Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2788452 C 2017/02/21

(11)(21) **2 788 452**

(12) **BREVET CANADIEN**
CANADIAN PATENT

(13) **C**

(86) **Date de dépôt PCT/PCT Filing Date:** 2009/01/29
(87) **Date publication PCT/PCT Publication Date:** 2010/08/05
(45) **Date de délivrance/Issue Date:** 2017/02/21
(85) **Entrée phase nationale/National Entry:** 2012/07/26
(86) **N° demande PCT/PCT Application No.:** JP 2009/051512
(87) **N° publication PCT/PCT Publication No.:** 2010/086990

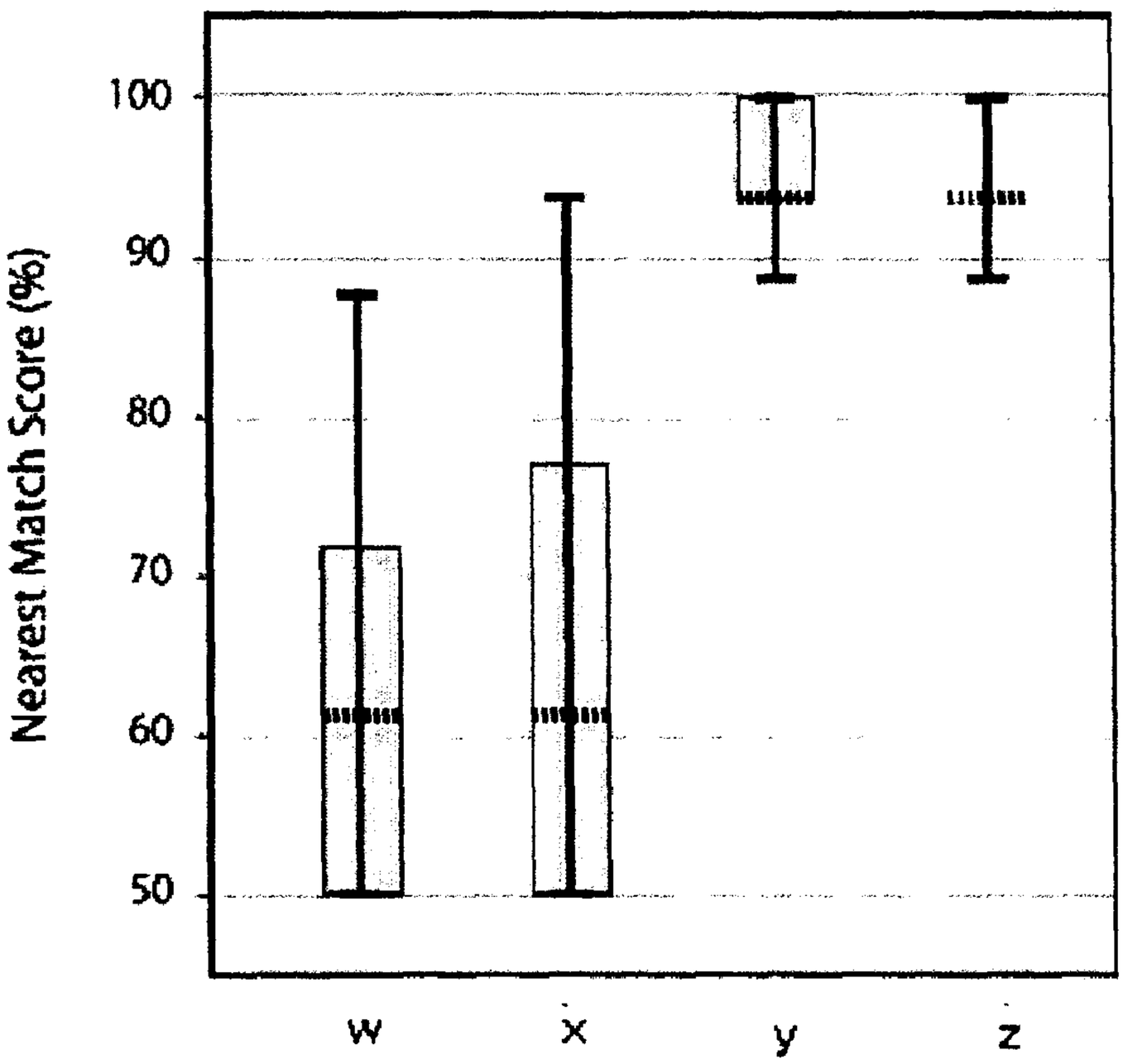
(51) **Cl.Int./Int.Cl.** ***C12N 15/09*** (2006.01),
C12N 1/15 (2006.01), ***C12N 1/19*** (2006.01),
C12N 1/21 (2006.01), ***C12N 5/10*** (2006.01),
C12Q 1/68 (2006.01)

(72) **Inventeurs/Inventors:**
SUGAHARA, JUNICHI, JP;
SEKIYAMA, KAZUHIDE, JP

(73) **Propriétaire/Owner:**
SPIBER INC., JP

(74) **Agent:** MARKS & CLERK

(54) **Titre : PROCEDE DE FABRICATION D'UNE ETIQUETTE ADN**
(54) **Title: METHOD OF MAKING DNA TAG**



(57) **Abrégé/Abstract:**
Provided is a method of determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence. The method includes: a step (S1) of obtaining

CA 2788452 C 2017/02/21

(11)(21) **2 788 452**

(13) **C**

(57) **Abrégé(suite)/Abstract(continued):**
a protein coding sequence of the genomic DNA sequence; a step (S2) of determining a region to be treated in the protein coding sequence; a step (S3) of fragmenting the coding sequence in the region to be treated; a step (S4) of subjecting the fragmented coding sequences to a silent mutation; a step (S5) of determining sequences suitable as primers from polynucleotides including the fragmented coding sequences or complementary base sequences thereof; a step (S6) of performing a homology search for the coding sequences thus determined; and a step (S8) of determining the determined coding sequence corresponding to a minimum value of an NMS as the DNA tag, and determining the site of the fragmented coding sequence corresponding to the determined coding sequence in the protein coding sequence as the introduction site.

Abstract

Provided is a method of determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence.

The method includes: a step (S1) of obtaining a protein coding sequence of the genomic DNA sequence; a step (S2) of determining a region to be treated in the protein coding sequence; a step (S3) of fragmenting the coding sequence in the region to be treated; a step (S4) of subjecting the fragmented coding sequences to a silent mutation; a step (S5) of determining sequences suitable as primers from polynucleotides including the fragmented coding sequences or complementary base sequences thereof; a step (S6) of performing a homology search for the coding sequences thus determined; and a step (S8) of determining the determined coding sequence corresponding to a minimum value of an NMS as the DNA tag, and determining the site of the fragmented coding sequence corresponding to the determined coding sequence in the protein coding sequence as the introduction site.

DESCRIPTION

METHOD OF MAKING DNA TAG

Technical Field

[0001]

The present invention relates to a method of determining a DNA sequence which shows high primer specificity in a group of unspecified DNA sequences and an introduction site of the DNA sequence into genomic DNA, and a method of introducing the sequence.

Background Art

[0002]

A treatment (cleanup) for restoring a natural environment polluted with a harmful substance to its unpolluted state, which includes no harmful substance, using a microorganism (such as a soil-improving bacterium) is called bioremediation. In the bioremediation, in order to grasp the degree of progress of the cleanup, it is essential to accurately monitor a microorganism spread in an environment. Further, it is necessary to accurately grasp the number of residual microorganism cells at a time when the cleanup is completed.

[0003]

An existing procedure includes preparing total DNA in an environment (environmental DNA) where a microorganism is spread and performing a quantitative polymerase chain reaction (PCR) using,

as a primer, a DNA sequence specific to the microorganism to roughly grasp the number of the microorganism cells (for example, see Non-patent Document 1). However, it is difficult to accurately grasp the number of the microorganism cells because the primer to be used in the quantitative PCR may react with non-specific DNA sequence (DNA sequences other than DNA sequences of interest in the environment, for example).

[0004]

Hitherto, many reports have been made on technologies for improving primer specificity in PCR and programs for the technologies (for example, see Non-patent Documents 2 and 3). However, all the technologies are specialized for an organism which has been isolated and cultured. If the organism which has been isolated and cultured is used, probability of a reaction of the primer designed with a region other than a region to be treated can be estimated by a search of the genome. Meanwhile, in the environmental DNA in which a large indefinite number of organisms coexist, it is very difficult to estimate the specificity of the primer.

[0005]

Further, the law required by the international treaty relating to a rule of handling of recombinant organisms (Cartagena Protocol) (hereinafter, sometimes referred to as "Cartagena Law") prohibits spread of recombinant microorganisms to open environments. Therefore, it is impossible to make an appropriate artificial DNA sequence which shows high primer specificity and introduce the

sequence into a microorganism to be used in bioremediation or the like without careful consideration. Meanwhile, there has not been proposed a technology for introducing an artificial primer sequence into a microorganism to be used in bioremediation in such a range that the microorganism does not fall within the category "recombinant organism" specified by the Cartagena Law.

[0006]

The bioremediation has advantages in that the cost and energy consumption are low as compared to a physicochemical cleanup method and that the method does not burden an ecosystem because the technology is mild, for example. On the other hand, sufficient findings on effects of spread of a microorganism for cleanup on an environment have not been obtained. Therefore, to grasp not only the degree of progress of cleanup but also the degree of dispersion of the microorganism spread and effects on the ecosystem in the field of spread as well, means for accurately monitoring has been strongly desired.

Prior Art Documents

[0007]

Non-patent Document 1: http://www.ritsumei.ac.jp/se/rc/staff/kubo/intro/biorem/index.html

Non-patent Document 2: Qu W, Shen Z, Zhao D, Yang Y, Zhang C (2009) MFEprimer: multiple factor evaluation of the specificity of PCR

primers. Bioinformatics. 15;25:276-278.

Non-patent Document 3: Miura F, Uematsu C, Sakaki Y, Ito T (2005) A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. Bioinformatics. 15;21:4363-4370.

Disclosure of the Invention

Problem to be solved by the Invention

[0008]

A main object of the present invention is to provide a technology for introducing a DNA sequence which is highly specific to a primer (hereinafter, sometimes referred to as "DNA tag") into a genome of a microorganism in an environment in which unspecified organisms coexist.

Means for solving the Problem

[0009]

The inventors of the present invention have made intensive studies to solve the above-mentioned problems, and as a result, have designed a program for selecting a DNA sequence which does not tend to be universally used by an organism and easily selecting a primer which can be universally used in environmental DNAs in which a large indefinite number of organisms coexist. The inventors of the present invention have further found that the degree of dispersion of a microorganism of interest or the degree of progress

of cleanup can be detected accurately and simply by using a microorganism in which the sequence (or a complementary sequence thereof) which is used at a low frequency in the organism obtained as described above has been integrated into a genome for bioremediation and by using a primer including a sequence which is used at a low frequency in the organism. The present invention has been completed by further studies based on such findings.

[0010]

The present invention provides the following method of determining a DNA tag and an introduction site.

[0011]

Item 1. A method of determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence, the method including:

a step S1 of obtaining a protein coding sequence from the genomic DNA sequence;

a step S2 of determining a first coding sequence as a region to be treated from the protein coding sequence;

a step S3 of obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and recording sites of the second coding sequences in the protein coding sequence;

a step S4 of obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences

obtained in the step S3 to a silent mutation;

a step S5 of judging whether or not the third coding sequences satisfy a predetermined condition and determining only the third coding sequences that satisfy the predetermined condition as fourth coding sequences;

a step S6 of performing a homology search for the fourth coding sequences and obtaining homologous base sequences;

a step S7 of determining an NMS for each of the fourth coding sequences; and

a step S8 of determining the fourth coding sequence corresponding to a minimum value of the NMS as the DNA tag and determining the site of the second coding sequence corresponding to the fourth coding sequence as the introduction site, in which:

in the step 2, the first coding sequence includes a coding sequence in a region having no influence on a biological function of the organism even if the sequence is subjected to the silent mutation;

the predetermined condition includes a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates the degree of homology.

[0012]

Item 2. A method of determining a DNA tag and an introduction site according to the item 1, further including, when the fourth coding sequences having a plurality of NMS's which are the same

exist, a step S9 of determining the fourth coding sequence corresponding to an NMS that appears at the minimum frequency out of the NMS's as the tag.

[0013]

Item 3. A method of determining a DNA tag and an introduction site according to the item 1 or 2, in which the condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer includes the following:

the complementary base sequence has a CG content of 45 to 55%;

the complementary base sequence has a tm value of 55 to 65°C; and

the complementary base sequence is free of four or more consecutive identified bases.

[0014]

Item 4. A program for allowing a computer to execute a function for determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence, the program including allowing a computer to execute:

a first function for obtaining a protein coding sequence from the genomic DNA sequence;

a second function for determining a first coding sequence as a region to be treated from the protein coding sequence;

a third function for obtaining a plurality of second coding

sequences each including a partial sequence in the first coding sequence and having a predetermined length;

a fourth function for obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the third function to a silent mutation;

a fifth function for judging whether or not the third coding sequences satisfy a predetermined condition and determining only the third coding sequences that satisfy the predetermined condition as fourth coding sequences;

a sixth function for performing a homology search for the fourth coding sequences and obtaining homologous base sequences;

a seventh function for determining an NMS for each of the fourth coding sequences; and

an eighth function for determining the fourth coding sequence corresponding to a minimum value of the NMS as the DNA tag and determining the site of the second coding sequence corresponding to the fourth coding sequence as the introduction site, in which:

in the step 2, the first coding sequence includes a coding sequence in a region having no influence on a biological function of the organism even if the sequence is subjected to the silent mutation;

the predetermined condition includes a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates the degree of homology.

[0015]

Item 5. A method of manufacturing a DNA-tag-introduced mutant, the method including the following steps:

(i) preparing the DNA tag determined by the method according to any one of Items 1 to 3 and cloning the DNA tag into a vector including a marker gene;

(ii) transforming a cell of an organism with the vector having the DNA tag cloned in the step (i); and

(iii) removing the marker gene by continuous culture or subculture of the cell transformed in the step (ii) and obtaining a mutant having the DNA tag introduced into a genome through homologous recombination.

[0016]

Item 6. A method according to the item 5, further including the step of confirming that the marker gene is absent in the genome of the organism.

[0017]

Item 7. A mutant, which is obtained by the method according to the item 5 or 6.

[0018]

Item 8. A vector, including a DNA tag determined by the method according to any one of Items 1 to 3.

[0019]

Item 9. A genomic DNA, including a DNA tag introduced by the vector according to the item 8.

[0020]

Item 10. A mutant, including the genomic DNA according to the item 9.

[0021]

Item 11. A method of monitoring progress of cleanup by quantifying the number of mutants each including a DNA tag determined by the method according to any one of Items 1 to 3 in an environment using a primer including a polynucleotide having a base sequence of the DNA tag or a complementary base sequence thereof.

[0022]

Item 12. A method of evaluating diffusion of a mutant including a DNA tag determined by the method according to any one of Items 1 to 3 in an environment using a primer including a polynucleotide having a base sequence of the DNA tag or a complementary base sequence thereof.

[0023]

Item 13. A method of labeling an organism, the method including introducing a DNA tag determined by the method according to any one of Items 1 to 3 into genomic DNA of an organism using a vector having the DNA tag cloned thereinto.

