Lecture 9:3 RNA Structure and Function

Day 9: Day June 4, 2003: 13:45 – 15:15

Marcel Turcotte, University of Ottawa

Key Concepts

- **Structure and function.**
- **Primary, secondary and tertiary structure.**
- **Structure prediction.**
- **Dynamic programming algorithm.**
- **Graph representation.**
- **Constraint satisfaction problem.**

What you will be able to do at end of this section

- **Discuss structure and function.**
- **Predict secondary structure from sequence.**
- **Formally define tertiary structure.**
- **Modelling RNA structure using MCSYM.**

Overview

Sequence

RNAs are polymers of nucleotides. A **nucleotide** is constituted by a 5'-phosphorylated cyclic furanoside of type β-D-ribose, where the C1' atom is substituted by one of the four heterocycles that is connected by a β-glycosyl C1'-N covalent bond. The heterocycles are the four nitrogen bases: two **purines**, **adenine**, **guanine**, and two **pyrimidines**, **cytosine** and **uracil**. The nucleotides are linked together by 3'-5'-phosphodiester bonds.

The primary structure or **sequence** of RNA is the sequence of its constituent nucleotides and is almost always represented by single-character codes: the purines **A** and **G**, and the pyrimidines **C** and **U**. Formally, an RNA sequence is represented by the ordered list, $S = s_1, s_2, \ldots, s_{n,}$ where n is the length of the sequence and s_i is the i^{th} nucleotide.

In tertiary structure, the flexibility of nucleotide conformations is conferred by seven torsion angles: six in the backbone (α, β, γ, δ, ε, and ζ) and one around the glycosyl bond linking the ribose sugar and nitrogen base (χ). An important nucleotide conformation attribute is the sugar puckering mode which is strongly dominated by the C3'- and C2'-endo conformations, as found respectively in the A- and B-forms of double helices. The cyclic furanoside sugar can be either puckered with a single atom pointing out of the plane formed by the four others, by up to 0.5Å, or twisted with two adjacent atoms out of a plane formed by the other three. Atoms out of the three or four atom plane on the side of the C5' are said to be **endo**, and those on the opposite side are said to be **exo**. As presented by Saenger, the χ torsion produces two main orientations: the **anti**, which places the bulk of the nitrogen base oriented in the direction opposite to the sugar group, and the **syn**, which places the bulk of the nitrogen base oriented towards the sugar group.

Secondary structure

The nucleotides contain several hydrogen bond (H-bond) donors and acceptors that form multiple intra-molecular H-bonds and, in particular, **Watson-Crick** base pairs, which involve N-H...N and N-H...O hydrogen bonds, and the wobble hydrogen-bonding pattern between G and U bases. Two hydrogen bonds can bifurcate between one acceptor, for instance C=O, and two donor atoms on its partner, for instance two N-H groups. Hydrogen bonds can be mediated by water molecules, for instance N-H...O H_2O...H-N, can involve the C-H bond, such as in C-H...O or C-H...N, and can involve the ribose oxygen's, such as in N-H...O2'(H)-C2'. Classifications of base pair types, experimentally observed and theoretically derived, can be found at the institut fur molekulare biotechnologie, and the *Laboratoire de Biologie Informatique et Théorique* Web sites.

The contact faces of the nitrogen bases and their respective backbone orientations characterize the geometry of base pairs. Two base pairs sharing the same characteristics, but in particular the backbone orientations, are said **isosteric**. Gautheret and Gutell proposed a geometric test for concluding that two base pairs are isosteric. They superimposed the N1 (or N9) and C1' atoms of two base pairs, and then they measured the root mean square deviations between the atomic coordinates of the two base pairs, as well as the angles between the vectors formed by the N1 (or N9) and C1' atoms in both nucleotides and base pairs. When one nucleotide involved in a base pair is mutated, often the other gets mutated to preserve structure and function. Such double-point mutations can be detected in a set of evolutionary related sequences by **co-variation** analysis. For instance, because A-U and G=C Watson-Crick base pairs are isosteric, the co-variations A-U from/to G=C are frequently observed.

