RNA: Secondary Structure Prediction and Analysis
Outline

1. RNA Folding
3. Covariance Model for RNA Structure Prediction
4. Small RNAs: Identification and Analysis
Section 1: RNA Folding
RNA Basics

- RNA bases: A, C, G, U

- Canonical Base Pairs
  - A-U
  - G-C
  - G-U

- Bases can only pair with one other base.
RNA Basics

- RNA bases: A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U
- Bases can only pair with one other base.
RNA Basics

- RNA bases: A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U
- Bases can only pair with one other base.

3 Hydrogen Bonds—more stable
RNA Basics

- RNA bases: A, C, G, U

- Canonical Base Pairs
  - A-U
  - G-C
  - G-U

- Bases can only pair with one other base.
RNA Basics

- Various types of RNA:
  - transfer RNA (tRNA)
  - messenger RNA (mRNA)
  - ribosomal RNA (rRNA)
  - small interfering RNA (siRNA)
  - micro RNA (miRNA)
  - small nuclear RNA (snRNA)
  - small nucleolar RNA (snoRNA)

http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Section 2: Dynamic Programming for RNA Secondary Structure Prediction
RNA Secondary Structure

- Single-Stranded
- Stem
- Bulge Loop
- Hairpin loop
- Interior Loop
- Junction (Multiloop)
- Pseudoknot

Image– Wuchty
Sequence Alignment to Determine Structure

• Bases pair in order to form backbones and determine the secondary structure.

• Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure.
Base Pair Maximization: Dynamic Programming

- $S(i,j)$ is the folding of the RNA subsequence of the strand from index $i$ to index $j$ which results in the highest number of base pairs.

- Recurrence:
Base Pair Maximization: Dynamic Programming

• $S(i, j)$ is the folding of the RNA subsequence of the strand from index $i$ to index $j$ which results in the highest number of base pairs.

• Recurrence:

$$S(i, j) = \max \begin{cases} 
\end{cases}$$

Images – Sean Eddy
Base Pair Maximization: Dynamic Programming

- \( S(i, j) \) is the folding of the RNA subsequence of the strand from index \( i \) to index \( j \) which results in the highest number of base pairs.

- Recurrence:

\[
S(i, j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{(if } (i, j) \text{ base pair)} \\
S(i, j) & \text{(if } (i, j) \text{ not base pair)} 
\end{cases}
\]
Base Pair Maximization: Dynamic Programming

- $S(i, j)$ is the folding of the RNA subsequence of the strand from index $i$ to index $j$ which results in the highest number of base pairs.

- Recurrence:

\[
S(i, j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{(if } (i, j) \text{ base pair)} \\
S(i + 1, j) & 
\end{cases}
\]

Images – Sean Eddy
Base Pair Maximization: Dynamic Programming

- $S(i, j)$ is the folding of the RNA subsequence of the strand from index $i$ to index $j$ which results in the highest number of base pairs.

- Recurrence:

$$S(i, j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{(if } (i, j) \text{ base pair)} \\
S(i + 1, j) & \\
S(i, j - 1) & 
\end{cases}$$
Base Pair Maximization: Dynamic Programming

- $S(i, j)$ is the folding of the RNA subsequence of the strand from index $i$ to index $j$ which results in the highest number of base pairs.

- Recurrence:

$$S(i, j) = \max \begin{cases} S(i + 1, j - 1) + 1 & \text{(if } (i, j) \text{ base pair)} \\ S(i + 1, j) \\ S(i, j - 1) \\ \max_{1 < k < j} S(i, k) + S(k + 1, j) \end{cases}$$

Images – Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension
Base Pair Maximization: Dynamic Programming

- **Alignment Method:**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- **Each score independent of overall structure**
- **Bifurcation adds extra dimension**

Initialize first two diagonals to 0

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Fill in squares sweeping diagonally
Base Pair Maximization: Dynamic Programming

- **Alignment Method:**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Fill in squares sweeping diagonally.
Base Pair Maximization: Dynamic Programming

- **Alignment Method:**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bases cannot pair

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bases can pair, similar to matched alignment

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Dynamic Programming—possible paths

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

**Dynamic Programming—possible paths**

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

\[
S(i + 1, j - 1) + 1
\]
Base Pair Maximization: Dynamic Programming

- **Alignment Method:**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Bifurcation—add values for all $k$

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Bifurcation—add values for all $k$

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bifurcation—add values for all \( k \)

Images—Sean Eddy
Base Pair Maximization: Dynamic Programming

• Alignment Method:
  • Align RNA strand to itself
  • Score increases for feasible base pairs
• Each score independent of overall structure
• Bifurcation adds extra dimension

Bifurcation—add values for all $k$

Images—Sean Eddy
Base Pair Maximization: Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure.
  - It may create structure with many interior loops or hairpins which are energetically unfavorable.

