Using multiplayer online videogames for structure prediction and design

Meera Srinivasan
Jesse Min
Seth Hildick-Smith

Table of Contents

1. **Protein Structure Prediction**

2. **RNA Design**

3. **Multiplayer online collaborative platforms**

Predicting protein structures with a multiplayer online game

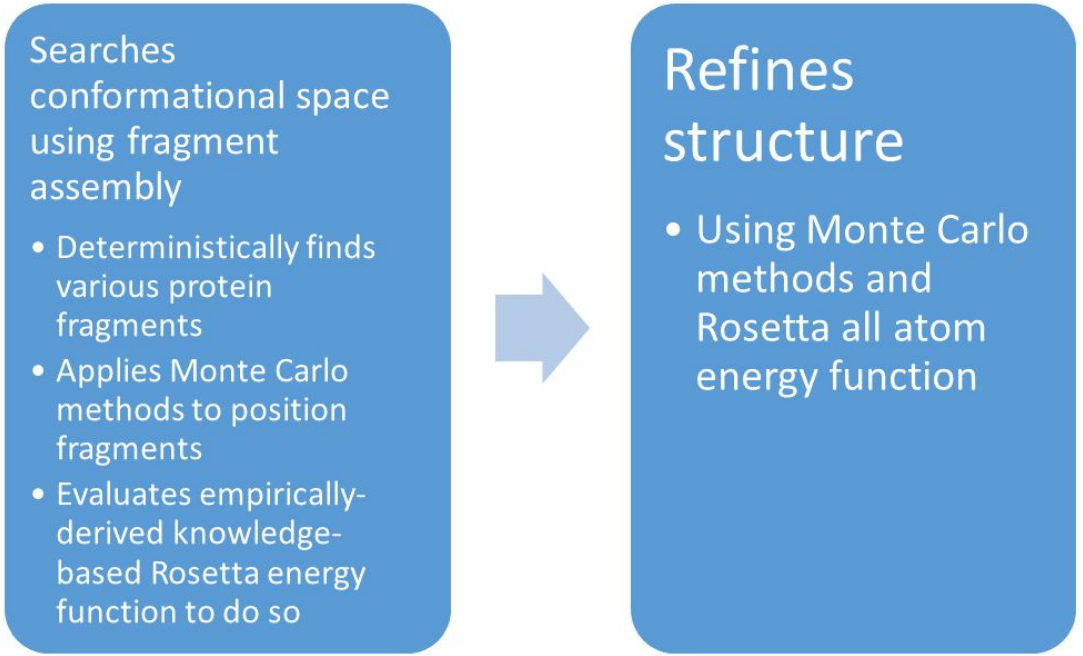
Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic & Foldit players

Presented by Meera Srinivasan

Protein Structure Prediction: How do you predict the structure of a protein from its amino acid sequence?

1. **Homology** or **template-based modeling**: given a particular protein sequence, search for a homolog
2. ***Ab initio* structure prediction**: when no homologs, combine deterministic & stochastic approaches to predict structure

Introduction: Rosetta for Protein Structure Prediction



This cycle is repeated many times.

Introduction: Rosetta's Components

- Deterministic: fragment finding and assembly
- Stochastic: energy minimization; structure perturbation & refinement

Challenges with Rosetta

- Free energy landscape is large
- Suboptimal conformations pursued in effort to minimize energy

Experimental Hypothesis:

What if we replaced stochastic components with human decision making and retained deterministic components?

FoldIt: Game Development for Protein Structure Prediction

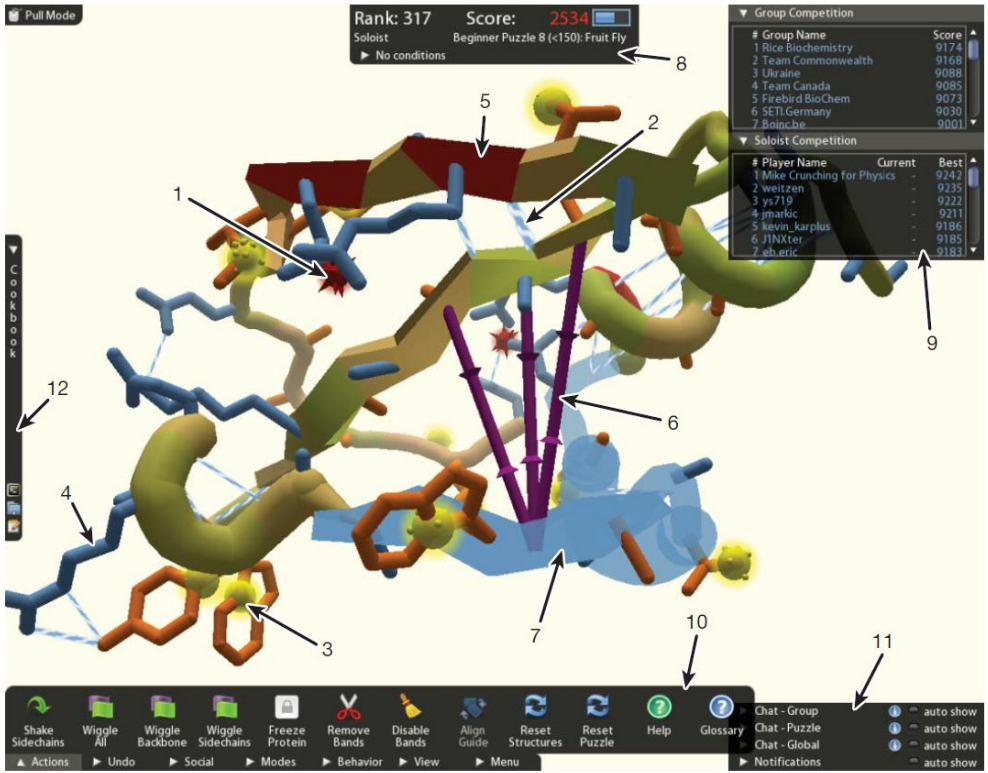


Image: (Seth et. al.

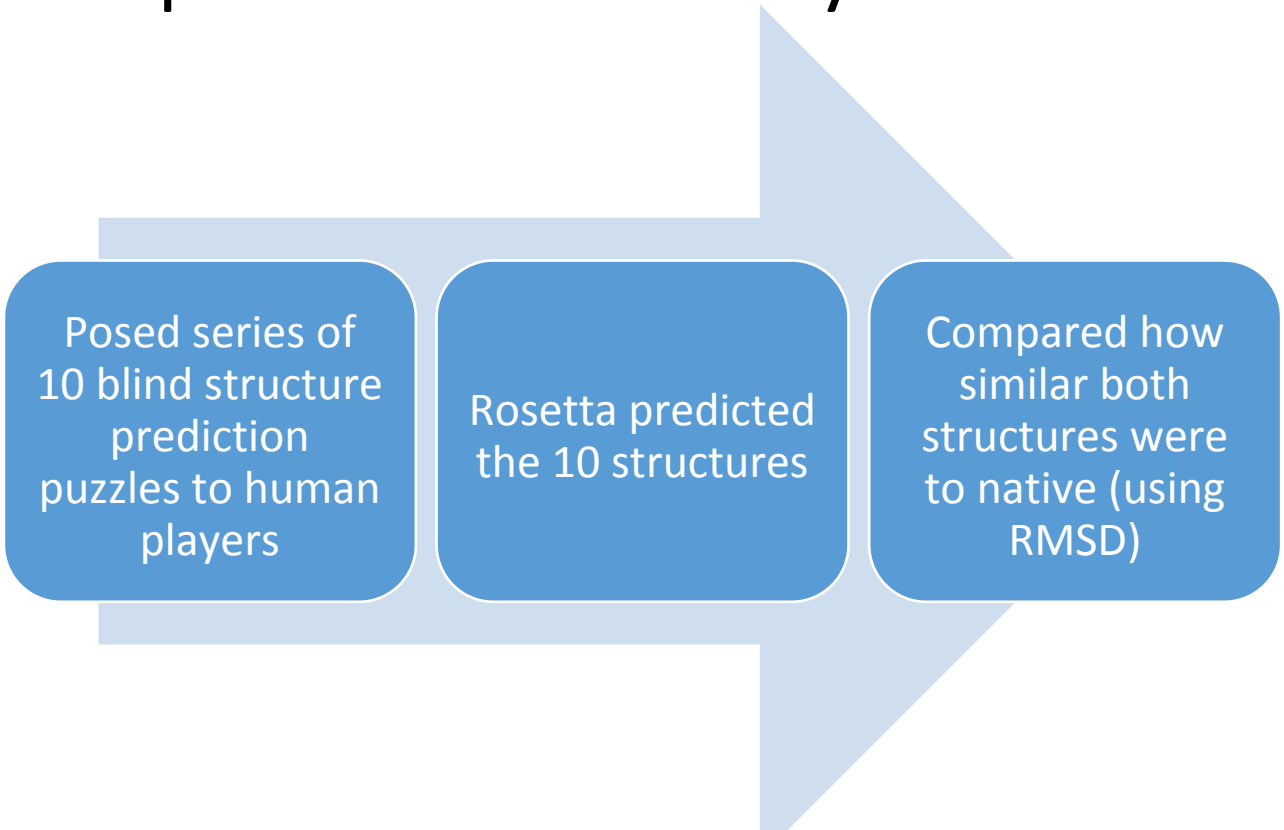
FoldIt: Game Development for Protein Structure Prediction

- User tools reflect deterministic Rosetta algorithms
  - Meant to minimize local energy
  - Scoring: computes energy same way Rosetta does, but multiplies by 100; goal is to get higher score.
- Player intentions parallel stochastic processes
  - Searches conformational space instead of computer doing so
- Meant to be approachable for players from all backgrounds

FoldIt: Game Development for Protein Structure Prediction

- https://www.youtube.com/watch?v=lGYJyur4FUA
- 1.40-1.55 important

Experimental Methods: Comparing human structure prediction accuracy to Rosetta's



Experimental Methods: RMSD (Root mean squared deviation)

- Measure of similarity between two structures for a given protein
- Must align the two structures being compared before measuring RMSD

Results: Human intuition is very effective

Table 1 | Blind data set

Puzzle ID	Foldit Cα r.m.s.d.	Rebuild and refine Cα r.m.s.d.	Native	Method	Number of residues	Figure(s)
986875	**1.4**	4.5	2kpo	NMR	99	3a–c, Supplementary 4
986698	**1.8**	3.7	2kky	NMR	102	3d, e
986836	**5.7**	6.6	3epu	X-ray	136	2c, Supplementary 6d
987088	**3.5**	4.3	2kpt	NMR	116	2a, b, Supplementary 6a, b
987162	**4.5**	5.2	3lur	X-ray	158	Supplementary 6c
987076	3.3	3.5	2kpm	NMR	81	2e, Supplementary 5c
986629	3.5	3.3	2kk1	NMR	135	Supplementary 5b
987145	2.6	2.3	3nuf	X-ray	105	2d, Supplementary 5a
986844	6.9	**5.8**	2ki0	NMR	36	Supplementary 10a
986961	10.6	**5.7**	2knr	NMR	118	Supplementary 10b

A listing of all the Foldit puzzles run in the blind data set. A Cα r.m.s.d. comparison to the native structure is given between the best-scoring model produced by Foldit players and the best-scoring model produced by the Rosetta rebuild and refine protocol, given the same starting model(s). Solutions considerably better with one method than the other are indicated in bold. The solved structures (which were released after each puzzle ended) are represented by their Protein Data Bank (PDB) codes. Results from these Foldit puzzles can be accessed on the Foldit website by replacing ID with the corresponding Foldit puzzle ID in http://fold.it/portal/node/ID. 2kky, 2kpt, 2kpm, 2kk1 and 2knr were taken from the CASD-NMR experiment[10]. 2kpo was provided by N. Koga and R. Koga. 2ki0 and 3epu were found by searching for unreleased structures on the PDB website (http://www.rcsb.org/pdb/search/searchStatus.do). 3lur and 3nuf were provided by the Joint Center for Structural Genomics (JCSG). The location of figures containing results for each puzzle are provided in the last column.

