Published in final edited form as:
*Proc IEEE Int Conf Escience*. 2011 December 31; 2011: 15–22. doi:10.1109/eScience.2011.11.

# rCAD: A Novel Database Schema for the Comparative Analysis of RNA

**Stuart Ozer**,
Microsoft Corporation 1 Microsoft Way Redmond, WA 98052 stuarto@microsoft.com

**Kishore J. Doshi**,
Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas 78712 kjdoshi@gmail.com

**Weijia Xu**, and
Texas Advanced Computing Center The University of Texas at Austin, Austin, Texas 78712 xwj@tacc.utexas.edu

**Robin R. Gutell**
Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712 robin.gutell@mail.utexas.edu

## Abstract

Beyond its direct involvement in protein synthesis with mRNA, tRNA, and rRNA, RNA is now being appreciated for its significance in the overall metabolism and regulation of the cell. Comparative analysis has been very effective in the identification and characterization of RNA molecules, including the accurate prediction of their secondary structure. We are developing an integrative scalable data management and analysis system, the RNA Comparative Analysis Database (rCAD), implemented with SQL Server to support RNA comparative analysis. The platformagnostic database schema of rCAD captures the essential relationships between the different dimensions of information for RNA comparative analysis datasets. The rCAD implementation enables a variety of comparative analysis manipulations with multiple integrated data dimensions for advanced RNA comparative analysis workflows. In this paper, we describe details of the rCAD schema design and illustrate its usefulness with two usage scenarios.

### Keywords

Biological Database; RNA Sequence Analysis; Bioinformatics; Database Schema

## I. Introduction

A new perspective is now emerging in Biology: RNAs have a dominant role in the structure, function and regulation of the cell. Like DNA, it has a well-defined set of rules for nucleotide base pairing – A pairs with U, and G pairs with C. These consecutive and antiparallel base pairs form canonical helices. Like protein, RNA is capable of forming a three-dimensional structure composed of helices, hairpin, internal, and multi-stem loops, and other structural motifs, and like proteins, RNA is capable of catalyzing chemical reactions

[1-7]. It is now widely appreciated that RNA, as a precursor to DNA and proteins, was essential to the origin of life and to establish the mechanism for the association of a cell's genotype to its phenotype [8-11].

Comparative analysis, used effectively by Darwin to compare and contrast anatomical features of animals [12], has the potential to facilitate the identification and characterization of the RNAs primary and higher-order structure, and patterns of variation and conservation for the set of analyzed sequences [13]. The utilization of comparative analysis is based on a very important discovery in molecular biology - the same RNA secondary and tertiary structure can have different RNA sequences [14, 15]. The predicted secondary structure models are generally conserved for each of the specific RNAs. The accuracy from these comparative studies is impressive. Approximately 98% of the base pairs identified with comparative analysis in the three ribosomal RNAs – 16S, 23S, and 5S that contain more than 4,500 nucleotides are in the high resolution crystal structures [16]. In addition to the prediction of an entire RNA structure model, comparative analysis has identified RNA structural motifs, the basic building blocks of RNA structure and biases in the distribution of nucleotides in the ribosomal RNAs secondary structure. These include: unpaired adenosines in the rRNA secondary structures [17, 18], preponderance of tetraloops - hairpin loops with four nucleotides [19] and other types of irregular structural elements in the ribosomal RNA [20].

Successful application of comparative analysis to an RNA dataset requires an interactive workflow that includes the acquisition, management and analysis of large amounts of biological information divided into multiple dimensions: 1) sequences and the alignments of those sequences based on a common structure model, 2) evolutionary relationships between sequences and 3) higher-order structure and structural motifs. The iterative nature of the comparative analysis workflow ultimately improves the predicted structure model and the quality of the sequence alignment.

However, the comparative analysis workflow does not lend itself to the software pipeline architecture currently favored by most computational biology and bioinformatics applications [21, 22]. The pipeline architecture assumes that relevant data (*e.g.*, sequence alignment) is primarily stored in flat-files. Different programs load the data from flat-files into memory, perform analysis and output the results to a different flat-file. The pipeline is created by chaining different programs together. Raw data enters at one end of the pipeline and the value-added analyses exit at the other end in an automated fashion.

The comparative analysis workflow requires a semiautomated iterative approach. An important part of the comparative analysis is the interaction between biologists and the data. For example, as more sequences become available, the existing multiple sequence alignment are updated and curated to improve the comparative model. Similarly, while comparative analysis leads to new discoveries about structure and function of RNA sequences, these findings subsequently need to be integrated into the comparative model. Therefore, the full comparative analysis workflow cannot be supported by software tools using non-interactive, in-memory pipeline architectures.

We propose a different infrastructure to implement the comparative analysis workflow that includes support for the rapid increase in size and number of suitable RNA datasets and allows a biologist to perform manipulation and analysis of RNA datasets across the different dimensions (sequence, structure and evolutionary relationships). At the core of our infrastructure is an efficient data storage layer (Figure 1), the RNA Comparative Analysis Database (rCAD), which persists both the data entities within each dimension in their natural form and the biologically relevant relationships between entities. The schema is applicable for any relational database management system and currently implemented using a scalable, high performance enterprise relational database system, Microsoft SQL Server 2008. Two novel features of the rCAD schema are to efficiently support evolving data, such as changes in alignment and the integration of multiple dimensions of information.

The rCAD schema enables direct storage of the data entities for an RNA dataset in rCAD. The rCAD system implements a primary objective of the RNA structure alignment ontology: the integration of structural (and phylogenetic) information with an RNA sequence alignment to facilitate analysis and exploration [23]. A centralized database has advantages over the use of flat-files. The rCAD system eliminates the translation of data formats, and minimizes customized I/O and memory management routines. Thus it is more efficient for analyzing large scale data. In this paper, we describe details of rCAD schema and two examples of the benefits obtained with rCAD.

## II. Related Works

The rCAD project integrates data curation, access and analysis within a centralized relational database system. While rCAD is a unique application, several other projects are related to the management and analysis of RNA information.

A well-known data source for RNA information is the RNA family database (Rfam) [24]. Rfam maintains a collection of RNA sequence families. Each RNA family is represented by multiple sequence alignments and covariance models. Rfam is primarily a data provider service. This database is frequently updated and curated from various external data sources. The web interface of Rfam enables users to browse, search and retrieve RNA sequence alignments and corresponding covariation models by taxonomy or keywords. Users can download data in flat file format for further analysis.

There are other projects focused on specific RNA types. For example the Ribosomal Database Project (RDP) contains bacterial and archaeal small subunit rRNA alignments and some analysis of this data[25]. The alignments maintained in the RDP project are aligned with a probabilistic model system. The probability parameters are trained from a set of representative sequences. Recent updates on RDP introduced new features for browsing and searching RNA sequence information, including a genome browser and a taxonomy visualization tool. A new analytic feature of RDP is its pyrosequencing pipeline for the analysis of bacterial rRNA composition from environmental samples. Another similar web service is the GreenGenes project which enable users to align and analyze their own 16S rRNA sequences [21].

