Cloning simulation in the cage environment

R.J. Douthait, J.J. Thomas, S.D. Rosier, J.E. Schmaltz and J.W. West

Battelle, Pacific Northwest Laboratories, Richland, WA 99352, USA

Received 29 August 1985

ABSTRACT

The CAGE/GEM™ software toolkit for genetic engineering is briefly described. The system functionally uses color graphics and is menu driven. It integrates genetics and features information ("Overlays") with information based on sequence analysis ("Representations"). The system is structured around CAD (Computer Aided Design) principles. The CAGE (Computer Aided Genetic Engineering) aspects of the software are emphasized and illustrated by a simulated cloning of the hepatitis B core antigen gene into the BamHI site of plasmid pBR322.

INTRODUCTION

CAGE/GEM™ is a software system developed to provide computer genetic engineering tools which closely match the development process used in the laboratory. The system provides an environment where the user has one interface to a broad range of tools all of which can be applied in any order and can be combined to see simultaneous results. Parameters are specified interactively and the results are visible immediately. No user management of intermediate results is required. The system functionally uses color graphics and is menu driven using single keystrokes and a cursor.

CAGE/GEM™ incorporates many techniques from traditional Computer Aided Design (CAD) systems for mechanical and electrical design. These concepts include dynamic graphics, automated file handling, single user interface, numerous complex as well as simple viewing representations, unstructured functional ordering, and knowledge integration.

The CAGE/GEM™ software is available for installation on existing computer systems. Information about licensing arrangements and cost can be obtained by contacting the principal author of this paper at Battelle, Pacific Northwest Laboratories, Battelle Blvd., Richland, Washington, 99352, (509) 375-2653.

In addition to information gained from direct sequence analysis, the location and definition of genetic elements and other features related to sequences are also of primary importance. In the EMBL and GENBANK sequence databases, information of this sort is given in the header sections of the sequence files. These features are related to the sequence by appropriate coordinates. In many instances, knowledge of the relationship of these features to appropriate restriction enzyme sites is sufficient information for planning and creation of new constructs.

Genetic and plasmid maps delineating genetics and features can be envisioned as low magnification views of sequences. A high magnification view of a given sequence or any subsequence would then be the actual string of alphabetic characters representing the actual order of the structural elements comprising the sequence. CAGE/GEM™ provides a zooming feature which allows dynamic and interactive viewing at various magnification levels. The sequence proper and information obtained from analysis of the sequence (Representations) and the genetics and features information (Overlays) are integrated and retained at all magnification levels.

The use of the system for creating mock genetic constructs is called Computer Aided Genetic Engineering (CAGE). In this paper, those aspects of the system essential for CAGE are briefly described and an example of "Simulated Cloning" of the Hepatitis B core into the TET region of plasmid pBR322 is given.

**REPRESENTATIONS**

Information gained from sequence analysis is displayed at all pertinent magnification levels by an appropriate coloring of a facsimile of the sequence called a **Representation**. Representations can be layered as concentric circles or stacked lines. Global views of various aspects of the sequence can be displayed simultaneously, giving immediate insights of inter-relationships sometimes encompassing thousands of base pairs. Zooming to various regions progressively discloses more analytical detail. Exact analysis over any designated region is available by using "instant functions." A list and descriptions of the Representations and their associated instant functions presently available are given in Table 1.

In the **neutral representation**, (Figure 1A), the sequence proper is represented by white tics at low magnification and appropriate colored tics or alphabetic characters at high magnification. The color coding is A (green), T or U (yellow), C (blue), and G (magenta) for nucleic acid
TABLE 1. GRAPHIC REPRESENTATIONS AND INSTANT FUNCTIONS