Image: (Seth et. al. 2010)

Results: Scenarios in which humans performed better

- Cases with substantial backbone remodeling so exposed hydrophobic residues faced inwards

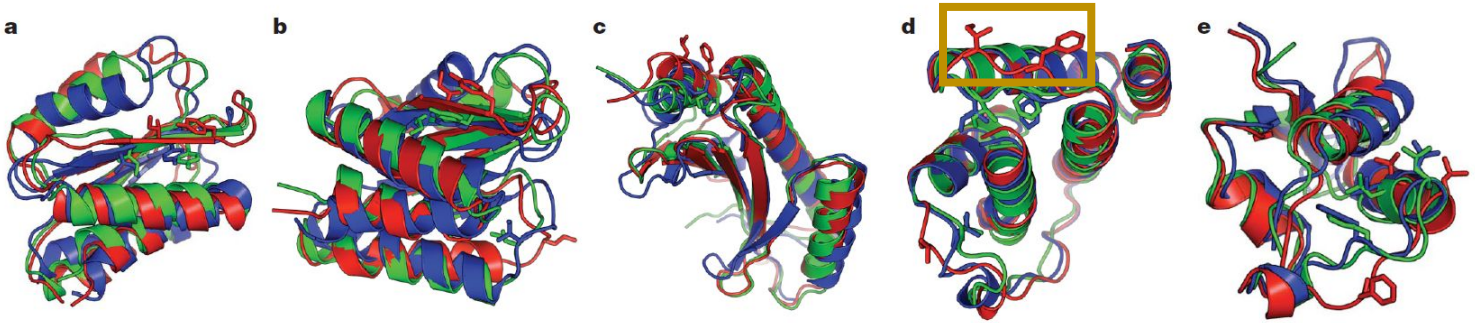


Figure 2 | Structure prediction problems solved by Foldit players. Examples of blind structure prediction problems in which players were successfully able to improve structures. Native structures are shown in blue, starting puzzles in red, and top-scoring Foldit predictions in green. **a**, The red starting puzzle had a register shift and the top-scoring green Foldit prediction correctly flips and slides the β-strand. **b**, On the same structure as above, Foldit players correctly buried an exposed isoleucine residue in the loop on the bottom right by remodelling the loop backbone. **c**, The top-scoring Foldit prediction correctly rotated an entire helix that was misplaced in the starting puzzle. **d**, The starting puzzle had an exposed isoleucine and phenylalanine on the top, as well as an exposed valine on the bottom left. The top-scoring Foldit prediction was able to correctly bury these exposed hydrophobic residues. **e**, Another successful Foldit helix rotation along with a remodelled loop that correctly buries an exposed phenylalanine. Images were produced using PyMOL software[11].

Image: (Seth et. al.

Results: Humans more likely to “take risks” sampling than Rosetta

- Rosetta’s stochastic Monte Carlo methods aim to **minimize energy** while humans are more likely to **create high energy intermediates**, knowing they will lead to **lower energy structures**.
- Humans can distinguish which starting point is most useful.
- Human players restructure significantly to improve hydrophobic burial and hydrogen bond quality.

Foldit's social & psychological dimensions

- Complexity, creativity, & collaboration in search process
- Inter-group competition
- Retaining engagement & "thrill" factor
- Motivation & rewards structure

Foldit's social & psychological dimensions

GROUP	PUZZLE	SCORE
Anthropic Dreams	1329: Unsolved ... 96	9,240
BIOL 4030 Winter 2017	Beginner Puzzle...ign	8,877
Beta Folders	1328: 85 Residu...int	10,621
Contenders	1327: Revisitin...one	9,327
Anthropic Dreams	Beginner Puzzle...nis	9,254
Kotocycle	Beginner Puzzle...yle	8,388
Gargleblasters	Beginner Puzzle...ign	10,942
Kotocycle	Beginner Puzzle...ity	14,439
Kotocycle	Beginner Puzzle...zle	8,582
		FULL

SOLOISTS | EVOLVERS | GROUPS | TOPICS

PLAYER	PUZZLE	SCORE
Bruno Kestemont 11 12	1329: Unsolved ... 96	9,239
AeonFluff 100 316	Beginner Puzzle...ign	8,879
bertro 23 7	1328: 85 Residu...int	10,620
gitwut 10 4	1327: Revisitin...one	9,327
jfryk 100 474	Beginner Puzzle...nis	9,254
d_chasy 100 3100	Beginner Puzzle...yle	8,598
cinnamonkitty 100 251	Beginner Puzzle...ign	11,068
cinnamonkitty 100 251	Beginner Puzzle...ity	14,530
florashaman 100 114	Beginner Puzzle...zle	8,680
		FULL

https://fold.it/portal/info/faq

Limitations of Foldit

- Players are reliant on FoldIt's user tools (e.g. visual cues, available moves)
  - Had trouble folding an extended protein chain
  - Tools may limit breadth of intuition
- Maintaining user engagement & sustaining interest

Possible Future Directions

- Psychological studies of player motives and behavior to improve existing PSP algorithms
- Applications to drug discovery-related research areas

Introduction to RNA Design and EteRNA

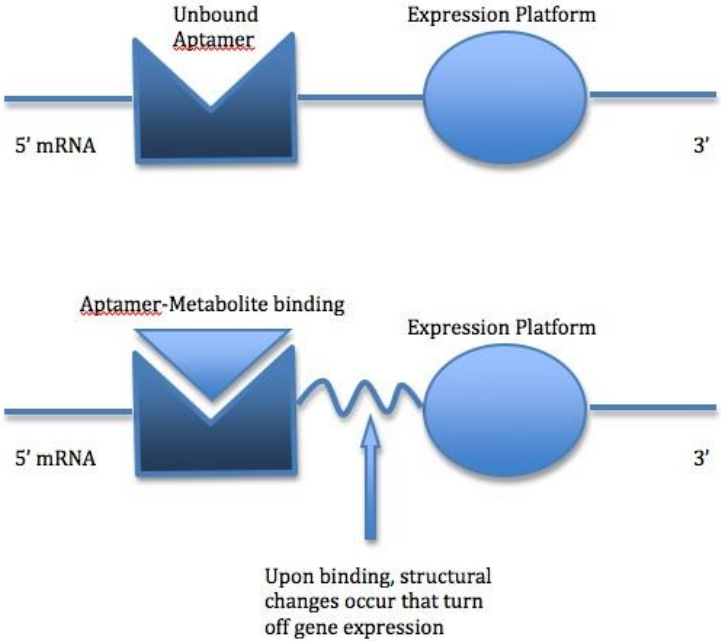
Table of Contents

1. **RNA Design**
    a. Importance of RNA Structure
    b. RNA Design: Computational Determination Methods

2. **EteRNA**

Importance of RNA Structure (1)

Riboswitch & “miRNA alterations are related to cancer” (Calin et al., 2006)



Wikibooks - Riboswitch

MicroRNA signatures in human cancers

George A. Calin and Carlo M. Croce

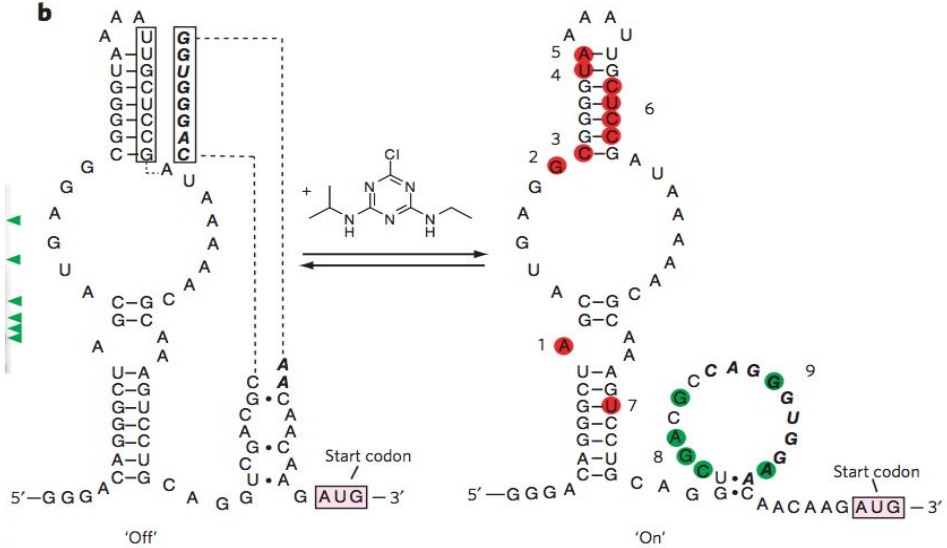
Abstract | MicroRNA (miRNA) alterations are involved in the initiation and progression of human cancer. The causes of the widespread differential expression of miRNA genes in malignant compared with normal cells can be explained by the location of these genes in cancer-associated genomic regions, by epigenetic mechanisms and by alterations in the miRNA processing machinery. MiRNA-expression profiling of human tumours has identified signatures associated with diagnosis, staging, progression, prognosis and response to treatment. In addition, profiling has been exploited to identify miRNA genes that might represent downstream targets of activated oncogenic pathways, or that target protein-coding genes involved in cancer.

