catalytic activities in families of enzymes can be grouped into two broad categories: 1) Those where the active site amino acids remain the same, but differences in the structural folds cause the enzymes to have different substrate specificities [Perona & Craik, *J. Biol. Chem.* 272: 29987-90 (1997)], and 2) Those where the enzyme structures are the same, but differences in the active site residues cause the enzymes to catalyze different reactions [Babbitt & Gerlt, *J. Biol. Chem.* 272: 30591-4 (1997)]. An example of the former is the serine protease family, and of the latter is the enolase superfamily.

[09] Although the differences between these categories might seem trivial, they have important implications for the method of molecular evolution and the concept of sequence space. For the enzymes in families with similar structural folds, a molecular evolution approach which samples sequence space throughout the length of the protein would likely be the optimal strategy to alter an enzyme's specificity, since the catalytic active site will likely require the same residues for the catalytic mechanism. However, for the second type of enzymes, increasing sequence space searching through the entire length of the protein is probably not necessary. More likely, increasing the sequence space sampling of the key catalytic domains will optimize the molecular evolution process. In this regard, it would be better to alter five key amino acids to each of the twenty amino acids and sample this more limited space (20^5) rather than to sample a sequence space of 20^5 spread out over the entire gene sequence. Additionally, sampling as many of the possible addition or deletion mutants in key regions would also contribute to possible success of the *in vitro* evolution protocol. Thus, a method which would optimize molecular evolution of a gene belonging to the second type of enzyme family would be a very important and robust technique.

[10] The swapping of genetic domains is an efficient means to evolve new or improved function in biomolecules. While the alteration of single nucleotide residues can affect gene and protein function, the wholesale exchange of multiple residues in a gene can have dramatic effects on protein function. For example, *E. coli* and *Salmonella* are highly related bacterial species, however the differences in their genetic content are due almost entirely to genetic swapping events as opposed to single residue changes. Additionally, swapping events where exchanges of large chunks of DNA create genes are thought to have occurred several times in pathways such as the clotting cascade, as well as to create novel transcription cassettes through transposition [Bell, (1997); Patthy, (1999)].

[11] A well known example of molecular evolution occurring naturally is that which underlies the production of antibodies in the immune system. In mammalian pre-lymphocytes natural molecular evolution occurs successfully on a daily basis. Antibodies are capable of binding a bewildering array of different antigens, yet have similar amino acid sequences and secondary structures. Antibody genes are arranged in the germline as gene segments (Figure 2). During lymphocyte maturation these segments (named variable or "V", diversity or "D", and junctional or "J") are juxtaposed to one another in the process termed V(D)J recombination to create a functional antibody or T-cell receptor gene. Multiple V, D, and J segments allow a substantial amount of diversity, and hence different antigen binding specificities, to be created in the final repertoire of lymphocytes. The diversity created by this mechanism is referred to as combinatorial diversity. Another type of diversity is also created during V(D)J recombination, which is as important as combinatorial diversity [Davis & Bjorkman, *Nature* 334: 395-402 (1988)]. This diversity is termed junctional diversity, and is created when nucleotides are lost or gained at the joints of the gene segments. Importantly, these joints encode the regions of the antibody molecule which contact antigen, so this type of diversity is critical to creating a diverse, but functional immune system.

[12] The two types of diversity utilized by the immune system might be characterized in the following way with regards to the practice of molecular evolution. Generation of combinatorial diversity in immunoglobulin genes allow a sampling of the total sequence space by providing multiple functional V, D, and J gene segments, each member of which is slightly different in sequence but still homologous to other members of the family of segments. In this respect, the combinatorial rearrangement of V, D, and J gene segments functions as a "domain swapping" event in order to generate novel antibody genes. Generation of junctional diversity allows a greater local sampling of sequence space at the

critical residues for contacting antigen through the mechanism of adding or deleting random nucleotides at the ends of the DNA that are to be ligated.

[13] Due to the aforementioned issues regarding genetic evolution; namely the difficulty in searching a vast sequence space, a preponderance of deleterious mutations in random mutagenesis, and amino acid interdependence, it has been difficult to devise robust methods for searching functional sequence space in the laboratory. Current methods in widespread use for creating mutant proteins in a library format are error-prone polymerase chain reaction [Caldwell & Joyce, (1992); Gram et al., *Proc Natl Acad Sci* 89: 3576-80 (1992)] and cassette mutagenesis [Arkin & Youvan, *Proc Natl Acad Sci* 89: 7811-5 (1992); Hermes et al., *Proc Natl Acad Sci* 87: 696-700 (1990); Oliphant et al., *Gene* 44: 177-83 (1986); Stemmer & Morris, *Biotechniques* 13: 214-20 (1992)], in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide. Alternatively, mutator strains of host cells have been employed to add mutational frequency [Greener et al., *Mol Biotechnol* 7: 189-95 (1997)]. In each case, a 'mutant cloud' [Kauffman, (1993)] is generated around certain sites in the original sequence.

[14] Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. Error prone PCR can also be used to mutagenize a mixture of fragments of unknown sequence. Error-prone PCR can randomly mutate genes by altering the concentrations of respective dNTP's in the presence of dITP [Caldwell & Joyce, (1992); Leung & Miyamoto, *Nucleic Acids Res* 17: 1177-95 (1989); Spee et al., *Nucleic Acids Res* 21: 777-8 (1993)].

[15] However, computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the block changes that are required for continued sequence evolution. The published error-prone PCR protocols are generally unsuited for reliable amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical application. Further, repeated cycles of error-prone PCR lead to an accumulation of neutral mutations, which, for example, may make a protein immunogenic.

[16] In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not significantly combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping into families, arbitrarily choosing a single family, and reducing it to a consensus motif, which is

resynthesized and reinserted into a single gene followed by additional selection. This process constitutes a statistical bottleneck, it is labor intensive and not practical for many rounds of mutagenesis.

[17] Methods of saturation mutagenesis utilizing random or partially degenerate primers that incorporate restriction sites have also been described [Hill et al., *Methods Enzymol* 155: 558-68 (1987); Oliphant et al., *Gene* 44: 177-83 (1986); Reidhaar-Olson et al., *Methods Enzymol* 208: 564-86 (1991)].

[18] "Cassette" mutagenesis is another method for creating libraries of mutant proteins [Bock et al., U.S. Patent # 5,830,720 (1995); Christou & McCabe, U.S. Patent # 5,830,728 (1998); Hill et al., *Methods Enzymol* 155: 558-68 (1987); Miller et al., U.S. Patent # 5,830,740 (1998); Shiraishi & Shimura, *Gene* 64: 313-9 (1988); Stemmer & Cramori, U.S. Patent # 5,830,721 (1998)]. Cassette mutagenesis typically replaces a sequence block length of a template with a partially randomized sequence. The maximum information content that can be obtained is thus limited statistically to the number of random sequences in the randomized portion of the cassette.

[19] A protocol has also been developed by which synthesis of an oligonucleotide is "doped" with non-native phosphoramidites, resulting in randomization of the gene section targeted for random mutagenesis [Wang & Hoover, *J Bacteriol* 179: 5812-9 (1997)]. This method allows control of position selection, while retaining a random substitution rate.

[20] Zaccolo and Gherardi (1999) describe a method of random mutagenesis utilizing pyrimidine and purine nucleoside analogs [Zaccolo & Gherardi, *J Mol Biol* 285: 775-83 (1999)]. This method was successful in achieving substitution mutations which rendered -lactamase with an increased catalytic rate against the cephalosporin cefotaxime. Crea describes a "walk through" method, wherein a predetermined amino acid is introduced into a targeted sequence at pre-selected positions [Crea, U.S. Patent # 5,798,208 (1998)].

[21] Methods for mutating a target gene by insertion and/or deletion mutations have also been developed. It has been demonstrated that insertion mutations could be accommodated in the interior of staphylococcal nuclease [Keefe et al., *Protein Sci* 3: 391-401 (1994)]. Examples of deletional mutagenesis methods developed include the utilization of an exonuclease (such as exonuclease III or Bal31) or through oligonucleotide directed deletions incorporating point deletions [Ner et al., *Nucleic Acids Res* 17: 4015-23 (1989)]. Additionally, Lietz describes a method whereby oligonucleotides with random sequences

may be combined with PCR to induce insertions and deletions. Enhancement of function by this technique has not been shown, and the capacity to overmutagenize (i.e. make too many insertions or deletions per polynucleotide) is substantial in this method [Lietz, U.S. Patent # 6,251,604 (2001)].

[22] The technique most often used to evolve proteins *in vitro* is known as "DNA Shuffling". In this method, a library of gene modifications is created by fragmenting homologous sequences of a gene, allowing the fragments to randomly anneal to one another, and filling in the overhangs with polymerase. A full length gene library is then reconstructed with polymerase chain reaction (PCR). The utility of this method occurs at the step of annealing, whereby homologous sequences may anneal to one another, producing sequences with attributes of both starting sequences. In effect, the method affects recombination between two or more genes that are homologous, but that contain significant differences at several positions. It has been shown that creation of the library using several homologous sequences allows a sampling of more sequence space than using a randomly mutated single starting sequence [Crameri et al., *Nature* 391: 288-91 (1998)]. This effect is likely due to the fact that years of evolution have already selected for different advantageous or neutral mutations amongst the homologs of the different species. Starting with homologs, then, appreciably limits the number of deleterious mutations in the creation of the library which is to be screened. Combinatorially rearranging the advantageous positions of the homologs can apparently allow for an optimized secondary protein structure for catalyzing a biochemical reaction. The resulting evolved protein appears to contain positive features contributed from each of the starting sequences, which results in drastically improved function following selection.

[23] Alterations to the DNA shuffling technique have been devised. One process is termed the 'staggered extension' process, or StEP. Instead of reassembling the pool of fragments created by the extended primers, full-length genes are assembled directly in the presence of the template(s). The StEP consists of repeated cycles of denaturation followed by extremely abbreviated annealing/extension steps. In each cycle the extended fragments can anneal to different templates based on complementarity and extend a little further to create "recombinant cassettes." Due to this template switching, most of the polynucleotides contain sequences from different parental genes (i.e. are novel recombinants). This process is repeated until full-length genes form. It can be followed by an optional gene amplification step [Arnold et al., U.S. Patent # 6,177,263 (2001)].

[24] In another technique, fragmentation of the initial DNA can be accomplished by premature termination of the polymerase in an extension reaction by inducing adduct formation in the target gene [Short, U.S. Patent # 5,965,408 (1999)]. In a different technique, a library is created by inducing incremental truncations in each of two homologs to produce a library of fusion genes, each of which contains domains donated from each homolog [Ostermeier et al., *Nat. Biotechnol.* 17: 1205-9 (1999)]. The advantage of this approach is that significant homology amongst the starting sequences is not required since the annealing step of previous methods is omitted. It is unclear, however, whether this modified technique actually will lead to generation of improved gene function after selection techniques are applied to the library.

[25] The previously described methods of gene shuffling using alleles of genes from different organisms allows combinatorial diversity to occur, but is limited by the homology found in the starting sequences. Additionally, these methods do not provide for a mechanism which would generate the junctional diversity formed through V(D)J joining of antibody gene segments. The present invention makes use of mechanisms analogous to junctional diversity by adding and deleting residues from protein or nucleic acid sequences in either a directed or a random fashion. The present invention also provides for "gene swapping" events analogous to the combinatorial diversity generated by combinatorial V(D)J recombination. This will greatly enhance the means by which genes are evolved *in vitro*.

SUMMARY OF THE INVENTION

[26] The present invention involves the directed molecular evolution of nucleic acid sequences by:

(a) adding or deleting nucleotide residues at random in a polynucleotide to produce a library of polynucleotides containing additions or deletions; and

(b) optionally subjecting the pool of polynucleotides in step (a) to a selection procedure capable of identifying polynucleotides encoding for a desired function or feature. Steps (a) and (b) can optionally be repeated. Libraries produced by the methods of the invention are also described and contemplated.

[27] Uniquely, the present invention allows a sampling of sequence space which will include sequences that significantly affect secondary protein structure, thus increasing the probability of identifying altered or improved function in an evolved gene.

Further, the present invention allows a sampling of sequence space which cannot be sampled by other current technologies. Moreover, libraries of polynucleotides created with the present invention cannot be obtained utilizing other current technologies.

[28] Several methods and compositions are described and contemplated below. One method of the invention generates a library of polynucleotide sequences having nucleotide deletions at differing positions in a sequence of a genetic element comprising the steps of:

(a) subjecting multiple copies of circular polynucleotides comprising the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5'end; and

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one of said DNA ends of said polynucleotides producing a library of deletion polynucleotide sequences, said library comprising multiple deletion polynucleotide sequences with deletions at different random positions.

Further, if desired, polynucleotides from step (b) may be subjected to a process that covalently joins the 3' and 5' ends to one another and the library of polynucleotides may be further subjected to a process that selects for a function of interest. The library of deletion polynucleotides may comprise more than two or more, for example, deletions of at least 10, 20 or 30 or more or even 50 to 100 individual polynucleotides each having a random deletion at a different position from the others may be obtained. The number of deletions made will depend upon the starting material and the goal of the technician. In some embodiments, the library of deletion polynucleotides comprises very short deletions of at least 1, 2, 3, 4, or 5 individual nucleotides or more. In different embodiments, the library may comprise larger deletions of 50-100 or more nucleotides. In another emmbodiment, the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element. Further, steps (a) and (b) may optionally be repeated. Another optional method includes a process for inserting nucleotides at the position of deletion in step (b).

[29] Substantially pure compositions comprising a library of multiple (preferably more than two, more preferably more than 5, most preferably more than 10)

linear polynucleotides each having a different 3' and a 5' end, but each linear polynucleotide being identical to the others if circularized are described and contemplated.

[30] Substantially pure compositions comprising a library of at least 2 (preferably more than 5, more preferably more than 10) deletion polynucleotides each differing from the other only by having a different random deletion are also described and contemplated. Optionally such deletion polynucleotides further comprise at least one nucleotide inserted at the position of deletion.

[31] Another method of the invention generates a library of polynucleotide sequences having nucleotide additions at random positions in a genetic element comprising the steps of:

(a) subjecting a composition of multiple copies of circular polynucleotides with the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which adds at least one nucleotide to one of said ends of said polynucleotides producing a library of addition polynucleotide sequences, said library comprising multiple addition sequences with additions at different random positions.

Further, if desired, the addition polynucleotides from step (b) may be subjected to a process that covalently joins said 3' and 5' DNA ends to one another. Optionally, the library of polynucleotides may be subjected to a process that selects for a function of interest.

In any of the methods described here, cleavage preferably occurs with the use of an endonuclease, preferably S1. This method permits the library of addition polynucleotides to comprise any number of different polynucleotides, for example, at least 5, 10, 20 or 30 individual polynucleotides each having a random addition of nucleotides at a different position from the others. In one embodiment of the claimed invention, the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element. Optionally, steps (a) and (b) of the method may be repeated. Another option includes a process for deleting nucleotides at the point of addition in step (b). Any number of nucleotides may be added in step (b) depending upon the starting molecule and the goal of the technician, for example, 1-3, 3-50, or 50-100 or more nucleotides may be added in step (b).

[32] Substantially pure compositions comprising a library of at least 2 (preferably at least 5, most preferably at least 10) addition polynucleotides each differing from the other only by having a different random addition are contemplated.

[33] Further, the present invention surprisingly provides a method to make short deletions at the end of a polynucleotide, producing a population of polynucleotides with short deletions (from 1 to 100), preferably from 1 to 35, most preferably 1 to 10 at the end. A DNA end having such deletions can then be covalently joined with other DNA ends, producing a library of polynucleotides containing deletions at a specific internal position. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product comprises circular polynucleotides. Such methods and compositions are important in the areas of protein engineering and directed evolution.

BRIEF DESCRIPTION OF THE DRAWINGS

[34] FIG. 1 is a diagram of the process of DNA shuffling, an earlier method of choice for molecular evolution. Homologs of a gene of interest are fragmented, subjected to denaturation and reannealing such that the single-strand fragments from the homologs can prime one another in an extension reaction. Amplification of the full length gene then produces a library of hybrid genes. A genetic screen is then applied to select an altered or improved gene.

[35] FIG. 2 is a diagram of the immunoglobulin heavy chain locus illustrating the process of V(D)J recombination which produces combinatorial diversity, and DNA end-joining which produces junctional diversity.

[36] FIG 3. is a diagram illustrating an example of a method which produces nucleotide deletions and insertions at random positions in a polynucleotide. A target gene is cleaved to produce a pool of genes each of which are fragmented at random positions in the gene. Residues can be deleted (left), or inserted (right) at the DNA ends to produce libraries containing deletions, insertions, or deletions and insertions at random positions.

[37] FIG 4 is a diagram illustrating the random cleavage of a polynucleotide. In panel A (Fig. 4A), the DNA plasmid pLacZi (Clontech, Palo Alto, CA) was either uncleaved (lane 2), cleaved with the single cutting restriction enzyme Cla I (lane 3), or increasing concentrations of S1 nuclease (lanes 4-7). Lanes 1 and 8 are lambda/Hind III DNA markers. In panel B (Fig. 4B), the pLacZi plasmid is uncut (lane 2), cleaved with Cla I (lane 3), or S1 nuclease (lane 4). A sample of the S1 cleaved pLacZi was gel purified

and run in lane 5, or further cleaved with Cla I and run in lane 6. Equal amounts of DNA were run in lanes 2-4 (1 μg), and lanes 5-6 (100 ng). The smear in lane 6 illustrates that cleavage by S1 was not site-specific. Lanes 1 and 7 contain lambda/Hind III DNA markers.

[38] FIG. 5 is a diagram illustrating an example of a method which produces short nucleotide deletions at a DNA end. Exonuclease III deletes nucleotides from the ends of a fluorescently labeled 232 bp DNA fragment in a salt dependent reaction. As salt is increased the number of deletions decreases.

[39] FIG 6 is a diagram illustrating the deletion of nucleotides in the LacZ gene. The plasmid pLacZi was cleaved with Cla I, treated with exonuclease III as described in FIG. 5, re-ligated, electroporated into *E. coli*, and plated on plates containing the colorimetric lactose analog X-Gal. Clones with either a blue or white color were picked, grown in LB, and DNA prepared. Plasmid was subjected to PCR with primers flanking the Cla I site, where one primer was fluorescently labeled. The PCR product was run on a 6% denaturing acrylamide gel in an ABI 373 DNA sequencer and analyzed with Genescan software (Perkin Elmer, Foster City, CA). The top panel shows PCR with the wild-type LacZ gene, producing a 312 bp fragment. Clones 1-6 had variable short deletions present. Clones 1 and 6 had blue phenotypes and 2-5 had white phenotypes.

[40] FIG. 7 is a 1.5% agarose gel showing 3 clones containing an insertion in pLacZi. CHO cell cDNA was fragmented with DNase I, ligated into the Cla I site of pLacZi, electroporated into *E. coli*, and plated on X-Gal plates. Three clones were analyzed by PCR of plasmid DNA using primers flanking the Cla I site. Lanes labeled 1-3 are clones containing different sized insertions, and lane 4 is pLacZi. The DNA in the first and last lanes are ΦX174/Hae III DNA markers with their sizes in basepairs indicated at the right.

DETAILED DESCRIPTION OF THE INVENTION

[41] Gene swapping events constitute a major driver in the evolution of macromolecules. Swapping events may include nucleotide insertions, deletions, or replacements. A swapping event may occur by means of homologous recombination, but may also occur by non-homologous means as they do in V(D)J recombination and the DNA-end joining mechanism used by antibody gene segments [Snider & Chu, *Sem. Immun.* 9: 189-97 (1997)]. Current technologies for molecular evolution do not provide a generally applicable non-homologous means for gene swapping.

[42] Applications of the current invention include producing novel genetic elements with improved or altered function. These genetic elements can have significant commercial value. For instance, the genetic element may enhance production of a protein pharmaceutical. The genetic element may encode a protein pharmaceutical such as a monoclonal antibody, or an enzyme used to treat a disease. Further, the genetic element may encode an enzyme important in industrial processes such as chemical manufacturing, or may be used in a product such as laundry detergent (i.e. proteases, lipases, or esterases). Further, the genetic element may have important uses in agriculture, such as to provide a means for pathogen resistance, or to allow production of novel nutrients by a plant species. Additionally, the genetic element may be used in microorganisms to produce novel products for human use, such as novel antibiotics, pigments or other small molecules. As can be seen, the modification of genetic elements in order to improve or alter their function has a myriad of applications in several diverse industries.

[43] For the purposes of describing this invention the following terms will be helpful and will have the following meanings:

Definitions

[44] The term "base" refers to a component of nucleic acid consisting of either adenine, guanine, thymine, cytosine, or uracil. Additionally, "purine" refers to either adenine or guanine, and "pyrimidine" refers to either thymine, cytosine, or uracil.

[45] The term "nucleoside" refers to a molecule comprising the covalent linkage of a pyrimidine or purine to a pentose ring (such as ribose or deoxyribose).

[46] The term "nucleotide" refers to the phosphate ester of a nucleoside.

[47] The term "polynucleotide(s)" refers to a molecule containing at least one 5' hydroxyl of one nucleotide covalently linked to one 3' hydroxyl of at least one other nucleotide through a bond such as a phosphodiester bond. A polynucleotide is necessarily composed of "positions" containing "residues" as defined below.

[48] The term "position" as it relates to a polynucleotide sequence or polypeptide sequence refers to the location of a given residue in the polynucleotide or polypeptide chain. For example, "position" in a polynucleotide sequence is defined as the location of a nucleotide in the polynucleotide chain in reference to at least one other nucleotide. For instance in the simple polynucleotide TG, the T is in position 1 (in reference to itself) and G is in position 2 (in reference to the T in position 1). Often it is convention to label the furthest 5' nucleotide as a reference and label it as position 1. In a double stranded

polynucleotide encoding a gene, such as DNA, often the translation start site of a gene is labeled as position 1. This is often the adenine in the ATG translation start sequence. Positions located 5' from the ATG would be given a negative position (such as -11, -35, etc.) and positions located 3' to the ATG would be given positive positions. Those skilled in the art will recognize the nature of the term "position" as it relates to the numbering scheme in sequences of polynucleotides. A "sequence" refers to the string resulting from the composition of the residues occupying each position. For example the sequence ATG means that the base adenine occupies a position which immediately precedes thymine, and thymine occupies a position which immediately precedes guanine. A "specific position" refers to a position in a polynucleotide between at least two nucleotides whose sequence and composition is known.

[49] The term "residue" as it relates to a polynucleotide or polypeptide refers to either a purine or pyrimidine nucleotide for polynucleotides, or an amino acid for a polypeptide.

[50] A "genetic element" means a sequence of polynucleotide encoding a function. For example, a "genetic element" may encode a polypeptide sequence, may encode a promoter function, an enhancer function, a transcription start or stop site, or RNA splice sites and the like. Genetic elements may be operatively linked to other genetic elements, for example a promoter may be operatively linked to a genetic element encoding a protein to allow expression of a protein in a given cell type. The term "gene" and "gene of interest" refer to a polynucleotide capable of encoding a polypeptide.