According to an aspect of the present invention, there is provided a method of determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence, the method comprising:

(A) performing, by a processing unit of a computer having stored thereon instructions executable by the computer, the following steps S1 to S8:

a step S1 of obtaining a protein coding sequence from the genomic DNA sequence;

a step S2 of determining a region within the protein coding sequence obtained in step S1 that has no influence on a biological function of the organism even if the sequence is subjected to a silent mutation, as a first coding sequence;

a step S3 of obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and recording sites of the second coding sequences in the protein coding sequence;

a step S4 of obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the step S3 to a silent mutation;

a step S5 of judging whether or not each of the third coding sequences obtained in step S4 satisfy a predetermined condition and determining one or more sequences among the third coding sequences judged to satisfy the predetermined condition as fourth coding sequences;

a step S6 of performing a homology search for each of the fourth coding sequences;

a step S7 of determining an NMS for each of the fourth coding sequences using the results of the homology search performed in step S6; and

a step S8 of determining the fourth coding sequence corresponding to the NMS with the minimum value among the NMSs determined for each of the fourth coding sequences in step S7 as the DNA tag, and determining the site of the second coding sequence corresponding to the fourth coding sequence determined as the DNA tag, as the introduction site, wherein:

the predetermined condition comprises a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates a degree of homology; and

(B) recording the determined DNA tag and the introduction site to a recording unit.

According to another aspect of the present invention, there is provided a method of manufacturing a DNA-tag-introduced mutant, the method comprising the following steps:

(i) preparing the DNA tag determined by the method as described herein and cloning the DNA tag into a vector including a marker gene;

(ii) transforming a cell of an organism with the vector having the DNA tag cloned in the step (i); and

(iii) removing the marker gene by continuous culture or subculture of the cell transformed in the step (ii) and obtaining a mutant having the DNA tag introduced into a genome through homologous recombination.

According to another aspect of the present invention, there is provided a method of labeling an organism, the method comprising:

(i) preparing the DNA tag determined by the method as described herein; and

(ii) introducing the DNA tag into the introduction site of the genomic DNA of the organism, the introduction site determined by the method as described herein, using a vector having the DNA tag cloned thereinto.

According to another aspect of the present invention, there is provided a method of manufacturing a vector, the method comprising:

(i) preparing the DNA tag determined by the method as described herein; and

(ii) cloning the DNA tag into a vector including a marker gene.

According to another aspect of the present invention, there is provided a computer readable medium having stored thereon instructions that when executed by a computer, perform a method of determining a DNA tag which is a base sequence to be introduced into a genomic DNA sequence of an organism and an introduction site of the DNA tag into the genomic DNA sequence, the method comprising:

a first function for obtaining a protein coding sequence from the genomic DNA sequence;

a second function for determining a region within the protein coding sequence obtained by the first function that has no influence on a biological function of the organism even if the sequence is subjected to a silent mutation, as a first coding sequence;

a third function for obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and

recording sites of the second coding sequences in the protein coding sequence;

a fourth function for obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the third function to a silent mutation;

a fifth function for judging whether or not each of the third coding sequences obtained by the fourth function satisfy a predetermined condition and determining one or more sequences among the third coding sequences judged to satisfy the predetermined condition as fourth coding sequences;

a sixth function for performing a homology search for each of the fourth coding sequences;

a seventh function for determining an NMS for each of the fourth coding sequences using the results of the homology search performed by the sixth function; and

an eighth function for determining the fourth coding sequence corresponding to the NMS with a minimum value among the NMSs determined for each of the fourth coding sequence in the seventh function as the DNA tag and determining the site of the second coding sequence corresponding to the fourth coding sequence determined as the DNA tag as the introduction site, wherein

the predetermined condition comprises a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates a degree of homology.

Effects of the Invention

[0024]

According to the method of the present invention, it is possible to artificially make the DNA sequence (sometimes referred to as

"DNA tag") which rarely appears in a natural environment. A primer having such DNA sequence (or a complementary strand thereof) can drastically improve primer specificity in quantitative PCR to be used, for example, for grasping the number of microorganism cells in the field of bioremediation or the like and can accurately and simply grasp the degree of progress in cleanup and the degree of dispersion of a microorganism.

[0025]

As mentioned above, the Cartagena Law forbids spread of a recombinant microorganism to an open environment. However, a recombinant obtained by self-cloning (including a silent mutation) does not fall within the category "recombinant organism" specified by the Cartagena Law. Therefore, a mutant, which does not undergo any alteration to a genomic DNA sequence in a cell other than a region into which the DNA tag is introduced, does not conflict with the Cartagena Law and can also be used in the open environment. That is, according to the method of the present invention, it is possible to introduce the DNA tag into a genome of an organism so that the recombinant does not fall within the category "recombinant organism" specified by the Cartagena Law.

[0026]

Meanwhile, for example, if the thus obtained microorganism including the DNA tag introduced is used for bioremediation, the microorganism can be monitored at high accuracy.

Brief Description of the Drawings

[0027]

FIG. 1 is a block diagram illustrating a configuration of a device to be used in a method of determining a tag and an introduction site according to an embodiment of the present invention.

FIG. 2 is a flow chart illustrating a method of determining a tag and an introduction site according to an embodiment of the present invention.

FIG. 3 is a diagram illustrating a relationship of CDSs to be produced.

FIG. 4 is a diagram illustrating a relationship between a CDS to be treated and fragmented CDSs.

FIG. 5 is a diagram illustrating one example of a silently mutated CDS and results of a homology search using the sequence.

FIG. 6 is a graph illustrating lengths of bases and probability of occurrence thereof in a metagenome (the vertical axis has a logarithmic scale) in the respective trials.

FIG. 7 is a boxplot illustrating results of a homology search of the respective data sets on an environmental DNA database. In this figure, the dotted lines show mean values, the both ends of the lines show the maximum values and minimum values, and the both ends of the boxes represent the first quartiles and third quartiles.

FIG. 8 is a scheme illustrating examples of a method of introducing a DNA tag. In the item (1) of this figure, DNA including a base sequence corresponding to a first coding sequence is cloned

into a vector including an antibiotic-resistant marker for a host organism. In the item (2), a point mutation is introduced into DNA including a base sequence corresponding to the first coding sequence cloned. In the item (3), the vector is introduced into the host, and selection is performed using an antibiotic (in this procedure, a mutant produced by homologous recombination is obtained). In the item (4), the thus obtained mutant is cultured in a medium containing no antibiotic. In the item (5), a mutant which has lost its antibiotic resistance by the homologous recombination is obtained. In the item (6), a mutant which includes the DNA tag introduced properly and has no vector sequence is obtained by sequencing.

FIG. 9 is a schematic diagram illustrating pBS4106ID constructed in Reference Test Example 3.

FIG. 10 is a diagram illustrating results of sequencing (first half) performed in Reference Test Example 3.

FIG. 11 is a diagram illustrating results of sequencing (latter half) performed in Reference Test Example 3.

Modes for carrying out the Invention

[0028]

Hereinafter, the present invention is described in detail.

[0029]

(1) Method of determining DNA tag and introduction site

Hereinafter, embodiments according to the present invention are described based on the accompanying drawings.

FIG. 1 illustrates a block diagram illustrating a configuration of a device to be used for a method of determining a DNA tag and an introduction site thereof according to an embodiment of the present invention. Here, the "DNA tag" refers to a DNA sequence which rarely appears in a natural environment (the sequence is a DNA sequence that may be a positive strand or an opposite strand, and the sequence per se or a complementary sequence thereof is used as a "primer"). The "introduction site" of the DNA tag refers to a site in a DNA sequence of an organism into which the DNA tag is introduced (the DNA tag is substituted for a partial sequence).

[0030]

In the present invention, the species of the organism into which the DNA tag is introduced is not particularly limited as long as the organism has a DNA where the tag can be introduced by a known method, and examples thereof include prokaryotes, archaea, and eukaryotes. In the present invention, it is preferred to use a microorganism such as *Escherichia coli*, *Bacillus subtilis*, or yeast which is known for its culture condition and storage condition and which can be stored for a relatively long period of time. Moreover, a microorganism to be used for environmental cleanup is preferably used because the microorganism can be easily monitored when used in bioremediation. Examples thereof include bacteria each having an ability to degrade petroleum, bacteria each having an ability to degrade a variety of harmful chemical substances, and bacteria which absorb salts in environments. Specific examples thereof

include: *Bacullus* bacteria such as *Bacullus* sp. ODM157 and *Bacullus* sp. ODNM4, *Bacullus* sp. F31; *Rhodococcus* bacteria such as *Rhodococcus* sp. ODNM2B, *Rhodococcus* sp.NDMI144, *Rhodococcus* sp.NDKK48, *Rhodococcus* sp.NDKK7, *Rhodococcus* sp.NDKK6, *Rhodococcus* sp. NDKK5, *Rhodococcus* sp. NDKK2, *Rhodococcus* sp. NDKK1, *Rhodococcus* sp. NDMI54, *Rhodococcus* sp. ODNM1C, *Rhodococcus* sp. NDKY3D, and *Rhodococcus* sp. NDKY72A; *Gordonia* bacteria such as *Gordonia* sp. NDKY76A, *Gordonia* sp. NDKK46, *Gordonia* sp. NDKY2B, and *Gordonia* sp. NDKY2C; *Acientobacter* bacteria such as *Acientobacter* sp. ODYM1, *Acientobacter*sp. ODYM2, *Acientobacter* sp. ODYM5, *Acientobacter* sp. ODDK71, *Acientobacter* sp. ODMI29, *Acientobacter* sp.ODNM6, *Acientobacter* sp.NDMI119, *Acientobacter* sp. A132, *Acientobacter* sp. NDMI78, and *Acientobacter* sp. YM3; and *Pseudomonas* bacteria such as *Pseudomonas* sp. F721 and *Pseudomonas* sp. F722. Meanwhile, the present invention also includes introduction of a DNA sequence tag into a plasmid.

[0031]

This device includes a computer 1, a display unit 2, and an operation unit 3. Examples of the display unit 2 include a liquid-crystal display device and a CRT display device. Examples of the operation unit 3 include a keyboard and a mouse for computer. The computer 1 includes a processing unit (hereinafter, referred to as "CPU") 11, a rewritable memory 12 capable of temporarily storing data (hereinafter, referred to as "RAM"), a recording unit 13 such as a rewritable hard disk drive capable of continually storing data,

an interface unit (hereinafter, referred to as "IF unit") 14, and an internal bus 15. In the recording unit 13, a program and data performed by the CPU 11 are recorded. The IF unit 14 plays a role as an interface between the computer 1 and an external apparatus. That is, the CPU 11 allows the display unit 2 to display processing results or the like via the IF unit 14, and in the case where the operation unit 3 is operated by, for example, a man, information relating to the operation is obtained via the IF unit 14. Meanwhile, the computer 1 is connected to a network 4 via the IF unit 14. The network 4 may be a public network such as internet or a local network. A database (hereinafter, referred to as "DB") 5 where genomic DNAs are recorded is also connected to the network 4. The respective components of the computer 1 exchange the data via the internal bus 15. This procedure enables the computer 1 to perform not only the arithmetic processing of the computer itself but also processing such as acquirement of instructions from the operation unit 3 or use of DB5. FIG. 1 illustrates one DB5, but the apparatus may have a plurality of DBs.

[0032]

Next, a specific description is made on the method of determining the tag and introduction site of the tag according to the embodiment of the present invention with reference to the flow chart illustrated in FIG. 2. In addition, FIG. 3 is a block diagram illustrating a relationship of sequences prepared by treatment as mentioned below.

[0033]

Hereinafter, the processing is performed by the CPU 11 unless otherwise specified. Further, the CPU11 appropriately reads necessary data from the recording unit 13 to the RAM 12, performs processing using a predetermined region of the RAM 12 as a work region, and appropriately records temporary processing results or final processing results in the recording unit 13. In addition, initial data is recorded in advance in the recording unit 13.

[0034]

In a step S1, genomic DNA, i.e., a base sequence (single-stranded sequence) information of the organism where the tag is introduced is obtained. For example, genomic DNA information recorded in advance in the recording unit 13 (for example, text data) is read out. Alternatively, the genomic DNA information may be obtained from DB5.

[0035]

From the genomic DNA information thus obtained, all DNA sequence regions encoding proteins (CDSs: Coding Sequences) are determined. A query is performed for existing DB5 to perform annotation of the CDS, i.e., determination of a DNA region (positional information) encoding the protein using CDS information, if present, or by a known informational procedure. The top of FIG. 3 illustrates a determined protein CDS. FIG. 4 illustrates a more specific example of the CDS. In FIG. 4, each codon in the base sequence is surrounded by a square, and amino acids encoded are illustrated on the respective

squares.

[0036]

In a step S2, regions to be treated in the following steps are determined from all the protein coding sequences determined in the step S1. That is, a region which is considered to have a large effect on proliferation or phenotype of the host cells (organisms) by addition of the silent mutation is excluded from the protein coding sequences. Examples of the regions to be excluded include the whole sequences of essential genes, functional sequences, and sequences which are located at near a translation starting point and are considered to be significantly involved in control of efficiency of protein translation (about 50 bases located on the downstream), and a sequence located at near a stop codon (about 50 bases located on the upstream). Further, in the case of organisms having industrial values, examples thereof include genes which may impair the industrial values by mutations (for example, in the case of microorganisms capable of producing useful substances, examples thereof include the whole sequences of genes which are directly involved in production of the useful substances). Examples of the functional sequences include sequences which recognize restriction enzymes and sequences which recognize nucleic acid binding proteins. Therefore, usually, a plurality of CDSs to be treated having different lengths are determined. The thus-determined protein coding sequences are sometimes referred to as "CDSs to be treated" (or "first coding sequences"). FIG. 3 illustrates the determined CDS

to be treated.

[0037]

The determined CDSs to be treated are recorded in the recording unit 13 in relation to the genomic DNA information. In this procedure, the positional information of the CDSs to be treated in the protein coding sequence (hereinafter, also referred to as "first positional information") is also recorded in the recording unit 13 in relation to the CDS to be treated. That is, the {genomic DNA information, CDS to be treated, and first positional information} are recorded as one set.

[0038]

In a step S3, each of the CDSs to be treated determined in the step S2 is fragmented into a designated size. That is, a continuous CDS having a predetermined length (hereinafter, also referred to as "fragmented CDS" or "second coding sequence") is selected from the CDSs to be treated. For example, as illustrated in FIG. 4, fragmented CDSs each having a designated size (for example, 12 bases) are selected with shifting the starting position one base by one base. Therefore, in the case where each fragmented CDS includes m base sequences (designated size: m), n-m+1 fragmented CDSs are determined from one CDS to be treated which includes n base sequences. FIG. 3 illustrates the fragmented CDSs determined. It should be noted that this procedure does not judge whether or not a plurality of the same fragmented CDSs are produced. This is because even if the same fragmented CDSs are used, final sequences

produced by the silent mutation may be different depending on insertion sites into a target organism (that is, depending on the upstream and downstream sequences in the fragmented CDS). In addition, if the upstream and downstream sequences are the same, this procedure does not limit one sequence because the positional information from which the fragmented CDSs are obtained becomes an important factor in the future. It should be noted that a plurality of the same fragmented CDSs are less likely to be produced.

[0039]

The size of each of the fragmented CDSs may be arbitrarily set as long as the size is a length which allows the CDSs to act as primers. Usually, the size is desirably 15 to 30 bases, more desirably 17 to 25 bases.

[0040]

In this procedure, the original CDS to be treated from which the fragmented CDSs are produced, and positional information of the fragmented CDSs in the CDS to be treated (hereinafter, also referred to as "second positional information") are recorded in the recording unit 13 in relation to the fragmented CDSs. That is, the {CDS to be treated, fragmented CDSs, and second positional information} are recorded as one set. Therefore, if one fragmented CDS is designated, the second positional information corresponding to the sequence, and the CDS to be treated and first positional information corresponding to the sequence are determined. From the first positional information and second positional information,

the position of the fragmented CDS in the genome sequence is specified.

