Performance Prediction for RNA Design Using Parametric and Non-Parametric Regression Models

Denny C. Dai and Kay C. Wiese

Bioinformatics Research Lab
School of Computing Science, Simon Fraser University
Canada

CIBCB, Mar. 30th 2009

Motivations

Empirical algorithm study

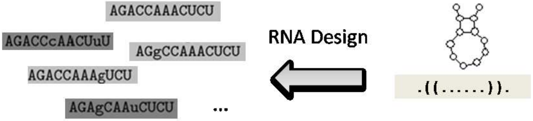
- Parameter tuning
    - Performance varies across parameter settings
- No free lunch theorem
    - Performance varies across problem instances

Therefore, we would like to

- Build a model to predict algorithm performance
    - prediction under different parameter settings
    - prediction on different problem instances
- This leads to robust algorithm design
    - automatic parameter adjustment

Motivations

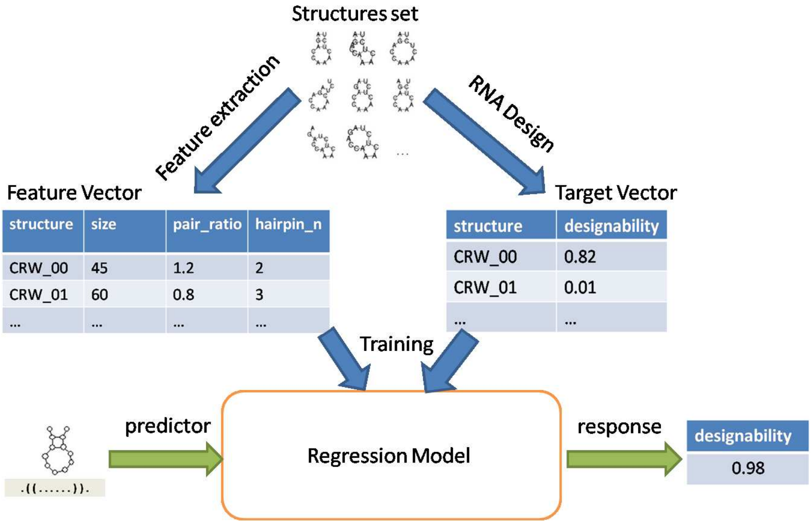
- RNA Design Problem
  - finding RNA primary sequence folded into target shape
  - NP hard
    - heuristic search algorithms



Motivations

- RNA Design Problem
    - finding RNA primary sequence folded into target shape
    - NP hard
        - heuristic search algorithms
- Performance prediction for RNA Design
    - Predict structure designability
    - Why some structures are hard/easy to design
        - correlate structure pattern with designability
        - identify structure components contributing to design difficulty
    - Empirical comparison among RNA design algorithms
        - predict expected algorithm performance on a given structure instance

Method



Structures set

Feature extraction

RNA Design

Feature Vector

structure	size	pair_ratio	hairpin_n
CRW_00	45	1.2	2
CRW_01	60	0.8	3
...	...	...	...

Target Vector

structure	designability
CRW_00	0.82
CRW_01	0.01
...	...

Training

.((......)).

predictor

Regression Model

response

designability
0.98

Regression Models

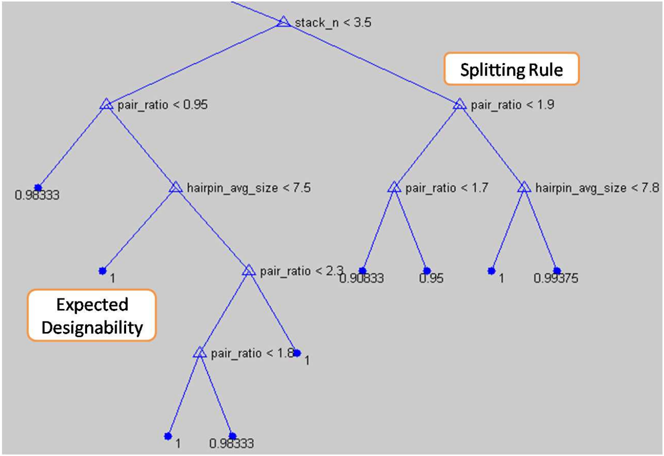
- Ridge regression
  - Linear regression with regularization
  - Build a linear function of input feature vectors
    - Find the optimal combining parameters
    - Minimize prediction error
- Kernel method
  - Using the whole training set for prediction
  - Build a kernel function
  - Prediction by taking a weighted average of the whole training set
    - weight is determined by the structure distance among training points
    - then scaled by kernel bandwidth h