WATERS is a workflow tool that bundles a suite of publicly available 16S rDNA analysis tools to construct analysis pipeline [22]. It doesn't address data curation or include a dedicated data management component. For example, the alignment has to be computed on the fly or read from external sources into memory for processing.

The rCAD project maintains a comprehensive collection of RNA sequences and multiple sequence alignments for all types of rRNAs – 5S, 16S, and 23S, from the three primary forms of life: bacteria, archaea, and eukaryotes (including their nuclear, chloroplast, and mitochondrial encoded rRNAs) as well as other dimensions of information. The RNA sequences are retrieved daily from NCBI. The alignment data is curated and available at the Comparative RNA Web (CRW) Site[13]. The rCAD schema also maintains other dimensions of information and supports various analysis features in addition to storing raw sequences and their alignments.

Chado is a relational database schema for biological sequences published in recent literature [26]. Chado manages biological knowledge for a wide variety of organisms with information that can be directly or indirectly associated with genome sequences or the primary RNA and protein products encoded by a genome. Chado is based on existing ontology and interoperability between open access model organism databases or any biological databases that conform to this schema. The primary focus of Chado is to support genome analysis instead of comparative sequence analysis. Alignments and comparative features of sequences are typically stored as pairs. For example, features of a sequence alignment may include hits and high-scoring pairs of the alignment. Such schema is a generic way for storing comparative features and enables results from external analysis programs (e.g. Blast) to be stored in a unified way. However, it is less efficient for storing evolving multiple sequence alignments at large scale. The rCAD is designed to provide an integrative platform for sequence data management access and analysis. Therefore, the database schema used in rCAD is to facilitate common analysis tasks within the relational database system and is different from Chado.

## III. Schema Design and Implementations

The design of rCAD is motivated by several requirements: 1) persist the different elements of an RNA dataset (sequences and sequence alignments, structures, evolutionary relationships and alignments) that mimics their natural relationships; 2) efficiently support frequent updates to the biological data, and 3) provide a central data repository supporting the manipulation and efficient execution of computational algorithms involved in comparative analysis.

Our RNA dataset for comparative analysis is decomposed into three interrelated dimensions of information: sequence (metadata and nucleotides), structure (2-D) and evolutionary relationships. The organization of our dimensions of information is congruent with the proposed RNA structure ontology [23]. The unique capabilities of SQL Server allow for different comparative analysis algorithms and analyses to be executed on very large RNA datasets **in-process**, avoiding the need to export large amounts of data to separate analysis applications (Figure 1). Relationships between the data entities persisted in rCAD are coordinated with the biological relationships and are queried directly. RNA sequence alignments are stored as 2-D grids (sparse matrices) as proposed by the RNA structure alignment ontology [23]. SQL queries can be used to access any column of the sequence alignment, without loading the entire alignment into memory, using the indexing provided by the database. RNA secondary structure elements (*e.g.*, base pairs, helices, etc.) are directly related to the relevant columns of the sequence alignment. The sequences in the alignment are mapped onto the Tree of Life, and queries can be devised to retrieve specific fragments of the alignment containing sequences from specific taxonomy subsets.

The database schema is de-normalized into four compartments extracted from three major dimensions of an RNA dataset used for Comparative Analysis: Sequence Alignment, Sequence Metadata, Evolutionary Relationships and Structure Relationships (Figure-2).

### A. Sequence Metadata Compartment

Metadata describing RNA sequences is stored in the Sequence Metadata compartment (Figure 3). Types of metadata include external database identifiers (*e.g.*, Genbank Accession and revisions) stored in the **SequenceAccession** table, the sequence classification (*e.g.*, 16S rRNA) stored in **SequenceType** table and the organellar location within the cell (*e.g.*, Nucleus or Mitochondrion) stored in **CellLocationInfo** table. The design of the **SequenceAccession** table allows the rCAD instance to include revisions of a sequence introduced into Genbank.

### B. Evolutionary Relationships Compartment

Evolutionary relationships (taxonomy) in rCAD are stored in the Evolutionary Relationships compartment (Figure 4), and are obtained from the NCBI Taxonomy database [27]. The evolutionary relationships among sequences are stored in the **Taxonomy** table as a set of parent-child pairs (*TaxID*, *ParentTaxID*). The *TaxID* field is the primary key for the **Taxonomy**, **TaxonomyNames** and **TaxonomyNamesOrdered** tables and acts as a foreign key to link between the Sequence Metadata and the Evolutionary Relationships compartments. The scientific name of each taxon is stored in **TaxonomyNames** table while all other names are stored in the **AlternateNames** table for references. The **TaxonomyNamesOrdered** table stores the full lineage to any leaf in the *LineageName* field (*e.g.*, root/cellular organisms/Bacteria) in order to facilitate depth-based queries of the evolutionary relationships.

### C. Sequence Alignment Compartment

All biological sequences in rCAD are stored in the Sequence Alignment compartment, organized into one or more logical sequence alignments. A sequence alignment for a specific RNA molecule type can be described as a two-dimensional matrix. All sequences within the alignment are juxtaposed with one another to identify equivalent positions. The juxtaposition is accomplished through the addition and removal of gaps. Thus, the contents of any individual cell in the alignment matrix is either a nucleotide or a gap (‘-’). Unlike in other schema where alignments are stored at the sequence level, rCAD directly stores multiple sequence alignment at nucleotide level.

Each sequence alignment stored in rCAD is assigned a unique key, the *AlnID*, and is catalogued in the **Alignment** table (Figure 5). Members of a sequence alignment are stored in the **AlignmentSequence** table through a mapping of the *AlnID* to *SeqID*. The **AlignmentData** table maps each nucleotide in a sequence aligned to a specific column within an alignment as a row record. In **AlignmentData** table, each nucleotide is associated to a specific *PhysicalColumnNumber* which does not change unless that specific nucleotide's relative position within the sequence alignment is modified. The columns of any sequence alignment are identified and managed through the **AlignmentColumn** table.

This particular schema design for storing sequence alignment has two advantages, space saving and efficient updates when changing the alignment data.

First, in this schema, there is no need to store any gap values as they are inferred at query time. As the number of sequences and observed sequence variation in an alignment increases, the number of gaps also increases significantly. The increasing sequence variation is partially a result of insertions and deletions observed in specific branches of the Tree of Life. The result is that the total number of columns in the alignment can greatly exceed the number of nucleotides for any sequence within the alignment (Table 1). For example, for the small subunit Ribosomal RNA (16S rRNA) sequence alignment spanning the entire Tree of Life [13], the average per row ratio of gaps to nucleotides is 85% (Table 1). Thus, for any given row of the 16S rRNA sequence alignment, on average 85% of the columns have gaps. Therefore our database schema, which doesn't store gaps, results in a significant space reduction for storing large alignments and permits the rCAD database to scale and support very large (>$10^6$ rows × >$10^4$ columns) sequence alignments.