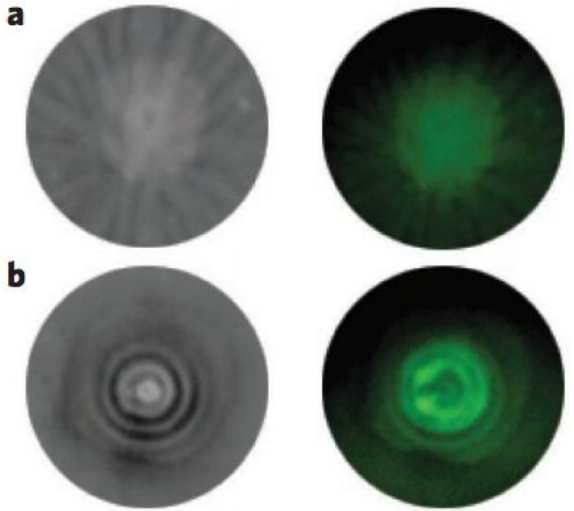
Calin et al., (2006)

Importance of RNA Structure (2)

Sinha et el., (2010) Reprogramming bacteria to seek and destroy an herbicide

- **Manipulate E. Coli's RNA to follow herbicide**





Sinha et al., (2010)

RNA Design: Computational Determination

SIAM J. APPL. MATH.
Vol. 35, No. 1, July 1978

© Society for Industrial and Applied Mathematics
0036–1399/78/3501–0006 $01.00/0

ALGORITHMS FOR LOOP MATCHINGS*

RUTH NUSSINOV,† GEORGE PIECZENIK,‡ JERROLD R. GRIGGS¶
AND DANIEL J. KLEITMAN§

Abstract. A simplified (two-base) version of the problem of planar folding of long chains (e.g., RNA and DNA biomolecules) is formulated as a matching problem. The chain is prescribed as a loop or circular sequence of letters A and B, n units long. A matching here means a set of A-B base pairings or matches obeying a planarity condition: no two matches may cross each other if drawn on the interior of the loop. Also, no two adjacent letters may be matched. We present a dynamic programming algorithm requiring $O(n^3)$ steps and $O(n^2)$ storage which computes the size of the maximum for the given A-B base sequence and which also allows reconstructing a particular folded form of the original string which realizes the maximum matching size. The algorithm can be adapted to deal with sequences with larger alphabets and with weighted matchings.

An algorithm is also presented for a modified problem closer to the biochemical problem of interest: We demand that every match must be adjacent to another match, forcing groups of two or more parallel matches.

Some results on the expected maximum matching size are presented. As $n \to \infty$, at least 80% of the vertices can be matched on the average on an A-B string of size n.

We briefly discuss the practical application of the algorithm by using contracted versions of very long molecules with a preliminary block construction. A maximum matching is presented for the J-gene of the ϕX174 DNA virus. We conclude by stating some problems requiring further study.

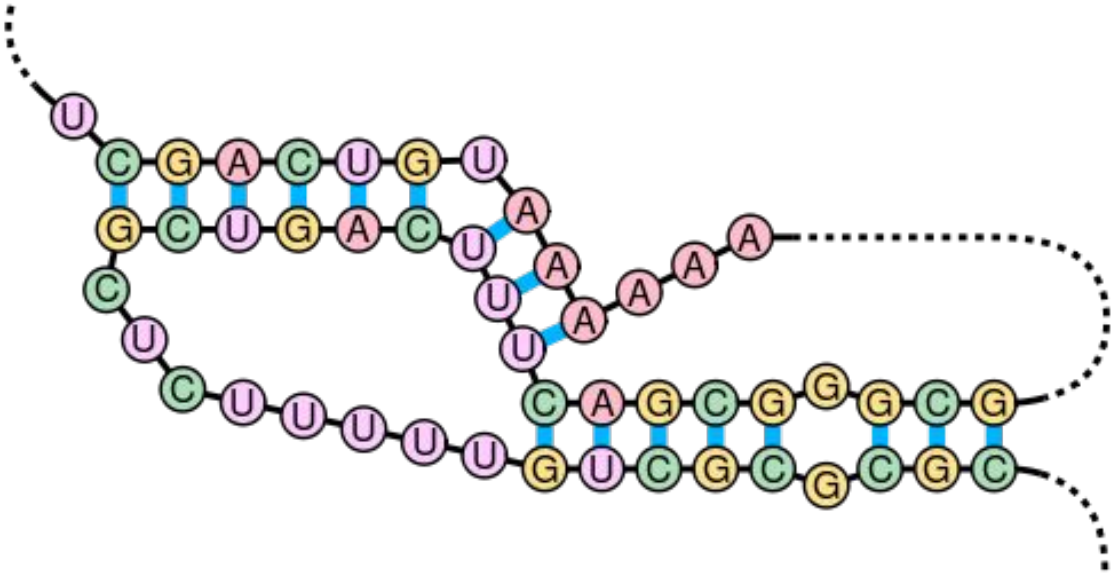
Nussinov et al., 1978

Table 4. Nearest neighbor parameters for G·U base-pairs

Nearest-neighbor	ΔG°_{37} (kcal/mol)	Error	ΔH° (kcal/mol)	Error	ΔS° (eu)[c]	Error[c]
5′ AG 3′ 3′ UU 5′	−0.55	0.32	−3.21	2.76	−8.6	8.45
5′ AU 3′ 3′ UG 5′	−1.36	0.24	−8.81	2.10	−24.0	6.44
5′ CG 3′ 3′ GU 5′	−1.41	0.24	−5.61	2.13	−13.5	6.53
5′ CU 3′ 3′ GG 5′	−2.11	0.25	−12.11	2.22	−32.2	6.81
5′ GG 3′ 3′ CU 5′	−1.53	0.27	−8.33	2.33	−21.9	7.14
5′ GU 3′ 3′ CG 5′	−2.51	0.25	−12.59	2.18	−32.5	6.67
5′ GA 3′ 3′ UU 5′	−1.27	0.28	−12.83	2.44	−37.3	7.47
5′ GG 3′ 3′ UU 5′	+0.47 (−0.5)[b]	0.96	−13.47	8.37	−44.9	25.65
5′ GU 3′ [a] 3′ UG 5′	+1.29	0.56	−14.59	4.92	−51.2	15.08
5′ GGUC 3′ [a] 3′ CUGG 5′	−4.12	0.54	−30.80	8.87	−86.0	23.70
5′ UG 3′ 3′ AU 5′	−1.00	0.30	−6.99	2.64	−19.3	8.09
5′ UG 3′ 3′ GU 5′	+0.30	0.48	−9.26	4.19	−30.8	12.86
Each Terminal G·U[d]	+0.45	-	+3.72	-	+10.5	-

Matthews et al., 1999

RNA Design: Pseudoknots



Pseudoknot found in telomerase
(Wikipedia)

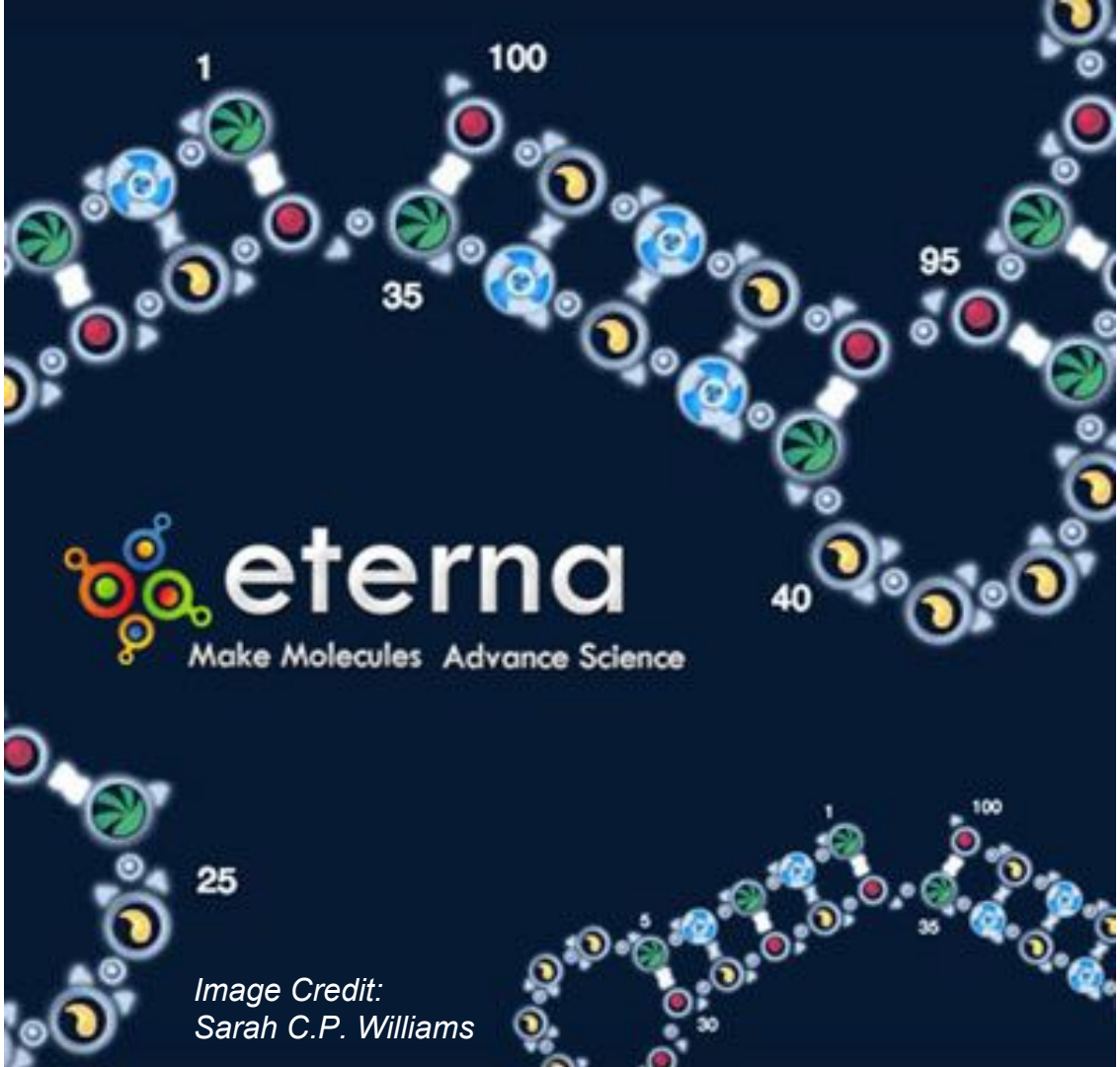
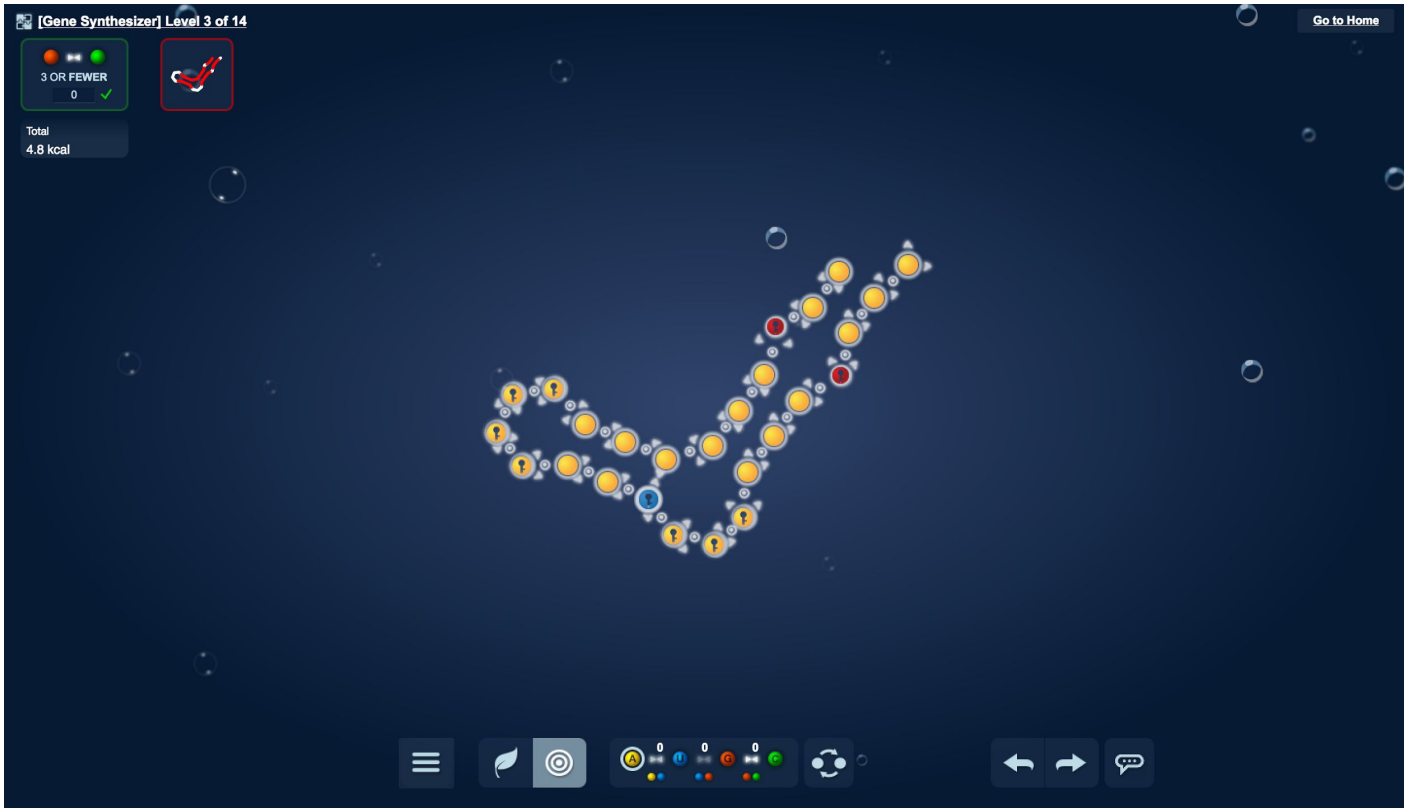


