An investigation into the structural basis for nucleic acid small molecule binding

by

José Miguel Cruz Toledo

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biology

Carleton University
Ottawa, Ontario

Abstract

The past 30 years of RNA research have seen a fundamental shift in our understanding of the biological roles that these macromolecules play. The sea change from our initial conception of RNA as a mere messenger of genetic information between DNA and proteins to our current understanding of RNA as a key player in various genetic and metabolic roles through their non-coding counterparts has motivated attempts to elucidate the structural underpinnings of their various biological functions. In this thesis, I describe research to develop and implement methods that formally represent nucleic acid structure, query and reason over their properties and computationally identify RNA structural motifs that are predictive of ligand binding. Chapter 1 presents the motivation, overall hypothesis and main objectives for this doctoral research, as well as a brief overview of the principles of nucleic acid structure and their representation using Semantic Web technologies. In Chapter 2, I present the RNA Knowledge Base (RKB), an instantiated ontology about RNA structure that provides machine understandable descriptions of nucleotide base pairs as observed in solved 3D structures. In Chapter 3, I describe Aptamer Base, a collaborative online knowledge base to describe aptamers and the details of the SELEX experiments that created them. In Chapter 4, I describe a methodology and implementation for the computational extraction of RNA motifs from a graph representation of their structures, and demonstrate that features of these motifs are predictive of ligand binding. In Chapter 5, I discuss future directions and present a summary of the contributions of this thesis.

Acknowledgements

I would like to express my gratitude towards my supervisor Dr. Michel Dumontier for his guidance and support during my PhD studies. As a member of Dr. Dumontier's lab, I was also given the opportunity to work alongside several talented and motivated fellow graduate and undergraduate students from whom I learned immensely. In particular I would like to acknowledge Natalia Villanueva-Rosales for your invaluable advice and contagious positive energy that kept me going. Special thanks also to Glen Newton, not only for patiently answering my countless questions, but also for your friendship. I am also grateful for the opportunity of having a scholarship as an international student at Carleton University.

To my parents, José Miguel and María Angélica, for having shown me the value of perseverance and hard work, and for everything that you have done to facilitate my education and personal development, for this and so much more, I am deeply thankful!

Lastly, I am greatly indebted to Alison Callahan, for having been a constant source of motivation and guidance throughout my doctorate.

Preface

1.1 Co-authorship statement

Of the three data chapters that comprise the bulk of this thesis, two exist as published work. In particular, Chapter 2 was published as a journal article in the Journal of Biomedical Semantics in a work entitled: "RKB: a Semantic Knowledge Base for RNA"[1]. Dr. Michel Dumontier and Dr. Francois Major conceived the research idea, Dr. Michel Dumontier and I generated the results and crafted the manuscript. Dr. Marc Parisien provided technical guidance. Chapter 3 was published as a journal article in Database: The Journal of Biological Databases and Curation in a work entitled: "Aptamer Base: a collaborative knowledgebase to describe aptamers and SELEX experiments"[1]. Dr. Michel Dumontier and I conceived of the idea. Dr. Maureen McKeague and myself created and refined the data model used in Aptamer Base with technical guidance from Dr. Maria DeRosa. Extraction of SELEX experimental data from scientific literature was done by a group of 8 MSc and PhD candidates trained and supervised by Dr. Mckeague and myself. I constructed and executed all data queries described in the results. I wrote the manuscript which was edited by Dr. McKeague, Dr. Dumontier and Dr. DeRosa and approved by all other co-authors. Finally, the research described in Chapter 4 was conceived by Dr. Michel Dumontier and myself. I created and executed all computer programs for this work and authored this chapter.

[1] Complete bibliographic information is provided at the beginning of each chapter.

1.2 Contribution of published works to thesis

For all Chapters herein presented both published and unpublished, in a section entitled "*Contribution to Thesis*" that follows each abstract, I provide a short description of each Chapter that situates the work in term of how it has contributed to my overall doctoral research objectives.

Table of Contents

List of Tables

List of Figures

List of Appendices

1 Chapter: **General Introduction**

1.1 Motivation for thesis

The central dogma of molecular biology describes the functional and informational relationships that have been thought to exist between ribonucleic acids (RNAs), deoxyribonucleic acids (DNAs) and proteins. Specifically, the central dogma posits that the first self-replicating systems relied on RNA as an information carrier and on proteins to synthesize RNA through enzymatic activities [1]. Notably, the central dogma of molecular biology as introduced by Crick [2] requires that two different types of macromolecules (*e.g.* nucleic acids and proteins) were synthesized by independent chemical reactions in the same place at the same time [3, 4]. In theory, it is more likely that a single type of molecule possessed the necessary functionality to synthesize itself and carry out its other key functions as informational vessels between generations. This is the basis of the "RNA World" hypothesis [4]. There are several lines of evidence that support the idea of RNA having been the first biocatalyst. RNA molecules have been found to exhibit self-replicating behaviours [5], and have the capability to catalyze the synthesis of proteins [6]. There is also evidence that RNA preceded the existence of DNA in that the biochemical reactions required for the synthesis of deoxyribonucleotides can be derived from those of ribonucleotide synthesis. Lastly, all living organisms use RNA catalysis to synthesize proteins and thus the Last Universal Common Ancestor (LUCA) must also have had this biochemical capability [4].

Starting with the RNA World hypothesis, our knowledge of the biological function of ribonucleic acids has been drastically transformed in the past four decades from the simple conception of RNA playing a role in protein synthesis as a passive information carrier between DNA and proteins [2, 7] to our current understanding of RNA as a key player in multiple aspects of molecular biology through the execution of genetic regulatory roles [8, 9]. In this thesis, the primary interest is in non-coding RNA (ncRNA) molecules. Non-coding RNAs are distinct from their protein encoding counterparts, in that these types of ribonucleotides function directly as regulatory elements of genes, rather than expressing mRNAs that encode proteins [10]. While in recent history our knowledge of ncRNAs was limited to biochemically abundant species and anecdotal discoveries, today there is an abundance of evidence for the importance of their functions in a wide array of cellular processes [11]. Moreover, the discovery of catalytic RNAs has expanded our knowledge of the possible biological functions of ribonucleic acids through a multitude of studies that consider both naturally occurring and artificially evolved and selected ribozymes. These initial descriptions of both aptamers [12], short artificial nucleic acids that fold into higher-ordered structures to form complexes with small molecules, proteins and cells, and of natural examples of non-coding small molecule binding ribonucleic genetic control elements called riboswitches [13, 14], have not only unveiled novel functionalities for nucleic acids that were previously assumed to only exist in proteins but have also opened new research opportunities regarding the numerous applications of small molecule binding nucleic acids.

One such application is the creation of bio-molecular recognition systems (bio-sensors) that would enable rapid multi-analyte sampling of food, air or drinking water contaminants that are capable of causing intoxication, diseases or chronic illness [15]. Antibody-based detection methodologies are still considered the standard in environmental, food and clinical analysis. The usage of antibodies in detection methods, however, is limited by the nature and synthesis of these protein receptors in that their generation necessarily requires either animals or cell lines [15-18]. The resulting antibodies can therefore only recognize targets under stringent physiological conditions, thus limiting the extent to which antibodies can be functionalized and applied. Moreover, the generation of antibodies against molecules that do not generate an immune response is difficult. Conversely, the *in vitro* selection of aptamers through the usage of Systematic Evolution of Ligands by EXponential enrichment (SELEX) not only enables researchers to obtain unlimited amounts of aptamer sequences of identical affinity but more importantly enables the creation of bio-sensors without the usage of animals. Furthermore, the wide range of molecules for which aptamers can be selected, including amino acids [19], antibiotics [20], organic dyes [21] , peptides [22], vitamins [23] and whole cells [24], make aptamers rival antibodies in their diagnostic potential. An additional advantage to the use of aptamers as biosensors is the recently emerging research in the rational design of the pools of random sequences from which aptamers are selected in SELEX experiments [25].

Given the disposition of certain nucleic acids (aptamers and riboswitches) to form complex 3D shapes and selectively recognize ligand targets with extremely high affinities

[26, 27], knowledge of the 3D structural features of nucleic acids is crucial for the elucidation of their molecular behavior in the cell. Experimental techniques currently used to determine the 3D structures of nucleic acids, such as X-Ray crystallography of single crystals of purified RNA or DNA, NMR spectroscopy and cryo-electron microscopy, are both time consuming and expensive. The cost of these experimental methods has resulted in a disparity between the number of known RNA sequences and three dimensional RNA structures [28]. It is this informational incongruity that is one of the primary motivations for developing a novel computational approach that would allow us to bridge the informational gap between the sequence, structure and the respective functional roles of previously uncharacterized nucleic acids.

1.2 Hypothesis

Binding of nucleic acids to small molecules has been shown to induce structural rearrangements that maximize favourable interactions between ligands and the macromolecule [11, 29-33]. The nature of this structural rearrangement depends upon both the structural composition of the nucleic acid and the type and features of the small molecule. We hypothesize that the repertoire of covalent and non-covalent inter-nucleotide residue interactions that define nucleic acid structures is predictive of their biochemical behaviour, and more specifically we postulate that an abstraction of nucleic acid structure that incorporates these features will enable the identification of functional units that govern nucleic acid-small molecule binding.

1.3 Objectives

The long term goal of my research is to facilitate the rational design of small molecule binding RNA aptamers. To this end, my research has three primary objectives: **Objective #1** is to collect and catalogue existing nucleic acid sequences and structures using state of the art biological information management. **Objective #2** is to carry out a quantitative analysis of known small molecule bound nucleic acids in terms of their sequence composition and recurring three dimensional structural motifs. Finally, **Objective #3** is to implement a machine learning method to predict small molecule-binding RNA structures.

1.4 Thesis Overview

In the remainder of this chapter, I provide a brief background into the characteristics of nucleic acids including their nomenclature, chemistry and structural properties. I also introduce state of the art technologies used throughout this research for knowledge representation in the life sciences specifically, Semantic Web standards for Linked Data and ontologies.

In **Chapter 2**, I provide a detailed description the RNA Knowledge Base (RKB), a knowledge base for RNA structure that provides a rich ontology for structural features of inter-nucleotide residue interactions, such as base pairs and base stacks and descriptions of sugar conformations. The RKB is populated with annotations of RNA structural data found in the Protein Data Bank [34] and enables querying and reasoning over RNA structural features.

In **Chapter 3**, I describe Aptamer Base, a collaboratively built knowledge base for aptamers and SELEX experiments. Aptamer Base provides rich descriptions of over 1800 RNA and DNA aptamers including the experimental details of the procedures used to generate them. Aptamer Base provides aptamer scientists with a community updatable and queryable data source for aptamer based knowledge.

In **Chapter 4** I describe a novel method for identifying recurring structural motifs extracted from symbolic annotations of three dimensional RNA structures deposited in PDB. Using these motifs two Support Vector Machine (SVM) classifiers were trained to predict the structures and motifs that participate in ligand binding.

Finally, **Chapter 5** provides a summary of the main contributions of my doctoral research, discusses the implications of my findings and explores future research directions.

1.5 On the representation of Nucleic Acid structure and function

In this section, I present a selective account of the structural and functional characteristics of nucleic acid molecules with special interest in their ribonucleic constituent. I first introduce the concept of nucleic acids by describing their basic components at the molecular level and their respective spatial arrangements, which are needed to define high-order inter and intra molecular RNA interactions used to describe RNA tertiary structure. I will then proceed to introduce the state of the art in secondary and tertiary structure representational methods for nucleic acids.

1.5.1 Structural characteristics of Nucleic Acids

1.5.1.1 What is a Nucleic Acid?

Nucleic acids are polymeric macromolecules composed of chains of monomeric repeating units called nucleotides or, more precisely, nucleotide residues, which are covalently linked through a backbone of alternating monosaccharide and phosphate units. Naturally occurring nucleic acids have an extraordinarily wide range of sizes, from 60-80 nucleotides, as in transfer RNAs (tRNAs), to over 10^8 nucleotide pairs in a single eukaryotic chromosome. Structurally, nucleotide residues can be characterized by their three component parts: (i) a five-membered sugar ring which is a ribose for RNA and deoxyribose for DNA (ii) one of four possible planar aromatic heterocyclic nucleobases attached to the sugar moiety through a β-oriented glycosidic bond, and (iii) the 3'-5' phosphodiester linkage joining the individual nucleoside units.

The four major naturally occurring nucleobases are adenine (A), guanine (G), cytosine (C) and uracil (U) which are in turn partitioned into two families: pyrimidines

(Y) cytosine, thymine and uracil, which are composed of a single pyrimidine ring ($C_4H_4N_2$), and the purines (R) adenine and guanine, which are constituted by a heterocyclic nucleobases composed of an imidazole ring ($C_3H_4N_2$) and a pyrimidine ring (where Uracil is replaced by Thymine in DNA). The major tautomeric forms of the nucleobases are shown in Figure 1.

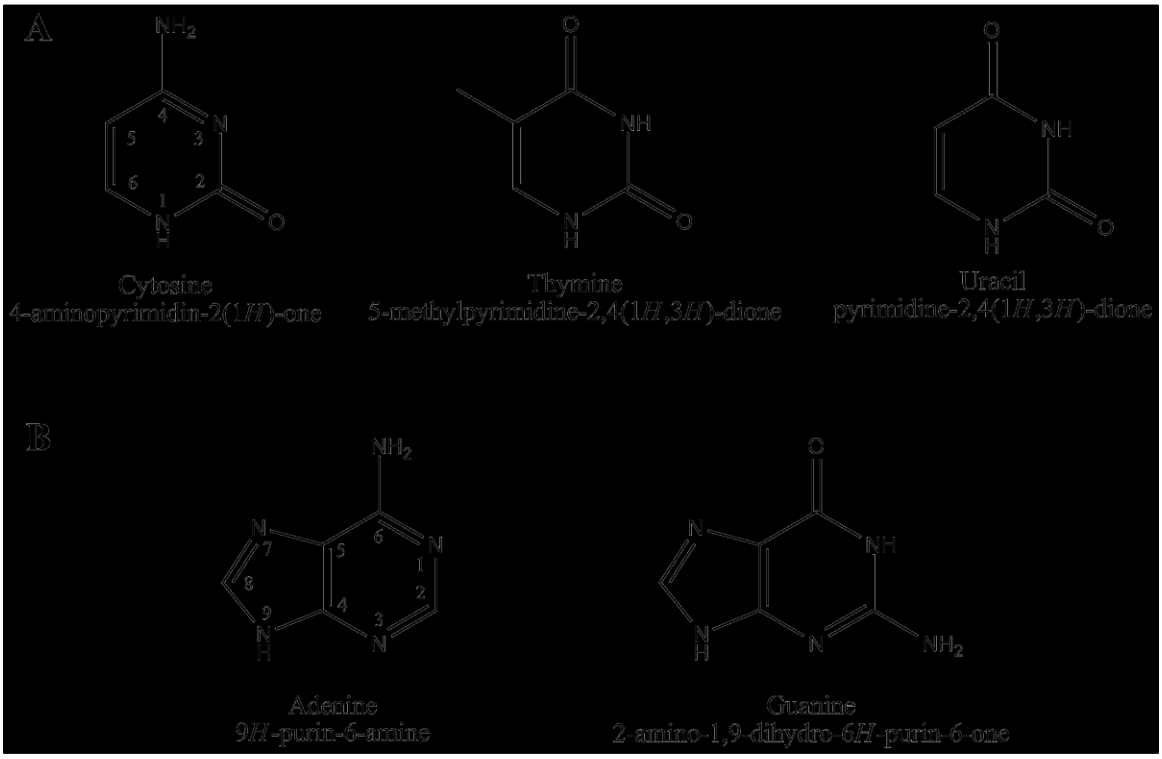


Figure 1 Structures of the five major purine and pyrimidine nucleobases in their dominant tautomeric forms. The International Union of Pure and Applied Chemistry (IUPAC) numbering and nomenclature for pyrimidines (A) and purines (B) is also shown.

Nucleosides are nucleobases that are covalently bonded to a sugar moiety. The sugar moiety (ribose or deoxyribose) is linked to the nucleobases through a glycosidic bond (described by torsion angle χ) between C1'-N9 for the case of purines and C1'-N1

for pyrimidine (Figure 2). The standard reference frame for assigning the absolute stereochemistry of other substituent groups on the sugar moiety of nucleic acids is defined such that when viewed end on, with the sugar ring oxygen atom O4' at the rear, the hydroxyl group at the 3' prime position is below the ring and the hydroxymethyl group at the 4' position is above. In naturally occurring nucleic acids, the glycosidic bond is usually found in its characteristic β stereochemistry, which is to say that the nucleobase is above the plane described by the ribose and therefore on the same face of the ribose plane as the C5' hydroxymethyl substituent [35, 36].

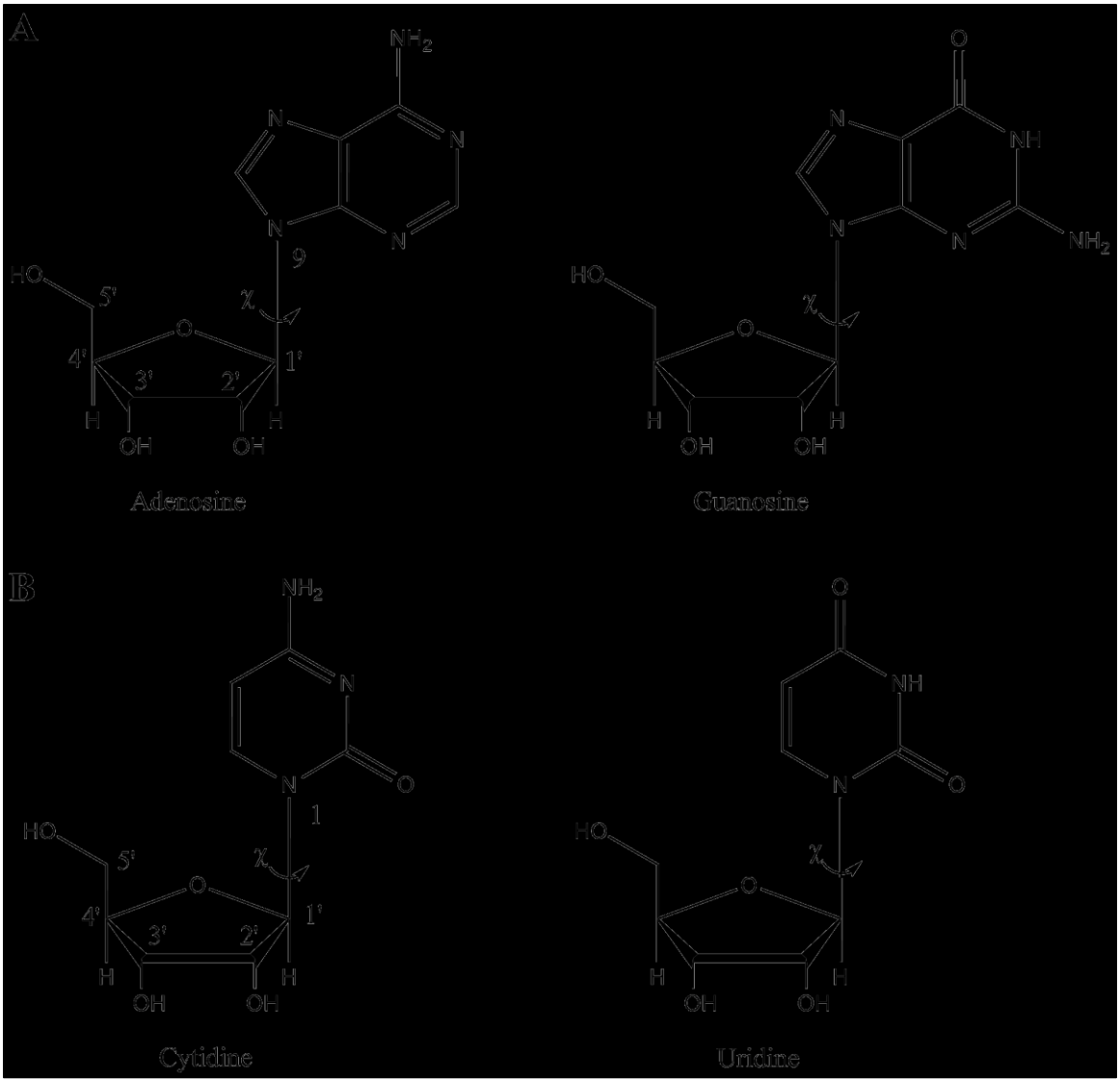


Figure 2 Stereochemistry of the four major naturally occurring α -pyrimidine (A) and α-purine (B) ribonucleosides. The nucleobases retain the same numbering system as shown in Figure 1 and the pentose carbons are numbered C1' through C5'.

Nucleotides are phosphate esters of nucleosides and these are the component parts of both ribonucleic acids (RNAs) and deoxyribonucleic acids (DNAs) where RNA and DNA molecules are composed primarily of ribonucleotides and 2'-deoxyribonucleotides respectively (Figure 3). The ribose and the phosphate moieties are the constituents of the

backbone and are linked through diester bonds C5'-O5' and C3'-O3'. The chain C3'-O3'-P-O5'-C5' from one ribose to another is referred to as the phosphodiester linkage that ties two nucleotide residues together.

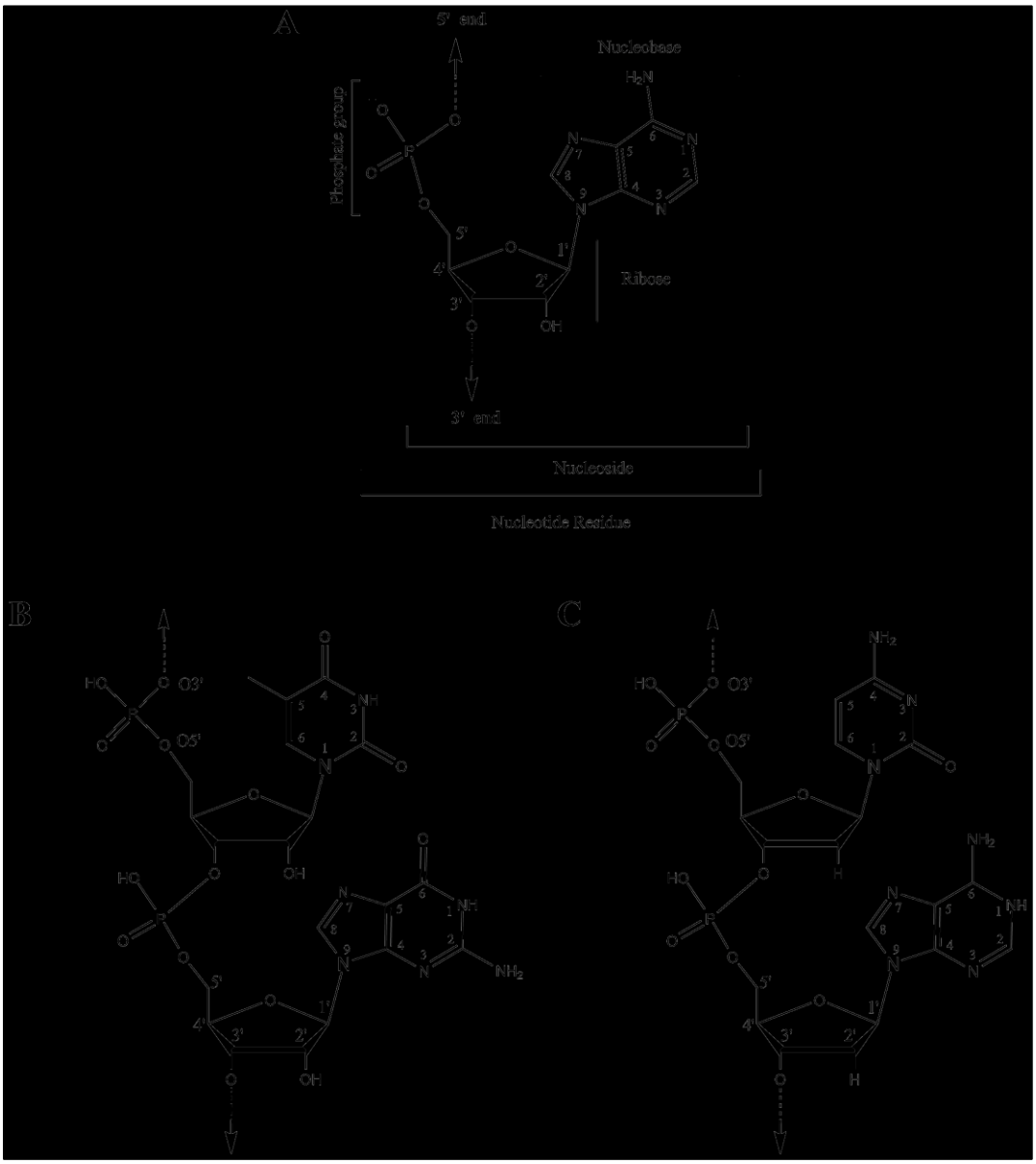


Figure 3 Sample structures of nucleosides and nucleotides. (A) Diagram of the structure of an adenosine monophosphate residue. The constituent nucleotide, nucleoside and nucleobases parts are depicted. (B) Diagram of the structure for a Uridine, Guanosine diribonucleotide showing the C3'-O3'-P-O5'-C5' backbone linkage between both nucleotide residues. (C) Diagram of the structure of a di-deoxyribonucleotide showing the backbone linkage between a Cytidine and an Adenine nucleotide residues.

1.5.1.2 Conformations of Nucleic Acids

Nucleic acids have rather compact shapes with several interactions between non-bonded atoms. Their molecular geometry can be accurately defined by the description of torsion angles α, β, γ, δ, ε and ζ in the phosphate backbone θ_0-θ_4 in the ribose ring and χ for the glycosidic bond (Figure 4). Given the innate interdependence of these torsional angles, the conformations of nucleic acids can be described in terms of four parameters: the sugar pucker, the anti-syn conformation of the glycosidic bond, base pair interactions and the shape of the phosphate sugar backbone.

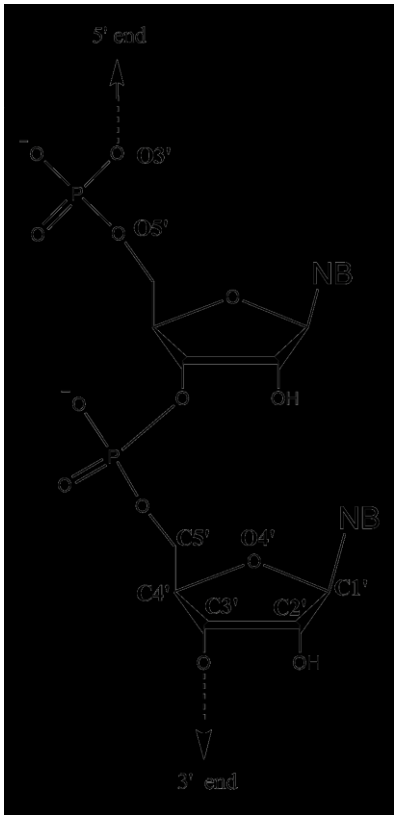


Figure 4 RNA backbone with six torsion angles across a dinucleotide residue. Nucleobases are represented as NB.

1.5.1.2.1 Sugar Puckers

The ribose rings in nucleic acids have the characteristic structural tendency of having carbon- 2' (C2') and C3' to be displaced from the plane described by C1', O4'and C4', thus forming a puckered or non–planar conformation [35, 37]. The relative rearrangement of the ring atoms arises as an effect of the non-covalent interactions that exist between the substituents at the four carbon atoms; energetically, the most stable conformation for the ring is that in which all substituents are as far apart as possible [38]. The puckering has been traditionally described by either a simple qualitative report of the conformation in terms of atoms deviating from ring co-planarity, or by specifying the torsion of its five endocyclic dihedral angles described within the monosaccharide (Figure 5).