Secondly, the mapping of *PhysicalColumnNumber* to *LogicalColumnNumber* in the **AlignmentColumn** table supports efficient (from a database perspective) global operations on a sequence alignment by placing a layer of indirection between the physical storage in the database and the logical view of the sequence alignment. As new RNA datasets are added to rCAD, the **AlignmentData** table quickly becomes the largest table, even with a minimum of one entry per nucleotide for each sequence within each alignment in an rCAD instance. Without this columnNumber mapping, any further updates of an existing alignment, such as inserting or deleting a column will require updating a significant portion of the existing rows. For example, when the sequence alignment topology is modified by column insertion (Figure 6), the data structure only adds columns to the "end" of the alignment, regardless of where the column is logically inserted. The mapping of *PhysicalColumnNumber* to *LogicalColumnNumber* in the **AlignmentColumn** table is updated to reflect the topology change, and **no** rows in the **AlignmentData** table are modified (Figure 6). Without the column indirection in the data structure, global operations on large sequence alignments, such as column insertions, would be expensive operations, requiring a significant number of row updates to the **AlignmentData** table.

To simplify queries that obtain data from a sequence alignment stored in rCAD, two views are created: ***vAlignmentGrid*** and ***vAlignmentGridUngapped*** that return sequence alignment with or without gaps respectively. The ability to retrieve specific sub-grids (rows × columns) of a sequence alignment through SQL queries drives many of the advanced applications of the rCAD system such as structural statistics and evolutionary event counting. Figure 7 depicts the power of the rCAD system through an example query for retrieving a subset of a sequence alignment using ***vAlignmentGrid*** -- leveraging the integration between the sequence alignment and the evolutionary relationships among sequences within the alignment. First, the rows of the subset are identified by evolutionary relationships (Figure 7a). The example query in Figure 7 is designed to only retrieve rows that contain Bacterial sequences. A recursive query (SQL Server common table expression) is used to identify all rows in the sequence alignment that contain Bacterial sequences (Figure 7b). The subset of the alignment is then retrieved on a column by column basis (Figure 7c) for the rows that contain Bacterial sequences.

### D. Structure Relationships Compartment

The Structural Relationships compartment (Figure 8) stores secondary structure relationships. RNA secondary structure for a given sequence, at a minimum, is the set of base pairs between different positions of an RNA sequence. From the set of base pairs, other structural elements/extents can be inferred such as helices (consecutive base pairs) and loops (un-paired nucleotides) classified as hairpin, internal or multi-stem.

The **SecondaryStructureBasePairs** table holds the set of known or predicted base pairs for any RNA sequence in a sequence alignment. Each base pair is stored with two fields *FivePrimeElementSequenceIndex* and *ThreePrimeElementSequenceIndex*. The more complicated secondary structural elements are identified in the **SecondaryStructureExtents** table and split into their subelements depending on extent type enumerated in the **SecondaryStructureExtentTypes** table. For example, Helices and internal loop structural elements are split into two extents for their 5′ and 3′ halves. Multi-stem loops are split into *n* extents depending on the number of stems in the loop. An extent is defined by its first and last nucleotide (index), *ExtentStartIndex* and *ExtentEndIndex*. The entire structural element is given an identifier, *ExtentID*, and each extent of the structural element has its own *ExtentOrdinal*.

The mapping of a simple stem-loop RNA secondary structure into the Structural Relationships compartment is depicted in Figure 9. The stem-loop structure has two helices (**5′** 1-4, **3′** 32-35 & **5′** 10-18, **3′** 20-25) an internal loop (**5′** 5-9, **3′** 26-31) and a hairpin loop (16-19). All base pairs from the two helices are entered in the **SecondaryStructureBasePairs** table. The first helix (**5′** 14, **3′** 32-35) has two entries in **SecondaryStructureExtents** table (*ExtentID*, *ExtentOrdinal*, *ExtentStartIndex*, *ExtentEndIndex*, *ExtentTypeID*) where the **5′** half is (1, 1, 1, 4, 1) and the 3’ half is (1, 2, 32, 35, 1) (Figure 9). The internal loop (*ExtentID* 2) also has two extents for its **5′** and **3′** half, but the hairpin loop has only one extent (*ExtentID* 4). The extent model enables SQL queries to characterize the patterns of sequence variation for different elements of a structural motif, across the phylogenetic tree.

## IV. Use case examples

Algorithms that operate on sequence alignments benefit from rCAD streamlining and managing their access to the other dimensions of data. Thus the development of programs for rCAD is not encumbered with data that is already organized and cross-indexed. Two representative examples of applications utilizing the rCAD system are presented below: structural statistics and evolutionary event counting.

## A. Structural Statistics

Here, we present a common task of finding base pair frequencies. The frequency of different base pair types observed for a base pair in a reference secondary structure model has many applications in comparative analysis. This includes the prediction of base pairings in an RNAs higher-order structure with covariation analysis and the evaluation of the alignment accuracy for any new sequence within an existing sequence alignment [13, 28].

The selected base pair, labeled (**1**) in Figure 10, is projected across the sequence alignment by identifying the column ordinals associated with the **5′** and **3′** nucleotides from a reference row (Figure 10 middle). The rows of the sequence alignment included in the frequency tabulation are selected using evolutionary relationships. Using the two column ordinals for the 5′ and 3′ half of the base pair, a SQL query is defined to retrieve nucleotides from the sequence alignment, and a grouping statement is applied to determine frequencies of observed base pair types (Figure 10 bottom).

Different structural statistics have been developed as SQL queries or compiled applications. Visit the CRW Site (http://www.rna.ccbb.utexas.edu/SAE/2D) for more presentations on structural statistics. Computing statistical potentials for RNA folding programs is one specific application of our structural statistics[29, 30].

## B. Evolutionary Event Counting

Positions in a sequence alignment that have similar patterns of variation (covariation) are usually base paired in the RNAs higher-order structure [17, 31]. The traditional methods for identifying these positions with covariation determine the frequency of each base pair type for all of the sequences in two columns of the alignment [32, 33]. While these methods have been very effective in the accurate prediction of an RNAs higher-order structure [16], they do not explicitly determine the number of covariations during the evolution of the RNA. It has been known since at least 1983 that the evolutionary history for each putative base pair will enhance the accuracy and resolution of the covariation analysis [31]. We utilized the phylogenetic tree to identify the first tertiary interaction in the rRNA [34]. However the number of covariations based on the phylogenetic tree (called evolutionary event counting) was identified manually and thus a rigorous attempt to identify and quantify these phylogenetic events was not possible at that time. Attempts to identify phylogenetic events with computational statistical algorithms of the phylogenetic trees improved the sensitivity of the covariation methods [35, 36].