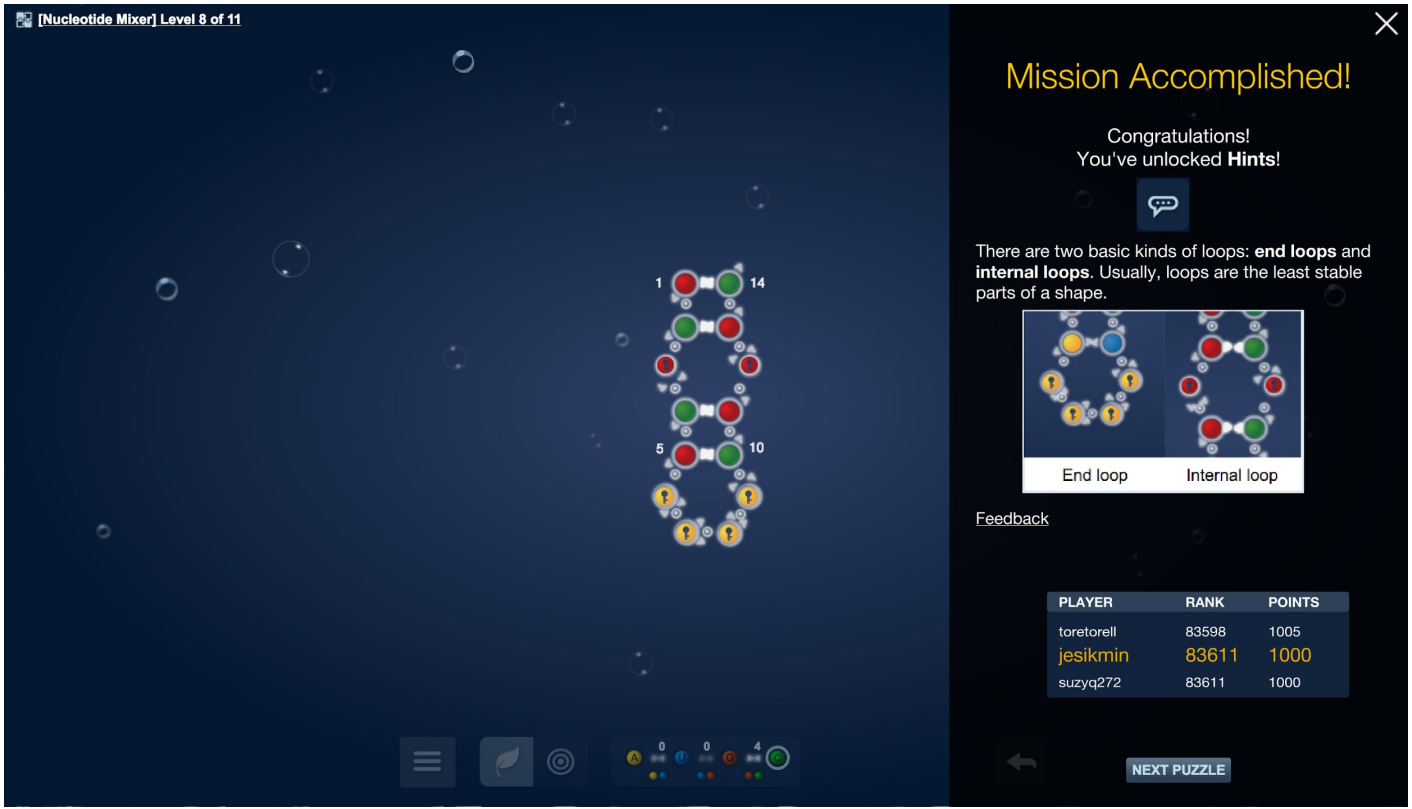
Image Credit:
Sarah C.P. Williams

Image Credit: Ivy Lily Crafts
http://bit.ly/2jpip8m

Design Interface: Tutorial / Badge Collections



Design Interface: Lesson



Design Interface: Challenge Puzzles

Demo!

RNA Design Rules from a Massive Open Laboratory

Jesse Min

Table of Contents

1. Emergence of RNA Folding Models: Vienna RNA Package
2. Citizen Science: Massive Open Laboratory
3. Interfaces
4. Results
5. Significances
6. Limitations

Emergence of RNA Folding Models: ViennaRNA

Hofacker et al., (2004) RNA secondary structure analysis using the Vienna RNA package

1) **Classical minimum free energy algorithm** (Zuker and Stiegler., 1981)

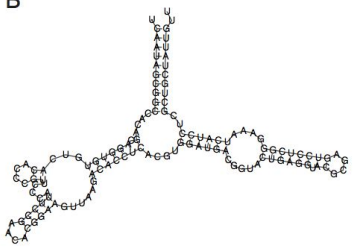
2) **Partition function algorithm** (McCaskill., 1990)

3) **Suboptimal folding algorithm** (Wuchty et al., 1999)

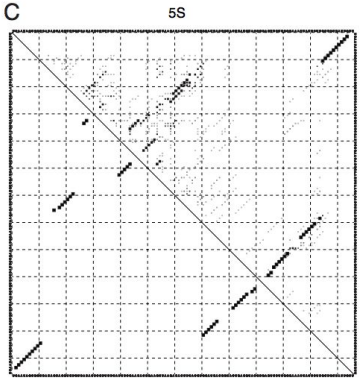
A

```
% cat 5s.seq
> 5s
UCAAUAGCGGCCACAGCAGGUGUGUCACACCCGUUCCCAUUCCGAACACGGAAGUUAAGACACCUCACGUGGAUGACGGUACUGAGGUACGCGAGUCCU
% RNAfold -p < 5s.seq > 5s.fold
RNAfold -p < 5s.seq > 5s.fold
moritz ~/TeX/CP/fig> cat 5s.fold
% 5s
UCAAUAGCGGCCACAGCAGGUGUGUCACACCCGUUCCCAUUCCGAACACGGAAGUUAAGACACCUCACGUGGAUGACGGUACUGAGGUACGCGAGUCCU
.((((((((((....(.((((((...((....)).....(((((....))))).....))))))).....((((((.....((((((.((....))))))))....))))))..)))))))))).. (-47.70)
.((((((((((....{..(({((.[.......({{{.......|}...|}}..)}....))))})}..(.((((((.....((((((.((....))))))))....))))))..)))))))))).. [-52.08]
 frequency of mfe structure in ensemble 0.000814848       0.000814848
% gv 5s_ss.ps
% gv 5s_dp.ps
% mountain.pl 5s_dp.ps | xmgrace -p1pe
```

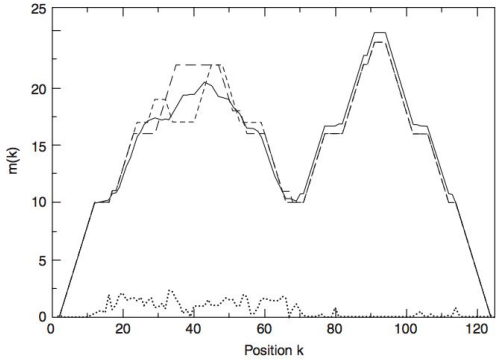
B



C



D



Hofacker et al., (2004)

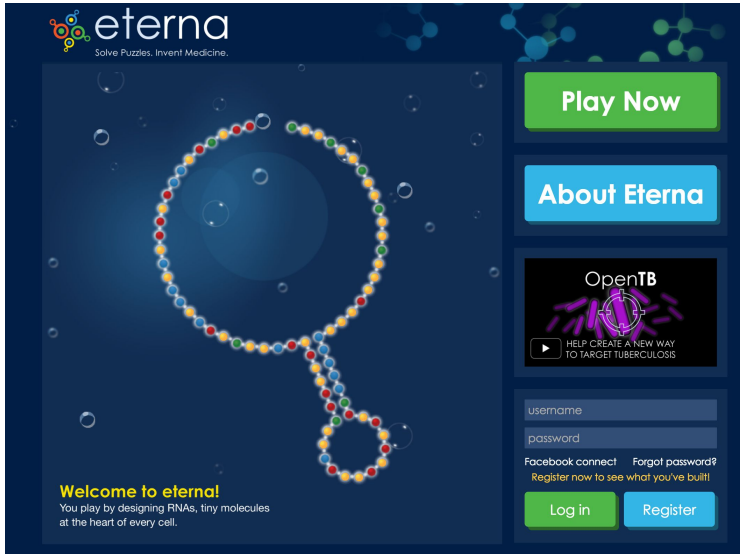
Citizen Science: Massive Open Laboratory

Limitations of current approach

- Small group of professional scientists must interpret empirical data
- How about with ML and visualization tools?

