Bioinformatics Tool for Genomic Era: A Step towards the In Silico Experiments - Focused on Molecular Cloning

Akira Ohyama(1,2)*, Ken Kurokawa(2), Kyotetsu Enai(1), Hitohiro Saitoh(1), Shigehiko Kanaya(2), Md.Altf-Ul-Amin(2), Naotake Ogasawara(2)

(1) In silico biology, inc., SOHO Station 706, Yamashita-cho 24-8, Naka-ku, Yokohama 231-0023, (2) Graduate School of Information Science, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma, Nara 630-0101

(Received July 10, 2006; Accepted August 1, 2006)

We are making an attempt to perform the molecular biological experiments in silico, one of the most promising alternative means of the biological experiments in vitro. Among the various fields of experiments, this paper is especially focused on molecular cloning, because DNA information is more systematically and exhaustively collected than that of other molecules. There are some advantages in in silico experiments in view of planning molecular biological experiments, usage as lab notebook, educational tool to learn molecular biological experiments. However, performing in silico cloning requires some formulations such as recording of the end shapes of digested products by restriction enzymes or amplified products by PCR. For this purpose, we introduce some extensions to GenBank/EMBL database annotation convention, and incorporate them into existing convention as new feature keys and qualifiers. In addition, features on a DNA sequence are occasionally truncated in the case of amplified products of PCR or digested fragments by restriction enzymes, therefore the annotations about the truncated features should be also formulated. The ambivalent nature of DNA requires frequent changes of the interested strand to and fro, therefore a reverse complementary operation of large size DNA is necessary to be implemented for such software. According to these definitions or data descriptions, we have developed a software tool for in silico experiments, named in silico MolecularCloning, and perform a few of typical molecular cloning experiments on computer, and verified that this approach would be effective.

Key Words: molecular cloning, in silico experiment, bioinformatics, software, restriction enzyme, PCR, ligation

1. Introduction

Cloning is the most popular experiment routinely performed in the molecular biology field. The experiments consist of basic routines such as DNA digestion by restriction enzyme, ligation and PCR, and so forth and routines for detection, such as gel electrophoresis (Figure 1). In practical way, a variety of combination of these basic routines, are repeated in laboratories.

Experiments in molecular biology are defined as the series of basic routines starting from the sample DNA molecules, after receiving various chemical or physical changes such as purification, amplification,
modification or decomposition, until obtaining final target materials which are regarded as more feasible to be detected in amount or in property. Namely, when DNA samples are applied into a test tube with given conditions, the molecules in the tube undergo chemical changes. It is suggested that a variety of consecutive experiments would be performed until the final target molecules are obtained.

In this paper, we are making an attempt to perform the molecular biological experiments in silico, in other words, in computer, one of the most promising alternative means of performing the biological experiments in vitro. However, the word, in silico experiments, may cause an ambiguity, or might have a meaning too broad. Thus, this word should be confined within the molecular cloning in which experiments are handled with mainly on DNA molecules. Therefore, this paper is focused on in silico molecular cloning. How far is in silico cloning feasible or effective? How should we formulate each component of the virtual cloning, such as DNA samples, reagents or reactions? In the rest of the paper, we try to provide answers to the above questions.

Most software tools, when handling DNA, define it as a single sequence. In nature, the DNA is, actually, consisting of two separate sequences which are entwine themselves each other by many hydrogen bonds, which make double helix configuration. Most of the regions in DNA have bases complementary to each other. Thus, bases on one strand correctly suggest the corresponding hydrogen bonded bases of another strand. Accordingly, it seems to be natural for such software to adopt a single sequence for the expression of DNA. However, as for the both ends of the molecule, it is quite different. They likely take single strand shapes at the end of DNA which takes an important role especially in ligation against other DNA end points. Most of such software tools do not

---

**Figure 1** A flow of cloning experiments. (a) Digestion of DNA by a restriction enzyme BamHI produces linear DNA fragments with 5' sticky ends. A circular vector is linearized by digestion. (b) Amplification of DNA by PCR produces DNA fragments between two priming sites recognized by the given PCR primers. Occasionally, ambiguous priming produces different size PCR products which can be seen by 1D gel electrophoresis. (c) The linear vector and a DNA fragment are ligated together with the presence of an enzyme ligase if the end shapes of the both ends are matched. This reaction produces two different circular vector with the insert is reversely ligated to the vector. (d) The vector with the insert is transformed into a host cell. The host cells are cultured until a number of clones are obtained. (e) 1D gel electrophoresis separates many DNA fragments by size.
have an expression about the end shapes of DNA molecules. Without information on the ends of DNA molecules, in silico cloning experiments are not feasible.

Here, we emphasize that it is necessary to express the end shapes of DNA before developing in silico cloning software. Therefore, we propose extra expression about the end forms of DNA molecules at first. Then, we discuss the in silico cloning software.