Base stacking involves dipole-dipole and dipole-induced dipole interactions (London dispersion), and hydrophobic forces, which, from a thermodynamic point of view, are weaker interactions than hydrogen bonds. Base stacking occurs more frequently between two nitrogen bases adjacent in the sequence, and is a dominant feature of the double-helix conformation, where the planes defined by the nitrogen bases are nearly parallel and their overlapping is maximized. Because weaker forces stabilize them, the conformational freedom and flexibility of two stacked bases is larger than that of base pairs. Nevertheless, information about base stacking reduces the conformational space of RNA tertiary structures.

Consider the yeast tRNA^{Phe} anticodon stem-loop. Base stacking can be observed between all nitrogen bases, except between those of U33 and G34. Base stacking restrains greatly the conformational freedom of the stem-loop, here the double-helical conformation for almost all nucleotides. At the double-helical breaking point, between U33 and G34, the particular conformation is called a U-turn because it involves a uracil, and it allows the RNA backbone to abruptly change direction. In general, stabilizing forces between two un-stacked nitrogen bases are necessary to maintain this conformation, here, base stacking between G34 and A35 and between U33 and C32. Nevertheless, the conformational space defined by un-stacked nucleotides is large and, most of the time, undesirable in tertiary structure prediction, except in rare occasions, such as the in the U-turn motif.

The **secondary structure** of RNA is a (sub-) set of its base pairs. The secondary structure can include base pairs whose formation results in a knotted diagram when the sequence is represented on a circle (**pseudoknot**). The computer prediction of secondary structures is based on the calculation of the free energy (ΔG) of different base pairing patterns. Thermodynamic parameters derived from oligonucleotides have permitted the approximation of free energy contributions of double-helical regions, dandling ends, terminal mismatches, and various loops.

A secondary structure on *S* is an ensemble of ordered pairs, *i.j, 1 <= i < j <= n* that satisfies:

1. j – i > 4
2. Given *i.j* and *i'.j'*, two base pairs, then either:
   a) *i = i'* and *j = j'* (they are the same)
   b) i < j < i' < j' (i.j precedes i'.j')
   c) i < i' < j' < j (i.j includes i'.j')
   d) i < i' < j < j' (pseudoknot)

Often, the condition 2.d (pseudoknot) is not allowed by the algorithms, because of the added complexity. Rivas and Eddy have proposed an implementation, based on dynamic programming, that allows for pseudonots, however, such algorithms generally require longer execution time. Pseudoknots are genuine and are important structural features of RNAs, they can sometimes be detected at a later step of the structure analysis.

Dynamic programming is a technique that is used to solve optimization problems that can be decomposed into sub-problems that have the same structure as the initial one. Nussinov was amongst the first to describe a dynamic programming algorithm for finding the maximum number of base pairs that can be formed by a sequence. Although the secondary structure that maximizes the number of base pairs do not generally correspond to the native structure, this simpler algorithm shares several characteristics with the more complex algorithms, such as **mfold.**

Nussinov's computation is recursive, *i.e.* the solution to a larger problem, here a larger sub-segment of the sequence, is obtained by combining the solutions of smaller problems, shorter sub-segments. The algorithm proceeds from the shortest allowable segments to the largest, and eventually finds a solution that encompasses the entire sequence. The key idea is to observe that the computation of the maximum number of base pairs that can be formed for the segment *i, j* is the maximum of 4 basic cases:

1. nucleotides *i* and *j* are paired, and the number of base pairs is 1 plus the number of base pairs that can be formed by the segment *i+1, j-1*, or;
2. nucleotide *i* is unpaired, and the number of base pairs is that of the segment *i+1, j*, or;
3. nucleotide *j* is unpaired, and the number of base pairs is that of the segment *i, j-1,* or;
4. the maximum number of base pairs is the sum of the number of base pairs for the segment *i, k*, and that of the segment *k+1, j*, for some *k*, where *i<k<j.*

At first, it would seem that these four cases are not enough to describe all possible secondary structures; in particular, it would seem that a segment could contain a maximum of two secondary structures (case 4). However, the number base pairs for each sub-segment is obtained by applying the same computation, and therefore each sub-segment can be made of two or more secondary structure elements. What makes this algorithm efficient is the fact that once a solution for a particular *i,j* has been found it is saved and will never be recomputed. All the intermediary solutions are saved in a two dimensional table of size *n* by *n*, when the algorithm terminates, the cell *1,n* will contain the maximum number of base pairs that can be formed for the entire sequence.

