Sequence to Shape

Predicting non-Coding RNA Structures

Jim Beck

Boise State University

May 1, 2020

Predicting ncRNA shape from sequence is scientifically important

NSF grand challenge - Genotype → Phenotype

1. Understand how organisms develop, interact, and adapt
2. ncRNA play critical roles in these processes Alberts et al. 2013
   - Roles: regulatory, structural, catalytic, ...
   - Diseases: cancer, Alzheimer's, COVID-19, ...
3. Predicting shape is a step towards predicting function
4. Connecting sequence to function → biological engineering

Empirical challenges make computational methods appealing

1. X-ray crystallography is inefficient and expensive
2. Growing inventory of ncRNA sequences

Fundamentals of the genetic code Alberts et al. 2013

Definitions

- Nucleotides: Guanine, Cytosine, Adenine, Thymine (Uracil)
- Canonical Base Pairs (Watson-Crick): G··C, A··T(U)

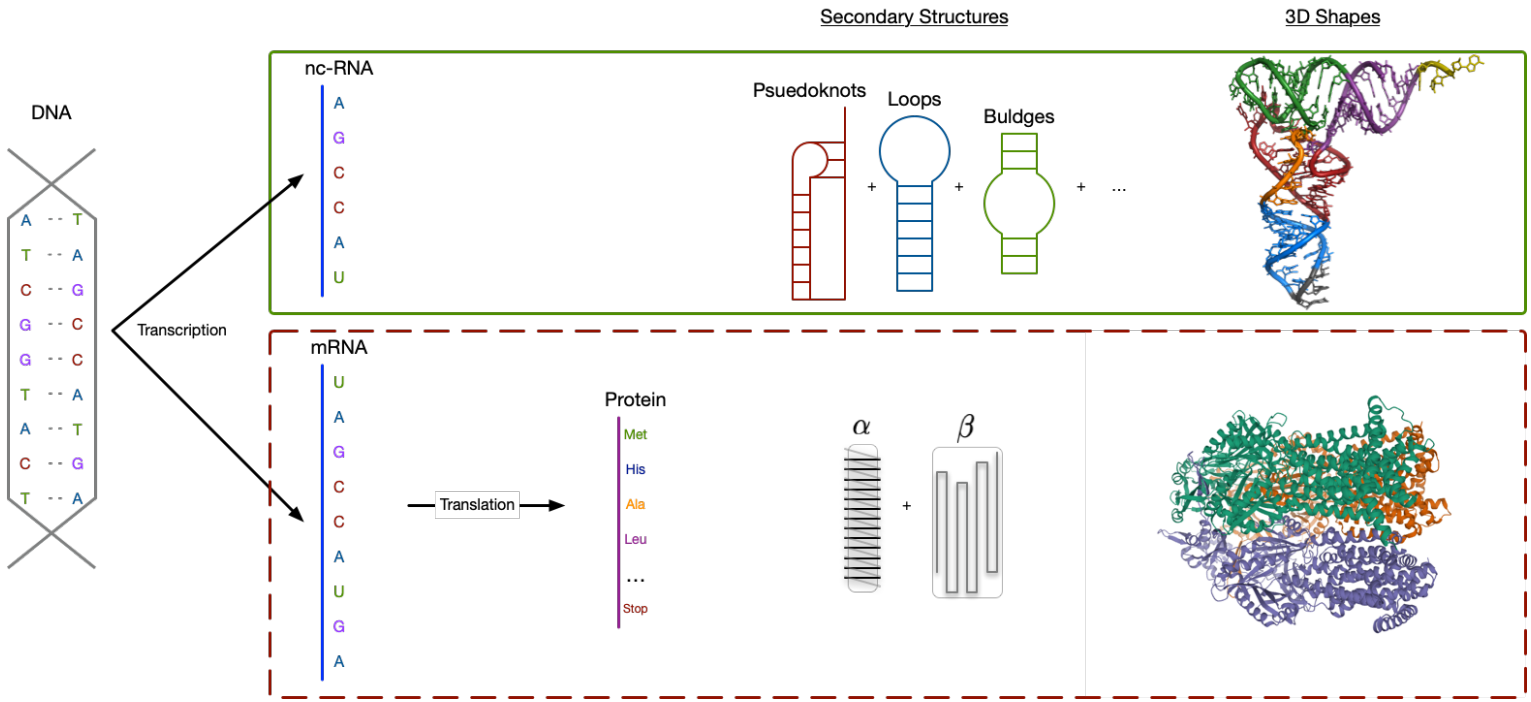
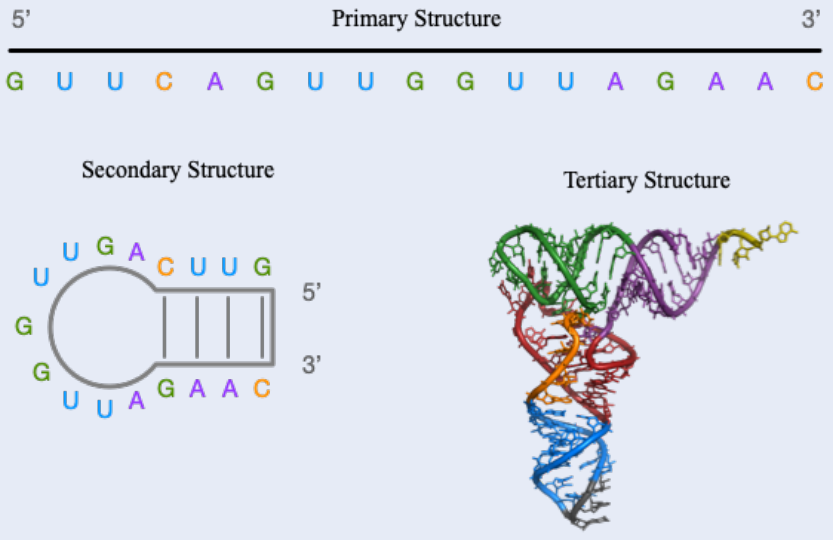