Genome projects such as the human genome, or many microbial genomes, have targeted the DNA sequencing across whole genomes, and most of these genomic DNA sequences are publicly available [1,2]. DNA sequences of artificial vectors and amino acid sequences of restriction enzymes, where both materials are usually used in cloning experiments, are also available in nucleotide or amino acid databases [3]. Thus, the infrastructure for launching in silico experiments has already been established by these preceding databases. For an organism, when its DNA sequence is determined, it becomes possible to make a logical plan for performing experiments, such as cloning of specific genes on the genome. In the present study, we introduce simulation of experiments for molecular biology referred to as in silico cloning focused on molecular cloning based on genome information.

2. Requirements for in silico cloning software

Taking such advantages of in silico experiments into consideration, we have developed software for in silico experiments for molecular biology focused on DNA cloning. In this section, we describe the definition of in silico cloning and requirements for the software. Typical functions of in silico cloning are implemented on this newly developed software on the basis of the required definitions and format for the DNA samples to describe state of these molecules. Then the feasibilities of performing such experiments in silico are discussed. Although many software tools have been developed to implement cloning experiments, few of them have focused on this treatment of consecutive operations of cloning. A common consensus about software requirements for

---

**Figure 2.** An example of GenBank format file. The end shapes of the DNA is described by a feature key “endtype” with its qualifier “endtype=n, m”. The above example shows that the DNA has two single stranded bases at both 5’ ends. The bases are recorded as the first two and the last two bases of its nucleotide sequence part.
*in silico* experiments should be established before describing about the software itself. That is, such software would satisfy at least the minimum requirements such that consecutive experimental routines could be performed and could describe and visualize the states of molecules such as single stranded ends of DNA, circular double stranded DNA with staggered and blunt ends and so on as described in Section 2.1. To satisfy these requirements, we propose the formats for sample DNA, reagents as described in Section 2.2.

### 2.1 Design of *in silico* cloning software

So as to make a consensus for what are *in silico* experiments, we, hereafter, propose the requirements, although this is, in the same time, regarded as definitions on in silico experiments, for software to be implemented as follows:

1. A recording unit of a DNA molecule should be one-to-one correspondence to each molecule participating in a reaction. Namely, information concerning the molecule should be stored as a single corresponding recording unit.

2. After a reaction has finished, information on input molecules would be inherited to the resulting products. Namely, whenever DNA molecules are fragmented into several smaller and partial ones or amplified as many clones, or aggregated into fewer molecules, information on the corresponding regions of them must be exactly transferred to those of newly produced molecules.

3. Whenever a DNA molecule is aggregated together from two different DNA molecules by covalent bond, the combined molecule should have a single DNA sequence without trace of reaction. That is, any evidence can not exist to show where or when the bonding reaction occurred.

4. In a circular DNA, visualization, detection of recognition sites and reaction, which occurs on any bases of the DNA, should be equivalently handled. Namely, this means that in circular DNA there should not be any gap between any bases.

5. An enzyme, as a protein, can be treated as a catalyst, i.e. nothing of these molecules should be affected or changed after any reaction. However, this is a temporary requirement adaptable only for the time being, until adequate information on such enzymes would be established as publicly available databases.

6. As for the selection of the viewing strand of DNA, there should not be any significant weight between the selected strand and the complementary one. And moreover, there should be a function of switching to the reverse complementary strand from current strand, with all the features kept on it.

These might lead to realize closer similarity

<table>
<thead>
<tr>
<th>Restriction Enzymes</th>
<th>Fragments</th>
<th>Annotation by IMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) BamHI</td>
<td>5′GATCC...G3′</td>
<td>/endtype=-4,4</td>
</tr>
<tr>
<td></td>
<td>3′G...CCTAG 5′</td>
<td></td>
</tr>
<tr>
<td>(b) BclI</td>
<td>5′GATCA...T3′</td>
<td>/endtype=-4,4</td>
</tr>
<tr>
<td></td>
<td>3′T...ACTAG 5′</td>
<td></td>
</tr>
<tr>
<td>(c) EcoRI</td>
<td>5′AATCC...G3′</td>
<td>/endtype=-4,4</td>
</tr>
<tr>
<td></td>
<td>3′G...CTTAA 5′</td>
<td></td>
</tr>
<tr>
<td>(d) AatII</td>
<td>5′C...GACAT5′</td>
<td>/endtype=4,-4</td>
</tr>
<tr>
<td></td>
<td>3′TAGAG...C5′</td>
<td></td>
</tr>
<tr>
<td>(e) PvuI</td>
<td>5′CG...CGAT3′</td>
<td>/endtype=2,-2</td>
</tr>
<tr>
<td></td>
<td>3′TAGC...GC5′</td>
<td></td>
</tr>
<tr>
<td>(f) EcoRV</td>
<td>5′ATC...GAT3′</td>
<td>/endtype=0,0</td>
</tr>
<tr>
<td></td>
<td>3′TAG...CTA5′</td>
<td></td>
</tr>
<tr>
<td>(g) SmaI</td>
<td>5′GGG...CCC3′</td>
<td>/endtype=0,0</td>
</tr>
<tr>
<td></td>
<td>3′CCC...GGG5′</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3.* Examples of DNA fragments digested by restriction enzymes. (a) 5′ sticky ends of a BamHI digested DNA fragment have four single stranded bases. The end shapes are annotated as “/endtype=-4,4” by IMC. (b) 5′ sticky ends of a BclI digested fragment. (c) 5′ Sticky ends of an EcoRI digested fragment. (d) 3′ sticky ends of an AatII digested fragment. (e) 3′ sticky ends of a PvuI digested fragment with two single stranded bases. (f) Blunt ends of an EcoRV digested fragment have no single stranded base. The end shapes are annotated as “/endtype=0,0”. (g) Blunt ends of a SmaI digested fragment.
between *in vitro* and *in silico* experiments.

