

A Relational Model for Visualizing Codon Usage and Palindrome Distributions in Genome Sequences

Brian J. d'Auriol

Department of Computer Science
The University of Texas at El Paso
El Paso, TX, USA 79968
Email: dauriol@acm.org

Abstract—A relational model for Genome Sequence Visualization is proposed in this paper. A relational characterization of genome sequences in terms of bases, codons, and patterns such as close inversions is developed and described. The proposed model is applied to the visualization of codon usage distribution and close inversion distribution. These applications are accompanied by visualizations of a 269 base RNA molecule, a part of the Hepatitis C virus NS5 gene, Locus AY769711. This presentation illustrates the extent and flexibility of the approach.

Keywords: Genome sequence visualization, Relationship visualization, Codon usage distribution, Close inversion distribution.

I. INTRODUCTION

Genome sequencing is an important science effort nowadays. Interesting aspects of sequences include codon usage, identification of palindromes and other similar kinds of subsequences, the distribution of palindromes, and secondary structures in the molecule. Determining a sequence and its properties provides information about the genetic structure of the organism and leads to better understanding of the genetic processes.

Visualization is generally accepted as a powerful approach in enabling understanding of scientific phenomenon. In this case, visualizations on sequence scale, gene scale and genomic scale help to see the structure of DNA or RNA, and its constituent components and properties thereof.

The relational model proposed in this paper is based on the Advanced Relation Model (ARM) reported in our earlier work [1]–[3]. The ARM has been applied to program visualization: ARM 4 PV. The ARM 4 PV is really a two-part decoupled model specification: in the first part, the ARM

abstracts information as a relation hierarchy whereas in the second part, the 4 PV applies visualization techniques over the relation hierarchy. In a sense, genome sequencing is a biological equivalent of computer programs. The relational visualization approach is adapted so that the ARM is coupled with the specifics of genome sequencing. The two foci of this paper are: first, to present a relational hierarchy that is meaningful in genome sequencing and which reflects from the basis of the ARM, second, to present visualizations based on a relational abstraction.

This paper is organized as follows. Section II provides a brief overview of genome sequencing and which is given so the casual reader may be able to follow the proposed model. The relational model is proposed in Section III. Section IV presents several visualization examples. Conclusions and discussion are given in Section V.

II. BACKGROUND

Molecular Biology and Bioinformatics have seen dramatic rise of interests over the recent past. The subject area is quite extensive and a number of excellent books are available on the subject, for example [4], [5]. This section specifically concentrates on the RNA molecules and sequences of nucleotides contained therein. (Information is cited from [4], [5]).

A nucleotide is a three-component unit made up of a base, a sugar and a phosphate. Nucleotides are identified by a single letter that corresponds with the base: A – adenine, G – guanine, C – cytosine, U – uracil and T – thymine. A linear sequence of nucleotides is identified by a string of these letters,

for example, AGC means the three nucleotides in the order of adenine, guanine and cytosine. An RNA molecule is composed of sequences of A, G, C and U while a DNA molecule is composed of sequences of A, G, C and T.

Nucleotides may be grouped in a number of ways. Codons are groupings of three nucleotides that specify amino acids during protein synthesis. For example, the sequence GAAAGG would refer to the two amino acids, in order, glutamic acid (Glu) and arginine (Arg). An interesting issue is that decoding a sequence could begin at any index, in the above example, starting at the second nucleotide, the codon would be AAA which refers to the amino acid lysine (Lys). The same acid may be specified by more than one codon. A sequence of nucleotides may specify a gene.

Base-pairing occurs due to hydrogen bonding between the two bases adenine and thymine for DNA, between the two bases adenine and uracil for RNA, or between the two bases cytosine and guanine. Hence, in terms of sequences, this occurs between A-T, A-U and C-G, respectively.

One effect of base-pairings in RNA is the construction of secondary structures, including, pseudo-knots. Understanding secondary structures in RNA helps in understanding the RNA's functions [6]. A number of methods have been proposed to predict and locate such secondary structures; comments regarding these methods are made in [6].