Figure: ncRNA from http://cen.xraycrystals.org/transfer-rna.html, protein from https://www.rcsb.org/3d-view/5T0O/1

Sequence drives shape which dictates function Schuster et al. 1994

Primary → Secondary → Tertiary



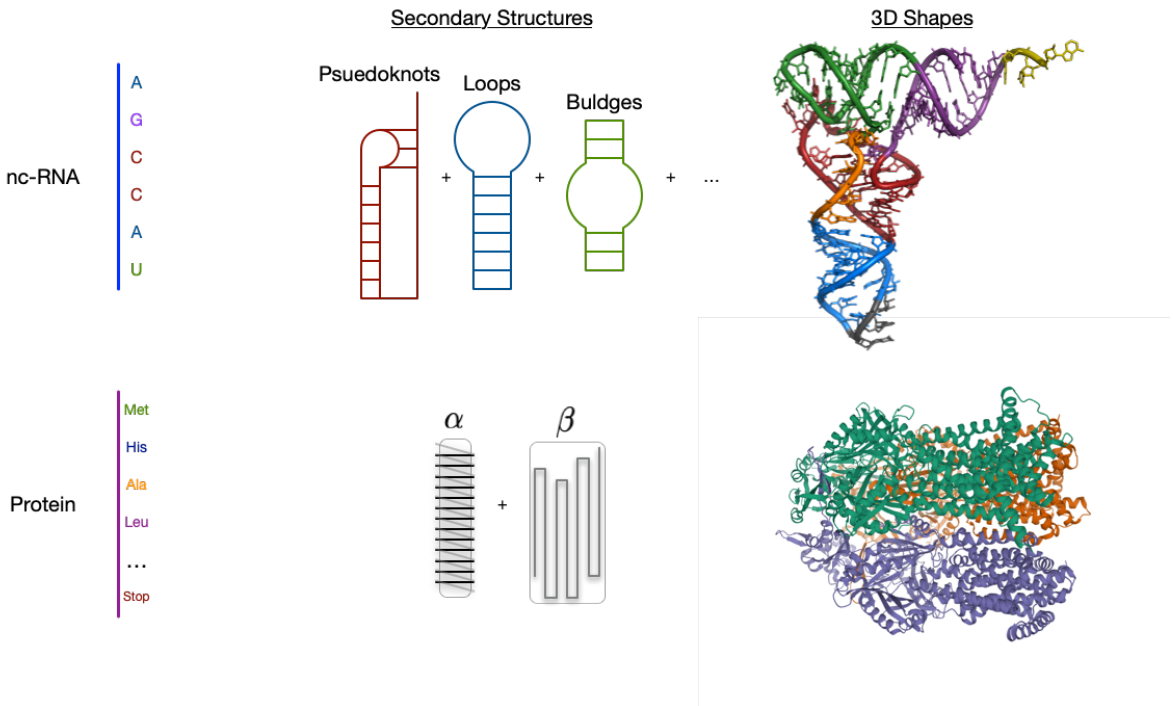
The relationship is analytically important

1. Isolated sequence mutations impact function
2. Many sequence mutations change/destroy function
3. Sequence change is faster than structural change
4. Different nucleotides can make the same secondary structure