### 2.2 Definitions of formats for sample DNA, reagents

Nucleic acids and proteins are well-known to be expressed as a sequence of characters. DNA or RNA are coded with four characters, T or U, C, A and G (T for thymine, U for uracil, C for cytosine, A for adenine and G for guanine, respectively) as one dimensional character sequences, while proteins are coded with 20 different characters and make one dimensional sequences. International nucleotide databases, such as GenBank, EMBL and DDBJ were established to systematically collect DNA and RNA nucleotide sequences, while Uni-Prot and others are collecting protein sequence information. Recently, most of the amino acid sequence entries are derived from translation of the coding regions of the genomic sequence of DNA.

In the international nucleotide database conventions, most of biological characteristics, such as genes, are expressed as feature keys and their qualifiers as shown in Fig. 2. Feature keys are defined as units which have functions or structures of biological meanings, while their qualifiers providing values or contents. When implementing software for *in silico* experiments, it is desirable to use these descriptions as much as possible. The reason is due to the fact that a large amount of nucleotide sequence is submitted to the international databases and anyone can access such databases and obtain the requested data anytime. Therefore, utilization of these international nucleotide database format records increases the data availability in a great extent. However, current conventions of GenBank/EMBL do not always satisfy the requirements to perform *in silico* experiments. Therefore, some extension for the existing conventions would be necessary. Software implementation is discussed in the following examples.

#### Example 1. Description of the end shapes of DNA fragments

Extension required for qualifiers, is description of the end shapes of DNA fragments after digestion by restriction enzymes or PCR amplification. The end shapes of DNA fragment after RE digestion or PCR are likely to take sticky ones, namely to be single stranded at the end. To describe these end shape, categories of feature keys and qualifiers must be extended as shown in Fig. 2. The annotation based on “endtype” is introduced as a new feature key to describe this molecule has an expression about the end shapes, and new qualifier named “endtype=m,n” are introduced to describe the type of ends created after reaction (Fig. 3). Actually, it is not necessary to describe about the cause of reaction for the new feature key, because we generally do not know the cause.

<table>
<thead>
<tr>
<th>Examples</th>
<th>Fragment A</th>
<th>Fragment B</th>
<th>End shape Check</th>
<th>Complementary Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) BamHI + BclI</td>
<td>BamHI</td>
<td>BclI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>endtype</td>
<td>5' GATCC........G 3' , 5' CCTAG 3'</td>
<td>5' GATCA........T 3' , 5' ACTAG 3'</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>(2) BamHI + EcoRI</td>
<td>BamHI</td>
<td>EcoRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>endtype</td>
<td>5' GATCC........G 3' , 5' CCTAG 3'</td>
<td>5' AATTG........G 3' , 5' CTAA 3'</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>(3) BamHI + AatII</td>
<td>BamHI</td>
<td>AatII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>endtype</td>
<td>5' GATCC........G 3' , 5' CCTAG 3'</td>
<td>5' FACAG........C 3'</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 4.* Examples of ligation reactions. Two fragments digested by BamHI and BclI can be ligated by End shape check and complementary check. Two fragments digested by BamHI and EcoRI cannot be ligated by complementary check. Two fragments digested by BamHI and AatII cannot be ligated by End shape check and complementary check.
cause of reaction.