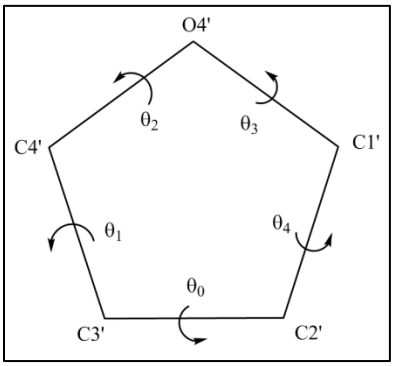


Figure 5 The five internal torsion angles of a ribose ring.

In principle, there is a continuum of interconvertible puckers, separated by energy barriers. These various puckers are produced by systematic changes in the ring torsion angles. A simple framework [39] for the understanding of the conformational characteristics of the sugar moieties in nucleic acids has been in use for the past three

decades. This framework introduced the concept of pseudorotation[2] to simplify the description of the conformation of the ribose ring in terms of only two variables: the pseudorotation phase *P* and angle amplitude τ_m [41]. *P* is defined in terms of the five torsion angles $\theta_0 - \theta_4$ and so the values of the pseudorotation phase angle therefore indicate which of the characteristics shapes the pucker is found: envelope or twist conformations (Figure 6) [39]. Within the continuum of possible values of *P*, the pucker is in the envelope form when only one of the five atoms of the ribose (C1', C2', C3', C4' or O4') is out of the plane formed by the four others. Conversely, the twist conformation is realized when two atoms are out of the plane formed by the remaining three, with these two on either side of the plane [36, 42, 43]. A commonly used pucker mode nomenclature scheme makes use of the direction of atomic displacement of the atoms in the ribose. This scheme classifies puckering atoms into one of two categories *endo* and *exo.* If the displaced atom is on the same side as the base and C4'-C5' bond, then the atom involved is termed *endo*. If it is on the opposite side of the plane, the atom is called *exo* [42].

[2] Pseudorotation is defined as a superimposable stereoisomerization that results in a structure that has been produced by a simple rotation of the initial molecule 40. McNaught, A.D., *Compendium of chemical terminology*. Vol. 1669. 1997: Blackwell Science Oxford.

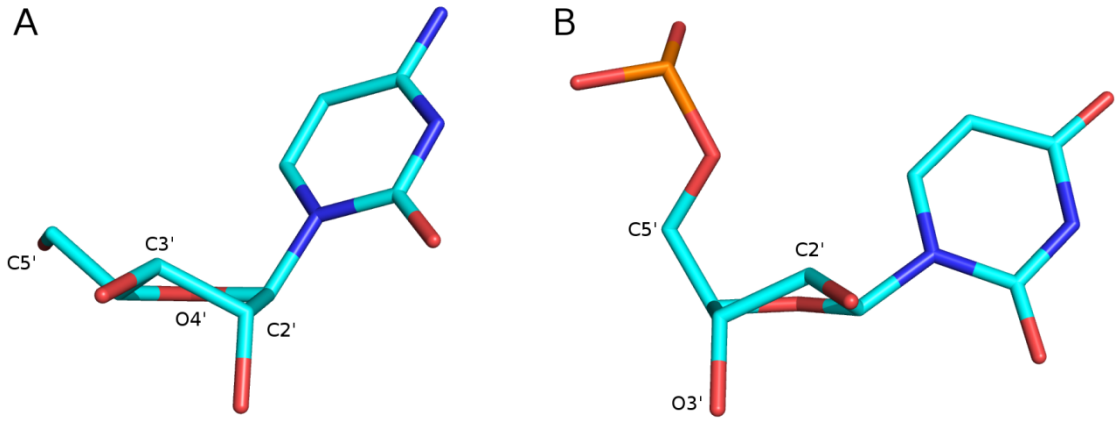


Figure 6 Twist A and Envelope B conformations of sugar puckers of a cytosine and a uracil nucleotide reside.

The plane described by the nucleobases is almost perpendicular to the plane described by the cognate ribose atoms C4'-O4'-C1' and approximately bisects the O4'-C1'-C2' angle (Figure 6) [36]. The angular relationship between these planes is best described by the glycosidic torsion angle χ and which, depending on its amplitude, describes the two nucleoside conformations: the *anti* and the *syn* conformations (Figure 7).

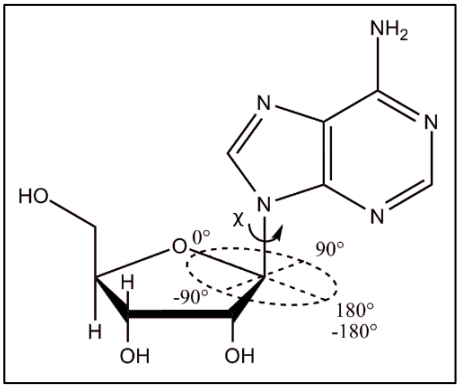


Figure 7 Glycosidic bond torsion. The anti conformation is realized when the nucleobases is pointing away from the ribose, conversely when the nucleobases is pointing towards the ribose the syn conformation is realized.

The *anti* conformation is realized when the plane of the nucleobases is oriented *away* from the ribose, that is, when $\chi = 180°$ the plane described by the nucleobase (N9-C8-N7 for purines and N1-C6-C5 for pyrimidines) is aligned with the O4'-C1' ribose bond in a direction away from the O4' atom. A nucleoside is mostly considered to be in an *anti* conformation for a range of χ angles, specifically a rotation of +/- 90°, for values of χ between -180° and -90° and between 90° and 180°. Conversely, the *syn* conformation is realized when $\chi = 0°$, that is, when the plane described by the nucleobase is aligned with the riboses' O4'-C1' bond in a direction towards O4. More generally, a nucleoside is considered to be in a *syn* conformation for χ angles ranging between +/- 90°, for values of χ that lie between either -90° and 0° or between 0° and 90° [42].

1.5.1.2.2 Base Stacking

Base stacking between the nucleobases of nucleotide residues stabilizes the double helical structure of both DNA and RNA. Base stacking interactions involve London dispersion forces [44], interactions between partial charges within the participating nucleobases [45]. More recent calculations suggest that the hydrophobic effect and van der Waals interactions drive the creation of base stacking interactions [47]. Stacking of nucleobase component parts of nucleic acids, occurs more frequently between adjacent nucleotide residues than between non-adjacent nucleotide residues, mostly in double-stranded helical regions [42].

In order to accurately characterize the type of base stack, a vector that is normal to the plane (or face) of the nucleobase (C6-N1-C2 for pyrimidines and C8-N9-C4 for purines) that is participating in a base stack can be defined so that any base in a classical A-form helix has their normal vectors oriented in the direction of the 3'-strand endpoint. In pyrimidines, the normal vector to the nucleobase is the rotational vector Ny obtained by a right-handed rotation, perpendicular to the nucleobase's face, from N1 to C6 around the pyrimidine ring.

1.5.1.2.3 Base Pairing

Base pair interactions are characterized by the formation of hydrogen bonds between exocyclic donor groups (mainly NH and NH_2) and acceptor groups (mainly CO and N) of the nucleobase component of nucleotide residues. More specifically, base pairs form when two nucleotide residues present their respective nucleobases in a roughly co-planar

orientation where the two bases present complementary edges. Complementary edges require hydrogen bond donors on the interacting edge of the first base pair to be juxtaposed with hydrogen acceptors on the edge of the second nucleobase or vice versa [48].

Base pairing interactions can be described in terms of both the edges that are presented by complementary arrangements of hydrogen donor and acceptor atoms found between the nucleobase constituents participating in the interaction (Adenosine, Cytosine, Guanosine, Thymine or Uracil) and the relative orientation of their respective glycosidic bonds (Figure 8).

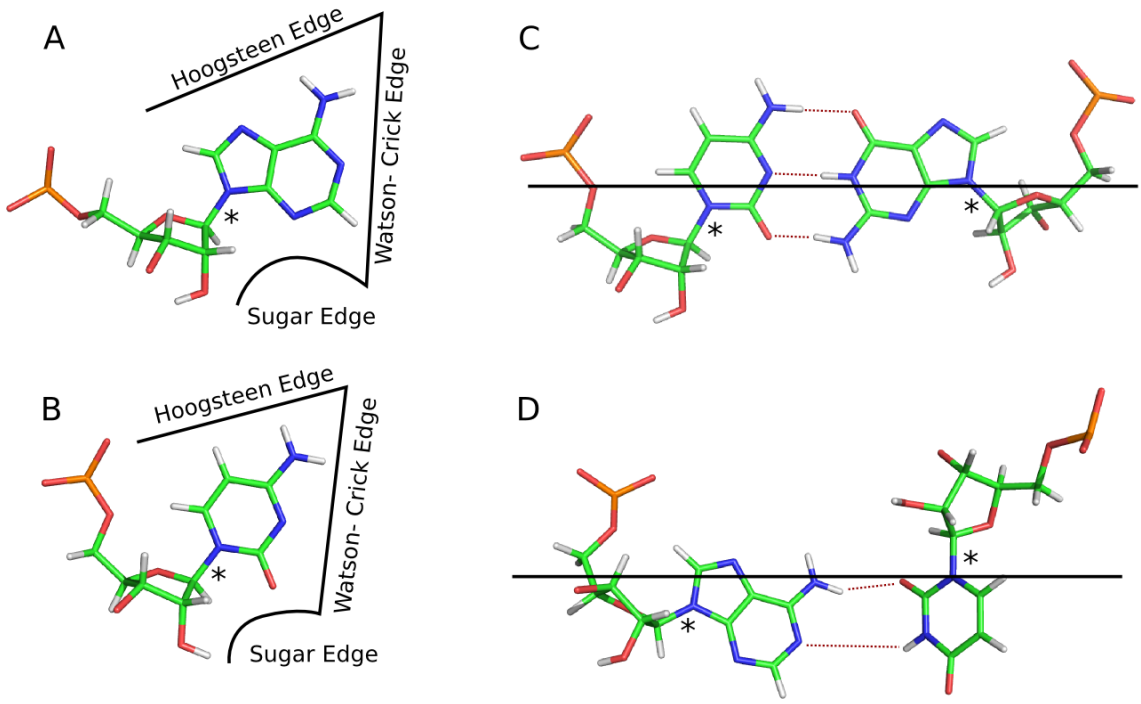


Figure 8 Chemical structures of representative purine and pyrimidine (A and B) nucleotide residues participating in base pair interactions (C and D). Purine (A - Adenosine) and pyrimidine (B - Cytidine) nucleotide residues are labeled with their respective Leontis-Westhof "edges", the Hoogsteen, Watson-Crick and Sugar edges. Cis (C) and trans (D) base pairing. These two kinds of base pairs can form which differ in the relative orientation of the glycosidic bonds (*). In (C) a cis WC-WC base pair is shown and (D) depicts a trans WC-WC base pair.

In order to facilitate accurate descriptions of base pairing interactions Leontis and Westhof [49] have devised a nomenclature scheme that enables their classification into geometrically similar classes. In this scheme, nucleobase interactions are described via their participating edges as one of 12 distinct geometric families of base pairs. Each family or class is described by indicating the interacting edges of the two bases, Watson-Crick (W or WC), Hoogsteen (H) or Sugar (S), and whether the interaction is cis (c) or trans (t) (Figure 8 C & D). The Leontis-Westhof nomenclature allows for distinguishing base pairs that involve the same nucleotide residues but present different edge-edge interactions. For example, consider the following LW descriptions of three different GA base pairs: a cis Watson-Crick/Hoogsteen Guanine-Adenine base pair (cWH), a cis Watson-Crick/Watson-Crick Guanine-Adenine (cWW) or a cis Watson-Crick/Hoogsteen Adenine-Guanine base pair (cWH), each of which would be assigned different LW classes.

1.6 Introduction to the Semantic Web: Data Integration and Knowledge Representation for Life Sciences

A key activity for life scientists in this post "-omics" age is searching for and integrating biological data obtained from a variety of sources and formats. The advent of the World Wide Web (WWW) has given scientists the ability to easily disseminate and share their work primarily in three ways: i) by provisioning web based interfaces, ii) by providing downloadable dumps of entire databases and more recently, iii) through application programmatic interfaces (APIs) that enable direct access to the data. Web based interfaces are often well designed and highly accessed but they are limited to the inherent

capabilities of the HyperText Markup Language (HTML)[3], which can only effectively make links between documents and is agnostic to the nature of the data contained in the documents being linked. Databases enable easy storage and fast retrieval of information but are typically not tailored for integration across databases and providers, essentially making biological databases standalone silos of information which must be further processed to discover links between them. Lastly, while APIs enable programmatic access to databases, the main two protocols used for this purpose (SOAP[4] and REST [50]) still do not provide any explicit semantics about the data being transferred. While the first 25 years of the WWW have seen a deluge of new databases and online based tools that have significantly improved the dissemination of new discoveries and ideas in the life sciences domain, the existing Web links documents and not data. This is unintentionally increasing the integration barrier between biological data sources thereby hampering our ability to reuse data to discover novel biologically significant relationships.

To address the limitations of the Web of documents, the Semantic Web (SW) [51] was conceived as a collection of standards and best practices for exposing data and its meaning over the WWW for applications to consume [52]. On the SW the semantics of both the information and services available on the Web are defined, thus making it possible for programs to interpret the intended meaning, or semantics of a request and process data or other Web services accordingly [53]. The SW aims to make information self-describing through the adoption of several standards: the Resource Description

[3] http://www.w3.org/TR/html401/
[4] http://www.w3.org/TR/soap/

Framework (RDF) [54][5], RDF Schema (RDFS) [55][6] and the Ontology Web Language (OWL) [56][7]. These technologies are organized in the Semantic Web Stack (Figure 9).

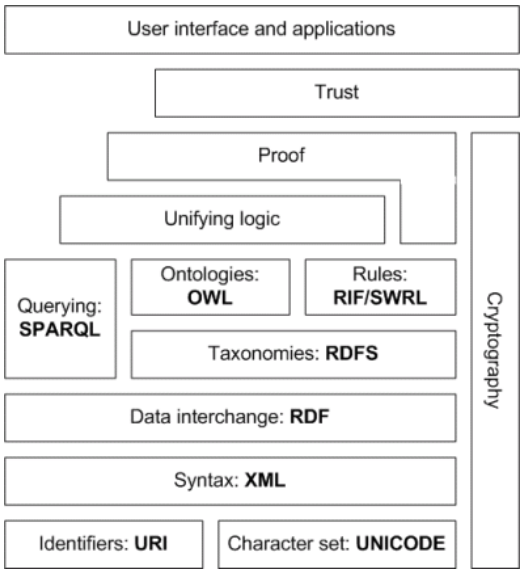


Figure 9 The Semantic Web technology stack[8].

RDF provides a flexible and expressive graph based data model for describing *relationships* between *things* including Web documents, digital content and also tangible real world entities, qualities and relations [57]. With RDF, every *thing* and the *relationships* between them are uniquely identifiable via Unique Resource Identifiers (URIs) [58]. URIs are strings of characters that used to identify any resource of the web. The HyperText Transfer Protocol (HTTP) [59], an application protocol for retrieving content on the Web, is used to make URIs and the content they refer to accessible on the

[5] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[6] http://www.w3.org/TR/2004/REC-rdf-mt-20040210/
[7] http://www.w3.org/TR/owl-features/
[8] Source: https://en.wikipedia.org/wiki/File:Semantic-web-stack.png

Web. RDF is at the core of a collection of standards and best practices for sharing data across the WWW as defined by Berners-Lee [60], known as *Linked Data*. Linked Data is characterized by the following principles: i) ise URIs to name things, ii) use HTTP URIs so that those names can be resolved on the web, iii) provide useful information in RDF about all named things and iv) link every named thing to others.

The basic unit of representation in RDF is the *triple*, a statement connecting a subject and an object through a relation or predicate. For example, to express the statement "Gene encodes Protein" in RDF the triple would consist of "Gene" as subject, "encodes" as predicate and "Protein" as object. The subject of one triple may be the object of another and as such one may construct highly connected graphs of varying size and complexity. One of the primary strengths of this data model is that it can be used to represent practically any kind of information, independent of its domain, thus making it well suited to become the *lingua franca* for data communication across both applications and computing platforms.

Following Linked Data principles, the triple described above would take the following form:

subject: <http://example.org/myGene>

predicate: <http://example.org/encodes>

object: <http://example.org/myProtein>

Each of these URIs would resolve on the Web to an RDF document describing the properties of the things that they represent and would also be linked to other things that they are related to. For example, <http://example.org/myProtein> could be linked to its biological function.

Linked Data resources can be further specified as instances or classes. In continuing with our example, we can further define the conceptualization of a protein by making the following statements:

<http://example.org/myProtein> rdf:type <http://example.org/Protein>

<http://example.org/Protein> rdf:type owl:Class

Here, we have used the RDF relation rdf:type [61][9] property to state that the "myProtein" resource is an instance of a "Protein" class as defined in OWL[10]. Through the usage of SW vocabularies, we can define machine understandable conceptualizations of any domain of knowledge by creating OWL ontologies. Specifically, an OWL ontology enables granular descriptions of classes as well as object and datatype properties that link instances to other instances or literals such as strings and numbers. Classes may be formally defined using axioms that specify the necessary and sufficient conditions to be

[9] Here we use the CURIE syntax for expressing shortened versions of URIs. In this case we have abbreviated http://www.w3.org/1999/02/22-rdf-syntax-ns#type as rdf:type and http://www.w3.org/2002/07/owl#Class as owl:Class

[10] http://www.w3.org/TR/owl-ref/#Class

satisfied by instances of that class [62]. For example, the <http://example.org/Protein> class could have the following axiomatic description (using the Manchester Syntax[11]):

'Macromolecule' and 'has part' some 'amino acid'

Where "Macromolecule" and "amino acid" refer to distinct OWL classes and "has part" is an object property. Many biological domains are formally described in ontologies in this way. The NCBO BioPortal [63] currently holds over 300 OWL ontologies for a wide range of domains of biological knowledge. For example, a widely used set of bio-ontologies is the Gene Ontology (GO), which consists of three interconnected ontologies that describe biological processes, cellular locations and molecular functions [64]. In using GO, biologists can annotate proteins and genes in an unequivocal manner that is also machine understandable. For instance the UniProt entry for the human Hexokinase-2[12] is annotated as having 'hexokinase activity' (GO:0004396) and is located in the 'cytosol' (GO:0005829).

1.7 Summary of chapter

This introduction provides the conceptual foundations for the research presented in this thesis. Specifically, I have described commonly used nomenclature regarding descriptions of nucleic acid structures and provide a succinct summary of the semantic web technologies used throughout this work. In the remaining chapters I present research concerning the machine understandable descriptions of nucleic acid structure, an effort to

[11] http://www.w3.org/TR/owl2-manchester-syntax/
[12] http://purl.uniprot.org/uniprot/P83776

publish online existing experimental data about SELEX derived aptamers, and describe a machine learning approach for classifying ligand binding based on motifs extracted from graph representations of RNA structures.

2 Chapter: **RKB: A Semantic Web Knowledge Base for RNA**

2.1 Abstract

Increasingly sophisticated knowledge about RNA structure and function requires an inclusive knowledge representation that facilitates the integration of independently – generated information arising from such efforts as genome sequencing projects, microarray analyses, structure determination and RNA SELEX experiments. While RNAML, an XML-based representation, has been proposed as an exchange format for a select subset of information, it lacks domain-specific semantics that are essential for answering questions that require expert knowledge. Here, we describe an RNA knowledge base (RKB) for structure-based knowledge using RDF/OWL Semantic Web technologies. RKB extends a number of ontologies and contains basic terminology for nucleic acid composition along with context/model-specific structural features such as sugar conformations, base pairings and base stackings. RKB (available at http://semanticscience.org/projects/rkb) is populated with PDB entries and MC-Annotate structural annotation. We show queries to the RKB using description logic reasoning, thus opening the door to question answering over independently-published RNA knowledge using Semantic Web technologies.

2.2 Contribution to thesis

In this chapter, I describe the RNA Knowledge Base (RKB) which contributes to the completion of Objective #1, to collect and catalogue existing nucleic acid sequences and structures using state of the art biological information management technologies, by creating machine understandable descriptions of base pair and base stacking interactions

identified from RNA molecular structure files deposited in the Protein Data Bank. This work also contributes to Objective #2, to carry out a quantitative analysis of known small molecule bound nucleic acids in terms of their recurring 3D structural motifs and sequence composition, as rich descriptions of base pair and base stacking interactions in the RKB are used as the basis for identifying structural motifs in RNA as described in Chapter 4.

2.3 Copyright notice

2.4 Background

The ability to accurately capture biomolecular behaviour is critical to our understanding of cellular systems. With biophysical instruments that measure everything from bond vibrations to fluorescence as a result of molecular interactions, scientists carefully translate these observations into a set of positive statements about the entities under investigation. The set of entities, objects and relations used by scientists, through their *lingua franca*, defines a *conceptualization* of their subjects of study. The explicit commitment to a conceptualization not only enables scientists to easily share knowledge, but also permits the creation of machine-understandable knowledge bases. An ontology is an explicit specification of a conceptualization of a particular domain of knowledge [65], in which the set of objects and their relations define its scope.

In some cases, the conclusions drawn about numerous experimental results do not necessarily apply universally, but instead appear as a result of a context-dependent experimental system. Biological situational modeling [66] has been used as a methodology to capture this knowledge in a precise and accurate manner, so that conflicting statements about biochemical entities may be tolerated provided there exists some circumstantial qualification. Hence, a long term solution for knowledge representation in the life sciences must consider context, in addition to identity and action.

Ribonucleic acids (RNAs) are essential cellular components with significant roles in protein synthesis and gene regulation. Increasingly sophisticated knowledge about

RNA structure and function is being revealed as a result of innovative biochemical investigations such as genome sequencing projects, sequence alignments, microarray analyses, structure determination and RNA SELEX experiments. Yet, our capacity to capture this knowledge by existing systems is limited in several important respects. First, RNAML [67], an XML-based exchange format for a select subset of information about RNAs, does not provide explicit formalization of the domain either from a logically or philosophical perspective. As an example, base stacking can be described with a natural language comment associated with the base-stack element, but we cannot specify a machine understandable type – what kind of thing is base stacking and what specializations of it exist (*e.g.* adjacent stacking or upward stacking). Second, XML Schema is primarily interested in the validation of the document structure, as opposed to the semantics of the domain terminology therein contained, thus language extensions cannot be properly validated. In contrast, RDF/OWL are formal logic based languages which enable the logical formalization of the domain, and as such can be used to infer new knowledge using some information system. Moreover, as languages of the Semantic Web, researchers may also publish their knowledge so as to further enhance structural and functional annotation in a machine accessible, but de-centralized manner.

Here, we describe an RNA knowledge base (RKB) for structure-oriented knowledge using RDF/OWL Semantic Web technologies. RKB extends the RNAO, an RNA ontology jointly developed with the RNA Ontology Consortium [68], and builds on other Open Biomedical Ontologies (OBO) for information content entities (*e.g.* PDB files, structure models), real world entities (*e.g.* base pairs, base stacks) and their qualities

(*e.g.* nucleoside/sugar conformations). RKB is populated from RNA-specific PDB entries and base pairing/stacking identified by MC-Annotate. We demonstrate how the resulting knowledge base supports powerful question answering over OWL-DL ontologies using a description logic system.

2.5 Results

This project pursued four main objectives: i) to unequivocally represent basic biochemical knowledge about nucleic acids and their structural characteristics, ii) to accurately capture the knowledge generated by a nucleic acid structural feature annotator such as MC-Annotate in such a way that it complemented other structural or functional knowledge, iii) to implement a scheme for the representation of knowledge obtained as information from a computational procedure, iv) to maximize interoperability with a set of trusted external ontologies. A high quality representation should facilitate data integration and enable question answering with a reasoning-capable knowledge base.

2.5.1 RNA Structure Ontology

The RNA knowledge base ontology[13] extends the RNAO[14] and provides a core set of hierarchically organized terminology for the accurate representation of RNA and their structural features. The RKB builds on material entities and qualities as defined by the BFO upper level ontology, the RO relation ontology for reusable domain independent

[13] Available at : http://semanticscience.org/projects/rkb/
[14] RNA Ontology: http://code.google.com/p/rnao/

relations, the IAO[15] for information content entities and ChEBI for specific chemical entities and their parts.

Material entities are spatially extended entities whose identity is independent and can be maintained through time. Material entity is the top level class for nucleic acids, base pairs, base stacks, chemical bonds / interactions and fiat parts of nucleic acids (nucleotide residues, sugar moieties, nucleobases) where bonds extend into another part from certain terminal atoms base pairs.

Qualities are categorical properties that existentially depend on, among other things, material entities. This forms the top level class for the syn- or anti- quality, a conformation[16] borne specifically by the nucleoside[17] part of a nucleotide residue[18] and imparts knowledge of the orientation of its respective base and sugar parts. Similarly, the envelope conformation[7] is a quality that is solely borne by the sugar part of a nucleotide residue.

The Information Artifact Ontology's Information Content Entities (ICEs) generically depend on at least one, but possibly more material entities. ICEs are the top level class for structure models, PDB records, coordinates and measurement values (Figure 10).

---

[15] Information Artifact Ontology: http://code.google.com/p/information-artifact-ontology
[16] Class URI: http://purl.obofoundry.org/obo/rnao/RNAO_0000123
[17] Class URI: http://semanticscience.org/rkb:RKB_000027
[18] Class URI: http://bio2rdf.org/chebi:50319
[7] Class URI: http://purl.obofoundry.org/obo/rnao/RNAO_0000124

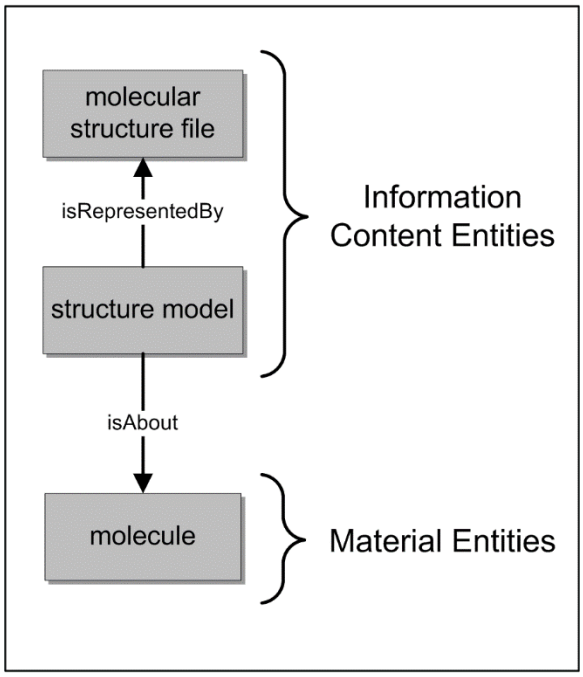


Figure 10 Information content entities are about material entities in the RKB

Illustration of the RDF-based representation used to relate Information Content Entities with their corresponding Material Entities. Molecular structure files are specific manifestations of structure models therefore structure models are represented by their corresponding PDB files. Structure models are also about molecules and other real entities (atoms, base pairs, etc).

We applied the Minimum Information to Reference an External Ontology Term (MIREOT) guidelines [69] to augment the RKB ontology with relevant classes, their annotations and minimal hierarchy from the RNAO, ChEBI [70], IAO,RO and BFO. The MIREOT document consists of 45 classes from the RNAO including nucleotide base pairs, base stacks and their structural qualities and the externally connected to spatial relation, 10 classes from the IAO for molecular structure data, 24 classes from ChEBI

which referred to nucleobases, nucleotides and their related sugar moieties. Conceptual overlap was captured using OWL's class equivalence relation.