ncRNA similarities with proteins motivate a computational method

Similarities within ncRNA and Proteins Rother et al. 2011

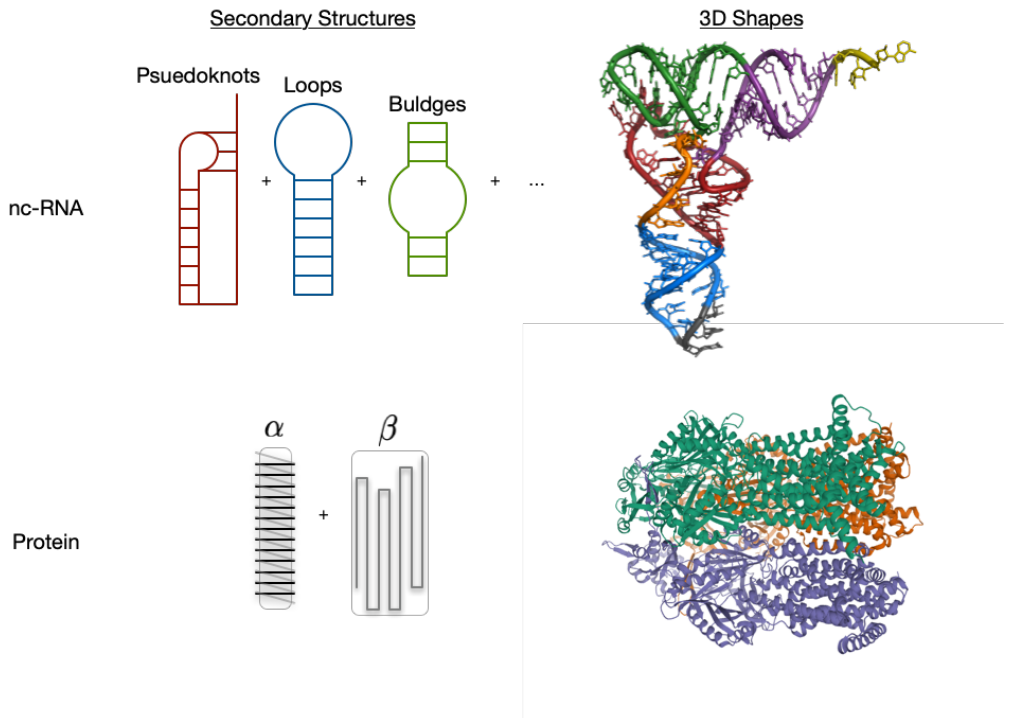
- Continuous molecule: backbone and side chains
- Chains held together covalently
- Secondary structures stabilized by Hydrogen bonds
- Spontaneously contorts into 3D configuration



ncRNA differences make prediction less accurate than in proteins

Differences between ncRNA and proteins Rother et al. 2011

- Each use different mechanisms to form structure
- ncRNA's have many more secondary structures
- ncRNA's compact more tightly



ncRNA prediction rely on methods used to predict protein shapes Rother et al. 2011

All methods apply some form of a two step process

1. Accumulate known secondary structures (building blocks)
2. Evaluate for lowest energy configuration

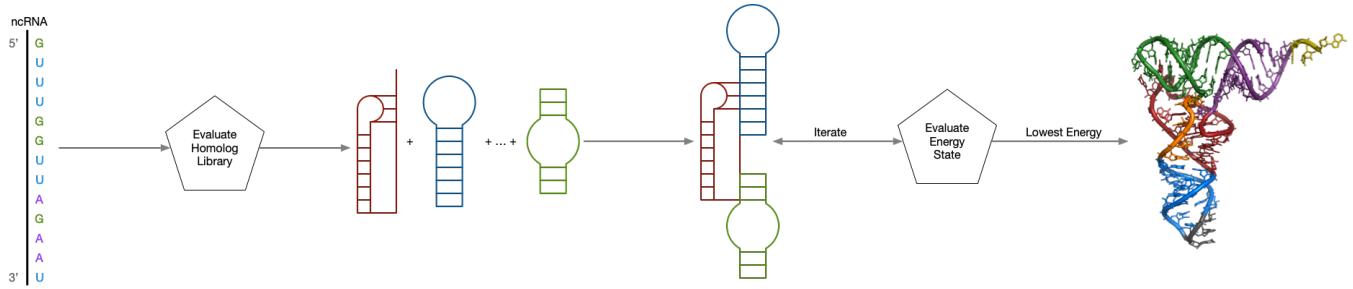


Figure: ncRNA Prediction Process. Homolog - related by descent from a common ancestral DNA sequence