The more recent algorithms, such as mfold, are similar, but instead of maximizing the number of base pairs, these algorithms are finding a secondary structure that minimizes the free energy. A free energy table of all sequence fragments is implemented and used to find the optimal arrangement. The algorithm can be modified to generate sub-optimal structures. The simplest way to compute RNA structure prediction is to assign simple energy rules, such as an energy value to each base pair through a function, *e(i,j)*. Typical values of *e* at 37ºC are –3, -2, and –1 kcal/mole for the pairs CG, AU, and GU, respectively. The energy of the entire structure is the sum:

$$E(S) = \sum_{i.j \in S} e(i,j)$$

A recursive algorithm allows computing the minimum energy structure. Let W = min WS, where S ranges over all secondary structures. The energy for pairing s_i with s_j is given by *e(i,j)*. W_{ij} are computed for all fragments, *i…j* of the RNA.

$$W_{ij} = 0 \text{ for j-i} \leq 4$$

$$W_{ij} = \min\left\{\, W_{i+1,j},\ W_{i,j-1},\ e(i,j) + W_{i+1,j-1},\ \min_{k\,=\,i+1..j-1} (W_{i,k} + W_{k+1,j})\,\right\}$$

Either bases s_i or s_j do not pair, or s_i and s_j pair with each other, or there is a *k, i<k<j,* such as the sum of the contribution for each segment is minimum. More sophisticated energy rules are necessary to capture the destabilizing effects of various loops, or the nearest neighbor interactions in helices and loops. Zuker has implemented such rules in his computer program *mfold.*

How good is secondary structure prediction with *mfold* these days? Applied to the tRNA sequences, *mfold* predicts the correct cloverleaf structure 32 times out of 95. For 63 tRNAs *mfold* identifies the cloverleaf as a sub-optimal solution. Recently, improvements of the energy parameters were made (Mathews et al. 1999). For domains of 700 nucleotides or less, 73% of known base pairs were correctly predicted, as compared to 64% before the improvement.

The most recent approach of predicting secondary structure is to combine the thermodynamics approach with comparative sequence analysis methods. This will be discussed in Lecture 10:3.

Tertiary structure

Detailed knowledge about the three-dimensional structure is essential to understand, and eventually manipulate, RNA function. The **tertiary structure** (3-D) of RNA is defined by the set of its atomic coordinates. X-ray crystallography and NMR spectroscopy produce precise 3-D structures, although they are not completely free from technical and interpretation uncertainties. These two methods are now applied as commonly to RNA as to proteins.

Structure information can be represented as a graph, $G = (V, E)$, where V is the set of vertices, labeled by one of {A, C, G, U} representing each type of nucleotides, and E, is the set of edges representing the structural relations between pairs of nucleotides. In the case of adjacent nucleotides in the sequence, the direction of the edges follows the phosphodiester linkage from 5' to 3', and undirected edges are used for non-adjacent nucleotides, such as for H-bonding relations. An RNA structure graph can be projected in three-dimensions using a **discrete search method**.

Structural relations are encoded by homogeneous transformation matrices that correspond to the transformations, combinations of translations and rotations, of a nitrogen base local referential into another. The local referential of a nucleotide can be thought of as its local axis system and is computed from its atomic coordinates. Consider the arbitrary choice: use the N1 atom as the origin; align the C4 atom with the Y axis; and, align the C6 atom with the X axis The local referential of a nitrogen base B1, R_{B1}, and a nitrogen base B2, R_{B2} is now easily expressed in a homogeneous transformation matrix: $T_{B1\rightarrow B2} = R_{B1}^{-1} \times R_{B2}$. A nitrogen base spatial relation between B1 and B2 can be reproduced between any pair of nitrogen bases, say B1' and B2', by applying the homogeneous transformation matrix $R_{B2'}^{-1} \times T_{B1\rightarrow B2} \times R_{B1'}$ to the atomic coordinates of B2' to position and orient B2' with respect to B1'. Symmetrically, $R_{B1'}^{-1} \times T^{-1}_{B1\rightarrow B2} \times R_{B2'}$ applied to atomic coordinates of B1' will position and orient B1' relative to B2', according to $T_{B1\rightarrow B2}$. In this way, any nitrogen base spatial relation found in a 3-D structure can be extracted and used afterwards as a building block.