Some definitions used in this paper follow. A *close inversion* is a sequence (p, s, p') where a subsequence p is separated by an arbitrary subsequence s from its complement subsequence p' , $|p| = |p'|$. A *palindrome* is a special case of a close inversion where $|s| = 0$. A *close repeat* is a sequence psp .

III. GENOME SEQUENCE VISUALIZATIONS

The primary datum in this model is a *relation* in a *relation hierarchy*. Individual relations are denoted by R_i^l where l refers to the relation's level in the hierarchy. Multi-relations are defined. In the following, the discussion limits to binary relations without loss in generality. The hierarchy consists of a set of connected relations: $\{R_{i_1}^1, R_{i_2}^2, \dots, R_{i_l}^l\}$ where $1 \leq i_j \leq n_j$ are integers that denote particular relations in the specified level; $1 \leq j \leq l$. Relations may be one of three types. *Semantic relations* bind both

semantics and properties to an object of interest: $R = (o, d, p)$ where o is an object of interest, d represents semantic information about o and p is a property set associated with o . *Constructive relations* are higher-order relations that combine the semantics and properties of two existing lower level relations: $R^l = (R^i, R^j, d_\alpha, p_\alpha)$ where $i, j < l$ and d_α, p_α are obtained by combining the respective information together. *Semantic Constructive relations* combine prior semantic information along with the combining of lower level information. The constructive type relations establish the means in which to *propagate* low-level abstractions about individual objects of interest to higher-level abstractions about groups of objects of interest.

The relational model is instantiated for genome sequences as follows. First, the selection of objects of interest, next, the determination of relational abstractions including semantics and property sets over these objects of interest, and last, propagation functions used to generate constructive relations. The decoupled nature of the relational model from the visualization models allows the various existing visualization models to be immediately applicable here.

A nucleotide is selected as the object of interest. In addition, for semantic constructive relations, both amino acids as well as the base, sugar and phosphate units of a nucleotide are also selected as objects of interest.

Figure 1 illustrates the general form of the relational abstraction. Although not used in the subsequent visualizations, Level 0 relations that abstract the base, sugar and phosphate molecules that make up a nucleotide are defined. Properties of these relations include the base, the molecular structure, bonding energy, and a short text description. In the figure, Level 0 relations are denoted as B, P and S. Level 1 relations are ternary relations that abstract the nucleotides. Relation properties include those of the previous level, and, in addition, the index of the nucleotide in the original sequence. In the figure, these relations are denoted as N. Level 2 relations abstract codons. These are also ternary relations that combine three nucleotides. Properties include the three nucleotide bases that describe the codon, the molecular structure, bonding information and a short text description. Due to multiple reading

frames, a set of Level 2 relations is correspondingly defined and denoted as C_1 , C_2 and C_3 respectively. The figure shows these sets. Level 3 relations abstract the amino acids coded by the codon with properties including the amino acid's name and molecular structure. In the figure, A denotes these relations. Level 4 relations are multi-relations that abstract subsequences of nucleotides, denoted by s, p or p' in the figure. Properties of these relations include a sequence designation (i.e., some identifying text), its starting index and length, together with the appropriate combinations of the molecular structure from the participating lower-level relations. Level 5 relations abstract the secondary structures of the sequence, denoted by n in the figure. Lastly, Level 6 relations abstract a gene, denoted by G in the figure. Level 3 relations may be connected to Level 4 relations, and in so doing, the necessary requirement of connecting Level 2 relations to Level 4 as well. This provides information about the sequence of amino acids coded by the respective subsequences and is useful for the subsequent visualization of codon usage. In the figure, these connections are shown by the dashed lines (for visual clarity, these dashed lines are connected to a different subsequence). The figure is linearly ordered in ascending levels for identification in the relation hierarchy. Its three dimensional structuring comes from the fact that codons and amino acids are non-necessary components of sequences, hence, can be considered to 'fill-in' the third dimensional (similar w.r.t. the pseudo-structures defined by Level 5 relations). In addition, the limited genome alphabet constrains the possible types of relations, for example, there are exactly five types of Level 0 relations, five types of Level 1 relations, 20 types of Level 2 relations and 20 types of Level 3.