Regression Models

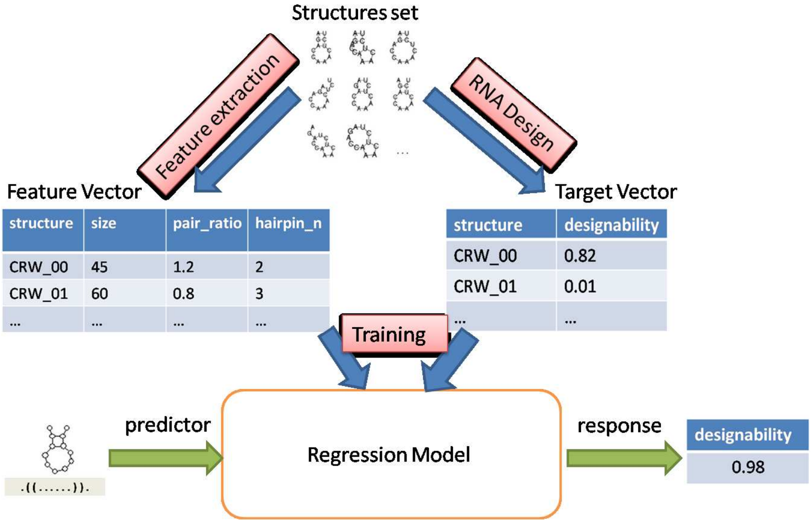
- Classification & Regression Tree (CART)
  - Non-parametric regression
    - avoid explicit functional assumption among feature vectors
    - reduce model complexity and training cost
  - Tree-structure classifier
    - Binary regression tree
    - Splitting rules (on feature vectors) at each internal node
    - Cluster of training points at leaf node

Regression Models

- Classification & Regression Tree (CART)



Experiment - Review



structure	size	pair_ratio	hairpin_n
CRW_00	45	1.2	2
CRW_01	60	0.8	3
...	...	...	...

structure	designability
CRW_00	0.82
CRW_01	0.01
...	...

designability
0.98

Experiment - Model Training

- Combination of input features
  - 15 structure features in total
  - 3 training sets including two biological sets and one random structure set
  - 10-fold cross validation for each feature combinations on each model
- Prediction accuracy
  - root mean square error (RMSE)
  - correlation coefficient (CC)

Empirical Results

- Prediction accuracy does not grow monotonically with feature numbers
  - Beneficial features
    - *hairpin_avg_size*, *pair_ratio*
  - Misleading features
    - *bulge_avg_size*, *stack_n*

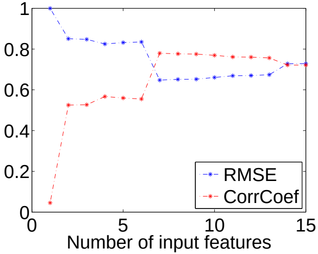


Figure: SyntheicRNA & Ridge Regression

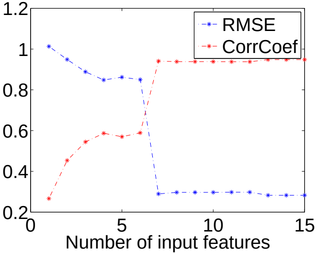


Figure: SyntheticRNA & CART Regression

Empirical Results

- Prediction accuracy does not grow monotonically with feature numbers
  - Beneficial features
    - *hairpin_avg_size*, *pair_ratio*
  - Misleading features
    - *bulge_avg_size*, *stack_n*