[51] The term "swap" or "gene swapping" in reference to a polynucleotide means either: 1) the occurrence of a deletion of at least two residues occupying consecutive positions in a polynucleotide, or 2) the occurrence of an addition of at least two residues occupying consecutive positions into a polynucleotide, or 3) the replacement of at least two residues occupying consecutive positions in a polynucleotide with other residues.

[52] The term "nucleotide deletions" as applied to a polynucleotide means that a polynucleotide has had one or more specific residues removed from one or more positions in the polynucleotide chain when the resulting polynucleotide is compared to the parental, wild-type, or other reference sequence.

[53] The term "nucleotide insertions" or "nucleotide additions" means that a polynucleotide has had specific residues added to the polynucleotide chain, such that at least one of the original residues now occupies a new position in the polynucleotide when compared to the parental, wild-type, or other reference sequence.

[54] The term "library of polynucleotide sequences" refers to a mixture of polynucleotides, wherein at least one of the sequences differs from at least one other sequence in the mixture by sequence composition or length, for example, where at least one position is occupied by a different nucleotide when the two sequences are compared or at least one nucleotide position is absent in one sequence when compared with the other sequence.

[55] The term "DNA" refers to deoxyribonucleic acid. It will be understood by those of skill in the art that where manipulations are described herein that relate to DNA they will also apply to RNA.

[56] The term "DNA ends" or ends refers to the position in a DNA strand wherein a phosphodiester bond is broken. In a single-stranded DNA end a nucleotide is only covalently linked with one other nucleotide. A "double-stranded DNA or RNA end" refers to the position in a double-stranded DNA or RNA molecule wherein the molecule is no longer double-stranded. Generally DNA ends are recognizable to those skilled in the art. Double-stranded DNA ends are characterized as blunt, having a 5' overhang, a 3' overhang, or a hairpin structure. A DNA end may or may not contain a 5' phosphate group.

[57] The term "cleavage" as used herein refers to the breakage of a bond between two nucleotides, such as a phosphodiester bond.

[58] The term "circular polynucleotide" refers to a polynucleotide wherein no double-stranded DNA ends are present. A circular polynucleotide may be single-stranded or double-stranded. A circular polynucleotide may, however, contain single-stranded DNA ends. A circular polynucleotide will be present if single-stranded DNA ends exist but hydrogen bonding keeps the two strands of the double-stranded molecule hybridized to one another such that a double-stranded DNA end is not created by the presence of two single-stranded ends in proximity to one another. Such a circular double-stranded polynucleotide is often referred to as "nicked".

[59] The term "linear polynucleotide" is a polynucleotide which contains at least one, but most often two DNA ends. A linear polynucleotide may be either single-stranded or double-stranded.

[60] The term "random" or "random position" as applied to a polynucleotide refers to a process by which any of the specific residue positions may be selected. Random as used here does not mean that all points or point of cleavage or nucleotides or positions are selected or chosen with equal frequency. Rather random focuses on the unpredictable nature of the process, i.e. the worker cannot predict *a priori* where an

event will occur or what position any base will have. Finally, not all positions need be available for cleavage for the process to be random as to the available positions or bases. For example, a polynucleotide with a length of N may have any or all of its positions (i.e. 1, 2,...N) affected by a manipulation. In the addition (insertion) or deletion of residues, a polynucleotide necessarily must have covalent bonds (such as phosphodiester bonds) cleaved, thereafter which residues are deleted or added (i.e. the total number of positions is decreased or increased, respectively). In describing "deletions at random positions" in a polynucleotide of length N, it is meant that any or all of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between nucleotides (i.e. phosphodiester bonds) are broken, and at least one nucleotide at the end is removed prior to re-ligation. Thus, in a process causing "deletions at random positions" the final length of the polynucleotide (N, or the number of positions) necessarily decreases. Similarly, In describing "insertions at random positions" in a polynucleotide of length N, it is meant that any or all of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between nucleotides (i.e. phosphodiester bonds) are broken, and at least one new nucleotide (i.e. a new position) is added at the end prior to re-ligation. Thus, in a process causing "insertions at random positions" the final length of the polynucleotide (N, or the number of positions) necessarily increases. It is recognized that a combination of processes involving "deletions at random positions" and "insertions at random positions" may allow the final length of the polynucleotide to remain unchanged (i.e. the additions cancel out the deletions and the final number of positions remains the same, however the nucleotides occupying the positions may be different). In describing "random cleavage" or a "single random break" in a polynucleotide of length N, it is meant that any one of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between residue positions in a single polynucleotide molecule are cleaved. Accordingly, in one vessel containing many copies of a polynucleotide, a single random break can occur at different positions in different molecules.

[61] As used herein, "substantially pure" means an object species is the predominant species present (i.e., on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably a substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional

detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

[62] The term "homologous" or "homeologous" means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentration as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp. Thus, "homologs" are nucleic acid molecules that are not identical but are capable of hybridizing to one another under physiological conditions. Double-stranded homologs are capable of hybridizing to one another following denaturation.

[63] The term "heterologous" means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that nucleic acid fragments or polynucleotides have areas or regions in the sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are, for example, areas of mutations.

[64] The term "identical" or "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a nucleic acid fragment or polynucleotide are identical or complementary to another polynucleotide or nucleic acid fragment.

[65] The term "amplification" means that the number of copies of a nucleic acid fragment is increased.

[66] The term "wild-type" means that the nucleic acid fragment does not comprise any mutations. A "wild-type" protein means that the protein will be active at a comparable level of activity found in nature and typically will comprise the amino acid sequence found in nature. In an aspect of the invention, the term "wild type" or "parental sequence" can indicate a starting or reference sequence prior to a manipulation of the sequence.

[67] The term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

[68] The term "chimeric polynucleotide" means that the polynucleotide comprises nucleotide regions which are wild-type and regions which are mutated. It may also

mean that the polynucleotide comprises wild-type regions from one polynucleotide and wild-type regions from another related polynucleotide.

[69] The term "population" as used herein means a collection of components such as polynucleotides, nucleic acid fragments or proteins. A "mixed population" means a collection of components which belong to the same family of nucleic acids or proteins (i.e. are related) but which differ in their sequence (i.e. are not identical) and hence in their biological activity. A "library" necessarily implies a population wherein at least two of the components is different in some aspect (chemical composition, length, etc.)

[70] The term "specific nucleic acid fragment" means a nucleic acid fragment having certain end points and having a certain nucleic acid sequence. Two nucleic acid fragments wherein one nucleic acid fragment has the identical sequence as a portion of the second nucleic acid fragment but different ends comprise two different specific nucleic acid fragments. Two nucleic acid fragments with identical sequences but different 5' or 3' ends comprise two different specific nucleic acid fragments.

[71] The term "mutations" means changes in the sequence of a wild-type nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications.

[72] In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

[73] The term "naturally-occurring" as used herein as applied to an object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring. Generally, the term naturally-occurring refers to an object

as present in a non-pathological (undiseased) individual, such as would be typical for the species.

[74] As used herein the term "physiological conditions" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45°C. and 0.001-10 mM divalent cation (e.g., Mg^{++}, Ca^{++}); preferably about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v). Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or nonionic detergents and/or membrane fractions and/or antifoam agents and/or scintillants.

[75] As used herein, "linker" or "spacer" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide, and serves to place the two molecules in a preferred configuration, e.g., so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

[76] As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

Producing Libraries of Evolving Random Molecules

[77] The present invention provides a method to create libraries of polynucleotides containing either nucleotide deletions, insertions or combinations of

deletions and insertions at random positions. In effect this invention provides a means to "swap" genetic elements without the need for homology or amplification techniques. The swapping of genetic elements is known to be a driving force in evolution of macromolecules, cells, and organisms [Ostermeier & Benkovic, *Adv Protein Chem* 55: 29-77 (2000)]. Current techniques, such as PCR based gene shuffling, do not allow significant swapping of genetic elements independent of homology.

Deletions

[78] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. One method of the invention, for example, comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of the ends of said polynucleotides; and

(c) optionally subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion at one position.

[79] Further, the invention provides a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. It is contemplated that deletions will allow removal of detrimental or unwanted functions of a genetic element. These functions might include protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

[80] In a further embodiment, the invention provides a method, for example, to generate polynucleotides wherein the polynucleotides contain deletions at more than one position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides; and

(c) optionally, subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion at one position.

A function of interest may then be selected for if desired (step(d)). Further, if desired, steps (a) to (c) or steps (a) to (d) may be repeated from 1 to 50 times or more.

[81] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain deletions at more than one position. It is contemplated that deletions at multiple positions will allow removal of multiple detrimental or unwanted functions of a genetic element. These functions might include any combination of protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins or other functions of interest as will be well appreciated by those of skill in the art.

Insertions

[82] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which inserts at least one nucleotide to at least one end of said polynucleotides;

(c) optionally subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by an insertion at one position.

[83] Further, the invention provides a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. This embodiment of the invention will allow novel fusion of genetic elements to occur. For example, a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases)

could be fused in new ways, or new functions like binding domains (i.e. nucleic acid binding domains, small molecule or ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

[84] Likewise in another embodiment, the invention provides a method to generate polynucleotides wherein the polynucleotides contain insertions at more than one position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions;

(b) subjecting said polynucleotides from step (a) to a process which inserts at least one nucleotide to at least one end of said DNA ends of said polynucleotides; and

(c) optionally, subjecting said polynucleotides from step (b) to a process which covalently joins said DNA ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by an insertion at one position; and

(d) optionally selecting for a function of interest. Steps (a)-(b), (a)-(c) or (a)-(d) may be repeated from 1 to 50 times or more.

[85] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions at more than one position. It is contemplated that this embodiment of the invention will allow multiple novel fusions of genetic elements to occur. For example, the following could be fused to a gene of interest in a combinatorial fashion: a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or new functions like multiple binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

Combinations of insertions and deletions

[86] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of deletions and insertions at a single random position. This method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides;

(c) subjecting said polynucleotides from step (b) to a process which inserts at least one nucleotide to at least one end of said DNA ends of said polynucleotides from step (b);

(d) optionally subjecting said polynucleotides from step (c) to a process which covalently joins said DNA ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion and insertion at one position.

[87] Further, the invention provides a population of polynucleotides, with members differing from one another by a combination of deletions and insertions at a single random position. It is contemplated that this embodiment will allow for new heterologous domains to replace domains in the gene of interest. In this regard, new functions, such as ligand binding or enzymatic catalysis could be conferred upon a genetic element. Also, native function could be enhanced utilizing this embodiment.

[88] In another embodiment, the invention provides a method to generate polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. In this regard deletions may occur at different positions than insertions, or deletions and insertions can occur at the same position. Further, deletions and/or insertions can occur at multiple positions. This method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides;

(c) optionally subjecting said polynucleotides from step (b) to a process which inserts at least one nucleotide to at least one end of said ends of said polynucleotides;

(d) optionally subjecting said polynucleotides from step (c) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion and insertion at one position;

(e) optionally selecting for a function of interest; and optionally repeating any of steps (a) to (d) from 1 to 50 times or more.

[89] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. It is contemplated that this embodiment of the invention will allow for classical directed evolution, wherein multiple rounds of insertions at random positions, deletions at random positions, and combinations of insertions and deletions, are produced with the genetic element being optionally subjected to selection between each round. This embodiment allows for the improvement or alteration of the function of a genetic element.

Starting material

[90] The present invention can be applied to any polynucleotide of interest to the researcher. The polynucleotide can be nucleic acid, i.e. RNA or DNA. Often the polynucleotide will be DNA consisting of genetic elements or one or more genes of interest. The starting material may be obtained through natural sources, or may be polynucleotides which have been synthesized in a laboratory (e.g. gene synthesis), or may be polynucleotides derived from natural sources which have been manipulated in a laboratory. Several sources of polynucleotides are available through publicly held databanks such as Genbank (http://www.ncbi.nlm.nih.gov:80/Genbank/index.html) or available commercially (Celera, Rockville, MD; Incyte, Palo Alto, CA; Clontech, Palo Alto, CA; Invitrogen, Carlsbad, CA).

[91] The nucleic acid may be obtained from any source. for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) [Mullis, U.S. Patent # 4,683,202 (1987); Mullis et al., U.S. Patent # 4,683,195 (1987)]. Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

[92] The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids, phagemids, YACs,

and BACs are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large nucleic acid fragments.

[93] If the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of mutants does not increase.

[94] Starting material should be in substantially pure form. The polynucleotide may be double-stranded or single-stranded, but more preferably is double-stranded. Further, the polynucleotide may be linear or circular, but in a preferred embodiment the polynucleotide is circular. Polynucleotides in circular form may be prepared by preparation of plasmid DNA from organisms such as bacteria, yeast, plants, or mammalian cells by techniques well known to those skilled in the art [Maniatis et al., (1989)]. The number of different specific nucleic acid fragments in the reaction vessel will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

[95] The starting material (i.e. the polynucleotide), while in substantially pure form, can also be present without homologs or related sequences. In other words, the polynucleotides in the initial vessel may all be identical, although they may also be related, unrelated or heterologous. In fact, performance of the present invention will be unaffected by the sequence of the starting material. Furthermore, the sequence of the starting material may be known or unknown. For directed evolution purposes, all that is required is a method to detect the function of the polynucleotide (such as a screening assay).

Cleaving the polynucleotide at a random position

[96] In general, a nucleic acid fragment may be cleaved by a number of different methods. The nucleic acid fragment may be digested with a nuclease such as DNAse I, S1 nuclease, P1 or mung bean nuclease, or RNAse, which are readily available. Other enzymes, such as RAG1 and RAG2, topoisomerases, and integrases are capable of cleaving polynucleotides. The nucleic acid may be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides. Chemical agents, such as bleomycin or methyl methanesulfonate (MMS) can also cleave polynucleotides.

[97] Of substantial importance to the generation of functionally mutated genes containing insertions or deletions is to cleave the polynucleotide a small number of

times, usually between 1 and 10, preferably between 1 and 5, and most preferably once. The present invention provides a means to cleave a polynucleotide such that cleavage occurs only at one position per polynucleotide in the reaction vessel. Of importance is that the present invention provides a means for a near random cleavage of a polynucleotide (i.e. cleavage at several different positions in different molecules). Cleavage can be double-stranded or single-stranded (i.e. produce single-stranded ends or double-stranded ends). Examples of enzymes which can cleave polynucleotides include DNase I, S1 nuclease, P1 nuclease, as well as topoisomerases, transposons, and integrases. Cleavage can occur transiently with enzymes such as topoisomerases, transposons, and integrases. These enzymes may cleave the polynucleotide once, or more than once. S1 nuclease can be used to cleave double or single-stranded polynucleotides in a generally random fashion. In a preferred embodiment, with circular double-stranded DNA, S1 nuclease will cleave the polynucleotide only once, producing two DNA ends (FIG 4).

[98] It is also contemplated that the nucleic acid may also be partially digested with one or more restriction enzymes which cleave DNA at a high frequency (i.e. at several positions within a polynucleotide), such that certain polynucleotides are cleaved only once, and that the resulting population contains polynucleotides cleaved one time, but with different polynucleotides cleaved at different positions. The cleavage with a restriction enzyme may not be entirely random, but if the genetic element of interest has enough specific restriction sites at different positions, the cleavage pattern may be useful enough to generate substantial diversity.

[99] It is contemplated that single cleavage of a polynucleotide can be accomplished through other alternative mechanisms which normally cleave polynucleotides several times. A polynucleotide can be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides. If any of these modalities is carefully titrated and a means of purification is utilized, the singly cleaved molecules can be obtained in substantially pure form (i.e. singly cleaved molecules can be purified away from uncleaved or multiply cleaved molecules).

[100] Furthermore, enzymes which act to cleave and rejoin DNA, such as topoisomerases, transposons, and integrases can be utilized to effectively cleave a polynucleotide [Singh et al., *Proc Natl Acad Sci* 94: 1304-9 (1997)]. In these cases the cleavage and rejoining steps may be coupled. Preferably the DNA ends are linked, or are in physical proximity to one another, following cleavage. This is in order to prevent the re-

ligation of the wrong ends to one another following deletional or insertional events. One mechanism to keep the ends linked is through the use of a circular polynucleotide as a starting material. In this case, the ends are linked by the intervening polynucleotide chain. Thus, the re-ligation will be an intramolecular event as opposed to intermolecular, and will proceed with greater efficiency. Other mechanisms to keep the ends in proximity is through a protein bridge, such as through chromatin (i.e. histones, or other DNA binding proteins), or through enzymes which couple cleavage with rejoining, such as transposons, integrases, or topoisomerases. Alternatively, ends could conceivably be left in proximity to one another through the linkage of opposite ends (the non-cleaved ends) to solid supports.

[101] Cleavage of a circular polynucleotide consisting of supercoiled plasmid DNA can be accomplished by incubating from 0.1 to 100 μg, preferably from 1 to 10 μg with a nuclease such as S1 nuclease. The nuclease can be present in amounts from 0.1 to 1000 units, but preferably from 1 to 100 units in a reaction of 10 μl. The temperature of the reaction can occur from between 0 and 100°C, but preferably between 4 and 50 °C. The reaction time can vary from 30 seconds to 1 hour, but preferably is between about 1 and 30 minutes. The degree of linearization can be measured by analyzing the plasmid DNA on an agarose gel as in FIG. 4. The linear DNA should preferably be purified from the uncut DNA by any of a number of methods well known to those skilled in the art. Such methods include utilization of agarose gel purification kits (Qiagen, Valencia, CA), HPLC, column chromatography and the like.

Deletion of nucleotides

[102] Nucleotide deletions can be generated at a DNA end by a variety of means. For instance, an exonuclease, such as exonuclease III, can be used to remove nucleotides in a 3' to 5' direction from a DNA end. The resulting DNA end then contains a 5' overhang which can be removed by digestion of the DNA with a single-stranded endonuclease such as P1 nuclease, S1 nuclease, or mung bean nuclease. Bal 31 nuclease is an enzyme which possesses 5' to 3' as well as 3' to 5' nucleolytic activity and can be used to delete nucleotides from a DNA end. Furthermore, several polymerases, like DNA polymerase I from *E. coli*, Klenow fragment, and Taq polymerase contain exonuclease activity and could conceivably be used to make deletions from a DNA end. Cell extracts from all organisms contain DNA repair enzymes which can act to delete nucleotides, thus unpure cell extract could conceivably be used as a source for exonuclease activity. Other nucleases, which may not have exonuclease activity under certain conditions may be capable

of producing deletions at a DNA end under other conditions. For example, S1 nuclease can produce short deletions when used at high enzyme concentrations. Furthermore, it is contemplated that mild denaturation of a DNA molecule, such that the DNA ends become "frayed", will allow deletions to occur upon application of a single-stranded endonuclease, such as S1, P1, or mung-bean nuclease.

[103] In a preferred embodiment the conditions of the deletion reaction are set such that the number of individual deletions occurring at each DNA end may be well controlled. For example, altering the salt concentration, altering the pH, altering the temperature, or altering any of the other biochemical parameters of the reaction can change the activity of the nuclease enzyme such that more or less deletions will occur depending on the intent of the investigator (for instance decreasing temperature or increasing salt may lower the processivity of the exonuclease and cause fewer deletions). Figure 5 shows altering conditions allowing differing numbers of deletions to occur on a DNA end. In some cases large deletions might be warranted (i.e. to completely remove a large domain in a genetic element), in other cases small deletions might be preferable (i.e. to remove a single amino acid, or a few amino acids such as those that comprise a protease site). Generally deletions could be obtained numbering from 1 to 1000, more preferably they would be from 1 to 100. In certain instances, as described, the deletions may number from 1 to 10.

[104] Due to cleavage at a random position in the polynucleotide, the location of the deletions in the resulting polynucleotide will also be located at a random position. Also, since residues are deleted from either end of the molecule, the total number of deletions will equal the sum of the deletions occurring on the 5' end and the 3' end.

Adding nucleotides

[105] In order to make additions to a polynucleotide in random positions, the polynucleotide is necessarily cleaved at a random position, as described above. Prior to insertion, nucleotides may be deleted from the DNA ends produced during the cleavage event. Alternatively, the DNA ends formed by the cleavage reaction can be used as substrates to which new nucleotides or polynucleotides are added.

[106] Several different mechanisms exist to add nucleotides to the ends of a polynucleotide. For example, nucleotides can be added by chemical coupling. A polymerase, such as terminal deoxynucleotidyl transferase can be utilized to add nucleotides in a semirandom fashion to a DNA end [Gauss & Lieber, *Mol Cell Biol 1996* 16: 258-69

(1996)]. Alternatively, the cleavage step may be coupled to the insertion event, as can be the case when employing transposons or integrases to the insertion event.

[107] A ligase such as *E. coli* ligase or phage T4 ligase can be utilized to covalently couple a new polynucleotide to the parent polynucleotide. In a preferred embodiment the polynucleotide is a genetic element or a fragment of a genetic element. A genetic element predisposes the resulting polynucleotide to have function since genetic elements are functional in some way by definition. The genetic element may be a gene, the regulatory element of a gene, or a genetic element encoding a useful domain. The genetic element may be a library of genetic elements such as a cDNA library or genomic DNA library. Fragments of a genetic element can be produced by digesting the polynucleotide with a nuclease, such as DNAse I, S1 nuclease, P1 or mung bean nuclease, or RNAse. Other enzymes, such as restriction enzymes and topoisomerases, can also cleave polynucleotides into fragments. The polynucleotides may be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides into fragments. Chemical agents, such as bleomycin or MMS can also cleave polynucleotides into fragments.

[108] It is contemplated that the mixture of a parent polynucleotide cleaved at a random position, with a population of genetic elements or fragments of genetic elements, and a ligase such as T4 DNA ligase, under the appropriate salt, buffer, and temperature conditions, will allow covalent coupling of the genetic elements with the parent polynucleotide at the position of the original cleavage event. Thus, a mixture of polynucleotides is produced comprising an insertion at a random position within the parent polynucleotide. The content (i.e. the sequence) of each insertion may be identical if the genetic elements or fragments of genetic elements are identical, or different if the fragments of genetic elements were non- identical.

Rejoining the DNA ends

[109] DNA ends may be rejoined covalently by incubating the DNA ends with an enzyme like a DNA ligase which will form phosphodiester bonds between nucleotides at the DNA end. Examples of ligases include *E. coli* DNA ligase, phage T4 DNA ligase, or human DNA ligases. These enzymes can be used under conditions well known to those skilled in the art to ligate DNA. Other enzymes are also capable of creating covalent linkages (like phosphodiester bonds) between nucleotides at DNA ends. Such enzymes are topoisomerases, transposons, integrases, and other recombination enzymes. Other

mechanisms can be used to join DNA ends such as the utilization of an oligonucleotide whose sequence can hybridize to sequences on either end (i.e. both the 5' and 3' ends) to "bridge" the ends with hydrogen bonds. The intervening sequence on the opposite strand could be filled in with a polymerase, such as *E. coli* polymerase, Klenow fragment, phage T4 polymerase, or Taq polymerase. Nicks could then be repaired by a DNA ligase as described above. Cellular extracts also contain ligase activities and cell or nuclear extracts could be used to rejoin DNA ends. Alternatively, DNA molecules could be introduced into intact cells and the cell's machinery could rejoin DNA ends by homologous or non-homologous means.