[0041]

In a step S4, each of the fragmented CDSs determined in the step S3 is subjected to the silent mutation, to thereby obtain one or more silently mutated CDSs (hereinafter, also referred to as "third coding sequences"). That is, CDSs are prepared by substituting synonymous codons (which have different base sequences but are translated into the same amino acid) for codons in each fragmented CDS. For example, in the case where one fragmented CDS includes ATTCTGCACGAT and the position of the base sequence on the 5'-end of a protein coding sequence to be treated is 0, if the position of the base on the 5'-end of the fragmented CDS of the protein coding sequence to be treated is a multiple of three (calculated from the first and second positional information), the sequence includes four complete codons and is translated into the amino acids: Ile-Leu-His-Asp. For the amino acids, there are synonymous codons as shown in Table A.

[0042]

[Table A]

Amino acids	Synonymous codons
Ala	GCU, GCC, GCA, GCG,
Arg	AGA, AGG, CGU, CGC, CGA, CGG
Asn	AAU, AAC
Asp	GAU, GAC
Cys	UGU, UGC
Gln	CAA, CAG
Glu	GAA, GAG
Gly	GGU, GGC, GGA, GGG
His	CAU, CAC

Ile	AUU, AUC, AUA,
Leu	CUU, CUC, CUA, CUG, UUA, UUG
Lys	AAA, AAG
Phe	UUU, UUC
Pro	CCU, CCC, CCA, CCG
Ser	AGU, AGC, UCU, UCC, CUA, CUG
Thr	ACU, ACC, ACA, ACG
Tyr	UAU, UAC
Val	GUU, GUC, GUA, GUG

[0043]

Therefore, in the case of the above-mentioned example, there are 3×6×2×2=72 possible sequences which include codons encoding four amino acids. Of those, sequences including at least one original codon (specifically, ATT, CTG, CAC, and GAT) are excluded, and hence there are 2×5×1×1=10 sequences. For example, from a fragmented CDS including ATTCTGCACGAT, silently mutated CDSs such as ATCTTACATGAC and ATACTCCATGAC are produced.

[0044]

Meanwhile, in the case where one fragmented CDS includes ATTCTGCACGAT and the position of the base sequence on the 5'-end of the protein coding sequence is 0, if the position of the base on the 5'-end of the fragmented CDS in the protein coding sequence to be treated is not a multiple of three, the fragmented CDS is a sequence including three complete codons and two incomplete codons (because of lack of base information of the first and last codons). In this case, deficient bases in the two incomplete codons can be supplied with reference to genomic DNA information recorded in the recording unit 13.

[0045]

In this procedure, to decrease the homology to the sequence before introduction of the silent mutation as much as possible, it is desirable to exclude fragmented CDSs including codons before introduction of the mutation and to add the silent mutation as much as possible.

[0046]

The silently mutated CDSs thus produced are recorded in the recording unit 13 in relation to the original fragmented CDSs. That is, the {silently mutated CDS, and fragmented CDSs} are recorded as one set.

[0047]

Meanwhile, the silent mutation may be introduced into one or more codons which form the fragmented CDSs, but preferably, the silent mutation is introduced into all codons which form the fragmented CDSs. In each codon, the silent mutation may be introduced into only any base into which the silent mutation can be introduced, more preferably only the third base. However, from the point of view of improvement of primer specificity, it is preferred to introduce all the silent mutations into the fragmented CDSs.

[0048]

Meanwhile, for example, usage of all codons in genomic DNA of an organism where the tag is introduced may be calculated in advance, and based on the information, use of codons which are very less frequently used in the genomic DNA of the organism where the

tag is introduced may be avoided at the time of introduction of the silent mutations.

[0049]

In a step S5, functions of polynucleotides including the silently mutated CDSs per se produced in the step S4 or complementary base sequences thereof as primers are evaluated, and silently mutated CDSs which do not satisfy predetermined criteria are excluded. The exclusion can be performed by setting a flag added to each fragmented CDS to a value (e.g., "1") different from the initial value (e.g., "0"). In this procedure, the {silently mutated CDS, fragmented CDSs, and flag} are recorded as one set.

[0050]

Examples of evaluation criteria of functions of the nucleotides as primers include the following conditions (a) to (c):

(a) the nucleotide has a GC content of 40 to 60%, preferably 45 to 55%;

(b) the nucleotide has a tm value of 55 to 65°C; and/or

(c) the nucleotide does not have five or more, preferably four or more of consecutive identical bases.

If a silently mutated CDS per se or a complementary base sequence thereof does not satisfy one or more of the conditions, the silently mutated CDS is excluded.

[0051]

Here, the GC content is a ratio (%) of the number of GC bases in a base sequence. The tm value is a temperature (melting

temperature) at which 50% of double-stranded DNAs dissociate into single-stranded DNAs. Methods of calculating the GC content and tm value are known, and hence detail descriptions of the methods are omitted. It should be noted that the criterion values may be recorded in advance as initial data in the recording unit 13, but may be designated from outside via the IF unit. Further, the evaluation criteria of the primers may be known ones and are not limited to the above-mentioned conditions (a) to (c).

[0052]

The silently mutated CDSs having functions as primers in this step S5 are sometimes referred to as "fourth coding sequences".

[0053]

In a step S6, the silently mutated CDSs remaining after the step S5 (fourth coding sequence; that is, the flag is "0") is subjected to a homologous search for a genome sequence database previously prepared depending on purposes using a sequence alignment procedure. For example, if the resultant primers are used for bioremediation, the homologous search is performed for an environmental DNA database for bioremediation using a BLAST method. More specifically, searches are performed on a genome database provided by DDBJ, GENBANK, or the like, or a database storing fragmented base sequences of genes or the like using gene homology search software including Blast algorism or the like, and in the case where there is a region having homology equal to or higher than a threshold, the score of the homology is recorded (the score of the homology may be, for

example, a ratio of the number of matched bases to the length of the silently mutated CDS. For example, in the BLAST method, an index such as Identity corresponds to the score). Meanwhile, the number of regions having homology equal to or higher than the threshold is counted. It should be noted that the threshold of the homology may be set so that the ratio of the number of matched bases to the length of the query base is, for example, 80% or more, preferably 50% or more, more preferably 20% or more, still more preferably 10% or more when a gap is not taken into account. In addition, the homology search can be performed not only by the BLAST method but also by a known method such as a FASTA method.

[0054]

In a step S7, appearance frequency and Nearest Match Score (NMS) of each silently mutated CDS are determined using the results of the homology search performed in the step S6, that is, responses (response data) from the database to the genome sequence database for query. In this procedure, the homology search may be performed by a variety of known methods, and one of the methods may be used. Detail descriptions of the methods are omitted here. Hereinafter, a region having 50% or more homology is defined as a homologous region. Meanwhile, the genome sequence database used may be produced from a database provided by National Center of Biotechnology Information (NCBI) in USA depending on the purpose. In addition, a variety of genome sequence databases are known and may be used.

[0055]

The appearance frequency is the number of genomes hit (the homologous region has been detected) by the homology search. Therefore, the initial value of the appearance frequency of a specific silently mutated CDS is defined as "0", and if there is a genome hit by the homology search, the appearance frequency is increased by "1". However, even in the case where a plurality of regions are hit in one genome, the appearance frequency is increased by "1". The NMS is the maximum homology value when a plurality of regions are hit.

[0056]

More detail descriptions are shown below. FIG. 5 illustrates ATCCATCATGAC as one example of the silently mutated CDSs. It is assumed that two homologous regions (hits 1 and 2) to the silent mutated CDS are obtained by the homology search on a database. In FIG. 5, the hit 1 is a partial sequence of ATCTTAGATAAC in the first genomic DNA sequence No. 1, and the hit 2 is a partial sequence of ATCGTACATCTA in the second genomic DNA sequence No. 2 (different from the first genomic DNA sequence). In this case, the identity of the hit 1 is 83.3%, while the identity of the hit 2 is 66.7%. Therefore, the identity 83.3% of the hit 1 having the maximum identity is defined as an NMS of the silently mutated CDS. In addition, the appearance frequency of the silently mutated CDS is "2". It should be noted that, as mentioned below, in the case where the NMSs are the same for a plurality of different silent mutations, the appearance frequency is used as a screening index for determining a sequence

which does not appear in a natural environment.

[0057]

As a result, the NMS and appearance frequency for one silently mutated CDS are determined and are recorded in the recording unit 13 so that the information is in relation to each other. That is, the {silently mutated CDS, NMS, and appearance frequency} are recorded as one set.

[0058]

In a step S8, the minimum value of the NMSs recorded in the step S7 is determined, a silently mutated CDS corresponding to the value is determined as a tag, and from positional information corresponding to the sequence, an introduction site (site in the genomic DNA sequence) is determined. The tag and introduction information determined are displayed in the display unit 2, for example.

[0059]

If there is one minimum NMS, the silently mutated CDS corresponding to the value is determined as the tag. However, in the case where the NMSs are the same value for a plurality of different silent mutations, a step of determining a silently mutated CDS having the minimum appearance frequency is determined as the tag as a step S9. This is because an object of the present invention is to determine a sequence (tag, primer) which does not appear in a natural environment.

[0060]

According to the above-mentioned procedures, it is possible to determine the tag which is highly likely to be used universally even in environmental DNAs in which a large indefinite number of organisms coexist and the introduction site into the genomic DNA sequence. Therefore, a complementary base sequence of the tag can be used as a primer having high specificity in quantitative PCR.

[0061]

Is should be noted that the invention of the present application is not limited to the above-mentioned embodiments and may be modified by changing the order of the treatments, eliminating a part of the treatments, or substituting another treatment for a part of the treatments, or the like.

[0062]

For example, the above-mentioned description shows the case where the genomic DNA of one organism is treated, but in the case where the genomic DNAs of a plurality of organisms are treated, the treatments of the step S1 to S8 may be performed for each genomic DNA.

[0063]

Moreover, in the step S2, a region to be excluded may be designated from the outside.

[0064]

Further, the step S2 may include not only excluding a region which is considered to largely affect host cells (organism) by the silent mutation but also excluding another region. Specifically,

examples thereof include a region encoding a protein having unknown functions and a region encoding a protein having functions important for achieving the purpose (such as an enzyme capable of degrading petroleum in the case where the present invention is used for bioremediation). If the number of CDSs to be treated is large, the treatment requires longer time because the number of the fragmented CDSs and silently mutated CDSs becomes larger. Therefore, to reduce the treatment time, the number of the CDSs to be treated may be reduced as far as a certain level of accuracy can be achieved, and usually, about 50 to 1,000 CDSs are treated.

[0065]

Further, the above-mentioned descriptions show the case where, in the step S3, the fragmented CDSs are determined by shifting the position of the top of a coding sequence having a predetermined length one by one, but regions having continuous predetermined lengths may be selected at random so that the regions do not overlap with each other, to thereby obtain a plurality of fragmented CDSs.

[0066]

Moreover, the criteria for evaluation of functions as primers in the step S5 are not limited to the above-mentioned ones. For example, instead of the condition (c), or in addition to the conditions (a) to (c), a condition where the base on the end is G or C may be added. That is, a sequence where the base on the end is A or T may be excluded from an object to be treated.

[0067]

Further, in the homology search in the step S6, the degree of the homology may be designated from the outside instead of use of a predetermined degree of the homology (50% or more in the above-mentioned case).

[0068]

Further, although the above-mentioned descriptions show the case where the NMS and appearance frequency are used as criteria of the output in the step S7, the criteria are not limited thereto, and the GC content or the like may be considered. In general, primers to be used in PCR preferably have GC contents of about 55 to 65%. In addition, the primers more preferably include G and C in large amounts on the 3'-end side. Moreover, in PCR, a bond between the 3'-end on the primer side and the 5'-end on the DNA side is important, and hence candidates having a smaller number of matches in the regions corresponding thereto (the number of homologous bases detected by alignment) are more preferred. Depending on the purpose, final candidates may be determined on the basis of the criteria.

[0069]

(2) Method of manufacturing DNA tag-introduced mutant

A mutant into which the DNA tag determined by the above-mentioned method of determining the DNA tag and introduction site has been introduced may be obtained based on a known genetic engineering procedure, and for example, the mutant may be obtained as follows.

[0070]

A method of manufacturing a DNA-tag-introduced mutant including the following steps:

(i) preparing the DNA tag determined by the method of determining a DNA tag and an introduction site and cloning the DNA tag into a vector including a marker gene;

(ii) transforming a cell of an organism with the vector having the DNA tag cloned in the step (i); and

(iii) removing the marker gene by continuous culture or subculture of the cell transformed in the step (ii) and obtaining a mutant having the DNA tag introduced into a genome through homologous recombination.

[0071]

Hereinafter, the respective steps are described in detail. It should be noted that a method of manufacturing a mutant based on homologous recombination by single crossover is described here as an example.

[0072]

In the step (i), the DNA tag determined by the method of determining a DNA tag and an introduction site is prepared and cloned into a vector.

[0073]

To determine the base sequence of the DNA tag to be introduced into cells and the introduction site of the base sequence, first, the steps S1 to S8 are performed as mentioned above. To synthesize DNA fragments based on the resultant DNA tag sequence, a

conventionally known synthesizer may be used.

[0074]

To clone the DNA tag into the vector, (m) a DNA including a base sequence corresponding to the first coding sequence may be synthesized and integrated into the vector, and then a silent mutation may be introduced into a predetermined site by a point mutation introduction method. Alternatively, (n) a DNA including a base sequence corresponding to the fourth coding sequence (DNA tag) may be synthesized and integrated into the vector.

[0075]

The point mutation introduction method may be performed in accordance with a conventionally known method such as GeneTailor Site-Directed Mutagenesis System (Invitorogen) KOD-Plus-Mutagenesis Kit (TOYOBO).

[0076]

It should be noted that the vector is desirably constructed by using a bacterium other than a target bacterium into which the DNA tag is introduced. Moreover, the transformation of host cells is performed by homologous recombination, and hence upstream and downstream base sequences each including a region into which the DNA tag sequence is introduced are constructed on the vector. The length of the former sequence of the DNA tag is appropriately set depending on efficiency of the homologous recombination of the bacterial strain where the tag is introduced, and the lengths of the upstream and downstream sequences are preferably 50 bases or

more, more preferably 200 bases or more, still more preferably 500 bases or more. When the regions are constructed on the vector, the regions may be obtained by PCR using genomic DNA of target host cells as a template. In this procedure, the upstream/downstream regions of the first coding region are obtained so that the base sequence corresponding to the first coding sequence is included, and then introduction of the point mutation may be performed on the vector (corresponding to the item (m)). Further, a base sequence obtained by replacing the first coding sequence sites on the upstream/downstream regions including the base sequence corresponding to the first coding sequence with a base sequence corresponding to the fourth coding sequence may be synthesized by a known artificial gene construction method (such as overlap extension PCR) and constructed on the vector (corresponding to the item (n)).

[0077]

According to the above-mentioned method, a mutant including only the DNA tag introduced into the genome by homologous recombination in the step (iii) mentioned below can be obtained.

[0078]

The vector preferably includes a marker gene so that the gene can be expressed in cells into which the DNA tag is introduced, and examples of the marker gene include an antibiotic-resistant marker gene, a gene encoding a fluorescent protein, and a gene encoding an enzyme which catalyzes a coloring reaction.