The example in Figure 2 illustrates the construction of a relational hierarchy for a portion of the Hepatitis C virus NS5 gene. The sequence was taken from the NCBI Nucleotide database, Locus AY769711, and consists of 269 bases. In the example, a close inversion is shown at indices 1–6 and 253–258.

Propagation of some of the properties of the relations is natural, for example, the molecular structure of the three Level 0 relations that support a Level 1 relation are propagated by combining these

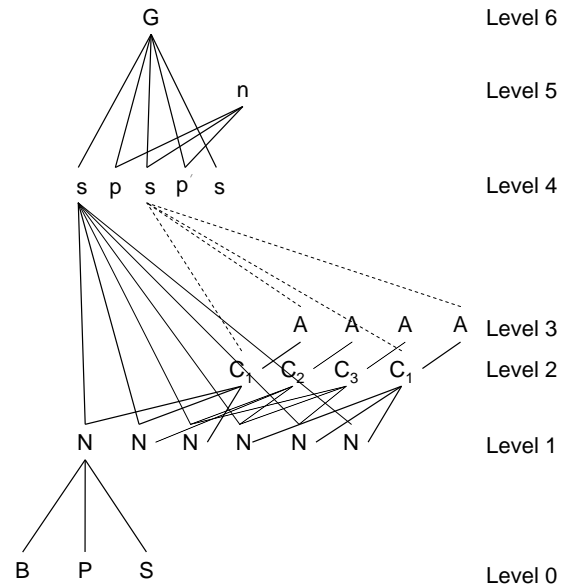


Fig. 1. Generic structure of the ARM 4 GSV relation hierarchy.

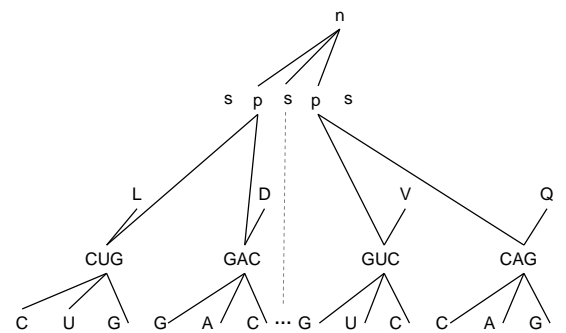


Fig. 2. Relational hierarchy illustration for a portion of the Hepatitis C virus NS5 gene.

into a single molecular structure for the designated nucleotide. Chemical rules may be applied to ensure the correct propagation.

IV. VISUALIZATION

The visualization of information in the relational hierarchy may be accomplished by any of the visualization models defined in the ARM 4 PV. Four such models are described in [1]–[3]. In keeping with the focus of this paper, one of these visualization models is selected.

The Conceptual Crown Visualization (CCV) model provides concept visualizations to facilitate the viewer's better understanding of the concepts inherent in the relational hierarchy. Here, concepts are defined by the relative abstractness of the rela-

tion, that is, the semantics of relations at a higher level define higher level concepts. There are two basic types of visualizations that are defined: a line structure and a space structure visualization. The former renders selected relations in the relation hierarchy as either single vertical lines (Level 1) or multiple piece-wise single point-connected lines (higher levels) whereas, the latter renders concepts as a convex hull of the participating lower-level relations.