<table>
<thead>
<tr>
<th>REPRESENTATION</th>
<th>PROGRESSIVE DISCLOSURE (ZOOMING)</th>
<th>INSTANT FUNCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>Representation tics, white colored at low magnification. Color coded tics or alpha numeric charac-</td>
<td>Base and Amino Acid Coordinates. Molecular weight over designated range.</td>
</tr>
<tr>
<td></td>
<td>ters at high magnification.</td>
<td></td>
</tr>
<tr>
<td>Base composition</td>
<td>GC calculated over designated window and increment presented as color coded tics.</td>
<td>GC calculation given over designated range.</td>
</tr>
<tr>
<td>Open Reading Frame</td>
<td>Direction and stops indicated by color coded arrowheads. Potential starts by color coded dots.</td>
<td>Translation to amino acids over designated range.</td>
</tr>
<tr>
<td>Homology</td>
<td>Homology starts designated as color coded tics. Stringency indicated by proper color coding.</td>
<td>Exact matching and % h displayed for homology beginning with chosen tic.</td>
</tr>
<tr>
<td>Restriction Enzyme Mapping</td>
<td>Arrowheads with labels and/or color coded tics.</td>
<td>Listing of cuts sites arranged according to site no., or enzyme name, or fragment size.</td>
</tr>
</tbody>
</table>

sequences. For amino acid sequences the coloring is hydrophobic (red), basic (blue), neutral (white) and acidic (magenta). Ambiguous symbols are also allowable with default coloring.1

In the base composition representation (Figure 1A), percent GC is calculated over a designated window and displayed as appropriately colored tics. The default values are 0-40% GC (yellow), 40-50% GC (green), 50-60% GC (blue), and 60-100% GC (magenta). This coloring pattern corresponds to that of the individual bases at high magnification of the neutral representation.

The open reading frame representation presents a display of colored arrowheads for stop codons and colored dots as putative methionine starts. All six reading frames (reading in both directions) are given in the display (Figure 1B).

The homology representation allows user selection of colors to set stringencies on the display. The beginning base of each region found is designated by an appropriate color tic. Both a simple percent match or a variation of the Queen and Korn3 algorithm allowing looping and gapping is available for homology searches. Figure 1C shows low stringency homology searches to consensus promoter sequences.

The restriction enzyme representation presents chosen restriction sites as arrows with enzyme names or as colored tics coded to particular enzyme sites. The color tic representation is limited to a maximum of five enzymes. Figure 1A shows graphed sites for single cut enzymes contained in the TET region of plasmid pBR322.
FIGURE 1. pBR322, CAGE/GEM REPRESENTATIONS AT LOW MAGNIFICATION

FIGURE 1A INSIDE: Neutral Representation (white tic marks) with genetics overlay.


NEXT: Restriction Enzyme Representation of single cutters in TET region, cut site indicated by colored tics defined in lower left portion of screen.

FIGURE 1B Open Reading Frame Representation. Color arrows designate stop codons reading in direction indicated. Potential starts (ATG,) given as colored dots. Three reading frames in either direction are given.

FIGURE 1C INSIDE: Homology Representation low stringency homology to -10 promoter consensus TATAAT

OUTSIDE: Low stringency homology to -35 promoter consensus TTGACA. Stringency color coding, lower left corner of screen.

FIGURE 1D Zooming to region preceding TET gene to examine homologies.
OVERLAYS

Graphical depicting of genetic elements and features of the sequence is accomplished by overlays. In Figure 1 the genetics overlay for the plasmid pBR322 including the TET, AMP, and ORI regions is displayed. This overlay is part of the Battelle Northwest data base. Overlays of about 50 of the more common plasmids, viruses, eukaryotic and prokaryotic genetic elements are part of this data base.

The overlays most often contain information such as genetics that cannot be directly obtained from analysis of the sequence. However, any information about the sequence such as restriction sites, location of symmetrical regions and control regions can be put in an overlay. The system allows the creation of multiple overlays to a given sequence. The overlay capability provides an extremely powerful data management tool as features and important regions are immediately available and graphically representable.

Overlays can be queried from the system in a unique, compact way. The user can ask for information about the sequence from the menu and a priority window will open giving specific information about the displayed sequence. The labels on the overlay can also be queried by either typing in the...
FIGURE 3. FRAGMENT END MODIFICATIONS

appropriate label or pointing with the cursor. A priority window will open giving information about that region. Overlays can be queried at all magnification levels.

