PMFastR: A New Approach to Multiple RNA Structure Alignment

Dan DeBlasio
Jocelyn Braund
Shaojie Zhang

University of Central Florida, Orlando, FL 32816 USA
{deblasio,shzang}@eecs.ucf.edu

University of California, San Diego, La Jolla, CA 93093 USA
jbraund@ucsd.edu
Background
RNA secondary structure is more important to function

- ncRNA secondary structure is very important to function
- Structure is more conserved than sequence
- For example, tRNA (shown) conserves secondary structure between samples more than it conserves sequence
- Alignments of ncRNA should take into account this secondary structure because it is so important

PMFastR: A New Approach to Multiple RNA Structure Alignment
Current methods for multiple alignment

- **ClustalW**
  - Uses profile construction and pairwise alignments to create a sequence based multiple alignment
  - Sequence only multiple alignment

- **RNACAD**
  - Brown. (ISMB 2000)
  - Uses a CM to build a multiple alignment from a seed alignment
  - Used for RDP
  - SCFG based, requires good seed alignment

- **LARA**
  - Bauer, *et al.* (BMC Bioinformatics 2007)
  - Using a graph and an ILP solver, they take structural probability into consideration
  - Does not output structure

PMFastR: A New Approach to Multiple RNA Structure Alignment
Different approaches to RNA alignment

- sequence with structure alignment is an extension of RNA structure prediction
- three major types of alignments
  - structure-structure — both sequences have structure
  - sequence-structure — only one sequence has the associated structure given

PMFastR: A New Approach to Multiple RNA Structure Alignment
Experimentally finding the sequence structure for RNA is expensive.

Assume only one sequence in a family has known structure.

Align all others to infer the structure.
Methods
PMFastR

- **Given:**
  - One sequence with structure
  - Database of sequences without
  - Output multiple alignment with structure
- Align the sequence with structure to one sequence from the database
- This becomes the input to the next alignment
- Align this profile with another sequence from the DB
- Repeat until all sequences have been aligned
- Run `CMBuild –refine` to repair unpaired regions removed in the alignment procedure
Binarized trees encode the RNA sequence and structure

- From Bafna, *et al.* 1995(a), and Zhang, *et al.* 2004(b)
- Each base pair is a solid node
- Parental structure because no pseudoknots
- Dotted nodes indicate bifurcation points or unpaired bases
Alignment is done by traversing the binarized tree

- Each node has its own dynamic programming table
- Each cell represents a segment (two locations) in the target that align to that particular node for solid nodes
- Uses the best calculated results from its children to find its answer
- Trace from the root of the tree to find the alignment
PMFastR: A New Approach to Multiple RNA Structure Alignment
For solid nodes, find the max score of
- Matching bases
- Match left, insert right
- Insert left, match right
- Delete left
- Delete right
- Gap left and right

For unpaired base nodes (dotted with one child), find the max score of
- Match base
- Delete
- Insert left
- Insert right

For Bifurcation nodes (dotted with two children), find the max score of
- For each split of the covered area, sum the score for the two children
PMFastR is doing a global alignment

We can assume that the location of the node $v$ in the target will be in a similar location in the target

Search and store only those locations

There is still a 2-D array for each node

This array reduces from $n^2$ to $\text{band}^2$
Bandwidth reduces running time and memory consumption

- This becomes very important for large sequences such as 16S and 23S rRNA
- Where b is the banding constant, most times set to be < 300
- b needs to be larger than the difference in lengths of the sequences
- Nawrocki and Eddy (PLoS, 2007) are using a similar idea to align 16S rRNA using Covariance Models

<table>
<thead>
<tr>
<th></th>
<th>Without Banding</th>
<th>Banded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Bases</td>
<td>16K</td>
<td>16K</td>
</tr>
<tr>
<td>Space Consumption</td>
<td>O(MN^2)</td>
<td>O(Mb^2)</td>
</tr>
<tr>
<td>(Order)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space Consumption</td>
<td>~3.8 GB</td>
<td>~137 MB</td>
</tr>
<tr>
<td>(Theoretical)</td>
<td></td>
<td>(Assuming b=300)</td>
</tr>
</tbody>
</table>
Issues that arise using banding

- PMFastR does a global alignment
- ncRNA within the same family has a similar length
- The length can be quite large
- As the profile grows the in height (number of sequences) it also expands in length
- For banding to be effective, the sequence lengths need to be similar
Compacting the unpaired regions to increase quality and normalize the length

- Improve quality and reduce the size of the profile by removing columns in unpaired regions
- A predefined quality metric is the percentage of a column that must be present to not be removed
- Present meaning not a blank character
- Columns that encode structure are never removed
Multithreading to increase the wall time

- each node depends on only its child nodes if they exist
- each of the children does not depend on each other
- the child nodes can then be run independently
- once both children have been computed, the parent can be computed
- this is recursive down the tree
Results
Memory Consumption Reduction

- Ran PMFastR and FastR on the same input set
- Cobalamin data
- 300 random sequence pairs
- X-axis: length of input
- Y-axis: memory
- Cubic regressions shown
A carefully constructed set of sequence groups from the Rfam 7 release

- Groups contain 2, 3, 5, 7, 10 or 15 sequences
- Range in APSI from 39 to 50
SCI (Structure Conservation Index)
- Measure of the percentage of bases that conserve structure
- Uses AliFold to produce structure

SCR (Structure Conservation Rate)
- Similar to SCI
- Uses input structure from the alignment

SPS (Sum-of-Pairs Score)
- Comparing with some reference, the ratio of the number of bases that are aligned in both the reference and the test alignment
- Score of 1 means the alignments are exactly the same

Compalign
- Similar to SPS
- Scores the bases not the locations in the sequence
BRAliBase Benchmarking

- LARA
  - Bauer, *et al.* 2007
  - Results to follow from supplemental data
- FoldAlign
  - Torarinsson, *et al.* 2007
- MAFFT
  - Katoh, *et al.* 2005
- STRAL
  - Dalli, *et al.* 2006
- MARNA
  - Siebert and Backofen, 2005
BRAliBase Benchmarking
SCI Results

K2

K3

K5

K7

K10

K15
BRAliBase Benchmarking
SPS Results

K2

K3

K5

K7

K10

K15
BRAliBase Benchmarking
Compalign Results

K2

K3

K5

K7

K10

K15
Reconstructing the Rfam database to show that PMFastR produces high quality alignments

- Downloaded all families from Rfam 8.1
- Assigned the structure to one of the sequences
- Aligned remaining sequences using PMFastR
- SPS and SCR benchmarks shown
Conclusion
Conclusion

- Multiple alignment in a low amount of space using structure information for only one sequence
- Results comparable to hand made alignments
- Publicly available along with detailed results at http://genome.ucf.edu/PMFastR
Future Work

- Remove the Refinement step and work that into each iteration
- Apply the PMFastR algorithm to a database search
- What if the input could be a multiple alignment rather than a single sequence?