[0079]

In the present invention, the antibiotic-resistant marker gene may be appropriately selected from conventionally known ones as long as the host cells have no antibody against the antibiotic, and the gene can be used as a marker. Specific examples of the antibiotic-resistant marker include ampicillin-resistant gene, streptomycin-resistant gene, tetracycline-resistant gene, erythromycin-resistant gene, puromycin-resistant gene, blasticidin S-resistant gene, hygromycin-resistant gene, kanamycin-resistant gene, gentamicin-resistant gene, chloramphenicol-resistant gene, and neomycin-resistant gene. Examples of the gene encoding a fluorescent protein include green fluorescent protein (GFP) gene, red fluorescent protein (RFP) gene, yellow fluorescent protein (YFP) gene, and luciferase gene. Examples of the gene encoding an enzyme which catalyzes a coloring reaction include β- glucuronidase (GUS) gene and lacZ gene.

[0080]

To integrate the marker gene into the vector so that the gene can be expressed, known promoter and terminator may be appropriately added depending on the type of the host cells. In the case where the promoter and terminator of the host cells are unknown, the sequences thereof may be determined before use in accordance with a conventionally known informatics procedure.

[0081]

The synthesized DNA fragments may include not only a region

encoding the sequence of the DNA tag but also a variety of regions. Examples of the region include: introduction of a transcription-terminating sequence; introduction of a known restriction enzyme-recognizing sequence for cleaving the site; and introduction of a methylase-recognizing sequence.

[0082]

The synthesized DNA fragments may be introduced into target cells after the fragments are inserted into an appropriate vector. The vector can be appropriately selected depending on the type of cells into which the vector is introduced. Examples of the vector include: plasmid DNAs including YCp-type *Escherichia coli*-yeast shuttle vectors such as pRS413, pRS414, pRS415, pRS416, YCp50, pAUR112, and pAUR123, YEp-type *Escherichia coli*-yeast shuttle vectors such as pYES2 and YEp13, YIp-type *Escherichia coli*-yeast shuttle vectors such as pRS403, pRS404, pRS405, pRS406, pAUR101, and pAUR135, plasmids derived from *Escherichia coli* (e.g., ColE-type plasmids such as pBR322, pBR325, pUC18, pUC19, pUC118, pUC119, pTV118N, pTV119N, pBluescript, pHSG298, pHSG396, and pTrc99A, p15A-type plasmids such as pACYC177 and pACYC184, and pSC101-type plasmids such as pMW118, pMW119, pMW218, and pMW219), plasmids derived from *Agrobacterium* (e.g., pBI101), and plasmids derived from *Bacillus subtilis* (e.g., pUB110 and pTP5); phage DNAs including λ phages (Charon4A, Charon21A, EMBL3, EMBL4, λ gt10, λ gt11, and λ ZAP), ΦX174, M13mp18, and M13mp19; retrotransposons including a Ty factor; and YAC vectors including pYACC2. In addition, there

can be also used animal virus vectors such as a retrovirus and a vaccinia virus, and insect virus vectors such as a baculovirus.

[0083]

It should be noted that the present invention may include a vector including the above-mentioned DNA tag cloned.

[0084]

In this step, the ori sequence in the vector to be introduced desirably does not act in an organism into which the DNA tag is introduced. In the case of using a shuttle vector to be used between a host organism in which the vector is constructed and an organism into which the DNA tag is introduced, the ori sequence for the organism into which the DNA tag is introduced is desirably removed in advance. The ori sequence is also called a replication starting point, and in the case where the vector has the sequence, replication is performed in the host cells into which the DNA tag is introduced. If the vector can be replicated in cells of a target organism in transformation of the organism by homologous recombination in the step (ii), selection using the marker gene may cause appearance of not only a mutant of interest but also a bacterial strain selected because the strain has the vector. In the step (ii), it is desirable to obtain a mutant which is transformed into the genome of the organism by homologous recombination and has no vector in the cells, and hence, it is desirable to use a vector including no ori sequence or to remove the ori sequence of the vector in advance. However, in the case where the ori sequence of a host organism (such as

Escherichia coli) in which the vector is constructed does not act as a replication-starting point in the organism into which the DNA tag is introduced, it is not necessary to remove the ori sequence.

[0085]

In the step (ii), transformation by homologous recombination is performed at a predetermined site on genomic DNA of cells using a vector including the DNA tag cloned, to thereby obtain cells which have been transformed into the genome by homologous recombination based on single crossover.

[0086]

As a transformation method, a conventionally known procedure may be employed. For example, in the case where the cells are plant cells, the above-mentioned vector may be introduced into the plant cells by a usual transformation method such as a vacuum wet method (agrobacterium method), a particle gun method, a PEG method, or an electroporation method. Tumor tissues, shoots, hairy roots, or the like obtained by such method may be used for cell culture, tissue culture, or organ culture without further treatment and can be regenerated into plant bodies by administration of appropriate concentrations of phytohormones (such as auxin, cytokinin, gibberellin, abscisic acid, ethylene, and brassinolide) using conventionally known methods for plant tissue culture. Meanwhile, in the case where the vector is introduced into a bacterium such as *Escherichia coli* or *Bacillus subtilis*, for example, a method involving using a calcium ion [Cohen, S.N. et al.:Proc. Natl. Acad.

Sci., USA, 69:2110 (1972)] or an electroporation method may be employed. Further, in the case where the vector is introduced into yeast, for example, an electroporation method [Becker, D.M. et al.:Methods. Enzymol., 194: 182(1990)], a spheroplast method [Hinnen, A. et al.:Proc. Natl. Acad. Sci., USA, 75: 1929(1978)], or a lithium acetate method [Itoh, H.:J. Bacteriol., 153:163(1983)] method may be employed. Further, in the case where the vector is introduced into animal cells, for example, an electroporation method, a calcium phosphate method, or a lipofection method may be employed. In the case where the vector is introduced into insect cells, for example, a calcium phosphate method, a lipofection method, or an electroporation method may be employed.

[0087]

As mentioned above, a combination including: a plasmid vector as the vector; an antibiotic-resistant gene as the marker gene; and a *Bacillus* bacterium, a *Rodococcus* bacterium, or a *Gordonia* bacterium as the host cells for introducing the DNA tag is exemplified. If the DNA tag is introduced in such combination, the DNA tag can be stably introduced into the genome.

[0088]

The host cells including the vector introduced are cultured depending on the type of the cells. In this case, cells including the vector introduced by a marker gene are selected. For example, in the case where transformation is performed using a vector including an antibiotic-resistant marker gene integrated, an antibiotic

suitable for a medium is added, and survival cells are selected.

[0089]

In the step (iii), the cells transformed in the step (ii) are appropriately cultured so that the generation number increases (continuous culture) or subcultured in a medium containing no medicament such as an antibiotic, to thereby obtain a mutant in which the DNA tag has been subjected to homologous recombination.

[0090]

Further, in the case where a gene encoding a fluorescent protein is used as the marker gene, a solution obtained by diluting a culture medium is plated in an appropriate agar medium or the like so that single colonies can be obtained, and UV irradiation may be performed to select colonies showing fluorescence. In the case of a gene encoding an enzyme which catalyzes a coloring reaction, as appropriate, a culture medium is plated in an appropriate agar medium or the like containing a substrate for the coloring reaction in the same way as above, and colonies can be selected by the coloring reaction. Even if another marker gene is used, a conventionally known detection method can be appropriately employed depending on the type of the marker gene used.

[0091]

The cells including the DNA tag introduced are appropriately cultured depending on the type of the cells. In this procedure, the culture is repeated 10 to 100 times, preferably 30 to 100 times. In this procedure, subculture may be performed. As mentioned above,

when the generation number increases for a long period of time to culture the cells, the marker gene and vector sequence may be removed by single crossover, to thereby obtain a mutant of interest where only the DNA tag has been subjected to homologous substitution. In this case, in the homologous regions added to the upstream and downstream of the DNA tag sequence, it is impossible to guess the region where the single crossover occurs (the former region or latter region). However, bacterial strains in which the single crossover occurs in opposite positions of the single crossover in the mutants obtained in the step (ii) are mutants of interest, and hence 50% of the mutants are desired strains. The desired strains can be easily obtained by screening using a procedure such as colony PCR using the DNA tags introduced as primers or by colony PCR using primers designed in the vector. In a final step, sequencing may be performed to confirm whether introduction of the DNA tag into regions to be treated is performed properly. The flow of the method is illustrated in FIG. 8.

[0092]

As a method of storing mutant cells, an appropriate method may be selected from conventionally known methods of storing cells in consideration of the type of the cells and preservation period. Examples of the method include refrigerated storage, frozen storage, freeze-dry storage, and slant storage. Meanwhile, in the case where cells including the DNA tag introduced are preserved for several tens of years, it is particularly preferred to use a microorganism

having an ability to form spores as host cells and store the cells in spore states.

[0093]

To confirm that the DNA tag is introduced properly by homologous substitution in a mutation introduction region of the mutant including the DNA tag introduced, obtained by the above-mentioned steps and whether a vector sequence does not remain in the genome, sequencing may be performed in accordance with a conventionally known method. The present invention includes the thus obtained mutant including the DNA tag introduced.

[0094]

The above-mentioned descriptions show the procedures using homologous recombination by the single crossover, but the method of manufacturing the mutant including the DNA tag introduced of the present invention is not limited thereto and includes a variety of procedures based on the procedure. For example, in the step (i), a DNA tag region is cloned using a vector including replication-competent ori in an organism into which the DNA tag is introduced, and in the step (ii), a mutant in which homologous recombination has been performed by double crossover is obtained. For example, a mutant of interest can be obtained by PCR using a primer including the DNA tag sequence and a primer designed on a further upstream in the upstream region of the DNA tag sequence cloned into the vector in the genomic DNA sequence of the host. In the step (iii), the bacterium is appropriately cultured in a

medium containing no antibiotic used as the marker, to thereby obtain a mutant in which the plasmid has been removed. As mentioned above, a variety of introduction methods can be performed with ingenuities in the respective steps, but the procedure of the homologous recombination by the single crossover is simple and accurate.

[0095]

(3) Monitoring method using DNA tag

The present invention provides a method of monitoring the cells into which the DNA tag has been introduced in the section (2) described above using the primer that recognizes the DNA tag designed in the section (1) described above. According to this method, it is possible to accurately grasp progress of cleanup and dispersion of microorganisms using the DNA tag as an indicator. For example, in the case where microorganisms including the DNA tag introduced are used for bioremediation, environmental genomic DNAs (whole DNA obtained from natural environment) are obtained from an environment where the microorganisms may be spread, and the microorganisms may be detected and quantified by quantitative PCR using a primer which recognize the DNA tag.

[0096]

More specifically, microorganisms including the DNA tag in their genome are spread to perform bioremediation, and the number of cells of the microorganism is appropriately quantified with time. Then, depending on the increase or decrease in the number of the cells, additional spread of the microorganisms may be appropriately

performed. Meanwhile, for example, at the time when soil improvement is completed, soil or water is collected in a region excluding the spread region and, if the microorganisms are not detected, it is possible to judge that the microorganisms are not dispersed to the outside.

[0097]

(4) Method of labeling organism

The present invention provides a method of labeling an organism, which includes introducing a base sequence obtained based on the above-mentioned method of determining the DNA tag and introduction site into a predetermined site in genomic DNA of the organism. The method of obtaining the DNA tag and the method of determining the introduction site and the method of introducing the DNA tag into the genomic DNA of the organism are as described above.

Examples

[0098]

Hereinafter, the present invention is described in more detail by way of test examples and the like. However, the present invention is not limited thereto.

[0099]

Reference Test Example 1

Using data of the whole genomic sequences of bacteria, archaea, and viruses registered in the NCBI, base sequences which are less likely to be used in the living world (1 to 12 bases) were obtained using a program P1 (hereinafter, may be abbreviated as "specific

sequences"). In a search of the base sequences which are less likely to be used, the maximum length of each sequence was set to 12 bases from the viewpoint of the computation time. Therefore, if the computation time is not considered, further longer sequences may be obtained.

[0100]

Trial (a): From data of genomic sequence of *Rhodococcus sp.* to be used for bioremediation, 15 to 30 bases (base sequences having lengths which may be used as primers for quantitative PCR) were obtained at random and aligned for environmental metagenome data currently registered in the NCBI, and the probability of the appearance of each of the sequences in the environmental metagenome was calculated (the trial was repeated ten thousand times).

[0101]

Trial (b): From the data of the genomic sequence of *Rhodococcus sp.*, 15 to 30 bases were obtained at random, and a data set was prepared by adding silent mutations at random to the third bases in codons of the base sequences. The sequences were aligned for environmental metagenome data currently registered in the NCBI, and the probability of the appearance of each of the sequences in the environmental metagenome was calculated (the trial was performed ten thousand times).

[0102]

Trial (c): From the data of the genomic sequence of *Rhodococcus sp.*, 15 to 30 bases were obtained at random, and a data set including

a 12-base specific sequence incorporated into a region was prepared by adding silent mutations to the third bases in codons of the base sequences to search the region that was able to be designed so as to contain the specific sequence. The sequences were aligned for environmental metagenome data currently registered in the NCBI, and the probability of the appearance of each of the sequences in the environmental metagenome was calculated (the trial was performed one thousand times).

[0103]

[Results]

The base lengths and appearance ratios in metagenomes in each of the trials (a) to (c) are illustrated in FIG. 6. In general, the maximum length of a primer is about 30 bases. FIG. 6 illustrates that the primer having a base sequence having a base length of 29 or 30 bases did not appear at all in the environmental metagenome in each of the trials (b) and (c) as the appearance ratio is zero for the primer. That is, the results have shown that, in the case where primers are designed using the method of the present invention, the probability of nonspecific binding of the primers to genome sequences in the environment can be drastically reduced in a range of primers having lengths usually designed as compared to the case where the primers are designed at random.

[0104]

Test Example I

1. Procedure and target

(1-1) Determination of type of target organism

As a target organism, *Bacillus subtilis* strain 168 (*Bacillus subtilis*) was selected. Some of *Bacillus* bacteria have an ability to degrade petroleum and are widely used in the field of bioremediation. Therefore, *Bacillus subtilis*, whose whole genome sequence had been determined and many of whose essential genes had been identified, was considered to be suitable as a model organism of this analysis.

[0105]

(1-2) Determination of target gene

Information of base sequences of a total of 4,106 protein coding genes of *Bacillus subtilis* was obtained from GenBank (http://www.ncbi.nlm.nih.gov/) of NCBI. Subsequently, 270 essential genes (Kobayashi et al. Proc. Natl. Acad. Sci. USA. 003 Apr 15; 100(8): 4678-83.) were removed from the genes.

[0106]

Meanwhile, the whole genome sequence information of the organism where a tag is to be introduced is not always obtained. If not, DNA sequences on the periphery of 16S rRNA gene may be appropriately determined using universal primers to introduce the tag. To verify that primers with sufficiently high specificity can be designed even in such case, only peripheral genes of rRNA gene were used in this analysis. Specifically, only genes located in positions within 10 kbp of the periphery of the rRNA gene (5 kbp from the 5'-end of rRNA gene toward the 5' direction, 5 kbp from the 3'-end toward the 3' direction) were selected as genes where

the tag was to be introduced in this analysis, and the other genes were eliminated. As a result, 93 genes (a total of about 70,000 bp) were selected as targets of this analysis.

[0107]

(1-3) Preparation of fragmented sequences

The 93 genes thus determined were each fragmented into a window size having 18 bases to prepare 67,953 fragmented sequences.