However, instead of using statistical methods, a more precise count of the number of changes and locations of the changes on the phylogenetic tree can be determined with rCAD. Our method builds an in-memory tree structure, populated with the nucleotides obtained from projecting the candidate positions across the sequence alignment (Figure 11). The in-memory tree structure is obtained from the Evolutionary Relationships compartment of rCAD, which is based on the NCBI Taxonomy database [27]. The application traverses the in-memory data structure using recursion to deduce the nucleotide composition of ancestor nodes and compute the evolutionary event counts. An earlier version of our method was published previously [37]. Our current event counting algorithm successfully identifies more true strong and weakly covariant positions with less false positives when compared to other methods (Shang, Xu, Ozer, & Gutell, manuscript in preparation).

## V. Conclusions

Presented here is a new database schema for the application of comparative analysis to RNA datasets. It is predicated on the integration and simultaneous manipulation of three primary dimensions of information - sequence and sequence alignments, associations between primary, secondary, and three-dimensional structure, and evolutionary relationships.

The rCAD system is significantly different from the traditional computational biology workflows where RNA sequence alignments are primarily stored in flat-files and manipulated in-memory by different programs. The rCAD schema directly persists and indexes RNA sequence alignments as sparse matrices.

Using a column indirection technique, rCAD is capable of performing global alignment operations including the insertion and deletion of columns efficiently for very large sequence alignments. The rCAD schema stores both evolutionary and secondary structure entities and relates them directly to the sequences and individual nucleotides in an RNA sequence alignment. Comparative sequence analysis algorithms can be implemented as declarative SQL queries. For example, structural statistics, relationships between sequences, different RNA structural elements and their occurrence on different parts of the phylogenetic tree are easily determined with rCAD.

The rCAD database implemented on the Microsoft SQL Server enables the development of more complicated algorithms as compiled applications that execute within the memory space of the database engine. These programs do not require significant amounts of custom programming for resource allocation, parallel processing, memory management and input/output optimization. The rCAD system places the responsibility for managing the data elements on the enterprise database engine, which is optimized for manipulating large quantities of information, and is scalable to support extremely large RNA datasets.

Readers who are interested in RNA sequence data can visit CRW Site (http://www.rna.ccbb.utexas.edu/) which uses rCAD for data management. Readers interested in building customized database can visit http://rcat.codeplex.com/ for available codes and utilities.

## Acknowledgments

This project has been funded by an External Research grant from Microsoft Research, National Institutes of Health (GM067317 and GM085337) and the Welch Foundation (#1427).

## References

[1]. Noller HF, Chaires JB. Functional modification of 16S ribosomal RNA by kethoxal. Proc. Natl. Acad. Sci. USA. Nov.1972 69:3115–8. [PubMed: 4564202]

[2]. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell. Nov.1982 31:147–57. [PubMed: 6297745]

[3]. Takada, C. Guerrier; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. Dec.1983 35:849–57. [PubMed: 6197186]

[4]. Noller HF, Hoffarth V, Zimniak L. Unusual resistance of peptidyl transferase to protein extraction procedures. Science. Jun 5.1992 256:1416–9. [PubMed: 1604315]

[5]. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science. Aug 11.2000 289:920–30. [PubMed: 10937990]

[6]. Wimberly BT, Brodersen DE, Clemons WM Jr. Warren R. J. Morgan, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. Structure of the 30S ribosomal subunit. Nature. Sep 21.2000 407:327–39. [PubMed: 11014182]

[7]. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF. Crystal structure of the ribosome at 5.5 A resolution. Science. May 4.2001 292:883–96. [PubMed: 11283358]

[8]. Woese CR, Fox GE. The concept of cellular evolution. J. Mol. Evol. Sep 20.1977 10:1–6. [PubMed: 903983]

[9]. Woese, CR. The genetic code; the molecular basis for genetic expression. Harper & Row; New York: 1967.

[10]. Orgel LE. Evolution of the genetic apparatus. J. Mol. Biol. Dec.1968 38:381–93. [PubMed: 5718557]

[11]. Crick FH. The origin of the genetic code. J. Mol. Biol. Dec.1968 38:367–79. [PubMed: 4887876]

[12]. Darwin, C. The origin of species. Barnes & Noble Classics; New York, NY: 2008.

[13]. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 2002; 3:2. [PubMed: 11869452]

[14]. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. Structure of a ribonucleic acid. Science. Mar 19.1965 147:1462–5. [PubMed: 14263761]

[15]. Fox GE, Woese CR. 5S RNA secondary structure. Nature. Aug 7.1975 256:505–7. [PubMed: 808733]

[16]. Gutell RR, Lee JC, Cannone JJ. The accuracy of ribosomal RNA comparative structure models. Curr. Opin. Struct. Biol. Jun.2002 12:301–10. [PubMed: 12127448]

[17]. Gutell RR, Weiser B, Woese CR, Noller HF. Comparative anatomy of 16-S-like ribosomal RNA, Prog. Nucleic Acid Res. Mol. Biol. 1985; 32:155–216.

[18]. Gutell RR, Cannone JJ, Shang Z, Du Y, Serra MJ. A story: unpaired adenosine bases in ribosomal RNAs. J. Mol. Biol. Dec 1.2000 304:335–54. [PubMed: 11090278]

[19]. Woese CR, Winker S, Gutell RR. Architecture of ribosomal RNA: constraints on the sequence of “tetra-loops”. Proc. Natl. Acad. Sci. USA. Nov.1990 87:8467–71. [PubMed: 2236056]

[20]. Gutell RR, Larsen N, Woese CR. Lessons from an evolving rRNA: 16S and 23 S rRNA structures from a comparative perspective. Microbiol. Rev. Mar.1994 58:10–26. [PubMed: 8177168]

[21]. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. Jul.2006 72:506–972. [PubMed: 16391085]

[22]. Hartman AL, Riddle S, McPhillips T, Ludascher B, Eisen JA. Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. BMC Bioinformatics. 2010; 11:317. [PubMed: 20540779]

[23]. Brown JW, Birmingham A, Griffiths PE, Jossinet F, Lafond R. Kachouri, Knight R, Lang BF, Leontis N, Steger G, Stombaugh J, Westhof E. The RNA structure alignment ontology. RNA. Sep.2009 15:1623–31. [PubMed: 19622678]

[24]. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Jones S. Griffiths, Eddy SR, Bateman A. Rfam: updates to the RNA families database. Nucleic Acids Res. Jan.2009 37:D136–40. [PubMed: 18953034]

[25]. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Mohideen A. S. Kulam-Syed, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. Jan.2009 37:D141–5. [PubMed: 19004872]

[26]. Mungall CJ, Emmert DB. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics. Jul 1.2007 23:i337–46. [PubMed: 17646315]

[27]. Benson DA, Mizrachi I. Karsch, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. Jan.2009 37:D26–31. [PubMed: 18940867]

[28]. Doshi, KJ. Ph.D. dissertation. The University of Texas at Austin. 2007.