Library compositions

[110] The present invention provides for novel libraries of which the following compositions are examples:

Deletions

[111] The invention provides a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. Such single deletion libraries can contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. Deletion libraries should contain at least one molecule that differs from at least one other molecule by the deletion of at least one nucleotide at one random position. The number of deletions at each position could be from 1 to 1000, but should be at least one. It is contemplated that deletions will allow removal of detrimental or unwanted functions of a genetic element. These functions might include protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

[112] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain deletions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These multiple deletion libraries should contain at least one molecule that differs from at least one other molecule by the deletion of at least one nucleotide at more than one random position. It is contemplated that deletions at multiple positions will allow removal of multiple detrimental or unwanted functions of a genetic element. These functions might include any combination of multiple protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

Insertions

[113] The invention provides a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. Insertion libraries can contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. Insertion libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at one random position. The number of insertions at each position could be from 1 to 10,000, but preferably will be at least one. For example, a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or a new function like binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

[114] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These multiple insertion libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at more than one random position. It is contemplated that this embodiment of the invention will allow novel fusion of genetic elements to occur. It is contemplated that this embodiment of the invention will allow multiple novel fusions of genetic elements to occur. For example the following could be fused to a gene of interest in a combinatorial fashion: a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or new function like binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

Combinations of insertions and deletions

[115] The invention provides a population of polynucleotides, with members differing from one another by a combination of deletions and insertions at a single random position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These combination libraries should contain at least one molecule that differs from at least one other molecule by the insertion of

one nucleotide and the deletion of at least one nucleotide at one random position. It is contemplated that this embodiment will allow for heterologous domains to replace domains in the gene of interest. In this regard, new functions, such as ligand binding or enzymatic catalysis could be conferred upon a genetic element. Also, native function could be enhanced utilizing this embodiment.

[116] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These combination libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at one random position and the deletion of at least one nucleotide at one random position. This embodiment of the invention will allow for classical directed evolution, wherein multiple rounds of insertions at random positions, deletions at random positions, and combinations of insertions and deletions, are produced with the gene of interest being optionally subjected to selection between each round. This embodiment allows for the improvement or alteration of function of a genetic element.

Analyzing the composition

[117] The composition of such libraries can be determined by mechanisms well known to those in the art. In order to determine whether a library contains insertions or deletions, the library can be analyzed by agarose or acrylamide gel electrophoresis and size can be compared to the parental sequence. Other methods, like HPLC, mass spectrometry, column chromatography can be used to identify size differences between polynucleotides. Because the present invention relates to random positions of insertions and deletions, the most definitive method to determining the composition of a library is to subject representative polynucleotides within the composition to sequencing, a method well known to those skilled in the art. Comparison of sequences of representative clones would allow one to determine if deletions or insertions occurred at random positions in different molecules in the library.

[118] The resulting library could be ligated into an expression vector for use as a vehicle to express the resulting variants contained within the library. The nature of the expression vector is described below in the "screening" section.

Screening for a function of interest

[119] In testing a library of polynucleotides for a function of interest, the library should be inserted in an appropriate expression vector. Alternatively, the library can be constructed in an expression vector (i.e. the library comprises an expression vector). The vector used for cloning is not critical provided that it will accept a DNA fragment of the desired size. If expression of the DNA fragment is desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the DNA fragment to allow expression of the DNA fragment in the host cell. For screening in bacterial cells, preferred vectors include the pUC series and the pBR series of plasmids.

[120] The resulting bacterial population will include a number of recombinant DNA fragments having random mutations. This mixed population may be tested to identify the desired recombinant nucleic acid fragment. The method of selection will depend on the DNA fragment desired.

[121] The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids, phagemids, YACs, and BACs are preferred where the specific nucleic acid sequence is larger because these vectors are able to stably propagate large nucleic acid fragments.

[122] If a DNA fragment which encodes for a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the DNA fragments in the population or library may be tested for their ability to bind to the ligand by methods known in the art (i.e. panning, affinity chromatography). If a DNA fragment which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the DNA fragments in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify DNA fragments which confer the desired properties onto the protein.

[123] In the context of the present invention the term "positive polypeptide variants" means resulting polypeptide variants possessing functional properties which has been improved in comparison to the polypeptides producible from the corresponding input DNA sequences. Examples, of such improved properties can be as different as, for example, enhanced or lowered biological activity, increased wash performance, thermostability, oxidation stability, substrate specificity, antibiotic resistance or others that may be of interest.

[124] Consequently, the screening method to be used for identifying positive variants depend on which property of the polypeptide in question it is desired to change, and in what direction the change is desired.

[125] A number of suitable screening or selection systems to screen or select for a desired biological activity are described in the art. For example, Strausberg et al. [Strausberg et al., *Biotechnology (N Y)* 13: 669-73 (1995)] describes a screening system for subtilisin variants having calcium-independent stability. Bryan et al. [Bryan et al., *Proteins* 1: 326-34 (1986)] describes a screening assay for proteases having an enhanced thermal stability.

[126] It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee Wis.). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein, a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

Methods of Effecting Targeted Short Deletions in Nucleic Acids

[127] The ability to make short deletions in a polynucleotide is generally hampered by the high activity and processivity of exonucleases that act at a DNA end. Several methods exist to make large (i.e. more than 100 base) deletions at DNA ends [Sambrook et al., (1989)]. However, methods to create short deletions, such as from 1 to 100 bases or very short deletions like from 1 to 10 bases in a controlled fashion have not been possible. The ability to make such deletions at specific sites is important in the field of protein engineering [Altamirano et al., *Nature* 403: 617-22 (2000)] and is highlighted in the end-joining mechanism of V(D)J recombination, the method which produces the substantial diversity in antibody genes [Smider & Chu, *Sem. Immun.* 9: 189-97 (1997)].

Starting material

[128] The deletion generating mechanism can be applied to any polynucleotide of interest to the researcher. The polynucleotide can be nucleic acid, i.e. RNA or DNA. Often the polynucleotide will be DNA consisting of genetic elements or one or

more genes of interest. The starting material may be obtained through natural sources, or may be polynucleotides which have been synthesized in a laboratory (e.g. gene synthesis), or may be polynucleotides derived from natural sources which have been manipulated in a laboratory. Several sources of polynucleotides are available through publicly held databanks such as Genbank (http://www.ncbi.nlm.nih.gov:80/Genbank/index.html) or available commercially (Celera, Rockville, MD; Incyte, Palo Alto, CA; Clontech, Palo Alto, CA; Invitrogen, Carlsbad, CA).

[129] The nucleic acid may be obtained from any source, for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) [Mullis, U.S. Patent # 4,683,202 (1987); Mullis et al., U.S. Patent # 4,683,195 (1987)]. Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Deletion of nucleotides

[130] Nucleotide deletions can be generated at a DNA end by a variety of means. For instance, an exonuclease, such as exonuclease III, can be used to remove nucleotides in a 3' to 5' direction from a DNA end. Often the resulting DNA end contains a 5' overhang which can be removed by digestion of the DNA with a single-stranded endonuclease such as P1 nuclease, S1 nuclease, or mung bean nuclease. Other exonucleases could also be used in the present invention. Bal 31 nuclease is an enzyme which possesses 5' to 3' as well as 3' to 5' nucleolytic activity and can be used to delete nucleotides from a DNA end. Exonuclease T can remove nucleotides in a 3' to 5' direction. Exonuclease 7 can remove nucleotides in a 5' to 3' direction, and can act at single-stranded ends such as nicks or gaps. Exonuclease I catalyzes the removal of nucleotides from single-stranded DNA in the 3′ to 5′ direction. Lambda exonuclease is a highly processive enzyme that acts in the 5′ to 3′ direction, catalyzing the removal of 5′ mononucleotides from duplex DNA. RecJ is a single-stranded DNA specific exonuclease that catalyzes the removal of deoxynucleotide monophosphates from DNA in the 5′ to 3′ direction. Furthermore, several polymerases, like DNA polymerase I from *e. coli*, Klenow fragment, and Taq polymerase contain exonuclease activity and could conceivably be used to make deletions from a DNA end. Cell extracts

from all organisms contain DNA repair enzymes which can act to delete nucleotides, thus impure cell extract could conceivably be used as a source for exonuclease activity. Other nucleases, which may not have exonuclease activity under certain conditions may be capable of producing deletions at a DNA end under other conditions. For example, S1 nuclease can produce short deletions when used at high enzyme concentrations. Furthermore, it is contemplated that mild denaturation of a DNA molecule, such that the DNA ends become "frayed", will allow deletions to occur upon application of a single-stranded endonuclease, such as S1, P1, or mung-bean nuclease.

[131] In a preferred embodiment, the conditions of the deletion reaction are set such that the number of individual deletions occurring at each DNA end may be well controlled. For example, altering the salt concentration and the temperature, altering the pH, or altering any of the other biochemical parameters of the reaction can change the activity of the nuclease enzyme such that more or less deletions will occur depending on the intent of the investigator. Most particularly and surprisingly we have found that decreasing temperature and/or increasing salt lowers the processivity of the exonuclease and results in more controlled small deletions. Salts used in the reaction may be any salt. Examples of salts include sodium chloride, sodium acetate, potassium chloride, or potassium acetate. Preferably the salt is either sodium chloride or potassium chloride. Salt concentrations can range from 10 mM to 1.0 M, but preferably is between 50 mM and 500 mM. Temperature of the reaction can also vary in the present invention. The temperature can range from 0°C to 30°C, but preferably is between 0°C and 24°C. Figure 5 shows altering conditions allowing differing numbers of deletions to occur on a DNA end. In some cases large deletions might be warranted (i.e. to completely remove a large domain in a genetic element), in other cases small deletions might be preferable (i.e. to remove a single amino acid, or a few amino acids such as those that comprise a protease site). The resulting population of polynucleotides contain variable amounts of deletions at the ends of the starting sequence. Generally deletions could be obtained numbering from 1 to 1000, more preferably they would be from 1 to 100. In a preferred embodiment, the deletions may number from 1 to 30 or even 1 to 10.

Rejoining the DNA ends

[132] In some cases it might be useful to join the DNA ends of a molecule containing a deletion with a second DNA end, such that the deletion now occurs at an internal position. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product is a circular polynucleotide. DNA ends may be rejoined by

incubating the DNA ends with an enzyme like a DNA ligase which will form phosphodiester bonds between nucleotides at the DNA end. Examples of ligases include *E. coli* DNA ligase, phage T4 DNA ligase, or human DNA ligases. These enzymes can be used under conditions well known to those skilled in the art to ligate DNA. Other enzymes are also capable of creating covalent linkages (like phosphodiester bonds) between nucleotides at DNA ends. Such enzymes are topoisomerases, transposons, integrases, and other recombination enzymes. Other mechanisms can be used to join DNA ends such as the utilization of an oligonucleotide whose sequence can hybridize to sequences on either end (i.e. both the 5' and 3' ends) to "bridge" the ends with hydrogen bonds. The intervening sequence on the opposite strand could be filled in with a polymerase, such as *e.coli* polymerase, Klenow fragment, phage T4 polymerase, or Taq polymerase. Nicks could then be repaired by a DNA ligase as described above. Cellular extracts also contain ligase activities and cell or nuclear extracts could be used to rejoin DNA ends. Alternatively, DNA molecules could be introduced into intact cells and the cell's machinery could rejoin DNA ends by homologous or non-homologous means.

Deletion compositions

[133] In one embodiment the current invention provides for a composition of polynucleotides, wherein members of the population differ from one another by the presence of deletions at one or both ends of the polynucleotide. The number of deletions may range from 1 to 100 at each end, but more preferable is from 1 to 30.

[134] Additionally, the current invention provides for a composition of polynucleotides differing from one another by short deletions at a specific internal position (i.e. not at an end). This composition is obtained by joining the composition of polynucleotides with deletions at the ends to other DNA ends, such that the deletion now occurs internally. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product is a circular polynucleotide. The number of deletions may range from 1 to 100 at each end, but more preferable is from 1 to 30.

[135] All references and patent publications referred to herein are hereby incorporated by reference herein.

[136] As can be appreciated from the disclosure provided above, the present invention has a wide variety of applications. Accordingly, the following examples are offered for illustration purposes and are not intended to be construed as a limitation on the invention in any way.

EXAMPLES

Example 1: Random cleavage of a plasmid

[137] Molecular evolution techniques utilizing insertions or deletions require a gene to be cleaved, at least transiently, a small number of times. Optimally, each molecule within a mix is cleaved once, at different random positions. There is significant difficulty in preparing singly cleaved DNA, wherein cleavage occurs at random positions. Biondi, et.al. described a cumbersome method using DNase I and DNA polymerase to induce nicks, followed by further cleavage of these nicks to produce a double stranded break [Biondi et al., *Nucleic Acids Res* 26: 4946-52 (1998)]. This process required tedious and time consuming cesium chloride gradient purification and linker ligation steps, and is not generally applicable to high throughput molecular biology techniques like molecular evolution.

[138] The strategy of utilizing a single-stranded endonuclease to induce double-stranded breaks at random positions in DNA has heretofore not been utilized. It was reasoned that a single-stranded nuclease, like S1, P1, or mung bean nuclease, would specifically cleave single-stranded regions in tightly supercoiled DNA, thus producing a nick. A nick is the natural substrate for these enzymes, so cleavage to produce a double-stranded break may then occur in the same reaction. Following cleavage, the single-stranded regions are no longer present since the plasmid is no longer supercoiled, so the DNA is no longer a substrate for the enzyme. Thus, cleavage would occur once and only once. This example illustrates the utility of this hypothesis.

[139] The plasmid pLacZi (Clontech, Palo Alto, CA) was used to illustrate the mechanism by which a polynucleotide can be cleaved at random positions. The plasmid was propagated in DH10B *E. coli* cells (Invitrogen, Carlsbad, CA) and plasmid was prepared by Qiagen maxiprep columns (Qiagen, Valencia, CA). Plasmid DNA at 200 ng/μl was incubated with 0.4, 2.0, 10, or 50 units of S1 nuclease (Promega, Madison, WI) in 1X S1 buffer (50 mM sodium acetate pH 4.5, 280 mM NaCl, 4.5 mM $ZnSO_4$) for 10 minutes at room temperature. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C for 10 minutes. Protein was removed by twice extracting with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate and resuspended in water.

[140] Cleaved pLacZi was analyzed by 1.5% agarose gel electrophoresis (Figure 4, panel A). S1 nuclease cleaved plasmid was seen to co-migrate with pLacZi cleaved with Cla I, which cuts pLacZi once. Thus, S1 nuclease can linearize a circular DNA

molecule. Although S1 nuclease is not known to cut DNA in a sequence specific manner, it was important to determine that the cleavage of plasmid by S1 was not site specific. To this end, linear plasmid produced by S1 cleavage was gel purified (Figure 4, panel B, lane 5), or purified and further cleaved with Cla I (lane 6). Controls included supercoiled plasmid (lane 2), plasmid linearized with Cla I (lane 3), or plasmid linearized with S1 nuclease and un-purified (lane 4). The S1/Cla I cleaved plasmid is seen as a smear, showing that S1 is cleaving in several different positions in the plasmid. If S1 cleaved at only one position, then the S1/Cla I cleaved plasmid would migrate as two bands; if S1 cleaved at two positions, then the S1/Cla I plasmid would migrate as three bands, and so on. The importance of this example is that a polynucleotide is able to be cleaved once (i.e. linearization of a circle), and only once, at different positions.

Example 2: Deletions at a site in LacZ

[141] Nucleotide deletions have been made for structural analysis of genes, and for nucleotide sequence analysis. Generally these deletions are large, in the range of well over 100 nucleotides. Under normal conditions, for example, exonuclease III removes over 100 bases per minute [Sambrook et al., (1989)]. The ability to create small deletions, however, would be useful to alter small domains in proteins or remove deleterious functions. In order to make small deletions at the end of a polynucleotide, exonuclease III was utilized under various conditions of salt (Figure 5) and temperature. A fluorescently labeled 232 base pair PCR product from pLacZi was exposed to 100 mM, 150 mM, and 200 mM NaCl in the presence of 10 U exonuclease III (New England Biolabs, Beverly, MA) in 10 µl of 66 mM Tris-Cl (pH 7.4), 0.66 mM $MgCl_2$ at 15°C in a 5 minute reaction. The reaction was stopped by the addition of EDTA to 0.025 M, and extracted once with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. DNA was resuspended in 20 µl deionized formamide, and 0.5 µl was run on a 6% polyacrylamide denaturing gel in ABI 373 sequencer (Perkin-Elmer, Foster City, CA) set to the genescan setting according to the manufacturers recommendation.

[142] Nearly 25 nucleotides can be removed under conditions of 100 mM NaCl (Figure 5, second panel), up to 15 nucleotides with 150 mM NaCl, and a few nucleotides with 200 mM NaCl (bottom panel).

[143] The Cla I site in pLacZi exists in the coding region of the LacZ gene. This site was utilized to make short deletions within the gene itself, which could then be analyzed further by PCR to determine the extent to which deletions were made. Additionally,

plasmids containing deletions were selected on LB agar plates containing 40 μg/ml X-Gal to determine the functionality of the LacZ gene. The pLacZi plasmid (10 μg) was linearized with Cla I in 200 μl, then incubated with 20 U of S1 nuclease in 400 μl to remove the 2 bp 5' overhangs. Further, the linearized plasmid was concentrated and filtered through an ultrafree MC membrane (30 kD cutoff, Millipore, Bedford, MA), then brought to a volume of 400 μl in 1X calf intestinal phosphatase buffer containing 100 U of calf intestinal phosphatase (New England Biolabs, Beverly, MA) and incubated for 45 minutes at room temperature. Plasmid was extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate, and resuspended in water. The plasmid was then incubated with exonuclease III as described in example 1, in the presence of either 100 mM, 150 mM or 200 mM NaCl for 5 minutes at 15°C in a 10 μl reaction. In a control arm, plasmid was not incubated with exonuclease III, to test for the frequency of religation of the dephosphorylated plasmid in the absence of deletions. After 5 minutes of exonuclease III reaction, a mix containing S1 nuclease 50 U in 1X S1 buffer was added. This mix was further incubated at room temperature for 15 minutes. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C for 10 minutes. The DNA was then extracted once with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate and resuspended in 10 μl of 1X ligase buffer containing 1.0 U of T4 DNA ligase (Invitrogen, Carlsbad, CA). Ligation reactions were incubated at 15°C for 12 hours. Electroporation of *E. coli* strain DH10B (Invitrogen, Carlsbad, CA) was accomplished with 1.0 μl of ligation mix. Cells were plated on LB agar plates containing 40 μg/ml X-Gal and 100 μg/ml ampicillin and incubated overnight at 30°C. Table 1 illustrates the results of the plating experiment.

Table 1. Colony characteristics after site directed deletions.

	Blue Colonies	White Colonies	Blue/White
No Exo III	0	0	-
Exo III, 100 mM NaCl	177	66	0.37
Exo III, 150 mM NaCl	340	140	0.41
Exo III, 200 mM NaCl	77	34	0.44

[144] Notably, no background is realized when dephosphorylated plasmid is not exposed to exonuclease III (first row, Table 1). Several blue and white colonies are evident with exonuclease III treatment under different salt concentrations. Interestingly, the theoretical maximum of the blue/white ratio is 0.33, since at least 2/3 of religations should be out of frame. However, the blue/white ratio in this experiment is slightly more than 0.33, and appears to increase as salt concentration increases. This bias may be due to the fact that a one basepair deletion from one end would allow in-frame religation to occur, and fewer deletions are favored as salt is increased. The statistical significance of this result has not been analyzed, so the true frequency may actually be nearer to 0.33.

[145] Six of the colonies were analyzed by PCR with primers flanking the Cla I site. FIG 6 shows these results. In the upper panel the wild-type 312 basepair fragment from pLacZi is shown. Clone 1 contains an in frame deletion of 291 bases (PCR product of 291 bases) and retains a blue phenotype. Clone 2 contains a 4 basepair out of frame deletion (PCR product of 308 bases) and has a white phenotype. Clone 3 contains a 9 basepair in frame deletion (PCR product of 303 bases) and has a white phenotype. Clone 4 contains a 6 basepair in frame deletion (PCR product of 306 bases) and has a white phenotype. Clone 5 contains a 7 basepair out of frame deletion (PCR product of 305 bases) and has a white phenotype. Clone 6 has a 3 basepair deletion (PCR product of 309 bases) and has a blue phenotype. Although it may be thought that shorter deletions would lead to less severe phenotype, this experiment illustrates that this is not necessarily the case. Clone 1 contains a deletion encompassing 7 amino acids but retains function whereas clones 3 and 4 contain in frame shorter deletions but do not retain function. Furthermore, this example illustrates the ability of deletional technology to search functional sequence space.

Example 3: Insertions in LacZ

[146] Insertions of random DNA in the LacZ gene was accomplished by employing DNase I to fragment cDNA derived from CHO cells, followed by ligation of these fragments into linearized pLacZi. Since cDNA is by definition functional, it is contemplated that the use of cDNA will optimize the likelihood of obtaining functional proteins. CHO cell cDNA (5 μg) was fragmented with 0.001 units of DNase I in a buffer containing 40 mM Tris-Cl pH 7.4 and 10.0 mM $MgCl_2$ for 5 minutes at room temperature. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C in the presence of 10 μg of protease K. DNA was extracted with an equal volume of phenol:chloroform:isoamyl alcohol

(25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. Plasmid linearized with Cla I or S1 nuclease were dephosphorylated as described above, then again extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. To insert random cDNA fragments into plasmid DNA, 0.2 mg of linearized, dephosphorylated plasmid was incubated with 1 ng of cDNA fragments in the presence of T4 DNA ligase (1.0 U) in a reaction volume of 10 ml at 15°C for 12 hours. As controls, linearized plasmid was incubated with ligase in the absence of cDNA fragments, and cDNA fragments were incubated with ligase in the absence of linearized vector. DH10B *E. coli* were then electroporated with 1.0 µl of each ligation mix.

[147] Several *e. coli* colonies were identified in the vector plus insert arms of the experiment which exhibited either white, intermediate, or blue phenotype on X-Gal plates. PCR across the Cla I site in the colonies which arose from vector linearized with Cla I ligated to cDNA fragments revealed several clones containing inserts of sizes from 100-300 basepairs. Three of these are illustrated in FIG 7. Thus, the insertion of fragments of cDNA into a genetic element can be accomplished with the present invention.

Example 4: Functional changes at random positions

[148] The lac operon is a model system by which genetic elements are easily studied. The enzyme β-galactosidase is encoded by the LacZ gene, but is normally only produced when lactose is present in the environment. Control of enzyme levels is accomplished at the level of transcription. The lac repressor protein binds to the operator sequence upstream from the ATG start site of LacZ, and inhibits transcription by RNA polymerase. In the presence of lactose, however, the repressor is removed from the operator and transcription can proceed. The mechanism of promotor activation is through the binding of lactose, the inducer, to the lac repressor and causing an allosteric change that causes its affinity for the operator to decrease dramatically. In the laboratory setting, LacZ transcription can be assessed by plating *E. coli* on the colorimetric substrate X-Gal, which causes colonies to turn blue when hydrolyzed by β-galactosidase. The operator can be de-repressed by utilizing the lactose analog IPTG, which is non-hydrolizable, and strongly induces LacZ transcription by binding the lac repressor.