In the case of ligation, it is possible to simulate a ligated DNA fragment derived by two DNA fragments if each end of DNA fragments matches in a complementary style. This is discriminated by using the "endtype=mn", where m means 5’ single stranded base length and n means 3’ single stranded base length as shown in Fig. 3. In the case that ligation would occur, a pair of ends match in length along single stranded bases and also match in complementary sequences. Let’s take the ligation between the end digested by BamHI, and the end digested by BclI, as an example (Fig.4a). The ends digested by BamHI, are expressed by “endtype=4,-4”, while the ends digested by BclI are also expressed by “endtype=4,-4”. Ligation occurs between 5’ end of one DNA fragment and 3’ end of another fragment, so the discriminating process goes on like this. (1) Among the four possible combinations (if the reverse complementary fragments are considered), the first one is taken to be examined. (2) The first combination (4, -4), (m, n) means the testing of 5’ end of a BamHI fragment and 3’ end of a BclI fragment. (3) Simple sum is performed between the digits in the bracket, if the answer is zero this means shapes are matched for ligation between the two ends. (4) Then, reverse complementary sequence between the two single stranded bases are examined, and the BamHI product has single strand of “GATC” at the 5’ end, while the BclI has single strand of “CTAG” at the 3’ end. Thus, the combination is proved to be ligated. (5) Remaining three combinations are examined and all the combinations are proved to be ligated. Further examples are shown in Figure 4. This means two circular DNA would be produced by ligation reaction between the two DNA fragments. In addition, it is better to implement suppression of ligation if phosphate group is removed from one or both ends by some enzymes like phosphatases. Thus, the implementation of the feature key “dephosphorylation” is necessary in view of the higher visibility, integrity and compactness for describing the state of DNA molecule.

Example 2. Description of recognition site sequences, cleavage patterns, and affection by methylation of recognition sites

The information on recognition site sequences and cleavage patterns of each restriction enzyme is registered in REBASE, a restriction enzyme database. However, according to the requirements of in silico experiments, data entries from amino acid databases would be applied as input data for restriction enzyme. Namely, it is desirable to use rather amino acid data entries which describe RE recognition site sequences, cleavage patterns and possibility of affection by methylation at the site.

![Figure 5](image-url)

**Figure 5.** Overview of IMC. This is the main window of IMC, in silico MolecularCloning. On top of the window, various buttons for operation of the software are placed. By clicking one of them enable users to operate the program. (a) The button for reading DNA sequence files. (b) Zooming buttons. (c) Cloning buttons include restriction enzyme button, primer design button, PCR button, ligation button, plasmid map drawing button, blunting button and dephosphorylation button.
3. Development of software for in silico experiment: in silico Molecular Cloning (IMC)

We have developed a software tool for in silico experiment referred to as in silico MolecularCloning (IMC), to implement capability of such experiments on computers. IMC imports DNA samples which are simply available in international nucleotide databases, and can draw any “feature keys” on the given DNA sequence, as well as implementing in silico experimental functionality in it. The software is written in Java Application language, and can be installed on both Windows Xp and Mac OS X. IMC can be downloaded via internet from http://www.insilicobiology.co.jp/ and we provide a trial license to use this software for two week.

3.1 Outlines of software operation

IMC can import nucleotide sequences as target DNA samples outlined in Figure 5. Imported DNA sequences are viewed as different icons depending on its linear or circular shape. By clicking these icons, selected DNA sequence is viewed and printed as colored and multi-figured feature map, where a feature map is defined as a map expression of biological annotation, namely, viewing of features on the exact base position with selection of drawing options. Sequence is called current one if it is currently shown in the feature map. For this current sequence, in silico cloning experiments could be performed. Namely, when performing restriction enzyme digestion, calling all or any of the enzymes from enzyme database and resulting digested DNA fragments are registered as new files with inheriting its ancestors’ annotation on them. Accordingly, consecutive experiments against newly produced molecules could be performed.

3.2 Importing samples

A GenBank/EMBL format file can be imported as a DNA sample. Any region of the imported nucleotide sequence becomes targets of in silico cloning. DNA sequence without annotation such as FastA format files could also be imported. Support of multiple format data enables to import a large number of entries by single operation.

3.3 Handling enormous size chromosome

In in vitro experiments, even large size chromosomes, such as chromosome no.1 of human genome whose size amounts to 250 Mbp, should be commonly handled likewise a much smaller DNA is handled. Although it had been difficult for many programs to draw such huge map of DNA sequence with all the features on it, IMC implements this requirement and can scroll the map fast enough along the large chromosome.

3.4 Drawing and viewing of circular DNA

When describing circular genome by using a GenBank/EMBL format file, a difficulty is encountered because of its linear presentation of character sequence. In in vitro molecule, there is no gap among the bases, namely, this sequence has no limit on it like a loop. IMC has implemented reaction across the gap, namely bridging the first base and the last base of the publicly recorded sequence text.

3.5 Editing of features on DNA sequences

Features on the sequence on a GenBank/EMBL format file, are exactly drawn on the feature map of the DNA. When viewing it from another strand, all the features are rearranged to opposite strand with reverse direction. If there are single stranded ends, these styles are also expressed in reverse way. When any region of such a DNA is amplified by PCR or digested by restriction enzymes, these may start or end within several features which make the affected features incomplete ones. IMC controls these operations automatically.