Massive Open Laboratory

1) **Simulated molecules**
2) **Remote experimental pipeline**
3) **Design RNA structures**

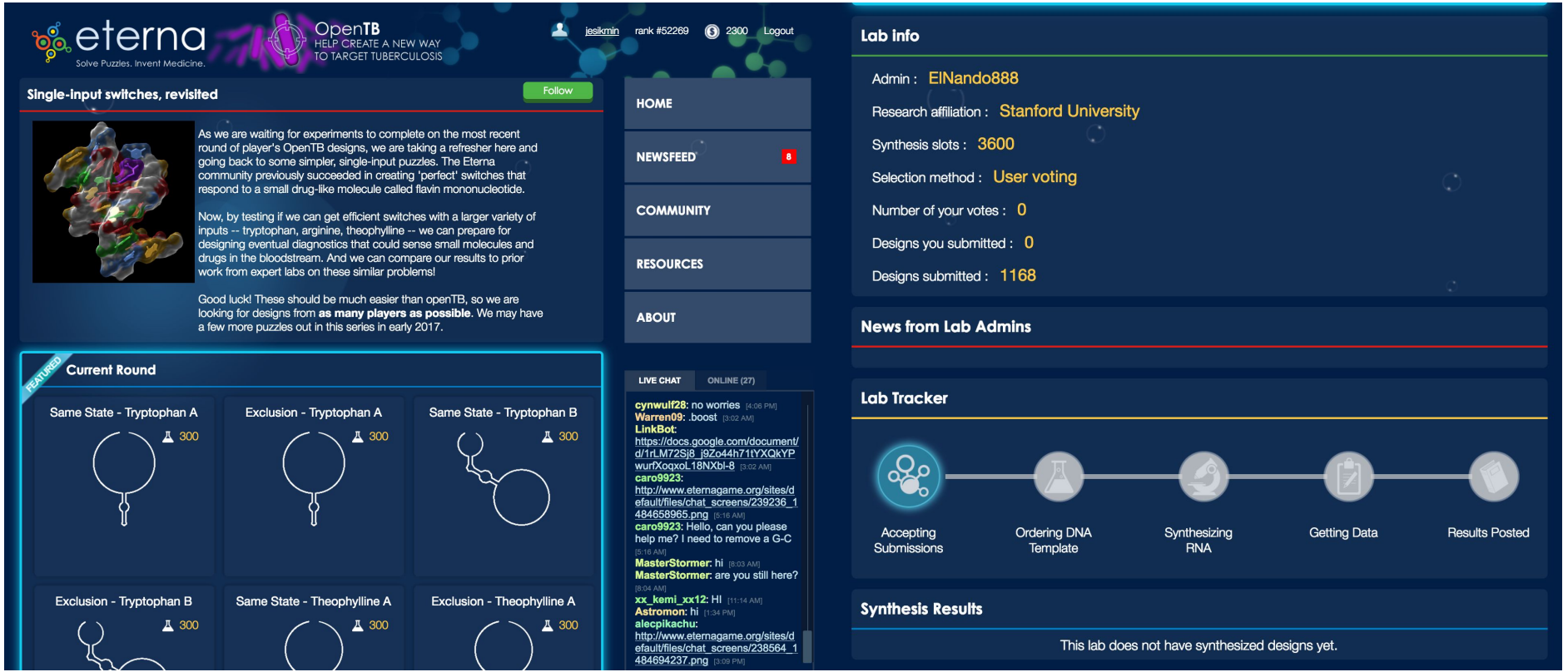


http://www.eternagame.org/web/

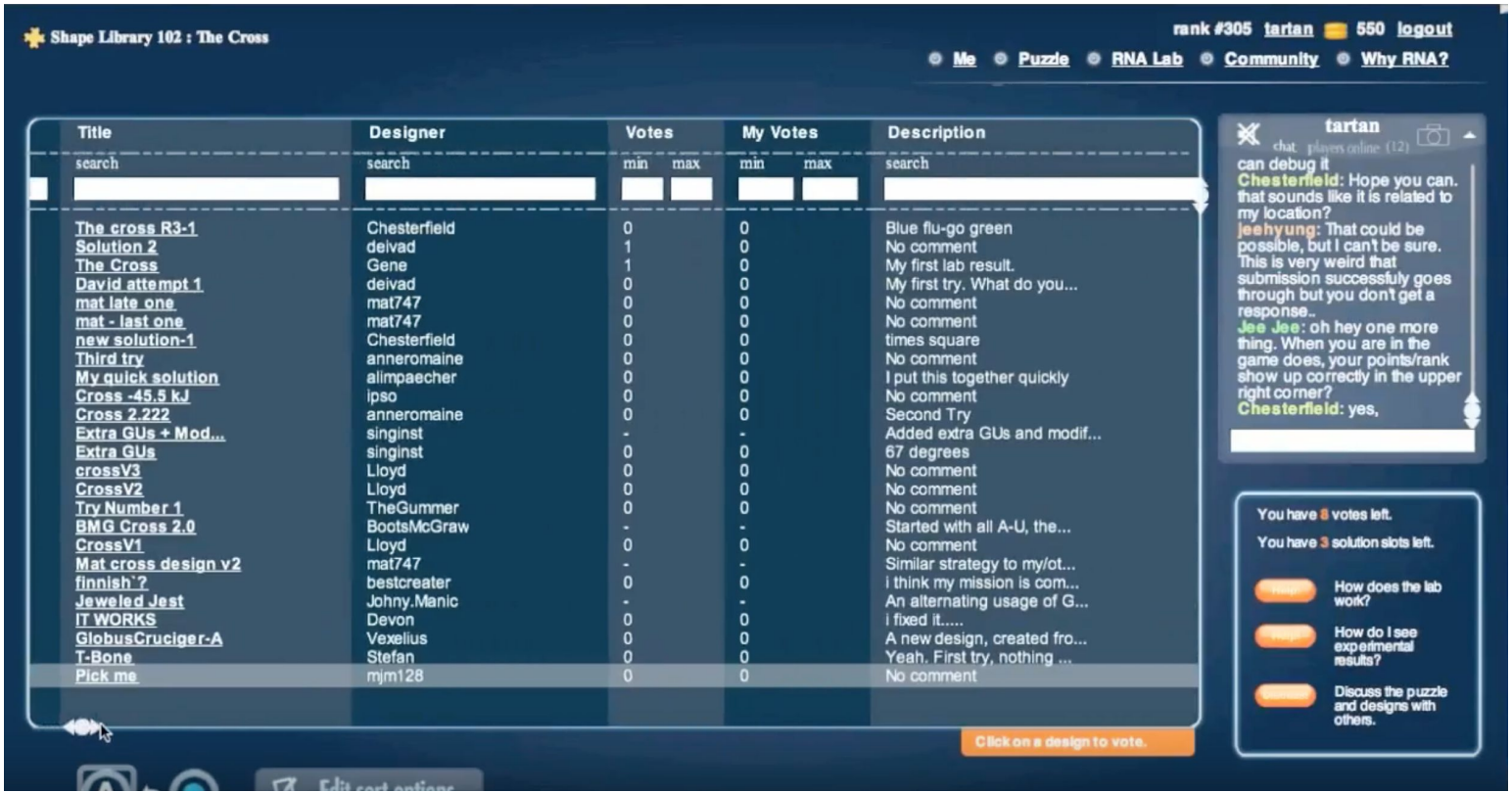
Design Interface: Challenge Puzzles

Lab Interface (1)

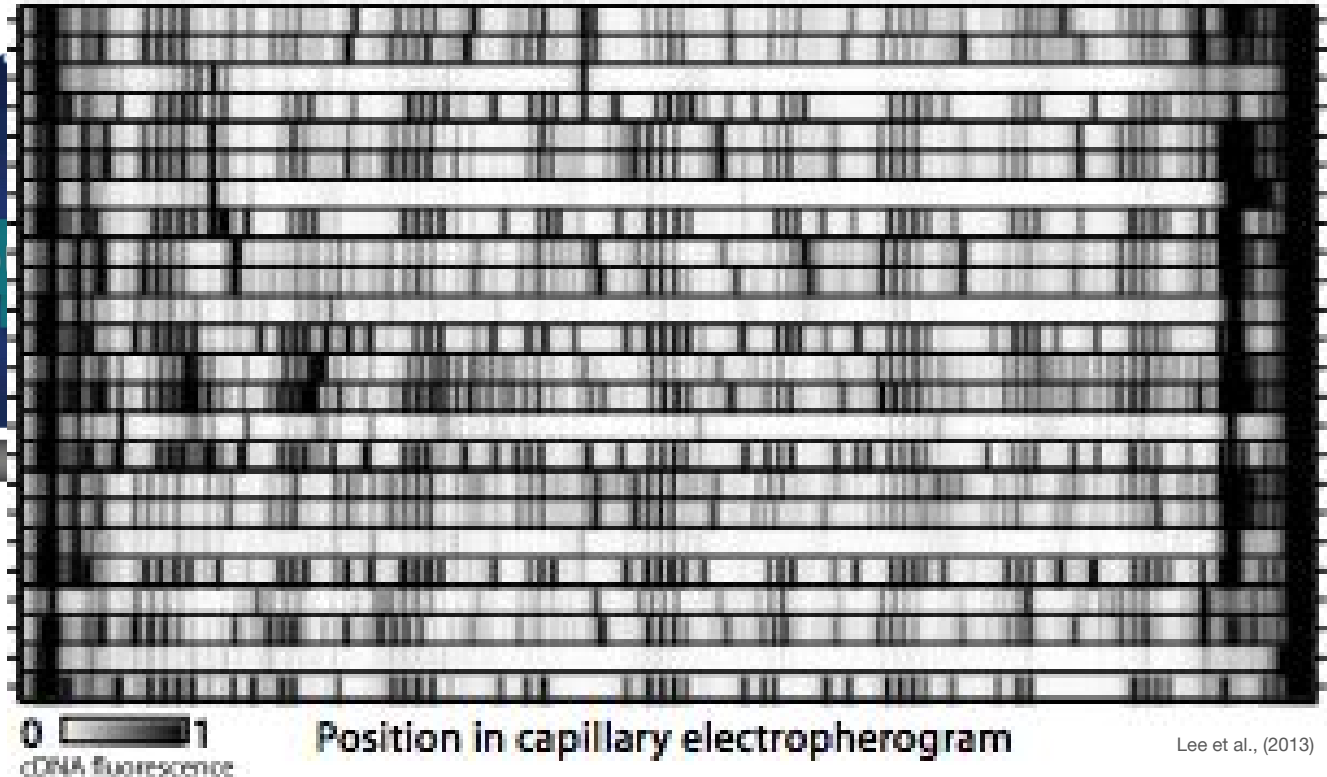
Lab Interface (2)