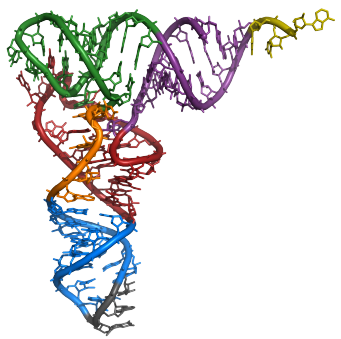
Caution

1. Many similarly low energy state configurations
2. Combinatorially complex

Combinatorial complexity is an impediment to most predictions

Causes of complexity

1. 4^L nucleotide combinations
2. $3N_{atoms} - 5$ Degrees of Freedom



Responses to complexity

1. Homolog libraries
2. Pseudo-atom
3. Limit range of degrees of freedom

No definitive measure of predictive accuracy exists Miao et al. 2017

Root Mean Square Deviation - most prominent

1. Difference between model and known crystal structure in Å
2. Global measurement of accuracy

Problem 3

Group[a]	Number[b]	RMSD[c]	Rank[d]	DI all[e]	Rank[d]	INF all[f]	Rank[d]	INF wc[g]	Rank[d]	INF nwc[h]	Rank[d]	INF stack[i]	Rank[d]	Clash Score[j]	Rank[d]	*P*-value[k]	Rank[d]
Chen	1	7.24	1	9.84	1	0.74	2	0.86	5	0	6	0.73	1	1.1	3	2.01E-05	1
Dokholyan	2	11.46	2	16.1	2	0.71	6	0.82	9	0	9	0.71	6	41.21	10	3.90E-02	2
Das	5	11.97	3	16.42	3	0.73	5	0.9	1	0.36	5	0.71	3	1.1	4	6.92E-02	3
Bujnicki	1	12.19	4	17.49	5	0.7	7	0.82	10	0	10	0.7	7	14.72	8	8.71E-02	4
Das	2	12.2	5	16.6	4	0.74	3	0.86	6	0.4	2	0.73	2	0.74	2	8.83E-02	5
Major	2	13.7	6	23.33	10	0.59	11	0.67	11	0	8	0.61	10	93.52	12	3.03E-01	6
Bujnicki	2	14.06	7	22.51	7	0.62	10	0.83	8	0	7	0.59	11	5.15	7	3.75E-01	7
Das	1	15.48	8	20.9	6	0.74	1	0.87	4	0.57	1	0.71	5	0	1	6.81E-01	8
Dokholyan	1	15.92	9	23.28	9	0.68	9	0.9	2	0	12	0.66	9	39.37	9	7.629E-01	9
Das	3	16.95	10	23.17	8	0.73	4	0.89	3	0.4	3	0.71	4	1.47	5	9.02E-01	10
Das	4	18.3	11	26.55	11	0.69	8	0.85	7	0.38	4	0.67	8	2.21	6	9.79E-01	11
Major	1	22.99	12	45.27	12	0.51	12	0.39	12	0	11	0.59	12	75.11	11	1.00E+00	12
Mean		14.37		21.79		0.68		0.80		0.18		0.68					
Standard deviation		3.99		8.69		0.07		0.14		0.22		0.05					
												X-Ray Model		1.83			

Figure: Miao et al. 2017

No single winner for all accuracy measures

No measure of function is employed to assess model accuracy.

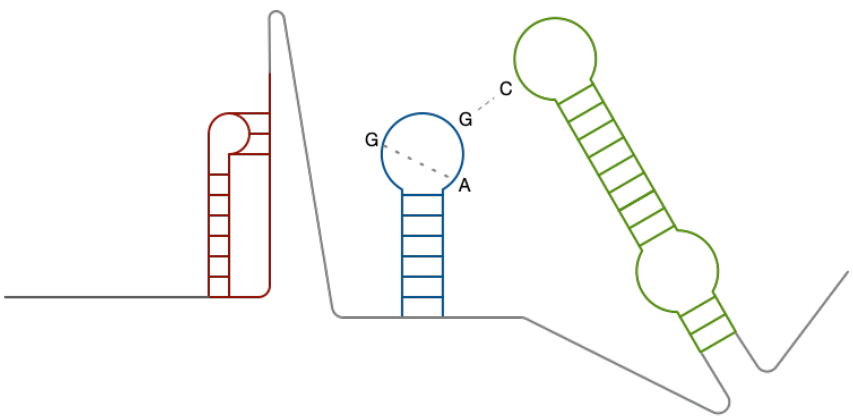
Despite some success, significant barriers remain Miao et al. 2017

RNA-Puzzles

A blind experiment in RNA 3D structure prediction - comparing various labs predictive models against known crystal structures.

Problem Areas

1. Homolog availability affects model accuracy
2. Long sequences affects accuracy
3. Non-Canonical pairings
4. Pairings at a distance



For even small sequences, complexity is a problem Watkins et al. 2018

Non-Canonical are Iteratively Solved

1. Complete atomic configuration is borderline intractable
2. Stepwise Monte Carlo is better (2 orders) but not acceptable.
3. Threshold is Metropolis Criterion
4. Å performance does not mean lowest energy state