[29]. Wu JC, Gardner DP, Ozer S, Gutell RR, Ren P. Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. J. Mol. Biol. Aug 28.2009 391:769–83. [PubMed: 19540243]

[30]. Gardner DP, Ren P, Ozer S, Gutell RR. Statistical Potentials for Hairpin and Internal Loops Improve the Accuracy of the Predicted RNA Structure. Journal of Molecular Biology. 2011 vol. in press.

[31]. Woese CR, Gutell R, Gupta R, Noller HF. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. Dec.1983 47:621–69. [PubMed: 6363901]

[32]. Chiu DK, Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. Comput. Appl. Biosci. Jul.1991 7:347–52. [PubMed: 1913217]

[33]. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res. Nov 11.1992 20:5785–95. [PubMed: 1454539]

[34]. Gutell RR, Noller HF, Woese CR. Higher order structure in ribosomal RNA. EMBO J. May.1986 5:1111–3. [PubMed: 3720727]

[35]. Dutheil J, Pupko T, Marie A. Jean, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol. Biol. Evol. Sep.2005 22:1919–28. [PubMed: 15944445]

[36]. Yeang CH, Darot JF, Noller HF, Haussler D. Detecting the coevolution of biosequences--an example of RNA interaction prediction. Mol. Biol. Evol. Sep.2007 24:2119–31. [PubMed: 17636042]

[37]. Xu W, Ozer S, Gutell RR. Covariant Evolutionary Event Analysis for Base Interaction Prediction Using a Relational Database Management System for RNA. Lecture Notes in Computer Science. May.2009 5566:200–216.

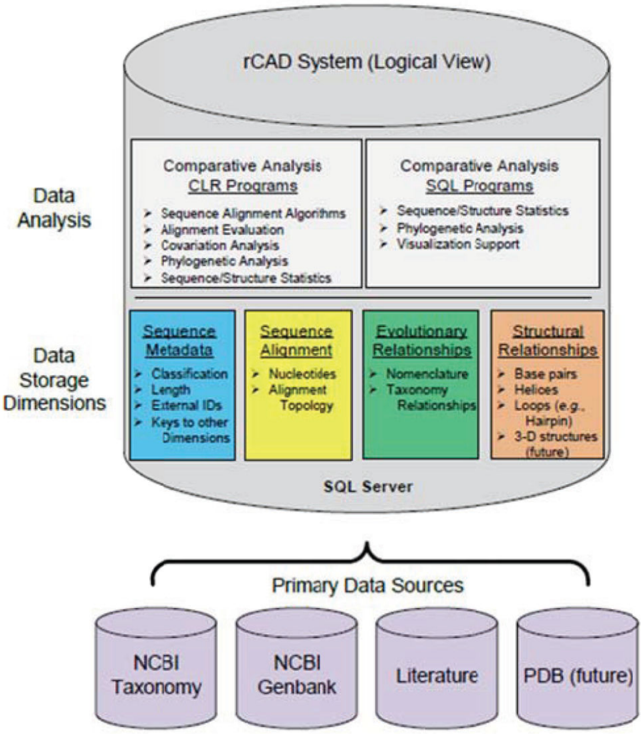


**Figure 1.**

Overview of rCAD system

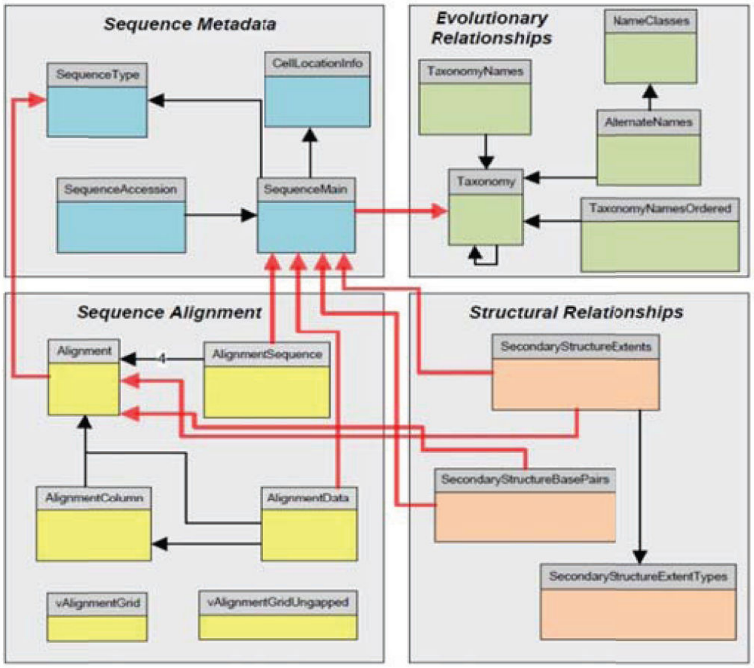
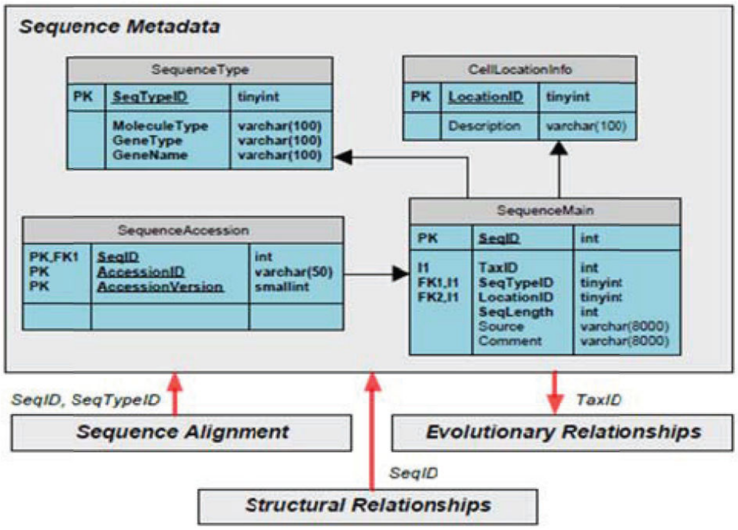


**Figure 2.**

rCAD database schema overview.



| SequenceType | | |
|---|---|---|
| PK | SeqTypeID | tinyint |
| | MoleculeType<br>GeneType<br>GeneName | varchar(100)<br>varchar(100)<br>varchar(100) |

| CellLocationInfo | | |
|---|---|---|
| PK | LocationID | tinyint |
| | Description | varchar(100) |

| SequenceAccession | | |
|---|---|---|
| PK,FK1<br>PK<br>PK | SeqID<br>AccessionID<br>AccessionVersion | int<br>varchar(50)<br>smallint |
| | | |

| SequenceMain | | |
|---|---|---|
| PK | SeqID | int |
| I1<br>FK1,I1<br>FK2,I1 | TaxID<br>SeqTypeID<br>LocationID<br>SeqLength<br>Source<br>Comment | int<br>tinyint<br>tinyint<br>int<br>varchar(8000)<br>varchar(8000) |

**Figure 3.**

Schema for sequence metadata.