Voting Interface

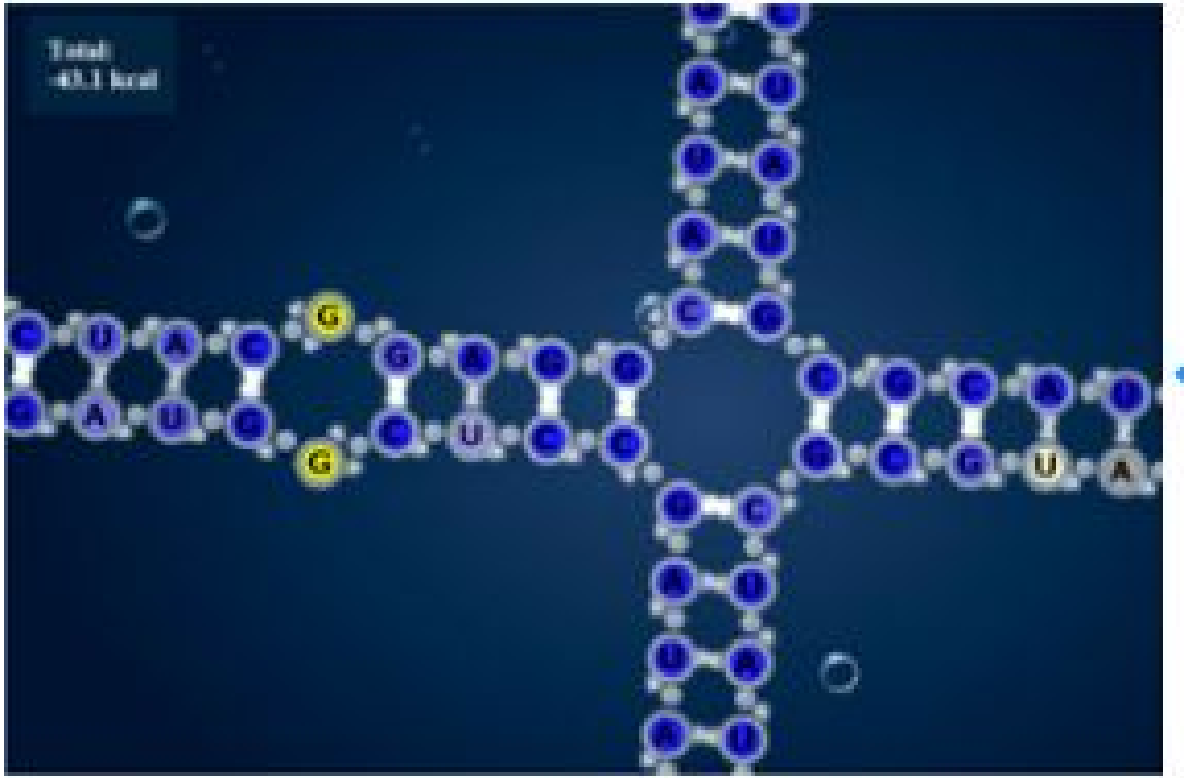


Remote Lab



Lee et al., (2013)

Result Viewer



Lee et al., (2013)

Results & Rule Collections

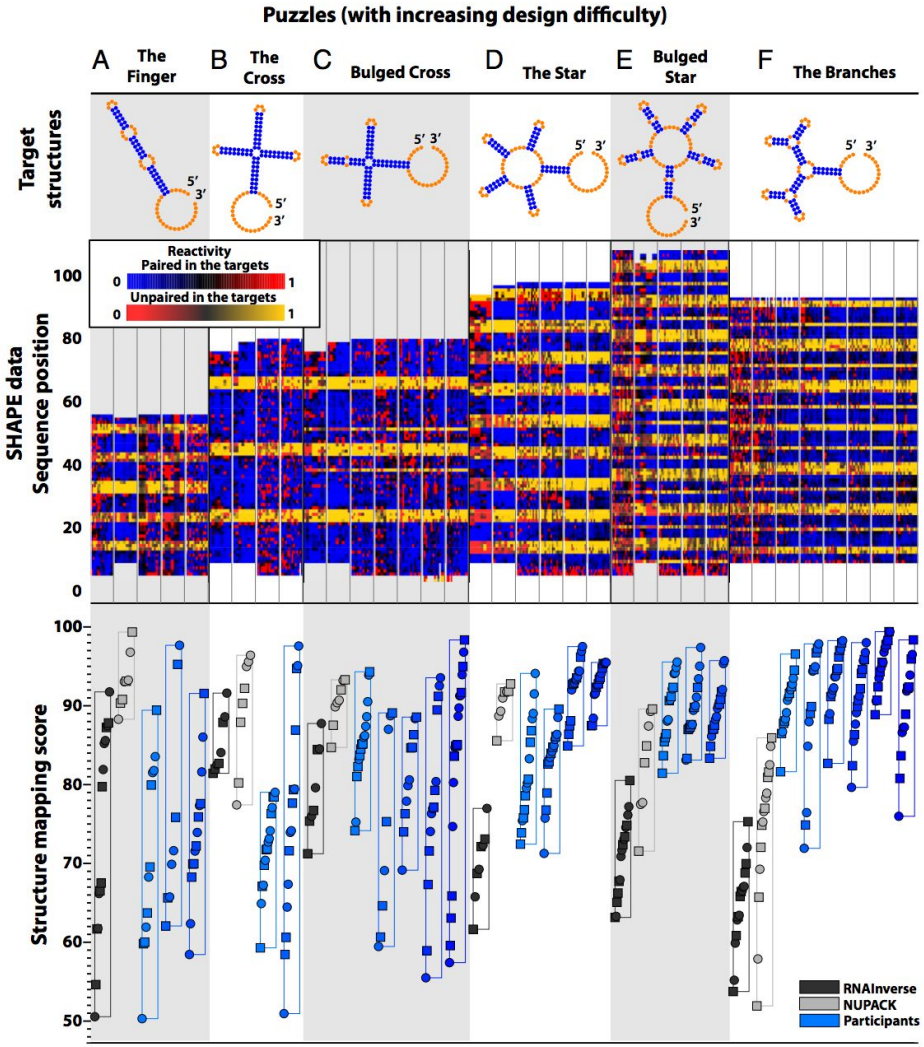
Winners (22 out of 34)

User	Name	Puzzle Title	Solution Title	Score
	Eli Fisker	[A]/[C] - DEC	Rocketdog 42 mod 3	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 G53+G85 Eli Fisker mod #MUTATE2	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 G52+U53 Eli Fisker mod #MUTATE2	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 U52+U53 Eli Fisker mod #MUTATE2	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 G53+U56 Eli Fisker mod #MUTATE2	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 G52+G85 Eli Fisker mod #MUTATE2	100
	jandersonlee	[A]/[C] - INC	jl-acinc 7061024 U52+G53 Eli Fisker mod #MUTATE2	100
	joy45	[B]/[C] - INC	B/C INC J1	100
	mat747	[A]/[C] - INC	Single #MUTATE 7061024 G85 Eli Fisker mod	100
	mat747	[A]/[C] - INC	Single #MUTATE 7061024 U53 Eli	100

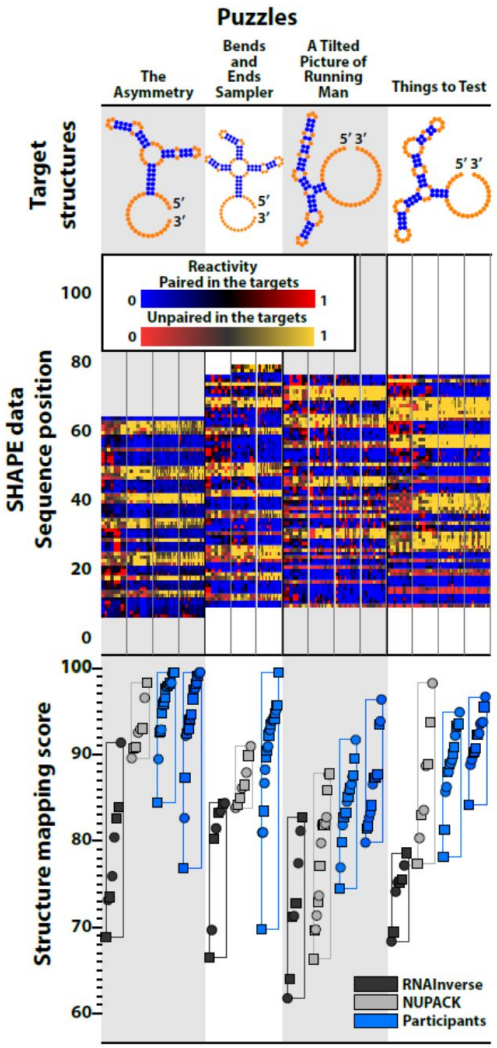
Review results | Lab Info

Lee et al., (2013)

Result: Phase(1)



Result: Phase(1) to Phase(2)



EteRNABot

A Basic Test

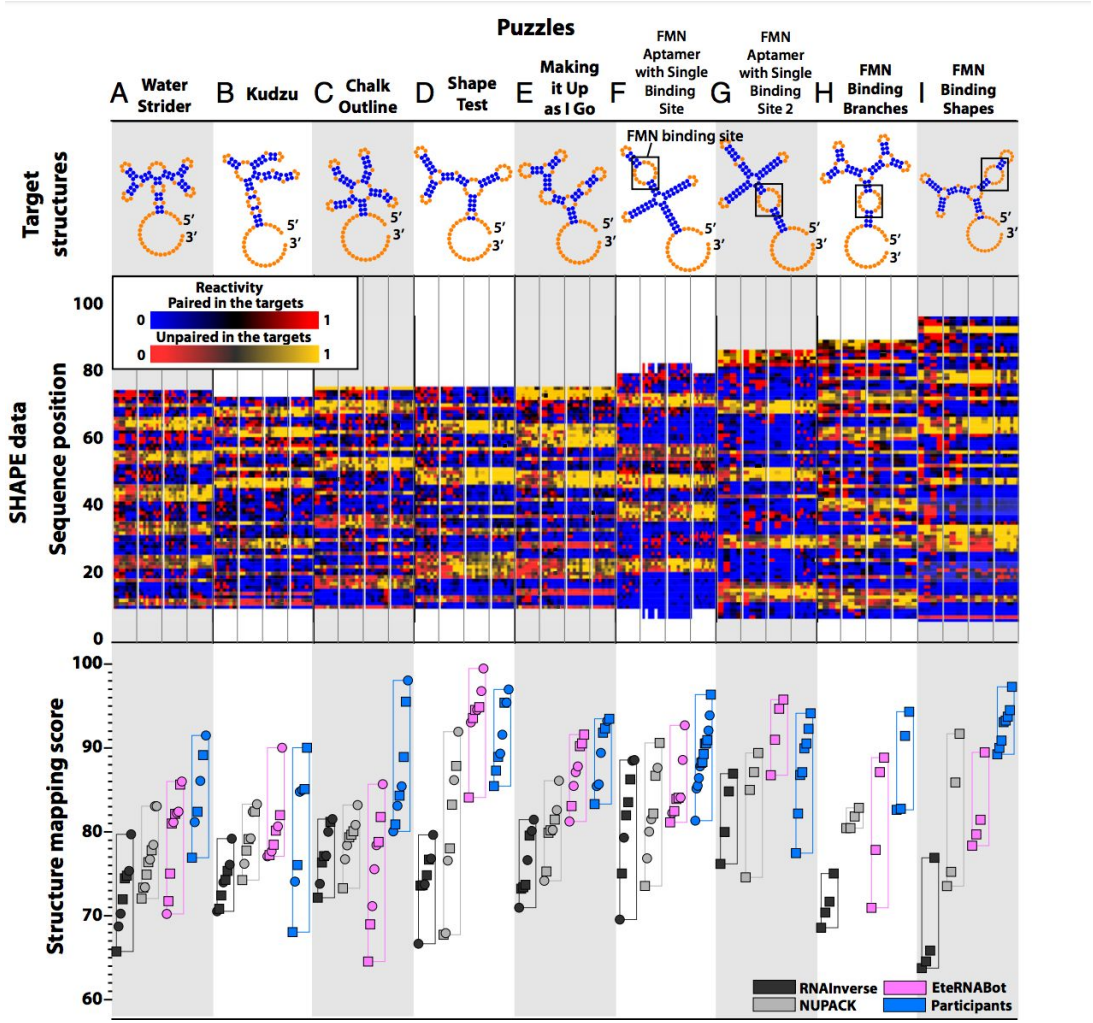
Clean Plot, stack caps, and safe GC

Direction of GC-pairs in multiloops + neckarea

Berex Test

Numbers of Yellow Nucleotides per Length of String

Result: Phase(2)