The EMBL and GENBANK DNA sequence, and the NBRF protein sequence data bases are available on CAGE/GEM™. The majority of these sequences have not as yet been abstracted into overlays. These data bases, however, can be queried and the sequence descriptions and headers will appear in a priority window.

CAGE/GEM™ provides a complete capability for the user to create Overlays for nucleic acids and amino acids sequences using a simple menu, single key strokes, and the cursor. The overlay items that can be selected include outlined boxes, dotted outlined boxes, filled boxes, arrows, and labels. Color size, width, and location of these items are also under user control.

Text description of any overlay label can be done within the CAGE/GEM™ system. When text modification is chosen, CAGE/GEM™ spawns the operating system text editor. Text can then be written that is associated with a particular label. This label can be queried at any time and a priority window opens displaying the textual information associated with the label.

Labels can also be placed in the upper left-hand area of the graphics display away from the sequence representation. Textual information can now be queried by typing the name of the label. The first item of this sort is
<table>
<thead>
<tr>
<th>SEQUENCE MODIFICATION</th>
<th>EFFECT ON OVERLAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELETION AND INSERTION</td>
<td>1. Label of total sequence changes to A-LABEL (red).</td>
</tr>
<tr>
<td></td>
<td>2. Specific overlay regions and regions directly altered change to A-LABEL (red) and red colored drawings.</td>
</tr>
<tr>
<td></td>
<td>3. Coordinates and file name of modified region indicated by inside automatic draw.</td>
</tr>
<tr>
<td></td>
<td>4. In cloning simulation mode, overlay identifies nature of ends.</td>
</tr>
</tbody>
</table>

**TRANSFORMATIONS**

No Effect on Overlays

1. DNA > RNA
2. Inverted complementation
3. One for one symbol exchange allowed by symbol ambiguity rules

Overlays Dropped

1. Inversion and complementation as separate commands
2. Editing in the alphanumeric rather than the graphic mode.

---

1 Symbol ambiguities are handled by lookup in a symbol equivalence table during the execution of analytical routines. Allowed exchanges include R (purine) for A and G and Y pyrimidine for T and C. Unknown bases are represented by the symbol X which is ignored by analytical routines. Undefined bases are represented by the symbol N which is totally ambiguous but analytically active.

reserved for information about the sequence. When the operating menu item, information about the sequence is chosen, the first description is displayed to the user.

As many as ten additional labels that are not part of the sequence representation overlay, with associated text, can be entered. Labels such as date, investigator, location, references, etc. with associated text can be extremely useful as part of the overlay database.

These labels can be colored by the operator. This flexibility allows labeling of information away from the graphics display of sequence and reduces clutter on the screen. For example, as shown in Figure 2, pC194 symmetry regions either taken from the literature or analytically determined from CAGE/GEM™ are represented by magenta colored boxes in the Battelle overlay database. The labels on the boxes for corresponding regions are numbered 1, 2, 3, 4, etc. and become clearly discernable only
FIGURE 4 SIMULATED CLONING: DELETION OF C GENE: BAMHI FRAGMENT FROM HEPATITIS B VIRUS GENOME.

FIGURE 4A Neutral representation, overlay and restriction enzyme representation for hepatitis B virus subtype AYW. Restriction sites determined from list of single cutters in TET region of pBR322 (Figure 1A) BAMHI (487, 1399, 2903), SPH1 (1235) sites found.

FIGURE 4B Fragment remaining after deletion of fragment between BAMHI sites at 1399 and 2903. Yellow and red label projecting inside of neutral representation designates a "break point." Regions of overlay effected by deletion change to red A-LABEL designation.

FIGURE 4C Zoom to high magnification at the "break point" showing the automatic overlay designation of the nature of the ends.

at intermediate or high magnification levels (Figure 3B). A description of each symmetrical region is given by querying these numbers. A magenta label, Symmetry regions, in the left-hand corner of the display, clearly identifies the nature of the magenta colored boxes.