Although RNA structure is mainly constrained by nitrogen base relations, the ribose and phosphate groups still allow one to prune several inconsistent conformations. For instance, atomic coordinates of rigid nucleotide conformations can be used. The flexibility of nucleotide conformations is conferred by seven free torsion angles: six in the backbone (α, β, γ, δ, ε, and ζ) and one around the glycosyl bond linking the sugar pucker and nitrogen base. The other important parameter is the sugar puckering mode, which is strongly dominated by the C3'- and

C2'-***endo*** conformations, as found respectively in the A- and B-form double-helix. The cyclic furanoside sugar can be either puckered with a single atom pointing out of the plane formed by the four others, by up to 0.5 Å, or twisted with two adjacent atoms out of a plane formed by the other three. Atoms out of the three or four atom plane on the side of the C5' are said to be **endo**, and those on the opposite side are said to be **exo**.

An RNA **conformational search space** (CSS), is a set of tertiary structures that we are interested to explore. The CSS is defined by a series of parameters. The product of the numbers of allowed values per parameter defines the **size** of the CSS. As an example, consider the CSS defined by the seven torsion angles of nucleotides, where each angle is assigned one of 360 degrees. The size of such a CSS for a RNA of *N* nucleotides is thus 360^{7N}. **Operators** that modify the value of each parameter relate the tertiary structures of a CSS. In the above example, the operator that modifies the ζ torsion relates in the CSS two structures that differ by their values of the ζ torsion. In searching a CSS, we have a particular **goal** in mind, which describes what is searched for. Here, the goal corresponds to a subset of tertiary structures that satisfy what is described in the structural graph. Most search methods proceed by systematically applying a set of operators and verifying whether the resulting tertiary structures are elements of the goal. When it is impossible to perform an exhaustive search of the CSS, probabilistic methods, such as Monte Carlo or simulated annealing, can be applied.

A **metric** over a CSS allows us to compute some measure of the value of a given structure. The best example of such a metric is the model potential energy function defined for molecular mechanics and dynamics simulations. This metric can direct the search based on the assumption that applying the operators to a structure estimated to be closer to the goal will lead to it more rapidly than applying operators to more distant structures. For instance, the model potential energy function serves as an **evaluation function,** which assigns a value to each structure according to its potential energy. In practice, however, partly due to the **local minima** problem, the model potential energy function is rather employed to refine tertiary structures of the goal. Search methods guided by metrics are called **heuristics**, and are based on algorithms supposed to perform well in practice, although they provide no guarantees they will find the goal.

An **inference engine** searches the CSS for the goal. Computational techniques for predicting RNA tertiary structure differ by what information from the structural graph they use and how they use it, how they define the CSS, and what algorithm they employ to find the goal. An example is the **backtracking** algorithm implemented in the *MC-Sym* computer program, which generates tertiary structures for most types of RNA structural data. *MC-Sym* backtracks on systematic assignments of rigid base pairs and base stacking examples, as extracted from the tertiary structures deposited in databases such as the Protein Data Bank (PDB) and the Nucleic

Acid Database (NDB). Tertiary structures generated with the *MC-Sym* engine can be refined, and made thermodynamically sound, with the use of the *AMBER* or *CHARMM* packages and force fields.

The backtracking algorithm organizes the search space as a tree where each node corresponds to the application of an operator, which appends a new nucleotide to the structure. At each application, the consistency of the partial structure is evaluated. If consistent, the next operator is applied and the process continues. If inconsistent, this node and attached branches are pruned from the search tree and the algorithm ``backtracks'' to the previous node.

This algorithm was most popularized as it was used as the inference engine of logic programming languages, such as Prolog. The computational complexity of the backtracking algorithm was analyzed by Haralick in 1980, and only allows us to solve small problems. The backtracking search space grows exponentially with the number of variables, the number of allowed values for each variable; however, it decreases exponentially with the constraints. The backtracking search procedure is sound and complete, and is used to solve the discrete **constraint satisfaction problem**.

Key Computational Challenge

- Determine structure from sequence information.

Appendix

1. Resources

i) Original Papers

- Bouthinon D, Soldano H, *Bioinformatics* **15**, pp. 785-798 (1999).

- Cedergren R, Major F, RNA Structure and Function, R.W. Simons and M. Grunberg-Manago eds., Cold Spring Harbor Press, Cold Spring Harbor, NY., pp. 37-75 (1998).