The formalization used here departs from our previous work [71], by considering a more *information-oriented* representation. Thus, instead of referring to those qualities, roles, or parts of the molecule that are involved in a situationally-dependent base pairing process, we instead consider PDB structures composed of coordinates as information content entities which are generically dependent on the material entities to which they pertain to, such as molecules and base pairs. Thus, since there is always *some* pairing/stacking process that existentially depends on the pair/stack, we name only the latter.

2.5.2 RKB Population

The RKB is populated with RNA structures from the PDB and results of MC-Annotate using in house scripts. The conversion of PDB structures follows our previous work on small molecule chemistry [72]. Ontology population involved three basic steps: assigning names, asserting class membership, and assigning relations between entities. Having a consistent naming scheme makes data integration from PDB entries with MC-Annotate information straightforward. Unique names were generated as valid Uniform Resource Identifiers (URI) where each name consisted of the PDB identifier followed a different naming convention for objects and qualities:

Material Entities:

a. Structure Model: PDBID_cCHAIN

b. Nucleotide residue: PDBID_cCHAIN_rRESIDUE

c. Atom: PDBID_cCHAIN_rRESIDUE_aATOM

Qualities:

PDBID_mMODEL_cCHAIN_rRESIDUE_QUALITY

2.5.3 RNA Structure Representation

RNA structures obtained through experimental procedures and computational model building and refinement yields a file containing information about a molecule or collection of molecules. More specifically, the file is a serialization of a data structure and contains a description of the structure model in terms of the spatial positioning of atoms as a set of coordinates in three-dimensional space. In NMR, multiple structure models may be obtained, each of which captures a significantly populated conformation. The key relation is that information content entities (*e.g.* coordinates and collections of coordinates) are about real world entities (*e.g.* atoms, molecules). Importantly, structure models provide the means by which more information about the structure and function may be determined through additional analysis.

We used MC-Annotate over the set of PDB files that contain RNA structures to identify base pairing and base stacking in terms of their adjacency and relative orientation. A different individual was generated for each structural feature (base pairing, base stacking) of each model in the PDB file. This maintains provenance, in that entity

assertions are related to the model from it was derived and also allows comparison of structures from different models.

Base Pairing: Nucleotide base pairs may occur between any pair of nucleotide residues, and involve any number of atoms. Canonical base pairs, as described by Leontis and Westhof [49], occur as a result of the hydrogen bonding between the *edges* of nucleobases. Since edges are composed of multiple atoms and hydrogen bonds occur between pairs of atoms, Lemieux and Major [73], developed a system of finer granularity that refers to *sub-edges* or so-called *faces*. The sub-edges extend from the nucleobase along the ribose sugar and hence includes two new atoms, the O2' and N9 or N1 for the case of purines or pyrimidines respectively. Hence, nucleotide base pairs can be represented in the RKB using either the Leontis and Westhof (LW) or the Lemieux and Major (LW+) specifications, both of which contain at least one edge or sub-edge interaction respectively. The RKB uses the RNAO's "externally connected to" relation to represent the interacting edges and the sub-edges in nucleotide base pairs, thus suggesting qualitative spatial reasoning across the regions they occupy using Region Connection Calculus (RCC-8).

Different models of the same RNA sequence may suggest flexibility through structural rearrangements. Models 5 and 10 of chain A in NMR structure PDB:1AJU suggests a difference in sub-edge interactions . Where model 5 shows a single sub-edge interaction between the Watson-Watson sub-edge of G34 residue and the O2' sub-edge of

the G36 residue base pair, model 10 pairs indicates two sub-edge interactions between the Ww/O2' sub-edges and the Ss/O2' sub-edges.

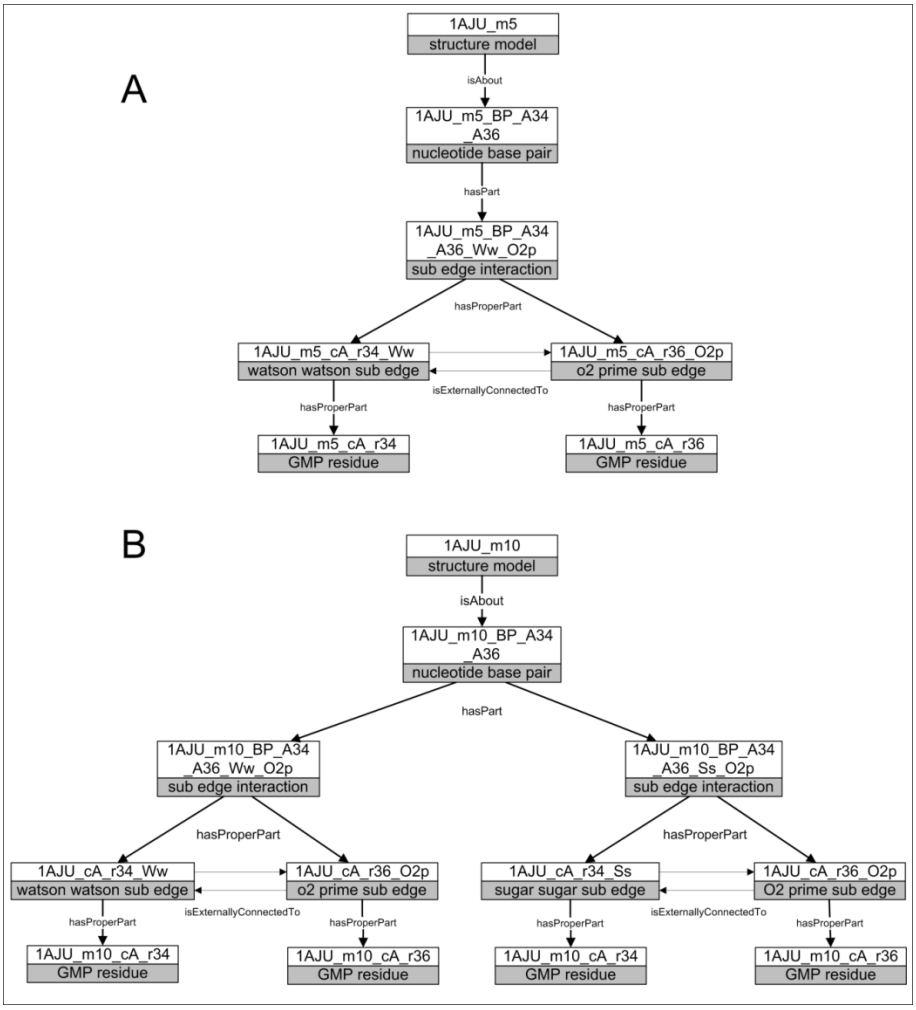


Figure 11 RKB nucleotide base pairs with varying sub-edge interactions.

Illustration of the RDF-based representation of molecular structure obtained from PDB files and from structure feature analysis of MC-Annotate. (A) Structure model 5 of PDB 1AJU is about a nucleotide base pair that is composed of a sub-edge interaction between the Watson-Watson sub-edge of the guanine residue at position 34 of chain A and the O2' sub-edge of the guanine residue at position 36 of chain A. (B) Structure model 10 of PDB 1AJU is about a nucleotide base pair between guanine residue at position 34 in chain A and guanine residue at position 36 in chain A, which is composed of two sub-edge interactions – a Watson-Watson sub-edge and O2' sub-edge, as well as a Sugar-sugar sub-edge and the O2' sub-edge.

Base Stacking: Base stacking involves a proximate spatial orientation of bases. For the RKB's representation of base stacks, we have chosen to make use of the RNAO's "externally connected to" object property to represent the non-covalent inter-molecular interactions existing between participating nucleobases. Base stacks also bear the RNAO's "base stack base-normal Orientation" and "base stack sequence adjacency" qualities for an accurate description of the relative directionality of the nucleobase normal vectors and for the description of adjacent and non-adjacent stacks.

Puckering: The ribose ring represents two main puckering modes, "envelope" and "twist". The "envelope" geometry is observed when one atom is located over or below the plane formed by the four others, whereas the "twist" geometry is observed when one atom is over and another is below the plane formed by the three others. The classification of a ribose, into either geometry, is dependent on the relative position of the carbon atoms of the ribose to its C5' atom. Hence while the carbon atoms in a ribose bear either the endo or exo qualities with respect to the plane formed by the other atoms, the ribose ring bears more specific envelop or twist qualities.

2.5.4 Question Answering

Example questions expressed using the Manchester OWL syntax are described below. They were formulated with the rdfs:label annotation properties using the DL Query plugin for Protégé 4 and answered using the embedded Pellet/FaCT++ reasoners. Queries (A)-(D) were performed on model 7 of PDB: 1AM0, queries (E) and (F) were performed on all models of the latter.

(A) *Find all nucleotide base pairs:*

'nucleotide base pair'

This query illustrates a simple retrieval of individuals that are instances of the class of *nucleotide base pairs,* as defined by the RNAO. In this case, 6 individuals were asserted to be members of this class and 6 are returned.

(B) *Find all base pairs that involve a Hoogsteen edge:*

'nucleotide base pair' that 'has part' some (('hoogsteen edge' or 'part of' some 'hoogsteen edge') and externally_connected_to some ('nucleotide edge' or 'part of' some 'nucleotide edge')

This conjunctive query involves existentially qualified variables. The individual must be an instance of the *nucleotide base pair* class, but this base pair must further be specified by a *hoogsteen edge.* Yet, no individuals are asserted to be instances of Hoogsteen edges, rather since MC-Annotate generates assertions at the sub-edge level, it is also necessary to ask for any parts of the edges. The mereological inference that a sub-edge is part of an edge makes use of agglomerating sub-edge classes, where for example the "hoogsteen sub-edge"[19] is equivalent to {"C8 sub-edge" or "hoogsteen hoogsteen sub-edge" or "hoogsteen watson sub-edge" or "bifurcated hoogsteen sub-edge"}, that establish a parthood mapping to edges, whereby the existential restrictions for the 3 edges are restricted to differing subsets of sub-edges. Finally, it is necessary to specify that the edge/sub-edge we are interested in is externally connected to another nucleotide edge/sub-edge. Two base pairs are retrieved in which the hoogsteen sub-edge is

[19] Class URI: http://semanticscience.org/rkb:RKB_000092

existentially known to be part of a hoogsteen edge and is also externally connected to another nucleotide edge/sub-edge.

(C) *Find all base pairs with a Hoogsteen edge that is part of a guanine residue*

'nucleotide base pair' that 'has part' some (('hoogsteen edge' or 'part of' some 'hoogsteen edge') and 'part of' some 'GMP residue [chebi:50324]' and externally_connected_to some ('nucleotide edge' or 'part of' some 'nucleotide edge'))

This query further refines Query (B), in that the Hoogsteen edge must be attached to a guanine residue and must be externally connected to another edge/sub-edge. Two base pairs are found.

(D) Find all base pairs involving a Watson-Watson sub-edge and a Hoogsteen-Hoogsteen sub-edge

'nucleotide base pair' that ('has part' some ('watson watson sub edge' and externally_connected_to some 'hoogsteen hoogsteen sub edge'))

This query aims to discover sub-edge interactions that are uniquely identified by MC-Annotate and are specified in RKB/RNAO using the *externally connected to* relation. Two results are obtained in model 7.

(E) Find all nucleotide base pairs involving at least one Hoogsteen sub edge interaction which is contained in a structure model from the PDB file 1AM0.

'nucleotide base pair' that 'has part' some (('hoogsteen edge' or 'part of' some 'hoogsteen edge') and externally_connected_to some ('nucleotide edge' or 'part of' some 'nucleotide edge')) and inv('is about') some ('structure model' that 'is represented by' some {'Molecular Structure File PDB:1am0'})

This conjunctive query with undistinguished variables (hoogsteen/nucleotide edge/part of edge, structure model) and a distinguished variable (the 1AM0 PDB file) identified 14 nucleotide base pairs out of a total of 40 across all models in the 1AM0 that involve at least one Hoogsteen edge.

(F) *Find how many structure models were defined in the pdb file for 1AM0*

'structure model' that 'is represented by' some {'Molecular Structure File PDB:1am0'}

This conjunctive query uses a distinguished variable to find all 7 structure models represented by the 1AM0 structure file.

2.6 Discussion

2.6.1 RNA on the Semantic Web

The aim of the RNA Ontology Consortium is "to create an integrated conceptual framework, an RNA Ontology (RNAO), with a common, dynamic, controlled, and structured vocabulary to describe and characterize RNA sequences, secondary structures, three dimensional structures, and dynamics pertaining to RNA function" [68]. The work described here on RNA structure and structural features provides the basis for a concrete

path towards which other essential RNA structural and functional features may be added in the future. With contextual modelling in hand, we are able to represent highly dynamic features of RNA structure and function as is observed in NMR and other experiments that characterize molecular dynamics. While the RKB requires an OWL2 compliant reasoning system to obtain all the expected inferences, there exists substantial value in being able to publish the knowledge base as a collection of Semantic Web documents which are also accessible through a SPARQL endpoint[20]. Future work involves provisioning the knowledge base through the Bio2RDF linked data network [74], thus enabling entity resolution and web-based interlinking between datasets.

2.6.2 OWL modelling

Modelling knowledge using OWL is challenging for a number of reasons. The first is that relations between objects are binary, of the form *relation*(x,y), which precludes temporal qualification as a third argument in a ternary relation. Thus n-ary relations must be converted into n-ary objects, and this approach is exemplified in our representation of base pairing and base stacking. A second challenge is that OWL imposes certain non-structural restrictions on properties in order to remain decidable. These restrictions ensure that properties which are either transitive or part of a role chain may not be involved in cardinality restrictions (min, exactly, max), and may not also be declared as functional, inverse functional, irreflexive, antisymmetric or disjoint with another property. Role chains involving *part of* / *is about* and their inverses are useful in finding all entities that are described by information content entities. But this precludes the use of these roles in

[20] See project page: http://semanticscience.org/projects/rkb/

cardinality restrictions. In order to overcome this restriction, more specific sub-properties such as *has grain* / *has quality* / *has role* could be used to make knowledge base assertions, and these roles are then used to make queries with cardinality restrictions. Thus cardinality restrictions can be placed in the ontology, and also in the instance base. For instance, the AA base pair class is equivalent to a Nucleotide base pair that has proper part exactly 2 adenine monophosphate residues. In this way, we can discover all such instances in RNA structure data.

2.6.3 Future directions

The RKB makes use of the MIREOT scheme to incorporate selected parts of trusted external ontologies. RKB's coverage will continue to grow alongside other resources represented with Semantic Web technologies such as the 30+ databases provided by the Bio2RDF [74], including UniProt [75], and the PDB. Yet a major challenge exists in ensuring that the raw linked data is massaged into more sophisticated knowledge representation schemes, such as the one described here. Ultimately, integration at both the syntactic and semantic levels across domains will allow maximum interoperability between the RKB and other relevant knowledge.

2.7 Conclusions

The RKB facilitates RNA knowledge discovery using a set of expressive OWL ontologies instantiated with PDB structure data and annotations from MC-Annotate. The resulting knowledge base can be used for simple information retrieval and more sophisticated ontology-based knowledge discovery. Our work demonstrates the

representation of information content entities such as PDB files and structure models, and how these relate to real world entities and their qualities. Continued collaboration with other members of the RNA Ontology Consortium should maximize interoperability of RNA-related information, particularly with sequence alignments, motifs and other structural and functional knowledge. Together, we will provide new avenues for biological knowledge discovery powered by the standards provided by the W3C Semantic Web effort.

2.8 Methods

2.8.1 Ontology Design

The RKB ontology was designed using the OWL editor Protégé Ontology Editor[21] (v4 Build 113) using Pellet or FaCT++ [76, 77] reasoners for consistency checking. Nucleic acid structures were obtained from the PDB [34], and MC-Annotate [73] was used to identify base pairings, base stackings, and various spatial conformations including sugar puckering. Our design approach followed a well-used methodology [78]. RNA structural feature terminology was obtained from literature [42],and new terminology created to group together classes related by subsumption. Subclasses are homogenous and increasingly specialized, while each child term can be easily differentiated from its parent with clear human readable labels, accurate and concise definitions and existential / universal / cardinal axiomatic descriptions where feasible. Upper level ontologies suggests increased interoperability and semantic coherency between domain ontologies due to grounding of the basic types of domain entities and the

[21] http://protege.stanford.edu/

imposing of restrictions on the relationships that these entities may specify. Our New Upper Level Ontology (NULO), inspired by the Basic Formal Ontology (BFO) [79], offers a simple framework that enables the distinction of objects, qualities, processes and spatial regions and also features object-process, object-quality, parthood, spatial, temporal relations drawn from foundational work [80].

Classes defined in the RKB are mapped to NULO concepts. For example, when considering the horizontal plane on a ribose, and the C5' atom is positioned to the left side, the location of the atoms with respect to the plane define either an "exo" or "endo" quality (below or above the plane, respectively) which is a quality of the corresponding atom of the ribose.

New object properties were added to further describe some of the more specific relations required in (but not restricted to) this domain. The pair *isImmediatelyAfter/isImmediatelyBefore* provides a relation between any two entities that are spatially related by adjacency. These properties permit the description of the relative positioning of nucleotide bases within a nucleic acid. They also allow for the description of the relative positioning of nucleobases that participate in either adjacent or non-adjacent stacking interactions.

2.9 Authors' contributions

MD and FM conceived of the study and participated in its design. MD and JCT generated the results and drafted the manuscript. FM and MP provided technical guidance and participated in the preparation of the manuscript.

2.10 Acknowledgements

This paper was written with the support of NSERC Discovery Grant for MD. The authors would like to thank Alison Callahan for her assistance with MC-Annotate RDFization.

3 Chapter: **Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX Experiments**

3.1 Abstract

Aptamers are short single-stranded nucleic acids or amino acid polymers that recognize and bind to targets with high affinity and selectivity. Nucleic acid aptamers are typically isolated from large combinatorial libraries through the application of Systematic Evolution of Ligands by Exponential enrichment (SELEX). With thousands of aptamers reported in the literature, having this information in a database would facilitate rapid retrieval of current aptamer knowledge for experimental researchers and enable new research in computational methods for predicting nucleic acid small molecule interactions. Here we present the Aptamer Base (http://aptamer.freebase.com), a collaborative knowledgebase about aptamers, their interactions and detailed experimental conditions with citations to primary scientific literature. Users can formulate keyword or structured queries over a set of types and single-value attributes. Available as a Freebase database, users can contribute, create and share aptamer experiments with the scientific community.

3.2 Contribution to thesis

This chapter contributes to the completion of thesis Objective #1, to collect and catalogue existing nucleic acid sequences and structures using state of the art biological information management technologies, by providing over 1600 (as of Mach 2014) aptamer sequences and respective molecular targets as extracted from primary scientific literature. This work

also provides the aptamer community with rich descriptions of the experimental conditions in which these aptamers were generated.

3.3 Copyright notice

3.4 Introduction

Over the last few decades, rapid developments in both molecular and information technology have collectively increased our ability to understand molecular recognition, leading to major implications for drug discovery [81]. *In vitro* screening techniques and adaptive molecular evolution methods, such as phage display [82], have made it easier to rapidly screen enormous molecular libraries to find promising binding ligands. Concurrently, advances in biomedical informatics have made it possible to harness the power of large datasets and make influential predictions for molecular recognition. Together, the scalability of information systems coupled to the massive reduction in the cost, efficiency and time of techniques such as DNA sequencing [83] will continue to support an exponential growth in information gain for molecular biology.

One emerging area of interest in the quest to understand molecular recognition is focused around the isolation of aptamers. Aptamers are single-stranded nucleic acid or amino acid polymers that recognize and bind to targets with high affinity and selectivity. Nucleic acid aptamers are typically isolated from large combinatorial libraries containing approximately 10^{15} different sequences. This *in vitro* selection process is usually performed using the Systematic Evolution of Ligands by Exponential enrichment (SELEX) process [12, 84]. The resulting selected aptamers often bind to their cognate ligands with dissociation constants in the picomolar range[85], and can be selected for a wide variety of targets including small molecules [86], organic dyes [87], toxins [88], proteins [8, 89], viruses[90] and whole cells[91]. As a result, aptamers have emerged as

attractive molecular recognition agents that rival antibodies in both therapeutic [92, 93] and diagnostic [94] applications.

While several efforts have been focused on collecting aptamers and their interactions [95-97], most of the information regarding experimental methods still remains in the unstructured and textual format of peer reviewed publications. Furthermore, access to the databases is normally limited to HTML forms, thereby hindering the type and number of queries that can be posed against these datasets. Additionally, major sequence data providers such as GenBank [98], EMBL [99], and DDBJ [100] do not maintain a list of artificially created sequences. Consequently, the community lacks a data resource in which information about aptamers, their sequences and the experimental conditions used in their selection can be both stored and queried.

To address this, we present the Aptamer Base, a collaborative knowledge base focused around aptamers in which detailed information is available regarding the nature of their interactions and the experimental conditions used in their selection. The open collaborative nature of the Aptamer Base provides the community with a unique resource that can be updated and curated in a decentralized manner, thereby accommodating the ever evolving field of aptamer research.

3.5 Materials and Methods

The Aptamer Base was created using Freebase (http://www.freebase.com), a free, openly licensed community-built resource for structured data that provides information on over

22 million topics. Freebase organizes over 360 million facts into bases, which are collections of thematically related topics. The Aptamer Base is one such collection that has been built by a group of expert curators who manually extracted information on over 157 SELEX based experiments from the primary literature published between 1990 and 2006. Entries in the literature from 2006 to present are added on a weekly basis.

3.5.1 The Aptamer Base data model

Based on the entity relationship model [101], data in Freebase is stored as a structured graph in which entities or *topics* represent single concepts or real world things. Every topic in Freebase is identified by a unique identifier and is accessible via a Uniform Resource Locator (URL). Topics can in turn be categorized or "typed", representing an *IS A* relationship between topics and their respective types. For example, typing a topic with both the *RNA* and *Aptamer* types indicates that the particular topic *IS AN RNA* and *IS AN Aptamer*. Every type in Freebase may hold any number of *properties* indicating a *HAS A* relationship between the topic and a value of the property. For example, a topic that is typed as a *Nucleic Acid HAS A* "secondary structure" property that may have a value of "pseudoknot". Properties in Freebase may point to other topics and can be restricted by typing the range of the relation (*e.g.* the *Dissociation Constant* type *HAS A* property called "is dissociation constant of" that links to topics only of the *Interactor* type, as in http://www.freebase.com/view/m/0cjchc8), thus controlling the topics that can be associated using relations with type restrictions.

The Aptamer Base data model (Figure 12) is composed of an *Interaction Experiment* which represents the results and details of experimental procedures used to identify biomolecular interactions. In the case of aptamers derived from SELEX, the *SELEX Experiment* captures the SELEX method used to generate the sequences, the partitioning method used to discard non-binding sequences and the recovery method used to isolate the aptamers from the aptamer-target complex. Additional details regarding the SELEX experiment, such as the number of selection rounds, template sequences and details on the selection solution may also be added. *Interaction Experiments* may report one or more *Interactions*. Each *Interaction* is related to the participants (*Aptamers* and *Aptamer Targets*) of the reported interaction, through the "has participant" property.

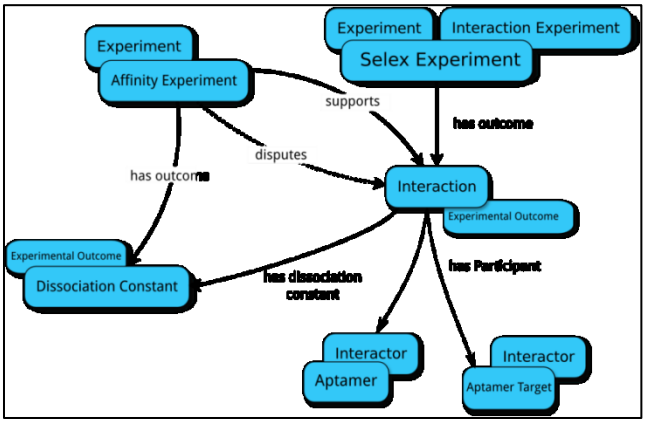


Figure 12 Basic type relation map used by the Aptamer Base to describe SELEX experiments. The Interaction Experiment type "has outcome" an Interaction. Each Interaction "has participant" at least two Interactors (Aptamer and Aptamer Target). The Affinity Experiment type "has outcome" a Dissociation Constant that either "confirms" or "disputes" an interaction. Blue ellipses denote types and arrows represent properties between topics. Overlapping ellipses represent the multiple types associated with topics.

Where possible, an *Interaction* is associated with the *Affinity Experiment* that quantitatively assesses the binding affinity between the aptamer and the corresponding target through a dissociation constant. *Interaction Experiments* can either "confirm" or "dispute" *Interactions* under a set of experimental conditions, thus enabling searching for binding and/or non-binding sequences.

*Minimal Aptamer*s are sub-sequences or variants of larger sequences that exhibit binding to a target ligand. Sequences obtained from SELEX are typically 30 or more nucleotides in length [102], but it is widely established that conserved consensus sequences play an important role in biological function. Therefore, a minimal sequence of the full length aptamer that retains the characteristic function for which the full length aptamer was originally selected may be identified. The reported minimal aptamers participate in *Interactions* that are distinct from those of their parent aptamers, thereby permitting the linking of a minimal aptamer-target interaction to a dissociation constant topic that is distinct from that of its parent (Figure 13).

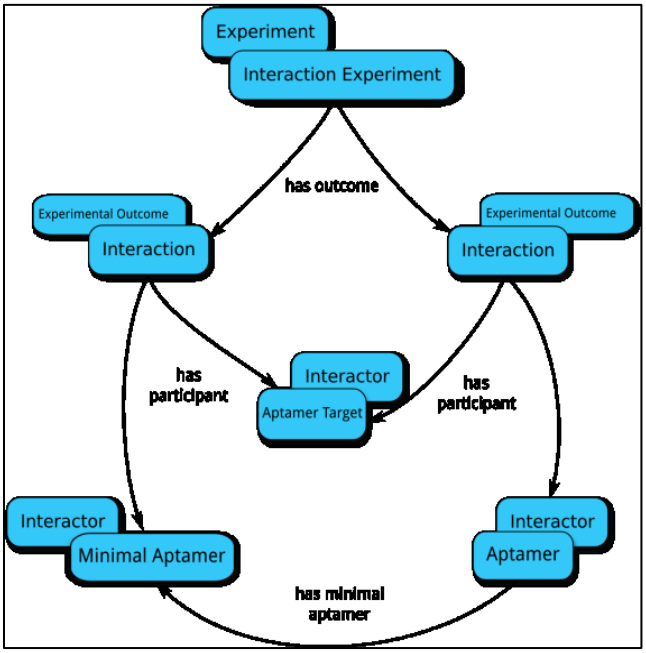


Figure 13 Minimal aptamers are captured in the Aptamer Base by creating a new Interaction for each individual Minimal Aptamer, permitting the description of unique Interactions corresponding to distinct Affinity Experiments with Dissociation Constants for both Aptamers and their Minimal Aptamer.

For every aptamer reported in the Aptamer Base we have also included secondary structure as predicted by RNAfold [103]. The dot bracket notation, minimum free energy and version of RNAfold are provided for each prediction (*e.g.* http://www.freebase.com/view/m/0gl4r9m). Given that the "has predicted secondary structure" property is a one-to-many relation, the Aptamer Base could accommodate the incorporation of alternative secondary structure predictions or even tertiary structures made by other prediction programs [104-106].

3.6 Results

3.6.1 Aptamer Base Content

The Aptamer Base currently contains information about 676 interactions between 928 aptamers and 131 targets (Figure 14) organized into 4143 topics, comprising in total almost 30,000 facts. This information has been compiled from 156 SELEX based experiments published in the primary literature between 1990 and 2006.