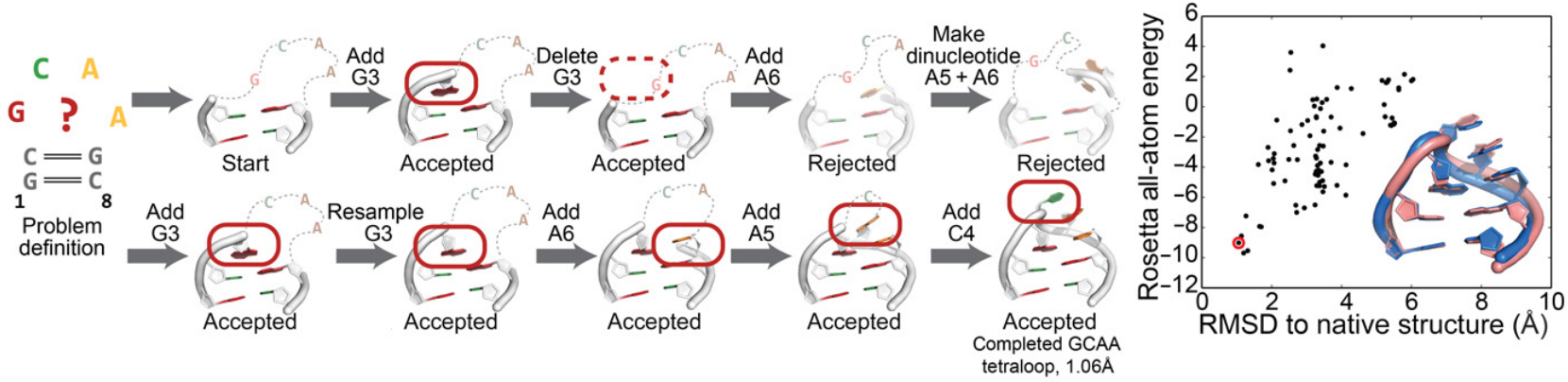


Figure: Watkins et al. 2018

Finding distance pairings is very difficult Weinreb et al. 2016

Identification of prospective pairs

- Look for co-variation to explain conserved shapes
- Deal with transitivity ($A \implies C$ & $B \implies C$, $A \not\Rightarrow B$)
- Recognize co-variation as a network of interactions.

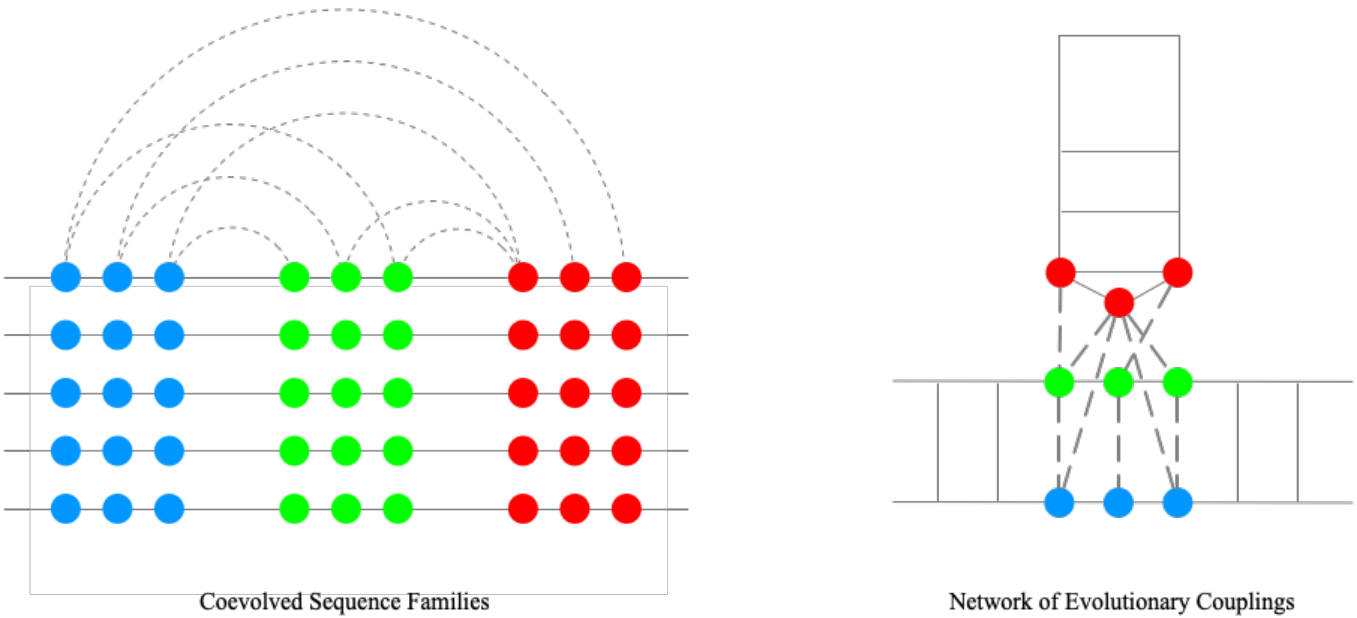


Figure: Weinreb et al. 2016

Options for improving prediction accuracy

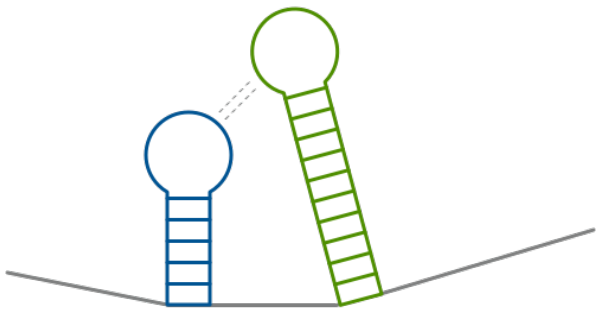
Three areas for improvement

1. Expand the utility of homolog libraries
2. Identify networks of connections and structure within data
3. Include other experimental data