4. Results and Discussion

4.1 In silico vs. in vitro experiments

In comparison of in silico experiments with in vitro experiments, let’s consider upon the digestion of a DNA sample by restriction enzymes. Information on
DNA is collected in the international nucleotide database such as GenBank or EMBL. As each sequence of DNA is recorded as one individual entry, it is also appropriate to use it as a sample to in silico experiments. In in vitro experiments, DNA samples are applied to a micro tube, followed by applying solution of restriction enzymes, then a reaction is allowed to progress in the tube, then digested fragments are obtained.

In silico version of this experiment is processed as software programs read the corresponding data file which describes a DNA sample sequence and annotations, and stores these data into a folder designed as a simulated micro tube. As soon as the sample is read, the corresponding feature map is drawn with many features indicated by a variety of figures. Typical experiments can be done against this sample. All the implemented experiments are applied to the DNA sample. For example, digestion by restriction enzymes is performed as follows. In in silico version, the restriction enzyme database is searched for cutting appropriate sites of DNA sequence and target enzymes are selected to be applied to the tube. Then, the reaction would be started until completely digested fragments are obtained. In the first step of digestion, a list of restriction sites are shown, then selection of sites leads to final digested DNA fragments, each of them is described and saved as a GenBank or EMBL format file.

Actually, IMC consists of a set of experiment functions ranging from digestion by restriction enzyme, PCR, ligation, and so forth as shown in Table 1. An experimental operator is defined as in silico process corresponding to the equivalent experiment in vitro. Thus, experiments in silico are regarded as computing of DNA sequences by some of the experimental operators to obtain the resulting sequences. There are some advantages in in silico experiments in view of (i) planning molecular biological experiments, (ii) usage as lab notebook, (iii) educational tool to learn molecular biological experiments.

(i) Clarification of procedure by predicting the results of actual experiments: A large cost is required in in vitro experiments handling DNA samples because reagents or disposable materials such as plastic micro titer plates, micro tubes and pipette tips are consumed in large amount. Genome-wide projects require much higher rates in resource consumption. in silico experiments can provide low cost alternatives for verifying the effectiveness of in vitro experiments prior to starting them actually.

For example, in a molecular cloning experiment, there is a case that insert DNA is actually inserted into vector in reverse way. To detect this event, an optimal selection of restriction enzyme is necessary, and leads to a good method to determine the direction of insert using 1D gel electrophoresis.

(ii) Usage as lab notebook: If there is a series of experiments that researchers are planning to perform, there are always some modifications against the standard protocols, so these procedures should be recorded just at the time the plan is ready. The history of these records could be utilized as a lab note, which is also important for researchers to verify the finished experiments. Recording such as comprehensive drawing of a plasmid map with DNA insert with features on it, will increase efficiency of the experiments.

(iii) Educational tool to learn molecular biological experiments: It might be difficult for beginners to understand what are going on in the micro tube in molecular biology experiments in vitro because molecules in the tube are only visible with indirect means like gel electrophoresis, in silico experiments, which can show any molecule at any time, help them to understand the micro phenomena and process of experiments.

An example of confirmation viewer which shows the target molecule as it is

Most of plasmids and prokaryotic chromosomes have circular structure in vitro or in vivo. Such circular structure has actually no end point along its nucleotide sequence. However, when it comes to record the nucleotide sequence in a text file, it can not be avoided to describe them as a string of characters with the start character and the end one. A seamless viewing and drawing of circular DNA sequence is thus required as well as for recognition or reaction across the point in in silico experiments.

Most of the ends of a DNA fragment resulting from digestion by restriction enzymes or from amplification by PCR are likely to have shapes of single strand bases protruded, instead of simply double stranded in most of DNA regions. This means that some additional expressions are necessary for describing such end shapes of DNA. A DNA fragment would change its form from symmetrical to
asymmetrical structure, and *vis a versa*. This is an important feature of double strands DNA whose ends have same activities against reactions.

**4.2 Operation of DNA cloning by IMC**

**DNA cloning** means the extraction of a particular region of DNA (often that of a particular gene) from a genomic DNA or from other DNA sources, then the extracted DNA fragment is inserted into a plasmid vector. After transformed them into host cells, a lot of clones are obtained as the host cells are cultured. Cloning any designated segment of DNA from a genome is one of the most important techniques of recombinant DNA technology, as it is the starting point for understanding the function of any region of DNA within the genome. Figure 1 shows protocol for DNA cloning. DNA cloning can be carried out by mainly three experimental protocols, digestion of DNA by restriction enzyme (RE digestion), polymerase chain reaction (PCR), ligation reaction, and transformation of the vector constructed to bacterial cells. In addition, to check the band patterns of gel electrophoresis are generally used to confirm whether or not DNA fragment of interest is obtained. In this section, we describe how efficiently IMC works in DNA cloning in combination of experimental units such as (i) RE digestion, (ii) PCR, (iii) ligation reaction, (iv) transformation and (v) gel electrophoresis.