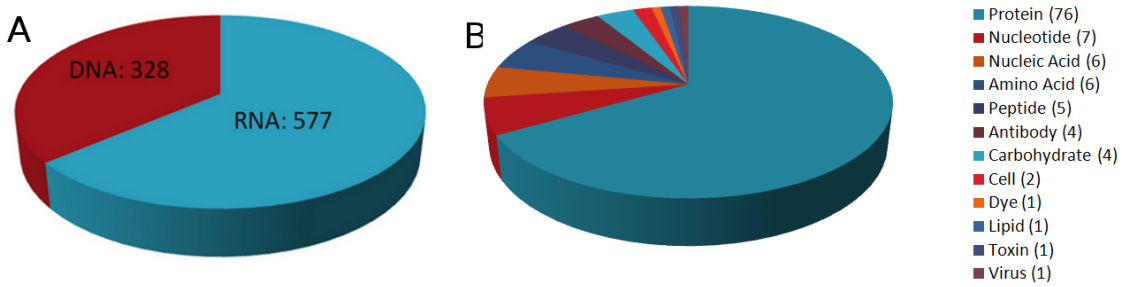


Figure 14 Summary of target types and aptamer types found in the Aptamer Base. A) Distribution of the over 900 aptamer types described by the Aptamer Base. B) Distribution of the 131 aptamer targets found in the Aptamer Base.

3.6.2 Using the Aptamer Base

Freebase utilizes the Metaweb Query Language (MQL) as its application programming interface (API) for providing programmatic access to its data. Both queries and responses are sent to Freebase through a Representational State Transfer (REST) interface that utilizes a plain text data interchange format serialized as JavaScript Object Notation (JSON). For example, to retrieve all sequences for all RNA aptamers in the Aptamer base, users must specify all required types as seen in Box 1, along with an abridged result set in Table 1.

```
[{
    "id":null,
    "type":"/base/aptamer/aptamer",
    "a:type":"/base/aptamer/linear_polymer",
    "b:type":"/base/aptamer/interactor",
    "c:type":"/base/aptamer/rna",
    "/base/aptamer/linear_polymer/sequence":null,
  "/base/aptamer/interactor/is_participant_in":[{
  "/base/aptamer/interaction/has_participant":[{
  "d:type":"/base/aptamer/aptamer_target",
  "name":null
  }]
    }]
  }]
```

Box 1 Sample MQL query. The set of all unique identifiers, sequences and aptamer target names for all RNA aptamers in the Aptamer Base are returned. MQL queries can be posed on Freebase data using their query editor found at http://www.freebase.com/queryeditor.

Table 1 Abridged list of Aptamer Base topics returned by query from Box 1. Topics can be viewed by visitinghttp://www.freebase.com/view/[TOPIC_ID].

Topic ID	Target Name	Aptamer Sequence
/m/0cvjvjp	Tetracycline	GGCCUAAAACAUACCAGAUUUCGAUCUGGAG AGGUGAAGAAUUCGACCACCUAGGCCGGU
/m/0cx03px	Dopamine	GGGAAUUCCGCGUGUGCGCCGCGGAAGACGU UGGAAGGAUAGAUACCUACAACGGGGAAUAU AGAGGCCAGCACAUAGUGAGGCCCUCCUCCCA AGGUCCGUUCGGGAUCCUC
/m/0cysc2w	Human epidermal growth factor receptor 3	CAGCGAAAGUUGCGUAUGGGUCACAUCGCAG GCACAUGUCAUCUGGGCG
/m/0czndxr	Human activated protein C	GUGAGACCAGCCGAGUGGUGUCUGGCUAUUC ACUGGAGCGUGGGUGGAACCCCUGCGCACUCG UUUGGCUGUCCGGGCCUUCGGGCCGGGAUUA UCUCU

3.7 Discussion

While other efforts have been previously directed at creating databases about aptamers [95-97], the Aptamer Base provides significant benefits beyond these in several respects. First, anyone can contribute to the Aptamer Base. Data entry and curation can be

undertaken by any registered member of Freebase. This unique feature enables researchers, educators and students within and outside of the aptamer community to contribute and make use of the knowledge in the Aptamer Base. Like other successful open data projects such as Wikipedia, Freebase relies on community collaboration to maintain a complete and accurate dataset. Users can correct or augment facts in a base, but may be suspended if in violation of basic Freebase content creation guidelines (see http://wiki.freebase.com/wiki/Contribution_guidelines). Furthermore, administrators of the Aptamer Base can at any point modify, remove or revert any entry to ensure data quality and consistency.

The Aptamer Base provides detailed, structured information about the experimental conditions under which the aptamers in the dataset were selected and tested. For example, pH, temperature, salt concentration and the buffering agent are recorded. Providing access to these experimental conditions is crucial in aiding the reproducibility of the reported experiments [107]. For the case of SELEX based experiments, our data model provides the required types and properties to capture experimental details of both the partitioning of non-binding sequences from the library and the recovery of binding sequences from the aptamer-target complex. This information can be leveraged by aptamer researchers wishing to either reproduce or select new aptamers. For example, researchers can query the conditions used in successful SELEX experiments for targets that are either identical or similar to the target that they wish to develop new or improved aptamers.

The over 900 aptamers described in the Aptamer Base, which have been added on a weekly basis between September 2010 and July 2011 (http://activity.freebaseapps.com/domain?id=/base/aptamer), comprise a set of sequences that have been asserted to participate in an interaction with an aptamer target in their respective publications. It is often the case that scientists report on aptamer sequences that have not been tested for their ability to recognize targets and as a result there is always a possibility that these sequences stored in other databases have little or no affinity for the target. The Aptamer Base contains only sequences that have been verified using an *Affinity Experiment* to either bind ("confirm") or not to bind to the target ("dispute") in question.

The Aptamer Base also reuses knowledge that is already in Freebase. Our dataset includes types and topics that have been contributed by the community. For example, we reuse information from Chemistry Commons (http://freebase.com/view/chemistry) to provide links to PubChem and Wikipedia for most of our aptamer target topics. By choosing to use types and topics already found in Freebase, our curators can save time in the creation of new relevant topics. In addition, curators can ascribe to the community consensus on particular topics and consequently enhance these topics by providing new contextualized knowledge about them. Similarly, the 12 registered members of the Aptamer Base have created types that have been reused by the Freebase staff to annotate 285 Wikipedia articles (http://www.freebase.com/view/base/aptamer/views/rna).

Programmatic access to Freebase data through either the REST interface or their hosted development environment Acre (http://wiki.freebase.com/wiki/Acre) enables its users to create data "mashups" via web applications that use and combine data from different bases in Freebase. These can be combined with external resources to further enhance the knowledge hosted on Freebase. For example, when possible, we provide ChEBI identifiers for our small-molecule ligand *Interactors*, these identifiers can be used to query ChEBI's Simple Object Access Protocol (SOAP) interface to extract further information.

The expandable nature of the data model used by the Aptamer Base has been tailored to provide a core set of the necessary details required to generically reproduce SELEX experiments [107]. However, the relatively young stage of aptamer *in vitro* experiments has inevitably produced publications of varied levels of detail. Consequently, the data stored in the Aptamer Base is only as detailed as the publication of origin. The intricacies of our data model could then be initially considered as a common set of requirements for reporting future *in vitro* aptamer experiments and subsequently expanded to accommodate results of other *in vitro* evolution experiments that involve biomolecular interactions such as phage display [82] and the development of ribozymes.

With the Aptamer Base, we hope to address the gap that exists in available data resources for *in vitro* selected sequences. As the biological significance of aptamers continues to be discovered, we predict that such resources will become increasingly

important. The rapid advances in structural biology efforts relating to the discovery of novel biochemical functions and in the elucidation of structure of nucleic acids and proteins which are involved in biomolecular interactions (*i.e.* protein-protein and protein-small molecule interactions) can be serialized into similar community-built knowledge bases which will provide the community with open programmatic access to their experimental results. Through the open framework for accessing and contributing data that is at the core of Freebase, the Aptamer Base promises to not only provide an up to date and quality data resource for aptamer scientists, but one which has the potential to persist and grow alongside the state of the art in this field.

Knowledge of these protein-small molecule interactions is important to understand how proteins function in biological systems. There have been rapid advances in structural biology and relating structure to biochemical function and mechanism. However, knowledge of protein structure alone does not ensure accurate prediction of function and biological activity. The complete characterization of any binding interaction requires a quantification of the affinity, number of binding sites, and the thermodynamics.

3.8 Funding

This work was supported by Natural Sciences and Engineering Research Council Discovery Grants to both MCD and MD.

3.9 Acknowledgements

We would like to acknowledge Matthew Chan, Alexander Wahba, Michael Beking and Alison Callahan for their support, useful discussions and technical insight in the creation of the Aptamer Base and of this manuscript.

4 Chapter: **Structural components of Ribonucleic Acids involved in small molecule binding**

4.1 Abstract

The recent growth in the number of experimentally determined RNA structures deposited in the PDB provides the data that scientists need to analyze nucleic acid-ligand interactions. However, we currently lack computational approaches for using this data to predict ligand-RNA binding sites. Such an approach would aid in the rational design of non-coding RNA molecules that bind to specific ligands of interest. Motivated by this problem, we describe a method for identifying recurring elements of RNA structures and use these elements to classify and predict RNA structural motifs that are involved in ligand binding. Specifically, we computed the minimum cycle bases of the tertiary structure of a set of RNA-only structures based on graphs constructed from symbolic annotations of base pair and backbone interactions. We show that cycles composing the nucleic acid structures can be used as features for training a Support Vector Machine (SVM) classifier that predicts ligand-RNA binding sites with high sensitivity and specificity.

4.2 Contribution to thesis

The work presented in this chapter is primarily focused on the completion of both Objectives #2 and #3 which are respectively concerned with the collection and cataloguing of existing nucleic acid sequences and structures using state of the art biological information management and with execution of a quantitative analysis of known small molecule bound nucleic acids in terms of their recurring 3D structural

motifs and sequence composition. Specifically, I describe a novel method for the quantitative analysis of recurring structural motifs in ligand containing RNA structures. I also describe the results obtained from training two classifiers to accurately predict RNA-ligand binding structures in terms of their minimum cycle bases.

4.3 Introduction

The structure function relationship in macromolecules has been well studied and is ubiquitous in our understanding of biological systems [108-111]. With the recent growth in the characterization of new non-coding RNAs, and the *in vivo* and *in vitro* [112, 113] study of riboswitches and aptamers, the need to understand how nucleic acid structure is related to function is even more pressing [9, 11]. Moreover, as new nucleic acid structures are solved, novel RNA functions will likely emerge. The rate at which new structures of nucleic acids have been deposited in the PDB has been steadily growing [114] and thus there is a need for tools that efficiently identify and categorize nucleic acid structure. More specifically, given the increasing interest in developing aptamers for various therapeutic applications [92, 93], biologists would benefit from computational models that describe and predict nucleic acid-small molecule interactions.

Recurring substructures of RNA, or RNA motifs, have been characterized in terms of their 3D atomic coordinates or secondary structure information. For example, FR3D [115] and ARTS [116] describe RNA structural motifs found in PDB submissions in terms of their atomic coordinates. RNA FRABASE [117] and RNA-Bricks [118] are databases that describe RNA motifs in terms of secondary structures such as loops, stems, hairpins and pseudoknots. RNA-Bricks also includes descriptions of protein RNA contacts. Importantly, these resources catalog fragments of RNA motifs that have been observed to repeat across structures, and in some cases which confer specific functionality [119, 120]. However, the arbitrary nature and lack of standards for describing RNA motifs means that though they are growing in size and scope, these

databases do not facilitate a reproducible analysis of RNA structural motifs. More specifically, it is difficult to integrate data about motifs and their properties across resources or to contribute additional data to existing motif annotations. Lastly, these resources do not enable *de novo* motif discovery, as they only support queries for already defined RNA motifs [121].

Lemieux and Major [122] were the first to address these challenges using a graph representation of RNA structure to computationally extract and identify RNA motifs. They identified graph fragments, or cycles, used to define RNA motifs in terms of their nucleotide residue composition and annotation of their covalent and non-covalent interactions. Specifically, a cycle is defined as a directed path where the first and last vertex (nucleotide) of the path (composed of edges that are covalent and non-covalent nucleotide residue interactions) is the same. Importantly, they identified indivisible cycles that repeated within the high resolution crystal of the large ribosomal subunit of *Haloarcula marismortui* (PDB ID: 1FFK [123]) and that, in some cases, correspond to previously defined RNA motifs. By hierarchically clustering these cycles based on their 3D geometries they generated groups of similar cycles, including cycles that contained isosteric base pairs.

Motivated by their approach, we adopted a graph based representation of nucleic acid structure to describe all *RNA-only* structures in PDB, where the nodes and edges of our graphs are derived from base pair annotations generated by DSSR [124, 125], an RNA structure annotator tool. Using this graph based representation, we computed the

Minimum Cycle Bases (MCB) [126, 127] of the *RNA-only* structures from PDB, and in doing so developed an approach to automatically group sets of minimum cycles that have identical topologies and component parts. We demonstrate that the set of generated minimum cycles represent a structurally redundant partitioning of ribononucleotides and that minimum cycles with the same nucleotide residue composition and identical relative covalent and non-covalent respective interactions recur *within* a given RNA structure and *across* homologous and non-homologous portions of the macromolecular structure. Moreover, we have used the minimum cycle bases of RNA structures to train Support Vector Machine (SVM) classifiers to distinguish structures that have binding sites and to identify minimum cycles that are in the vicinity of a binding site.

4.4 Methods

We developed a data processing pipeline that enables the identification of the complete set of minimum cycles computed from a non-redundant set of RNA X-Ray crystals and used their features to construct two binary classifiers. The first classifier considers the composition of minimum cycles as its main set of features, while the second classifier takes into account minimum cycle centric features and their proximity to the ligand (Figure 15). In the following sections we describe in more detail each of these steps.

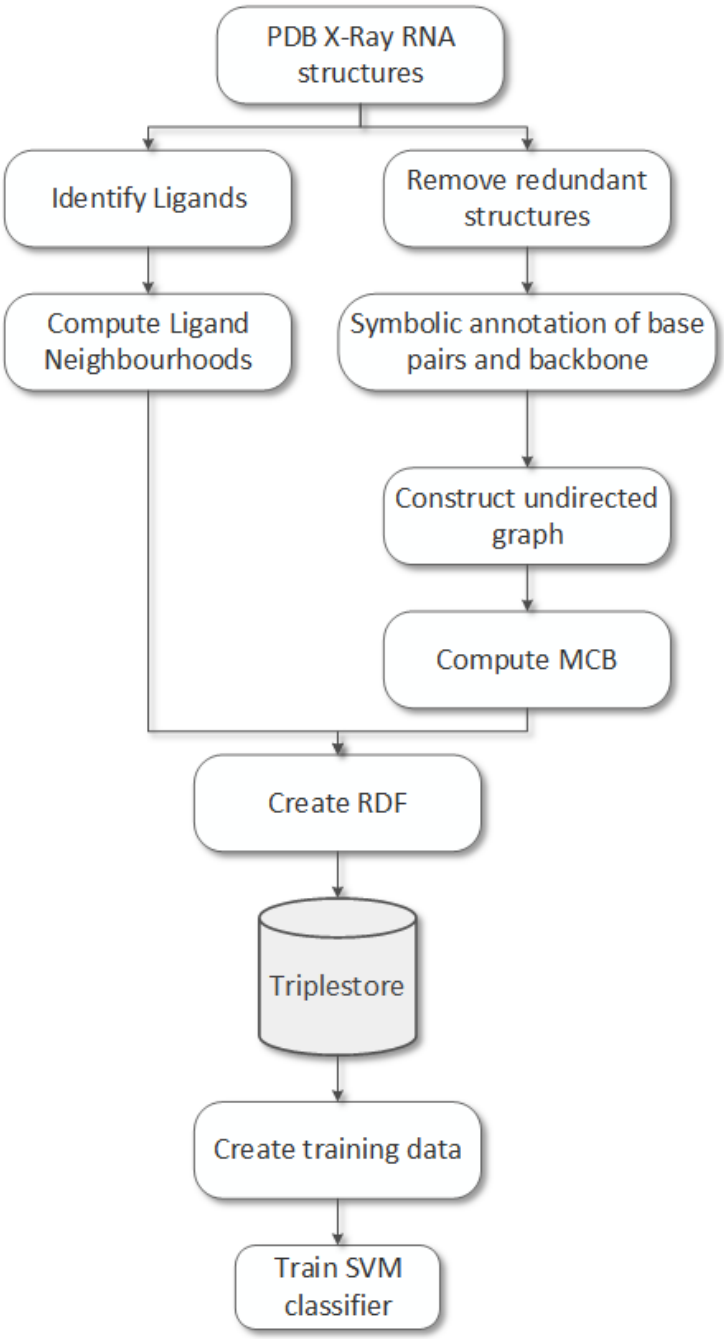


Figure 15 Workflow diagram of data acquisition and processing followed by machine learning and classification.

4.4.1 Construct a structurally heterogeneous set of *RNA-only* X-Ray crystals deposited in PDB

As of November of 2013, the Protein Data Bank [34] database hosted over 2,200 experimentally determined 3D structures *containing* RNA. X-Ray crystallography remains the most popular experimental method used, accounting for over 70% of the deposited submissions to this database, the remaining nucleic acid structures were solved by either Nucleic Magnetic Resonance or cryoelectron microscopy. As is the case with protein structures [128-130], the growing set of ribonucleic acid structures deposited in the PDB are often not fundamentally distinct from previous submissions in two ways: Firstly, depending on the nature of the asymmetric unit (ASU) described in the structure file, PDB structure records may contain redundancies *within* each individual ASU [131]. Secondly, it is often the case that more than one research group may submit experimental results of structure determination experiments on substantially similar structures to the PDB [132].

An ASU is the smallest unit that can be rotated and translated to generate one *unit cell* which can in turn be copied and translated to generate the entire crystal structure. The PDB provisions structure files as ".pdb" files which contain a single ASU, however depending on the position and conformation of the crystalized macromolecule within the *unit cell* the ASU may contain one of: a portion of the *biological assembly*, a complete *biological assembly* or multiple *biological assemblies* [133]. The *biological assembly* refers to the macromolecular structure that is believed to be the functional form of the structure being studied and is generally the structure of interest [133, 134]. As an

example of multiple biological units being presented in a single ".pdb" file, the PDB record for the crystal structure of the *Geobacillus kaustophilus* T box riboswitch (PDB ID: 4MGN [135]) contains exactly 2 copies of the biological unit in the published ASU.

The sources of structural redundancy *between* PDB submissions are varied [115, 128, 129, 133]: Firstly, structural biologists often times are interested in solving structures containing various mutations from a previously submitted wild type structure to test functional and/or structural hypotheses. For example, the structure of the prokaryotic S-Adenosyl methionine riboswitch has been submitted to PDB several times containing various point mutations (PDB IDs: 3GX5 [136], 3GX3 [136] and 4KQY [137]). Secondly, structural investigations may involve measuring the effect on a macromolecule when incorporating a structurally similar yet distinct ligand. Consider for example the structure of the Thiamine Pyrophosphate riboswitch from *Arabidopsis thaliana* (PDB ID: 2CKY [138]) which has been solved bound to oxythiamine pyrophosphate (PDB ID: 3D2X [139]). Despite structural similarities between the ligand, the resulting 3D structures of these riboswitches have a high level of structural overlap and relatively low RMSD measures of 96.1% and 0.62 Å respectively as measured by CLICK [140]. Additional sources of structural redundancies between PDB records include submissions of highly conserved substructures (PDB IDs: 430D [141], 480D [142] and 1Q96 [143]) of larger nucleic acid macromolecules and submissions of homologous RNA molecules which exhibit high levels of sequence similarity.

As the focus of this study is to gain insight into the structural patterns that govern small molecule-RNA interactions, we constructed an *RNA-only* set of X-ray crystals which was obtained from the PDB via their REST API. In total, 533 RNA structures containing no DNA, nor Proteins, nor DNA-RNA hybrid component parts were identified. This initial structurally redundant set of *RNA-only* structures was refined via a two-step process. We first removed intra-structure file redundancies by obtaining a curated set of nucleic acid containing X-Ray crystal structures containing no ASU redundancies from the Nucleic Acid Database (NDB) [144]. Secondly, we removed inter-structure file redundancies by cross referencing our 533 PDB *RNA-only* structures with a computationally inferred non-redundant list of *RNA-containing* clusters of 3D structures, or *equivalence classes* downloaded from the RNA Structure Atlas [133, 145]. Each *equivalence class* is assigned one representative structure. The resulting inter-structure non-redundant set of RNA X-Ray crystals was augmented by including 26 additional structures belonging to *equivalence classes* but that have a ligand that is not identical to that of its representative structure, thus resulting in a set of 206 distinct RNA-only X-Ray structures (Table 2).

Table 2 Complete list PDB ids of non-redundant *RNA-only* X-Ray crystals used in this study.

157D	1QC0	2G91	2XNZ	3FO4	3R4F	466D
165D	1QCU	2G9C	2XSL	3FU2	3RG5	472D
1CSL	1RNA	2GDI	2YIE	3G4M	3RKF	486D
1D4R	1SAQ	2GRB	2ZY6	3GAO	3S49	488D
1DQH	1SDR	2HO7	353D	3GCA	3SD3	4AOB
1DUQ	1T0D	2HOJ	354D	3GES	3SJ2	4B5R
1ET4	1T0E	2HOM	361D	3GLP	3SKW	4E5C
1F1T	1U9S	2HOO	364D	3GOT	3SKZ	4E6B
1F27	1X8W	2HOP	377D	3GVN	3SLM	4ENB
1FIR	1XJR	2IL9	387D	3GX2	3SLQ	4FNJ
1FUF	1XPE	2JLT	397D	3GX3	3SUH	4FRG
1GID	1Y0Q	2NOK	3B31	3GX5	3SUX	4FRN
1H1K	1Y26	2O3V	3BNL	3IBK	3SYW	4GMA
1I9V	1YFG	2O3X	3BNP	3IVN	3SZX	4GXY
1I9X	1YKV	2OE5	3BNQ	3JXR	3TD0	4IQS
1J9H	1YZD	2OIU	3CGP	3LA5	3TD1	4JF2
1JZV	1ZEV	2PN4	3CGR	3LOA	3TZR	4JRC
1KD5	1ZX7	2Q1O	3CGS	3MEI	3U5F	4JRD
1KH6	259D	2QBZ	3CJZ	3MIJ	3U5H	4JRT
1KXK	280D	2QEK	3D0U	3NJ7	402D	4K27
1L2X	2A0P	2QUW	3D2V	3NPN	405D	4K31
1LNT	2A43	2QWY	3D2X	3OWZ	406D	4KQY
1MHK	2A64	2R20	3DIG	3P22	409D	4KYY
1MWL	2AO5	2R22	3DIQ	3P4B	420D	4KZ2
1NBS	2B57	2TRA	3DIR	3P4D	422D	4MGM
1NLC	2CKY	2V6W	3DJ0	3P59	429D	4MGN
1NTB	2D2L	2V7R	3DVZ	3Q3Z	433D	
1NUV	2EEU	2VAL	3E5C	3Q50	435D	
1Q96	2FQN	2VUQ	3E5E	3R1C	438D	
1QBP	2G3S	2W89	3E5F	3R1E	439D	

4.4.2 Symbolic annotation of base pair and backbone connectivity

The set of non-redundant *RNA-only* PDB X-ray crystals were analyzed using DSSR, a new component of the 3DNA suite software programs [124, 125]. We used this command line program to identify all base pairs in the RNA-only set of structures. Base pair

annotations were directly parsed from DSSR's output. Thirteen distinct base pair classes were identified (listed in Table 3 of Results section).

The backbone connectivity between each nucleotide residue was inferred by parsing the PDB file's Primary Structure Section's SEQRES directive [146]. In this section of the structure file, the authors provide the nucleic acid sequence of residues in each chain of the macromolecule. The chain identifiers and residue sequence positions were used to infer backbone connectivity. Taken together, we consider 3 major classes of inter nucleotide residue interactions in this study: link (L), link-pair (LP) and any of the 12 base pair geometric classes.

4.4.3 Identifying ligand neighbourhoods in *RNA-only* structures

In order to determine the complete span of nucleic acid-small molecule interactions in our dataset, we manually inspected the 206 non-redundant *RNA-only* set of structures in search for molecular ligands. In total, we identified 49 non-covalently bound carbon containing ligands (Appendix A, Table 10). This list is composed of molecules that are varied in molecular weight (75-1500 Da) and biological roles; there are 20 naturally occurring or modified amino acids, 13 naturally occurring or derivatives of nucleotide residues and 16 other molecules including 11 co-factors, 3 vitamin complexes, 3 antibiotics and 1 organic dye.

We used BioPython's[22][147, 148] PDB package[23] to compute all inter-atomic distances to every nucleotide residue in our *RNA-only* set of structures. Here we regard the set of nucleotide residues that have any atoms within 5 Å of a ligand's atom as the *ligand neighbourhood* of a structure (Figure 16), *i.e.* all nucleotide residues in a structure that are hypothetically within range of a non-covalent interaction [149]. In our *RNA-only* set, a total of 1030 nucleotide residues were found to form part of one the 49 distinct ligand neighbourhood regions in our non-redundant set of *RNA-only* structures.

[22] http://biopython.org
[23] http://biopython.org/DIST/docs/api/Bio.PDB-module.html

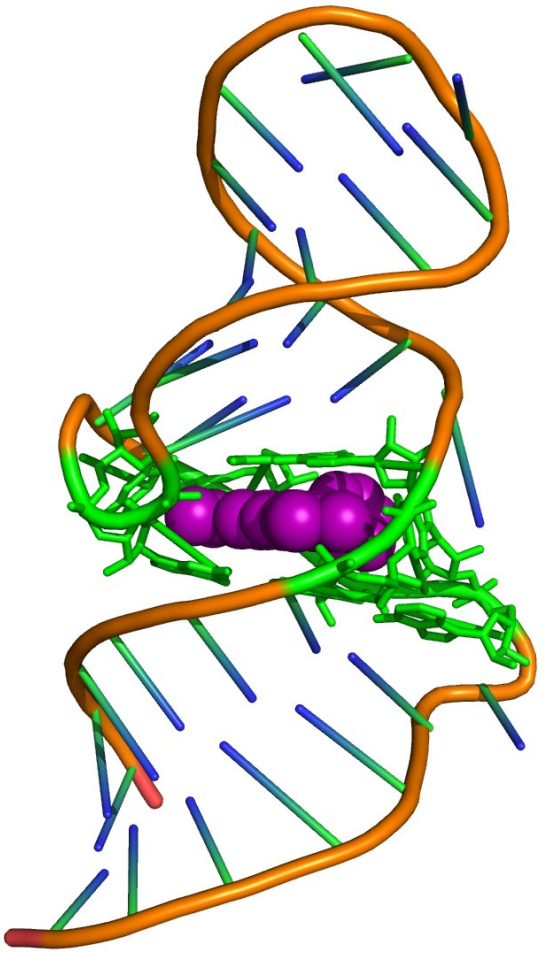


Figure 16 A ligand neighbourhood of Tetramethylrosamine (ROS) in the malachite green aptamer. The set of eight nucleotide residues that are within 5 Å of any atom of the ligand ROS are coloured in green and represented as sticks.

4.4.4 Identifying the Minimum Cycle Basis (MCB) of a nucleic acid graph

The set of base pair interactions between nucleotide residues and backbone connectivity, identified for each of the *RNA-only* X-Ray crystals, was utilized to construct a graph representation of nucleic acid structure. In graph theory, a graph $G = (V, E)$ represents a collection of points and lines connecting some subset of them [127, 150, 151]. The set points are also known as nodes or vertices, while the lines connecting the vertices are commonly referred to as edges. The constructed graphs consider the nucleotide residues of RNA X-Ray crystals as the set of vertices, and the set of inter-nucleotide interactions (base pairs and phosphodiester backbones) as the set of edges connecting them.