- Gautheret D, Major, Cedergren R, *Journal of Molecular Biology* **229**, pp. 1049-1064 (1993).

- Gendron P, Lemieux S, Major F, *Journal of Molecular Biology* **308**, pp. 919-936 (2001).

- Lemieux S, Oldziej S, Major F, The Encyclopedia of Computational Chemistry, Schleyer, P. v. R.;Allinger, N. L., Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R. Eds., John Wiley & Sons: Chichester (1998).

- Leontis BN, Westhof E, *RNA* **7**, pp. 499-512 (2001).

- Major F, Griffey R, *Current Opinion in Structural Biology* **11**, pp. 282-286 (2001).

- Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E, Cedergren R, *Science* **253**, pp. 1255-60 (1991).

- Major F, Gautheret D, Encyclopedia of Molecular Biology and Molecular Medicine, - Robert A. Meyers ed., VCH Publishers Inc. New-York, NY, pp. 371-388 (1996).

- Major F, Gautheret D, Cedergren R, *PNAS (USA)* **90**, pp. 9408-9412 (1993).

- Mathews DH, Sabina J, Zuker M, Turner DH, *Journal of Molecular Biology* **288**, pp. 911-940 (1999).

- Moulton V, Zuker M, Steel M, Pointon R, Penny D, *Journal of Computational Biology* **7**, pp. 277-292 (2000).

- Storz G *Science* **296**, pp. 1260-1263 (2002).

- Rivas E, Eddy, *Journal of Molecular Biology* **285**, pp. 2053-2068 (1999).

- Walter A, Turner D, Kim J, Lyttle M, Müller P, Mathews D, Zuker M, *PNAS (USA)* **91**, pp. 9218-9222 (1994).

ii) Software

- Mfold, PKNOTS, RNAfold, RnaViz, MC-SYM, and Rasmol.

iii) Text books:

- RNA Structure and Function. R.W. Simons and M. Grunberg-Manago eds., Cold Spring Harbor Press, Cold Spring Harbor, NY. 1998.

- Stryer L. *Biochemistry*. Fourth edition. W.H. Freeman and Company, NY. 1995.

- Mathews C., van Hole K.E. *Biochemistry*. The Benjamin Cummings Publishing Company, Redwood City, CA. 1990.

- Mount, D.W. Bioinformatics, Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, NY 2001 (chapter 5).

- Durbin R., Eddy S., Krogh A. and Mitchison G. Biological Sequence Analysis. Cambridge University Press, Cambridge. 2000.

iv) Web Sites:

- bioinfo.math.rpi.edu/~zukerm/ (*mfold*)
- www.genetics.wustl.edu/eddy/software/#pk (*PKNOTS)*
- www.tbi.univie.ac.at/~ivo/RNA *(Vienna RNA Package)*
- rrna.uia.ac.be/rnaviz/ *(RnaViz)*
- www-lbit.iro.umontreal.ca/mcsym/ *(Mc-Sym)*

2. Presentation Overheads

Lecture 10:3 RNA Evolution

Day 10: Day June 5, 2003: 13:45 – 15:15

Marcel Turcotte, University of Ottawa

Key Concepts

- Sequence alignment.
- Comparative sequence analysis.
- Motifs.

What you will be able to do at end of this section

- Determine secondary and tertiary interactions.
- Define a motif from sequence alignment.
- Define a motif from structure.
- Search a motif in sequence databases.

Overview

Alignment

Despite constant progress estimating ΔG parameters for secondary structure prediction, comparative sequence analysis remains the gold standard in secondary structure determination. Statistical analyses of evolutionary related RNA sequences allow to determine secondary and tertiary structure interactions with great accuracy. Data gathered by Gutell and co-workers suggesting the existence of a large number of non Watson-Crick base pairs in both helical regions and tertiary interactions are particularly useful leading to a deeper understanding of RNA structure. Extension of non-canonical base pair data has a major conceptual impact, leading to important inferences on the overall conformation of a molecule by efficient pruning of a set of sound conformations. In particular, **co-variation** indicates **isosteric** base pairs that can be used in 3-D modelling.