Improve the utility of homolog template libraries Laing and Schlick 2010

Features of higher order structures

1. Structures occur together
2. Possess collaborative function



Benefits to improved homolog libraries

1. Typically improved accuracy
2. Reduce computational complexity
3. Highlight functionally important sub-sequences

Evaluate structure as a network Laing and Schlick 2010

Graph representations are common in structural models

1. Useful in comparing RNA structures for isomorphisms
2. Reduces the problem size

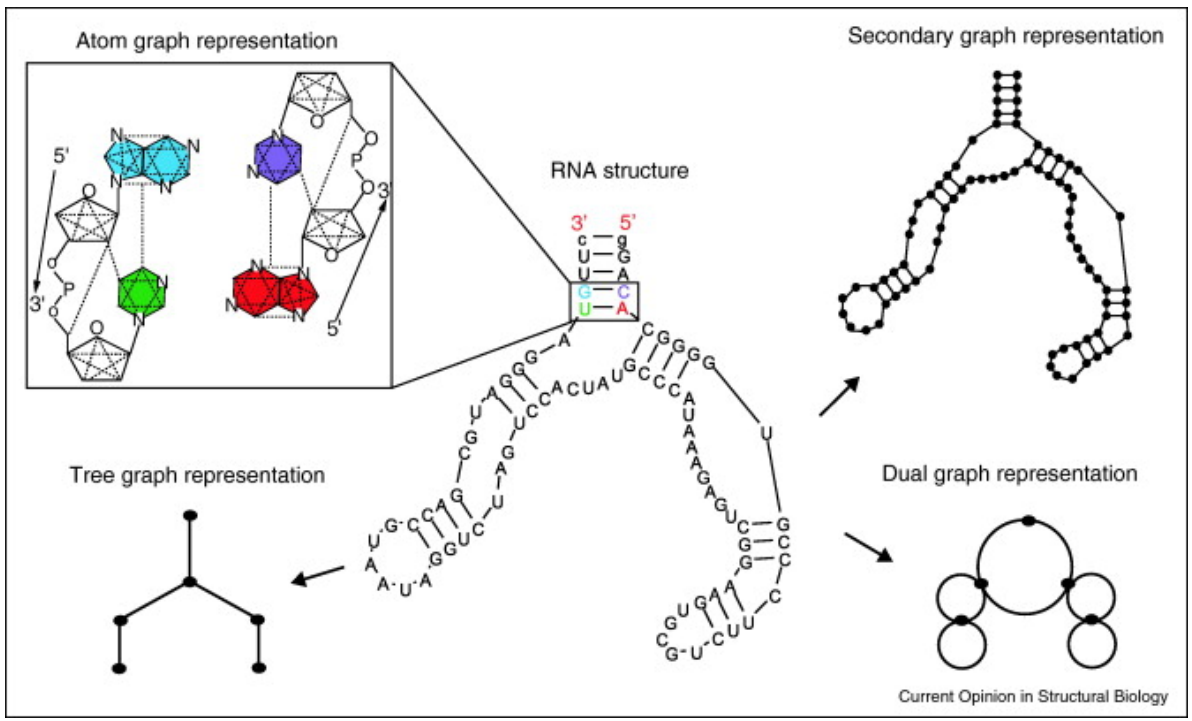
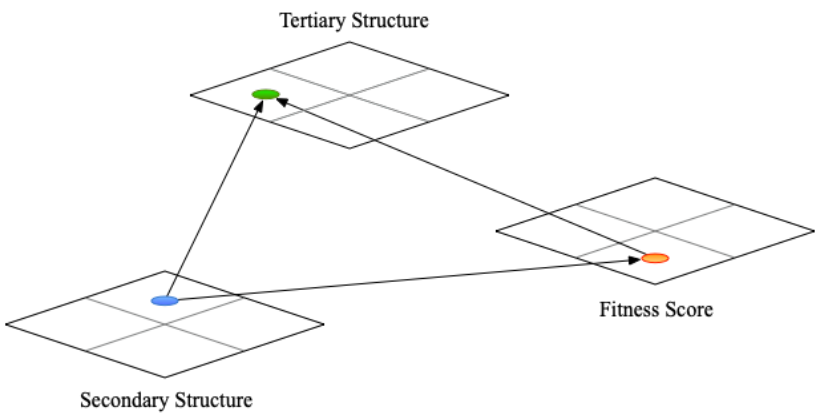


Figure: Laing and Schlick 2010

Higher-order structure in network Benson, Gleich, and Leskovec 2016

Structures are common between different data types

1. Small network subgraphs as network motifs
2. Important relationships may be revealed using different motifs
3. Expand this motif assessment across data of different type and granularity