We used JGraphT[24] version 0.9.0, an open source Java class library that provides graph-theory objects and algorithms, to construct simple undirected graph representations [151] based on the symbolic annotations of inter-nucleotide residue interactions identified from RNA X-Ray crystal structures. The resulting graphs have no edges that are self-loops (*i.e.* containing edges that start and end at the same vertex), and there is at most a single directionless edge between any two vertices of the graph.

We computed the Minimum Cycle Basis (MCB) of each nucleic acid structure graph to identify the set of smallest indivisible cycles that define each graph. A minimum cycle shares an edge with at most one other cycle in the basis of the nucleic acid structure from which it was computed. We made use of Berger's implementation [127] of the

[24] https://github.com/jgrapht/jgrapht

Horton algorithm [126] available in the Chemical Development Toolkit (CDK)[152] Java library[25] (Figure 17, B and Figure 19).

[25] https://github.com/jctoledo/cdk/tree/master/src/main/org/openscience/cdk/ringsearch

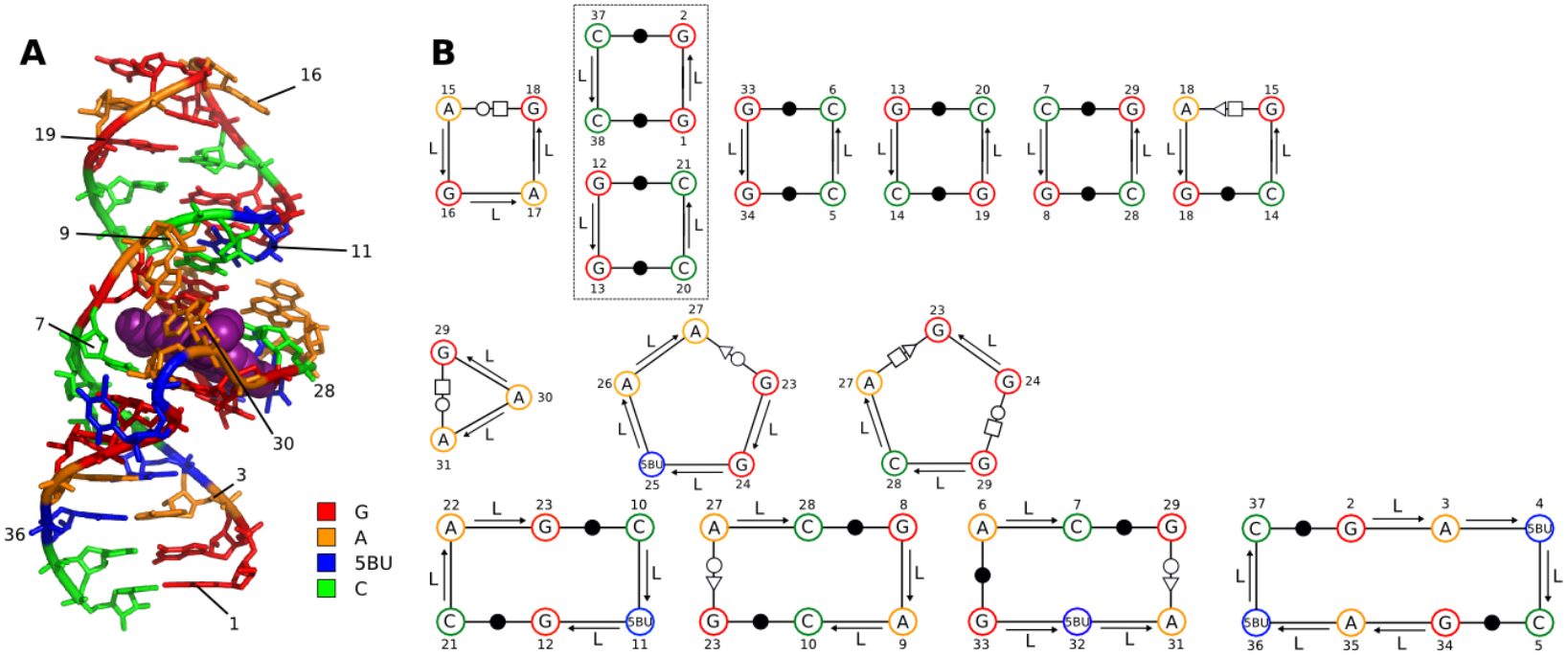


Figure 17 Diagram of the Minimum Cycle Basis of the Malachite green aptamer complexed with tetramethylrosamine. A) Stick representation of the Malachite green aptamer (PDB ID: 1F1T [153]). Backbone of this structure has been represented as a thick ribbon. Tetramethylrosamine is represented as spheres (purple). Nucleotide residues have been coloured by residue type: adenosines (orange), guanosines (red), 5-Bromo-uridines (blue) and cytidine residues (green). B) Complete set of cycles belonging to the MCB of 1F1T. Leontis/Westhof symbolic nomenclature of base pair families is used [49] (Table 3). Phosphodiester bonding is represented with an arrow, directionality of the arrow depicts the backbone orientation. In this structure, there are two four membered minimum cycles have been grouped together in a dashed box, to indicate their topological equivalence.

In order to identify recurrent structural motifs in RNA structure, for each minimum cycle we computed two distinct profile classes or identifiers which are based on the connectivity and identity of the cycle's nucleotide residues and their respective interactions. The first profile considers only phosphodiester backbone linkages and a primitive base pair (*i.e.* inter-nucleobase edge-edge interaction and relative positioning of the glycosidic bond are not given) type, while the second profile distinguishes between all Leontis-Westhof geometric families of base pairs. We call these cycle level 1 profile annotation (L1PA) and cycle level 2 profile annotation (L2PA) classes, respectively. If two cycles have the same number and types of nucleotide residues, and the same inter-nucleotide non-covalent interactions between them, then both of these isomorphic cycles are assigned identical L1PA and L2PA classes, which are identified via unique alphanumeric strings (Figure 17 B). In this way, we utilize minimum cycles to describe recurring sub-structural components of RNA structures.

4.4.5 Generating linked data of macromolecular structures and their annotations

We developed a representation of molecular structure that enables accurate annotation of arbitrary component sub-structures of PDB entries. We made use of the Resource Description Framework (RDF) [54] to implement a data model that enables querying across all non-redundant RNA only structures, their minimum cycles and ligand neighbourhoods using the SPARQL [154] query language.

In generating the RDF, we follow the approach taken for Biological Situational Modeling [66] so as to distinguish between processes, objects and their qualities.

Processes have temporal parts and unfold in time, whereas objects are spatiotemporally extended and exist in whole whenever they exist, thus they have no temporal parts. Here, for example, the process of structure determination, as described in a PDB structure file, involves the generation of a model, which can be considered a temporal "snapshot" of the molecular structure in question. Atoms, ligands, and modified nucleic acid residues are types of objects that are spatially extended, maximally self-connected and self-contained and can bear any number of qualities. Qualities are categorical properties that are intrinsically associated with their bearing entity at all times, but whose observed or measured values may change. For example, the interatomic distances between hydrogen donor-acceptor atoms, the relative orientation of the glycosidic bonds of the two bases or the orientation of the backbone structure are all qualities of a base pair interaction.

Using this modelling approach, we developed 'PDB2RDF'[26], a suite of multiple open-source licensed Java programs to process PDBML [155][27] files to RDF. The generated RDF follows the principles of linked data and ascribes to Bio2RDF's Release 2 practices [156]. Specifically, no blank nodes are used and every resource is typed with a corresponding vocabulary term (when available) from the PDB Exchange data dictionary [155]. RDF representations of every minimum cycle and ligand neighbourhood were also generated following Bio2RDF Release 2 guidelines and loaded into a single instance of Virtuoso Open Source Edition build 6.01.3127[28].

---

[26] https://github.com/bio2rdf/bio2rdf-scripts/tree/release3/pdb
[27] http://pdbml.pdb.org/
[28] https://github.com/openlink/virtuoso-opensource/tree/stable/7

We further annotated every minimum cycle with additional computable information *e.g.* GC content calculation, cycle size and all first degree neighbouring minimum cycles. We consider two cycles as first degree neighbours of each other if they share at least one vertex with another cycle from the structure's computed MCB (Figure 18). In this work, we refer to the set of cycles that share at least one vertex with another cycle *C* that belongs to the same MCB as the First Degree Neighbourhood (FDN) of cycle *C*.

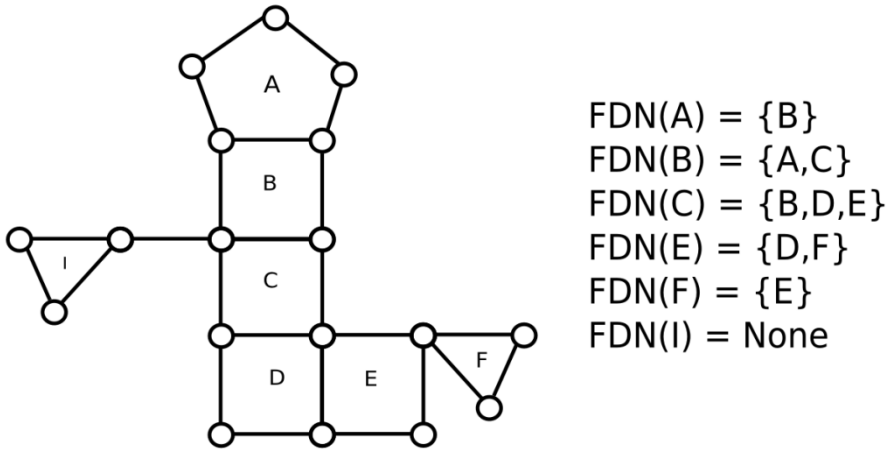


Figure 18 Visual depiction of first degree neighbours (FDN) of a cycle. Two minimum cycles that share at least one vertex are considered to be first degree neighbours. A cycle may have zero or more first degree neighbours.

Each of these attributes is given a unique URI and is directly linked with the minimum cycle they describe (Figure 19).

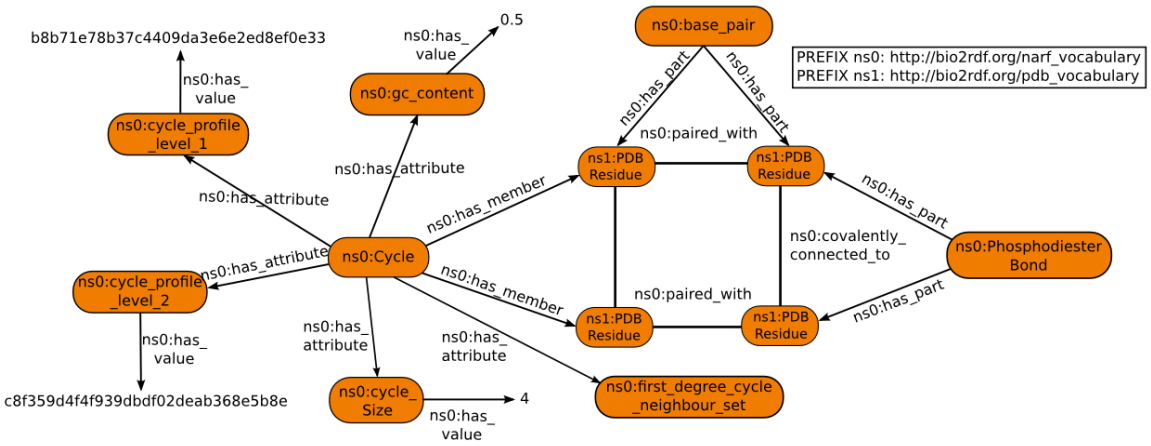


Figure 19 Linked data representation of minimum cycles and their attributes. Every minimum cycle has a unique resource URI that enables further annotation of additional attributes such as GC-content, cycle size, L1PA, L2PA and first degree neighbour set membership.

4.4.6 Supervised learning of RNA structure MCBs and their membership to ligand neighbourhoods

Based on the computation of all Minimum Cycle Bases from every non-redundant X-Ray RNA structure in PDB, we trained SVMs to: i) Predict whether a structure has a ligand neighbourhood or not, based on the structures' MCB composition (herein referred to as the *'Contains ligand neighbourhood' Classifier* or *CLNC*) and ii) Predict whether or not a cycle belongs to a ligand neighbourhood, based on features of the cycle (herein referred to as the *'In ligand neighbourhood' Classifier* or *ILNC*). We made use of Weka (version 3.6.10), a data mining toolkit for analysis of the data and classification experiments [157] to train two Support Vector Machines (SVMs) both running an implementation of the Sequential Minimization Optimization (SMO) [158].

The CLNC was constructed for the purpose of distinguishing RNA structures that contain ligand neighbourhoods from those that do not, based on the cycle composition of their minimum cycle basis. Specifically, for each RNA structure, we generated feature vectors from the frequency of occurrence of each of L1PA or L2PA class in the structure, with an additional class label indicating whether the structure had a ligand neighbourhood ('+') or not ('-'). We used these vectors to train two instances of the CLNC: CLNCi used the L1PA classes as features and CLNCii used the L2PA classes as features. In total, from the computed minimum cycle basis of the 203 non-redundant RNA X-Ray crystals we identified 1714 distinct L1PA classes and 2343 L2PA classes.

The ILNC was trained to distinguish between minimum cycles that are in ligand neighbourhoods and those that are not. For each cycle we thus created feature vectors that consisted of the following: Cycle level profile annotation class (level 1 or level 2), cycle size, Leontis-Westhof geometric family membership of all base pairs in cycle, GC-content, number of phosphodiester backbones in the cycle, the complete set of first degree neighbours of every cycle and an additional label indicating whether the particular cycle was in a ligand neighbourhood ('+') or not ('-'). In this way we trained one ILNC instance that used L1PA classes (ILNCi) and a second ILNC instance that used L2PA classes (ILNCii) for both the individual cycles and their first degree neighbours. The training data consisted of 560 cycles in ligand neighbourhoods ('+') and 7718 cycles not in ligand neighbourhoods ('-').

4.4.7 Assessment of prediction quality

We evaluated the results of the CLNC and ILNC instances with commonly used classification metrics: Sensitivity (*e.g.* true positive rate, TPR), Specificity (*e.g.* true negative rate, TNR), F-Measure and the Mathews Correlation Coefficient:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$F\ Measure = \frac{2 * TP}{(2 * TP + FP + FN)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, FP, FN and TN represent the numbers of true positives, false positive, false negative and true negative respectively. F-measure values provide an geometric average parameter based on the precision (how many of the entities found classified are correct) and recall (how many of the positive classes did the classifier identify). The Matthews Correlation Coefficient (MCC) is used as a measurement of the agreement between the predictions made by a classifier and the actual values. The MCC produces values that range between -1 and 1, where an MCC of -1 indicates an absolute disagreement between the classifier's prediction and the real value, a value of 0 describes random predictions and +1 indicates a strong agreement between the predicted values and their actual class assignments. We also report classifier results with the area under ROC curve (AUC). The value of the AUC score ranges from zero to one, where a score of 1.0 indicates perfect

separation between the classes and 0.5 corresponds to random guesses regarding membership between different classes.

4.5 Results

In this section, we first present general metrics identified in computing the minimum cycle bases of 206 structurally non-redundant RNA X-Ray crystals. We then describe SVM classification results. Specifically, we investigate the discriminatory power of an MCB computed on an RNA structure in terms of its ability to distinguish structures that contain *ligand neighbourhoods* from those that do not, and evaluate an additional classifier constructed to differentiate between minimum cycles that form part of *ligand neighbourhoods* from those that do not.

4.5.1 Descriptive statistics of minimum cycles of RNA X-Ray crystals

The set of *RNA-only* PDB structures used in this study was comprised of 206 X-Ray crystals composed by 536 distinct macromolecular chains. We identified 7606 base pair interactions and 22797 distinct phosphodiester backbone bonds.

Table 3 describes the thirteen distinct base pair classes identified by DSSR. The Leontis-Westhof nomenclature scheme describes base pair interactions in terms of the positioning of at least *two* hydrogen bonds around the edges of each participating nucleotide residue's structure. Three distinct edges for hydrogen bond interactions, two relative orientations of the glycosidic bonds (*cis* and *trans*) and the corresponding orientation of the nucleotide backbone are used in this scheme to produce 12 distinct geometric classes that categorize 93.41% of all identified base pair interactions. Notably, DSSR annotations also provide descriptions of an additional 501 non-canonical base pair interactions that cannot be described using the LW nomenclature. These base pairs contain least *one* hydrogen bond between two nucleotide residues. This set includes reverse Hoogsteen and sheared G-A, among others [124, 125]. Here, we categorize these base pair interactions simply as "base pairs" and as such make use of the corresponding coarse RNAO class [68] [29] for a generic base pair interaction in which there are no annotations for any nucleobase edge-edge interactions, relative positioning of glycosidic bond or strand orientation. On average *RNA-only* structures are composed by 34.49 base pairs (with a range of 3 to 1455 base pair interactions, found in G-quadruplex PDB ID:3MIJ [159] and *Saccharomyces cerevisiae* 60S ribosomal PDB ID:3U5H [160], respectively) and 110.665 phosphodiester bonds (with a range of 10 to 5529 backbone interactions also for PDB ID:3MIJ and PDB ID:3U5H, respectively) per structure.

[29]RNAO's Base pair URI: http://purl.obolibrary.org/obo/RNAO_0000001

Table 3 Counts of the thirteen distinct base pair classes identified by DSSR in the set of 206 RNA-only PDB structures used in this study

Base Pair Class	Bond orientation	Interacting edges	Strand orientation	Symbol	RNAO Class URI	% total	Count
Base Pair	N/A	N/A	N/A	N/A	http://purl.obolibrary.org/obo/RNAO_0000001	6.59	501
LW 1	Cis	WC/WC	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000003	77.52	5896
LW 2	Trans	WC/WC	Parallel		http://purl.obolibrary.org/obo/RNAO_0000004	0.24	18
LW 3	Cis	WC/H	Parallel		http://purl.obolibrary.org/obo/RNAO_0000005	1.92	146
LW 4	Trans	WC/H	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000006	1.47	112
LW 5	Cis	WC/S	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000007	2.12	161
LW 6	Trans	WC/S	Parallel		http://purl.obolibrary.org/obo/RNAO_0000008	2.08	158
LW 7	Cis	H/H	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000009	0.05	4
LW 8	Trans	H/H	Parallel		http://purl.obolibrary.org/obo/RNAO_0000010	0.75	57
LW 9	Cis	H/S	Parallel		http://purl.obolibrary.org/obo/RNAO_0000011	1.38	105
LW 10	Trans	H/S	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000012	4.29	326
LW 11	Cis	S/S	Antiparallel		http://purl.obolibrary.org/obo/RNAO_0000013	0.49	37
LW12	Trans	S/S	Parallel		http://purl.obolibrary.org/obo/RNAO_0000014	1.12	85
						Total	**7606**

The complete set of minimum cycle bases for our *RNA-only* structures were composed of a total of 7738 distinct cycles. While the computed minimum cycles had a wide range of sizes (between 3 and 143), most cycles (96.4%) had 10 or fewer nucleotide residues with the majority (77.83%) of computed cycles were of size[30] four (Figure 20). On average there are 37.56 minimum cycles per *RNA-only* structure.

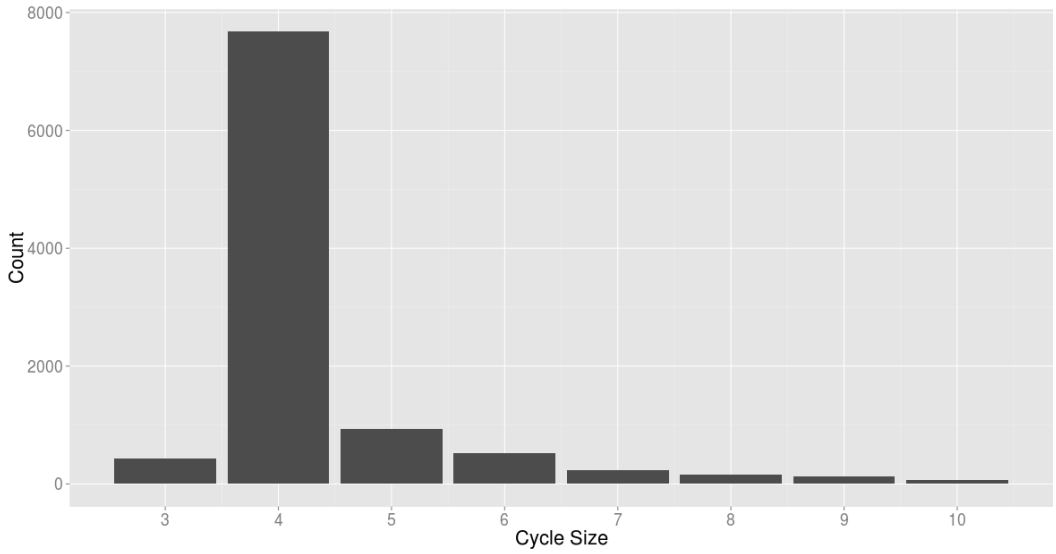


Figure 20 Minimum cycle size distribution in RNA-only X-ray crystals. Only cycles of size 10 or less are shown which corresponds to 96.4% of all cycle sizes.

For each minimum cycle we also identified the set of first degree neighbor cycles (FDN). As expected, there was a single FDN set for each of the 7738 computed minimum cycles, due to the connected nature[31] of the graphs from which the minimum cycles were

[30] Here, we refer to the number of nucleotide residues that compose a cycle as its "size".
[31] A graph is considered connected if there is a path (*e.g.* a set of edges that connect vertices) from *any* vertex to *any* other vertex in the graph.

computed. The majority (~32.3 %) of computed set of FDN have exactly 2 members (Figure 21).

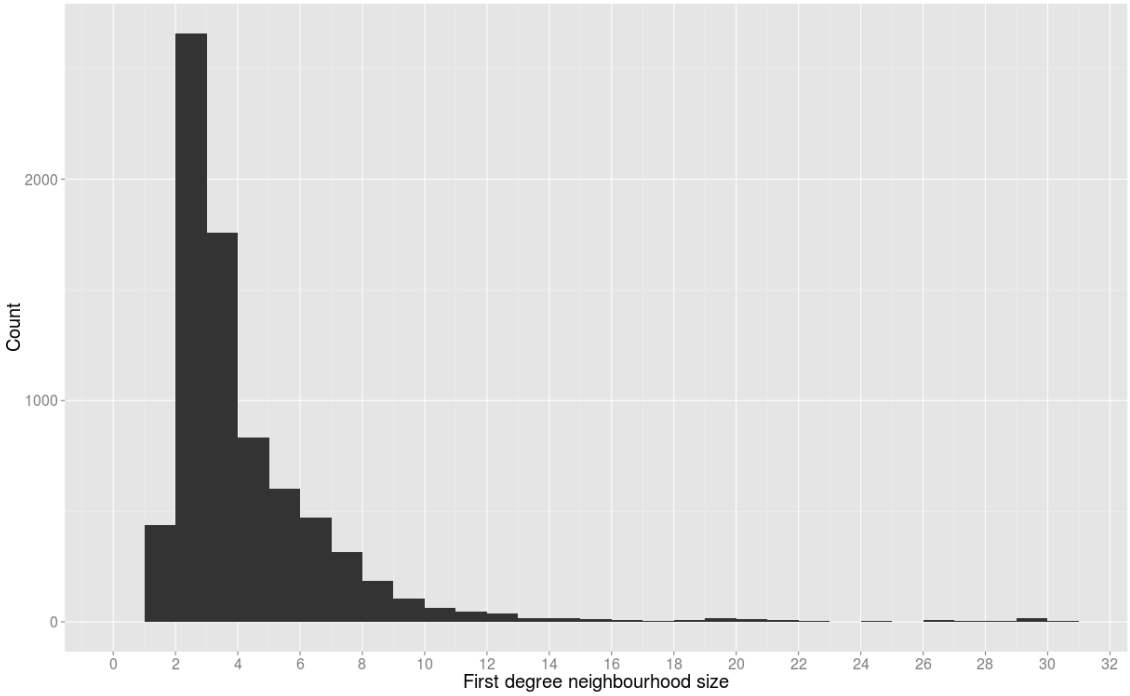


Figure 21 Frequency plot of minimum cycle first degree neighbourhood size

For the set of 49 ligands that were identified (Appendix A, Table 10), we computed 54 distinct ligand neighbourhoods which were composed by 651 nucleotide residues. This set of ligand neighbourhoods are described by a total of 560 minimum cycles which ranged in size from 3 to 45 nucleotide residues per cycle.

4.5.2 Minimum cycles are structurally redundant partitions of nucleic acid structures

Computing the minimum cycle bases of ribonucleic acid structures produced sets of structurally redundant partitions. Minimum cycle redundancy was found both *within* the

MCB of a given structure and *between* the MCBs of homologous and non-homologous RNA structures. For example, the four membered cycle composed of two guanine nucleotide residues that are backbone covalently connected to each other, form two anti-parallel stacked canonical G-C base pairs with two backbone covalently connected stacked cytosine residues (Figure 22 A), occurs 404 times across 131 different RNA-only non-homologous structures. We identify this minimum cycle using its L1PA and L2PA identifiers "aceea01cdae796d8c15cb6c4421c32d5" and "445c69d2714e43b727182b7496657eb1" respectively. Instances of this type of cycle can be observed in functionally diverse RNA structures (*e.g.* this cycle may be observed as part of a helical stem in the human thymidlate synthase mRNA (PDB ID:3MEI [161] and as a part of the viral RNA pseudoknot structure (PDB ID:2A43 [162]). Similarly, different minimum cycles of varying sizes and topologies recur within the RNA-only dataset. Figure 22 B shows a five member cycle that contains a C-G-A high-order base pair pivoted around a guanine residue and two parallel, independently stacked, backbone links between an adenine and a cytosine residue and between a uracil and an adenine nucleotide residue enclosed by an antiparallel cis Watson-Watson A-U base pair that occurs in four purine riboswitches (*e.g.* PDB IDs: 2XNZ [163], 3GAO[164], 1Y26[165] and 3LA5[166]). Minimum cycles are also representationally equivalent to commonly used partitions of nucleic acid structure. Figure 22 C shows the minimum cycle that describes a hairpin loop structure composed of 5 nucleotide residues enclosed by a trans A-U Hoogsteen-Watson base pair which occurs in 13 RNA-only structures (*e.g.* PDB IDs:2HOO [167] and 2A64 [168] among others).