Significance: User-created Design Rules

User-created design rules

- Player-created Guides
- Multi-loop Guides
- **Negative Design Rules**
  - Penalties on special structures (repetitions, tetraloop similarity, and twisted base pairs)

Significance: Feedback from Real Experiments

Previous Efforts

Year	Paper	Authors
2001	Can distributed volunteers accomplish massive data analysis tasks?	Kanefsky et al.
2008	Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey	Land K et al.
2011	Algorithm discovery by protein folding game players	Khatib F et al.
2012	Increased Diels-Alderase activity through backbone remodeling guided by Foldit players	Eiben CB et al.

Lee et al., (2013)

- Collectively **generate and test hypotheses through actual experiments**

Limitations: What is the next step?

Engineering Cost

Theoretical Explanation

More Incentive

Social Prejudice

Limitations: What is the next step?

Treuille et al., (2014) Scientific rigor through videogames

- Budget
- Career risk
- Generalization
  - Schindel et al., (2005) DNA barcoding a useful tool for taxonomists
  - Helmstaedter et al., (2013) Cellular-resolution connectomics: Challenges of dense neural circuit reconstruction

Principles for Predicting RNA Secondary Structure Design Difficulty

...

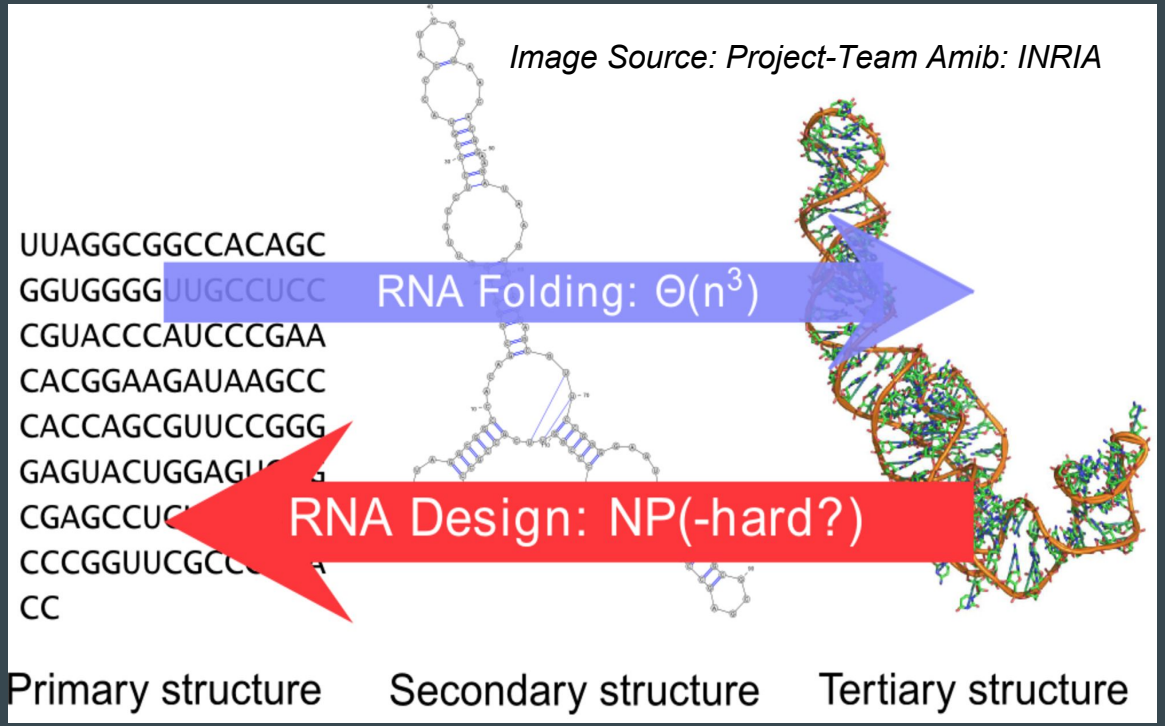
Jeff Anderson-Lee, Eli Fisker, Vineet Kosaraju1, Michelle Wu1, Justin Kong1, Jeehyung Lee1, Minjae Lee1, Mathew Zada1, Adrien Treuille1, Rhiju Das1, Eterna Players

Presentation Overview

- Paper Focus/Main Idea
- Difficult Structures
  - Short Stems
  - Symmetry
  - Structural Motifs
- Impact
- Weaknesses

Inverse Folding Problem

- Know to be hard
- Unknown which structures are tractable
- Paper set forth measures for difficulty



Research Process

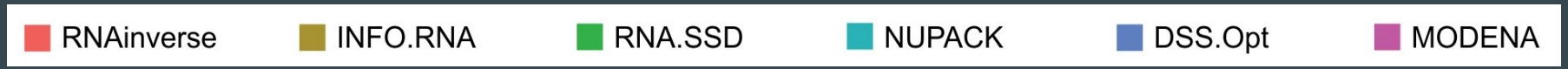
- Although “RNA design rules from a massive open laboratory” was revolutionary with 37,000 authors, the lead authors were scientists
- In “Principles for Predicting RNA Secondary Structure Design Difficulty” the lead authors are Eterna players: the revolution continues
- Jeffrey Anderson-Lee: computer systems manager
  - “It's really a bit amusing, I guess it's stretching [the journal’s] boundaries a bit.”
- Eli Fisker: Librarian
  - “When the scientist is away, the lab rat will play”

Overview

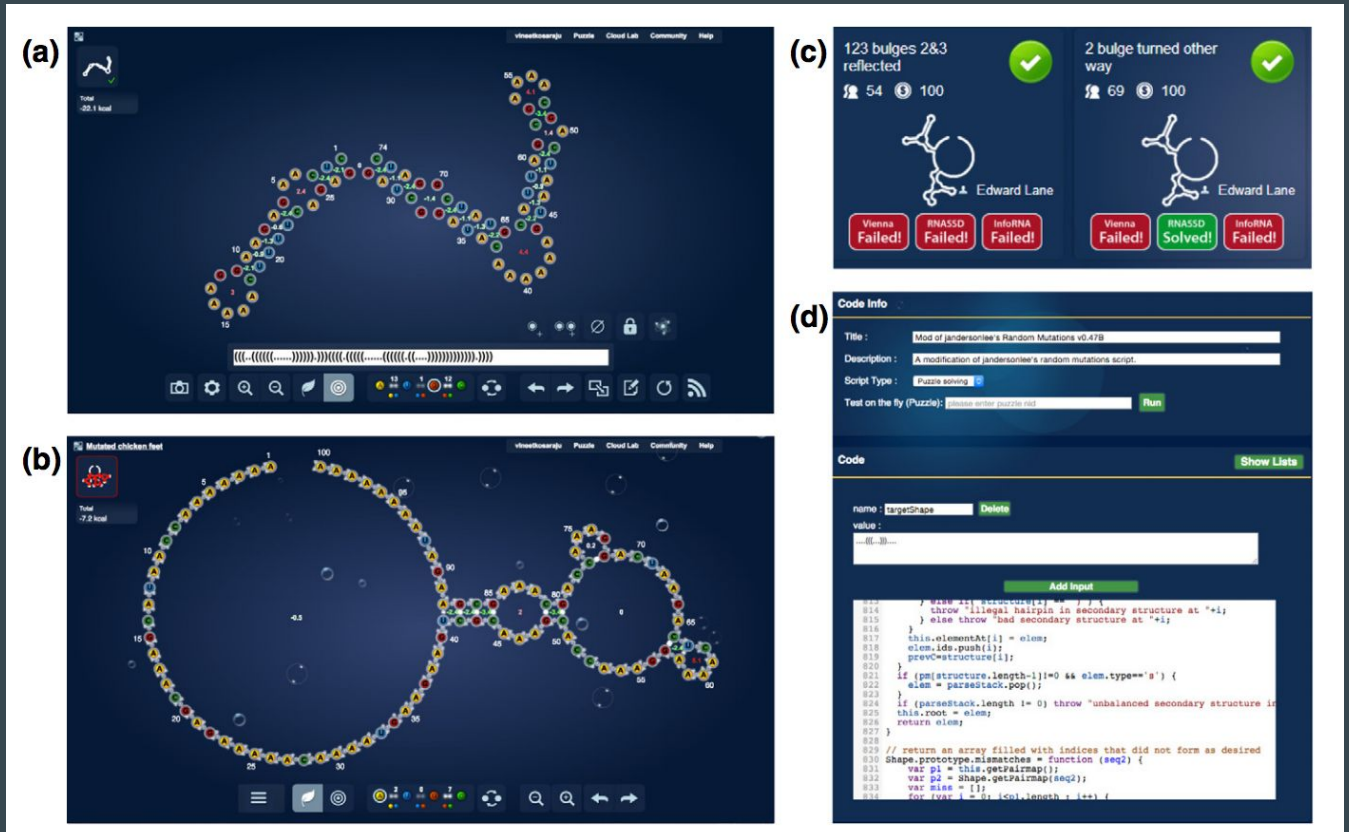
- Paper Structure/Main Idea
- **Difficult Structures**
  - Short Stems
  - Symmetry
  - Structural Motifs
- Impact
- Weaknesses

Difficult Structures

- The paper gathers knowledge from the EteRNA player community
- Compares player input to automated design algorithm performance