[0108]

(1-4) Preparation of negative control data

To examine the appropriateness of the present invention, a fragmented sequence including no DNA tag introduced by a silent mutation was used as a control. From the 67,953 fragmented sequences determined by the procedure (1-3), a total of 2,238 candidates, which were judged to have sufficient functions as primer sequences based on the following function evaluation criteria (a) to (c), were selected as controls (hereinafter, referred to as negative control data). It should be noted that all the following criteria (a) to (c) are usually used for design of primers.

(a) Having a GC content of 40% or more and 60% or less.

(b) Having a Tm value of 55°C or more and 65°C or less.

(c) Having terminal bases of G or C.

[0109]

(1-5) Preparation of DNA tag-introduced sequence

Each of the 67,953 fragmented sequences was subjected to silent mutations to prepare all patterns of sequences different from the

original sequences. Moreover, after the silent mutations, functions of the sequences as primer sequences were evaluated based on the above-mentioned function evaluation criteria (a) to (c), and candidates were screened. As a result, 812,864 tag-introduced sequences were prepared.

[0110]

(1-6) Calculation of Nearest Match Score (NMS) and appearance frequency of DNA tag-introduced sequence

The 812,864 DNA tag-introduced sequences were used as query sequences, and homology searches were performed for the whole genome database to calculate NMSs and appearance frequency. The whole genome database includes the genome base sequences of all prokaryotes, archaea, plasmids, and viruses the whole genome sequence of each of which has been determined. All the data of the genome base sequences were obtained from the FTP site of the NCBI (http://www.ncbi.nlm.nih.gov/) (Table 1). In the homology searches, NCBI Blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used. It should be noted that, in this analysis, regions each having less than 50% homology were not counted as homologous regions.

[0111]

[Table 1]

Contents of whole genome database

Type of genome	Target	Total file number	Total base number (bp)
Complete genome	Bacteria	714	2,269,972,014
	Archaea	55	113,127,648
	Plasmid	501	67,408,122
	Viruses	3,173	61,226,978

[0112]

2. Control Experiment

(2-1) To examine the appropriateness of the present invention, first, the following sequences (w) to (z) were prepared.

(w) A total of 234 all tag-introduced sequences each having an NMS of less than 50%.

The sequences correspond to about 0.025% from the bottom of NMS rank of all the 812,864 tag-introduced sequences prepared in the procedure (1-6).

(x) 404 sequences corresponding to about 0.05% from the bottom of NMS rank of all the 812,864 tag-introduced sequences prepared in the procedure (1-6).

(y) 409 tag-introduced sequences corresponding to about 0.05% from the top of NMS rank of all the 812,864 tag-introduced sequences prepared in the procedure (1-6).

(z) 2,238 tag-unintroduced sequences (negative control data created in the procedure (1-4)).

[0113]

(2-2) Subsequently, to virtually imitate an environment where a large indefinite number of DNA sequences were present as soil or the like used as a target of bioremediation, environmental metagenome sequences registered in the NCBI were obtained and provided as an environmental DNA database (Table 2).

[0114]

[Table 2]

Contents of environmental DNA database

Type of genome	Target	Total file number	Total base number (bp)
Environmental genome	coral	15	101,879,793
	Fish	8	49,317,878
	Fossil	2	53,822,347
	Freshwater	8	102,870,126
	Human gut	13	462,989,954
	Human lung	2	10,818,592
	Marine	15	4,948,763,548
	Microbial mat	10	84,253,870
	Mine drainage	3	70,335,918
	Mosquito	3	164,887,115
	Mouse gut	6	53,922,645
	Saltern	18	148,587,256
	Soil	1	612,702
	Stromatolite	6	154,900,267
	Termite gut	1	55,038,603
	Others	10	357,259,566

[0115]

A homologous search using Blast as an environmental DNA database was performed using the sequences (w) to (z) as query sequences, to thereby calculate NMS for each sequence. The distribution of the NMSs in the respective data sets (w) to (z) is illustrated by a boxplot (FIG. 7).

[0116]

Comparison of w, x, and y indicates that, as the NMSs and appearance frequency calculated based on the whole genome database become lower, the number of homologous regions that appear in the environmental DNA database becomes smaller. This indicates that DNA of an organism includes a sequence which is universally hardly appears or universally easily appears and means that the specificity of primers in the environmental DNA can be estimated to a certain degree on the basis of NMSs used as indices calculated from restricted

genome sequences (in this analyses, the whole genome database).

[0117]

Next, it is found that the sequences (w and x) having low NMSs and appearance frequency in the whole genome database have a small number of homologous regions in the environmental DNA database as compared to the original sequences (z) where no DNA tag was introduced. This result indicates the appropriateness of the present invention, that is, the fact that the specificity of the primers can be improved by inserting the silent mutations so that the NMSs are lowered.

[0118]

The dataset (z) shows similar values in the first quartiles, third quartiles, and medians. It should be noted that, in this analysis, regions each having less than 50% homology were not counted, and hence all the NMSs in sequences where NMS=0 was calculated were defined as 50. Parameters of w, x, y, and Z are 234, 404, 409, and 2,238, respectively.

[0119]

In general, primers each having about 80% sequence homology are known to cause nonspecific reactions. FIG. 7 reveals that all the sequences (z) where no DNA tag was inserted each have a region having 80% or more homology in the environmental DNA database. This means that there is no region in which a specific primer can be designed on the 93 genes selected from *Bacillus subtilis* unless the DNA tag technology is used. On the other hand, the figure reveals that 75% or more of the sequences (w or x) including the DNA tag

inserted so as to reduce NMSs have specificity as primers even in the environmental DNAs (calculated from the third quartiles in the following boxplot).

[0120]

Test Example II

To obtain a mutant which is not included in the "recombinant organisms" specified by the Cartagena Law and includes the DNA tag introduced, the following test was performed.

[0121]

(II-1) Determination of target organism and region where DNA tag is introduced, and preparation of vector for transformation

In bioremediation, a variety of microorganisms capable of performing environmental cleanup is used in open environments, and hence only organisms which are not included in the "recombinant organisms" specified by the Cartagena Law can be used industrially. For example, in the case where cleanup of soil polluted with petroleum is performed, *Bacillus* bacteria are generally used. Therefore, *Bacillus subtilis* strain 168 was used in the test.

[0122]

As a target gene into which the DNA tag was to be introduced, a gene which was considered to have little influence on proliferation of host cells was selected. The gene is a gene having a locus tag of BSU03680, a gene locus of 417561-419315, and unknown functions. Before a test of insertion of the DNA tag into the gene, a region of 418431-418547, which corresponds to an intermediate region of

the gene, was defined as a region into which the DNA tag was to be introduced, and a sequence including a random silent mutation introduced was defined as a pseudo DNA tag. An artificial gene having a full length of 520 bases, obtained by adding the upstream sequence of 200 bases of the sequence and the downstream sequence of 200 bases of the sequence, was synthesized by a known method of constructing an artificial gene (GenScript), and the gene was constructed on a plasmid vector pHASH203 using an EcoRI restriction enzyme site added on the upstream and downstream and was named pBS4106ID (FIG. 9). As a resistant gene, an erythromycin-resistant gene was used. Further, the plasmid was constructed using *Escherichia coli* DH5α. The results of sequencing are shown in FIGS. 10 and 11.

[0123]

Shown below are the sequence of the region into which the DNA tag is introduced (SEQ ID NO: 1), the DNA tag (SEQ ID NO: 2), and the sequence after introduction of the DNA tag (SEQ ID NO: 3). It should be noted that, in SEQ ID NO: 2, the shaded parts each show that a silent mutation has been introduced.

[0124]

[Table 3]

5' - CCT
GAAGACGCGGCGGCTGCGCCAAGAGTCGGGATAACCGGACTCGGTGTATCACAAGGAAAAGAGACCAGTGATGCAGTGAT
CGCTGGCAACCTCATTACCGGATTTTCAACTGGA
-3' (SEQ ID NO: 1)

[0125]

[Table 4]

CCCGAAGATGCGGCAGCAGCGCCTAGAGTCGGTATAACTGGACTGGGAGTGTCACAAGGCAAGGAAACCAGTGACGCAG

TGATTGCAGGCAACCTCATAACCGGCTTGTCTACAGGG

-3' (SEQ ID NO: 2)

[0126]

Shown below is the gene sequence including the DNA tag introduced. The DNA tag regions are surrounded by squares (including about 200 bases in each of the upstream and the downstream).

[0127]

[Table 5]

5' -

AAGGATATGGCGAAGGTGACGTTGATTATGAGGAACCGATTAATGTATCAATTCGCAATAACCACTTTGTTGGAAACGTT
TCAAGTTCTGTGACCAATTTTAACGGGTATGGCATTTTAATAGAAGGAAATCACTCAGACAATACAATCAGCTACGGATA

TGGGACGCAAACAGTGATCAAGGGCAATATTCTGAGACGCCCCGAAGATGCGGCAGCAGCGCCTAGAGTCGGTATAACT

GGACTGGGAGTGTCACAAGGCAAGGAAACCAGTGACGCAGTGATTGCAGGCAACCTCATAACCGGCTTGTCTACAGGGA

TTGATGTCAGGGGAAAGAGCGTTCTTGTGACGAACAACAAAATCAGCAACTTTGAAAACACAGGGATATTGGTTTATCAG
TCCTCCGACGTAAAGGTAGACGGAAACCAAATTCAAAACGGACTGTCTGAAACAAGGCGCAGCATCGGTCTTCGCGCAGT
GCTGTCAGATGACATCGCATTTCTGAATAACTGTCTCATTCA

-3' (SEQ ID NO: 3)

[0128]

(II-2) Transformation of *B. subtilis* strain 168

B. subtilis strain 168 was transformed with pBS4106ID. First, as preculture, *B. subtilis* strain 168 was inoculated into any LB agar medium (Tripton 10g/L, Yeast Extract 5g/L, NaCl 10g/L, 1.5% agarose) using a platinum needle and cultured at room temperature. Next, a single colony was inoculated into 5 mL of a CI medium (1×MM

medium 5 mL, 50% glucose 50 μL, 1M MgSO4 25 μL, L-leucine 5mg/mL 50 μL, L-tryptophan 5 mg/mL 50 μL, 5% Yeast Extract 50 μL) so that OD660 was 0.1 and cultured with shaking at 37°C. When OD660 reached 1.5, 500 μL of the culture solution was separated in a centrifugal tube and centrifuged at 15,000 rpm for 2 minutes, and the supernatant was removed, to thereby obtain a bacterial cell pellet.

[0129]

The bacterial cell pellet was suspended in 1 mL of a CII medium (1×MM medium 5 mL, 50% glucose 50 μL, 1M MgSO4 25 μL, L-leucine 5mg/mL 5 μL, L-tryptophan 5 mg/mL 5 μL, 5% Yeast Extract 25 μL) using Vortex, and 100 μL of the suspension was dispensed in a small test tube. Thereafter, 100 ng of a vector was added, and the bacteria were cultured with shaking at 37°C for 90 minutes. After culture, 300 μL of an LB medium was added, and the bacteria were further cultured for 60 minutes, plated in an LB agar medium containing 5 μg/ml erythromycin, and cultured at 37°C overnight. In all resultant colonies, a sequence of pBS4106ID was introduced into the genome of the host by homologous recombination (single crossover). The transformant of *B. subtilis* strain 168 thus obtained was named strain BS4106A.

[0130]

(III-3) Acquirement of mutant of interest

To remove extra sequences derived from the vector pBS4106ID by exchanging only the DNA tag regions for gene regions in the genome of the host, strain BS4106A was inoculated into 5 mL of an LB medium

containing no reagent (erythromycin) and cultured with shaking at 37°C for 36 hours (if the bacteria was divided once per 30 minutes, about 70th generation of the bacteria was obtained). The culture solution was diluted 1000-fold with an LB medium and plated in an LB agar medium. The resultant colonies were patched on each of an LB agar medium containing 5 μg/ml erythromycin and an LB agar medium containing no reagent using a platinum needle, to thereby obtain an erythromycin-sensitive strain. For 200 colonies of the bacterial strain thus obtained, primers (SEQ ID NOS: 4 and 5) were designed from an inner sequence of pBS4106ID, and colony PCR was performed. A bacterial strain which was not amplified by the colony PCR were subjected to sequencing, to thereby obtain a bacterial strain including the DNA tag introduced properly. The bacterial strain thus obtained was named BS4106ID. It should be noted that, in strain BS4106ID, only the DNA tag regions were exchanged for the gene regions in the genome of the host, and hence the strain was not included in the "recombinant organisms" specified by the Cartagena Law.

[0131]

Primer sequences used in the colony PCR are as follows.

ERM-F (SEQ ID NO: 4)

5'-CGTAGAGCACACGGTTTAACG-3'

TET-R2 (SEQ ID NO: 5)

5'-GCCATAGTGACTGGCGATGC-3'

[0132]

Primer sequences used in the sequencing are as follows.

bs4106id_F (SEQ ID NO: 6)

5'-AGGATATGGCGAAGGTGACG-3'

bs4106id_R (SEQ ID NO: 7)

5'-GTCATCTGACAGCACTGCGC-3'

Reference Signs List

[0133]

1 computer

2 display unit

3 operation unit

4 network

5 database (DB)

11 computer processing unit (CPU)

12 rewritable memory (RAM)

13 recording unit

14 interface unit (IF unit)

15 internal bus

Sequence Listing Free Text

[0134]

SEQ ID NO: 2 represents a DNA tag to be introduced.

SEQ ID NO: 3 represents a gene locus 417561-419315 of *Bacillus subtilis* strain 168 including the DNA tag introduced.

SEQ ID NO: 4 represents an ERM-F primer.

SEQ ID NO: 5 represents a TET-R2 primer.

SEQ ID NO: 6 represents a bs4106id_F primer.

SEQ ID NO: 7 represents a bs4106id_R primer.

The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows:

1. A method of manufacturing a DNA-tag-introduced mutant, the method comprising the following steps:

(i) preparing a DNA tag by the following steps:

a step S1 of obtaining a protein coding sequence from a genomic DNA sequence of a microorganism;

a step S2 of determining a first coding sequence as a region to be treated from the protein coding sequence;

a step S3 of obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and recording sites of the second coding sequences in the protein coding sequence;

a step S4 of obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the step S3 to a silent mutation;

a step S5 of determining whether or not the third coding sequences satisfy a predetermined condition and determining only the third coding sequences that satisfy the predetermined condition as fourth coding sequences;

a step S6 of performing a homology search for the fourth coding sequences and obtaining homologous base sequences;

a step S7 of determining a Nearest Match Score (NMS) for each of the fourth coding sequences; and

a step S8 of determining the fourth coding sequence corresponding to a minimum value of the NMS as the DNA tag and determining the site of the second

coding sequence corresponding to the fourth coding sequence as the introduction site, wherein:

in step S2, the first coding sequence comprises a coding sequence in a region having no influence on a biological function of the organism even if the sequence is subjected to the silent mutation;

the predetermined condition comprises a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates a degree of homology;

(ii) cloning the DNA tag into a vector including a marker gene;

(iii) transforming a cell of a microorganism with the vector having the DNA tag cloned in the step (ii); and

(iv) removing the marker gene by continuous culture or subculture of the cell transformed in the step (iii) and obtaining a mutant having the DNA tag introduced into a genome through homologous recombination.

2. The method according to claim 1, further comprising the step of confirming that the marker gene is absent in the genome of the microorganism.