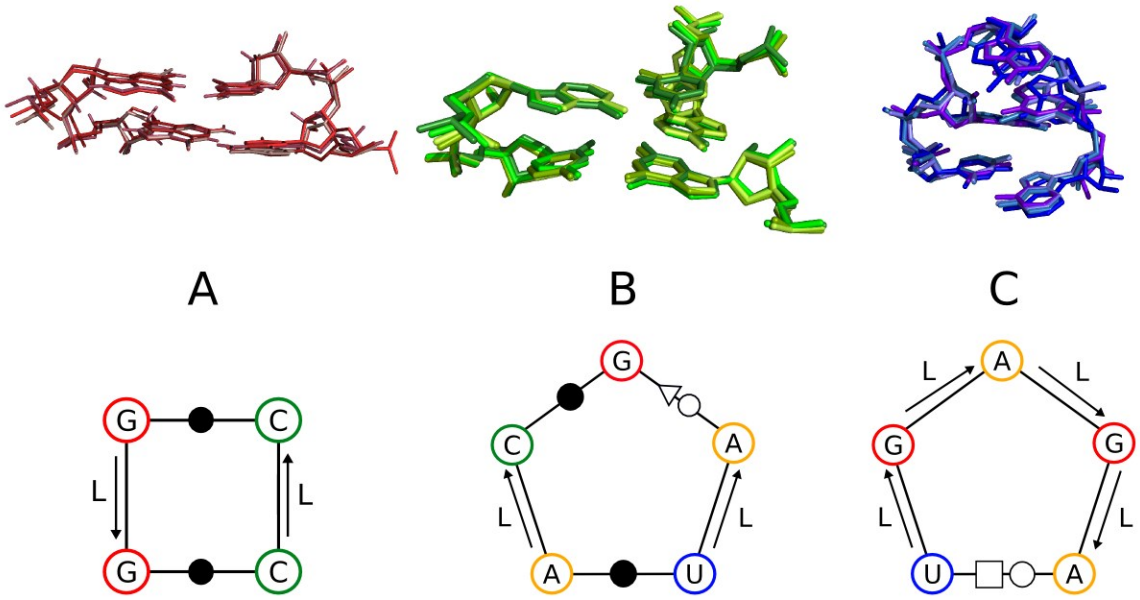
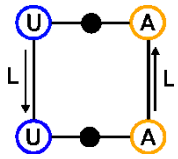
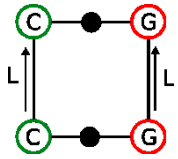
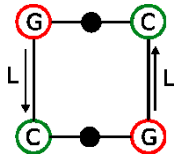
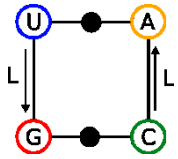
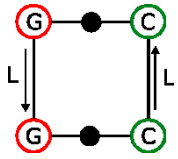


Figure 22 Examples of minimum cycle structural redundancy. A, B and C show the superimpositions of three distinct cycles classes each across 3 different RNA-only structures. Stick representations of the 3D conformation of each cycle are shown above a 2D representation of the corresponding cycle features including LW base pair interactions, identity of residues and directional backbone interactions.

The complete set of 7738 minimum cycles is described by 2346 unique L2PA cycle classes and 1717 distinct L1PA classes. Table 4 lists the ten most frequently occurring L2PA cycle classes, the total number of times they occur across all structures and the total number of distinct structures in which they were observed. Table 5 shows the same metrics for the top ten L1PA cycle classes. Notably, the most frequently occurring L2PA cycle occurs in more than 20% of all RNA structures and several other cycle classes occur across more than 50% of structures. Similar trends are observed for L1PA cycle classes, thus quantitatively demonstrating the recurrent nature of classes of minimum cycles across RNA structures.

Table 4 Ten most frequently occurring Level 2 Profile Annotation (L2PA) minimum cycle classes.

2D Cycle representation	L2PA identifier	Cycle size	Total # of occurrences of cycle in RNA-only structures	Total # of different structures in which cycle occurs
	9416250e526a6acde942c769ba8997d2	4	493	46
	445c69d2714e43b727182b7496657eb1	4	404	131
	569b5cb77a752990858da2e000ccbca9	4	327	142
	067637a49b5c6ae1da07a199cd31857e	4	318	110
	d31dd7f5d2954c3b77ae96cbcd620797	4	297	109

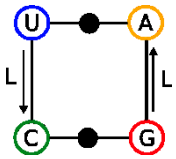
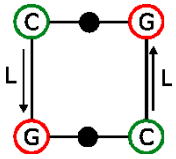
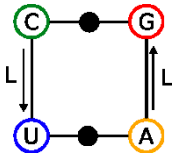
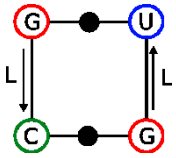
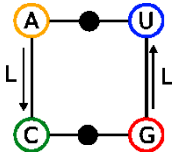
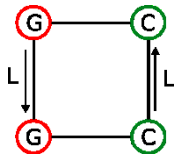
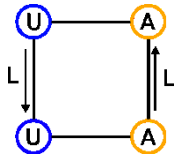
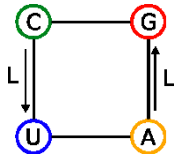
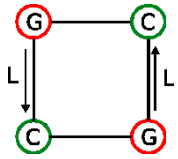
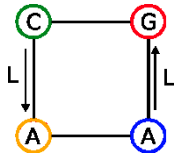
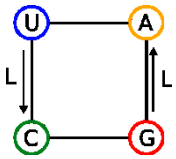
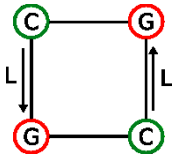
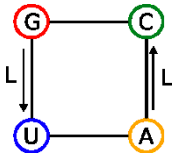
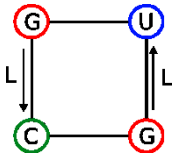
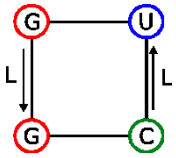
2D Cycle representation	L2PA identifier	Cycle size	Total # of occurrences of cycle in RNA-only structures	Total # of different structures in which cycle occurs
	6d8dafded5e229079feda37224b1659f	4	273	97
	fd8442c54f213e1ed9c707c049e97566	4	266	120
	6c91ba0d4f95e89e7c5929cd6027bd53	4	186	76
	62eb27df7094cc957b77b7ed31ebca96	4	139	65
	f0fb386df526c7f38f12ddbaadff5e8c	4	135	78

Table 5 Ten most frequently occurring Level 1 Profile Annotation (L1PA) minimum cycle classes.

2D Cycle representation	L1PA identifier	Cycle size	Total # of occurrences of cycle in RNA-only structures	Total # of different structures in which cycle occurs
	aceea01cdae796d8c15cb6c4421c32d5	4	705	148
	0cfa77f58ebffecf8c4c35cbc67acf63	4	596	57
	053234f5dbc296fae20e7339e1bbb605	4	331	101
	180cc1c7d9ac3a8126c378a15c6fa4c1	4	328	142
	5405b2aa15ea41f0f3a84b19f6645515	4	322	111

2D Cycle representation	L1PA identifier	Cycle size	Total # of occurrences of cycle in RNA-only structures	Total # of different structures in which cycle occurs
	4030e1763b588205a904a1dba1412dfd	4	275	97
	dfcb33bcb31dd679e0f357391bdffb85	4	267	120
	2c6cc2a4c268b20b327d6ee551544fe2	4	265	101
	833ab01947ebc04c886c27cc5de2293b	4	141	67
	365bd0babfb6ba1136cd5aa61fc4275e	4	115	45

4.5.3 Performance of binary classification for presence of a ligand neighbourhood in RNA structures

Based on the L1PA and L2PA class composition of the complete minimum cycle bases computed for each RNA-only structure, we trained two SVM classifiers – CLNCi and CLNCii – and evaluated their performance using 10 fold cross validation to classify RNA structures into one of two classes: contains a ligand neighbourhood (+) or does not contain a ligand neighbourhood (-). The CLNCi used the individual counts of each of the 1717 distinct L1PA classes as features in each structure's computed MCB and CLNCii used the frequency of each of the 2343 unique L2PA classes. In total, the set of 206 RNA structures contained 54 examples of ligand neighbourhood containing structures ("+" class) and 152 that did not contain a ligand neighbourhood ("-" class).

Table 6 shows the results of both ligand neighbourhood classifiers. Both CLNCi and CLNCii achieved precision values greater than 0.89 and recall rates of ~ 0.9, and also had significant values for the Mathews Correlation Coefficient of 0.72 or better.

Table 6 Results of CLNC classification of small molecule containing RNA structures.

Classifier	CLNCi	CLNCii
TP Rate	0.898	0.908
FP Rate	0.215	0.212
Precision	0.896	0.908
Recall	0.898	0.908
F-Measure	0.895	0.904
ROC	0.841	0.848
MCC	0.727	0.753

4.5.4 Performance of cycle centric classification of RNA minimum cycles

We also assessed the performance of a cycle centric classifier, ILNC, that predicts whether or not a given cycle is in a ligand neighbourhood or not. Given the uneven distribution of instances of the 1717 L1PA and 2143 L2PA cycle classes, where 73.15% of L1PA classes and 77.62% of L2PA classes have only one instance, we wanted to assess the effect of training the ILNC on five different sets of training data which vary the minimum number of instances of each profile annotation (L1PA or L2PA) class included in the training data from 0, 1, 10 up to a minimum of 100 instances of each cycle class. The '0' training set contains all cycle classes regardless of their number of instances, the '10' training set consists of only cycle classes that have at least 10 instances across the RNA structures, and so on.). Table 7 shows the classification results when varying the minimum number of instances of each profile annotation classes. Classifier performance is better when trained on the data that includes all cycle classes, compared to those that only contain more frequently occurring cycle classes. Importantly, classifier performance

deteriorates considerably in the reduced datasets which include only profile annotation classes that occur 10 or more times.

Table 7 Performance of the ILNC in 10-fold cross validation for both cycle profile annotation levels (L1PA and L2PA cycle classes).

Redundancy Level	**All**		**> 1**		**> 10**		**> 100**	
ILNC	**i**	**ii**	**i**	**ii**	**I**	**ii**	**i**	**ii**
# of class instances	1717	2346	461	525	74	64	11	12
TP Rate	0.979	0.981	0.974	0.972	0.955	0.959	0.963	0.969
FP rate	0.199	0.186	0.287	0.304	0.823	0.767	0.964	0.969
Precision	0.979	0.98	0.973	0.971	0.946	0.949	0.93	0.939
Recall	0.979	0.981	0.974	0.972	0.955	0.959	0.963	0.969
F-Measure	0.979	0.98	0.972	0.97	0.939	0.948	0.946	0.954
ROC	0.89	0.897	0.843	0.834	0.566	0.596	0.5	0.5
MCC	0.838	0.849	0.787	0.772	0.303	0.343	-0.006	-∞

We also used Weka to evaluate the performance of our classifiers with reduced sets of features. In particular, we ran a correlation feature selection procedure [169] to reduce the set of 1717 and 2343 features used in CLNCi and CLNCii to 20 and 23 features respectively (Table 11 and Table 12). The performance of these classifiers is shown in Table 8. Performance is slightly improved for both, compared to CLNCi and CLNCii which used the complete feature set.

Table 8 SMO performance on reduced data for CLNC correlation feature selection

CLNC	i	ii
TP Rate	0.922	0.947
FP rate	0.207	0.15
Precision	0.927	0.95
Recall	0.922	0.947
F-Measure	0.918	0.945
ROC	0.858	0.898
MCC	0.795	0.862

Using CFS we also reduced the 1735 ILNCi and 2364 ILNCii features sets to 36 and 40 respectively (Table 13 and Table 14). The performance of these classifiers is shown in Table 9. The performance of these classifiers is marginally worse compared to their full feature set counterparts.

Table 9 SMO performance on reduced data for ILNC correlation feature selection

ILNC	i	ii
TP Rate	0.971	0.973
FP rate	0.279	0.246
Precision	0.969	0.971
Recall	0.971	0.973
F-Measure	0.969	0.972
ROC	0.846	0.863
MCC	0.764	0.783

4.6 Discussion

In this work we make use of a structurally non-redundant set of *RNA-only* structures deposited in the PDB to develop a methodology for identifying recurring sub-structural components of these macromolecules that may contribute to ligand binding. Inspired by the previous work of Lemieux & Major [122], we constructed graph representations of

nucleic acid structure that are based on the nucleotide residues that compose the structure and the symbolic annotation of all inter-nucleotide residue interactions. For every structure's graph representation we computed the set of indivisible fragments via the computation of the minimum cycle basis [126, 127]. While Lemieux & Major computed cycles for only one large ribosomal subunit, and hierarchically clustered these cycles by RMSD, we compute cycles across *all* RNA structures in PDB and use these cycles to predict ligand binding. Specifically, using minimum cycles extracted from these graphs in conjunction with annotations of their occurrence in ligand neighbourhoods, we trained two sets of SVM classifiers to predict i) structures that contain ligand neighbourhoods and ii) minimum cycles that occur in ligand neighbourhoods. In the following sections, we proceed to discuss our results and highlight the unique features of our approach in comparison to previous related work in naming and identifying repeating tertiary elements of RNAs. We also describe challenges encountered and areas for improvement in future extensions of our approach.

4.6.1 Evaluating classifier performance

The results obtained from our CLNC show that the MCBs of RNA structures that have ligand neighbourhoods are distinct from RNA structures that do not have them. This suggests that minimum cycles may be structurally related to ligand binding. Both CLNCi and CLNCii produced correct predictions in spite of the numeric imbalance of between number of positive ('+') and negative ('-') examples in our training data. The moderate value for the false positive rate of ~ 0.2 taken together with a strong MCC of ~ 0.7 indicates that predictions made by both CLNCi and CLNCii are reliable.

Our results from both ILNCs show that cycles that belong to ligand neighbourhoods are distinct from those that do not and thus their features can be used to correctly predict whether or not a cycle is in a ligand neighbourhood. Both level 1 and level 2 profile annotation classes were unevenly distributed in our data. Specifically, approximately 70% of cycle classes have only one instance and only ~6.5% of cycle classes have 100 or more instances. We were thus interested in determining the effect of *rare* cycle classes on classifier performance for both L1PA and L2PA classes. As shown in Table 7, using only cycle classes that had a frequency of occurrence of more than 1 resulted in a small decrease of classifier performance in comparison to using all cycle classes. This decrease in performance was amplified when using only cycle classes that occurred either more than 10 or more than 100 times in our dataset. This performance decrease is indicated by the reduction in MCC values despite the high values for precision and recall that are a result of correct identification of the negative ('-') class, which represents the majority of instances. These results suggest that *rare* cycle classes play an important role in predicting whether or not a cycle is in a ligand neighbourhood.

4.6.2 Cycles as RNA motifs

Traditionally the notion of RNA 3D motifs in the literature has been focused around the human understandable description and identification of elements of tertiary structure that have been repeatedly *observed* within and across various nucleic acids that possess biological activity [170, 171]. Examples of these motifs include: stems, hairpins, loops, junctions, pseudoknots, kissing-hairpins, kinks and G-quadruplexes. Lacking in the

prevailing descriptions of these motifs is a computable definition of the elements and the respective interactions that describe them, thereby hindering the reuse of their definitions by others who may be interested in reproducibly identifying these substructures of RNA. In contrast, minimum cycle bases provide granular and indivisible definitions of repeated substructures that are inherently computable and may be combined to define other higher order structures. Importantly, minimum cycles meet the criteria of canonical RNA motifs, in that they recur *within* and *across* structures and may be used to discriminate substructures that have biological activity. Minimum cycles also enable quantitative comparisons of substructures that they define.

Sarver et al. [115] and Petrov et al. [171] have identified potential shortcomings of using symbolic annotations for pairwise interactions of nucleic acids. Specifically, resolutions of X-Ray crystals can affect the symbolic annotations of base pair interactions, as they depend on the identification of a limited set of hydrogen bonding interactions the accuracy of which depends on resolution. It is difficult to quantify potential annotation errors introduced by low resolution structures, but we propose that any such errors may be balanced by the size and diversity of the RNA structure dataset that we have produced and used for classifier training especially given that 90% of the 206 *RNA-only* structures in our dataset are of moderate resolution (≤ 3 Å) or better. Sarver et al. [115] also highlight the potential issue that motifs can only contain symbolic annotations that are already known. We observed this phenomenon in our dataset where 501 base pairs (6.59%) were not classified as any of the 12 LW classes and so the cycles

containing these base pairs could only use a generic base pair class to label the interaction. However this only occurred in only a minority of cases.

4.6.3 The challenges of extracting knowledge from PDB data

The PDB format captures details of experimental procedures (X-Ray crystallography, solution NMR, solid state NMR and cryo-electron microscopy), information about the extraction of the macromolecules, references to scientific literature and information related to the internal book keeping of records. Arbitrarily extended over time, the current PDB specification is challenging to understand (and correctly parse). Despite the release of an XML version of the PDB file format (PDBML) and the corresponding PDB Exchange dictionary [155], the format does not explicitly capture the intended semantics, nor does it accurately capture the biological or informational relationships between described entities.

To augment PDB, the PDBeChem database was created to provide supplementary data including chemical definitions for standard and modified amino acids, nucleic acids, drugs, inhibitors, cofactors and other chemical species often included with other PDB records [172]. However, neither PDB nor PDBeChem explicitly identifies the biological (or otherwise) role of all monomeric chemical components that appear in PDB structures files. While PDBeChem provisions chemioinformatic annotations of monomeric components of structures, it heavily relies on their user base to keep up with the exponential growth of the PDB. As a result, we had to manually inspect all *RNA-only* structure files in conjunction with the corresponding PDBeChem records to

determine if any of the carbon-containing chemical components not covalently connected to the nucleic acid were in fact ligands as opposed to ions, experimental artifacts or modified residues, and then compute which nucleotide residues were in their vicinity (as described in our Methods section). Our ligand annotation process thus resulted in the contribution of 42 new entries to PDBeChem from Table 10. This poses a potential bottleneck for future extensions of this work.

4.6.4 Feature selection and inferring causality from SVM classifiers

The main goal of machine learning is to use classifier predictions as guides in an experimental setting, rather than as indicators of causality between the selected features and the identified class [173, 174]. Indeed, in the case of support vector machines, methods to extract causal features based on the weights used in trained models are a developing area of research. It has been shown that features which are given high weights by SVM models are not necessarily those with the most significant causal relationships to the output labels [173]. Thus, while it is tempting to infer causality for the higher weighted features in our CLNCs and ILNCs and thus suggest cycle features responsible for ligand binding, it is not correct to do so. It may be the case that the higher-order interactions of the features are responsible for their excellent classification performance, as opposed to any subset of features in isolation.

Nonetheless, feature selection is an important area of machine learning [175] concerned with increasing the cost-effectiveness of classification schemes constructed using high dimensional data by reducing the feature space (*e.g.* number of training

features used to classify training data) and thus providing faster and in some cases, better performing classifiers. In considering the high dimensional nature of the features used in both CLNC and ILNC, which ranged between 1717 and 2346 for both CLNCi and CLNCii and between 1735 and 2364 features for both ILNCi and ILNCii respectively, we made use of a correlation-based feature subset (CFS) selection [169] to identify a subset of features that would obtain similar performance levels but with a fraction of the time required to train both CLNC and ILNC (Table 8 and Table 9).

Our results show that using CFS selection on CLNC reduced the feature space to only 20 and 23 features for CLNCi and CLNCii respectively, which in turn moderately increased the predictive power of both of these classifiers as reflected by marginal increases in MCC. Specifically, CLNCi's MCC improved from 0.727 to 0.795 and CLNCii's MCC improved from 0.753 to 0.862. Similarly, we used CFS selection to identify a reduced set of only 36 and 41 features for ILNCi and ILNCii respectively. In this case performance of both ILNCs decreased – values of MCC were reduced by ~0.074 and ~0.065 for ILNCi and ILNCii respectively. Based on the minor differences in performance resulting from feature selection coupled with the significant decrease in processing time, it will be beneficial to perform feature selection for future versions of our classifiers as the amount of available training data in PDB continues to increase.

4.7 Future work

4.7.1 Alternative profile classes for minimum cycles

As is the case for proteins, the three dimensional structures of homologous RNAs are more conserved than their primary sequences [35, 36, 176, 177]. Specifically, when considering the C1'-C1' distances between each of the canonical base pairs' (*i.e.* A-U, U-A, G=C and C=G base pairs) nucleotide residues may be very similar when both residues have the same relative orientations of the glycosidic bond and as such, they can replace each other without drastically changing the three dimensional space they occupy and relative geometric orientations of their respective backbone interactions. These base pairs are referred to as 'isosteric' [170, 176]. Currently our cycle profile annotation classes level 1 and 2 (L1PA and L2PA) are defined in part by the nucleotide residues that compose them. Consequently, cycles containing isosteric base pairs are given different L1PA and L2PA class identifiers. However, the proposed methodology for annotating features of minimum cycles enables the creation of alternative cycle profile classes to contain serializations of isosteric base pairs provided the measurements of C1'-C1' distances and verification of relative orientation of glycosidic bonds fall within similar ranges.

Similarly, our cycle profile annotation classes do not currently capture either adjacent or non-adjacent nucleobase stacks. In the future we intend to capture these relationships in the cycle profile class definitions and assess the resulting effect on classifier performance.

4.7.2 Extrapolating features learned from minimum cycle classes into predictions of ligand binding in aptamers

Given that secondary structure predictors [103, 178] do not provide information about the interacting nucleotide residue edges or annotate the relative positioning of the glycosidic bonds in the provided predictions, we will use level 1 profile annotation (L1PA) classes generated from our ILNC to annotate secondary structure predictions of over 800 RNA aptamers found in Aptamer Base [179][32] and then predict which individual cycles may be part of ligand neighbourhoods. However, further analysis must be done in order to assess the structural overlap that exists between minimum cycles extracted from graphs constructed from secondary structure predictions and those extracted from three dimensional structures. We suspect that there may exist populations of L1PA cycles constructed from secondary structure predictions that do not have a correspondence in any observed RNA three dimensional structures.

4.8 Conclusions

This work provides a reproducible method for identifying structural redundancy in the three dimensional structure files of ribonucleic acids deposited in PDB. This set of redundant sub-structures of RNA molecules which were derived from the computation of all minimum cycle bases (MCB) of graph representations of nucleic acid structures, not only comprise a set of distinctly connected elements of nucleic acid structure that recur *within* and *across* different RNA molecules, but we also demonstrate that recurring

[32] http://aptamer.freebase.com

minimum cycles may be utilized to train a classifier that accurately categorize RNA structures in terms of their ligand binding characteristics.

5 Chapter: **Future directions and summary of contributions**

5.1 Future directions

As new structures of RNAs are solved, the known repertoire of structural motifs identified from these nucleic acids will also continue to grow. As noted in Chapter 4, the rich variety of RNA base pairs that may be observed in RNA structures, is more than the 12 basic geometric classes of base pair interactions proposed by Leontis-Westhof [49] can meaningfully represent [125]. In total we identified over 550 base pair interactions with either one hydrogen bond or bifurcated hydrogen bond interactions that were annotated generically as 'base pairs' and no further information is given about the relative positioning of the components of the participating nucleotide residues. This is because we lack an RNAO or RKB class to describe these particular base pair interactions. Thus, future work on the RKB will focus on extending the RKB ontology to provide machine understandable descriptions of these base pair interactions. In addition, a second future activity will be to extend the size of the knowledge base by instantiating new RNA structures as they are deposited to the PDB.

The current data model that is used in Aptamer Base enables its users to accurately describe the experimental results of SELEX experiments and the resulting aptamer sequences. The flexible nature of the data model used by Aptamer Base will enable me to further explore extensions to the types of data stored therein. Specifically, with the growing number of different types of non-coding RNA sequences [9, 180, 181] that are being identified I propose to extend the Aptamer Base to include descriptions of non-SELEX sourced aptamers and also include descriptions for other non-coding RNA

sequences. This would facilitate attempts to elucidate structure-function patterns in non-coding RNA sequences from multiple different sources.

As minimum cycles are derived from symbolic annotations of inter-nucleotide interactions that are observed in RNA structure files, the types of annotations that we collect will affect which minimum cycles may be identified. In our current implementation, we catalogue 14 unique types of edges that connect nucleotide residues – 13 base pair classes and a single backbone link class. We would like to investigate the effect of including annotations of inter-nucleobase stacking interactions. As not all stacking interactions occur exclusively between adjacent nucleotide residues [73], the addition of base stack edges may certainly result in minimum cycles that we are currently not able to identify. Specifically, as the graph representations from which minimum cycles are derived are dependent on the annotations of nucleotide residues and their respective interactions, adding a new type of edge will expectedly result in graphs of different topologies and consequently result in the identification of new minimum cycles.

In order to enable an accurate quantification of the extent of structural redundancy that we have observed in symbolically annotated RNA motifs via their minimum cycles, we would also like to compute the structural alignment of all instances of L1PA and L2PA classes that we identified from minimum cycles. Specifically, for each level profile class we propose computing the Root Mean Squared Deviation (RMSD) of superimpositions of atomic coordinates of nucleotide residues belonging to identical cycle classes.

Graph representations of nucleic acid structure described in this thesis can be generated from a variety of structural annotations, including symbolic annotations of 3D structure but also less granular annotations extracted from secondary structure descriptions such as dot-bracket notations (DBNs)[33]. Specifically, it is possible to identify minimum cycle bases for nucleic acid structures using graph representations constructed directly from secondary structure annotations, whether the structure was predicted or not. In this way, RNA structural motifs that are captured from annotations of three dimensional structures files may be topologically equivalent from ones identified from secondary structures. Aptamer sequences deposited in Aptamer Base do not have solved 3D structures equivalences in PDB and methods for 3D structure prediction directly from sequence information remain to be developed [182]. However, many secondary structure predictors exist, such as RNAfold [103], that achieve sensitivities as high as 70% in predicting base pair interactions for short (< 200 bases) nucleic acid sequences [183-185]. I propose computing minimum cycles from graphs constructed from secondary structure predictions of aptamer sequences deposited in Aptamer Base and to use the classifiers presented in Chapter 4 to identify minimum cycles that are predicted to belong to ligand neighbourhoods.

Secondary structure annotations can be used to construct L1PA cycle classes because these classes describe groups of minimum cycles in terms of only one base pair interaction, one type of backbone link and over 550 residue types. In L1PA classes I

[33] DBNs are a commonly used syntax to serialize base pairing and backbone interactions in nucleic acids, in this representation a base pairs are represented by balanced brackets "(" and ")" and single stranded regions or loops by dots ".".

exclusively use the RNAO class for a generic 'base pair' interaction[34], and as such create an L1PA identifier that does not consider the relative positioning of the glycosidic bond between the nucleoside residues participating in any base pair interaction. Using this level of annotation for minimum cycles enables the use predictions made by both CLNCi and ILNCi classifiers to be applicable to cycles constructed from predictions of secondary structures. With this approach, I will quantify the role of secondary structure based motifs in aptamer ligand interactions.

The instances of base pairs identified via RKB classes as described in Chapter 2, the aptamers and SELEX experiments in Aptamer Base as described in Chapter 3, and the minimum cycles and their features as described in Chapter 4 are amenable to being added to the Bio2RDF Linked Data network [156, 186]. In this way, this data may be reused in further extensions of this work, directly queried via publically available SPARQL endpoints and also linked to other Bio2RDF data such as PDB2RDF and to bio-ontologies that describe experimental settings such as the Ontology for Biomedical Investigations (OBI) [187].

Research done in aptamer biochemistry in the past 25 years has been primarily focused on identifying instances of short nucleotide sequences that bind to a given molecular target (*e.g.* small molecules, proteins, viruses, cells, etc.). Resulting publications typically describe only aptamers that were found to produce "acceptable" ranges of affinities usually measured in the nanomolar (10^{-6}) or lower dissociation

[34] http://purl.obolibrary.org/obo/RNAO_0000001

constant ranges. While the work reported in this body of scientific literature has enabled the elucidation of an ever growing number of aptamer based biosensors that have therapeutic applications, there is a void in the publication and reporting of aptamer sequences that were found to be poor binders. The resulting bias in aptamer related literature is two-fold: i) aptamer scientists do not usually report negative results, *i.e.* experimental details of SELEX experiments that did not produce "significant" results and ii) when publications are made about highly affine aptamers-target interactions reported from the results of a SELEX experiment, only a minute fraction of entire population of tested aptamer sequences that were experimentally determined to be good binders, is ever reported. There are two potential approaches to addressing these challenges: i) make it part of the practice of SELEX experimental protocol to record those nucleotide sequences that were not selected, and ii) develop a computational approach to simulate the sequence pools used in the initial phases of a SELEX experiment and thereby produce a theoretical negative set of sequences for any given aptamer to compare against. Both would be significant challenges in their own right, but would greatly advance the quality of data available to aptamer scientists.