[149] In order to test the ability of random deletions to affect gene function, the pBluescript II KS+ plasmid was linearized with S1 nuclease, gel purified, dephosphorylated, and subjected to exonuclease III digestion as described in examples 1 and

2. Linearized plasmid at 20 ng/μl was incubated with 10 U exonuclease III in 66 mM Tris-Cl pH 7.4, 0.66 mM $MgCl_2$ buffer at 15°C for 5 minutes, followed by addition of 1X S1 solution containing 50 mM sodium acetate pH 4.5, 280 mM NaCl, 4.5 mM $ZnSO_4$ and 10 U S1 nuclease, and incubation for 15 minutes at room temperature. The reaction was stopped by adding EDTA to 0.025 M, and extraction with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. DNA was resuspended in 1X T4 DNA ligase buffer containing 1.0 U T4 DNA ligase and incubated at 15°C for 12 hours. The ligation reaction (1 μl) was then used to electroporate *E. coli* strain TOP 10 F', which produces the lac repressor protein (Invitrogen, Carlsbad, CA). The *E. coli* were incubated on LB plates either with or without IPTG as inducer, and in the presence of X-Gal to measure β-galactosidase activity. Additionally, pBluescript plasmid was plated in the presence or absence of IPTG on X-Gal containing plates. Table 2 illustrates the results of the experiment.

Table 2. Functional changes in transcription of β-galactosidase

	+ IPTG		- IPTG	
	Blue	**White**	**Blue**	**White**
pBluescript	100%	0	0	100%
pBluescript/deletions	66%	34%	2%	98%

Several colonies gained the ability to transcribe LacZ in the absence of the inducer IPTG in the arm of the experiment where deletions were made at random positions. Additionally, several colonies lost their ability to produce functional β-galactosidase in the presence of IPTG. One white colony in the presence of IPTG from the pBluescript/deletions arm was sequenced and found to have an eight basepair deletion at the translation start site. This sequence is illustrated below, with the translation start site (ATG) encoding methionine codon underlined.

CACACAGGAAA--------ACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

(SEQ ID NO: 1 and SEQ ID NO:2, respectively)

Thus, random cleavage of a plasmid, followed by short deletions made by exonuclease III can cause functional changes in regulatory and protein coding regions of genetic elements. These changes can then be detected with a functional assay.

SEQUENCE LISTING

SEQ ID NO:1

Mutation in 5' end of gene encoding β- galactosidase

CACACAGGAAAACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

SEQ ID NO:2

5' end of wild type gene encoding β- galactosidase

CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGA

ACAA

WHAT IS CLAIMED IS:

1. A method for generating a library of polynucleotide sequences having nucleotide deletions at differing positions in a sequence of a genetic element comprising the steps of:

(a) subjecting multiple copies of circular polynucleotides comprising the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one of said ends of said polynucleotides producing a library of deletion polynucleotide sequences, said library comprising multiple deletion polynucleotide sequences with deletions at different random positions.

2. The method of claim 1, further wherein said polynucleotides from step (b) are subjected to a process that covalently joins said 3' and 5' ends to one another.

3. The method of claim 1, wherein said library of polynucleotides is further subjected to a process that selects for a function of interest.

4. The method of claim 1, wherein the cleavage occurs with an endonuclease.

5. The method of claim 4, wherein the endonuclease is S1.

6. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 5 individual polynucleotides each having a random deletion at a different position from the others.

7. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 10 individual polynucleotides each having a random deletion at a different position from the others.

8. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 30 individual polynucleotides each having a random deletion at a different position from the others.

9. The method of claim 1, wherein the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element.

10. The method of claim 1, wherein steps (a) and (b) are repeated.

11. The method of claim 1, wherein step (b) further includes a process for inserting nucleotides at the position of deletion.

12. The method of claim 1, wherein 1-3 nucleotides are deleted in step (b).

13. 13. The method of claim 1, wherein 50-100 nucleotides are deleted in step (b).

14. A substantially pure composition comprising a library of multiple linear polynucleotides each having a different 3' and a 5' end, but each linear polynucleotide being identical to the others if circularized.

15. The composition of claim 14, wherein said library comprises at least 5 polynucleotides having a different 3' and a 5' end.

16. A substantially pure composition comprising a library of at least 2 deletion polynucleotides each differing from the other only by having a different random deletion.

17. The substantially pure composition of claim 16, wherein said deletion polynucleotides further comprise at least one nucleotide inserted at the position of deletion.

18. The composition of claim 16, wherein the library has at least 5 polynucleotides each differing from the other only by having a different random deletion.

19. A method for generating a library of polynucleotide sequences having nucleotide additions at random positions in a genetic element comprising the steps of:

(a) subjecting a composition of multiple copies of circular polynucleotides with the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which adds at least one nucleotide to one of said ends of said polynucleotides producing a library of

addition polynucleotide sequences, said library comprising multiple addition sequences with additions at different random positions.

20. The method of claim 19, further wherein said addition polynucleotides from step (b) are subjected to a process that covalently joins said 3' and 5' ends to one another.

21. The method of claim 19, further subjecting said library of polynucleotides to a process that selects for a function of interest.

22. The method of claim 19, wherein the cleavage occurs with an endonuclease.

23. The method of claim 22, wherein the endonuclease is S1.

24. The method of claim 19, wherein the library of addition polynucleotides comprises at least 5 individual polynucleotides each having a random addition of nucleotides at a different position from the others.

25. The method of claim 19, wherein the library of addition polynucleotides comprises at least 10 individual polynucleotides each having a random addition at a different position from the others.

26. The method of claim 19, wherein the library of addition polynucleotides comprises at least 30 individual polynucleotides each having a random addition at a different position from the others.

27. The method of claim 19, wherein the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element.

28. The method of claim 19, wherein steps (a) and (b) are repeated.

29. The method of claim 19, wherein step (b) includes a process for deleting nucleotides at the point of addition.

30. The method of claim 19, wherein 1-3 nucleotides are added in step (b).

31. 31. The method of claim 19, wherein 3-50 nucleotides are added in step (b).

32. The method of claim 19, wherein 50-100 nucleotides are added in step (b).

33. A substantially pure composition comprising a library of at least 2 addition polynucleotides each differing from the other only by having a different random addition.

34. A substantially pure composition comprising a library of at least 5 addition polynucleotides each differing from the other only by having a different random addition.

35. A method for producing short deletions from the end of a polynucleotide by incubating a population of polynucleotides with an exonuclease at a temperature from 0°C to 24°C in the presence of 10 to 500 mM salt, thereby producing a population of polynucleotides containing deletions of 1-100 residues from at least one end of the polynucleotide.

36. The method of claim 35, wherein the polynucleotide is double-stranded.

37. The method of claim 35, wherein the exonuclease is exonuclease III.

38. The method of claim 36, wherein the double-stranded nucleic acid is incubated with a single-stranded endonuclease to produce a blunt end.

39. The method of claim 35, further wherein the resulting population of polynucleotides containing deletions at the ends are covalently joined to at least a second end, producing a population of polynucleotides containing a deletion at an internal position.

40. The method of claim 38, wherein the single-stranded endonuclease is S1 nuclease.

41. The method of claim 39, wherein the polynucleotides resulting from covalent joining are circular polynucleotides.

42. The method of claim 35 wherein the population of polynucleotides contains deletions of 1-50 residues from at least one end of the polynucleotide.

43. The method of claim 35, wherein the population of polynucleotides contains deletions of 1-30 residues from at least one end of the polynucleotide.

44. A substantially pure composition of at least two polynucleotides each having two ends and each differing from one another only by having different deletions of 1 to 100 residues at one or both ends.

45. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 50 residues at one or both ends.

46. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 30 residues at one or both ends.

47. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 10 residues at one or both ends.

48. A substantially pure composition of at least two polynucleotides each differing from one another only by deletions of 1 to 100 residues at a specific internal position within the polynucleotides.

49. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 50 residues at the specific internal position.

50. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 30 residues at the specific internal position.

51. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 10 residues at the specific internal position.

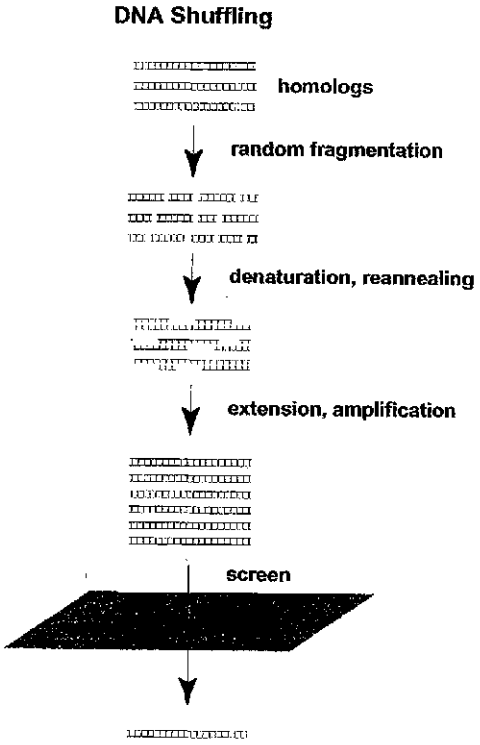

DNA Shuffling
homologs
random fragmentation
denaturation, reannealing
extension, amplification
screen

Fig 1

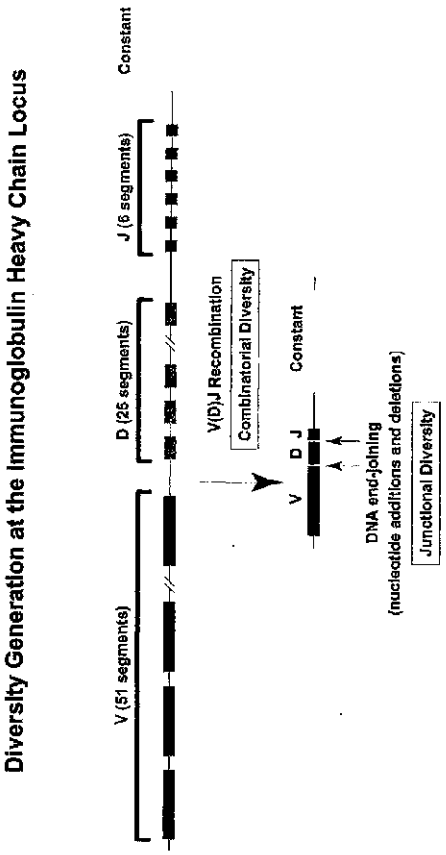

Diversity Generation at the Immunoglobulin Heavy Chain Locus
V (51 segments)
D (25 segments)
J (6 segments)
Constant
V(D)J Recombination
Combinatorial Diversity
V D J
Constant
DNA end-joining
(nucleotide additions and deletions)
Junctional Diversity

Fig 2

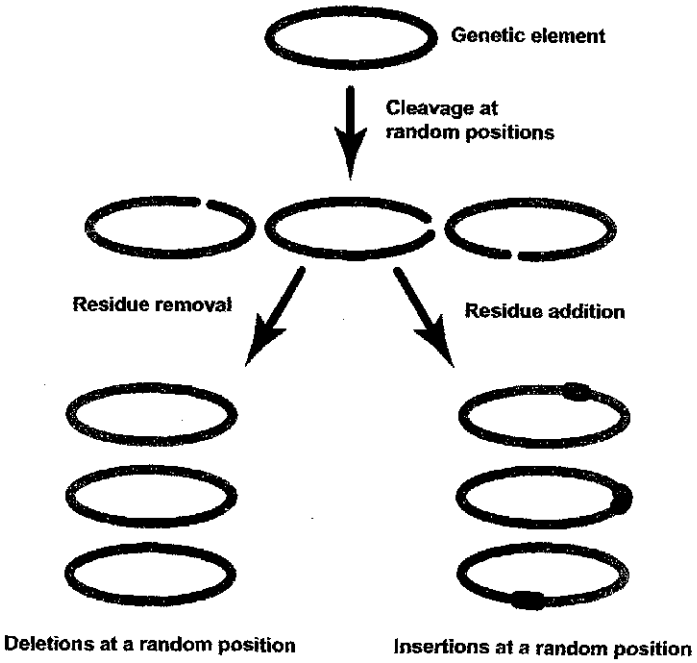

Genetic element
Cleavage at random positions
Residue removal
Residue addition
Deletions at a random position
Insertions at a random position

Fig 3

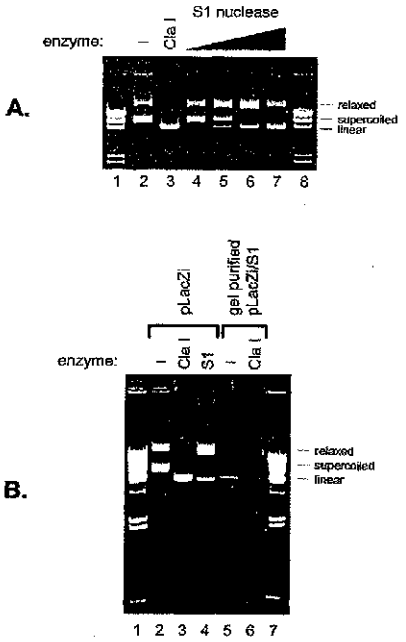

A.
S1 nuclease
enzyme:
–
Cla I
relaxed
supercoiled
linear
1 2 3 4 5 6 7 8
B.
pLacZI
gel purified pLacZI/S1
enzyme:
–
Cla I
S1
–
Cla I
relaxed
supercoiled
linear
1 2 3 4 5 6 7

Fig 4

Short Deletions at a DNA End

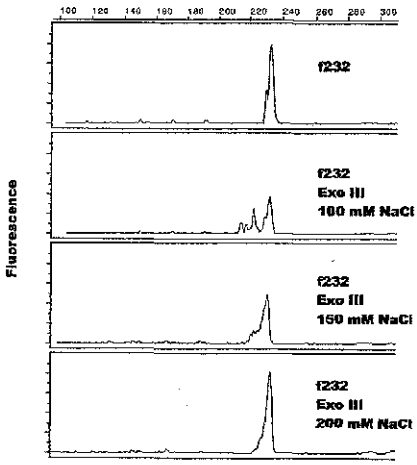


Fig 5

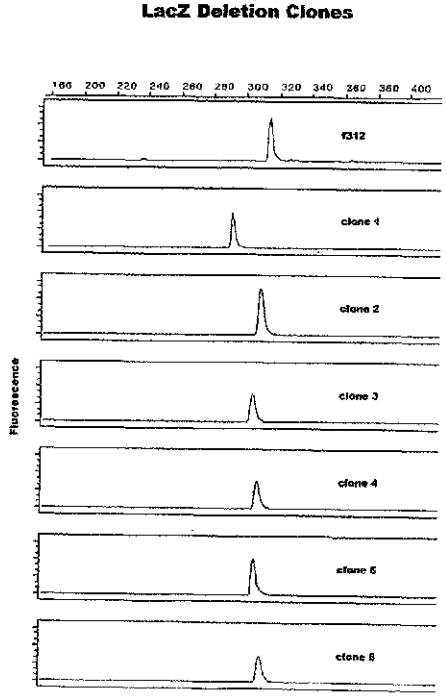

LacZ Deletion Clones
160 200 220 240 260 280 300 320 340 360 380 400
f312
clone 1
clone 2
clone 3
clone 4
clone 5
clone 6
Fluorescence

Fig 6

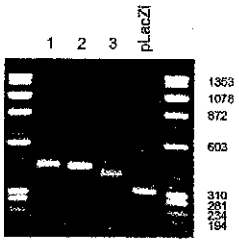


Fig 7

【国際公開パンフレット（コレクトバージョン）】

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 February 2002 (28.02.2002)

PCT

(10) International Publication Number
WO 02/016642 A1

(51) International Patent Classification⁷: C12Q 1/68, C12P 19/34, C12N 15/64, C07H 21/02

(21) International Application Number: PCT/US01/25788

(22) International Filing Date: 17 August 2001 (17.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/226,477 18 August 2000 (18.08.2000) US

(71) Applicant (*for all designated States except US*): INTEGRIGEN, INC. [US/US]; 42 Digital Drive, Unit 6, Novato, CA 94949 (US).

(72) Inventor; and

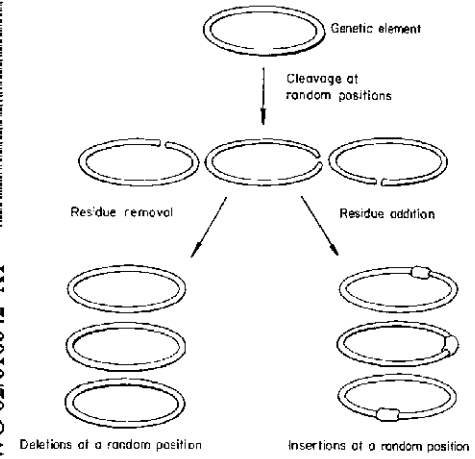
(75) Inventor/Applicant (*for US only*): SMIDER, Vaughn [US/US], 1823-A Pearl Street, Alameda, CA 94501 (US).

(74) Agents: WEBER, Kenneth, A. et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111-3834 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW

[Continued on next page]

(54) Title: METHODS AND COMPOSITIONS FOR DIRECTED MOLECULAR EVOLUTION USING DNA-END MODIFICATION



(57) Abstract: Methods as depicted in Figure 3, for directed evolution are described where genetic elements are randomly cleaved to permit the deletion or addition of polynucleotides or both to create a library of related genetic elements with additions or deletions. Corresponding library populations are also described. These processes allow a significant sampling of sequence space which is necessary for directed evolution of genes. Further described are methods for effecting very small nucleotide deletions in genetic elements of interest.

WO 02/016642 A1

WO 02/016642 A1

(84) Designated States *regionals*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(48) Date of publication of this corrected version:
27 March 2003

(15) Information about Correction:
see PCT Gazette No. 13/2003 of 27 March 2003, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette

METHODS AND COMPOSITIONS FOR DIRECTED MOLECULAR EVOLUTION USING DNA-END MODIFICATION

FIELD OF THE INVENTION

[01] The invention relates to directed evolution, which encompasses methods that can be applied to genetic engineering and protein engineering. Directed evolution is used to evolve gene sequences with the goal of improving or altering gene or protein function. Directed evolution can be applied to many areas including, but not limited to, pharmaceutical development, bioremediation, bioleaching, and the chemical industry.

BACKGROUND OF THE INVENTION

[02] Recently attempts have been made to simulate the process of evolution *in vitro*, thereby inducing genetic changes in specific genes to alter or improve their functions. Although techniques that alter genes have been known for several years, generally detailed features about the encoded protein's structure and function were required for these methods to be successful. The technique of DNA shuffling overcame this barrier to a certain extent, and has been applied to evolve several genes successfully in the past few years [Minshull & Stemmer, U.S. Patent # 5,837,458 (1998)].

[03] Natural evolution occurred over millions of years for genes in the environment. *In vitro* evolution attempts to mimic the natural process in days or weeks. In order for an *in vitro* strategy to succeed, several facets of evolutionary theory must be understood. First, the concept of sequence space defines the total number of possible sequences of a protein of a given length [Kauffman, (1993)]. Thus,

$$S = 20^N$$

where S, the sequence space, is the number of possible sequences, and N is the length of the protein. In *in vitro* evolution experiments, it would be optimal to search S sequences of a given protein in order to identify the fraction of those with the most improved or altered activity. It can quickly be seen that a protein with a modest 50 amino acids has an S of 20^{50} possible different sequences, a number which is virtually infinite in terms of analysis with current molecular biology techniques. Second, it is clear that most amino acid changes are deleterious to proteins. These changes may render the protein inactive, cause disruptions in proper folding, or cause instability to the protein or mRNA *in vivo*. It has been estimated that

the ratio of advantageous to deleterious mutations on average is 1 in 10^5 [Radman et al., *Ann. N.Y. Acad. Sci.* 870: 146-55 (1999)]. In this regard, mutation rate is an important parameter when mutating genes to improve their function. If the mutation rate is too high, deleterious mutations will occur in *cis* with the advantageous mutations, a condition which makes the genes with advantageous mutations impossible to identify since the resulting proteins that contain them will be inactive due to concomitant deleterious mutations. Third, in order to overcome the consequences of higher mutation rates, homologous recombination may be utilized to remove deleterious mutation through double crossover events. Fourth, any *in vitro* evolution technique requires a selection screen in order to identify those sequences which improve or alter the function of the protein.

[04] A major barrier to current molecular evolution is the inability to efficiently search sequence space for a given protein. In this regard, of critical importance is the ability to generate and identify sequences that differ at more than one residue, wherein those differences may have an additive effect upon protein function. This additive effect may be described at amino acid interdependence. For example, a protein with a single mutation at residue *i* may not have any detectable increase in function unless a concomitant mutation at *j* is also present. In this case, in order for evolution to be successful, all possible two-mutant variants of the target sequence should be sampled and tested for improved function. In general, the number of R-mutant variants of a protein of length N is given by:

$$S_R = \frac{20^R N!}{(N-R)! R!}$$

where R is the number of mutant variants, and 20 denotes the number of amino acids possible at each position. Thus, for a protein of length 50, there are 490,000 different two mutant variants.

[05] In these statistical analyses of sequence space, a critical value is the length of the protein (i.e. the number of R-mutant variants depends on the length of the protein). However, in nature the length of any given protein may not be as critical for its function as the arrangement of amino acid residues in three dimensional space. Indeed, a hypothetical concept of "catalytic task space" has been proposed to account for this principle (Kauffman, 1993). Alterations in amino acid residues without altering protein length, N, may not affect the three dimensional structure of the protein in the same ways that increasing or decreasing N would. Alternatively, changing N may not change the biological function of the protein at all. Analysis of virtually any family of homologous proteins reveals that members

have different lengths, sometimes with substantial insertions or deletions, but may retain indistinguishable biological function. Thus, the above formula probably does not give an accurate view of the various R-mutant variants that should be screened when searching for improved or altered biological function. In the laboratory it would be optimum to search all of the R-mutant neighbors of a protein plus a number of variants with nucleotides either added or deleted at every position.

[06] In the case of deletions, the number of D-mutant deletions would be given by,

$$S_D = \frac{N!}{(N-D)!\,D!}$$

where N is the initial length of the protein and D is the number of positions where a deletion occurs. In the case of amino acid additions, a similar formula for all possible additions accounts for the fact that any of the 20 amino acids may be added at any position:

$$S_A = \frac{20^A\ N!}{(N-D)!\,A!}$$

where A is the number of possible addition mutants. In the cases of the addition and deletion mutants, these formula both assume that only one amino acid is added or deleted at each position. In terms of *in vitro* molecular evolution, however, it would be optimal to search all 1, 2, 3....C number of amino acids added or deleted at each position. Thus, for deletion mutants, the number of sequences with variable amino acids deleted at each position,

$$S_{CD} = \frac{C_D\ N!}{(N-D)!\,D!}$$

where C_D denotes the number of amino acids deleted at each position and D is the number of positions where deletions occur. For addition mutants, the formula becomes:

$$S_{CA} = \frac{C_A\ 20^A\ N!}{(N-A)!\,A!}$$

where C_A is the number of amino acids added at each position, and A is the number of positions where additions occur.