(i) **RE digestion**

The restriction enzymes used in cloning technology are derived mainly from bacteria, and their recognition sequences are too short to accidentally occur in any long DNA molecule. Thus restriction enzymes can be used to analyze DNA from any source. The main reason why they are useful is that a given enzyme will always cut a given DNA molecule at the same sites.

There are two types of ends of DNA fragments cleaved by restriction enzymes, *sticky* or *blunt ends*. For example, restriction enzyme BamHI recognizes the site 5'-GGATCC-3' and cleave it as two DNA fragments with 5'-G-3' and 5'-GATCC-3' which carry sticky ends (Fig. 3a). On the other hand, blunt ends are produced by cleavage by restriction enzyme, such as EcoRV (Fig. 3b). Note that sticky ends enable the ends of the two fragments to base-paired correctly with each other. This ligation also reconstructs the original restriction enzyme recognition site, which allows DNA fragments to be easily inserted or removed. Information on restriction enzymes has been accumulated in REBASE [3].

IMC has functions concerning DNA fragmentation by restriction enzymes as shown in Table 1. One is to be able to recognize both two types of ends (staggered and blunt ends). This is important that the consecutive ligation is performed by using such information. Once a DNA fragment is digested by a restriction enzyme in IMC, the end type of the fragment is described with the qualifier described in Example 1 of Section 2.2.

Once, a DNA sample has been digested by restriction enzymes, an extra description is added or re-written so as to identify the exact shape of its end. When the end has a single strand or protruding bases (sticky end), the end shapes of a digested DNA are expressed as two integers with sign depending on the direction of protruding bases and with comma as the delimiter which separates 5' and 3' end digits. A DNA sequence with this description will be examined whether the ligation between two ends of any fragments is possible or not, when these fragments are specified as target molecules applied with ligase enzyme.

As mentioned above, the single stranded bases in the double stranded DNA can be recorded upon the information about each end shape. Along most of DNA sequence, a DNA molecule has double stranded shape so only one strand sequence is adequate to be recorded. In addition to that, one end with actually complementary to the recorded strand must be converted into complementary sequence before viewed. Therefore IMC keeps the description of the extra single stranded bases after digestion. The bases on the sticky end of the complementary strand to the current sequence, are immediately converted as the bases on current strand, therefore it is always enough to record only one strand sequence.

Digestion by a restriction enzyme sometimes occurs just on any feature of the DNA sequence. After digestion, the feature is divided into two separate fragments. Namely, the each fragment has one incomplete feature on it each other. IMC performs this operation precisely.

(ii) **PCR**

Using PCR technique, a given nucleotide sequence can be selectively and rapidly replicated in large amounts from any DNA sample that contains it. PCR is iterative reactions consisting of three steps by starting with a double-stranded DNA; separation of
two strands (Step 1), hybridization of two primers to complementary sequences in the two DNA strands (Step 2), and synthesis of DNA from the two primers (Step 3). In the case of \( N \)th iteration, DNA fragments specified by two primers are produced to \( 2^N \) times of them.

Like as digestion by a restriction enzyme, priming in PCR sometimes occurs just on any feature of the DNA sequence. After PCR, the feature on the PCR product becomes incomplete one with either side of the feature missing. As for the primers themselves, they are usually modified by insertion or substitution of a few bases. These modifications are inherited to the PCR product. IMC performs this operation precisely. Occasionally, different sizes of multiple DNA fragments are obtained in PCR because of a few base-mismatched priming to complementary DNA. These multiple DNA fragments can be viewed by 1D gel electrophoresis in IMC.

(iii) Ligation

The enzyme DNA ligase reseals also the nicks in the DNA backbone that arise during DNA replication and DNA repair, and has become one of the most commonly used tools of recombinant DNA technology, as it allows to combine any two DNA fragments. Isolated DNA fragments can be recombined in the test tube to produce DNA molecules.

IMC can tell whether two DNA fragments are ligated or not. This is important for constructing a vector including DNA fragment of interest. The results of in silico digestion by restriction enzymes or in silico PCR performed by IMC, are recorded as to reconstruct the shape of each products. When some of these products are the targets of in silico ligation, a test is performed if the end shapes are matched to be ligated. If not, ligation does not occur. The results are easily verified by drawing the plasmid map or one dimensional gel electrophoresis.

It should not identify the exact bonding site of ligation after the reaction. Namely, there is no trace of ligation on the DNA ligated in experiment in vitro. However, identification of the ligation site sometimes would be convenient because IMC users would like to draw the plasmid map with the insert sequence ballooned outside the map. In this case, the answer is that it records both the DNA fragments with different sources, instead of recording the ligation site. It has been a custom to describe that the ligated DNA sequence with various inserts should be recorded with the feature key “source”. Self-ligation could be
described in the same manner. A single DNA fragment would be ligated with both ends of its own if matched for ligation. This ligation produces one circular DNA molecule, and the ligation site would not be identified after reaction. In this case, it is difficult to hide the trace of ligation because the sequence itself is actually recorded as a string of characters with the start character and the last one.