Non-canonical base pairs are widespread in RNA structures. They form within **helices**, **loops**, or between different loops (GNRA-type loops, for instance, can function as anchors for outside helices). The G:U "wobble" mismatch is the most commonly observed mismatch. Wobbles induce specific distortions in double-helical structures that have been well characterized by crystallographic studies. G:A mismatches are frequent in helices and internal loops. NMR and crystallographic studies have allowed us to observe G:A mismatches in two different conformations. Some other non-canonical pairs of which structures have been observed include A:U pairs in the reverse Hoogsteen conformation, G:C pairs in the reverse Watson-Crick conformation or water-mediated C:U pairs. It has been established that many other non-canonical base- pairs exist (notably G:G, A:A and U:U).

Non-canonical base pairs are powerful 3-D constraints when they involve distant positions, such as 19:56 in tRNA. However, there are multiple ways of constructing single or double H-bond pairing conformations for a given base pair. An A:G pair, for instance, can adopt four possible geometries with 2 H-bonds. There is no reliable method to determine which conformation is adopted from one sequence alone. In modeling studies, the stability and **steric** feasibility of each conformation must be assessed in its specific environment. To perform this task, model builders also take into account the results of mutational and phylogenetic studies that determine the set of "acceptable" sequences at certain positions, thereby suggesting a unique base pair conformation.

Base-triples consist of three H-bonded bases. They have been observed in crystal structures of tRNA and, indirectly, by phylogenetic studies and mutational experiments, in group I introns and RNase P RNA. Base-triples are commonly composed of a Watson-Crick pair to which a third nucleotide is connected via non-canonical H-bond interactions. However, there are also base-triples where all interactions are non-canonical, such as in the *T.thermophilus* tRNA^{Ser}. As for non-canonical base pairs, the sequence of base-triple alone does not allow for the determination of their exact conformation. However, triple structures have been inferred from the interpretation of phylogenetic and mutational studies.

RNA sequences are difficult to align without structural information. In fact, it is the problem of capturing the essentials of RNA sequences that is computationally difficult. Only sophisticated data structures and algorithms allow us to represent an RNA sequence alignment. For this reason, Chteinberg has aligned the tRNA sequences manually using crystal structure information, and co-variation analysis. Gutell applied the same analysis for ribosomal RNA sequences, Brown and Pace for Rnase P RNA sequences, and Michel for group I and II intron sequences. In general, the "manual" alignments fit better the actual structural information, when compared to the alignments obtained by programs such as *clustalw*, or the probabilistic algorithm of *RNASA*. In fact, most of the predictions made by Gutell and co-workers concerning ribosomal RNA were verified in the recent crystal structure.

RNASA refines a rough input alignment using a parallel simulated annealing algorithm. The input specifies a set of connected base pairs, such as those in stem regions. After twelve-hour of annealing process beginning with a rough alignment computed from a parallel iterative algorithm, this system generated the final RNA alignment. The alignment identified well-known tRNA stems of the cloverleaf secondary structure. Co-variation data were considered at the generation of the rough, input, alignment.

The sequence alignment and modeling systems (*SAM* and *HMMER*) are a collection of flexible software tools for creating, refining, and using linear Hidden Markov Models (HMM) for biological sequence analysis. The model states can be viewed as representing the sequence of columns in a multiple sequence alignment, with provisions for arbitrary position-dependent insertions and deletions in each sequence. The models are trained on a family of protein or nucleic acid sequences using an expectation-maximization algorithm and a variety of algorithmic heuristics. A trained model can then be used to both generate multiple alignments and search databases for new members of the family. *SAM* is written in the C programming language for Unix machines, and includes extensive documentation. A regular HMM, such as the one used in *SAM*, cannot be used to represent secondary structure information.

In 1997, Notre-Dame and co-workers have clearly shown, using their program RAGA, that including base pair matching in sequence alignment improves much its accuracy, when compared to alignments made by experts.