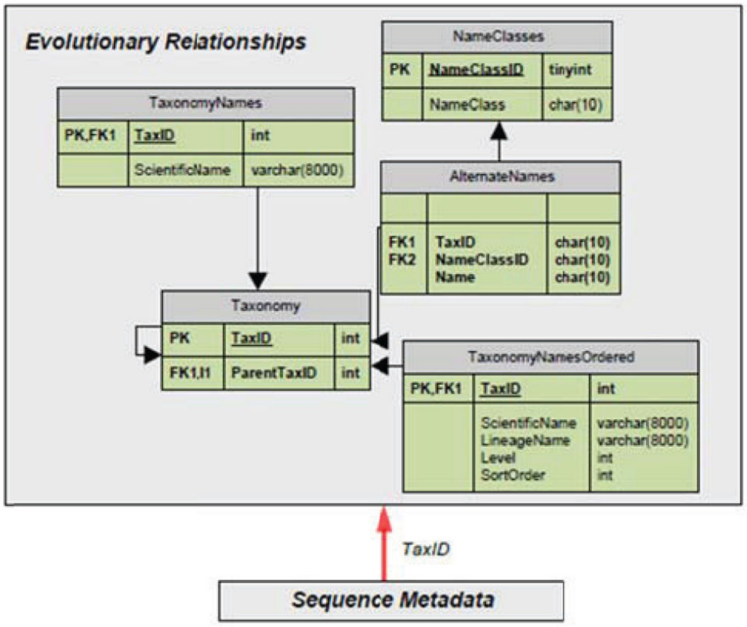
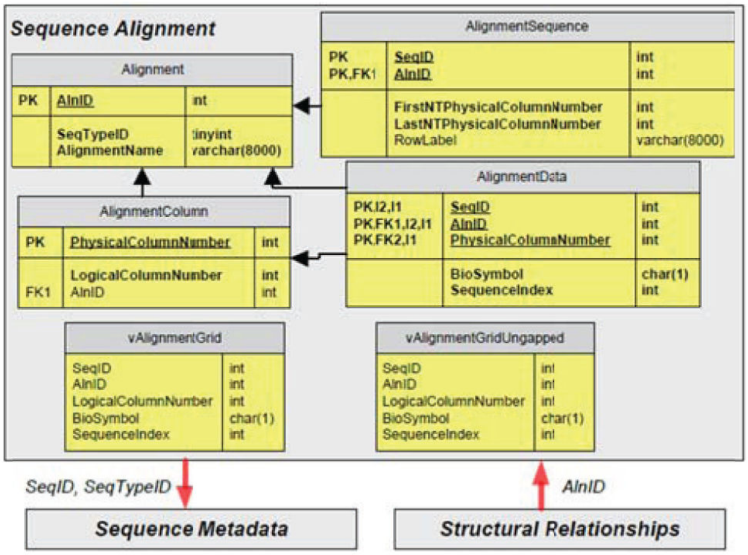


**Figure 4.**

Schema for evolutionary relationships



| Alignment | | |
|---|---|---|
| PK | AlnID | int |
| | SeqTypeID<br>AlignmentName | tinyint<br>varchar(8000) |

| AlignmentSequence | | |
|---|---|---|
| PK<br>PK,FK1 | SeqID<br>AlnID | int<br>int |
| | FirstNTPhysicalColumnNumber<br>LastNTPhysicalColumnNumber<br>RowLabel | int<br>int<br>varchar(8000) |

| AlignmentColumn | | |
|---|---|---|
| PK | PhysicalColumnNumber | int |
| <br>FK1 | LogicalColumnNumber<br>AlnID | int<br>int |

| AlignmentData | | |
|---|---|---|
| PK,I2,I1<br>PK,FK1,I2,I1<br>PK,FK2,I1 | SeqID<br>AlnID<br>PhysicalColumnNumber | int<br>int<br>int |
| | BioSymbol<br>SequenceIndex | char(1)<br>int |

| vAlignmentGrid | |
|---|---|
| SeqID<br>AlnID<br>LogicalColumnNumber<br>BioSymbol<br>SequenceIndex | int<br>int<br>int<br>char(1)<br>int |

| vAlignmentGridUngapped | |
|---|---|
| SeqID<br>AlnID<br>LogicalColumnNumber<br>BioSymbol<br>SequenceIndex | int<br>int<br>int<br>char(1)<br>int |

**Figure 5.**

Schema for sequence alignment

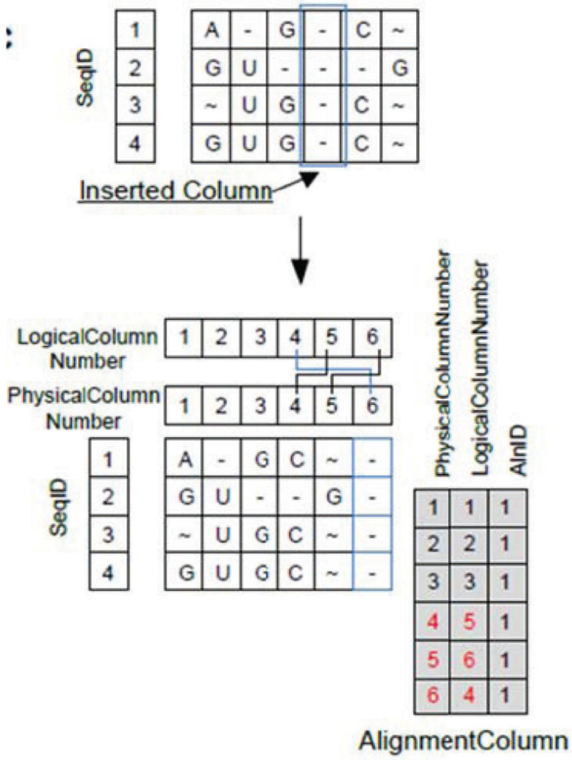
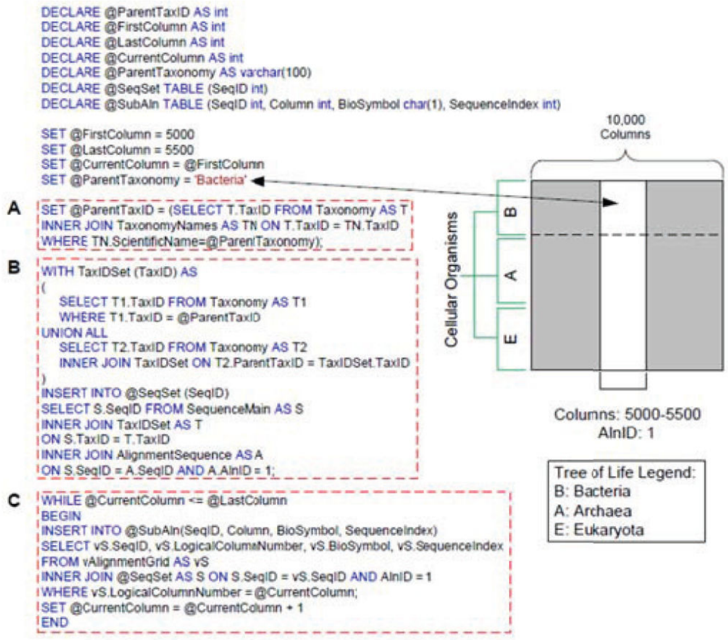