Relations are graphed in the $x - y$ plane. The relation's level is mapped to the y -axis. The linear x -axis is ordinal, that is, represents an ordering of instances of the objects of interest. With respect to genome sequence visualization as proposed in this paper, nucleotides are mapped to the x -axis such that the index of the nucleotide is identified with its value on the axis. For perceptive reasons, the x -axis itself is mapped to a circle in the viewer's coordinates thereby creating a cylinder shaped visual object. This enables: (a) immersive visualization by allowing the viewpoint to be placed in the center of the circle, (b) a zooming operation by providing greater forefront focus while compressing the information in the peripheral area, and (c) greater information density by providing up to twice the amount of information displayed on the screen.

Visualizations of sequences have been reported in the literature. The VISTA toolset [7], [8] includes the VISTA browser to view visualizations of sequences and comparisons between sequences. Additional visualization models and programs are also available. Some plotting and visualization tools are provided as part of the European Molecular Biology Open Software Suite (EMBOSS) [9]. A visualization system called GenomePlot is reported in [10] where genomic scale visualizations are illustrated.

The subsequent visualizations are based on the 269 base sequence for a portion of the Hepatitis C virus NS5 gene, Locus AY769711 (i.e., the same dataset used in Figure 2). Although the primary emphasis with respect to these visualizations is to show the application of our proposed visualization model and further, we lack the biological background to assess the realism and usefulness of the visualization, we feel that these example visualizations illustrate the potential of this approach.

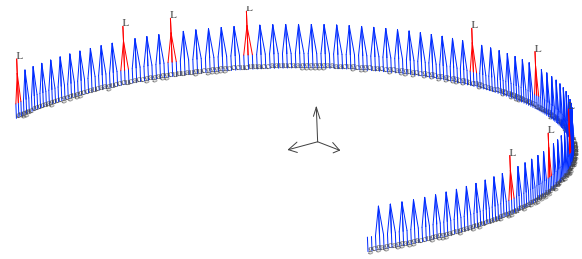


Fig. 3. Codon usage visualization: trinary relationships, angle factor of 12, Leucine coding highlighted in red, top front rotated view.

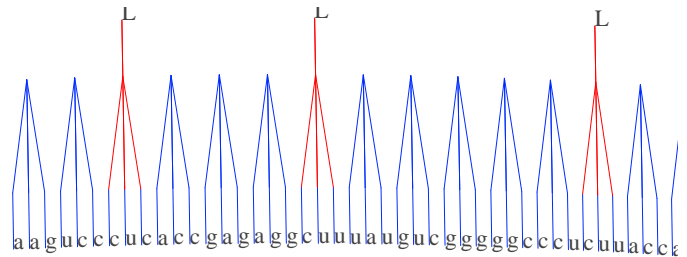


Fig. 4. Codon usage visualization: trinary relationships, angle factor of 12, Leucine coding highlighted in red, front zoomed view.

A prototype visualization system is implemented. AVS/Express is used to render the images and provide user interactive features.

Common aspects of the subsequent visualizations include: a circular x -axis at varying degrees of closure with the nucleotide sequence ordered left-to-right and (in most of the figures) displayed below a corresponding vertical line, various rotated views, a three dimensional axis located near the center of the image, and highlighting via colors. Specific aspects are described below.

A. Codon Usage Distribution

Various visualizations of the CCV model applied in the area of codon usage are presented in this section.

Figures 3 and 4 show the distribution of the Leucine (L) amino acid in the sequence. Figure 3 is a rotated tilted view while Figure 4 is a zoomed image of a portion of the sequence. The selected amino acid of interest is highlight in red. These figures allow all regions of the sequence to be visually inspected, including, both the begging and ending portions.

Figures 5 and 6 show a space-structure variation that highlights the Leucine acid. The angle is more closed than the earlier figures as well. Here, the

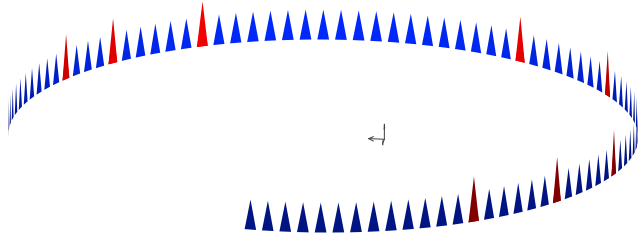


Fig. 5. Codon usage visualization: triangle represented relationships, angle factor of 16, Leucine coding highlighted in red, top front left-rotated view, unidirectional (interior) lighting.