- In comparison to aligning sequences with scattered matches—not biologically reasonable.
Energy Minimization

• Thermodynamic Stability
  • Estimated using experimental techniques.
  • Theory: Most Stable = Most likely
• No pseudoknots due to algorithm limitations.
• Attempts to maximize the score, taking thermodynamics into account.
• MFOLD and ViennaRNA
Energy Minimization Results

• Linear RNA strand folded back on itself to create secondary structure
• Circularized representation uses this requirement
  • Arcs represent base pairing

Images – David Mount
Energy Minimization Results

- All loops must have exactly three bases in them.
  - Equivalent to having at least three base pairs between arc endpoints.

Images – David Mount
Energy Minimization Results

• All loops must have exactly three bases in them.
  • Equivalent to having at least three base pairs between arc endpoints.

Images – David Mount
Energy Minimization Results

• All loops must have exactly three bases in them.
• **Exception**: Location where beginning and end of RNA come together in circularized representation.
Energy Minimization Results

- All loops must have exactly three bases in them.
- **Exception**: Location where beginning and end of RNA come together in circularized representation.
Trouble with Pseudoknots

- Pseudoknots cause a breakdown in the dynamic programming algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired—this breaks down the recurrence relations.

Images – David Mount
Trouble with Pseudoknots

- Pseudoknots cause a breakdown in the dynamic programming algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired—this breaks down the recurrence relations.
Trouble with Pseudoknots

- Pseudoknots cause a breakdown in the dynamic programming algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired—this breaks down the recurrence relations.

Images – David Mount
Energy Minimization: Drawbacks

• Computes only one optimal structure.

• Optimal solution may not represent the biologically correct solution.
Section 3: Covariance Model for RNA Structure Prediction
Alternative Algorithms - Covariation

- Incorporates Similarity-based method
  - Evolution maintains sequences that are important
  - Change in sequence coincides to maintain structure through base pairs (Covariance)
    - Cross-species structure conservation example – tRNA
- Manual and automated approaches have been used to identify covarying base pairs
- Models for structure based on results
  - Ordered Tree Model
  - Stochastic Context Free Grammar
Alternative Algorithms - Covariation

- Expect areas of base pairing in tRNA to be covarying between various species.
Alternative Algorithms - Covariation

- Expect areas of base pairing in tRNA to be covarying between various species.
- Base pairing creates same stable tRNA structure in organisms.
Alternative Algorithms - Covariation

- Expect areas of base pairing in tRNA to be covarying between various species.
- Base pairing creates same stable tRNA structure in organisms.
- Mutation in one base yields pairing impossible and breaks down structure.
Alternative Algorithms - Covariation

• Expect areas of base pairing in tRNA to be covarying between various species.
• Base pairing creates same stable tRNA structure in organisms.
• Mutation in one base yields pairing impossible and breaks down structure.
• Covariation ensures ability to base pair is maintained and RNA structure is conserved.
Binary Tree Representation of RNA Secondary Structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Binary Tree Representation of RNA Secondary Structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Binary Tree Representation of RNA Secondary Structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Binary Tree Representation of RNA Secondary Structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Covariance Model

- **Covariance Model**: HMM which permits flexible alignment to an RNA structure – emission and transition probabilities
- Model trees based on finite number of states
  - Match states – sequence conforms to the model:
    - MATP: State in which bases are paired in the model and sequence.
    - MATL & MATR: State in which either right or left bulges in the sequence and the model.
  - Deletion – State in which there is deletion in the sequence when compared to the model.
  - Insertion – State in which there is an insertion relative to model.
Covariance Model

- **Covariance Model**: HMM which permits flexible alignment to an RNA structure – emission and transition probabilities
- Transitions have probabilities.
  - Varying probability: Enter insertion, remain in current state, etc.
  - Bifurcation: No probability, describes path.
Covariance Model (CM) Training Algorithm

- $S(i, j) =$ Score at indices $i$ and $j$ in RNA when aligned to the Covariance Model.
  \[
  S(i, j) = \max \begin{cases} 
  S(i + 1, j - 1) + M(i, j) \\
  S(i + 1, j) \\
  S(i, j - 1) \\
  \max_{i < k < j} S(i, k) + S(k + 1, j)
  \end{cases}
  \]