**Figure 6.**

an example of updating alignment data.



```
DECLARE @ParentTaxID AS int
DECLARE @FirstColumn AS int
DECLARE @LastColumn AS int
DECLARE @CurrentColumn AS int
DECLARE @ParentTaxonomy AS varchar(100)
DECLARE @SeqSet TABLE (SeqID int)
DECLARE @SubAln TABLE (SeqID int, Column int, BioSymbol char(1), SequenceIndex int)

SET @FirstColumn = 5000
SET @LastColumn = 5500
SET @CurrentColumn = @FirstColumn
SET @ParentTaxonomy = 'Bacteria'

SET @ParentTaxID = (SELECT T.TaxID FROM Taxonomy AS T
INNER JOIN TaxonomyNames AS TN ON T.TaxID = TN.TaxID
WHERE TN.ScientificName=@ParentTaxonomy);

WITH TaxIDSet (TaxID) AS
(
    SELECT T1.TaxID FROM Taxonomy AS T1
    WHERE T1.TaxID = @ParentTaxID
UNION ALL
    SELECT T2.TaxID FROM Taxonomy AS T2
    INNER JOIN TaxIDSet ON T2.ParentTaxID = TaxIDSet.TaxID
)
INSERT INTO @SeqSet (SeqID)
SELECT S.SeqID FROM SequenceMain AS S
INNER JOIN TaxIDSet AS T
ON S.TaxID = T.TaxID
INNER JOIN AlignmentSequence AS A
ON S.SeqID = A.SeqID AND A.AlnID = 1;

WHILE @CurrentColumn <= @LastColumn
BEGIN
INSERT INTO @SubAln(SeqID, Column, BioSymbol, SequenceIndex)
SELECT vS.SeqID, vS.LogicalColumnNumber, vS.BioSymbol, vS.SequenceIndex
FROM vAlignmentGrid AS vS
INNER JOIN @SeqSet AS S ON S.SeqID = vS.SeqID AND AlnID = 1
WHERE vS.LogicalColumnNumber = @CurrentColumn;
SET @CurrentColumn = @CurrentColumn + 1
END
```

**Figure 7.**

An exemplar SQL query for retrieving partial alignment

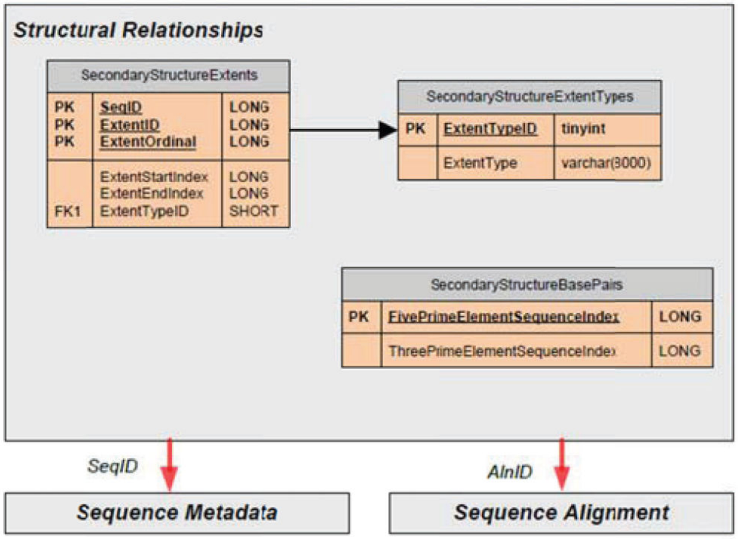
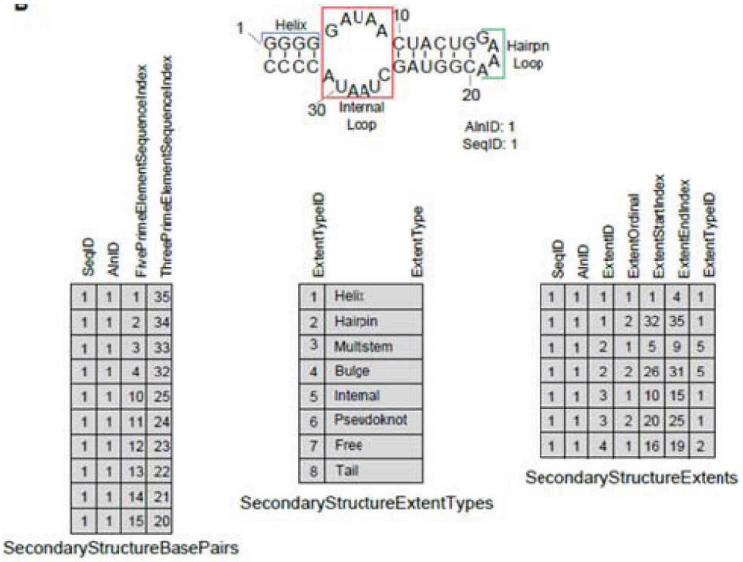


**Figure 8.**

Schema of Structure Relationships



| SeqID | AlnID | FivePrimeElementSequenceIndex | ThreePrimeElementSequenceIndex |
|---|---|---|---|
| 1 | 1 | 1 | 35 |
| 1 | 1 | 2 | 34 |
| 1 | 1 | 3 | 33 |
| 1 | 1 | 4 | 32 |
| 1 | 1 | 10 | 25 |
| 1 | 1 | 11 | 24 |
| 1 | 1 | 12 | 23 |
| 1 | 1 | 13 | 22 |
| 1 | 1 | 14 | 21 |
| 1 | 1 | 15 | 20 |