[07] Since current molecular biology techniques only allow a fraction of the total space to be generated and sampled for a given protein, a formula which describes an experimental space to be generated can be defined. This expression will also allow the monitoring of improvements in library construction techniques and allow analysis of the

space which is relevant to protein function. A total space to be searched experimentally can be defined as,

$$S_{EX} = S_R S_{CD} S_{CA}$$

where amino acids are mutated to other residues (S_R), are deleted (S_{CD}), or added (S_{CA}) in various combinations and permutations. Certainly current molecular biology techniques would allow a library to be created where R = 1 so $S_R = N$, D = 1 so $S_{CD} = N$, and A = 1 so $S_{CA} = 20$ N. For a protein of N = 50 then, this hypothetical library would contain $20 * N^2 = 2.5 \times 10^6$ different sequences where all permutations of one position changes, deletions, and additions are represented.

[08] The above discussion on the sequence space relevant to protein evolution may be applied in different ways to the *in vitro* engineering of evolved sequences. In nature, the evolution of different catalytic activities in families of enzymes can be grouped into two broad categories: 1) Those where the active site amino acids remain the same, but differences in the structural folds cause the enzymes to have different substrate specificities [Perona & Craik, *J. Biol. Chem.* 272: 29987-90 (1997)], and 2) Those where the enzyme structures are the same, but differences in the active site residues cause the enzymes to catalyze different reactions [Babbitt & Gerlt, *J. Biol. Chem.* 272: 30591-4 (1997)]. An example of the former is the serine protease family, and of the latter is the enolase superfamily.

[09] Although the differences between these categories might seem trivial, they have important implications for the method of molecular evolution and the concept of sequence space. For the enzymes in families with similar structural folds, a molecular evolution approach which samples sequence space throughout the length of the protein would likely be the optimal strategy to alter an enzyme's specificity, since the catalytic active site will likely require the same residues for the catalytic mechanism. However, for the second type of enzymes, increasing sequence space searching through the entire length of the protein is probably not necessary. More likely, increasing the sequence space sampling of the key catalytic domains will optimize the molecular evolution process. In this regard, it would be better to alter five key amino acids to each of the twenty amino acids and sample this more limited space (20^5) rather than to sample a sequence space of 20^5 spread out over the entire gene sequence. Additionally, sampling as many of the possible addition or deletion mutants in key regions would also contribute to possible success of the *in vitro* evolution protocol. Thus, a method which would optimize molecular evolution of a gene belonging to the second type of enzyme family would be a very important and robust technique.

[10] The swapping of genetic domains is an efficient means to evolve new or improved function in biomolecules. While the alteration of single nucleotide residues can affect gene and protein function, the wholesale exchange of multiple residues in a gene can have dramatic effects on protein function. For example, *E. coli* and *Salmonella* are highly related bacterial species, however the differences in their genetic content are due almost entirely to genetic swapping events as opposed to single residue changes. Additionally, swapping events where exchanges of large chunks of DNA create genes are thought to have occurred several times in pathways such as the clotting cascade, as well as to create novel transcription cassettes through transposition [Bell, (1997); Patthy, (1999)].

[11] A well known example of molecular evolution occurring naturally is that which underlies the production of antibodies in the immune system. In mammalian pre-lymphocytes natural molecular evolution occurs successfully on a daily basis. Antibodies are capable of binding a bewildering array of different antigens, yet have similar amino acid sequences and secondary structures. Antibody genes are arranged in the germline as gene segments (Figure 2). During lymphocyte maturation these segments (named variable or "V", diversity or "D", and junctional or "J") are juxtaposed to one another in the process termed V(D)J recombination to create a functional antibody or T-cell receptor gene. Multiple V, D, and J segments allow a substantial amount of diversity, and hence different antigen binding specificities, to be created in the final repertoire of lymphocytes. The diversity created by this mechanism is referred to as combinatorial diversity. Another type of diversity is also created during V(D)J recombination, which is as important as combinatorial diversity [Davis & Bjorkman, *Nature* 334: 395-402 (1988)]. This diversity is termed junctional diversity, and is created when nucleotides are lost or gained at the joints of the gene segments. Importantly, these joints encode the regions of the antibody molecule which contact antigen, so this type of diversity is critical to creating a diverse, but functional immune system.

[12] The two types of diversity utilized by the immune system might be characterized in the following way with regards to the practice of molecular evolution. Generation of combinatorial diversity in immunoglobulin genes allow a sampling of the total sequence space by providing multiple functional V, D, and J gene segments, each member of which is slightly different in sequence but still homologous to other members of the family of segments. In this respect, the combinatorial rearrangement of V, D, and J gene segments functions as a "domain swapping" event in order to generate novel antibody genes. Generation of junctional diversity allows a greater local sampling of sequence space at the

critical residues for contacting antigen through the mechanism of adding or deleting random nucleotides at the ends of the DNA that are to be ligated.

[13] Due to the aforementioned issues regarding genetic evolution; namely the difficulty in searching a vast sequence space, a preponderance of deleterious mutations in random mutagenesis, and amino acid interdependence, it has been difficult to devise robust methods for searching functional sequence space in the laboratory. Current methods in widespread use for creating mutant proteins in a library format are error-prone polymerase chain reaction [Caldwell & Joyce, (1992); Gram et al., *Proc Natl Acad Sci* 89: 3576-80 (1992)] and cassette mutagenesis [Arkin & Youvan, *Proc Natl Acad Sci* 89: 7811-5 (1992); Hermes et al., *Proc Natl Acad Sci* 87: 696-700 (1990); Oliphant et al., *Gene* 44: 177-83 (1986); Stemmer & Morris, *Biotechniques* 13: 214-20 (1992)], in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide. Alternatively, mutator strains of host cells have been employed to add mutational frequency [Greener et al., *Mol Biotechnol* 7: 189-95 (1997)]. In each case, a `mutant cloud` [Kauffman, (1993)] is generated around certain sites in the original sequence.

[14] Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. Error prone PCR can also be used to mutagenize a mixture of fragments of unknown sequence. Error-prone PCR can randomly mutate genes by altering the concentrations of respective dNTP's in the presence of dITP [Caldwell & Joyce, (1992); Leung & Miyamoto, *Nucleic Acids Res* 17: 1177-95 (1989); Spee et al., *Nucleic Acids Res* 21: 777-8 (1993)].

[15] However, computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the block changes that are required for continued sequence evolution. The published error-prone PCR protocols are generally unsuited for reliable amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical application. Further, repeated cycles of error-prone PCR lead to an accumulation of neutral mutations, which, for example, may make a protein immunogenic.

[16] In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not significantly combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping into families, arbitrarily choosing a single family, and reducing it to a consensus motif, which is

resynthesized and reinserted into a single gene followed by additional selection. This process constitutes a statistical bottleneck, it is labor intensive and not practical for many rounds of mutagenesis.

[17] Methods of saturation mutagenesis utilizing random or partially degenerate primers that incorporate restriction sites have also been described [Hill et al., *Methods Enzymol* 155: 558-68 (1987); Oliphant et al., *Gene* 44: 177-83 (1986); Reidhaar-Olson et al., *Methods Enzymol* 208: 564-86 (1991)].

[18] "Cassette" mutagenesis is another method for creating libraries of mutant proteins [Bock et al., U.S. Patent # 5,830,720 (1995); Christou & McCabe, U.S. Patent # 5,830,728 (1998); Hill et al., *Methods Enzymol* 155: 558-68 (1987); Miller et al., U.S. Patent # 5,830,740 (1998); Shiraishi & Shimura, *Gene* 64: 313-9 (1988); Stemmer & Crameri, U.S. Patent # 5,830,721 (1998)]. Cassette mutagenesis typically replaces a sequence block length of a template with a partially randomized sequence. The maximum information content that can be obtained is thus limited statistically to the number of random sequences in the randomized portion of the cassette.

[19] A protocol has also been developed by which synthesis of an oligonucleotide is "doped" with non-native phosphoramidites, resulting in randomization of the gene section targeted for random mutagenesis [Wang & Hoover, *J Bacteriol* 179: 5812-9 (1997)]. This method allows control of position selection, while retaining a random substitution rate.

[20] Zaccolo and Gherardi (1999) describe a method of random mutagenesis utilizing pyrimidine and purine nucleoside analogs [Zaccolo & Gherardi, *J Mol Biol* 285: 775-83 (1999)]. This method was successful in achieving substitution mutations which rendered -lactamase with an increased catalytic rate against the cephalosporin cefotaxime. Crea describes a "walk through" method, wherein a predetermined amino acid is introduced into a targeted sequence at pre-selected positions [Crea, U.S. Patent # 5,798,208 (1998)].

[21] Methods for mutating a target gene by insertion and/or deletion mutations have also been developed. It has been demonstrated that insertion mutations could be accommodated in the interior of staphylococcal nuclease [Keefe et al., *Protein Sci* 3: 391-401 (1994)]. Examples of deletional mutagenesis methods developed include the utilization of an exonuclease (such as exonuclease III or Bal31) or through oligonucleotide directed deletions incorporating point deletions [Ner et al., *Nucleic Acids Res* 17: 4015-23 (1989)]. Additionally, Lietz describes a method whereby oligonucleotides with random sequences

may be combined with PCR to induce insertions and deletions. Enhancement of function by this technique has not been shown, and the capacity to overmutagenize (i.e. make too many insertions or deletions per polynucleotide) is substantial in this method [Lictz, U.S. Patent # 6,251,604 (2001)].

[22] The technique most often used to evolve proteins *in vitro* is known as "DNA Shuffling". In this method, a library of gene modifications is created by fragmenting homologous sequences of a gene, allowing the fragments to randomly anneal to one another, and filling in the overhangs with polymerase. A full length gene library is then reconstructed with polymerase chain reaction (PCR). The utility of this method occurs at the step of annealing, whereby homologous sequences may anneal to one another, producing sequences with attributes of both starting sequences. In effect, the method affects recombination between two or more genes that are homologous, but that contain significant differences at several positions. It has been shown that creation of the library using several homologous sequences allows a sampling of more sequence space than using a randomly mutated single starting sequence [Crameri et al., *Nature* 391: 288-91 (1998)]. This effect is likely due to the fact that years of evolution have already selected for different advantageous or neutral mutations amongst the homologs of the different species. Starting with homologs, then, appreciably limits the number of deleterious mutations in the creation of the library which is to be screened. Combinatorially rearranging the advantageous positions of the homologs can apparently allow for an optimized secondary protein structure for catalyzing a biochemical reaction. The resulting evolved protein appears to contain positive features contributed from each of the starting sequences, which results in drastically improved function following selection.

[23] Alterations to the DNA shuffling technique have been devised. One process is termed the 'staggered extension' process, or StEP. Instead of reassembling the pool of fragments created by the extended primers, full-length genes are assembled directly in the presence of the template(s). The StEP consists of repeated cycles of denaturation followed by extremely abbreviated annealing/extension steps. In each cycle the extended fragments can anneal to different templates based on complementarity and extend a little further to create "recombinant cassettes." Due to this template switching, most of the polynucleotides contain sequences from different parental genes (i.e. are novel recombinants). This process is repeated until full-length genes form. It can be followed by an optional gene amplification step [Arnold et al., U.S. Patent # 6,177,263 (2001)].

[24] In another technique, fragmentation of the initial DNA can be accomplished by premature termination of the polymerase in an extension reaction by inducing adduct formation in the target gene [Short, U.S. Patent # 5,965,408 (1999)]. In a different technique, a library is created by inducing incremental truncations in each of two homologs to produce a library of fusion genes, each of which contains domains donated from each homolog [Ostermeier et al., *Nat. Biotechnol.* 17: 1205-9 (1999)]. The advantage of this approach is that significant homology amongst the starting sequences is not required since the annealing step of previous methods is omitted. It is unclear, however, whether this modified technique actually will lead to generation of improved gene function after selection techniques are applied to the library.

[25] The previously described methods of gene shuffling using alleles of genes from different organisms allows combinatorial diversity to occur, but is limited by the homology found in the starting sequences. Additionally, these methods do not provide for a mechanism which would generate the junctional diversity formed through V(D)J joining of antibody gene segments. The present invention makes use of mechanisms analogous to junctional diversity by adding and deleting residues from protein or nucleic acid sequences in either a directed or a random fashion. The present invention also provides for "gene swapping" events analogous to the combinatorial diversity generated by combinatorial V(D)J recombination. This will greatly enhance the means by which genes are evolved *in vitro*.

SUMMARY OF THE INVENTION

[26] The present invention involves the directed molecular evolution of nucleic acid sequences by:

(a) adding or deleting nucleotide residues at random in a polynucleotide to produce a library of polynucleotides containing additions or deletions; and

(b) optionally subjecting the pool of polynucleotides in step (a) to a selection procedure capable of identifying polynucleotides encoding for a desired function or feature. Steps (a) and (b) can optionally be repeated. Libraries produced by the methods of the invention are also described and contemplated.

[27] Uniquely, the present invention allows a sampling of sequence space which will include sequences that significantly affect secondary protein structure, thus increasing the probability of identifying altered or improved function in an evolved gene.

Further, the present invention allows a sampling of sequence space which cannot be sampled by other current technologies. Moreover, libraries of polynucleotides created with the present invention cannot be obtained utilizing other current technologies.

[28] Several methods and compositions are described and contemplated below. One method of the invention generates a library of polynucleotide sequences having nucleotide deletions at differing positions in a sequence of a genetic element comprising the steps of:

(a) subjecting multiple copies of circular polynucleotides comprising the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5'end; and

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one of said DNA ends of said polynucleotides producing a library of deletion polynucleotide sequences, said library comprising multiple deletion polynucleotide sequences with deletions at different random positions.

Further, if desired, polynucleotides from step (b) may be subjected to a process that covalently joins the 3' and 5' ends to one another and the library of polynucleotides may be further subjected to a process that selects for a function of interest. The library of deletion polynucleotides may comprise more than two or more, for example, deletions of at least 10, 20 or 30 or more or even 50 to 100 individual polynucleotides each having a random deletion at a different position from the others may be obtained. The number of deletions made will depend upon the starting material and the goal of the technician. In some embodiments, the library of deletion polynucleotides comprises very short deletions of at least 1, 2, 3, 4, or 5 individual nucleotides or more. In different embodiments, the library may comprise larger deletions of 50-100 or more nucleotides. In another emmbodiment, the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element. Further, steps (a) and (b) may optionally be repeated. Another optional method includes a process for inserting nucleotides at the position of deletion in step (b).

[29] Substantially pure compositions comprising a library of multiple (preferably more than two, more preferably more than 5, most preferably more than 10)

linear polynucleotides each having a different 3' and a 5' end, but each linear polynucleotide being identical to the others if circularized are described and contemplated.

[30] Substantially pure compositions comprising a library of at least 2 (preferably more than 5, more preferably more than 10) deletion polynucleotides each differing from the other only by having a different random deletion are also described and contemplated. Optionally such deletion polynucleotides further comprise at least one nucleotide inserted at the position of deletion.

[31] Another method of the invention generates a library of polynucleotide sequences having nucleotide additions at random positions in a genetic element comprising the steps of:

(a) subjecting a composition of multiple copies of circular polynucleotides with the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which adds at least one nucleotide to one of said ends of said polynucleotides producing a library of addition polynucleotide sequences, said library comprising multiple addition sequences with additions at different random positions.

Further, if desired, the addition polynucleotides from step (b) may be subjected to a process that covalently joins said 3' and 5' DNA ends to one another. Optionally, the library of polynucleotides may be subjected to a process that selects for a function of interest.

In any of the methods described here, cleavage preferably occurs with the use of an endonuclease, preferably S1. This method permits the library of addition polynucleotides to comprise any number of different polynucleotides, for example, at least 5, 10, 20 or 30 individual polynucleotides each having a random addition of nucleotides at a different position from the others. In one embodiment of the claimed invention, the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element. Optionally, steps (a) and (b) of the method may be repeated. Another option includes a process for deleting nucleotides at the point of addition in step (b). Any number of nucleotides may be added in step (b) depending upon the starting molecule and the goal of the technician, for example, 1-3, 3-50, or 50-100 or more nucleotides may be added in step (b).

[32] Substantially pure compositions comprising a library of at least 2 (preferably at least 5, most preferably at least 10) addition polynucleotides each differing from the other only by having a different random addition are contemplated.

[33] Further, the present invention surprisingly provides a method to make short deletions at the end of a polynucleotide, producing a population of polynucleotides with short deletions (from 1 to 100), preferably from 1 to 35, most preferably 1 to 10 at the end. A DNA end having such deletions can then be covalently joined with other DNA ends, producing a library of polynucleotides containing deletions at a specific internal position. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product comprises circular polynucleotides. Such methods and compositions are important in the areas of protein engineering and directed evolution.

BRIEF DESCRIPTION OF THE DRAWINGS

[34] FIG. 1 is a diagram of the process of DNA shuffling, an earlier method of choice for molecular evolution. Homologs of a gene of interest are fragmented, subjected to denaturation and reannealing such that the single-strand fragments from the homologs can prime one another in an extension reaction. Amplification of the full length gene then produces a library of hybrid genes. A genetic screen is then applied to select an altered or improved gene.

[35] FIG. 2 is a diagram of the immunoglobulin heavy chain locus illustrating the process of V(D)J recombination which produces combinatorial diversity, and DNA end-joining which produces junctional diversity.

[36] FIG 3. is a diagram illustrating an example of a method which produces nucleotide deletions and insertions at random positions in a polynucleotide. A target gene is cleaved to produce a pool of genes each of which are fragmented at random positions in the gene. Residues can be deleted (left), or inserted (right) at the DNA ends to produce libraries containing deletions, insertions, or deletions and insertions at random positions.

[37] FIG 4 is a diagram illustrating the random cleavage of a polynucleotide. In panel A (Fig. 4A), the DNA plasmid pLacZi (Clontech, Palo Alto, CA) was either uncleaved (lane 2), cleaved with the single cutting restriction enzyme Cla I (lane 3), or increasing concentrations of S1 nuclease (lanes 4-7). Lanes 1 and 8 are lambda/Hind III DNA markers. In panel B (Fig. 4B), the pLacZi plasmid is uncut (lane 2), cleaved with Cla I (lane 3), or S1 nuclease (lane 4). A sample of the S1 cleaved pLacZi was gel purified

and run in lane 5, or further cleaved with Cla I and run in lane 6. Equal amounts of DNA were run in lanes 2-4 (1 μg), and lanes 5-6 (100 ng). The smear in lane 6 illustrates that cleavage by S1 was not site-specific. Lanes 1 and 7 contain lambda/Hind III DNA markers.

[38] FIG. 5 is a diagram illustrating an example of a method which produces short nucleotide deletions at a DNA end. Exonuclease III deletes nucleotides from the ends of a fluorescently labeled 232 bp DNA fragment in a salt dependent reaction. As salt is increased the number of deletions decreases.

[39] FIG 6 is a diagram illustrating the deletion of nucleotides in the LacZ gene. The plasmid pLacZi was cleaved with Cla I, treated with exonuclease III as described in FIG. 5, re-ligated, electroporated into *E. coli*, and plated on plates containing the colorimetric lactose analog X-Gal. Clones with either a blue or white color were picked, grown in LB, and DNA prepared. Plasmid was subjected to PCR with primers flanking the Cla I site, where one primer was fluorescently labeled. The PCR product was run on a 6% denaturing acrylamide gel in an ABI 373 DNA sequencer and analyzed with Genescan software (Perkin Elmer, Foster City, CA). The top panel shows PCR with the wild-type LacZ gene, producing a 312 bp fragment. Clones 1-6 had variable short deletions present. Clones 1 and 6 had blue phenotypes and 2-5 had white phenotypes.

[40] FIG. 7 is a 1.5% agarose gel showing 3 clones containing an insertion in pLacZi. CHO cell cDNA was fragmented with DNase I, ligated into the Cla I site of pLacZi, electroporated into *E. coli*, and plated on X-Gal plates. Three clones were analyzed by PCR of plasmid DNA using primers flanking the Cla I site. Lanes labeled 1-3 are clones containing different sized insertions, and lane 4 is pLacZi. The DNA in the first and last lanes are ΦX174/Hae III DNA markers with their sizes in basepairs indicated at the right.

DETAILED DESCRIPTION OF THE INVENTION

[41] Gene swapping events constitute a major driver in the evolution of macromolecules. Swapping events may include nucleotide insertions, deletions, or replacements. A swapping event may occur by means of homologous recombination, but may also occur by non-homologous means as they do in V(D)J recombination and the DNA-end joining mechanism used by antibody gene segments [Smider & Chu, *Sem. Immun.* 9: 189-97 (1997)]. Current technologies for molecular evolution do not provide a generally applicable non-homologous means for gene swapping.

[42] Applications of the current invention include producing novel genetic elements with improved or altered function. These genetic elements can have significant commercial value. For instance, the genetic element may enhance production of a protein pharmaceutical. The genetic element may encode a protein pharmaceutical such as a monoclonal antibody, or an enzyme used to treat a disease. Further, the genetic element may encode an enzyme important in industrial processes such as chemical manufacturing, or may be used in a product such as laundry detergent (i.e. proteases, lipases, or esterases). Further, the genetic element may have important uses in agriculture, such as to provide a means for pathogen resistance, or to allow production of novel nutrients by a plant species. Additionally, the genetic element may be used in microorganisms to produce novel products for human use, such as novel antibiotics, pigments or other small molecules. As can be seen, the modification of genetic elements in order to improve or alter their function has a myriad of applications in several diverse industries.

[43] For the purposes of describing this invention the following terms will be helpful and will have the following meanings:

Definitions

[44] The term "base" refers to a component of nucleic acid consisting of either adenine, guanine, thymine, cytosine, or uracil. Additionally, "purine" refers to either adenine or guanine, and "pyrimidine" refers to either thymine, cytosine, or uracil.

[45] The term "nucleoside" refers to a molecule comprising the covalent linkage of a pyrimidine or purine to a pentose ring (such as ribose or deoxyribose).

[46] The term "nucleotide" refers to the phosphate ester of a nucleoside.

[47] The term "polynucleotide(s)" refers to a molecule containing at least one 5' hydroxyl of one nucleotide covalently linked to one 3' hydroxyl of at least one other nucleotide through a bond such as a phosphodiester bond. A polynucleotide is necessarily composed of "positions" containing "residues" as defined below.

[48] The term "position" as it relates to a polynucleotide sequence or polypeptide sequence refers to the location of a given residue in the polynucleotide or polypeptide chain. For example, "position" in a polynucleotide sequence is defined as the location of a nucleotide in the polynucleotide chain in reference to at least one other nucleotide. For instance in the simple polynucleotide TG, the T is in position 1 (in reference to itself) and G is in position 2 (in reference to the T in position 1). Often it is convention to label the furthest 5' nucleotide as a reference and label it as position 1. In a double stranded

polynucleotide encoding a gene, such as DNA, often the translation start site of a gene is labeled as position 1. This is often the adenine in the ATG translation start sequence. Positions located 5' from the ATG would be given a negative position (such as -11, -35, etc.) and positions located 3' to the ATG would be given positive positions. Those skilled in the art will recognize the nature of the term "position" as it relates to the numbering scheme in sequences of polynucleotides. A "sequence" refers to the string resulting from the composition of the residues occupying each position. For example the sequence ATG means that the base adenine occupies a position which immediately precedes thymine, and thymine occupies a position which immediately precedes guanine. A "specific position" refers to a position in a polynucleotide between at least two nucleotides whose sequence and composition is known.