- Frequencies obtained by aligning model to “training data”—consists of sample sequences.
  - Reflect values which optimize alignment of sequences to model.
Covariance Model (CM) Training Algorithm

- \( S(i, j) = \) Score at indices \( i \) and \( j \) in RNA when aligned to the Covariance Model.

\[
S(i, j) = \max \begin{cases}
S(i + 1, j - 1) + M(i, j) \\
S(i + 1, j) \\
S(i, j - 1) \\
\max_{i < k < j} S(i, k) + S(k + 1, j)
\end{cases}
\]

- \( M_{i,j} = \sum_{x_i,x_j} f_{x_i,x_j} \log_2 \frac{f_{x_i,x_j}}{f_{x_i} f_{x_j}} \)

Frequency of seeing the symbols (A, C, G, T) together in locations \( i \) and \( j \) depending on symbol.

- Frequencies obtained by aligning model to “training data”—consists of sample sequences.
- Reflect values which optimize alignment of sequences to model.
Covariance Model (CM) Training Algorithm

- \( S(i,j) = \) Score at indices \( i \) and \( j \) in RNA when aligned to the Covariance Model.

\[
S(i,j) = \max \begin{cases} 
S(i+1,j-1) + M(i,j) \\
S(i+1,j) \\
S(i,j-1) \\
\max_{i<k<j} S(i,k) + S(k+1,j)
\end{cases}
\]

- \( M_{i,j} = \sum_{x_i, x_j} f_{x_i,x_j} \log_2 \frac{f_{x_i,x_j}}{f_{x_i}f_{x_j}} \)

Independent frequency of seeing the symbols (A, C, G, T) in locations \( i \) or \( j \) depending on symbol.

- Frequencies obtained by aligning model to “training data”—consists of sample sequences.
- Reflect values which optimize alignment of sequences to model.
Covariance Model (CM) Training Algorithm

- \( S(i, j) = \) Score at indices \( i \) and \( j \) in RNA when aligned to the Covariance Model.

\[
S(i, j) = \max \begin{cases} 
S(i+1, j-1) + M(i, j) \\
S(i+1, j) \\
S(i, j-1) \\
\max_{i < k < j} S(i, k) + S(k+1, j)
\end{cases}
\]

- Frequencies obtained by aligning model to “training data”—consists of sample sequences.
  - Reflect values which optimize alignment of sequences to model.

\[
M_{i,j} = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}
\]

Independent frequency of seeing the symbols (A, C, G, T) in locations \( i \) or \( j \) depending on symbol.
Alignment to CM Algorithm

- Calculate the probability score of aligning RNA to CM.

- Three dimensional matrix—$O(n^3)$
  - Align sequence to given subtrees in CM.
  - For each subsequence, calculate all possible states.

- Subtrees evolve from bifurcations
  - For simplicity, left singlet is default.

Images—Eddy et al.
Alignment to CM Algorithm

- For each calculation, take into account:
  - Transition (T) to next state.
  - Emission probability (P) in the state as determined by training data.

\[
S_{i,j,y}(y = MATP) = \max_{y_{next}} [S_{i+1,j-1,y_{next}} + \log T(y_{next} | y) + \log P(x_i, x_j | y)]
\]

Images—Eddy et al.
Alignment to CM Algorithm

- For each calculation, take into account:
  - Transition (T) to next state.
  - Emission probability (P) in the state as determined by training data.

$$S_{i,j,y}(y = MATL, INSL) = \max_{y_{next}} [S_{i+1,j,y_{next}} + \log T(y_{next} | y) + \log P(x_i | y)]$$

Images—Eddy et al.
Alignment to CM Algorithm

• For each calculation, take into account:
  • Transition (T) to next state.
  • Emission probability (P) in the state as determined by training data.

\[
S_{i,j,y}(y = MATR, INSR) = \max_{y_{next}} \left[ S_{i,j-1,y_{next}} + \log T(y_{next} \mid y) + \log P(x_j \mid y) \right]
\]

Images—Eddy et al.
Alignment to CM Algorithm

- For each calculation, take into account:
  - Transition (T) to next state.
  - Emission probability (P) in the state as determined by training data.

- Deletion—does not have emission probability associated with it.

\[ S_{i,j,y}(y = DEL) = \max_{y_{next}} [S_{i,j,y_{next}} + \log T(y_{next} | y)] \]

Images—Eddy et al.
Alignment to CM Algorithm

• For each calculation, take into account:
  • Transition (T) to next state.
  • Emission probability (P) in the state as determined by training data.