IMC realize a circular DNA map by examining the definition line information, “linear” or “circular” of the international nucleotide databases. If the definition line is written as “circular”, IMC would draw one circular DNA feature map without identifying the start base or last base. That is, the map can be eternally scrolled to one direction or another without stop.

If a DNA fragment is removed by a phosphate group at its 5’ end by enzyme phosphatases, this end could not be ligated with any end of other DNA any more after the reaction. We have introduced a new feature key “dephosphorylation” for this purpose. If the DNA fragment has the feature key “dephosphorylation” as the expression according to the international nucleotide databases, ligation to this DNA fragment does not occur.

(iv) Other experiments

DNA can be introduced into bacteria by a mechanism called transformation. DNA fragment does not change in transformation. A DNA fragment transformed with a vector plasmid, could be methylated in certain sequences if the host is that of Escherichia coli. If methylated, the sequence might not be digested by some of the restriction enzymes which are affected not to digest their recognition sites on methylated DNA. IMC recognizes these affections by methylases in such cells. If some recognition sites are affected by methylation, IMC does not digest at these sites.

4.3 Examples of IMC software operation of consecutive routines

As described above, IMC implements fundamental in silico experiments concerning DNA cloning, that is, RE digestion, PCR, ligation reaction, transformation, and gel electrophoresis. We explain how IMC performs the consecutive routines by giving an example of TA cloning. In TA cloning [4-6], a DNA fragment with sticky end of a deoxyriboadenosine obtained in PCR procedure is cloned into the linear plasmid-vector with a deoxyribothymine (dT) addition at the 3’ end, which is the complementary to the 5’ of the PCR product.

This experiment is performed in silico with IMC, according to the procedure described below.

1. Applying a sample of DNA: Apply a sample of DNA, namely, read a GenBank format file of a DNA sequence. This is carried out by clicking “Read Sequence File” button. Then, “Read Sequence File(s) into Feature Map” dialog window is popped up, where one or more sequence files are selectable to be read and feature map of one of them is shown (Fig.6(1)).

2. Specify the region to be amplified on the map: After showing the region to be amplified by manipulating zoom or scroll buttons, dragging mouse across the region on the feature map results in change of the background color in red (Fig.6(2)).

3. Design of optimal primer sets: Design optimal primer sets to amplify the declared region of the DNA sequence. On the colored region, clicking of the right button of mouse pops up a menu. On the top of the menu, there is a submenu “Design PCR Primer”, then clicking of the submenu pops up “PCR Primer Design” window where a set of parameters for designing PCR primers are listed and can be changed. PCR runs immediately after clicking “set” button at the bottom of the window (Fig.6(3)).

4. PCR: Let PCR be performed after selecting one of the just designed primer sets. This is initiated by starting PCR with clicking “PCR” button after selection of the primer, then “Priming site search” window is popped up where one set of primers can be selected with allowance of mismatches. If one base mismatch is allowed, select the radio button of “1 base mismatch”, then clicking of “Reaction” button starts PCR. All or specified portion of the DNA sample sequence are searched for priming sites, and a list of PCR products is shown (Fig.6(4)).

5. Selection of PCR: After selection of PCR products from the list of PCR products, a click on “Reaction” button starts PCR. Perform one dimensional gel electrophoresis for all the above PCR products. One clear band is shown on the lane with many dark bands resulting from mismatched priming. On the bottom of “Priming site” window, there is a button for one dimensional gel electrophoresis which is used to start gel electrophoresis about all or selected PCR products (Fig.6(5)).

6. Registration of PCR products in DNA
sequence file: Any of products is registered in a new GenBank or EMBL format file. The shape of the both ends of a PCR product can be changed dependent on the kind of PCR enzymes. IMC lets users select type of shapes of one adenine base sticky or blunting end. The end shapes of the PCR product can be verified by scrolling the map until the both ends of DNA sequence are shown (this could be done one by one), after making the PCR product DNA as the current one. Each of the PCR products is recorded as a single entry of GenBank or EMBL format file. The contents of the file can be verified by a click on the “GenBank/EMBL Viewer” button(Fig.6(6)).

7. Selection of Restriction enzyme: Let us take a vector sequence digested by the restriction enzyme “EcoRV” which generates blunting ends. This is done by clicking “RE Recognition” button after the vector sequence is selected as current one. Then, “Enzyme Selection Window” is popped up and checking on “EcoRV” and clicking of “Show Recognition Site” button makes the “Recognition site” window to pop up on which a list of the recognition sites by the enzyme is shown. Clicking “Digestion” button starts the digestion process(Fig.6(7)).