Covariance analysis

Covariance analysis consists of measuring the degree of association between pairs of positions in a multiple sequence alignment. In nucleic acid sequences, the signal produced by compensatory changes is strong. Simultaneous changes that preserve Watson-Crick base pair (G.C becomes C.G, for example) are frequently observed in stems, but more complex replacement patterns are also possible. The degree of association between pairs of positions is often measured using **mutual information**; this allows detecting complex substitution patterns corresponding to secondary and tertiary structure interactions. Let's denote by **M** the multiple sequence alignment, containing *m* sequences and *n* positions, that serves as input for the covariance analysis. The first step consists of measure the degree of association between all possible pairs of positions *i, j, 1≤i<j≤n.* Let the random variables *I* and *J* denote the i^{th} and j^{th} columns, respectively, of the multiple sequence alignment **M**. The mutual information is defined

$$M(I,J) = H(I) + H(J) - H(I,J)$$

as:

where *H* is the entropy. This quantity will be at its maximum when the entropy at both positions is maximum, but the interdependence entropy for *I* and *J* is minimum – when two random variables are "perfectly correlated", knowing the outcome of one of the two variables allows inferring the outcome of the other, and the entropy/uncertainty is therefore zero. The entropy is defined as:

where the sample space *S* consists of *{A, C, G, U, -}*. The entropy attains its maximum when all

$$H(I) = -\sum_{\alpha \in S} P(i = \alpha) \log P(i = \alpha)$$

the events are equiprobable (occur with the same probability). The interdependence entropy is defined as follows:

$$H(I,J) = -\sum_{\alpha\beta} P(i = \alpha, j = \beta) \log P(i = \alpha, j = \beta)$$

The probability functions *P(I)* and *P(J)* may be estimated from the data by counting the number of occurrences of each base, at each of the two positions and dividing by the number of sequences, *m*. Similarly, *P(I,J)* may be estimated by calculating the observed frequency of each pair of bases at positions *i* and *j*. Since it has been shown that the mutual information has an asymptotic chi-square distribution, a statistical test may be devised to screen the statistically significant interactions. Finally, the statistically significant interactions can be divided into two sets, non-crossing and crossing interactions, where non-crossing interactions correspond to secondary structure interactions and crossing interactions correspond to tertiary interactions.

Motifs

A structural RNA **motif** is a recurrent subset of nucleotide arrangements in secondary and/or tertiary structure. RNA motifs are represented by **graphs of relations** where the **nodes** represent the nucleotides and the **edges** represent structural relations. A distinction is made between the identification of RNA motifs from known secondary and tertiary structures, which requires a sophisticated incremental search algorithm, and the search for a particular motif in sequence databases, which is accomplished using available programs, such as *RNAMOT*.

With an increasing number of important RNA structure, and the rapid growth of sequence databases, the detection of functional RNA genes in sequenced DNA has become an important task in many laboratories. RNA molecules, unlike DNA, are poorly described by simple sequence information, limiting the application of sequence search tools such as *Blast* or *Fasta*. Woese *et al.* showed in phylogenetic studies of 16S rRNA that many regions preserve base pairing patterns, even though the sequences vary considerably. Also, think of the divergent sequences, but common secondary structure, of telomere RNAs. Searching for a specific RNA motif in sequence databases, whether identified from the above incremental graph isomorphism procedure or otherwise, is a task of searching for well-conserved secondary structures rather than that of sequences. The use of information about secondary and tertiary structure in database searches is essential, for no other reason than the dominant role they play in RNA function. Such specific motif searches are performed using available computer programs, such as *tRNA-Scan* by Lowe and Eddy, and *FAStRNA* by El Mabrouk and Lisacek, which search specifically for tRNA motifs, and *RNAMOT* by Gautheret, Major and Cedergren, for general motif searches.

The input of general motif search procedures, like *RNAMOT*, requires a description of the motif in terms of its secondary and tertiary structure, the **descriptor** or **pattern** (see Figure 1). The descriptor lists the structural elements (SE), their position and length. "S" is used for single-stranded regions, and "H" for helical regions. Helix notations are presented twice to define both strands. Any SE arrangement is permitted, including knots, as long as every helix is delineated by a single strand (possibly of length zero). It is not possible to pair the nucleotides in the single-stranded regions, although this often occurs in RNAs. The list of SE is followed by their arguments. Helices (H) have minimum and maximum **lengths**, a number of allowed **mismatches**, and any primary **sequence** restrictions. The same arguments are defined for single-stranded regions, except, obviously, for the number of allowed mismatches. The IUPAC-IUB notation is used for sequence restrictions. Optional declarations of the search order "R", and the total number of allowed base mismatches for the motif "M" can be used.

```
H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

H1 3:5 0
H2 4:5 1 AGC:GCU
H3 4:5 1
S1 3:6 ucc
S2 5:7
S3 0:3
S4 5:8 gaag
S5 3:5

R H2 H3 H1
M 1
```

Figure 1. An RNAMOT input script.