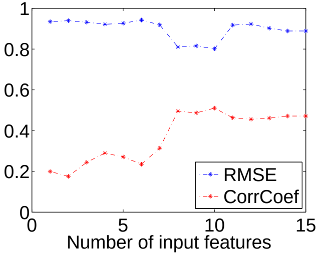


Figure: TransferRNA & Ridge Regression

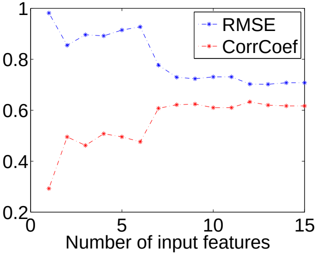
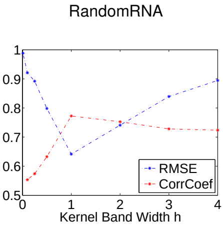
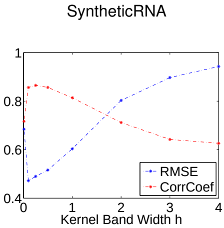


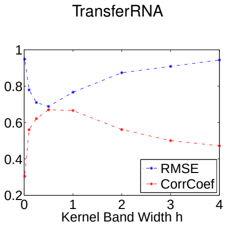
Figure: TransferRNA & CART Regression

Empirical Results

- Optimal kernel bandwidth is found at smaller values on biological set comparing to random training set.
  - kernel bandwidth scales the structure distance measurement
  - it controls how many training points are used towards the prediction on the given structure
  - smaller optimal kernel bandwidth means:
    - higher degree of structure similarity among biological structure set
    - smaller kernel scope is sufficient to achieve high prediction accuracy

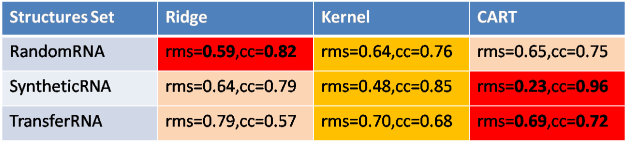






Empirical Results

- CART model achieves the highest prediction accuracy on biological data sets [1]
  - effective clustering of training data points
  - the regression tree captures the correlation between structure patterns and designability

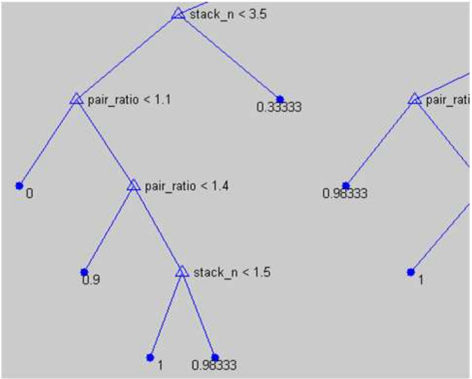


Structures Set	Ridge	Kernel	CART
RandomRNA	rms=**0.59**,cc=**0.82**	rms=0.64,cc=0.76	rms=0.65,cc=0.75
SyntheticRNA	rms=0.64,cc=0.79	rms=0.48,cc=0.85	rms=**0.23**,cc=**0.96**
TransferRNA	rms=0.79,cc=0.57	rms=0.70,cc=0.68	rms=**0.69**,cc=**0.72**

[1] *rms* is the root mean square error, *cc* is correlation coefficient

Empirical Results

- CART model achieves the highest prediction accuracy on biological data sets
  - effective clustering of training data points
  - the regression tree captures the correlation between structure patterns and designability



Conclusion and Discussion

Algorithm performance prediction for RNA Design

- Non-parametric models (kernel, CART) outperforms parametric (ridge) method on biological data set
    - 13% to 30% increase in prediction accuracy
- Biological data set has higher degree of structure similarity
    - optimal kernel bandwidth differs by one order of magnitude
- input features affect prediction accuracy
    - Beneficial & misleading features
    - greedy feature selection improves prediction by 12% to 18%

In our future work,

- extend the performance benchmarks
    - designability
    - runtime cost (local search steps, CPU seconds)
- integrate prediction model in RNA design algorithm
    - expected performance (per problem-instance based)
    - algorithm parameter self-adjustment

Question

- Thank you!