3. A method of labeling a microorganism, the method comprising:

(i) preparing the DNA tag by the following steps:

a step S1 of obtaining a protein coding sequence from a genomic DNA sequence of a microorganism;

a step S2 of determining a first coding sequence as a region to be treated from the protein coding sequence;

a step S3 of obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and recording sites of the second coding sequences in the protein coding sequence;

a step S4 of obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the step S3 to a silent mutation;

a step S5 of determining whether or not the third coding sequences satisfy a predetermined condition and determining only the third coding sequences that satisfy the predetermined condition as fourth coding sequences;

a step S6 of performing a homology search for the fourth coding sequences and obtaining homologous base sequences;

a step S7 of determining a Nearest Match Score (NMS) for each of the fourth coding sequences; and

a step S8 of determining the fourth coding sequence corresponding to a minimum value of the NMS as the DNA tag and determining the site of the second coding sequence corresponding to the fourth coding sequence as the introduction site, wherein:

in step S2, the first coding sequence comprises a coding sequence in a region having no influence on a biological function of the organism even if the sequence is subjected to the silent mutation;

the predetermined condition comprises a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates a degree of homology; and

(ii) introducing the DNA tag into the introduction site of the genomic DNA of the microorganism using a vector having the DNA tag cloned thereinto.

4. A method of manufacturing a vector, the method comprising:

(i) preparing the DNA tag by the following steps:

a step S1 of obtaining a protein coding sequence from a genomic DNA sequence of a microorganism;

a step S2 of determining a first coding sequence as a region to be treated from the protein coding sequence;

a step S3 of obtaining a plurality of second coding sequences each including a partial sequence in the first coding sequence and having a predetermined length and recording sites of the second coding sequences in the protein coding sequence;

a step S4 of obtaining one or more third coding sequences produced by subjecting the corresponding second coding sequences obtained in the step S3 to a silent mutation;

a step S5 of determining whether or not the third coding sequences satisfy a predetermined condition and determining only the third coding sequences that satisfy the predetermined condition as fourth coding sequences;

a step S6 of performing a homology search for the fourth coding sequences and obtaining homologous base sequences;

a step S7 of determining a Nearest Match Score (NMS) for each of the fourth coding sequences; and

a step S8 of determining the fourth coding sequence corresponding to a minimum value of the NMS

as the DNA tag and determining the site of the second coding sequence corresponding to the fourth coding sequence as the introduction site, wherein:

in step S2, the first coding sequence comprises a coding sequence in a region having no influence on a biological function of the organism even if the sequence is subjected to the silent mutation;

the predetermined condition comprises a condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer; and

the NMS indicates a degree of homology; and

(ii) cloning the DNA tag into a vector including a marker gene.

5. The method according to any one of claims 1 to 4, wherein the steps of preparing the DNA tag further comprise:

determining an appearance frequency for each of the fourth coding sequences in step S7, the appearance frequency indicating the number of genomes that a region homologous to the fourth coding sequence is detected by the homology search performed in step S6, and

when a plurality of fourth coding sequences corresponding to the NMS with the minimum value exists in step S8,

a step S9 of determining the fourth coding sequence corresponding to the appearance frequency with the minimum value among the appearance frequency of the plurality of fourth coding sequences corresponding to the NMS of the minimum value as the DNA tag, and determining as the introduction site, the site of the second coding sequence

corresponding to the fourth coding sequence determined as the DNA tag.

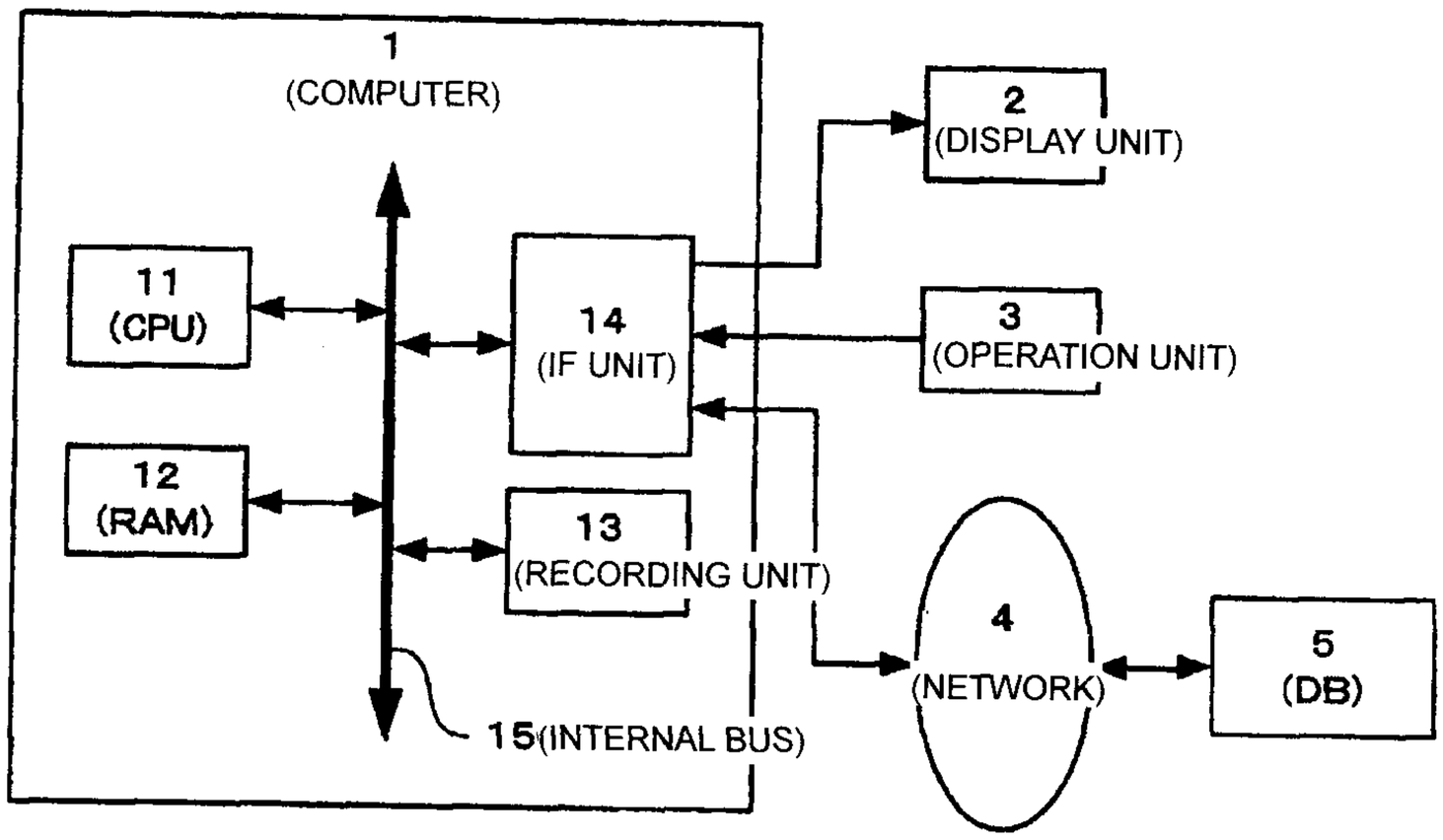
6. The method according to any one of claims 1 to 4, wherein the condition where a polynucleotide including the third coding sequence or a complementary base sequence thereof is suitable as a primer comprises the following:

the complementary base sequence has a CG content of 45 to 55%;

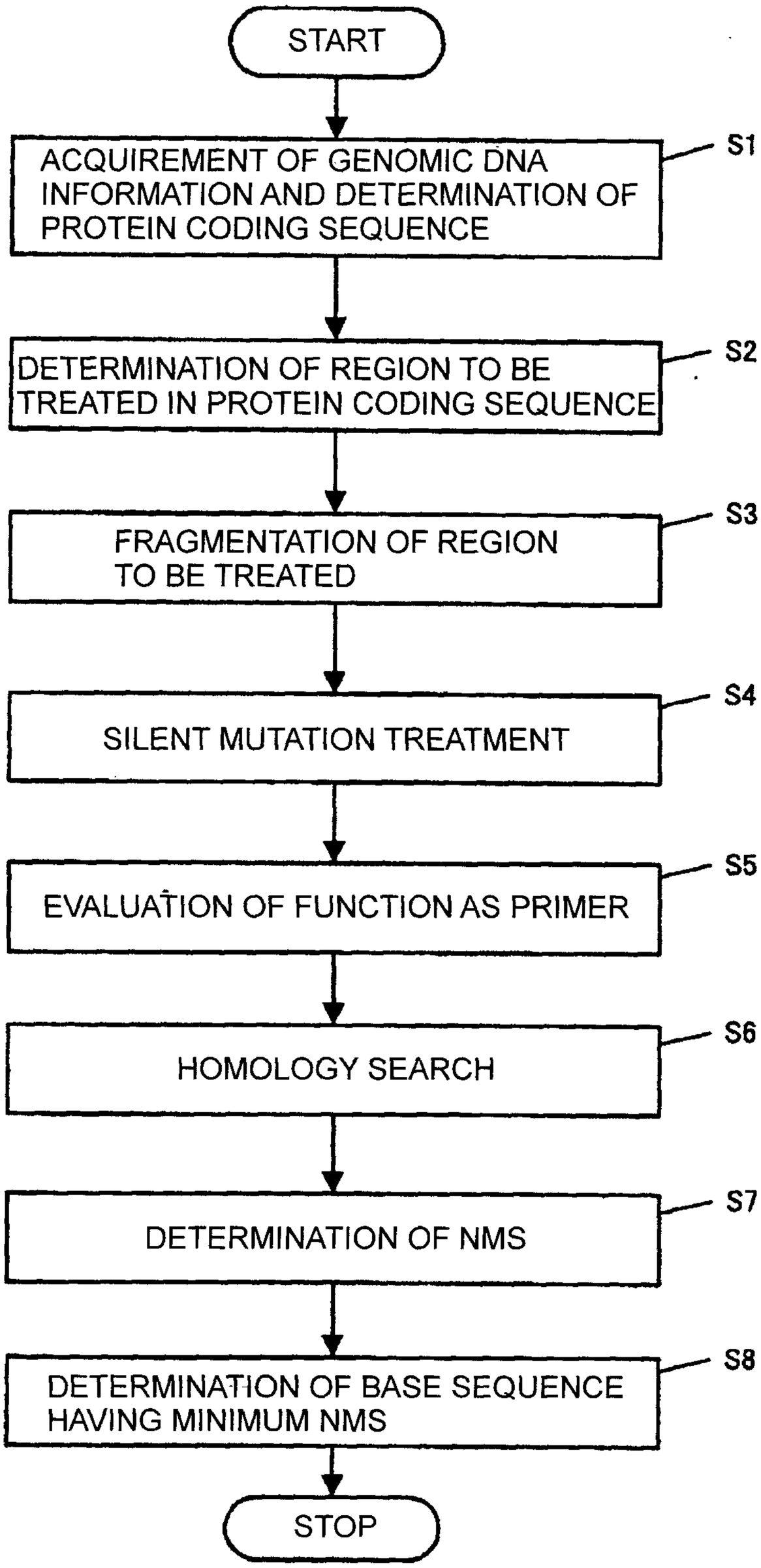
the complementary base sequence has a T_m value of 55 to 65° C; and

the complementary base sequence is free of four or more consecutive identical bases.

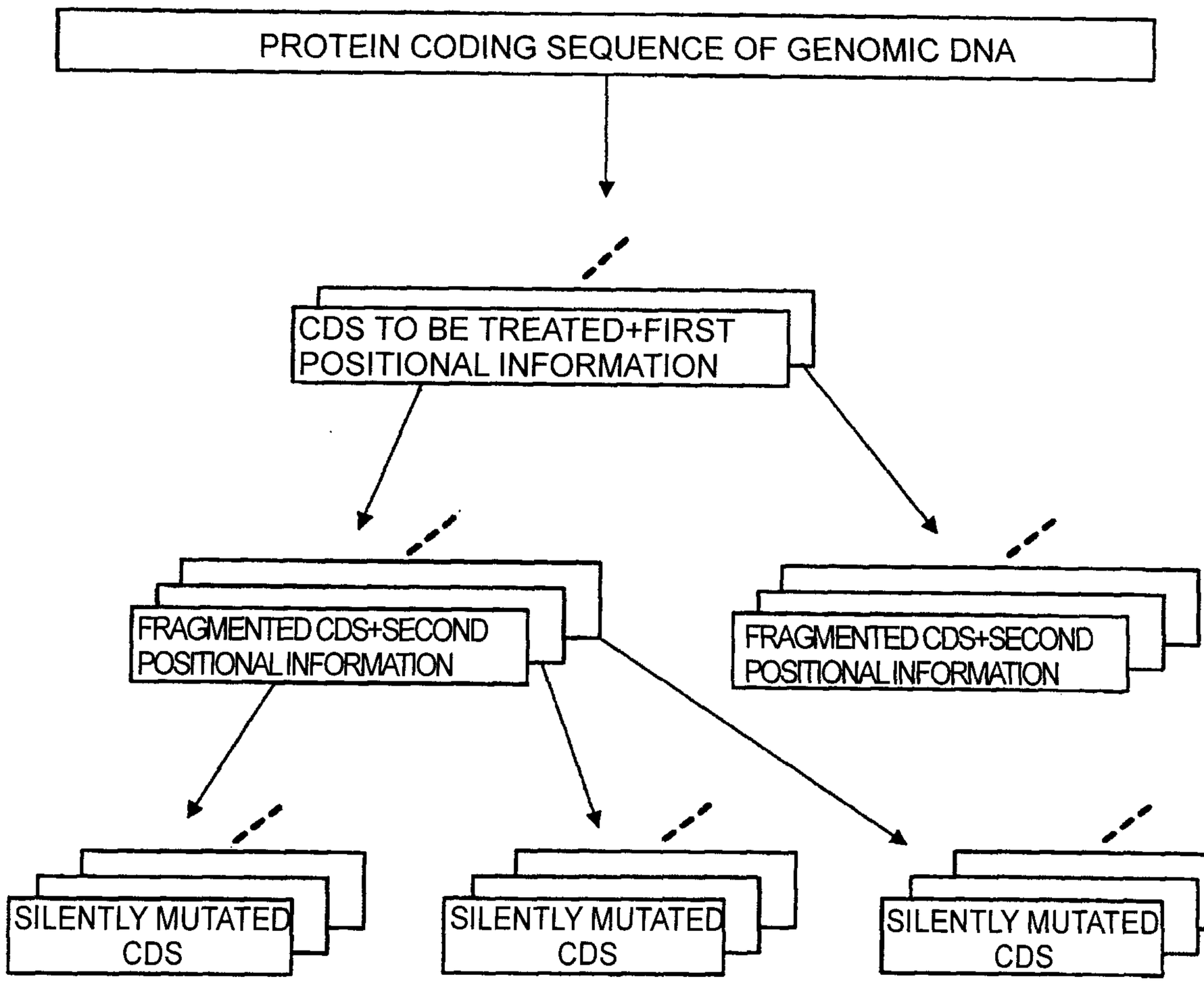
[FIG. 1]


1
(COMPUTER)
11
(CPU)
12
(RAM)
14
(IF UNIT)
13
(RECORDING UNIT)
15(INTERNAL BUS)
2
(DISPLAY UNIT)
3
(OPERATION UNIT)
4
(NETWORK)
5
(DB)