SecondaryStructureBasePairs

| ExtentTypeID | ExtentType |
|---|---|
| 1 | Helix |
| 2 | Hairpin |
| 3 | Multstem |
| 4 | Bulge |
| 5 | Internal |
| 6 | Pseudoknot |
| 7 | Free |
| 8 | Tail |

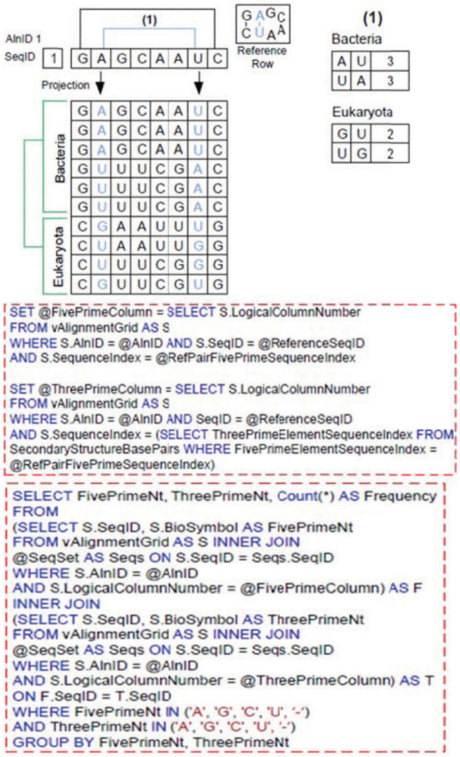
SecondaryStructureExtentTypes

| SeqID | AlnID | ExtentID | ExtentOrdinal | ExtentStartIndex | ExtentEndIndex | ExtentTypeID |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 4 | 1 |
| 1 | 1 | 1 | 2 | 32 | 35 | 1 |
| 1 | 1 | 2 | 1 | 5 | 9 | 5 |
| 1 | 1 | 2 | 2 | 26 | 31 | 5 |
| 1 | 1 | 3 | 1 | 10 | 15 | 1 |
| 1 | 1 | 3 | 2 | 20 | 25 | 1 |
| 1 | 1 | 4 | 1 | 16 | 19 | 2 |

SecondaryStructureExtents

**Figure 9.**

Example of mapping RNA structure into database



| SeqID | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | G | A | G | C | A | A | U | C |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Bacteria | G | A | G | C | A | A | U | C |
| | G | A | G | C | A | A | U | C |
| | G | A | G | C | A | A | U | C |
| | G | U | U | U | C | G | A | C |
| | G | U | U | U | C | G | A | C |
| | G | U | U | U | C | G | A | C |
| Eukaryota | C | G | A | A | U | U | U | G |
| | C | U | A | A | U | U | G | G |
| | C | U | U | U | C | G | G | G |
| | C | G | U | U | C | G | U | G |

(1)

Bacteria

| | | |
|---|---|---|
| A | U | 3 |
| U | A | 3 |

Eukaryota

| | | |
|---|---|---|
| G | U | 2 |
| U | G | 2 |

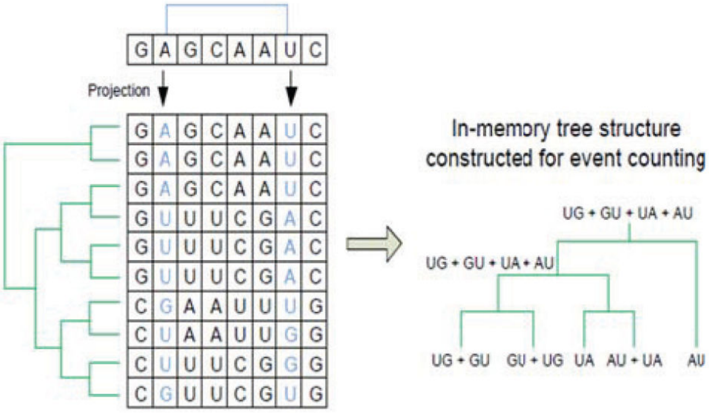
```
SET @FivePrimeColumn = SELECT S.LogicalColumnNumber
FROM vAlignmentGrid AS S
WHERE S.AlnID = @AlnID AND S.SeqID = @ReferenceSeqID
AND S.SequenceIndex = @RefPairFivePrimeSequenceIndex

SET @ThreePrimeColumn = SELECT S.LogicalColumnNumber
FROM vAlignmentGrid AS S
WHERE S.AlnID = @AlnID AND SeqID = @ReferenceSeqID
AND S.SequenceIndex = (SELECT ThreePrimeElementSequenceIndex FROM
SecondaryStructureBasePairs WHERE FivePrimeElementSequenceIndex =
@RefPairFivePrimeSequenceIndex)
```

```
SELECT FivePrimeNt, ThreePrimeNt, Count(*) AS Frequency
FROM
(SELECT S.SeqID, S.BioSymbol AS FivePrimeNt
FROM vAlignmentGrid AS S INNER JOIN
@SeqSet AS Seqs ON S.SeqID = Seqs.SeqID
WHERE S.AlnID = @AlnID
AND S.LogicalColumnNumber = @FivePrimeColumn) AS F
INNER JOIN
(SELECT S.SeqID, S.BioSymbol AS ThreePrimeNt
FROM vAlignmentGrid AS S INNER JOIN
@SeqSet AS Seqs ON S.SeqID = Seqs.SeqID
WHERE S.AlnID = @AlnID
AND S.LogicalColumnNumber = @ThreePrimeColumn) AS T
ON F.SeqID = T.SeqID
WHERE FivePrimeNt IN ('A', 'G', 'C', 'U', '-')
AND ThreePrimeNt IN ('A', 'G', 'C', 'U', '-')
GROUP BY FivePrimeNt, ThreePrimeNt
```

**Figure 10.**

Finding base pair frequency example.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| G | A | G | C | A | A | U | C |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| G | A | G | C | A | A | U | C |
| G | A | G | C | A | A | U | C |
| G | A | G | C | A | A | U | C |
| G | U | U | U | C | G | A | C |
| G | U | U | U | C | G | A | C |
| G | U | U | U | C | G | A | C |
| C | G | A | A | U | U | U | G |
| C | U | A | A | U | U | G | G |
| C | U | U | U | C | G | G | G |
| C | G | U | U | C | G | U | G |

**Figure 11.**

Evolutionary Event Counting example

## Table 1

Topology of Ribosomal RNA sequence alignments spanning the Tree of Life

| | Sequences | Total Alignment Columns | Max/Min Sequence Length | Avg. Gap Ratio (%) |
|---|---|---|---|---|
| 5S rRNA | 5355 | 404 | 242/14 | 73% ± 7% |
| 16S rRNA | 5762 | 9655 | 3316/506 | 85% ± 2% |
| 23S rRNA | 670 | 17047 | 5317/880 | 85% ± 5% |