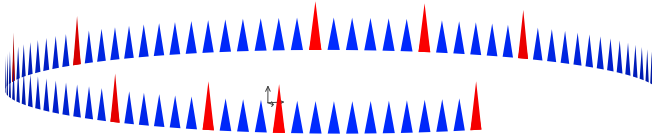


Fig. 6. Codon usage visualization: triangle represented relationships, angle factor of 16, Leucine coding highlighted in red, top side right-rotated view, bi-direction (interior and exterior) lighting.

relations are replaced by a corresponding triangle. Although difficult to see in the printed form, the first figure uses a unidirectional light source from the interior of the figure, thereby, the interior surface is rendered more brightly than the exterior surface. One advantage is that heightened depth-cues are provided to the user thereby making it easier for the user to maintain visual identification with the image. Figure 6 however, uses bi-directional lighting, thereby, both the interior and exterior surfaces are shown brightly.

Figure 7 is a perspective view of the distribution of Arginine (R) in the sequence. The selected acid is highlighted in red. The perspective view provides heightened immersive experiences to the user, that is, the image extends from in-front-of the user out along both sides. Peripheral vision is therefore enabled as well. In addition, this figure displays all of the amino acids in contrast with the previous two which displayed only the highlighted amino acid.

Figure 8 highlights the various codons that code for L. There are six such codons, of which, the following appear: two coded by ‘cug’ shown in cyan, four by ‘cuc’ shown in green, two by ‘cuu’ shown in red, and one by ‘uug’ shown in purple.

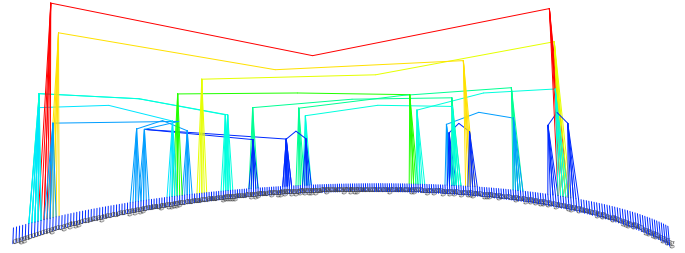


Fig. 9. Palindrome distribution visualization: palindrome length of 5 with no maximum gap between elements, no mismatches allowed, perspective front view, height and color coding ordered by gap between elements.

In addition, the sequence is displayed in a shallow curve thereby allowing better printed visualizations.

B. Close Inversion, Palindrome and Close Repeat Distribution

The modeling of close inversions, palindromes and close repeats also extends naturally from the relational hierarchy. A single Level 5 binary relation is used to connect p with p' or, for repeats, p with p . The figures shown in this section present these relations in various ways.

The EMBOSS ‘palindrome’ program is used to generate the close inversions displayed in this section. For convenience, the output of the EMBOSS software is filtered to provide the input into our visualization system.

Figures 9 and 10 show the distribution of close inversions of length five which occur anywhere in the sequence. Both color and height are used to highlight properties. In these figures, both color and height is based on $|s|$. Figure 9 uses perspective viewing while Figure 10 is a top view.

Figure 11 shows close inversions of lengths four and five for $|s| \leq 20$. Both color and height represent $|p|$.

A landscape visualization is an adaptation of the space structure variation of the CCV where the z -axis is used to plot additional information. Figure 12 shows a front-tilted-rotated view of close inversions of lengths four and five for $|s| \leq 20$. Color and height represent palindrome length, blue for $|p| = 4$ and red for $|p| = 5$.

The homogenous data abstraction of relations over the underlying domain data set of genome