[49] The term "residue" as it relates to a polynucleotide or polypeptide refers to either a purine or pyrimidine nucleotide for polynucleotides, or an amino acid for a polypeptide.

[50] A "genetic element" means a sequence of polynucleotide encoding a function. For example, a "genetic element" may encode a polypeptide sequence, may encode a promoter function, an enhancer function, a transcription start or stop site, or RNA splice sites and the like. Genetic elements may be operatively linked to other genetic elements, for example a promoter may be operatively linked to a genetic element encoding a protein to allow expression of a protein in a given cell type. The term "gene" and "gene of interest" refer to a polynucleotide capable of encoding a polypeptide.

[51] The term "swap" or "gene swapping" in reference to a polynucleotide means either: 1) the occurrence of a deletion of at least two residues occupying consecutive positions in a polynucleotide, or 2) the occurrence of an addition of at least two residues occupying consecutive positions into a polynucleotide, or 3) the replacement of at least two residues occupying consecutive positions in a polynucleotide with other residues.

[52] The term "nucleotide deletions" as applied to a polynucleotide means that a polynucleotide has had one or more specific residues removed from one or more positions in the polynucleotide chain when the resulting polynucleotide is compared to the parental, wild-type, or other reference sequence.

[53] The term "nucleotide insertions" or "nucleotide additions" means that a polynucleotide has had specific residues added to the polynucleotide chain, such that at least one of the original residues now occupies a new position in the polynucleotide when compared to the parental, wild-type, or other reference sequence.

[54] The term "library of polynucleotide sequences" refers to a mixture of polynucleotides, wherein at least one of the sequences differs from at least one other sequence in the mixture by sequence composition or length, for example, where at least one position is occupied by a different nucleotide when the two sequences are compared or at least one nucleotide position is absent in one sequence when compared with the other sequence.

[55] The term "DNA" refers to deoxyribonucleic acid. It will be understood by those of skill in the art that where manipulations are described herein that relate to DNA they will also apply to RNA.

[56] The term "DNA ends" or ends refers to the position in a DNA strand wherein a phosphodiester bond is broken. In a single-stranded DNA end a nucleotide is only covalently linked with one other nucleotide. A "double-stranded DNA or RNA end" refers to the position in a double-stranded DNA or RNA molecule wherein the molecule is no longer double-stranded. Generally DNA ends are recognizable to those skilled in the art. Double-stranded DNA ends are characterized as blunt, having a 5' overhang, a 3' overhang, or a hairpin structure. A DNA end may or may not contain a 5' phosphate group.

[57] The term "cleavage" as used herein refers to the breakage of a bond between two nucleotides, such as a phosphodiester bond.

[58] The term "circular polynucleotide" refers to a polynucleotide wherein no double-stranded DNA ends are present. A circular polynucleotide may be single-stranded or double-stranded. A circular polynucleotide may, however, contain single-stranded DNA ends. A circular polynucleotide will be present if single-stranded DNA ends exist but hydrogen bonding keeps the two strands of the double-stranded molecule hybridized to one another such that a double-stranded DNA end is not created by the presence of two single-stranded ends in proximity to one another. Such a circular double-stranded polynucleotide is often referred to as "nicked".

[59] The term "linear polynucleotide" is a polynucleotide which contains at least one, but most often two DNA ends. A linear polynucleotide may be either single-stranded or double-stranded.

[60] The term "random" or "random position" as applied to a polynucleotide refers to a process by which any of the specific residue positions may be selected. Random as used here does not mean that all points or point of cleavage or nucleotides or positions are selected or chosen with equal frequency. Rather random focuses on the unpredictable nature of the process, i.e. the worker cannot predict *a priori* where an

event will occur or what position any base will have. Finally, not all positions need be available for cleavage for the process to be random as to the available positions or bases. For example, a polynucleotide with a length of N may have any or all of its positions (i.e. 1, 2,...N) affected by a manipulation. In the addition (insertion) or deletion of residues, a polynucleotide necessarily must have covalent bonds (such as phosphodiester bonds) cleaved, thereafter which residues are deleted or added (i.e. the total number of positions is decreased or increased, respectively). In describing "deletions at random positions" in a polynucleotide of length N, it is meant that any or all of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between nucleotides (i.e. phosphodiester bonds) are broken, and at least one nucleotide at the end is removed prior to re-ligation. Thus, in a process causing "deletions at random positions" the final length of the polynucleotide (N, or the number of positions) necessarily decreases. Similarly, In describing "insertions at random positions" in a polynucleotide of length N, it is meant that any or all of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between nucleotides (i.e. phosphodiester bonds) are broken, and at least one new nucleotide (i.e. a new position) is added at the end prior to re-ligation. Thus, in a process causing "insertions at random positions" the final length of the polynucleotide (N, or the number of positions) necessarily increases. It is recognized that a combination of processes involving "deletions at random positions" and "insertions at random positions" may allow the final length of the polynucleotide to remain unchanged (i.e. the additions cancel out the deletions and the final number of positions remains the same, however the nucleotides occupying the positions may be different). In describing "random cleavage" or a "single random break" in a polynucleotide of length N, it is meant that any one of the N (in a circular polynucleotide) or N-1 (in a linear polynucleotide) covalent linkages between residue positions in a single polynucleotide molecule are cleaved. Accordingly, in one vessel containing many copies of a polynucleotide, a single random break can occur at different positions in different molecules.

[61] As used herein, "substantially pure" means an object species is the predominant species present (i.e., on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably a substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional

detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

[62] The term "homologous" or "homeologous" means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentration as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp. Thus, "homologs" are nucleic acid molecules that are not identical but are capable of hybridizing to one another under physiological conditions. Double-stranded homologs are capable of hybridizing to one another following denaturation.

[63] The term "heterologous" means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that nucleic acid fragments or polynucleotides have areas or regions in the sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are, for example, areas of mutations.

[64] The term "identical" or "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a nucleic acid fragment or polynucleotide are identical or complementary to another polynucleotide or nucleic acid fragment.

[65] The term "amplification" means that the number of copies of a nucleic acid fragment is increased.

[66] The term "wild-type" means that the nucleic acid fragment does not comprise any mutations. A "wild-type" protein means that the protein will be active at a comparable level of activity found in nature and typically will comprise the amino acid sequence found in nature. In an aspect of the invention, the term "wild type" or "parental sequence" can indicate a starting or reference sequence prior to a manipulation of the sequence.

[67] The term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

[68] The term "chimeric polynucleotide" means that the polynucleotide comprises nucleotide regions which are wild-type and regions which are mutated. It may also

mean that the polynucleotide comprises wild-type regions from one polynucleotide and wild-type regions from another related polynucleotide.

[69] The term "population" as used herein means a collection of components such as polynucleotides, nucleic acid fragments or proteins. A "mixed population" means a collection of components which belong to the same family of nucleic acids or proteins (i.e. are related) but which differ in their sequence (i.e. are not identical) and hence in their biological activity. A "library" necessarily implies a population wherein at least two of the components is different in some aspect (chemical composition, length, etc.)

[70] The term "specific nucleic acid fragment" means a nucleic acid fragment having certain end points and having a certain nucleic acid sequence. Two nucleic acid fragments wherein one nucleic acid fragment has the identical sequence as a portion of the second nucleic acid fragment but different ends comprise two different specific nucleic acid fragments. Two nucleic acid fragments with identical sequences but different 5' or 3' ends comprise two different specific nucleic acid fragments.

[71] The term "mutations" means changes in the sequence of a wild-type nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications.

[72] In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

[73] The term "naturally-occurring" as used herein as applied to an object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring. Generally, the term naturally-occurring refers to an object

as present in a non-pathological (undiseased) individual, such as would be typical for the species.

[74] As used herein the term "physiological conditions" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45°C. and 0.001-10 mM divalent cation (e.g., Mg^{++}, Ca^{++}); preferably about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v). Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or nonionic detergents and/or membrane fractions and/or antifoam agents and/or scintillants.

[75] As used herein, "linker" or "spacer" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide, and serves to place the two molecules in a preferred configuration, e.g., so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

[76] As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

Producing Libraries of Evolving Random Molecules

[77] The present invention provides a method to create libraries of polynucleotides containing either nucleotide deletions, insertions or combinations of

deletions and insertions at random positions. In effect this invention provides a means to "swap" genetic elements without the need for homology or amplification techniques. The swapping of genetic elements is known to be a driving force in evolution of macromolecules, cells, and organisms [Ostermeier & Benkovic, *Adv Protein Chem* 55: 29-77 (2000)]. Current techniques, such as PCR based gene shuffling, do not allow significant swapping of genetic elements independent of homology.

Deletions

[78] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. One method of the invention, for example, comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of the ends of said polynucleotides; and

(c) optionally subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion at one position.

[79] Further, the invention provides a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. It is contemplated that deletions will allow removal of detrimental or unwanted functions of a genetic element. These functions might include protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

[80] In a further embodiment, the invention provides a method, for example, to generate polynucleotides wherein the polynucleotides contain deletions at more than one position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides; and

(c) optionally, subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion at one position.

A function of interest may then be selected for if desired (step(d)). Further, if desired, steps (a) to (c) or steps (a) to (d) may be repeated from 1 to 50 times or more.

[81] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain deletions at more than one position. It is contemplated that deletions at multiple positions will allow removal of multiple detrimental or unwanted functions of a genetic element. These functions might include any combination of protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins or other functions of interest as will be well appreciated by those of skill in the art.

Insertions

[82] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which inserts at least one nucleotide to at least one end of said polynucleotides;

(c) optionally subjecting said polynucleotides from step (b) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by an insertion at one position.

[83] Further, the invention provides a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. This embodiment of the invention will allow novel fusion of genetic elements to occur. For example, a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases)

could be fused in new ways, or new functions like binding domains (i.e. nucleic acid binding domains, small molecule or ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

[84] Likewise in another embodiment, the invention provides a method to generate polynucleotides wherein the polynucleotides contain insertions at more than one position. One method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions;

(b) subjecting said polynucleotides from step (a) to a process which inserts at least one nucleotide to at least one end of said DNA ends of said polynucleotides; and

(c) optionally, subjecting said polynucleotides from step (b) to a process which covalently joins said DNA ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by an insertion at one position; and

(d) optionally selecting for a function of interest. Steps (a)-(b), (a)-(c) or (a)-(d) may be repeated from 1 to 50 times or more.

[85] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions at more than one position. It is contemplated that this embodiment of the invention will allow multiple novel fusions of genetic elements to occur. For example, the following could be fused to a gene of interest in a combinatorial fashion: a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or new functions like multiple binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

Combinations of insertions and deletions

[86] In one embodiment, the invention provides a method to create a population of polynucleotides, with members of the population differing from one another by the presence of deletions and insertions at a single random position. This method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides;

(c) subjecting said polynucleotides from step (b) to a process which inserts at least one nucleotide to at least one end of said DNA ends of said polynucleotides from step (b);

(d) optionally subjecting said polynucleotides from step (c) to a process which covalently joins said DNA ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion and insertion at one position.

[87] Further, the invention provides a population of polynucleotides, with members differing from one another by a combination of deletions and insertions at a single random position. It is contemplated that this embodiment will allow for new heterologous domains to replace domains in the gene of interest. In this regard, new functions, such as ligand binding or enzymatic catalysis could be conferred upon a genetic element. Also, native function could be enhanced utilizing this embodiment.

[88] In another embodiment, the invention provides a method to generate polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. In this regard deletions may occur at different positions than insertions, or deletions and insertions can occur at the same position. Further, deletions and/or insertions can occur at multiple positions. This method comprises the steps of:

(a) cleavage of a composition of multiple copies of polynucleotides at random positions to create two ends;

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one end of said ends of said polynucleotides;

(c) optionally subjecting said polynucleotides from step (b) to a process which inserts at least one nucleotide to at least one end of said ends of said polynucleotides;

(d) optionally subjecting said polynucleotides from step (c) to a process which covalently joins said ends to one another, producing a library of polynucleotides which contains at least one polynucleotide that differs from the others by a deletion and insertion at one position;

(c) optionally selecting for a function of interest; and optionally repeating any of steps (a) to (d) from 1 to 50 times or more.

[89] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. It is contemplated that this embodiment of the invention will allow for classical directed evolution, wherein multiple rounds of insertions at random positions, deletions at random positions, and combinations of insertions and deletions, are produced with the genetic element being optionally subjected to selection between each round. This embodiment allows for the improvement or alteration of the function of a genetic element.

Starting material

[90] The present invention can be applied to any polynucleotide of interest to the researcher. The polynucleotide can be nucleic acid, i.e. RNA or DNA. Often the polynucleotide will be DNA consisting of genetic elements or one or more genes of interest. The starting material may be obtained through natural sources, or may be polynucleotides which have been synthesized in a laboratory (e.g. gene synthesis), or may be polynucleotides derived from natural sources which have been manipulated in a laboratory. Several sources of polynucleotides are available through publicly held databanks such as Genbank (http://www.ncbi.nlm.nih.gov:80/Genbank/index.html) or available commercially (Celera, Rockville, MD; Incyte, Palo Alto, CA; Clontech, Palo Alto, CA; Invitrogen, Carlsbad, CA).

[91] The nucleic acid may be obtained from any source, for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) [Mullis, U.S. Patent # 4,683,202 (1987); Mullis et al., U.S. Patent # 4,683,195 (1987)]. Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

[92] The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids, phagemids, YACs,

and BACs are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large nucleic acid fragments.

[93] If the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of mutants does not increase.

[94] Starting material should be in substantially pure form. The polynucleotide may be double-stranded or single-stranded, but more preferably is double-stranded. Further, the polynucleotide may be linear or circular, but in a preferred embodiment the polynucleotide is circular. Polynucleotides in circular form may be prepared by preparation of plasmid DNA from organisms such as bacteria, yeast, plants, or mammalian cells by techniques well known to those skilled in the art [Maniatis et al., (1989)]. The number of different specific nucleic acid fragments in the reaction vessel will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

[95] The starting material (i.e. the polynucleotide), while in substantially pure form, can also be present without homologs or related sequences. In other words, the polynucleotides in the initial vessel may all be identical, although they may also be related, unrelated or heterologous. In fact, performance of the present invention will be unaffected by the sequence of the starting material. Furthermore, the sequence of the starting material may be known or unknown. For directed evolution purposes, all that is required is a method to detect the function of the polynucleotide (such as a screening assay).

Cleaving the polynucleotide at a random position

[96] In general, a nucleic acid fragment may be cleaved by a number of different methods. The nucleic acid fragment may be digested with a nuclease such as DNAse I, S1 nuclease, P1 or mung bean nuclease, or RNAse, which are readily available. Other enzymes, such as RAG1 and RAG2, topoisomerases, and integrases are capable of cleaving polynucleotides. The nucleic acid may be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides. Chemical agents, such as bleomycin or methyl methanesulfonate (MMS) can also cleave polynucleotides.

[97] Of substantial importance to the generation of functionally mutated genes containing insertions or deletions is to cleave the polynucleotide a small number of

times, usually between 1 and 10, preferably between 1 and 5, and most preferably once. The present invention provides a means to cleave a polynucleotide such that cleavage occurs only at one position per polynucleotide in the reaction vessel. Of importance is that the present invention provides a means for a near random cleavage of a polynucleotide (i.e. cleavage at several different positions in different molecules). Cleavage can be double-stranded or single-stranded (i.e. produce single-stranded ends or double-stranded ends). Examples of enzymes which can cleave polynucleotides include DNase I, S1 nuclease, P1 nuclease, as well as topoisomerases, transposons, and integrases. Cleavage can occur transiently with enzymes such as topoisomerases, transposons, and integrases. These enzymes may cleave the polynucleotide once, or more than once. S1 nuclease can be used to cleave double or single-stranded polynucleotides in a generally random fashion. In a preferred embodiment, with circular double-stranded DNA, S1 nuclease will cleave the polynucleotide only once, producing two DNA ends (FIG 4).

[98] It is also contemplated that the nucleic acid may also be partially digested with one or more restriction enzymes which cleave DNA at a high frequency (i.e. at several positions within a polynucleotide), such that certain polynucleotides are cleaved only once, and that the resulting population contains polynucleotides cleaved one time, but with different polynucleotides cleaved at different positions. The cleavage with a restriction enzyme may not be entirely random, but if the genetic element of interest has enough specific restriction sites at different positions, the cleavage pattern may be useful enough to generate substantial diversity.

[99] It is contemplated that single cleavage of a polynucleotide can be accomplished through other alternative mechanisms which normally cleave polynucleotides several times. A polynucleotide can be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides. If any of these modalities is carefully titrated and a means of purification is utilized, the singly cleaved molecules can be obtained in substantially pure form (i.e. singly cleaved molecules can be purified away from uncleaved or multiply cleaved molecules).

[100] Furthermore, enzymes which act to cleave and rejoin DNA, such as topoisomerases, transposons, and integrases can be utilized to effectively cleave a polynucleotide [Singh et al., *Proc Natl Acad Sci* 94: 1304-9 (1997)]. In these cases the cleavage and rejoining steps may be coupled. Preferably the DNA ends are linked, or are in physical proximity to one another, following cleavage. This is in order to prevent the re-

ligation of the wrong ends to one another following deletional or insertional events. One mechanism to keep the ends linked is through the use of a circular polynucleotide as a starting material. In this case, the ends are linked by the intervening polynucleotide chain. Thus, the re-ligation will be an intramolecular event as opposed to intermolecular, and will proceed with greater efficiency. Other mechanisms to keep the ends in proximity is through a protein bridge, such as through chromatin (i.e. histones, or other DNA binding proteins), or through enzymes which couple cleavage with rejoining, such as transposons, integrases, or topoisomerases. Alternatively, ends could conceivably be left in proximity to one another through the linkage of opposite ends (the non-cleaved ends) to solid supports.

[101] Cleavage of a circular polynucleotide consisting of supercoiled plasmid DNA can be accomplished by incubating from 0.1 to 100 µg, preferably from 1 to 10 µg with a nuclease such as S1 nuclease. The nuclease can be present in amounts from 0.1 to 1000 units, but preferably from 1 to 100 units in a reaction of 10 µl. The temperature of the reaction can occur from between 0 and 100°C, but preferably between 4 and 50 °C. The reaction time can vary from 30 seconds to 1 hour, but preferably is between about 1 and 30 minutes. The degree of linearization can be measured by analyzing the plasmid DNA on an agarose gel as in FIG. 4. The linear DNA should preferably be purified from the uncut DNA by any of a number of methods well known to those skilled in the art. Such methods include utilization of agarose gel purification kits (Qiagen, Valencia, CA). HPLC, column chromatography and the like.

Deletion of nucleotides

[102] Nucleotide deletions can be generated at a DNA end by a variety of means. For instance, an exonuclease, such as exonuclease III, can be used to remove nucleotides in a 3' to 5' direction from a DNA end. The resulting DNA end then contains a 5' overhang which can be removed by digestion of the DNA with a single-stranded endonuclease such as P1 nuclease, S1 nuclease, or mung bean nuclease. Bal 31 nuclease is an enzyme which possesses 5' to 3' as well as 3' to 5' nucleolytic activity and can be used to delete nucleotides from a DNA end. Furthermore, several polymerases, like DNA polymerase I from *E. coli*, Klenow fragment, and Taq polymerase contain exonuclease activity and could conceivably be used to make deletions from a DNA end. Cell extracts from all organisms contain DNA repair enzymes which can act to delete nucleotides, thus unpure cell extract could conceivably be used as a source for exonuclease activity. Other nucleases, which may not have exonuclease activity under certain conditions may be capable

of producing deletions at a DNA end under other conditions. For example, S1 nuclease can produce short deletions when used at high enzyme concentrations. Furthermore, it is contemplated that mild denaturation of a DNA molecule, such that the DNA ends become "frayed", will allow deletions to occur upon application of a single-stranded endonuclease, such as S1, P1, or mung-bean nuclease.

[103] In a preferred embodiment the conditions of the deletion reaction are set such that the number of individual deletions occurring at each DNA end may be well controlled. For example, altering the salt concentration, altering the pH, altering the temperature, or altering any of the other biochemical parameters of the reaction can change the activity of the nuclease enzyme such that more or less deletions will occur depending on the intent of the investigator (for instance decreasing temperature or increasing salt may lower the processivity of the exonuclease and cause fewer deletions). Figure 5 shows altering conditions allowing differing numbers of deletions to occur on a DNA end. In some cases large deletions might be warranted (i.e. to completely remove a large domain in a genetic element), in other cases small deletions might be preferable (i.e. to remove a single amino acid, or a few amino acids such as those that comprise a protease site). Generally deletions could be obtained numbering from 1 to 1000, more preferably they would be from 1 to 100. In certain instances, as described, the deletions may number from 1 to 10.

[104] Due to cleavage at a random position in the polynucleotide, the location of the deletions in the resulting polynucleotide will also be located at a random position. Also, since residues are deleted from either end of the molecule, the total number of deletions will equal the sum of the deletions occurring on the 5' end and the 3' end.

Adding nucleotides

[105] In order to make additions to a polynucleotide in random positions, the polynucleotide is necessarily cleaved at a random position, as described above. Prior to insertion, nucleotides may be deleted from the DNA ends produced during the cleavage event. Alternatively, the DNA ends formed by the cleavage reaction can be used as substrates to which new nucleotides or polynucleotides are added.

[106] Several different mechanisms exist to add nucleotides to the ends of a polynucleotide. For example, nucleotides can be added by chemical coupling. A polymerase, such as terminal deoxynucleotidyl transferase can be utilized to add nucleotides in a semirandom fashion to a DNA end [Gauss & Lieber, *Mol Cell Biol 1996* 16: 258-69

(1996)]. Alternatively, the cleavage step may be coupled to the insertion event, as can be the case when employing transposons or integrases to the insertion event.

[107] A ligase such as *E. coli* ligase or phage T4 ligase can be utilized to covalently couple a new polynucleotide to the parent polynucleotide. In a preferred embodiment the polynucleotide is a genetic element or a fragment of a genetic element. A genetic element predisposes the resulting polynucleotide to have function since genetic elements are functional in some way by definition. The genetic element may be a gene, the regulatory element of a gene, or a genetic element encoding a useful domain. The genetic element may be a library of genetic elements such as a cDNA library or genomic DNA library. Fragments of a genetic element can be produced by digesting the polynucleotide with a nuclease, such as DNAse I, S1 nuclease, P1 or mung bean nuclease, or RNAse. Other enzymes, such as restriction enzymes and topoisomerases, can also cleave polynucleotides into fragments. The polynucleotides may be randomly sheared by the method of sonication or by passage through a tube having a small orifice. The use of radiation, such as gamma radiation or ultraviolet radiation is also capable of cleaving polynucleotides into fragments. Chemical agents, such as bleomycin or MMS can also cleave polynucleotides into fragments.