One of the difficulties of using descriptors like these, or the graph of relations, is to find the best descriptor for a **class** (or **family**) of RNAs. Consider, for example, tRNAs. An *RNAMOT* descriptor was defined that found 95.8% of the tRNAs in all tRNA sequences, or 99% of the cytoplasmic tRNAs. However, before reaching this definition, another descriptor that seemed equally reasonable identified only 79.8% of all tRNAs, or 86.1% of the cytoplasmic tRNAs. One of the problems is to capture the most important structural features for any RNA function.

Transformational grammars provide a formal framework to describe string motifs. The grammars are organized into a hierarchy, called the **Chomsky hierarchy of transformational grammars** – named after the computational linguist that developed the bulk of this theory. The least expressive grammars are the **regular grammars** (PROSITE motifs are described with regular expressions, which is an example of a regular grammar). Regular grammars are modelling the primary sequence. In order to model long-range interactions, such as those of RNA secondary structure, a more general class of grammars must be used, **the context-free grammars** (CFG). CFGs allow modelling both primary and secondary structure but their computational cost (time and memory) is much higher.

Transformational grammars consist of symbols and rewriting rules. There are two kinds of symbols: non-terminal and terminal. The former type of symbols is abstract and used to model the structure of the string. The later kind corresponds to observed characters of the input string.

A grammar provides a formalism to describe a motif that can then be used to find new instances of this motif. The application of a motif to a string (sequence) is called **parsing**. It consists in finding a sequence of application of rewriting rules transforming the start non-terminal into the observed string – this process is represented as a **derivation** or **parse tree**. Stochastic CFG (SCFG) are CFG that have probabilities assigned to each production (rewriting) rule. Similarly, a HMM can be seen as a probabilistic regular grammar. These probabilities are useful to evaluate the joint probability of a derivation tree, or simply to compute the most probable derivation tree. Haussler and co-workers successfully applied SCFGs to the tRNA sequences. The SCFG they used was obtained automatically from an alignment of tRNA sequences. Eddy developed a similar approach.

Key Computational Challenge

- Automatically determine a higher-order motif to characterize a family of sequences.

Appendix

3. Resources

iii) Original Papers

- Barrette I et al. *Nucleic Acids Res*. **29**, pp. 753-758 (2001).
- Chen JH, Le SY, Maizel JV, *Nucleic Acids Research* **28**, pp. 991-999 (2000).
- Chiu DKY, Kolodziejczak, *Comput. Appl. Biosci*. **7**, pp. 347-352 (1991).
- Eddy SR, Durbin R, *Nucleic Acids Research*. **22**, pp. 2079-2088 (1994).
- Gautheret D, Major F, Cedergren R, *Comput Appl Biosci*. **6**, pp. 325-331 (1990).
- Juan V, Wilson C, *Journal of Molecular Biology* **289**, pp. 935-947 (1999).
- Knudsen B, Hein, J *Bioinformatics* **15**, pp. 446-454 (1999).
- Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, Underwood RC, Haussler D, *Nucleic Acids Research* **22**, pp. 5112-5120 (1994).
- Notre-Dame C, O'Brien EA, Higgins DG, *Nucleic Acids Research* **25**, pp. 4570-80 (1997).

iv) Software

- *SAM* and *HMMER*
- *RNASA*
- *RAGA*
- *RNAMOT* (and *RNABOB*)

iii) Text books:

- Durbin R., Eddy S., Krogh A. and Mitchison G. Biological Sequence Analysis. Cambridge University Press, Cambridge. 2000.

iv) Web Sites:

- www.cse.ucsc.edu/research/compbio/sam.html (*SAM*)
- hmmer.wustl.edu/ (HMMER)
- www.esil.univ-mrs.fr/~dgaut/download/ (*RNAMOT*)
- www.genetics.wustl.edu/eddy/software/#rnabob (*RNABOB*)
- www.genetics.wustl.edu/eddy/software/#trnascan (tRNA-Scan)
- www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/ (tRNA sequence and gene database)
- jwbrown.mbio.ncsu.edu/RNaseP/ (RNase P database)

- www.rna.icmb.utexas.edu/ (rRNA database)
- www-lbit.iro.umontreal.ca/RNA_Links/RNA.shtml(RNA links)

4. Presentation Overheads