Automatic Alteration of Overlays

One of the important features of the CAGE/GEM™ system is dynamic interaction involving representations and overlays at all levels of
FIGURE 5 SIMULATED CLONING: INSERTION OF C GENE BAMHI FRAGMENT INTO pBR322.BAMHI SITE.
The representations of the construct have been linearized for convenience at the ECORI site so all the layered representations fit on the graphics screen.

FIGURE 5A Insertion of fragment BAMBAM previously deleted from representation of hepatitis B virus (Figure 4) at BAMHI site in TET region of pBR322 while in simulated cloning mode. Overlay identifies (green box) automatically drawn below line. TET region (see Figure 1A) label change to A-TET (red). Box (TET) defining region changed to red and extended to accommodate insertion. Regions interrupted by deletion from hepatitis B virus (Figure 4) change to A-LABELS (red) and red outlines. C region (green fill) not changed.
Restriction enzyme representation verifies analytically the correctness of the insert. Blue tics verify restored BAMHI sites.
Open reading frame representation (red) verifies orientation of C gene. Notice, insert is out of phase with TET.

FIGURE 5B Insertion of fragment MABMAB, inverted complement of BAMBAM into TET BAMHI site. White labeled BAMHI sites verify correctness of insertion. Colored tics and labels are alternative ways of indicating restriction site locations found analytically. Open reading frame representation verifies reading orientation opposite of 5A for C region (magenta colored).

magnification. Sequence editing, including simulated cloning, not only modifies the sequence but also the overlays. CAGE/GEM™ automatically traces these alterations and modifies all appropriate labels to a red color prefaced by the letter A. Table 2 lists the alteration commands and their effect on the overlay.

Insertions and deletions are traced by automatically drawing of appropriate overlay items on the inside of a circular representation or on the bottom of a linear representation. The operator then has a complete graphical description of all editing that has been done to the sequence.
The space outside the circle in circular representations or above the line in linear representations is reserved for genetics and other features information. An insertion, deletion, or change in these regions alters them such that they no longer represent biologically the sequence as labeled. CAGE/GEM™ automatically alters the overlay graphic to a red color and changes the label to red A-LABEL (Figure 4).

**End Modification**

Fragments created by restriction enzyme site recognition in the cloning simulation mode have three possible end designations—5', 3', and blunt. The system remembers the end designation, and the sequence of the overhang (core). Figure 3 shows a decision-making flow chart for modification of these ends. End modification rules are based upon current techniques in the laboratory. The choice to modify a 5' end designation presents the user with three options. The SI cutback option simulates the use of SI nuclease in the laboratory. The core sequence is dropped, and the end designation is changed to blunt.

The T4, Klenow fill-in option emulates the use of T4 polymerase or Klenow fragment⁴ to fill in the 5' overhang. The program appends the core permanently to the fragment, and changes the end designation to blunt.

The Klenow partial fill-in option simulates the procedure of partial fill-in of N-M bases to a 5' overhang N bases long, where M≤N, using Klenow fragment and the proper nucleotide triphosphates.⁵ The user is given the option of appending 1.2.....N-1 bases of the core permanently to the fragment end. The option is then presented to further modify or not. The choice not to further modify keeps the 5' end designation and changes the core to the bases of the previous core that have not been appended permanently to the fragment end. The option to further modify drops the new core and changes the end designation to blunt. This simulates the removal of the remaining overhang with SI nuclease after partial fill-in.

There is only one modification option for 3' designated ends. This option simulates the use of the exonuclease activity of T4 polymerase or SI nuclease to remove the overhang.⁴ The program drops the core and redefines the end designation as blunt.

**CLONING SIMULATION EXAMPLE**

The neutral representation and overlay of hepatitis B subtype AYM virus⁶ is shown in Figure 4A. Of particular interest is the green fill-in region label C. Querying the label for this region (not shown) gives the
following information about the C gene..."Codes for the major nucleocapsid polypeptide P19 of 183 amino acids (hepatitis B core antigen-HBC Ag)."