[108] It is contemplated that the mixture of a parent polynucleotide cleaved at a random position, with a population of genetic elements or fragments of genetic elements, and a ligase such as T4 DNA ligase, under the appropriate salt, buffer, and temperature conditions, will allow covalent coupling of the genetic elements with the parent polynucleotide at the position of the original cleavage event. Thus, a mixture of polynucleotides is produced comprising an insertion at a random position within the parent polynucleotide. The content (i.e. the sequence) of each insertion may be identical if the genetic elements or fragments of genetic elements are identical, or different if the fragments of genetic elements were non- identical.

Rejoining the DNA ends

[109] DNA ends may be rejoined covalently by incubating the DNA ends with an enzyme like a DNA ligase which will form phosphodiester bonds between nucleotides at the DNA end. Examples of ligases include *E. coli* DNA ligase, phage T4 DNA ligase, or human DNA ligases. These enzymes can be used under conditions well known to those skilled in the art to ligate DNA. Other enzymes are also capable of creating covalent linkages (like phosphodiester bonds) between nucleotides at DNA ends. Such enzymes are topoisomerases, transposons, integrases, and other recombination enzymes. Other

mechanisms can be used to join DNA ends such as the utilization of an oligonucleotide whose sequence can hybridize to sequences on either end (i.e. both the 5' and 3' ends) to "bridge" the ends with hydrogen bonds. The intervening sequence on the opposite strand could be filled in with a polymerase, such as *E. coli* polymerase, Klenow fragment, phage T4 polymerase, or Taq polymerase. Nicks could then be repaired by a DNA ligase as described above. Cellular extracts also contain ligase activities and cell or nuclear extracts could be used to rejoin DNA ends. Alternatively, DNA molecules could be introduced into intact cells and the cell's machinery could rejoin DNA ends by homologous or non-homologous means.

Library compositions

[110] The present invention provides for novel libraries of which the following compositions are examples:

Deletions

[111] The invention provides a population of polynucleotides, with members of the population differing from one another by the presence of deletions at a single random position. Such single deletion libraries can contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. Deletion libraries should contain at least one molecule that differs from at least one other molecule by the deletion of at least one nucleotide at one random position. The number of deletions at each position could be from 1 to 1000, but should be at least one. It is contemplated that deletions will allow removal of detrimental or unwanted functions of a genetic element. These functions might include protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

[112] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain deletions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These multiple deletion libraries should contain at least one molecule that differs from at least one other molecule by the deletion of at least one nucleotide at more than one random position. It is contemplated that deletions at multiple positions will allow removal of multiple detrimental or unwanted functions of a genetic element. These functions might include any combination of multiple protease sites, ion binding domains, DNA binding sequences for inhibitory transcription factors, immunogenic domains of proteins and the like.

Insertions

[113] The invention provides a population of polynucleotides, with members of the population differing from one another by the presence of insertions at a single random position. Insertion libraries can contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. Insertion libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at one random position. The number of insertions at each position could be from 1 to 10,000, but preferably will be at least one. For example, a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or a new function like binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

[114] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These multiple insertion libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at more than one random position. It is contemplated that this embodiment of the invention will allow novel fusion of genetic elements to occur. It is contemplated that this embodiment of the invention will allow multiple novel fusions of genetic elements to occur. For example the following could be fused to a gene of interest in a combinatorial fashion: a toxin could be fused to a targeting molecule (like an antibody), enzyme modules in important metabolic pathways (such as polyketide synthetases) could be fused in new ways, or new function like binding domains (i.e. nucleic acid binding domains, ion binding domains, protease sites, or other post-translational modification modules) could be incorporated into existing genetic elements.

Combinations of insertions and deletions

[115] The invention provides a population of polynucleotides, with members differing from one another by a combination of deletions and insertions at a single random position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These combination libraries should contain at least one molecule that differs from at least one other molecule by the insertion of

one nucleotide and the deletion of at least one nucleotide at one random position. It is contemplated that this embodiment will allow for heterologous domains to replace domains in the gene of interest. In this regard, new functions, such as ligand binding or enzymatic catalysis could be conferred upon a genetic element. Also, native function could be enhanced utilizing this embodiment.

[116] Further, the invention provides a population of polynucleotides wherein the polynucleotides contain insertions and deletions at more than one position. Such a library should contain at least 2 molecules, but preferably 100 molecules, and most preferably at least about 1000 molecules. These combination libraries should contain at least one molecule that differs from at least one other molecule by the insertion of at least one nucleotide at one random position and the deletion of at least one nucleotide at one random position. This embodiment of the invention will allow for classical directed evolution, wherein multiple rounds of insertions at random positions, deletions at random positions, and combinations of insertions and deletions, are produced with the gene of interest being optionally subjected to selection between each round. This embodiment allows for the improvement or alteration of function of a genetic element.

Analyzing the composition

[117] The composition of such libraries can be determined by mechanisms well known to those in the art. In order to determine whether a library contains insertions or deletions, the library can be analyzed by agarose or acrylamide gel electrophoresis and size can be compared to the parental sequence. Other methods, like HPLC, mass spectrometry, column chromatography can be used to identify size differences between polynucleotides. Because the present invention relates to random positions of insertions and deletions, the most definitive method to determining the composition of a library is to subject representative polynucleotides within the composition to sequencing, a method well known to those skilled in the art. Comparison of sequences of representative clones would allow one to determine if deletions or insertions occurred at random positions in different molecules in the library.

[118] The resulting library could be ligated into an expression vector for use as a vehicle to express the resulting variants contained within the library. The nature of the expression vector is described below in the "screening" section.

Screening for a function of interest

[119] In testing a library of polynucleotides for a function of interest, the library should be inserted in an appropriate expression vector. Alternatively, the library can be constructed in an expression vector (i.e. the library comprises an expression vector). The vector used for cloning is not critical provided that it will accept a DNA fragment of the desired size. If expression of the DNA fragment is desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the DNA fragment to allow expression of the DNA fragment in the host cell. For screening in bacterial cells, preferred vectors include the pUC series and the pBR series of plasmids.

[120] The resulting bacterial population will include a number of recombinant DNA fragments having random mutations. This mixed population may be tested to identify the desired recombinant nucleic acid fragment. The method of selection will depend on the DNA fragment desired.

[121] The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids, phagemids, YACs, and BACs are preferred where the specific nucleic acid sequence is larger because these vectors are able to stably propagate large nucleic acid fragments.

[122] If a DNA fragment which encodes for a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the DNA fragments in the population or library may be tested for their ability to bind to the ligand by methods known in the art (i.e. panning, affinity chromatography). If a DNA fragment which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the DNA fragments in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify DNA fragments which confer the desired properties onto the protein.

[123] In the context of the present invention the term "positive polypeptide variants" means resulting polypeptide variants possessing functional properties which has been improved in comparison to the polypeptides producible from the corresponding input DNA sequences. Examples, of such improved properties can be as different as, for example, enhanced or lowered biological activity, increased wash performance, thermostability, oxidation stability, substrate specificity, antibiotic resistance or others that may be of interest.

[124] Consequently, the screening method to be used for identifying positive variants depend on which property of the polypeptide in question it is desired to change, and in what direction the change is desired.

[125] A number of suitable screening or selection systems to screen or select for a desired biological activity are described in the art. For example, Strausberg et al. [Strausberg et al., *Biotechnology (N Y)* 13: 669-73 (1995)] describes a screening system for subtilisin variants having calcium-independent stability. Bryan et al. [Bryan et al., *Proteins* 1: 326-34 (1986)] describes a screening assay for proteases having an enhanced thermal stability.

[126] It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee Wis.). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein, a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

Methods of Effecting Targeted Short Deletions in Nucleic Acids

[127] The ability to make short deletions in a polynucleotide is generally hampered by the high activity and processivity of exonucleases that act at a DNA end. Several methods exist to make large (i.e. more than 100 base) deletions at DNA ends [Sambrook et al., (1989)]. However, methods to create short deletions, such as from 1 to 100 bases or very short deletions like from 1 to 10 bases in a controlled fashion have not been possible. The ability to make such deletions at specific sites is important in the field of protein engineering [Altamirano et al., *Nature* 403: 617-22 (2000)] and is highlighted in the end-joining mechanism of V(D)J recombination, the method which produces the substantial diversity in antibody genes [Smider & Chu, *Sem. Immun.* 9: 189-97 (1997)].

Starting material

[128] The deletion generating mechanism can be applied to any polynucleotide of interest to the researcher. The polynucleotide can be nucleic acid, i.e. RNA or DNA. Often the polynucleotide will be DNA consisting of genetic elements or one or

more genes of interest. The starting material may be obtained through natural sources, or may be polynucleotides which have been synthesized in a laboratory (e.g. gene synthesis), or may be polynucleotides derived from natural sources which have been manipulated in a laboratory. Several sources of polynucleotides are available through publicly held databanks such as Genbank (http://www.ncbi.nlm.nih.gov:80/Genbank/index.html) or available commercially (Celera, Rockville, MD; Incyte, Palo Alto, CA; Clontech, Palo Alto, CA; Invitrogen, Carlsbad, CA).

[129] The nucleic acid may be obtained from any source, for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) [Mullis, U.S. Patent # 4,683,202 (1987); Mullis et al., U.S. Patent # 4,683,195 (1987)]. Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Deletion of nucleotides

[130] Nucleotide deletions can be generated at a DNA end by a variety of means. For instance, an exonuclease, such as exonuclease III, can be used to remove nucleotides in a 3' to 5' direction from a DNA end. Often the resulting DNA end contains a 5' overhang which can be removed by digestion of the DNA with a single-stranded endonuclease such as P1 nuclease, S1 nuclease, or mung bean nuclease. Other exonucleases could also be used in the present invention. Bal 31 nuclease is an enzyme which possesses 5' to 3' as well as 3' to 5' nucleolytic activity and can be used to delete nucleotides from a DNA end. Exonuclease T can remove nucleotides in a 3' to 5' direction. Exonuclease 7 can remove nucleotides in a 5' to 3' direction, and can act at single-stranded ends such as nicks or gaps. Exonuclease I catalyzes the removal of nucleotides from single-stranded DNA in the 3′ to 5′ direction. Lambda exonuclease is a highly processive enzyme that acts in the 5′ to 3′ direction, catalyzing the removal of 5′ mononucleotides from duplex DNA. RecJ is a single-stranded DNA specific exonuclease that catalyzes the removal of deoxynucleotide monophosphates from DNA in the 5′ to 3′ direction. Furthermore, several polymerases, like DNA polymerase I from *e. coli*, Klenow fragment, and Taq polymerase contain exonuclease activity and could conceivably be used to make deletions from a DNA end. Cell extracts

from all organisms contain DNA repair enzymes which can act to delete nucleotides, thus unpure cell extract could conceivably be used as a source for exonuclease activity. Other nucleases, which may not have exonuclease activity under certain conditions may be capable of producing deletions at a DNA end under other conditions. For example, S1 nuclease can produce short deletions when used at high enzyme concentrations. Furthermore, it is contemplated that mild denaturation of a DNA molecule, such that the DNA ends become "frayed", will allow deletions to occur upon application of a single-stranded endonuclease, such as S1, P1, or mung-bean nuclease.

[131] In a preferred embodiment, the conditions of the deletion reaction are set such that the number of individual deletions occurring at each DNA end may be well controlled. For example, altering the salt concentration and the temperature, altering the pH, or altering any of the other biochemical parameters of the reaction can change the activity of the nuclease enzyme such that more or less deletions will occur depending on the intent of the investigator. Most particularly and surprisingly we have found that decreasing temperature and/or increasing salt lowers the processivity of the exonuclease and results in more controlled small deletions. Salts used in the reaction may be any salt. Examples of salts include sodium chloride, sodium acetate, potassium chloride, or potassium acetate. Preferably the salt is either sodium chloride or potassium chloride. Salt concentrations can range from 10 mM to 1.0 M, but preferably is between 50 mM and 500 mM. Temperature of the reaction can also vary in the present invention. The temperature can range from 0°C to 30°C, but preferably is between 0°C and 24°C. Figure 5 shows altering conditions allowing differing numbers of deletions to occur on a DNA end. In some cases large deletions might be warranted (i.e. to completely remove a large domain in a genetic element), in other cases small deletions might be preferable (i.e. to remove a single amino acid, or a few amino acids such as those that comprise a protease site). The resulting population of polynucleotides contain variable amounts of deletions at the ends of the starting sequence. Generally deletions could be obtained numbering from 1 to 1000, more preferably they would be from 1 to 100. In a preferred embodiment, the deletions may number from 1 to 30 or even 1 to 10.

Rejoining the DNA ends

[132] In some cases it might be useful to join the DNA ends of a molecule containing a deletion with a second DNA end, such that the deletion now occurs at an internal position. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product is a circular polynucleotide. DNA ends may be rejoined by

incubating the DNA ends with an enzyme like a DNA ligase which will form phosphodiester bonds between nucleotides at the DNA end. Examples of ligases include *E. coli* DNA ligase, phage T4 DNA ligase, or human DNA ligases. These enzymes can be used under conditions well known to those skilled in the art to ligate DNA. Other enzymes are also capable of creating covalent linkages (like phosphodiester bonds) between nucleotides at DNA ends. Such enzymes are topoisomerases, transposons, integrases, and other recombination enzymes. Other mechanisms can be used to join DNA ends such as the utilization of an oligonucleotide whose sequence can hybridize to sequences on either end (i.e. both the 5' and 3' ends) to "bridge" the ends with hydrogen bonds. The intervening sequence on the opposite strand could be filled in with a polymerase, such as *e.coli* polymerase, Klenow fragment, phage T4 polymerase, or Taq polymerase. Nicks could then be repaired by a DNA ligase as described above. Cellular extracts also contain ligase activities and cell or nuclear extracts could be used to rejoin DNA ends. Alternatively, DNA molecules could be introduced into intact cells and the cell's machinery could rejoin DNA ends by homologous or non-homologous means.

Deletion compositions

[133] In one embodiment the current invention provides for a composition of polynucleotides, wherein members of the population differ from one another by the presence of deletions at one or both ends of the polynucleotide. The number of deletions may range from 1 to 100 at each end, but more preferable is from 1 to 30.

[134] Additionally, the current invention provides for a composition of polynucleotides differing from one another by short deletions at a specific internal position (i.e. not at an end). This composition is obtained by joining the composition of polynucleotides with deletions at the ends to other DNA ends, such that the deletion now occurs internally. Often the two ends to be ligated will be present on the same DNA molecule, such that the resulting ligation product is a circular polynucleotide. The number of deletions may range from 1 to 100 at each end, but more preferable is from 1 to 30.

[135] All references and patent publications referred to herein are hereby incorporated by reference herein.

[136] As can be appreciated from the disclosure provided above, the present invention has a wide variety of applications. Accordingly, the following examples are offered for illustration purposes and are not intended to be construed as a limitation on the invention in any way.

EXAMPLES

Example 1: Random cleavage of a plasmid

[137] Molecular evolution techniques utilizing insertions or deletions require a gene to be cleaved, at least transiently, a small number of times. Optimally, each molecule within a mix is cleaved once, at different random positions. There is significant difficulty in preparing singly cleaved DNA, wherein cleavage occurs at random positions. Biondi, et.al. described a cumbersome method using DNase I and DNA polymerase to induce nicks, followed by further cleavage of these nicks to produce a double stranded break [Biondi et al., *Nucleic Acids Res* 26: 4946-52 (1998)]. This process required tedious and time consuming cesium chloride gradient purification and linker ligation steps, and is not generally applicable to high throughput molecular biology techniques like molecular evolution.

[138] The strategy of utilizing a single-stranded endonuclease to induce double-stranded breaks at random positions in DNA has heretofore not been utilized. It was reasoned that a single-stranded nuclease, like S1, P1, or mung bean nuclease, would specifically cleave single-stranded regions in tightly supercoiled DNA, thus producing a nick. A nick is the natural substrate for these enzymes, so cleavage to produce a double-stranded break may then occur in the same reaction. Following cleavage, the single-stranded regions are no longer present since the plasmid is no longer supercoiled, so the DNA is no longer a substrate for the enzyme. Thus, cleavage would occur once and only once. This example illustrates the utility of this hypothesis.

[139] The plasmid pLacZi (Clontech, Palo Alto, CA) was used to illustrate the mechanism by which a polynucleotide can be cleaved at random positions. The plasmid was propagated in DH10B *E. coli* cells (Invitrogen, Carlsbad, CA) and plasmid was prepared by Qiagen maxiprep columns (Qiagen, Valencia, CA). Plasmid DNA at 200 ng/μl was incubated with 0.4, 2.0, 10, or 50 units of S1 nuclease (Promega, Madison, WI) in 1X S1 buffer (50 mM sodium acetate pH 4.5, 280 mM NaCl, 4.5 mM $ZnSO_4$) for 10 minutes at room temperature. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C for 10 minutes. Protein was removed by twice extracting with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate and resuspended in water.

[140] Cleaved pLacZi was analyzed by 1.5% agarose gel electrophoresis (Figure 4, panel A). S1 nuclease cleaved plasmid was seen to co-migrate with pLacZi cleaved with Cla I, which cuts pLacZi once. Thus, S1 nuclease can linearize a circular DNA

molecule. Although S1 nuclease is not known to cut DNA in a sequence specific manner, it was important to determine that the cleavage of plasmid by S1 was not site specific. To this end, linear plasmid produced by S1 cleavage was gel purified (Figure 4, panel B, lane 5), or purified and further cleaved with Cla I (lane 6). Controls included supercoiled plasmid (lane 2), plasmid linearized with Cla I (lane 3), or plasmid linearized with S1 nuclease and un-purified (lane 4). The S1/Cla I cleaved plasmid is seen as a smear, showing that S1 is cleaving in several different positions in the plasmid. If S1 cleaved at only one position, then the S1/Cla I cleaved plasmid would migrate as two bands; if S1 cleaved at two positions, then the S1/Cla I plasmid would migrate as three bands, and so on. The importance of this example is that a polynucleotide is able to be cleaved once (i.e. linearization of a circle), and only once, at different positions.

Example 2: Deletions at a site in LacZ

[141] Nucleotide deletions have been made for structural analysis of genes, and for nucleotide sequence analysis. Generally these deletions are large, in the range of well over 100 nucleotides. Under normal conditions, for example, exonuclease III removes over 100 bases per minute [Sambrook et al., (1989)]. The ability to create small deletions, however, would be useful to alter small domains in proteins or remove deleterious functions. In order to make small deletions at the end of a polynucleotide, exonuclease III was utilized under various conditions of salt (Figure 5) and temperature. A fluorescently labeled 232 base pair PCR product from pLacZi was exposed to 100 mM, 150 mM, and 200 mM NaCl in the presence of 10 U exonuclease III (New England Biolabs, Beverly, MA) in 10 µl of 66 mM Tris-Cl (pH 7.4), 0.66 mM $MgCl_2$ at 15°C in a 5 minute reaction. The reaction was stopped by the addition of EDTA to 0.025 M, and extracted once with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. DNA was resuspended in 20 µl deionized formamide, and 0.5 µl was run on a 6% polyacrylamide denaturing gel in ABI 373 sequencer (Perkin-Elmer, Foster City, CA) set to the genescan setting according to the manufacturers recommendation.

[142] Nearly 25 nucleotides can be removed under conditions of 100 mM NaCl (Figure 5, second panel), up to 15 nucleotides with 150 mM NaCl, and a few nucleotides with 200 mM NaCl (bottom panel).

[143] The Cla I site in pLacZi exists in the coding region of the LacZ gene. This site was utilized to make short deletions within the gene itself, which could then be analyzed further by PCR to determine the extent to which deletions were made. Additionally,

plasmids containing deletions were selected on LB agar plates containing 40 μg/ml X-Gal to determine the functionality of the LacZ gene. The pLacZi plasmid (10 μg) was linearized with Cla I in 200 μl, then incubated with 20 U of S1 nuclease in 400 μl to remove the 2 bp 5' overhangs. Further, the linearized plasmid was concentrated and filtered through an ultrafree MC membrane (30 kD cutoff, Millipore, Bedford, MA), then brought to a volume of 400 μl in 1X calf intestinal phosphatase buffer containing 100 U of calf intestinal phosphatase (New England Biolabs, Beverly, MA) and incubated for 45 minutes at room temperature. Plasmid was extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate, and resuspended in water. The plasmid was then incubated with exonuclease III as described in example 1, in the presence of either 100 mM, 150 mM or 200 mM NaCl for 5 minutes at 15°C in a 10 μl reaction. In a control arm, plasmid was not incubated with exonuclease III, to test for the frequency of religation of the dephosphorylated plasmid in the absence of deletions. After 5 minutes of exonuclease III reaction, a mix containing S1 nuclease 50 U in 1X S1 buffer was added. This mix was further incubated at room temperature for 15 minutes. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C for 10 minutes. The DNA was then extracted once with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, precipitated with sodium acetate and resuspended in 10 μl of 1X ligase buffer containing 1.0 U of T4 DNA ligase (Invitrogen, Carlsbad, CA). Ligation reactions were incubated at 15°C for 12 hours. Electroporation of *E. coli* strain DH10B (Invitrogen, Carlsbad, CA) was accomplished with 1.0 μl of ligation mix. Cells were plated on LB agar plates containing 40 μg/ml X-Gal and 100 μg/ml ampicillin and incubated overnight at 30°C. Table 1 illustrates the results of the plating experiment.

Table 1. Colony characteristics after site directed deletions.

	Blue Colonies	White Colonies	Blue/White
No Exo III	0	0	-
Exo III, 100 mM NaCl	177	66	0.37
Exo III, 150 mM NaCl	340	140	0.41
Exo III, 200 mM NaCl	77	34	0.44

[144] Notably, no background is realized when dephosphorylated plasmid is not exposed to exonuclease III (first row, Table 1). Several blue and white colonies are evident with exonuclease III treatment under different salt concentrations. Interestingly, the theoretical maximum of the blue/white ratio is 0.33, since at least 2/3 of religations should be out of frame. However, the blue/white ratio in this experiment is slightly more than 0.33, and appears to increase as salt concentration increases. This bias may be due to the fact that a one basepair deletion from one end would allow in-frame religation to occur, and fewer deletions are favored as salt is increased. The statistical significance of this result has not been analyzed, so the true frequency may actually be nearer to 0.33.

[145] Six of the colonies were analyzed by PCR with primers flanking the Cla I site. FIG 6 shows these results. In the upper panel the wild-type 312 basepair fragment from pLacZi is shown. Clone 1 contains an in frame deletion of 291 bases (PCR product of 291 bases) and retains a blue phenotype. Clone 2 contains a 4 basepair out of frame deletion (PCR product of 308 bases) and has a white phenotype. Clone 3 contains a 9 basepair in frame deletion (PCR product of 303 bases) and has a white phenotype. Clone 4 contains a 6 basepair in frame deletion (PCR product of 306 bases) and has a white phenotype. Clone 5 contains a 7 basepair out of frame deletion (PCR product of 305 bases) and has a white phenotype. Clone 6 has a 3 basepair deletion (PCR product of 309 bases) and has a blue phenotype. Although it may be thought that shorter deletions would lead to less severe phenotype, this experiment illustrates that this is not necessarily the case. Clone 1 contains a deletion encompassing 7 amino acids but retains function whereas clones 3 and 4 contain in frame shorter deletions but do not retain function. Furthermore, this example illustrates the ability of deletional technology to search functional sequence space.