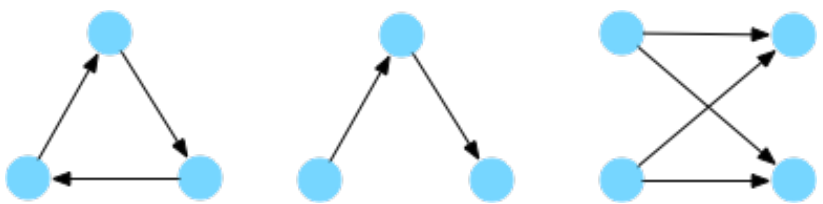


Figure: Benson, Gleich, and Leskovec 2016

Include other information to improve predictions Nichol et al. 2019

Fitness landscapes contain important information about evolutionarily advantaged sequences

1. Proximate sequence mutations possessing a fitness gradient.
2. Fitness is a measure of functional performance

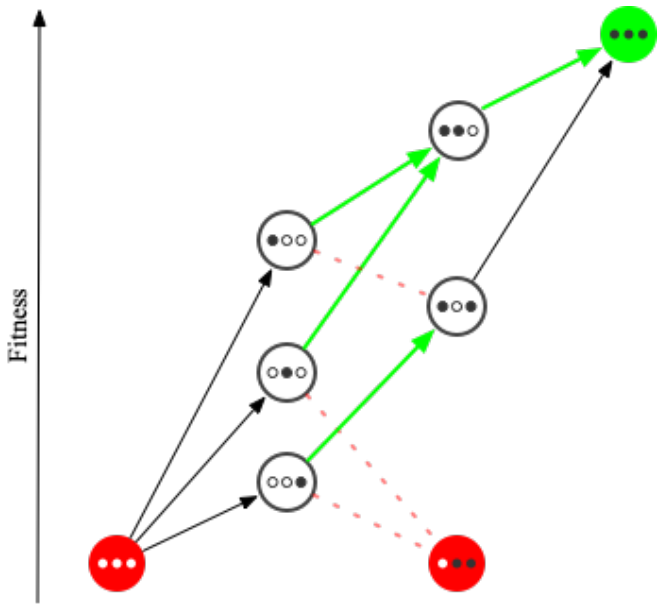


Figure: Poelwijk et al. 2007

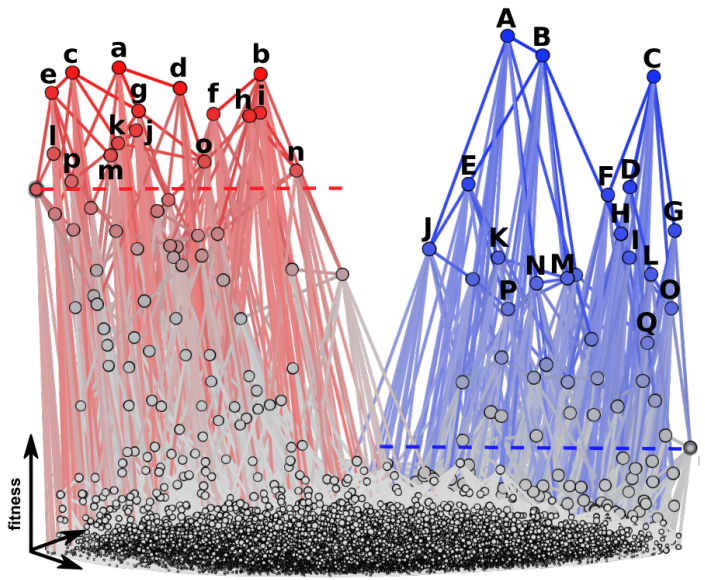


Figure: Bendixsen et al. 2019

Survey a fitness landscape for available mutational pathways

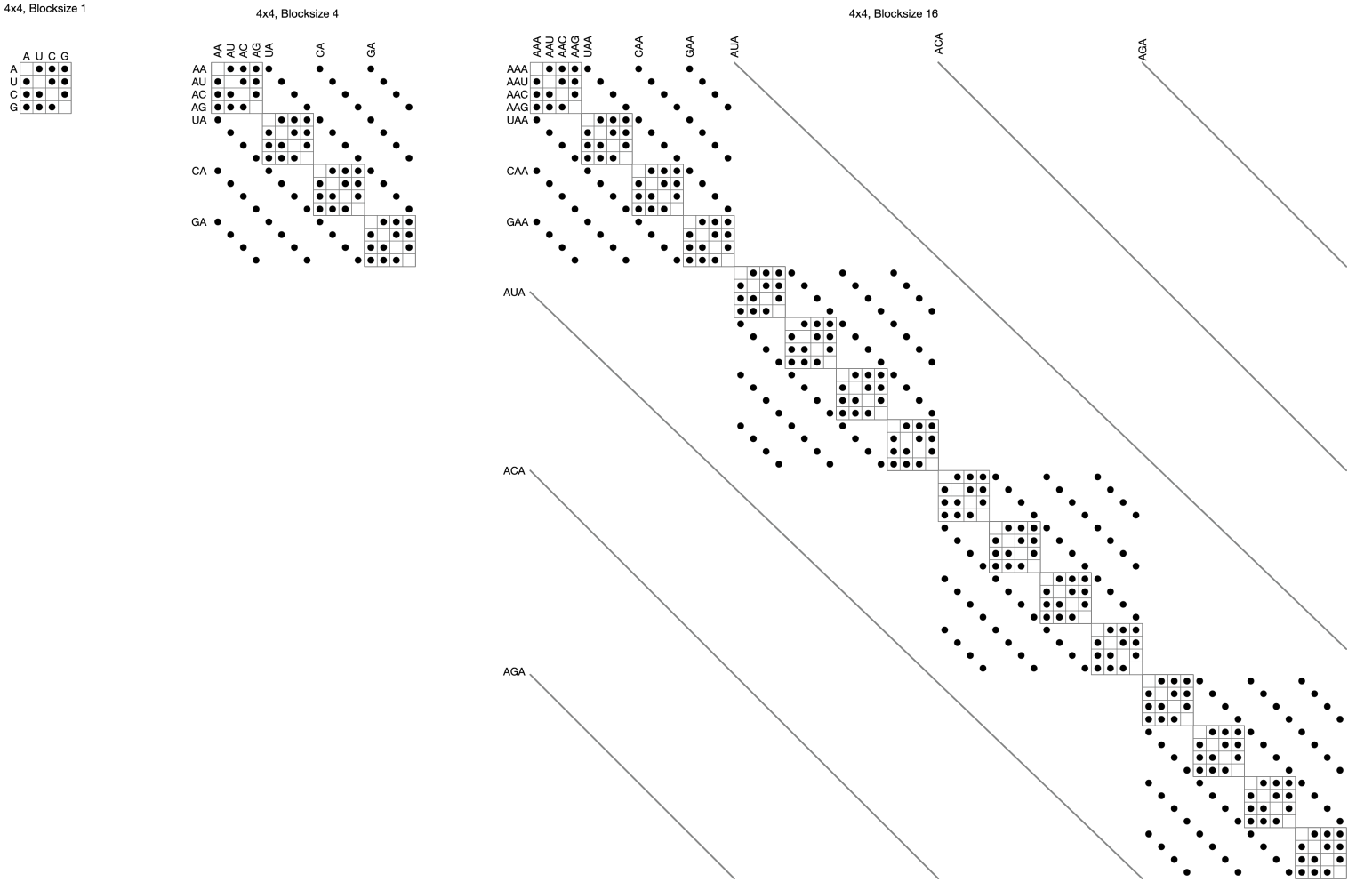
Goals for Coding Artifact

1. Randomly survey fitness landscape for features
2. Identify important nodes (sequence mutations)
3. Find important paths (connections between nodes)

Future Goals

1. Refine "important"
2. Relate fitness to shape
3. Create a relationship network
4. Make an efficient Julia package

Curious Observation: Adjacency Matrix is a Kronecker Graph



Conclusion

Things we've covered

1. Predicting ncRNA is important
2. Prediction is challenging
3. Opportunities exist to improve
4. Code to survey landscapes for features

References


Alberts, Bruce et al. (2013). *Essential Cell Biology*. Garland Science.


Bendixsen, Devin P et al. (2019). "Genotype network intersections promote evolutionary innovation". In: *PLoS biology* 17.5, e3000300.