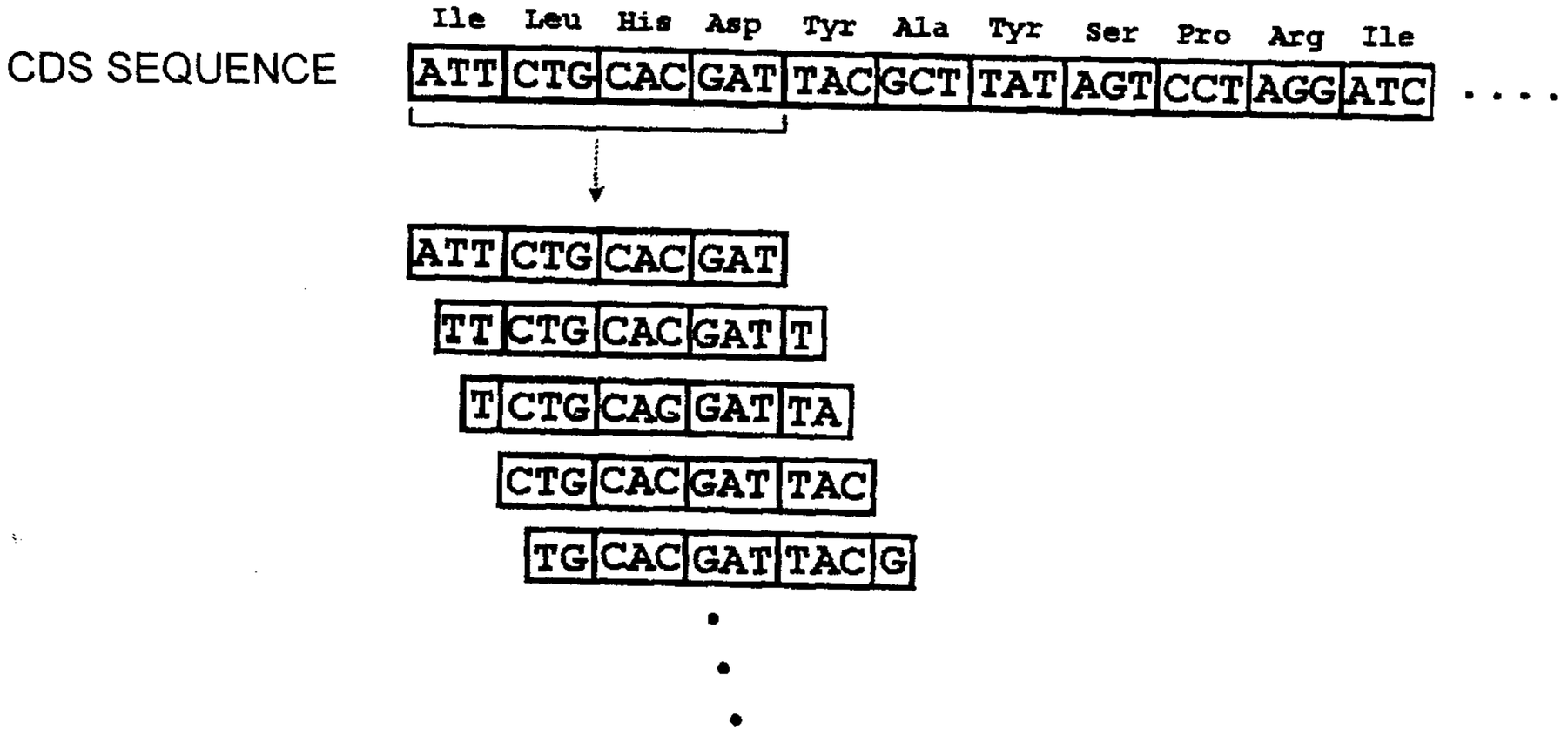
[FIG. 2]


START
ACQUIREMENT OF GENOMIC DNA INFORMATION AND DETERMINATION OF PROTEIN CODING SEQUENCE
S1
DETERMINATION OF REGION TO BE TREATED IN PROTEIN CODING SEQUENCE
S2
FRAGMENTATION OF REGION TO BE TREATED
S3
SILENT MUTATION TREATMENT
S4
EVALUATION OF FUNCTION AS PRIMER
S5
HOMOLOGY SEARCH
S6
DETERMINATION OF NMS
S7
DETERMINATION OF BASE SEQUENCE HAVING MINIMUM NMS
S8
STOP

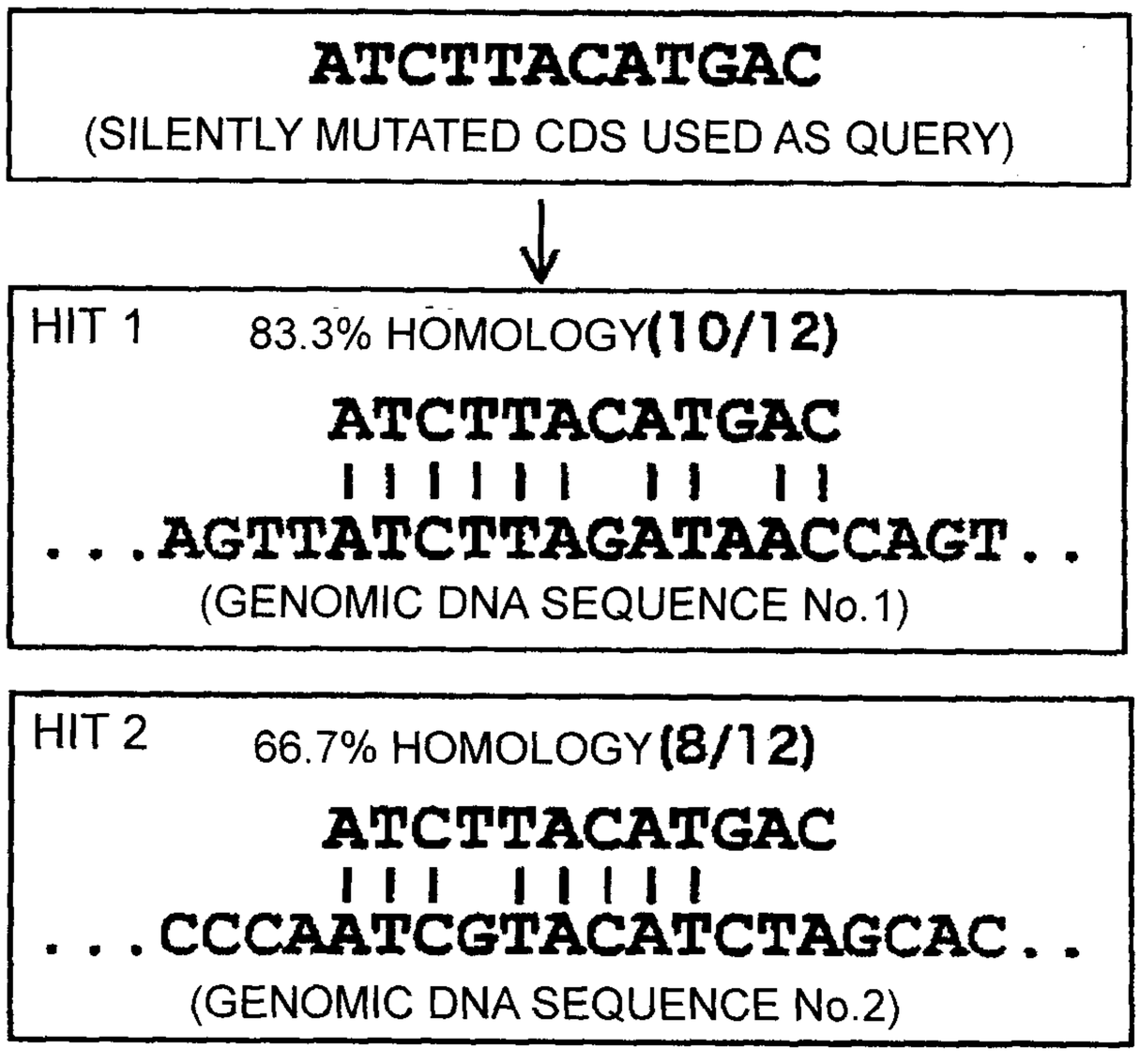
[FIG. 3]


PROTEIN CODING SEQUENCE OF GENOMIC DNA
CDS TO BE TREATED+FIRST POSITIONAL INFORMATION
FRAGMENTED CDS+SECOND POSITIONAL INFORMATION
FRAGMENTED CDS+SECOND POSITIONAL INFORMATION
SILENTLY MUTATED CDS
SILENTLY MUTATED CDS
SILENTLY MUTATED CDS

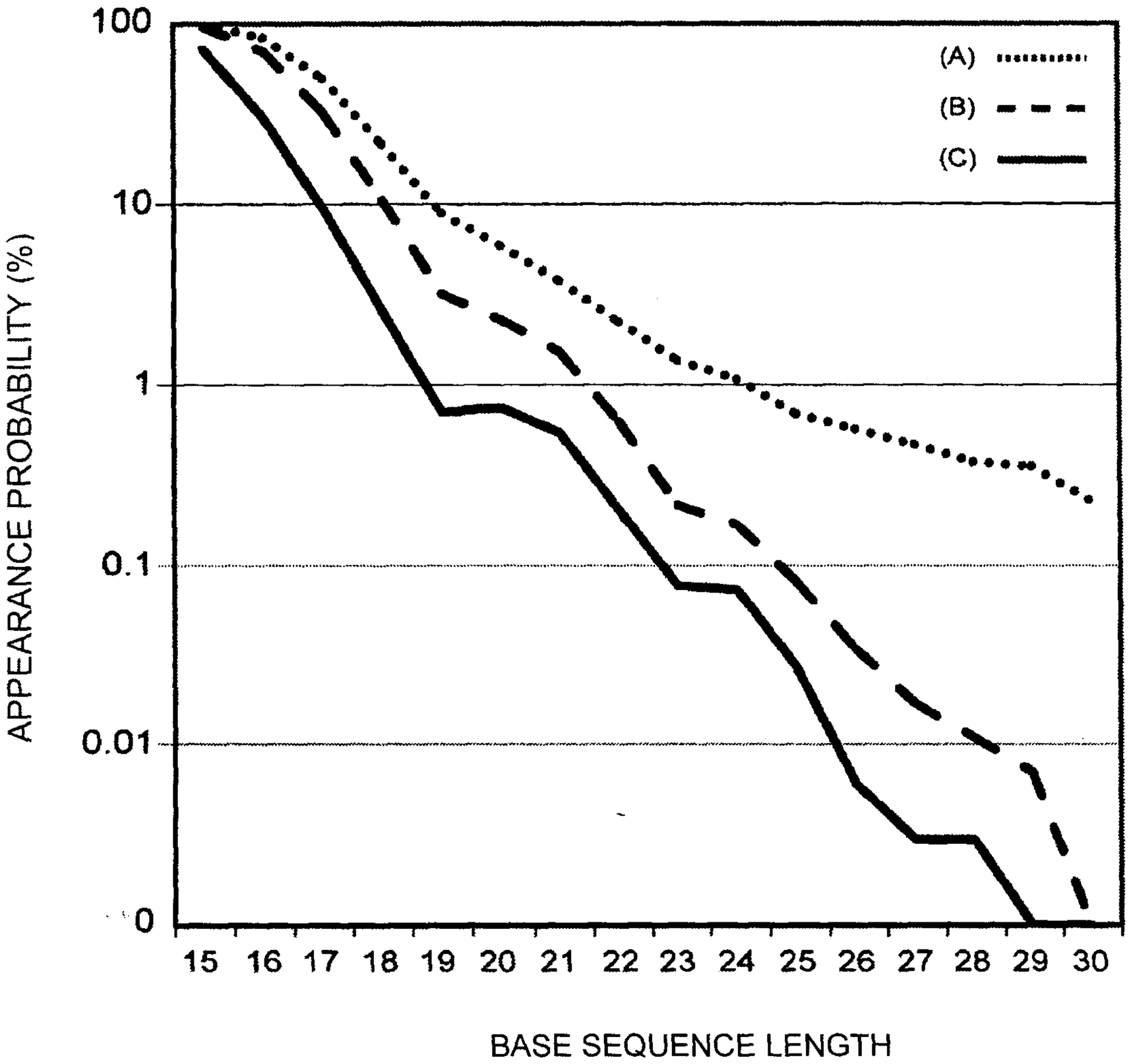
[FIG. 4]


CDS SEQUENCE
Ile Leu His Asp Tyr Ala Tyr Ser Pro Arg Ile
ATT CTG CAC GAT TAC GCT TAT AGT CCT AGG ATC ....
ATT CTG CAC GAT
TT CTG CAC GAT T
T CTG CAC GAT TA
CTG CAC GAT TAC
TG CAC GAT TAC G

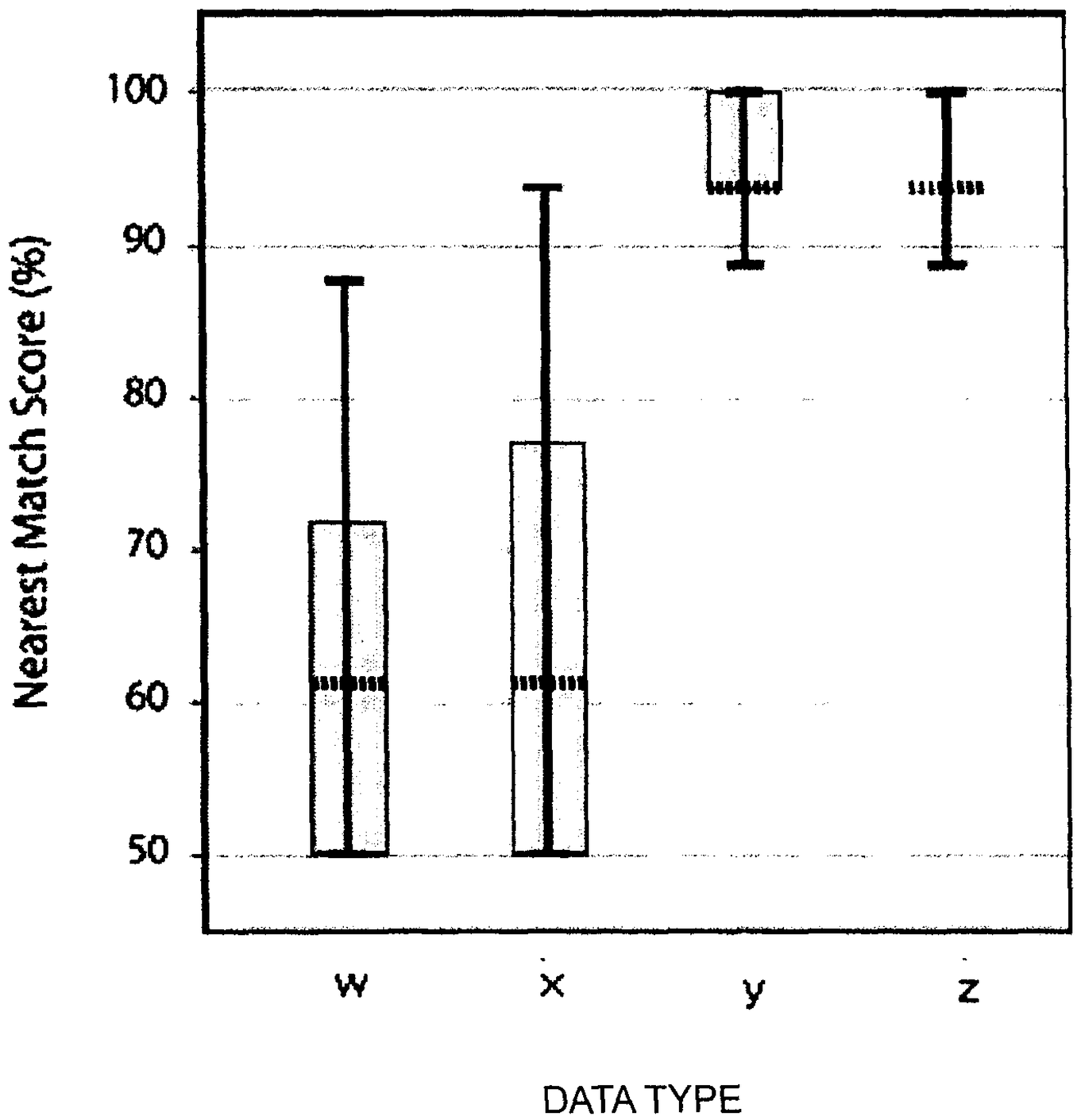
[FIG. 5]