• Deletion—does not have emission probability associated with it.
• Bifurcation—does not have state probability associated with it.

\[ S_{i,j,y}(y = BIFURC) = \max_{i-1 \leq mid \leq j} [S_{i,mid,y_{left}} + S_{mid+1,j,y_{right}}] \]

Images—Eddy et al.
Covariance Model Drawbacks

• Needs to be well trained.

• Not suitable for searches of large RNA.
  • Structural complexity of large RNA cannot be modeled
  • Runtime
  • Memory requirements
Section 4: Small RNAs: Identification and Analysis
Discovery of small RNAs

• The first small RNA:
  • In 1993 Rosalind Lee was studying a non-coding gene in *C. elegans*, lin-4, that was involved in silencing of another gene, lin-14, at the appropriate time in the development of the worm *C. elegans*.
  • Two small transcripts of lin-4 (22nt and 61nt) were found to be complementary to a sequence in the 3' UTR of lin-14.
  • Because lin-4 encoded no protein, she deduced that it must be these transcripts that are causing the silencing by RNA-RNA interactions.
  • The second small RNA wasn't discovered until 2000!
What are small ncRNAs?

- Two flavors of small non-coding RNA:
  1. micro RNA (miRNA)
  2. short interfering RNA (siRNA)
- Properties of small non-coding RNA:
  - Involved in silencing other mRNA transcripts.
  - Called “small” because they are usually only about 21-24 nucleotides long.
  - Synthesized by first cutting up longer precursor sequences (like the 61nt one that Lee discovered).
  - Silence an mRNA by base pairing with some sequence on the mRNA.
miRNA Pathway Illustration

1. The microRNA (miRNA) precursor folds back on itself, held together by hydrogen bonds.
2. An enzyme called Dicer moves along the double-stranded RNA, cutting it into shorter segments.
3. One strand of each short double-stranded RNA is degraded; the other strand (miRNA) then associates with a complex of proteins.
4. The bound miRNA can base-pair with any target mRNA that contains the complementary sequence.
5. The miRNA-protein complex prevents gene expression either by degrading the target mRNA or by blocking its translation.
Complementary base pairing facilitates the mRNA cleavage
Features of miRNAs

- Hundreds miRNA genes are already identified in human genome.

- Most miRNAs start with a U

- The second 7-mer on the 5' end is known as the “seed.”
  - When an miRNAs bind to their targets, the seed sequence has perfect or near-perfect alignment to some part of the target sequence.

- **Example:** UGAGCUUAGCAG...
Features of miRNAs

- Many miRNAs are conserved across species:
  - For half of known human miRNAs, >18% of all occurrences of one of these miRNA seeds are conserved among human, dog, rat, and mouse.

- As a rule, the full sequence of miRNAs is almost never completely complementary to the target sequence.
  - Common to see a loop or bulge after the seed when binding.
  - Loop/bulge is often a hairpin because of stability.

- The site at which miRNAs attack is often in their target's 3' UTR.
Hairpin is more stable than a simple bulge

The MRE is known as the “miRNA recognition element.” This is simply the sequence in the target that an miRNA binds to.
Locating miRNA Genes: Experimentally

- Locating miRNA experimentally is difficult.

- Procedure:
  1. Find a gene that causes down-regulation of another gene.
  2. Determine if no protein is encoded.
  3. Analyze the sequence to determine if it is complementary to its target.
Locating miRNA Genes: Comparative Genomics

• **Idea:** Find the seed binding sites.
  1. Examine well-conserved 3' UTRs among species to find well-conserved 8-mers (A + seed) that might be an miRNA target sequence.
  2. Look for a sequence complementary to this 8-mer to identify a potential miRNA seed. Once found, check flanking sequence to see if any stable hairpin structures can form—these are potentially pre-miRNAs.
  3. To determine the possibility of secondary RNA structure, use RNAfold.
Locating miRNA Genes: Example

• Suppose you found a well-conserved 8-mer in 3' UTRs (this could be where an miRNA seed binds in its target).
  • **Example**: AGACTAGG

• Look elsewhere in genome for complementary sequence (this could be an miRNA seed).
  • **Example**: TCTGATCC

• When TCTGATCC is found, check to see (with RNAfold) if the sequences around it could form hairpin; if so, this could be an miRNA gene.
Finding miRNA Targets: Method 1

• Now we know of some miRNAs, but where do they attack?
• **Goal:** Find the targets of a set of miRNAs that are shared between human and mouse.
  • Looking for the miRNA recognition element (MRE), not whole mRNA. This is just the part that the miRNA would bind to.
• **Basic Assumption:** Whole miRNA:MRE interactions (binding) are likely to have highly energetically favorable base pairing.
• **Basic Method:** Look through the conserved 3' UTRs—this is where the MREs are most likely to be located—and try to make an alignment that minimizes the binding energy between the miRNA sequence and the UTRs (most favorable).
Finding miRNA Targets: Method 1

- **Method:**
  - First look at the binding energies of all 38-mers of the mRNA when binding to the miRNA. Subsequently apply several filters to pick alignments that “look” like miRNA binding.
  - Why 38-mers? ~22 nt for the miRNA and the rest to allow for bulges, loops, etc.