5.2 Summary of contributions

In this thesis, I was most interested in investigating the structural basis for ligand binding by various non-coding RNAs. The discovery of novel functions of RNA molecules [9] via high throughput sequencing technologies, *in vitro* aptamer selection, and the *in vivo* study of ribozymes and riboswitches, along with the dramatic increase in number of RNA structures that are being solved and deposited into the Protein Data Bank has stimulated

the development of computational methods for characterizing nucleic acid structure in terms of their functional component parts. These approaches are motivated by the desire to design synthetic molecules with predefined functions, such as the ability to sequester drugs quickly and specifically, sense the presence of chemicals at low concentrations, or even regulate gene expression. The potential of RNA to realize such functions is demonstrated by the diversity of its biological roles. Towards the rational design of RNA aptamers for such applications, this doctoral research presents three main contributions: (1) the cataloguing of RNA sequence and structures using machine understandable descriptions that enable reasoning and retrieval, (2) quantitative analysis of ligand bound nucleic acids in terms of their structural motifs, including the identification of motifs that recur within and across structures and (3) the application of these motifs to classify nucleic acid-ligand binding and to predict these interactions. In the following sections, I summarize the work described in Chapters 2 – 4 in the context of these contributions and discuss future work.

Chapter 2 presented the RNA Knowledge Base, a collection of rich machine understandable descriptions of inter nucleotide residue interactions that are extant in RNA PDB structure files serialized as an instantiated ontology. This work was the first published demonstration of the use of Semantic Web technologies to accurately describe the conceptualizations of inter-nucleotide interactions between classes of RNA nucleotide residues as defined by the RNA scientific community in a machine understandable manner. The RKB OWL ontology was instantiated with symbolic annotations of PDB

structures computed using MC-Annotate [73], and can be used for simple information retrieval as well as ontology based reasoning about RNA structural features.

Chapter 3 presented Aptamer Base, a collaboratively created knowledge base about aptamers and their SELEX experiments. This work provides the aptamer community with a unique resource of data about aptamers and specific accounts of the experimental details used in their construction. Aptamer Base has continued to grow, and currently houses sequences for over 1600 aptamers that were manually extracted from over 500 SELEX experimental papers published between 1990 and 2013 in the scientific literature. This work provides an important contribution to the aptamer community: a representational model for serializing the results of SELEX experiments and their resulting aptamer sequences that exposes aptamer data previously available only in unstructured literature. Aptamer Base data is published in a machine understandable format that enables rich querying through the Metaweb Query Language (MQL). Moreover, the open nature of Aptamer Base, in that any member of Freebase.com can contribute to the knowledge base, provides the aptamer community with an open and collaborative environment for annotating SELEX experiments and their results.

In Chapter 4, I describe a methodology to identify and computationally describe recurring motifs of RNA structure that are predictive of ligand binding. In particular, the method involved annotating all base pair interactions and backbone links to create graph based representations of nucleic acid structures. These graphs were then partitioned into sets of indivisible fragments via the computation of each graph's Minimum Cycle Basis

(MCB). I automatically identified classes of minimum cycles based on their participating nucleotide residues and respective interactions, developed a machine readable representation for these classes, and found that they recur within and across RNA structures. Lastly, I demonstrate that the minimum cycle composition of an RNA structure is predictive of a ligand binding interactions and that features of individual minimum cycle classes are predictive of their presence or absence in a ligand neighbourhood.

Thousands of person hours have been dedicated to investigating the relationship between nucleic acid structure *in vivo* and its many biological functions, and to approaches for designing aptamers that bind to ligands with high sensitivity and specificity. However, a major barrier to the effective repurposing and analysis of this data remains – scientists lack a consistent nomenclature and methodology for describing and publicly sharing their findings. Together, the works described in this thesis are an effort to lower this barrier by leveraging Semantic Web technologies to develop and integrate formal descriptions of nucleic acid features including their sequences, secondary structures, orientation in three dimensional space, and structural motifs to enable the re-use of existing data towards the rational design of RNA aptamers.

Appendices

Appendix A

Table 10 list of ligands used in this study

Ligand Label	PDBeChem ID	ChEBI
9H-PURINE-2,6-DIAMINE	2BP	CHEBI:479072
6-METHYL-1,3,5-TRIAZINE-2,4-DIAMINE	3AW	CHEBI:72475
PYRIMIDINE-2,4,6-TRIAMINE	3AY	CHEBI:39857
6-AMINO-1,3,5-TRIAZIN-2(1H)-ONE	5AZ	CHEBI:72474
6-CHLOROGUANINE	6GU	CHEBI:72345
2-FLUOROADENINE	A2F	CHEBI:72457
ARGININEAMIDE	AAR	CHEBI:40477
6-AMINO-HEXANOIC ACID	ACA	CHEBI:16586
ADENINE	ADE	CHEBI:16708
DEOXY-METHYL-ARGININE	ARM	CHEBI:2412
ADENOSYLCOBALAMIN	B1Z	CHEBI:18408
S-BENZOYLTHIAMINE O-MONOPHOSPHATE	BFT	CHEBI:41039
BIOTIN	BTN	CHEBI:15956
CITRULLINE	CIR	CHEBI:18211
1-[(4-AMINO-2-METHYLPYRIMIDIN-5-YL)METHYL]-3-(2-{[HYDROXY(PHOSPHONOOXY)PHOSPHORYL]OXY}ETHYL)-2-METHYLPYRIDINIUM	D2X	CHEBI:72458
2'-DEOXYGUANOSINE-5'-MONOPHOSPHATE	DGP	CHEBI:16192
c-di-GMP	DGP	CHEBI:49537
THIOGUANINE	DX4	CHEBI:9555
[(3S)-3-AMINO-4-HYDROXY-4-OXO-BUTYL]-[[(2S,3S,4R,5R)-5-(6-AMINOPURIN-9-YL)-3,4-DIHYDROXY-OXOLAN-2-YL]METHYL]-METHYL-SELANIUM	EEM	CHEBI:9066
N-[4-({[(6S)-2-AMINO-5-FORMYL-4-OXO-3,4,5,6,7,8-HEXAHYDROPTERIDIN-6-YL]METHYL}AMINO)BENZOYL]-L-GLUTAMIC ACID	FFO	CHEBI:15640
FLAVIN MONONUCLEOTIDE	FMN	CHEBI:17621
N-[4-({[(6S)-2-AMINO-5-FORMYL-4-OXO-1,4,5,6,7,8-HEXAHYDROPTERIDIN-6-YL]METHYL}AMINO)BENZOYL]-L-GLUTAMIC ACID	FOZ	CHEBI:63606
GLYCINE	GLY	CHEBI:15428
GUANOSINE	GMP	CHEBI:16759

Ligand Label	PDBeChem ID	ChEBI
L-ARGININE	GND	CHEBI:16467
2'-DEOXY-GUANOSINE	GNG	CHEBI:17172
HYPOXANTHINE	HPA	CHEBI:17368
L-HOMOARGININE	HRG	CHEBI:27747
HYDROXOCOBALAMIN	I2A	CHEBI:27786
N~6~-[(1Z)-ETHANIMIDOYL]-L-LYSINE	IEL	CHEBI:63971
LYSINE	LYS	CHEBI:18019
O-(2-AMINOETHYL)-L-SERINE	OLZ	CHEBI:72341
2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDINE-5-CARBONITRILE	PQ0	CHEBI:45075
7-DEAZA-7-AMINOMETHYL-GUANINE	PRF	CHEBI:45126
1-[(4-AMINO-2-METHYLPYRIMIDIN-5-YL)METHYL]-3-(2-HYDROXYETHYL)-2-METHYLPYRIDINIUM	PYI	CHEBI:45395
RIBOFLAVIN	RBF	CHEBI:17015
N,N'-TETRAMETHYL-ROSAMINE	ROS	CHEBI:45358
1-DEOXY-1-[8-(DIMETHYLAMINO)-7-METHYL-2,4-DIOXO-3,4-DIHYDROBENZO[G]PTERIDIN-10(2H)-YL]-D-RIBITOL	RS3	CHEBI:72346
S-ADENOSYL-L-HOMOCYSTEINE	SAH	CHEBI:16680
S-ADENOSYLMETHIONINE	SAM	CHEBI:15414
SINEFUNGIN	SFG	CHEBI:45453
L-THIALYSINE	SLZ	CHEBI:497734
STREPTOMYCIN	SRY	CHEBI:17076
(8R)-8-[(DIMETHYLAMINO)METHYL]-1-[3-(DIMETHYLAMINO)PROPYL]-1,7,8,9-TETRAHYDROCHROMENO[5,6-D]IMIDAZOL-2-AMINE	SS0	CHEBI:72332
THIAMINE DIPHOSPHATE	TDP	CHEBI:9532
THEOPHYLLINE	TEP	CHEBI:28177
5-HYDROXYMETHYLENE-6-HYDROFOLIC ACID	THF	CHEBI:45981
THIAMIN PHOSPHATE	TPS	CHEBI:9533
XANTHINE	XAN	CHEBI:17712

Table 11 Correlation feature selection results on CLNCi features. Selected CLNCi annotation classes are shown.

L1PA	Residues	Base Pairs	Cycle Length	Number Of Links
053234f5dbc296fae20e7339e1bbb605	U,C,G,A	RNAO_0000003,RNAO_0000003	4	2
0c6ce050f828529a2ac273e9590c6777	G,U,C,G,G	RNAO_0000003,RNAO_0000003	5	3
0f6c5f77c83a7f100125b16d27668d01	A,U,G,C	RNAO_0000117,RNAO_0000003,RNAO_0000007,RNAO_0000013	4	0
218f55d1ea4c1d878e5b2eb93d8b4e7d	C,U,G,G,G,A,A	RNAO_0000003,RNAO_0000003	7	5
4030e1763b588205a904a1dba1412dfd	C,G,U,A	RNAO_0000003,RNAO_0000003	4	2
4af7758072db2c18ff7e8e1d3bd2b37e	C,C,A	RNAO_0000001,RNAO_0000001	3	1
4f5424954fabd9fc0e79261a68b400af	G,A,A,A,G,U	RNAO_0000014,RNAO_0000005	6	4
5405b2aa15ea41f0f3a84b19f6645515	G,A,C,U	RNAO_0000003,RNAO_0000003	4	2
7461d4ad6a37eae2f6636d8332f9c362	U,U,C,U,U	RNAO_0000006	5	4
79e76329907ad77d86d3e71d5f9aec73	G,C,G,U	RNAO_0000003,RNAO_0000003	4	2
816b6c9e10abee0114583b6a3c7cc4b8	A,G,A,G	RNAO_0000012,RNAO_0000012	4	2
820e951dd181e2be8bafaa7a35830c25	A,G,G,G,U	RNAO_0000012,RNAO_0000003	5	3
9d67c8325473655de6f3fe76a4c6fe02	U,A,U,A	RNAO_0000003,RNAO_0000003	4	2
9ea11cb1b18622d2c9fc454e0de0ae69	C,A,G,A,G	RNAO_0000003,RNAO_0000001	5	3
a6da9574c0e05bd21293e9d3f5237baf	U,G,C,A,U	RNAO_0000006,RNAO_0000003	5	3
aceea01cdae796d8c15cb6c4421c32d5	G,C,C,G	RNAO_0000003,RNAO_0000003	4	2
b4e977f5b2a0ef17602a3ded6a10932c	G,U,C,G	RNAO_0000003,RNAO_0000003	4	2
d47dd1686f5fb4e30a5b5ba44c778ea2	U,U,G,C	RNAO_0000003,RNAO_0000003	4	2
d7086882abb4c92cfbbd8378d483ed10	G,G,C,A	RNAO_0000012,RNAO_0000003	4	2
e1735e99cc3182d764dda2fcd31b6ace	U,G,A,C,C,G,G	RNAO_0000003	7	6
e61fc5d2f8232bfa66e7f0f9c1ff110b	A,C,G	RNAO_0000003,RNAO_0000001,RNAO_0000001	3	0

Table 12 Correlation feature selection results on CLNCii features. Selected CLNCii annotation classes are shown.

L2PA	Residues	Base Pairs	Cycle Length	Number Of Links
067637a49b5c6ae1da07a199cd31857e	U,A,G,C	RNAO_0000003,RNAO_0000003	4	2
0b738c5008a743627f5b794628c96d56	G,U,A,G,A,G,C	RNAO_0000012,RNAO_0000008,RNAO_0000003	7	4
0f077dfaff5262be64cdabb7a82b7571	G,A,U,A,A	RNAO_0000008,RNAO_0000003	5	3
1081df2b34ad2b23c8b9728b13923918	G,C,G,A	RNAO_0000012,RNAO_0000003	4	2
17be419f67d3b8d435186214f42a6bf9	G,A,A,C	RNAO_0000012,RNAO_0000003	4	2
1a9ad98e3f86796a0579f41b792647c5	G,A,A,G,G	RNAO_0000012,RNAO_0000014,RNAO_0000001	5	2
1b555e39ef6e9198d1d95da0f0582411	C,G,G,U,G	RNAO_0000003,RNAO_0000003	5	3
1cf7a03d4a92ffe6405c9b8e4c67c9ae	C,A,U,U	RNAO_0000007,RNAO_0000007	4	2
2616dedfa164a7578394927a3daada44	G,A,A,G	RNAO_0000008,RNAO_0000001	4	2
3b91e43753dbd3c3f2bfe0c7d90c7783	G,A,G,A	RNAO_0000008,RNAO_0000005	4	2
445c69d2714e43b727182b7496657eb1	G,C,G,C	RNAO_0000003,RNAO_0000003	4	2
53e10a2d7319c18baed26cfcdb795501	A,G,G,G,C,U,U	RNAO_0000117,RNAO_0000014,RNAO_0000003	7	4
5b9d58c29249c840c412cba34d952c9e	U,G,C,G	RNAO_0000003,RNAO_0000003	4	2
6d8dafded5e229079feda37224b1659f	C,G,U,A	RNAO_0000003,RNAO_0000003	4	2
6fa1871cb8f33c568904edb85f279d95	U,U,A,A	RNAO_0000003,RNAO_0000003	4	2
9a3fa41c46cb8eff174956a692d5fb0f	G,G,C,A,G	RNAO_0000003,RNAO_0000001	5	3
9a693ae9e5dfb3a889bbdb09fb82a2da	G,U,C,A	RNAO_0000003,RNAO_0000003	4	2
9ac07d7c3e00dc0cc9bb8ac4f61c040b	U,U,C,G	RNAO_0000003,RNAO_0000003	4	2
ac80d8cdae61f6cef6c895989d911c91	U,A,A,A	RNAO_0000117,RNAO_0000006	4	2
b0b5ede8c7744f3bba26d8a4393bb5cd	A,C,U,G,C,G,G	RNAO_0000003	7	6
e201d313e974351336a2ef1d46b01719	A,U,A,U	RNAO_0000003,RNAO_0000003	4	2

L2PA	Residues	Base Pairs	Cycle Length	Number Of Links
fa1397f2eb3bc520f72511485690f106	G,U,C,G	RNAO_0000003,RNAO_0000003	4	2

Table 13 Correlation feature selection results on ILNCi features. Selected ILNCi profile classes are shown. For brevity, we use only show the L1PA classes for the selected first degree neighbours.

L1PA	Residues	Base Pairs	Cycle Length	Number Of Links
094f9cbc695ea3d01fa00918a7fd4efa	A,C,G,G,G,U	RNAO_0000006,RNAO_0000003	6	4
0cd0cfedefe092eb7b7e8cd04a1f41bb	C,A,G,U	RNAO_0000003,RNAO_0000003	4	2
0dc24e7b414b2d34856b4d6a674a1770	A,C,G,G	RNAO_0000003,RNAO_0000003	4	2
1d41d06115d82944cb7c36681c5db5f2	C,A,G,G,A,A,A,U,G,U,A,A,A,G	RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000001	14	10
23501c789b9ef06528fae9e0f0025515	A,U,A,U,A,A	RNAO_0000117,RNAO_0000117,RNAO_0000005	6	3
268c7727c215fef4cbf827fa4619a990	A,C,C,A,G,U,G	RNAO_0000011,RNAO_0000003,RNAO_0000003	7	5
273a3a15367d90b184993f133e28afc9	C,A,A,C,C,A,G	RNAO_0000008,RNAO_0000003,RNAO_0000007	7	4
2a246452b9fd1b6bf40e5300753d6c71	G,U,A,G,A,G,C	RNAO_0000012,RNAO_0000008,RNAO_0000003	7	4
2a6f846f91e57e071948a7980d13b4b8	U,U,A,A	RNAO_0000117	4	3
2be1b26ac2f53de50d20a69cefb0d037	G,U,C,G,G,A,A,C,U,G,G,C,A,G,C,G,U	RNAO_0000014,RNAO_0000006,RNAO_0000001	17	14
309ed9a1ad924586086f1e316e373286	C,C,G,A,C,A,G,U,A,A,G	RNAO_0000117,RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000003	11	6
35ebff0fb85d027e108b017a8b61306d	A,U,C,G	RNAO_0000003,RNAO_0000003	4	2
3a44fb798de6c5325bc0643ad1eb1aaa	U,G,G,G,U,G,G,A,G,C,A	RNAO_0000014,RNAO_0000006,RNAO_0000003,RNAO_0000003,RNAO_0000001	11	6
42ec0baa1b8930a42ad989883384f5e5	G,U,G,C,C	RNAO_0000011,RNAO_0000003,RNAO_0000003	5	2
43dcaf2120b0477088f01d03b405420a	G,G,A,G,G,C,G,G	RNAO_0000116,RNAO_0000012,RNAO_0000013	8	5
4f5424954fabd9fc0e79261a68b400af	G,A,A,A,G,U	RNAO_0000014,RNAO_0000005	6	4
57d817650805c0e58887bd6040dab8d9	G,C,U,A,G,G,U	RNAO_0000117,RNAO_0000014,RNAO_0000003	7	4
613ef43c5e97009f9f	C,U,G,G,	RNAO_0000014,RNAO_0000001,R	18	15

L1PA	Residues	Base Pairs	Cycle Length	Number Of Links
df8c27e1576a7e	U,G,A,A,G,G,G,C,G,C,U,C,C,A	NAO_0000001		
6e70e8117d23ab3dfd1665f38616a38b	G,C,G,U,G,C,A	RNAO_0000012,RNAO_0000014,RNAO_0000003	7	4
73901499d457b4e6b6ca952e0e3f3eae	G,C,G,G,U,A,U,A,C,G,G,U,G	RNAO_0000014,RNAO_0000001,RNAO_0000001	13	10
7461d4ad6a37eae2f6636d8332f9c362	U,U,C,U,U	RNAO_0000006	5	4
74cd7162dc813058a28718ed63cb8f96	G,A,G,C,A	RNAO_0000006,RNAO_0000003	5	3
7a4fac04d3aeac0d70ed7faf3b7d04c8	G,G,A,A,C,A,G,G,A,G,A	RNAO_0000006,RNAO_0000003,RNAO_0000003,RNAO_0000007	11	7
7c68f5d14e58fbf5396cd3c3a4dbd389	A,C,C,G,C,U	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
7ce30c6d44935c0c2dcaa1446f358a47	G,G,C,G,A	RNAO_0000012,RNAO_0000012	5	3
7f1ea52b86f63f4f8cd5b3303f843423	U,A,A,U,C,U	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
9ada805c8d5464037f4e46e49ba9e130	U,C,A,C,G,U,C,A,U	RNAO_0000003,RNAO_0000007	9	7
9b5765218f36f37cd12af12c1dd44e97	C,C,A,G,G,U,C	RNAO_0000117,RNAO_0000008,RNAO_0000003	7	4
9c7c124537ae6420b8dfcb7e2a579d2e	U,G,A,G,A,G,C	RNAO_0000117,RNAO_0000003	7	5
a0d1cab0807cf80622e92b9764c8c944	C,G,C,G	RNAO_0000116,RNAO_0000003	4	2
c94753dc3d775fb01631713ae0f688ca	A,C,G,G,C,A,U,U	RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000003	8	4
d39b774e384b0cee442c8635fc92f82d	U,A,A,U,U,C	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
e94c91d7b721b6df7b40545419801583	C,A,A,C,G	RNAO_0000003,RNAO_0000007	5	3

Table 14 Correlation feature selection results on ILNCii features. Selected ILNCii profile classes are shown. For brevity, we use only show the L2PA classes for the selected first degree neighbours.

L2PA	Residues	Base Pairs	Cycle Length	Number Of Links
02bb526e87e3a6659eb1461552693f3b	G,G,A,A,C,A,G,G,A,G,A	RNAO_0000006,RNAO_0000003,RNAO_0000003,RNAO_0000007	11	7
04c227f8aa18a682a5c0e6addc24590b	G,C,G,G,U,A,U,A,C,G,G,U,G	RNAO_0000014,RNAO_0000001,RNAO_0000001	13	10
0b738c5008a743627f5b794628c96d56	G,U,A,G,A,G,C	RNAO_0000012,RNAO_0000008,RNAO_0000003	7	4
0dc24e7b414b2d34856b4d6a674a1770	A,C,G,G	RNAO_0000003,RNAO_0000003	4	2
1795ecbca06c0a82d7f2bd6e8a3867a9	G,G,C,C,C	RNAO_0000003,RNAO_0000003	5	3
19fa62a9498bd203c7b695c0121b938d	A,G,U,G,U	RNAO_0000003,RNAO_0000005	5	3
1cf7a03d4a92ffe6405c9b8e4c67c9ae	C,A,U,U	RNAO_0000007,RNAO_0000007	4	2
21999d691e1dcef76bd50b7baa0fb411	U,G,A,G,A,G,C	RNAO_0000117,RNAO_0000003	7	5
2cf5c1da9e99e2782151b37467f3bd6b	U,C,U,A,U,A	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
2e3d9dc63325dac2530e2f4c185e3db8	A,U,A,U,A,A	RNAO_0000117,RNAO_0000117,RNAO_0000005	6	3
370cf44efd140d110b8bf18f5fd28b41	G,U,C,G,G,A,A,C,U,G,G,C,A,G,C,G,U	RNAO_0000014,RNAO_0000006,RNAO_0000001	17	14
48c01fb57b91ade2f157da1493d3b5ee	G,C,G,U,G,C,A	RNAO_0000012,RNAO_0000014,RNAO_0000003	7	4
4a2f31c8b74862799f554cb4f2b07bb3	G,G,A,C,G	RNAO_0000006,RNAO_0000008	5	3
521888e01c1eec35bcacc1fece20ec57	C,C,G,G,C,C,C,A	RNAO_0000003	8	7
532aea1d2586be910bdaeac21d8a4a1b	G,G,A,G,G,C,G,G	RNAO_0000116,RNAO_0000012,RNAO_0000013	8	5
53e10a2d7319c18baed26cfcdb795501	A,G,G,G,C,U,U	RNAO_0000117,RNAO_0000014,RNAO_0000003	7	4
5586d55c15c2e9dad8d894840a10489e	U,A,A,U,U,C	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
5df56cec27ae126913470bfabefbd84d	C,G,G,A	RNAO_0000117,RNAO_0000005	4	2

L2PA	Residues	Base Pairs	Cycle Length	Number Of Links
5e195d56e263a626d4e1dc884d5b24b2	A,U,U,A,G	RNAO_0000003,RNAO_0000003	5	3
684175ca1a90649ccdc18f7975e87341	C,A,A,U,G	RNAO_0000003,RNAO_0000001	5	3
6dc4c431991f0167146d716b377eea60	G,G,A,U,C	RNAO_0000003,RNAO_0000005	5	3
6e6ed5ef865d770c7299264b6d4e58b6	G,U,G,G,A,G,G,U,A,C,G	RNAO_0000012,RNAO_0000014,RNAO_0000006,RNAO_0000003,RNAO_0000003	11	6
7144e5c5efff9dcb532861de930062bb	C,U,U,U,U	RNAO_0000006	5	4
73248f78a7ac98f9ae18db30cf410062	G,A,U,C,C,A,C,U,U	RNAO_0000003,RNAO_0000007	9	7
7817ba9e97c2abe59e165b84bb8aa5ad	G,A,G,C,A	RNAO_0000006,RNAO_0000003	5	3
896dafa0ec064403dbd6b9e461dc0a71	C,A,A,C,G	RNAO_0000003,RNAO_0000007	5	3
8b80c93cea0be17146751cf54591815e	A,C,G,G,G,U	RNAO_0000006,RNAO_0000003	6	4
90fcb0c4c3b8dbe1f08a68d1515e7478	G,U,G,C,C	RNAO_0000011,RNAO_0000003,RNAO_0000003	5	2
9416250e526a6acde942c769ba8997d2	U,U,A,A	RNAO_0000003,RNAO_0000003	4	2
97cc2a2bc06cb900a918890c0dd44689	C,A,G,G,A,A,A,U,G,U,A,A,A,G	RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000001	14	10
98b77ba1e68a77a12f00c04a467c17ef	C,G,G,G,A	RNAO_0000014,RNAO_0000003,RNAO_0000001	5	2
99a171ee594f39a318fbc62e3126a17a	A,C,G,U,G,C,A	RNAO_0000011,RNAO_0000003,RNAO_0000003	7	5
b3f997d65dc8dd2bebb2ab1747518f14	A,C,G,G,C,A,U,U	RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000003	8	4
b6ca19e8f39f1433b1f69e07998077e1	G,A,U,A,A,G	RNAO_0000014,RNAO_0000003	6	4
bdbc4a965aebfd40e03581b0519c6844	G,C,C,G,U,C,C,A,C,G	RNAO_0000003,RNAO_0000003,RNAO_0000003	10	7
c350dd20eaf1ab2160c388a6147371dc	C,A,A,C,C,A,G	RNAO_0000008,RNAO_0000003,RNAO_0000007	7	4
dc4fd2dfff28ddf8acba166be975720d	C,A,G,U	RNAO_0000003,RNAO_0000003	4	2
e0e78cdae8e4e9e1aeacbc97e3d2431e	C,C,A,G,G,U,C	RNAO_0000117,RNAO_0000008,RNAO_0000003	7	4
e201d313e974351336a2ef1d46b01719	A,U,A,U	RNAO_0000003,RNAO_0000003	4	2
e4f4af9fc750918801e	C,U,G,G,U,	RNAO_0000014,RNAO_000000	18	15

L2PA	Residues	Base Pairs	Cycle Length	Number Of Links
d58f21ebb6213	G,A,A,G,G,G,C,G,C,U,C,C,A	1,RNAO_0000001		
e9ebe23b699a317887d24155fcb32dd7	U,G,C,A,U	RNAO_0000006,RNAO_0000003	5	3
f63bf3d7d6a8c151afe481cbbae4693e	C,C,G,A,C,A,G,U,A,A,G	RNAO_0000117,RNAO_0000003,RNAO_0000003,RNAO_0000003,RNAO_0000003	11	6
f73a81368d397941f82176644e030617	A,C,C,G,C,U	RNAO_0000003,RNAO_0000003,RNAO_0000007	6	3
f78bf14243acec310eb37e2b35660f98	G,A,A,A,G,U	RNAO_0000014,RNAO_0000005	6	4
fb9e98e4142b663e6a943224c4fd2f8b	G,G,A,C	RNAO_0000013,RNAO_0000001	4	2

References

1. Gilbert, W., *Origin of life: The RNA world.* Nature, 1986. **319**(6055).
2. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
3. Robertson, M.P., *Origins of Life: Emergence of the RNA World.* Wiley Encyclopedia of Chemical Biology, 2008: p. 1–12.
4. Cech, T.R., *The RNA worlds in context.* Cold Spring Harb Perspect Biol, 2012. **4**(7): p. a006742.
5. Cech, T.R., *A model for the RNA-catalyzed replication of RNA.* Proc Natl Acad Sci U S A, 1986. **83**(12): p. 4360-3.
6. Yarus, M., *Getting Past the RNA World: The Initial Darwinian Ancestor.* Cold Spring Harb Perspect Biol, 2010. **3**(4): p. a003590-a003590.
7. Lacey Jr, J. and D. Mullins Jr, *Proteins and nucleic acids in prebiotic evolution*, in *Molecular Evolution*1972, Springer. p. 171-188.
8. Rhodes, A., et al., *The generation and characterization of antagonist RNA aptamers to human oncostatin M.* J Biol Chem, 2000. **275**(37): p. 28555-61.
9. Serganov, A. and E. Nudler, *A decade of riboswitches.* Cell, 2013. **152**(1): p. 17-24.
10. Eddy, S.R., *Non-coding RNA genes and the modern RNA world.* Nat Rev Genet, 2001. **2**(12): p. 919-29.
11. Cech, T.R. and J.A. Steitz, *The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones.* Cell, 2014. **157**(1): p. 77-94.

12. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**(6287): p. 818-22.

13. Mandal, M. and R.R. Breaker, *Adenine riboswitches and gene activation by disruption of a transcription terminator.* Nature Structural & Molecular Biology, 2004. **11**(1): p. 29-35.

14. Nudler, E. and A.S. Mironov, *The riboswitch control of bacterial metabolism.* Trends in Biochemical Sciences, 2004. **29**(1): p. 11-17.

15. Tombelli, S., M. Minunni, and M. Mascini, *Aptamers-based assays for diagnostics, environmental and food analysis.* Biomolecular engineering, 2007. **24**(2): p. 191-200.

16. Tombelli, S., M. Minunni, and M. Mascini, *Analytical applications of aptamers.* Biosens Bioelectron, 2005. **20**(12): p. 2424-34.

17. Swensen, J.S., et al., *Continuous, Real-Time Monitoring of Cocaine in Undiluted Blood Serum via a Microfluidic, Electrochemical Aptamer-Based Sensor.* Journal of the American Chemical Society, 2009. **131**(12): p. 4262-4266.

18. Guthrie, J., et al., *Assays for cytokines using aptamers.* Methods, 2006. **38**(4): p. 324-330.

19. Lin, C.H., et al., *Formation of an amino-acid-binding pocket through adaptive zippering-up of a large DNA hairpin loop.* Chem Biol, 1998. **5**(10): p. 555-72.

20. Jiang, L. and D.J. Patel, *Solution structure of the tobramycin-RNA aptamer complex.* Nat Struct Biol, 1998. **5**(9): p. 769-74.

21. Flinders, J., et al., *Recognition of planar and nonplanar ligands in the malachite green-RNA aptamer complex.* Chembiochem, 2004. **5**(1): p. 62-72.

22. Jiang, F., et al., *Anchoring an extended HTLV-1 Rex peptide within an RNA major groove containing junctional base triples.* Structure, 1999. **7**(12): p. 1461-72.

23. Nix, J., D. Sussman, and C. Wilson, *The 1.3 A crystal structure of a biotin-binding pseudoknot and the basis for RNA molecular recognition.* J Mol Biol, 2000. **296**(5): p. 1235-44.

24. Herr, J.K., et al., *Aptamer-conjugated nanoparticles for selective collection and detection of cancer cells.* Anal Chem, 2006. **78**(9): p. 2918-24.

25. Chushak, Y. and M.O. Stone, *In silico selection of RNA aptamers.* Nucleic Acids Res, 2009. **37**(12): p. e87.

26. Osborne, S.E. and A.D. Ellington, *Nucleic acid selection and the challenge of combinatorial chemistry.* Chemical Reviews, 1997. **97**(2): p. 349-370.

27. Gold, L., et al., *Diversity of Oligonucleotide Functions.* Annual Review of Biochemistry, 1995. **64**: p. 763-797.

28. Shapiro, B.A., et al., *Bridging the gap in RNA structure prediction.* Curr Opin Struct Biol, 2007. **17**(2): p. 157-65.

29. Lee, J.-O., et al., *Aptamers as molecular recognition elements for electrical nanobiosensors.* Analytical and bioanalytical chemistry, 2008. **390**(4): p. 1023-1032.

30. Henkin, T.M., *Riboswitch RNAs: using RNA to sense cellular metabolism.* Genes & development, 2008. **22**(24): p. 3383-3390.

31. Schwalbe, H., et al., *Structures of RNA switches: insight into molecular recognition and tertiary structure.* Angewandte Chemie International Edition, 2007. **46**(8): p. 1212-1219.

32. Edwards, T.E., D.J. Klein, and A.R. Ferre-D'Amare, *Riboswitches: small-molecule recognition by gene regulatory RNAs.* Curr Opin Struct Biol, 2007. **17**(3): p. 273-279.

33. Harvey, I., P. Garneau, and J. Pelletier, *Inhibition of translation by RNA-small molecule interactions.* RNA, 2002. **8**(4): p. 452-463.

34. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

35. Blackburn, G.M. and Royal Society of Chemistry (Great Britain), *Nucleic acids in chemistry and biology*. 3rd ed2006, Cambridge: RSC Pub. xxxi, 470 p.

36. Neidle, S., *Principles of nucleic acid structure*. First edition. ed2007, Amsterdam: Elsevier. xii, 289 p.

37. Spencer, M., *The stereochemistry of deoxyribonucleic acid. I. Covalent bond lengths and angles.* Acta Crystallographica, 1959. **12**(1): p. 59-65.

38. Murthy, V.L., et al., *A complete conformational map for RNA.* J Mol Biol, 1999. **291**(2): p. 313-27.

39. Altona, C. and M. Sundaralingam, *Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation.* J Am Chem Soc, 1972. **94**(23): p. 8205-12.

40. McNaught, A.D., *Compendium of chemical terminology*. Vol. 1669. 1997: Blackwell Science Oxford.

41. McNaught, A.D., A. Wilkinson, and International Union of Pure and Applied Chemistry., *Compendium of chemical terminology : IUPAC recommendations*. 2nd ed1997, Oxford England ; Malden, MA, USA: Blackwell Science. vii, 450 p.

42. Major, F. and P. Thibault, *Computer Modeling of RNA Three-Dimensional Structures*, in *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, R.A. Meyers, Editor 2005, Weinheim. p. 605-636.

43. Townsend, L.B., *Chemistry of nucleosides and nucleotides*. Vol. 1. 1988, New York: Springer.

44. Hanlon, S., *The importance of London dispersion forces in the maintenance of the deoxyribonucleic acid helix.* Biochem Biophys Res Commun, 1966. **23**(6): p. 861-7.

45. Sarai, A., et al., *Origin of DNA helical structure and its sequence dependence.* Biochemistry, 1988. **27**(22): p. 8498-502.

46. Tazawa, I., T. Koike, and Y. Inoue, *Stacking properties of a highly hydrophobic dinucleotide sequence, N6, N6-dimethyladenylyl(3' leads to 5')N6, N6-dimethyladenosine, occurring in 16--18-S ribosomal RNA.* Eur J Biochem, 1980. **109**(1): p. 33-8.

47. Luo, R., et al., *The physical basis of nucleic acid base stacking in water.* Biophys J, 2001. **80**(1): p. 140-8.

48. Hoehndorf, R., et al., *The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data.* Applied Ontology, 2011. **6**(1): p. 53-89.

49. Leontis, N.B. and E. Westhof, *Geometric nomenclature and classification of RNA base pairs.* RNA, 2001. **7**(4): p. 499-512.

50. Fielding, R.T., *Architectural styles and the design of network-based software architectures (Doctoral Thesis)*, 2000, University of California.

51. Berners-Lee, T., J. Hendler, and O. Lassila, *The semantic web.* Scientific american, 2001. **284**(5): p. 28-37.

52. DuCharme, B., *Learning SPARQL* 2013: O'Reilly Media, Inc.

53. Wilkinson, M.D., B. Vandervalk, and L. McCarthy, *The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation.* J Biomed Semantics, 2011. **2**(1): p. 8.

54. Brickley, D. and R.V. Guha, *Resource Description Framework (RDF) Schema Specification 1.0: W3C Candidate Recommendation 27 March 2000.* 2000.

55. Hayes, P. and B. McBride. *RDF Semantics*. 2004 03/01/2014]; Available from: http://www.w3.org/TR/2004/REC-rdf-mt-20040210.

56. McGuinness, D.L. and F. Van Harmelen, *OWL web ontology language overview.* W3C recommendation, 2004. **10**(2004-03): p. 10.

57. Heath, T. and C. Bizer, *Linked data: Evolving the web into a global data space*. First ed. Synthesis lectures on the semantic web: theory and technology. Vol. 1. 2011: Morgan & Claypool. 1-136.

58. Berners-Lee, T., R. Fielding, and L. Masinter, *Uniform resource identifiers (uri).* Generic Syntax. IEDTF Draft Standard (RFC 2396), 1998.

59. Fielding, R., et al. *Hypertext transfer protocol–HTTP/1.1 - RFC 2616*. 1999.

60. Berners-Lee, T. *Linked data-design issues* 2006 [cited 2014 01/02/2014]; Available from: http://www.w3.org/DesignIssues/LinkedData.html.

61. Birbeck, M. and S. McCarron *CURIE Syntax 1.0–A syntax for expressing Compact URIs. W3C Recommendation, January 2009*.

62. Hitzler, P., et al., *OWL 2 web ontology language primer.* W3C recommendation, 2009. **27**(1): p. 123.

63. Whetzel, P.L., et al., *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W541-5.

64. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

65. Gruber, T.R., *A translation approach to portable ontology specifications.* Knowledge acquisition, 1993. **5**: p. 199-199.

66. Dumontier, M., *Biological situational modeling: Defining Molecular Roles in Pathways and Reactions*, in *OWL Experiences and Design (OWLED-EU)*2008: Karlsruhe, Germany.

67. Waugh, A., et al., *RNAML: a standard syntax for exchanging RNA information.* RNA, 2002. **8**(6): p. 707-17.

68. Leontis, N.B., et al., *The RNA Ontology Consortium: an open invitation to the RNA community.* RNA, 2006. **12**(4): p. 533-41.

69. Courtot, M., et al., *MIREOT: The minimum information to reference an external ontology term.* Appl. Ontol., 2011. **6**(1): p. 23-33.

70. Degtyarenko, K., et al., *ChEBI: a database and ontology for chemical entities of biological interest.* Nucleic Acids Res, 2008. **36**(Database issue): p. D344-50.

71. Dumontier, M., et al. *RKB: A Semantic Web Knowledge Base for RNA*. in *Bio-ontologies 2009*. 2009. Stockholm, Sweden.

72. Koynk, M., A. De Leon, and M. Dumontier, *Chemical Knowledge for the Semantic Web*, in *Data Integration in the Life Sciences (DILS2008)*2008: Evry, France.

73. Lemieux, S. and F. Major, *RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire.* Nucleic Acids Res, 2002. **30**(19): p. 4250-63.

74. Belleau, F., et al., *Bio2RDF: towards a mashup to build bioinformatics knowledge systems.* J Biomed Inform, 2008. **41**(5): p. 706-16.

75. Consortium, U., *The Universal Protein Resource (UniProt) 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D169-74.

76. Siren, E., et al., *Pellet: a practical OWL-DL reasoner.* Journal of Web Semantics, 2007. **5**(2): p. 51-53.

77. Tsarkov, D. and I. Horrocks, *FaCT++ description logic reasoner: system description*, in *Proceedings of the Third international joint conference on Automated Reasoning*2006, Springer-Verlag: Seattle, WA. p. 292-297.

78. Villanueva-Rosales, N. and M. Dumontier, *yOWL: an ontology-driven knowledge base for yeast biologists.* J Biomed Inform, 2008. **41**(5): p. 779-89.

79. Grenon, P., B. Smith, and L. Goldberg, *Biodynamic ontology: applying BFO in the biomedical domain.* Stud Health Technol Inform, 2004. **102**: p. 20-38.

80. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.

81. Mannhold, R., et al., *Protein-ligand interactions: from molecular recognition to drug design*. Methods and Principles in Medicinal Chemistry, ed. R. M., H. K., and G. F.2006, Berlin: John Wiley & Sons.

82. Vodnik, M., et al., *Phage display: selecting straws instead of a needle from a haystack.* Molecules, 2011. **16**(1): p. 790-817.

83. Zhou, X., et al., *The next-generation sequencing technology and application.* Protein Cell, 2010. **1**(6): p. 520-36.

84. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**(4968): p. 505-10.

85. Jayasena, S.D., *Aptamers: an emerging class of molecules that rival antibodies in diagnostics.* Clin Chem, 1999. **45**(9): p. 1628-50.

86. Huizenga, D.E. and J.W. Szostak, *A DNA aptamer that binds adenosine and ATP.* Biochemistry, 1995. **34**(2): p. 656-65.

87. Wilson, C. and J.W. Szostak, *Isolation of a fluorophore-specific DNA aptamer with weak redox activity.* Chem Biol, 1998. **5**(11): p. 609-17.

88. McKeague, M., et al., *Screening and initial binding assessment of fumonisin b(1) aptamers.* Int J Mol Sci, 2010. **11**(12): p. 4864-81.

89. Sekiya, S., et al., *Characterization and application of a novel RNA aptamer against the mouse prion protein.* J Biochem, 2006. **139**(3): p. 383-90.

90. Jones, L.A., et al., *High-affinity aptamers to subtype 3a hepatitis C virus polymerase display genotypic specificity.* Antimicrob Agents Chemother, 2006. **50**(9): p. 3019-27.

91. Wang, C., et al., *Single-stranded DNA aptamers that bind differentiated but not parental cells: subtractive systematic evolution of ligands by exponential enrichment.* J Biotechnol, 2003. **102**(1): p. 15-22.

92. White, R.R., B.A. Sullenger, and C.P. Rusconi, *Developing aptamers into therapeutics.* J Clin Invest, 2000. **106**(8): p. 929-34.

93. Nimjee, S.M., C.P. Rusconi, and B.A. Sullenger, *Aptamers: an emerging class of therapeutics.* Annu Rev Med, 2005. **56**: p. 555-83.

94. Brody, E.N. and L. Gold, *Aptamers as therapeutic and diagnostic agents.* Journal of Biotechnology, 2000. **74**(1): p. 5-13.

95. Ponomarenko, J.V., et al., *SELEX_DB: a database on in vitro selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data.* Nucleic Acids Res, 2002. **30**(1): p. 195-9.

96. Lee, J.F., et al., *Aptamer database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D95-100.

97. Thodima, V., M. Pirooznia, and Y. Deng, *RiboaptDB: a comprehensive database of ribozymes and aptamers.* BMC Bioinformatics, 2006. **7 Suppl 2**: p. S6.

98. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2011. **39**(Database issue): p. D32-7.

99. Cochrane, G., et al., *Petabyte-scale innovations at the European Nucleotide Archive.* Nucleic Acids Res, 2009. **37**(Database issue): p. D19-25.

100. Kaminuma, E., et al., *DDBJ launches a new archive database with analytical tools for next-generation sequence data.* Nucleic Acids Res, 2010. **38**(Database issue): p. D33-8.

101. Chen, P.P., *The entity-relationship model---toward a unified view of data.* ACM Transactions on database systems, 1976. **1**(1): p. 9-36.

102. Silverman, S.K., *Artificial Functional Nucleic Acids: Aptamers, Ribozymes, and Deoxyribozymes Identified by In Vitro Selection.* Functional Nucleic Acids for Analytical Applications, 2009: p. 47-108.

103. Hofacker, I.L. and P.F. Stadler, *Memory efficient folding algorithms for circular RNA secondary structures.* Bioinformatics, 2006. **22**(10): p. 1172-6.

104. Mathews, D.H., et al., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J Mol Biol, 1999. **288**(5): p. 911-40.

105. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7287-92.

106. Parisien, M. and F. Major, *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.* Nature, 2008. **452**(7183): p. 51-5.

107. Cho, E.J., J.W. Lee, and A.D. Ellington, *Applications of aptamers as sensors.* Annu Rev Anal Chem (Palo Alto Calif), 2009. **2**: p. 241-64.

108. Hamdani, H.Y., et al., *NASSAM: a server to search for and annotate tertiary interactions and motifs in three-dimensional structures of complex RNA molecules.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W35-41.

109. Mayer, O., et al., *Protein-induced RNA switches in nature*, in *Nucleic Acid Switches and Sensors*2006, Springer. p. 75-86.

110. Gorbalenya, A.E. and E.V. Koonin, *Helicases: amino acid sequence comparisons and structure-function relationships.* Curr Opin Struct Biol, 1993. **3**(3): p. 419-429.

111. Levitt, M. and C. Chothia, *Structural patterns in globular proteins.* Nature, 1976. **261**(5561): p. 552-8.

112. Switzer, R., H. Zalkin, and H. Saxild, *Purine, pyrimidine, and pyridine nucleotide metabolism*, in *Bacillus subtilis and its closest relatives: from genes to cells.* , A.L. Sonenshein, J.A. Hoch, and R. Losick, Editors. 2002, American Society for Microbiology: Washington, DC. p. 255-269.

113. Christiansen, L.C., et al., *Xanthine metabolism in Bacillus subtilis: characterization of the xpt-pbuX operon and evidence for purine-and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism.* Journal of bacteriology, 1997. **179**(8): p. 2540-2550.

114. *RCSB PDB Statistics*. 2014 [cited 2014 01/01/2014]; Available from: http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html.

115. Sarver, M., et al., *FR3D: finding local and composite recurrent structural motifs in RNA 3D structures.* J Math Biol, 2008. **56**(1-2): p. 215-52.

116. Dror, O., R. Nussinov, and H. Wolfson, *ARTS: alignment of RNA tertiary structures.* Bioinformatics, 2005. **21 Suppl 2**: p. ii47-53.

117. Popenda, M., et al., *RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures.* BMC Bioinformatics, 2010. **11**: p. 231.

118. Chojnowski, G., T. Walen, and J.M. Bujnicki, *RNA Bricks--a database of RNA 3D motifs and their interactions.* Nucleic Acids Res, 2014. **42**(Database issue): p. D123-31.

119. Montange, R.K. and R.T. Batey, *Riboswitches: emerging themes in RNA structure and function.* Annu Rev Biophys, 2008. **37**: p. 117-33.

120. Kang, M., C.D. Eichhorn, and J. Feigon, *Structural determinants for ligand capture by a class II preQ1 riboswitch.* Proc Natl Acad Sci U S A, 2014. **111**(6): p. E663-71.

121. Zhong, C. and S. Zhang, *Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment.* Nucleic Acids Res, 2012. **40**(3): p. 1307-17.

122. Lemieux, S. and F. Major, *Automated extraction and classification of RNA tertiary structure cyclic motifs.* Nucleic Acids Res, 2006. **34**(8): p. 2340-6.

123. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.

124. Lu, X.J. and W.K. Olson, *3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures.* Nat Protoc, 2008. **3**(7): p. 1213-27.

125. Lu, X.J., W.K. Olson, and H.J. Bussemaker, *The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex.* Nucleic Acids Res, 2010. **38**(14): p. 4868-76.

126. Horton, J.D., *A polynomial-time algorithm to find the shortest cycle basis of a graph.* SIAM J. Comp, 1987. **16**.

127. Berger, F., P. Gritzmann, and S. Vries, *Minimum cycle bases for network graphs.* Algorithmica, 2004. **40**(1): p. 51-62.

128. Burra, P.V., et al., *Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure.* Proc Natl Acad Sci U S A, 2009. **106**(26): p. 10505-10.

129. Kallberg, Y. and B. Persson, *KIND-a non-redundant protein database.* Bioinformatics, 1999. **15**(3): p. 260-1.

130. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.

131. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.

132. Wallach, I. and R. Lilien, *The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding.* Bioinformatics, 2009. **25**(5): p. 615-20.

133. Leontis, N. and C. Zirbel, *Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking*, in *RNA 3D Structure Analysis and Prediction*, N. Leontis and E. Westhof, Editors. 2012, Springer Berlin Heidelberg. p. 281-298.

134. Dutta, S., R.K. Green, and C.L. Lawson. *Looking at Structures: Introduction to Biological Assemblies nad the PDB Archive*. 2008 [cited 2014 01/01/2014]; Available from: http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/bioassembly_tutorial.html.

135. Grigg, J.C. and A. Ke, *Structural determinants for geometry and information decoding of tRNA by T box leader RNA.* Structure, 2013. **21**(11): p. 2025-32.

136. Montange, R.K., et al., *Discrimination between closely related cellular metabolites by the SAM-I riboswitch.* J Mol Biol, 2010. **396**(3): p. 761-72.

137. Lu, C., et al., *SAM recognition and conformational switching mechanism in the Bacillus subtilis yitJ S box/SAM-I riboswitch.* J Mol Biol, 2010. **404**(5): p. 803-18.

138. Thore, S., M. Leibundgut, and N. Ban, *Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand.* Science, 2006. **312**(5777): p. 1208-11.

139. Thore, S., C. Frick, and N. Ban, *Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch.* J Am Chem Soc, 2008. **130**(26): p. 8116-7.

140. Nguyen, M.N. and M.S. Madhusudhan, *Biological insights from topology independent comparison of protein 3D structures.* Nucleic Acids Res, 2011. **39**(14): p. e94.

141. Correll, C.C., et al., *Crystal structure of the ribosomal RNA domain essential for binding elongation factors.* Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13436-41.

142. Correll, C.C., I.G. Wool, and A. Munishkin, *The two faces of the Escherichia coli 23 S rRNA sarcin/ricin domain: the structure at 1.11 A resolution.* J Mol Biol, 1999. **292**(2): p. 275-87.

143. Correll, C.C., et al., *The common and the distinctive features of the bulged-G motif based on a 1.04 A resolution RNA structure.* Nucleic Acids Res, 2003. **31**(23): p. 6806-18.

144. Coimbatore Narayanan, B., et al., *The Nucleic Acid Database: new features and capabilities.* Nucleic Acids Res, 2014. **42**(Database issue): p. D114-22.

145. group, B.R. *RNA Structure Atlas*. 2014 [cited 2014 03/03/2014]; Available from: http://rna.bgsu.edu/rna3dhub/pdb.

146. PDB. *wwPDB Format version 2.3: Primary Structure Section*. 2007 24/12/2014]; Available from: http://www.wwpdb.org/documentation/format23/sect3.html.

147. Hamelryck, T. and B. Manderick, *PDB file parser and structure class implemented in Python.* Bioinformatics, 2003. **19**(17): p. 2308-10.

148. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.

149. Grabowski, S.J. and SpringerLink (Online service), *Hydrogen Bonding--New Insights*. p. 1 online resource.

150. Diestel, R., *Graph theory*. 3rd ed. Graduate texts in mathematics,2005, Berlin ; New York: Springer. xvi, 410 p.

151. Griffin, C., *Graph Theory: Penn State Math 485 Lecture Notes*, P.S. University, Editor 2012. p. 173.

152. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics.* J Chem Inf Comput Sci, 2003. **43**(2): p. 493-500.

153. Baugh, C., D. Grate, and C. Wilson, *2.8 A crystal structure of the malachite green aptamer.* J Mol Biol, 2000. **301**(1): p. 117-28.

154. Prud'hommeaux, E. and A. Seaborne, *SPARQL Query Language for RDF. W3C Recommendation, January 2008*, 2008.

155. Westbrook, J., et al., *PDBML: the representation of archival macromolecular structure data in XML.* Bioinformatics, 2005. **21**(7): p. 988-92.

156. Callahan, A., et al., *Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data*, in *The Semantic Web: Semantics and Big Data*2013, Springer. p. 200-212.

157. Bouckaert, R., et al., *WEKA - Experiences with a Java Open-Source Project.* Journal of Machine Learning Research, 2010. **11**: p. 2533-2541.

158. Platt, J., *Sequential minimal optimization: A fast algorithm for training support vector machines*, 1998: Microsoft Research.

159. Collie, G.W., et al., *Structural basis of telomeric RNA quadruplex--acridine ligand recognition.* Journal of the American Chemistry Society, 2011. **133**(8): p. 2721-8.

160. Ben-Shem, A., et al., *The structure of the eukaryotic ribosome at 3.0 A resolution.* Science, 2011. **334**(6062): p. 1524-9.

161. Dibrov, S., J. McLean, and T. Hermann, *Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA.* Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 2): p. 97-104.

162. Pallan, P.S., et al., *Crystal structure of a luteoviral RNA pseudoknot and model for a minimal ribosomal frameshifting motif.* Biochemistry, 2005. **44**(34): p. 11315-22.

163. Daldrop, P., et al., *Novel ligands for a purine riboswitch discovered by RNA-ligand docking.* Chem Biol, 2011. **18**(3): p. 324-35.

164. Gilbert, S.D., et al., *Adaptive ligand binding by the purine riboswitch in the recognition of guanine and adenine analogs.* Structure, 2009. **17**(6): p. 857-68.

165. Serganov, A., et al., *Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs.* Chem Biol, 2004. **11**(12): p. 1729-41.

166. Dixon, N., et al., *Reengineering orthogonally selective riboswitches.* Proc Natl Acad Sci U S A, 2010. **107**(7): p. 2830-5.

167. Edwards, T.E. and A.R. Ferre-D'Amare, *Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition.* Structure, 2006. **14**(9): p. 1459-68.

168. Kazantsev, A.V., et al., *Crystal structure of a bacterial ribonuclease P RNA.* Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13392-7.

169. Hall, M.A., *Correlation-based feature selection for machine learning (Doctoral Thesis)*, 1999, The University of Waikato.

170. Leontis, N.B., A. Lescoute, and E. Westhof, *The building blocks and motifs of RNA architecture.* Curr Opin Struct Biol, 2006. **16**(3): p. 279-87.

171. Petrov, A.I., C.L. Zirbel, and N.B. Leontis, *Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas.* RNA, 2013. **19**(10): p. 1327-40.

172. Dimitropoulos, D., J. Ionides, and K. Henrick, *Using MSDchem to search the PDB ligand dictionary.* Curr Protoc Bioinformatics, 2006. **Chapter 14**: p. Unit14 3.

173. Statnikov, A., D. Hardin, and C. Aliferis. *Using SVM Weight-Based Methods to Identify Causally Relevant and Non-Causally Relevant Variables*. in *Proceedings*

of the NIPS 2006 Workshop on Causality and Feature Selection. 2006. Vancouver, B.C., Canada.

174. Domingos, P., *A few useful things to know about machine learning.* Communications of the ACM, 2012. **55**(10): p. 78-87.

175. Guyon, I., et al., *Gene selection for cancer classification using support vector machines.* Machine learning, 2002. **46**(1-3): p. 389-422.

176. Leontis, N.B., J. Stombaugh, and E. Westhof, *The non-Watson-Crick base pairs and their associated isostericity matrices.* Nucleic Acids Res, 2002. **30**(16): p. 3497-531.

177. Tinoco, I., Jr. and C. Bustamante, *How RNA folds.* J Mol Biol, 1999. **293**(2): p. 271-81.

178. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.

179. Cruz-Toledo, J., et al., *Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments.* Database (Oxford), 2012. **2012**: p. bas006.

180. Esteller, M., *Non-coding RNAs in human disease.* Nat Rev Genet, 2011. **12**(12): p. 861-74.

181. Costa, F.F., *Non-coding RNAs: lost in translation?* Gene, 2007. **386**(1-2): p. 1-10.

182. Cruz, J.A., et al., *RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction.* RNA, 2012. **18**(4): p. 610-25.

183. Andronescu, M., et al., *Efficient parameter estimation for RNA secondary structure prediction.* Bioinformatics, 2007. **23**(13): p. i19-28.

184. Lu, Z.J., J.W. Gloor, and D.H. Mathews, *Improved RNA secondary structure prediction by maximizing expected pair accuracy.* RNA, 2009. **15**(10): p. 1805-13.

185. Mathews, D.H. and D.H. Turner, *Prediction of RNA secondary structure by free energy minimization.* Curr Opin Struct Biol, 2006. **16**(3): p. 270-8.

186. Callahan, A., J. Cruz-Toledo, and M. Dumontier, *Ontology-Based Querying with Bio2RDF's Linked Open Data.* Journal of Biomedical Semantics, 2013. **4 Suppl 1**: p. S1.

187. Brinkman, R.R., et al., *Modeling biomedical experimental processes with OBI.* J Biomed Semantics, 2010. **1 Suppl 1**: p. S7.