A restriction enzyme representation of selected sites is also shown. In this example the current list of restriction enzymes was chosen by an analysis of pBR322 for single cutters and then further modifying this list to those that cut in the TET region. This list was used to map cut locations on the hepatitis genome. There are three BAMHI sites at coordinates 487, 1399, and 2903. The entire C region is flanked by two (1399-2903) BAMHI sites. In the Cloning Simulation Mode the region between these two BAMHI sites is selected for deletion and is saved to a file BAMBAM.

The altered hepatitis B viral genome after the deletion is shown in Figure 4B. In the laboratory this would be a linear fragment separable from the fragment containing the C region on an appropriate gel. In CAGE/GEM™ the fragment keeps a circular representation. A flag inside of the circle designates the name of the file containing the BAMBAM fragment. Figure 4C shows a zoom to high magnification of the region where the deletion was done. The overlay automatically designates the nature of the overhang, the overhang core sequence, and the enzyme name.

The point where the two designated ends are pseudo joined on the graphics representation is called a "break point." If more than one deletion is done while in the Cloning Simulation Mode, more than one "break point" is created. Insertions at "break points" are allowed only if corresponding ends are compatible.

Using "break points" and restriction sites in the Cloning Simulation Mode, any construct which can be done in the laboratory can be simulated on the computer. The user must keep in mind, however, that contiguous appearing sequences on the graphics display containing "break points" actually represent noncontiguous fragments in the laboratory.

When insertion is attempted at an appropriate restriction site or break point, the system first checks whether the end designations are compatible. A 5' end designation is compatible only with a 5' end designation, a 3' end designation is compatible only with a 3' end designation, and a blunt-end designation only with a blunt-end designation. The end designation is determined from the position of the cut site in the recognition sequence of the restriction enzyme used to create the end. If the ends are compatible, the system then checks the homology of the cores. If the cores are homologous, the splicing is completed and one core retained creating the appropriate splice junction.
The CAGE/GEM™ software works with only one DNA strand. Any fragment created in the Cloning Simulation Mode can be transformed to its inverted complement. When this operation is done, the fragment ends are appropriately adjusted. The creation of a fragment and its inverse complement allows insertions in more than one orientation.

The menu choices allow insertions only at restriction sites or "break points." If more than one site exists for a given enzyme, or more than one "break point" exists on the sequence, the user chooses the appropriate site with the cursor. The cursor need only be placed near the chosen site for a proper choice to be made. This allows complete cloning scenarios to take place at low magnification. The correctness of any procedure can be checked by analysis of the junctions for expected or created restriction sites using the restriction enzyme mapping representation or by zooming to high magnification and examining the actual base sequence at the junction points.

Figures 5A and 5B show the results of inserting the fragment BAMBAM into the BAMHI site of pBR322 and the fragment MABMAB which is the inverted complement of BAMBAM into the same site while in the Cloning Simulation Mode.

Figures 5A and 5B illustrate a number of important points about cloning simulation using CAGE/GEM™. The fragment from hepatitis B retains all of the overlay information associated with it in its original state. All those regions interrupted by the excision of the fragment now have the red (A-LABEL) designations to signify that an alteration has occurred. The C region proper was not altered in the excision and therefore remains unaltered on the new overlay.

In both Figures 5A and 5B a restriction enzyme representation is shown with only BAMHI on the current list. This verifies analytically that the BAMHI sites were restored for the cloned fragment in both possible orientations. Also, open reading frame representations are shown verifying the inversion of the reading frame of the C gene for the two possible cloning orientations.

THE CAGE/GEM™ SYSTEM

The CAGE/GEM™ software is written in Fortran for easy portability. It currently runs on VAXes and MICROVAXes utilizing VMS and on the HP9000 family using HP-VX. It requires a medium resolution color graphics terminal and a pointing device to manipulate the cursor.
It is impossible in this article to describe all of the features of the CAGE/GEM™ software. Additional information, including a quick reference guide to the menu structure, is available by contacting the principal author.

1To whom correspondence should be addressed

2Present address: United Nuclear Corp., Richland, WA 99352, USA

REFERENCES