- **Algorithm:** Use a modified dynamic programming sequence alignment algorithm to calculate the binding energies for each 38-mer.

- **Modifications:** Scoring and speedup
Finding miRNA Targets: Method 1

**Scoring:**
- Mismatches and indels allowed.
- Matrix based on RNA-RNA binding energies.
  - Use known binding energies of Watson-Crick pairing and wobble (G-U) pairing.
- Binding energy (score) calculated for every two adjacent pairings (unlike the standard alignment algorithm which just takes into account the “score” for one pair at a time).
  - Adds dimensions to scoring matrix.
  - Adds complexity to recurrence relation.
Finding miRNA targets: Method 2

- **Goal**: Find the set of miRNA targets for miRNAs shared across multiple species
  - Trying to identify which genes have 3' UTRs are attacked by miRNAs
- **Basic Assumptions**:
  1. There is perfect binding to the miRNA seed.
  2. Any leftover sequence wants to achieve optimal RNA secondary structure.
- **Basic Method**: For each species’ set of 3' UTRs, find sites where there is perfect binding of the miRNA seed and “optimal folding” nearby. Look for agreement among all the species.
Method 2: Example
Method 2: Steps

1. Find a perfect match to the miRNA seed.

2. Extend the matching region if possible.

3. Find the optimal folding for the remaining sequences.

4. Calculate the energy of this interaction.
Method 2: Details

- **Input**: A set of miRNAs conserved among species and a set of 3' UTR sequences for those species.

- **Method**: For each organism:
  1. Find all occurrences in the UTR sequences that match the miRNA seed exactly.
  2. Extend this region with perfect or wobble pairings.
  3. With the remaining sequence of the miRNA, use the program RNAfold to find optimal folding with the next 35 bases of the UTR sequence.
  4. Calculate a score for this interaction based on the free energy of the interaction given by RNAfold.
Method 2: Details

- **Method Cont.:**
  5. Sum up the scores of all interactions for each UTR.
  6. Rank all the organism's gene's UTRs by this score (sum of all interactions in that UTR).
  7. Repeat the above steps for each organism.
  8. Create a cutoff score and a cutoff rank for the UTRs.
  9. Select the set of genes where the orthologous genes across all the sampled species have UTR's that score and rank above this cutoff.
Method 2: Details

• Verification:
  • Find the number of predicted binding sites per miRNA.
  • Compare it to number of binding sites for a randomly generated miRNA.
  • The result is much higher.
Analysis of the Two Methods

• **Method 1:**
  • Good at identifying very strong, highly complementary miRNA targets.
  • Found gene targets with one miRNA binding site, failed to identify genes with multiple weaker binding sites.

• **Method 2:**
  • Good at identifying gene targets that have many weaker interactions.
  • Fails to identify single-site genes.
Analysis of the Two Methods

• **Both Methods:**
  • Speed is an issue.
  • Won't find targets that aren't in the 3' UTR of a gene.

• **We need more species sequenced!**
  • Conserved sequences are used to discover small RNAs.
  • Conserved small RNAs are used to discover targets.
  • Confidence in prediction of small RNAs and targets.
  • Allows for broader scope with different combinations of species.
Results

• Predicted a large portion of already known targets and provided direction for identifying undiscovered targets.

• Found that it is more common that genes are regulated by multiple small RNAs.

• Found that many small RNAs have multiple targets.
A Novel siRNA Mechanism

• Recently, a new mechanism of siRNA activity was discovered.
  • Two genes (called A and B here) that have a cis-antisense orientation (they are overlapping on opposite strands) have transcripts that produce an siRNA due to the dsRNA formed by their mRNA transcripts.
  • Gene A is constitutive, gene B is induced by salt stress
    • Normally, just B's transcript is present.
  • When both A and B are present, we get annealing to get dsRNA and this forms an siRNA.
  • Since the siRNA is complementary to gene A's transcript, the siRNA attacks gene A, silencing it.
  • These genes might provide direction to finding new siRNAs.
Annealing of transcripts nicely sets up the dsRNA to be used later in making the siRNA.

Both transcripts present if salt is present.

siRNA silences A.
References


References


References


References