8. Addition of thymine to 3’ end: Let one thymine be added at 3’ end of the above vector. This is a virtual reaction for IMC user’s convenience and could produce a T-vector for TA-cloning. This is carried out by clicking “Add T-base at both 3’ends” button. The end shapes of the digested product can be examined by scrolling the map until the both ends of sequence are shown, after making one of the digestion fragments as the current one(Fig.6(8)).

9. Ligation: Let one of the PCR products and a T-vector be ligated. This reaction generates a circular plasmid vector with the PCR product inserted. This is carried out by clicking “Ligation” button. On the pop up window for ligation, two DNA sequences can be specified. After setting the T-Vector sequence as the first fragment and setting the PCR product as the second, clicking of “Ligation” button starts ligation and produce two circular DNA after reaction(Fig.6(9)).

10. Plasmid map: Before drawing of the plasmid map of the vector, list of DNA sources which consist of the circular DNA, is shown for the selection of inserted DNA. The inserted DNA is clearly identified on the plasmid map. This is done by simply clicking the “Plasmid Map” button. On the popup, a list of sources on the ligation product is shown. After selecting one for the source of the inserted PCR product, clicking of “Set” button leads to draw the plasmid map with the insert of PCR product. Map sizes can be changed by modifying the parameters(Fig.6(10)).

5. Conclusion and Remarks

Performing of in silico cloning requires recording of the end shapes of digested products by restriction enzymes or amplified products by PCR. For this purpose, we introduce a new feature key “endtype” and its qualifier “endtype”, and incorporate them into GenBank/EMBL database annotation convention. Some features on a DNA sequence might be truncated by PCR or digestion by restriction enzymes, therefore the annotations about the truncated features should also be modified. The ambivalent nature of DNA also requires occasional switching to the interested strand from one to another. In addition, we redefine information about the RE recognition sequences, end shapes after digestion and affection of methylation at the sites, to a new feature key and qualifiers. According to these definitions or data descriptions, we have developed a software for in silico experiments, and perform a few of typical molecular cloning experiments on computer, and verified that this approach would be effective as a recording tool of a series of experiments as a lab notebook, training tools for beginners to molecular biology, prior simulating tool for time or cost consuming experiments.

In molecular biology experiments, it is important how to describe the functionalities or activities of enzymes and how to use such description. According to the requirements of in silico experiment, one data entry to one enzyme seems to be the best way. Descriptions on the enzymatic functional sites, are still poor in case of most of protein database entries. Therefore, in this stage, we assume that enzyme proteins act as only catalyst instead of multi-functional protein which has residue-specific activities around its amino acid sequence. Further
research is necessary to take these activities into consideration for in silico experiments. In addition, in most laboratory experiments, reaction temperature or reaction time is an important factor, while IMC has not considered these factors yet. This remains as a future work.

References and Notes


ゲノム時代のバイオインフォマティクスツール：インシリコ実験への第一歩（分子クローニングに着目して）

大山 彰(1,2)*、黒川 顕(2)、恵内 京徹(1)、斉藤 仁浩(1)、金谷 重彦(2)、モハマド・アルタフル・アミン(2)、小笠原 直毅(2)

(1)インシリコバイオロジー株式会社、横浜市中区山下町24番地8 SOHO STATION 706号、
(2)奈良先端科学技術大学院大学情報科学研究科、奈良県生駒市高山町8916-5

インシロト（試験管内）での生物学実験を代替する手段として有力と考えられているインシロ（コンピュータ内）での分子生物学実験を試みている。生物学の数多くの分野の中でも分子クローニングに着目し、生物分野でもっとも系統的にまた網羅的に情報が収集されているDNA情報を利用する。インシロ実験にはいくつかの利点があり、たとえば、分子生物学実験の計画に活用できる、実験ノートの代用となる、分子生物学実験の教育用ツールとして用いることができることなどが挙げられる。しかしながら、インシロ実験を実装するためには、制限酵素による消化切断生成物の末端形状やPCR生成物などの記録方法などの定式化が必要である。このため、GenBank/EMBLデータベース注釈規約にいくつかの拡張を行い、新規FeatureおよびQualifierとして定義した。さらに、DNA配列上に表現されたFeatureはPCRや制限酵素消化切断などの際に部分的に削られることもあり、これが定義した規約の定式化を行った。また、DNAの2面的な性質により、通常興味の対象としては片面のStrandのみを表示するが、インシロ実験を実装するソフトウェアにおいてはこの対象としての長鎖DNA配列をその上にFeatureを保持したまま逆相補鎖と頻繁に切り替える操作が必要であることを示した。これらの定義やデータ記述の従い、我々は「インシロモレキュラークローニング」と呼ぶインシロ実験ソフトウェアを開発し、いくつかの典型的な分子クローニング実験をコンピュータ上で実行し、この方法が有効であることを示した。

キーワード：分子クローニング、インシロ実験、バイオインフォマティクス、ソフトウェア、制限酵素、PCR、ライゲーション

* [E-mail address] akr@insilicobiology.co.jp