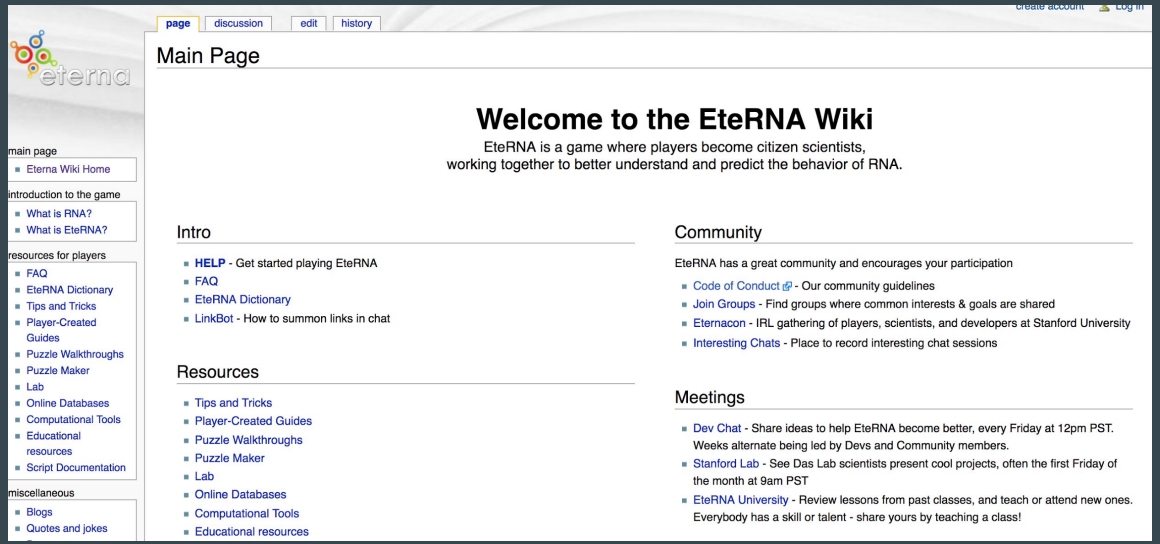


Identifying Difficult Problems: Tools



Identifying Difficult Problems

- Player's use problem solving, problem creation, algorithm integration and scripting tools to test hypotheses about difficulty
- Player insight is synthesized online in community wiki: crucial to research



Short Stems

- Structures joined by short stems are difficult
- Multiplicity of short stems affects free energy

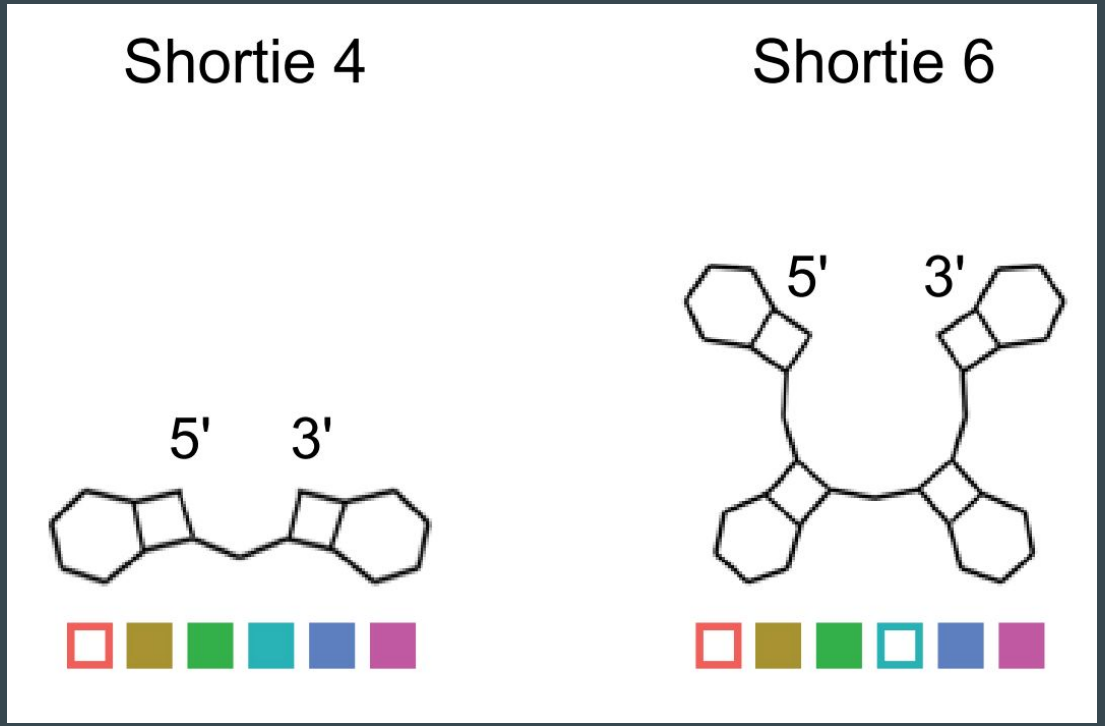


Image Source: Principles for Predicting RNA Secondary Structure Design Difficulty

Symmetry

- True symmetry is rare in naturally occurring RNA structures
- Authors postulate this is due to difficulty of design

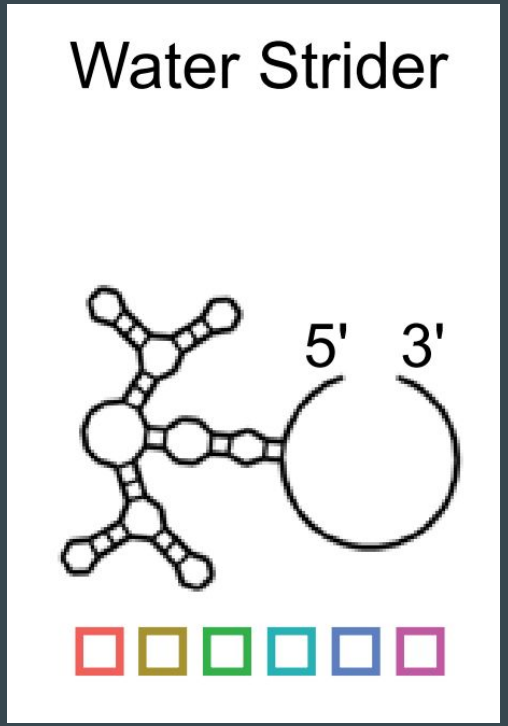


Image Source: Principles for Predicting RNA Secondary Structure Design Difficulty

Structural Motifs

- Player designated “zig-zags” are an example of player discovered difficult structural motifs
- “Zig-zags” close to multi-loops cause current design algorithms to fail
- Again, rare in nature

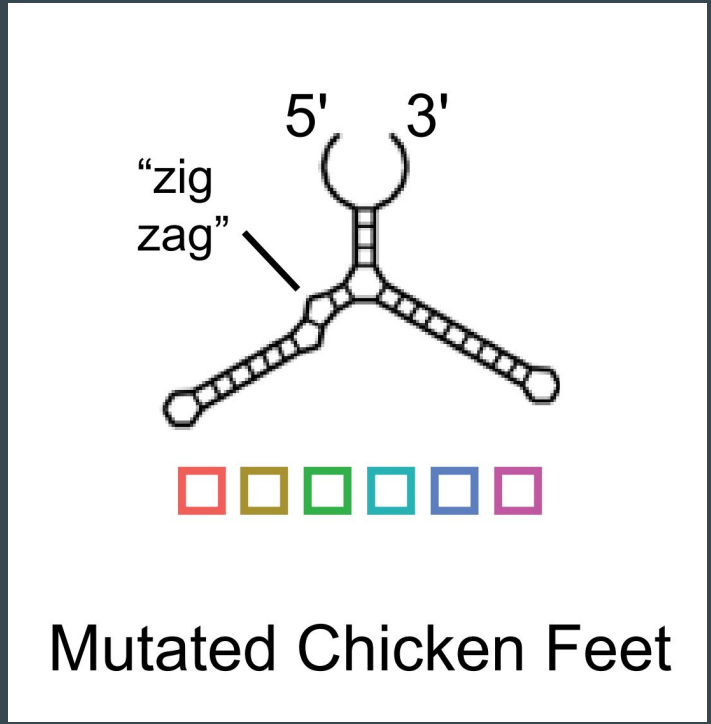


Image Source: Principles for Predicting RNA Secondary Structure Design Difficulty

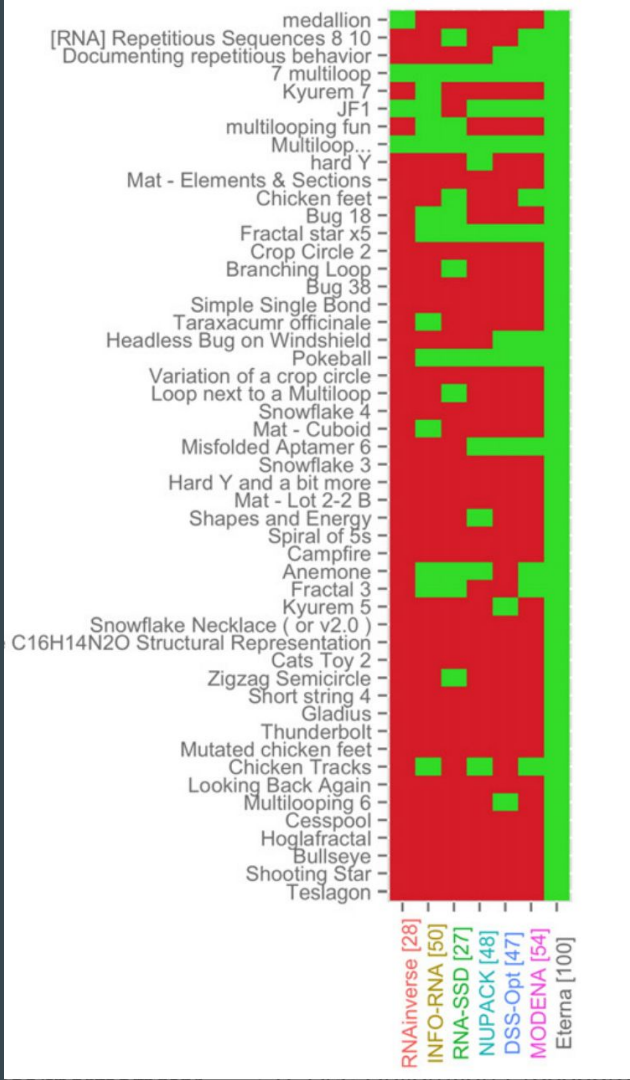
Overview

- Paper Structure/Main Idea
- Difficult Structures
  - Short Stems
  - Symmetry
  - Structural Motifs
- **Impact**
- Weaknesses

Impact

- Paper set forward the Eterna100 benchmark of secondary structure design challenges
- Allows for common benchmark of *in silico* design algorithms

Image Source: Principles for Predicting RNA Secondary Structure Design Difficulty



Overview

- Paper Structure/Main Idea
- Difficult Structures
  - Short Stems
  - Symmetry
  - Structural Motifs
- Impact
- **Weaknesses**

Weaknesses

- Focus on in-silico test
- No *in vitro* or *in vivo* testing
  - The authors do not include benchmark evaluation on EteRNABot, an algorithm designed for in vitro success

Questions?

Additional Slides