Benson, Austin R, David F Gleich, and Jure Leskovec (2016). "Higher-order organization of complex networks". In: *Science* 353.6295, pp. 163–166.


Laing, Christian and Tamar Schlick (2010). "Computational approaches to 3D modeling of RNA". In: *Journal of Physics: Condensed Matter* 22.28, p. 283101.


Miao, Zhichao et al. (2017). "RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme". In: *Rna* 23.5, pp. 655–672.


Nichol, Daniel et al. (2019). "Model genotype–phenotype mappings and the algorithmic structure of evolution". In: *Journal of the Royal Society Interface* 16.160, p. 20190332.


Poelwijk, Frank J et al. (2007). "Empirical fitness landscapes reveal accessible evolutionary paths". In: *Nature* 445.7126, pp. 383–386.


Rother, Kristian et al. (2011). "RNA and protein 3D structure modeling: similarities and differences". In: *Journal of molecular modeling* 17.9, pp. 2325–2336.


Schuster, Peter et al. (1994). "From sequences to shapes and back: a case study in RNA secondary structures". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255.1344, pp. 279–284.


Watkins, Andrew M et al. (2018). "Blind prediction of noncanonical RNA structure at atomic accuracy". In: *Science advances* 4.5, eaar5316.


Weinreb, Caleb et al. (2016). "3D RNA and functional interactions from evolutionary couplings". In: *Cell* 165.4, pp. 963–975.