Example 3: Insertions in LacZ

[146] Insertions of random DNA in the LacZ gene was accomplished by employing DNase I to fragment cDNA derived from CHO cells, followed by ligation of these fragments into linearized pLacZi. Since cDNA is by definition functional, it is contemplated that the use of cDNA will optimize the likelihood of obtaining functional proteins. CHO cell cDNA (5 μg) was fragmented with 0.001 units of DNase I in a buffer containing 40 mM Tris-Cl pH 7.4 and 10.0 mM $MgCl_2$ for 5 minutes at room temperature. The reaction was stopped by the addition of EDTA to 0.025 M and heated to 70°C in the presence of 10 μg of protease K. DNA was extracted with an equal volume of phenol:chloroform:isoamyl alcohol

(25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. Plasmid linearized with Cla I or S1 nuclease were dephosphorylated as described above, then again extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. To insert random cDNA fragments into plasmid DNA, 0.2 mg of linearized, dephosphorylated plasmid was incubated with 1 ng of cDNA fragments in the presence of T4 DNA ligase (1.0 U) in a reaction volume of 10 ml at 15°C for 12 hours. As controls, linearized plasmid was incubated with ligase in the absence of cDNA fragments, and cDNA fragments were incubated with ligase in the absence of linearized vector. DH10B *E. coli* were then electroporated with 1.0 µl of each ligation mix.

[147] Several *e. coli* colonies were identified in the vector plus insert arms of the experiment which exhibited either white, intermediate, or blue phenotype on X-Gal plates. PCR across the Cla I site in the colonies which arose from vector linearized with Cla I ligated to cDNA fragments revealed several clones containing inserts of sizes from 100-300 basepairs. Three of these are illustrated in FIG 7. Thus, the insertion of fragments of cDNA into a genetic element can be accomplished with the present invention.

Example 4: Functional changes at random positions

[148] The lac operon is a model system by which genetic elements are easily studied. The enzyme β-galactosidase is encoded by the LacZ gene, but is normally only produced when lactose is present in the environment. Control of enzyme levels is accomplished at the level of transcription. The lac repressor protein binds to the operator sequence upstream from the ATG start site of LacZ, and inhibits transcription by RNA polymerase. In the presence of lactose, however, the repressor is removed from the operator and transcription can proceed. The mechanism of promotor activation is through the binding of lactose, the inducer, to the lac repressor and causing an allosteric change that causes its affinity for the operator to decrease dramatically. In the laboratory setting, LacZ transcription can be assessed by plating *E. coli* on the colorimetric substrate X-Gal, which causes colonies to turn blue when hydrolyzed by β-galactosidase. The operator can be de-repressed by utilizing the lactose analog IPTG, which is non-hydrolizable, and strongly induces LacZ transcription by binding the lac repressor.

[149] In order to test the ability of random deletions to affect gene function, the pBluescript II KS+ plasmid was linearized with S1 nuclease, gel purified, dephosphorylated, and subjected to exonuclease III digestion as described in examples 1 and

2. Linearized plasmid at 20 ng/μl was incubated with 10 U exonuclease III in 66 mM Tris-Cl pH 7.4, 0.66 mM $MgCl_2$ buffer at 15°C for 5 minutes, followed by addition of 1X S1 solution containing 50 mM sodium acetate pH 4.5, 280 mM NaCl, 4.5 mM $ZnSO_4$ and 10 U S1 nuclease, and incubation for 15 minutes at room temperature. The reaction was stopped by adding EDTA to 0.025 M, and extraction with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), once with an equal volume of ether, and precipitated with sodium acetate. DNA was resuspended in 1X T4 DNA ligase buffer containing 1.0 U T4 DNA ligase and incubated at 15°C for 12 hours. The ligation reaction (1 μl) was then used to electroporate *E. coli* strain TOP 10 F', which produces the lac repressor protein (Invitrogen, Carlsbad, CA). The *E. coli* were incubated on LB plates either with or without IPTG as inducer, and in the presence of X-Gal to measure β-galactosidase activity. Additionally, pBluescript plasmid was plated in the presence or absence of IPTG on X-Gal containing plates. Table 2 illustrates the results of the experiment.

Table 2. Functional changes in transcription of β-galactosidase

	+ IPTG		- IPTG	
	Blue	**White**	**Blue**	**White**
pBluescript	100%	0	0	100%
pBluescript/deletions	66%	34%	2%	98%

Several colonies gained the ability to transcribe LacZ in the absence of the inducer IPTG in the arm of the experiment where deletions were made at random positions. Additionally, several colonies lost their ability to produce functional β-galactosidase in the presence of IPTG. One white colony in the presence of IPTG from the pBluescript/deletions arm was sequenced and found to have an eight basepair deletion at the translation start site. This sequence is illustrated below, with the translation start site (ATG) encoding methionine codon underlined.

CACACAGGAAA--------ACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

(SEQ ID NO: 1 and SEQ ID NO:2, respectively)

Thus, random cleavage of a plasmid, followed by short deletions made by exonuclease III can cause functional changes in regulatory and protein coding regions of genetic elements. These changes can then be detected with a functional assay.

SEQUENCE LISTING

SEQ ID NO:1

Mutation in 5' end of gene encoding β- galactosidase

CACACAGGAAAACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAA

SEQ ID NO:2

5' end of wild type gene encoding β- galactosidase

CACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGA

ACAA

WHAT IS CLAIMED IS:

1. A method for generating a library of polynucleotide sequences having nucleotide deletions at differing positions in a sequence of a genetic element comprising the steps of:

(a) subjecting multiple copies of circular polynucleotides comprising the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which removes at least one nucleotide from one of said ends of said polynucleotides producing a library of deletion polynucleotide sequences, said library comprising multiple deletion polynucleotide sequences with deletions at different random positions.

2. The method of claim 1, further wherein said polynucleotides from step (b) are subjected to a process that covalently joins said 3' and 5' ends to one another.

3. The method of claim 1, wherein said library of polynucleotides is further subjected to a process that selects for a function of interest.

4. The method of claim 1, wherein the cleavage occurs with an endonuclease.

5. The method of claim 4, wherein the endonuclease is S1.

6. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 5 individual polynucleotides each having a random deletion at a different position from the others.

7. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 10 individual polynucleotides each having a random deletion at a different position from the others.

8. The method of claim 1, wherein the library of deletion polynucleotides comprises at least 30 individual polynucleotides each having a random deletion at a different position from the others.

9. The method of claim 1, wherein the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element.

10. The method of claim 1, wherein steps (a) and (b) are repeated.

11. The method of claim 1, wherein step (b) further includes a process for inserting nucleotides at the position of deletion.

12. The method of claim 1, wherein 1-3 nucleotides are deleted in step (b).

13. 13. The method of claim 1, wherein 50-100 nucleotides are deleted in step (b).

14. A substantially pure composition comprising a library of multiple linear polynucleotides each having a different 3' and a 5' end, but each linear polynucleotide being identical to the others if circularized.

15. The composition of claim 14, wherein said library comprises at least 5 polynucleotides having a different 3' and a 5' end.

16. A substantially pure composition comprising a library of at least 2 deletion polynucleotides each differing from the other only by having a different random deletion.

17. The substantially pure composition of claim 16, wherein said deletion polynucleotides further comprise at least one nucleotide inserted at the position of deletion.

18. The composition of claim 16, wherein the library has at least 5 polynucleotides each differing from the other only by having a different random deletion.

19. A method for generating a library of polynucleotide sequences having nucleotide additions at random positions in a genetic element comprising the steps of:

(a) subjecting a composition of multiple copies of circular polynucleotides with the genetic element to random cleavage to obtain multiple linear polynucleotides each polynucleotide having at least one 3' and 5' end; and

(b) subjecting said polynucleotides from step (a) to a process which adds at least one nucleotide to one of said ends of said polynucleotides producing a library of

addition polynucleotide sequences, said library comprising multiple addition sequences with additions at different random positions.

20. The method of claim 19, further wherein said addition polynucleotides from step (b) are subjected to a process that covalently joins said 3' and 5' ends to one another.

21. The method of claim 19, further subjecting said library of polynucleotides to a process that selects for a function of interest.

22. The method of claim 19, wherein the cleavage occurs with an endonuclease.

23. The method of claim 22, wherein the endonuclease is S1.

24. The method of claim 19, wherein the library of addition polynucleotides comprises at least 5 individual polynucleotides each having a random addition of nucleotides at a different position from the others.

25. The method of claim 19, wherein the library of addition polynucleotides comprises at least 10 individual polynucleotides each having a random addition at a different position from the others.

26. The method of claim 19, wherein the library of addition polynucleotides comprises at least 30 individual polynucleotides each having a random addition at a different position from the others.

27. The method of claim 19, wherein the composition of multiple copies of circular polynucleotides is free of naturally-occurring homologs to the genetic element.

28. The method of claim 19, wherein steps (a) and (b) are repeated.

29. The method of claim 19, wherein step (b) includes a process for deleting nucleotides at the point of addition.

30. The method of claim 19, wherein 1-3 nucleotides are added in step (b).

31. 31. The method of claim 19, wherein 3-50 nucleotides are added in step (b).

32. The method of claim 19, wherein 50-100 nucleotides are added in step (b).

33. A substantially pure composition comprising a library of at least 2 addition polynucleotides each differing from the other only by having a different random addition.

34. A substantially pure composition comprising a library of at least 5 addition polynucleotides each differing from the other only by having a different random addition.

35. A method for producing short deletions from the end of a polynucleotide by incubating a population of polynucleotides with an exonuclease at a temperature from 0°C to 24°C in the presence of 10 to 500 mM salt, thereby producing a population of polynucleotides containing deletions of 1-100 residues from at least one end of the polynucleotide.

36. The method of claim 35, wherein the polynucleotide is double-stranded.

37. The method of claim 35, wherein the exonuclease is exonuclease III.

38. The method of claim 36, wherein the double-stranded nucleic acid is incubated with a single-stranded endonuclease to produce a blunt end.

39. The method of claim 35, further wherein the resulting population of polynucleotides containing deletions at the ends are covalently joined to at least a second end, producing a population of polynucleotides containing a deletion at an internal position.

40. The method of claim 38, wherein the single-stranded endonuclease is S1 nuclease.

41. The method of claim 39, wherein the polynucleotides resulting from covalent joining are circular polynucleotides.

42. The method of claim 35 wherein the population of polynucleotides contains deletions of 1-50 residues from at least one end of the polynucleotide.

43. The method of claim 35, wherein the population of polynucleotides contains deletions of 1-30 residues from at least one end of the polynucleotide.

44. A substantially pure composition of at least two polynucleotides each having two ends and each differing from one another only by having different deletions of 1 to 100 residues at one or both ends.

45. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 50 residues at one or both ends.

46. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 30 residues at one or both ends.

47. The composition of claim 44, wherein the composition of polynucleotides differs from one another by deletions of 1 to 10 residues at one or both ends.

48. A substantially pure composition of at least two polynucleotides each differing from one another only by deletions of 1 to 100 residues at a specific internal position within the polynucleotides.

49. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 50 residues at the specific internal position.

50. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 30 residues at the specific internal position.

51. The substantially pure composition of claim 48, wherein the polynucleotides differ from one another by deletions of 1 to 10 residues at the specific internal position.

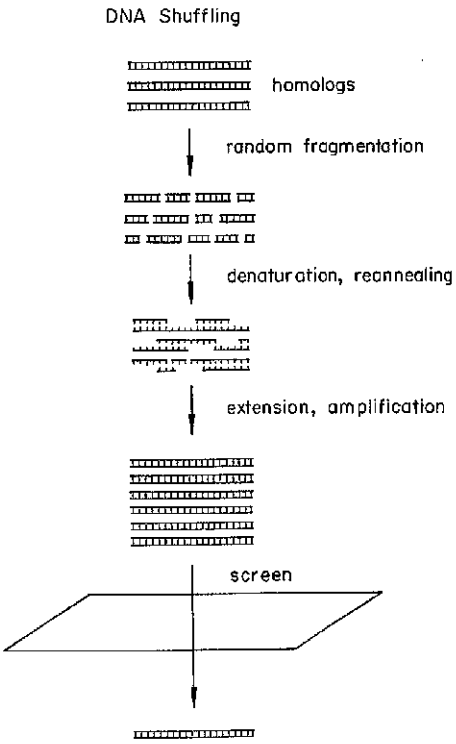

DNA Shuffling
homologs
random fragmentation
denaturation, reannealing
extension, amplification
screen

FIG. 1.

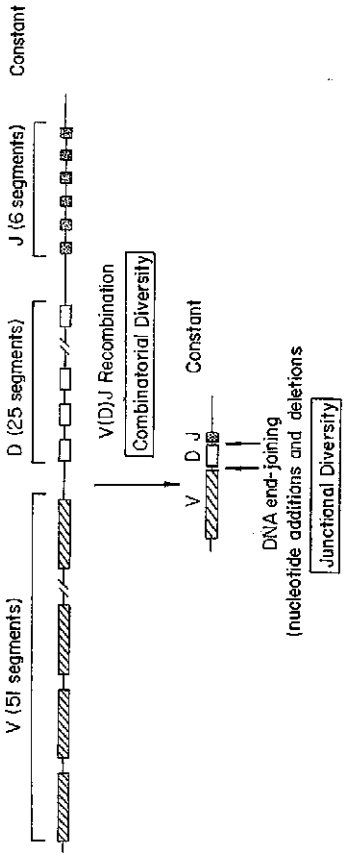

Constant
J (6 segments)
D (25 segments)
V (51 segments)
V(D)J Recombination
Combinatorial Diversity
V D J
Constant
DNA end-joining
(nucleotide additions and deletions
Junctional Diversity

FIG. 2.

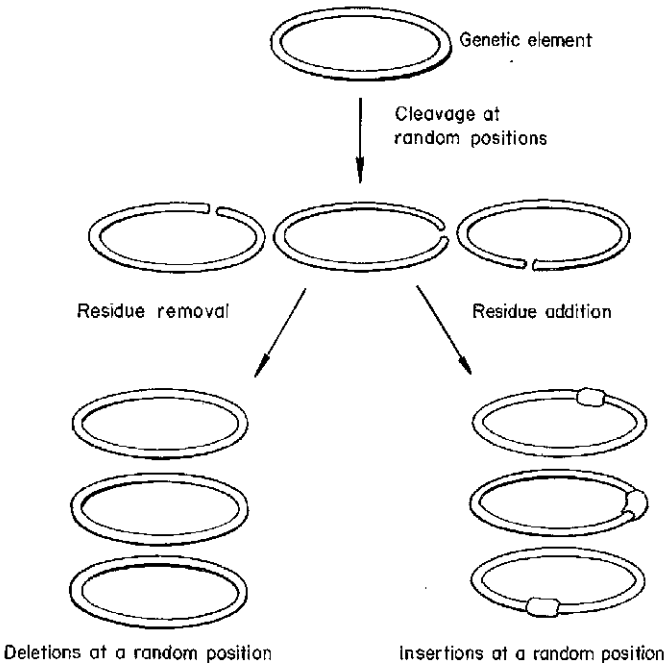

Genetic element
Cleavage at
random positions
Residue removal
Residue addition
Deletions at a random position
Insertions at a random position

FIG. 3.

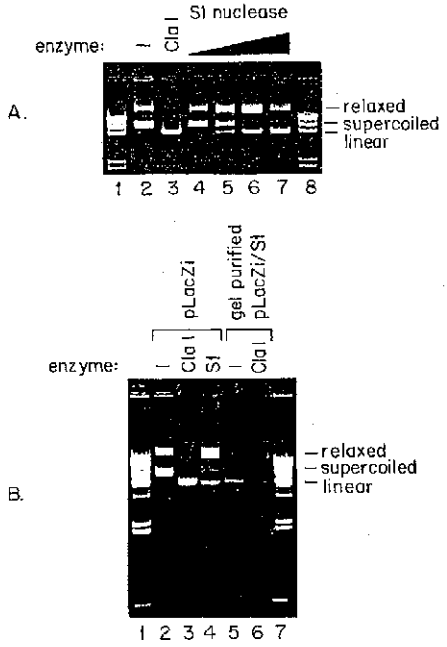

S1 nuclease
enzyme: – ClaI
A.
relaxed
supercoiled
linear
1 2 3 4 5 6 7 8
pLacZi
gel purified pLacZi/S1
enzyme: – ClaI S1 – ClaI
B.
relaxed
supercoiled
linear
1 2 3 4 5 6 7

FIG. 4.

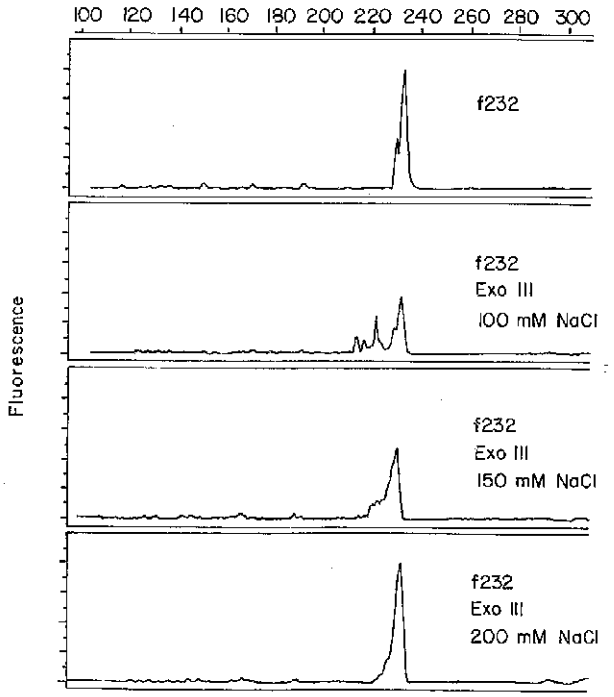

100 120 140 160 180 200 220 240 260 280 300
Fluorescence
f232
f232
Exo III
100 mM NaCl
f232
Exo III
150 mM NaCl
f232
Exo III
200 mM NaCl

FIG. 5.

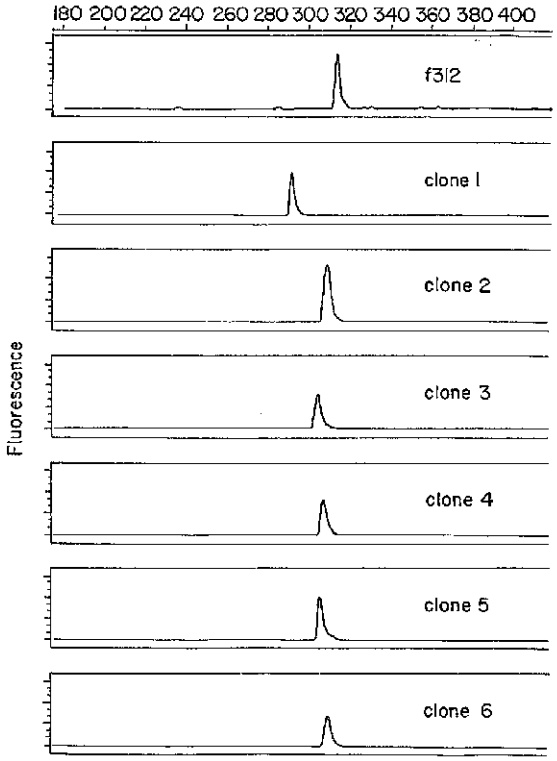

180 200 220 240 260 280 300 320 340 360 380 400
f3l2
clone l
clone 2
clone 3
clone 4
clone 5
clone 6
Fluorescence

FIG. 6.

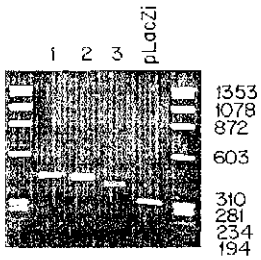


FIG. 7.

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/28788

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : C12Q 1/68; C12P 19/34; C12N 15/09; C07H 21/02
US CL : 435/6, 91.53, 91.1, 91.4, 91.42; 536/23.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 435/6, 91.53, 91.1, 91.4, 91.42; 536/23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WEST, DERWENT, CAPLUS, MEDLINE, EMBASE, SCISEARCH

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GREEN, C. et al. Targeted Deletions of Sequences from Closed Circular DNA. Proc. Natl. Acad. Sci. USA. May 1980, Vol. 77, No. 5, pages 2455-2459, entire document.	1-13
Y	WALDECK, W. et al. Random Cleavage of Superhelical SV40 DNA by S_1 Nuclease. Biochimica et Biophysica Acta. March 1976, Vol. 425, pages 157-167, entire document.	1-13

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"B" earlier document published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
06 NOVEMBER 2001

Date of mailing of the international search report
28 DEC 2001

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
JAISHRI PONNALURI
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/05765

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely.

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a)

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in the international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-15

Remark on Protest ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 (continuation of first sheet(1)) (July 1998)★

INTERNATIONAL SEARCH REPORT

International application No
PCT/US01/20782

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I, claim(s)1-13, drawn to a method for generating a library of polynucleotide sequences having nucleotide deletions at differing positions in a sequence of genetic element.
Group II, claim(s) 14-16, drawn to a substantially pure composition comprising a library of multiple linear polynucleotides.
Group III, claim(s) 19-32, drawn to a method for generating a library of polynucleotides sequences having nucleotide additions.
Group IV, claim(s)33-36, drawn to a substantially pure composition comprising a library of at least 2 addition polynucleotides each differing from each other only by having a different random addition.
Group V, claim(s)38-43, drawn to a method for producing a short deletions from end of a polynucleotide.
Group VI, claim(s) 44-47, drawn to a substantially pure composition of at least two polynucleotides each having two ends and each differing from one another.
Group VII, claim(s)48-51, drawn to substantially pure composition of at least two polynucleotides each differing from one another by deletion of residues at a specific internal position within the polynucleotides
The inventions listed as Groups I-VII do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: The special technical feature of group I is polynucleotides having nucleotide deletions at the ends; the special technical feature of group II is multiple polynucleotides having different ends; the special technical feature of group III inventions is polynucleotides having a nucleotide additions at random positions; the special technical feature of group IV is a library of polynucleotides having addition polynucleotides; the special technical feature of group V inventions is a deletions at the ends of the polynucleotides; the special technical feature of group VI is polynucleotides having deletions at either end; the special technical feature of group VII is a polynucleotides having deletions at random positions internally. The special technical feature of group I well known in the art (see Green et al PNAS, vol. 77, pp 3566-3569, May 1980) Since the special technical feature is known the art, the invention lacks unity.

Form PCT/ISA/210 (extra sheet) (July 1998)*

フロントページの続き

(81)指定国 AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,CH,CY,DE,DK,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NO,NZ,PH,PL,PT,RO,RU,SD,SE,SG,SI,SK,SL,TJ,TM,TR,TT,TZ,UA,UG,US,UZ,VN,YU,ZA,ZW

(72)発明者 スマイダー, ボーン
アメリカ合衆国 カリフォルニア ９４５０１, アラメダ, パール ストリート １８２３－エイ

Ｆターム(参考) 4B024 AA20 CA01 GA30 HA08