ATCTTACATGAC
(SILENTLY MUTATED CDS USED AS QUERY)
HIT 1
83.3% HOMOLOGY(10/12)
ATCTTACATGAC
...AGTTATCTTAGATAACCAGT..
(GENOMIC DNA SEQUENCE No.1)
HIT 2
66.7% HOMOLOGY(8/12)
ATCTTACATGAC
...CCCAATCGTACATCTAGCAC..
(GENOMIC DNA SEQUENCE No.2)

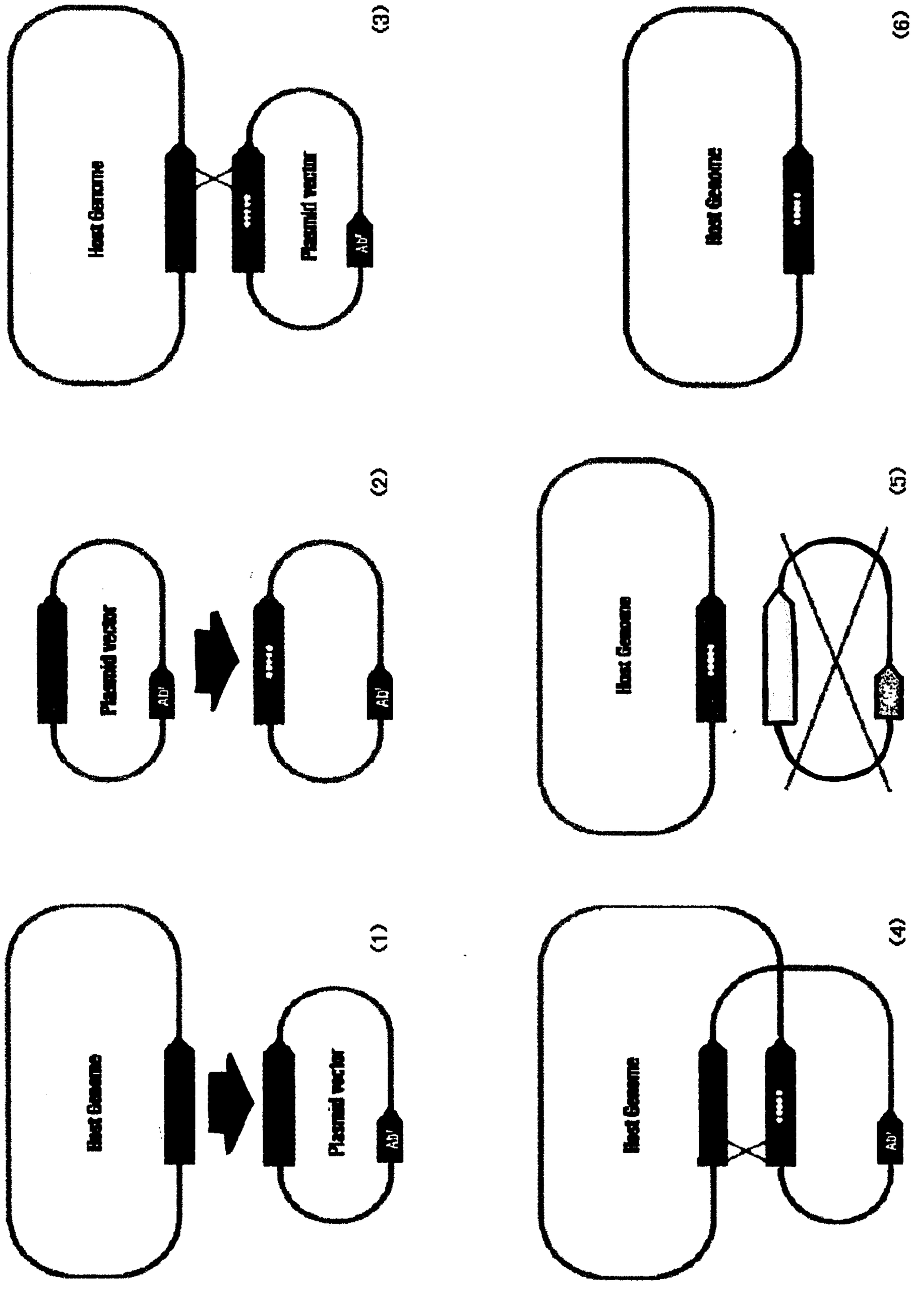
[FIG. 6]


100
10
1
0.1
0.01
0
APPEARANCE PROBABILITY (%)
(A)
(B)
(C)
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
BASE SEQUENCE LENGTH

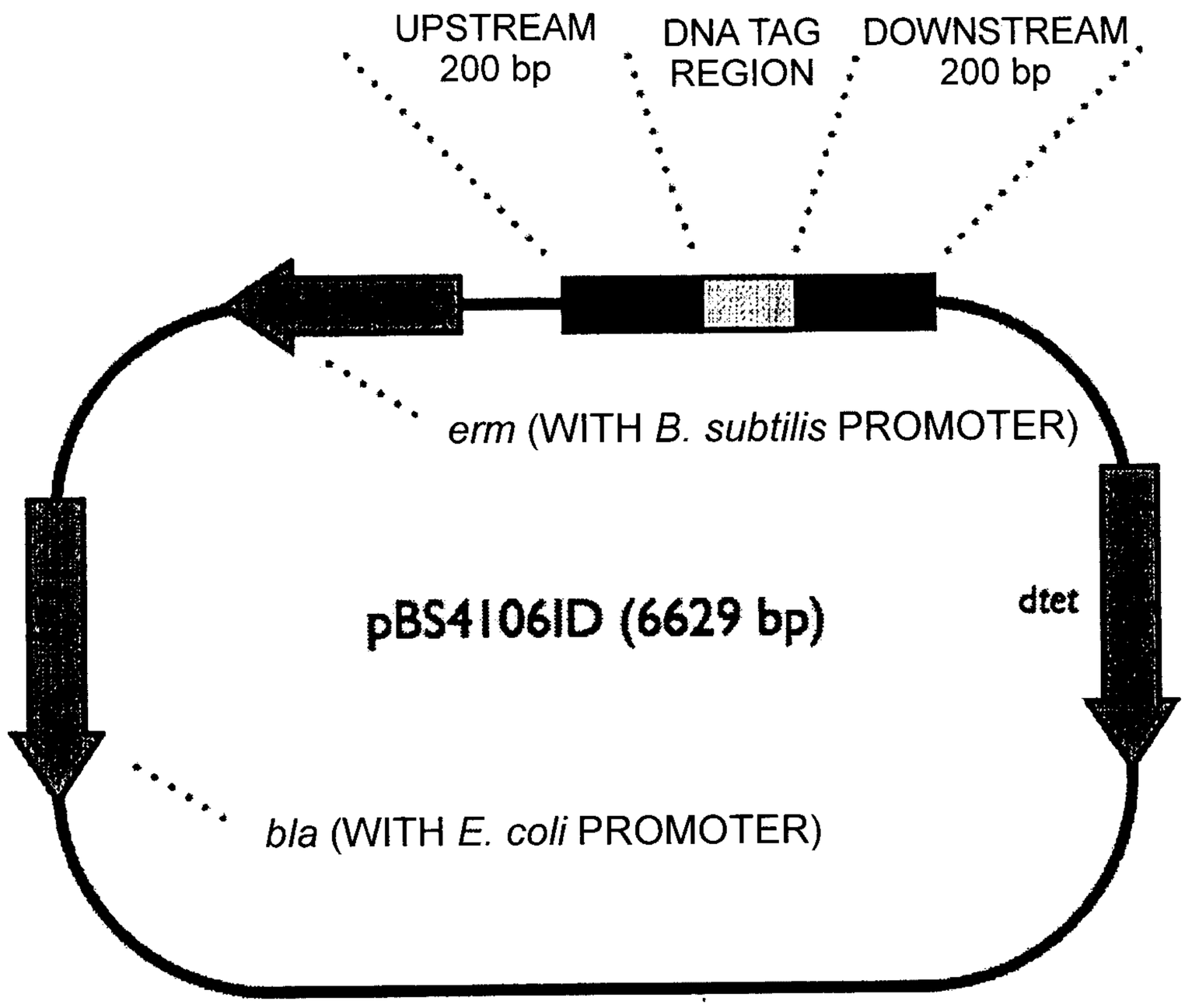
[FIG. 7]



[FIG. 8]


(1)
Host Genome
Plasmid vector
Ab'
(2)
Plasmid vector
Ab'
(3)
Host Genome
Plasmid vector
Ab'
(4)
Host Genome
Ab'
(5)
Host Genome
(6)
Host Genome

[FIG. 9]



erm : ERYTHROMYCIN-RESISTANT GENE

bla : AMPICILLIN-RESISTANT GENE:

d*tet* : TETRACYCLINE-RESISTANT GENE FRAGMENT SEQUENCE

[FIG. 10]

```
                                                
```

[FIG. 11]

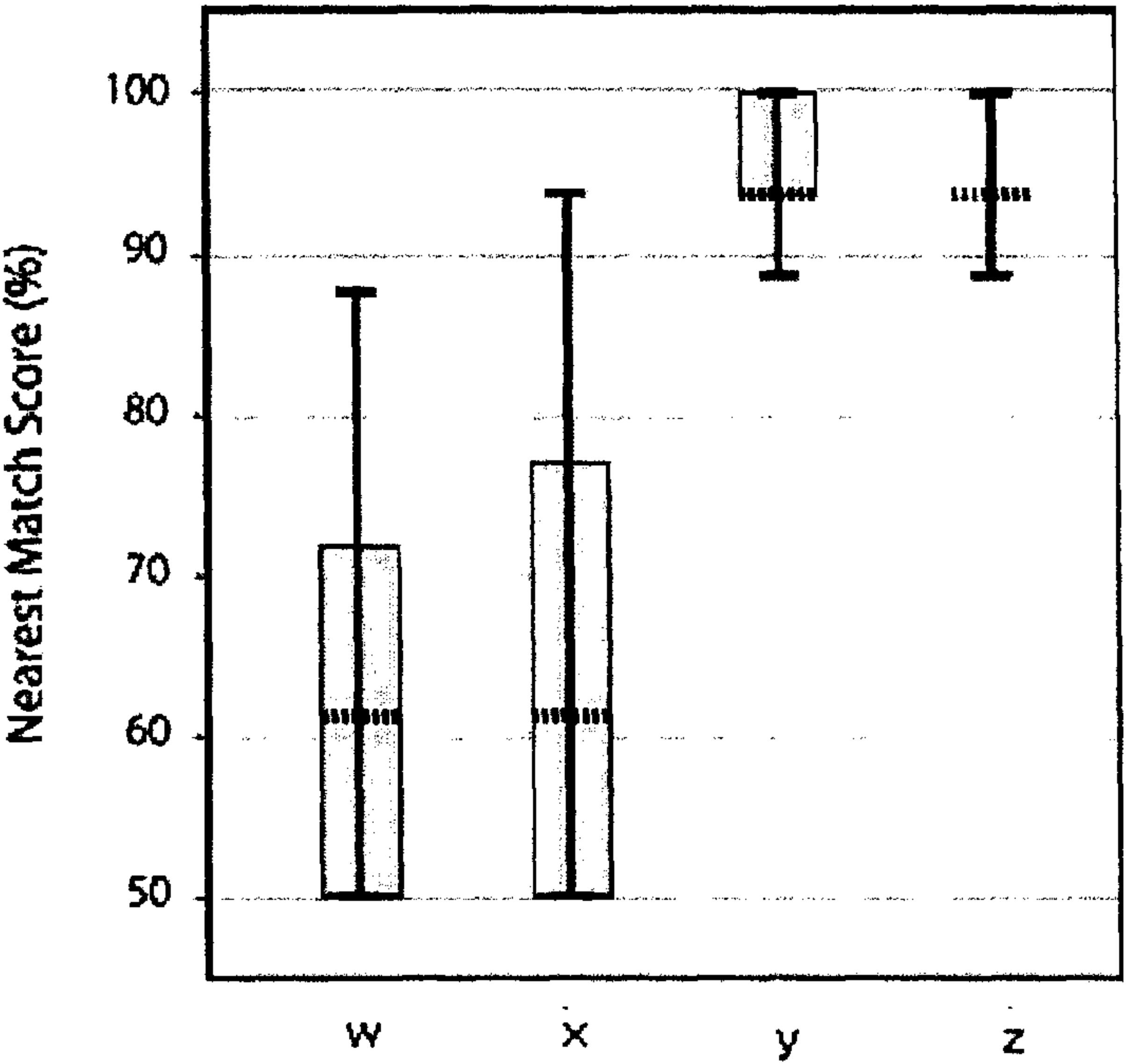
```
                                                370       380       390       400       410       420
                                                |         |         |         |         |         |
                                                GACGAACAACAAAATCAGCAACTTTGAAAACACAGGGATATTGGTTTATCAGTCCTCCGAC
C04-J23692-59737-1-M13R(-48).ab1(1>630)    ←    GACGAACAACAAAATCAGCAACTTTGAAAACACAGGGATATTGGTTTATCAGTCCTCCGAC
59737-1.seq(1>521)                         →    GACGAACAACAAAATCAGCAACTTTGAAAACACAGGGATATTGGTTTATCAGTCCTCCGAC

                                                430       440       450       460       470       480
                                                |         |         |         |         |         |
                                                GTAAAGGTAGACGGAAACCAAATTCAAAACGGACTGTCTGAAACAAGGCGCAGCATCGGTC
C04-J23692-59737-1-M13R(-48).ab1(1>630)    ←    GTAAAGGTAGACGGAAACCAAATTCAAAACGGACTGTCTGAAACAAGGCGCAGCATCGGTC
59737-1.seq(1>521)                         →    GTAAAGGTAGACGGAAACCAAATTCAAAACGGACTGTCTGAAACAAGGCGCAGCATCGGTC

                                                490       500       510       520       530       540
                                                |         |         |         |         |         |
                                                TTCGCGCAGTGCTGTCAGATGACATCGCATTTCTGAATAACTGTCTCATTCAGAATTCGAG
C04-J23692-59737-1-M13R(-48).ab1(1>630)    ←    TTCGCGCAGTGCTGTCAGATGACATCGCATTTCTGAATAACTGTCTCATTCAGAATTCGAG
59737-1.seq(1>521)                         →    TTCGCGCAGTGCTGTCAGATGACATCGCATTTCTGAATAACTGTCTCATTCA

                                                550       560       570       580       590       600       610
                                                |         |         |         |         |         |         |
                                                CTCGGTACCTCGCGAATGCATCTAGATATCGGATCCCGGGCCCGTCGACTGCAGAGGCCTG
C04-J23692-59737-1-M13R(-48).ab1(1>630)    ←    CTCGGTACCTCGCGAATGCATCTAGATATCGGATCCCGGGCCCGTCGACTGCAGAGGCCTG

                                                620       630
                                                |         |
                                                CATGCAAGCTTGGCgtAATC
C04-J23692-59737-1-M13R(-48).ab1(1>630)    ←    CATGCAAGCTTGGCgtAATC
```


Nearest Match Score (%)
100
90
80
70
60
50
w
x
y
z
DATA TYPE