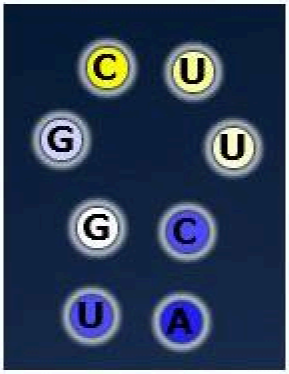
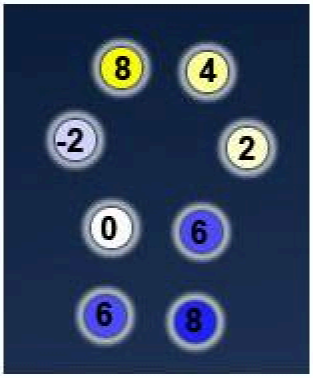
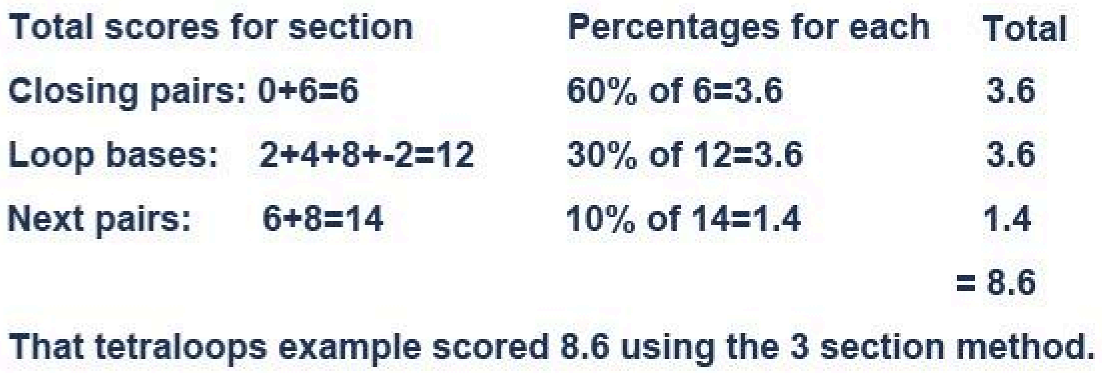
Computationally Selected Elements







Total scores for section		Percentages for each	Total
Closing pairs:	0+6=6	60% of 6=3.6	3.6
Loop bases:	2+4+8+-2=12	30% of 12=3.6	3.6
Next pairs:	6+8=14	10% of 14=1.4	1.4
			= 8.6

That tetraloops example scored 8.6 using the 3 section method.

This is my attempt to convert my intuition for lab analysis into formulas (strategies) that can be used for the analysis of lab results via computational methods. I used this method of analysis together with other strategies, for lab designing. ***Computationally selected elements*** is the idea for a tool with strategies to analyse the lab results, to aid the later designing process.

The idea for converting the colours to numbers started when I was thinking about how to help players with colorblindness view the Lab 101 - The Finger results. I spoke with, and sent images to Sneh (Dev) back around the time of Lab 101.

22 August 2011, I sent Jee the whole skeleton of my idea. Jee asked: So you are scoring patterns based on their experimental results, and try to use high scored patterns as much as possible? This is exactly my aim.

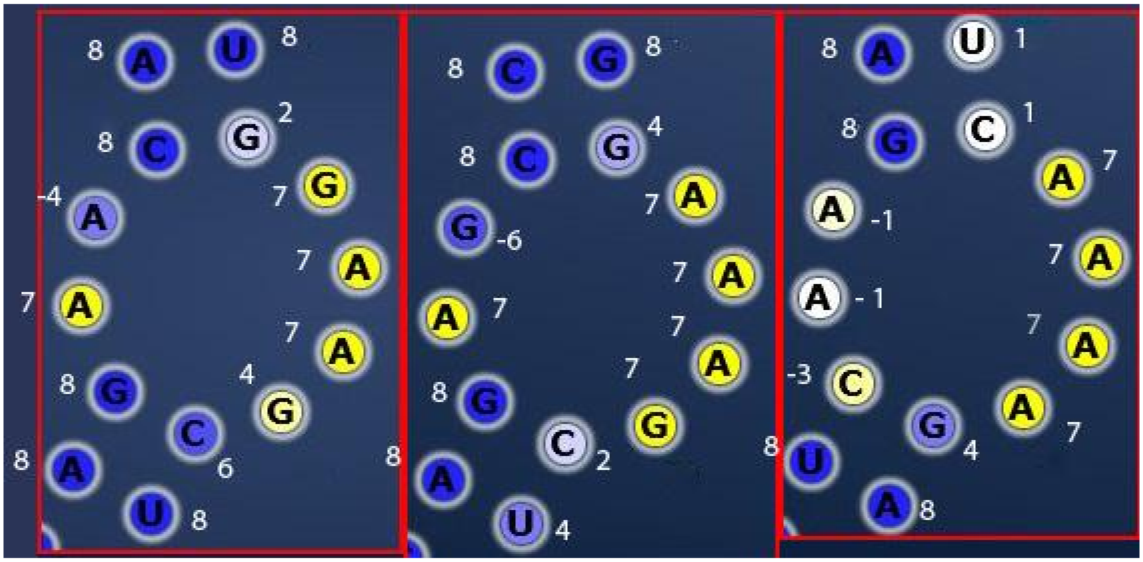
At the time of sending the email to Jee, my main analysis style was to look for tendencies in the results, with a method that uses three sections. This tries to balance the different areas of element/fragment when making a conclusion about that fragment compared with other fragments, which became easier in april 2011 with the arrival of the continuous experimental color mode.

Below is the main content from the mail sent to Jee 22 August, 2011, with just a few pictures added to it:

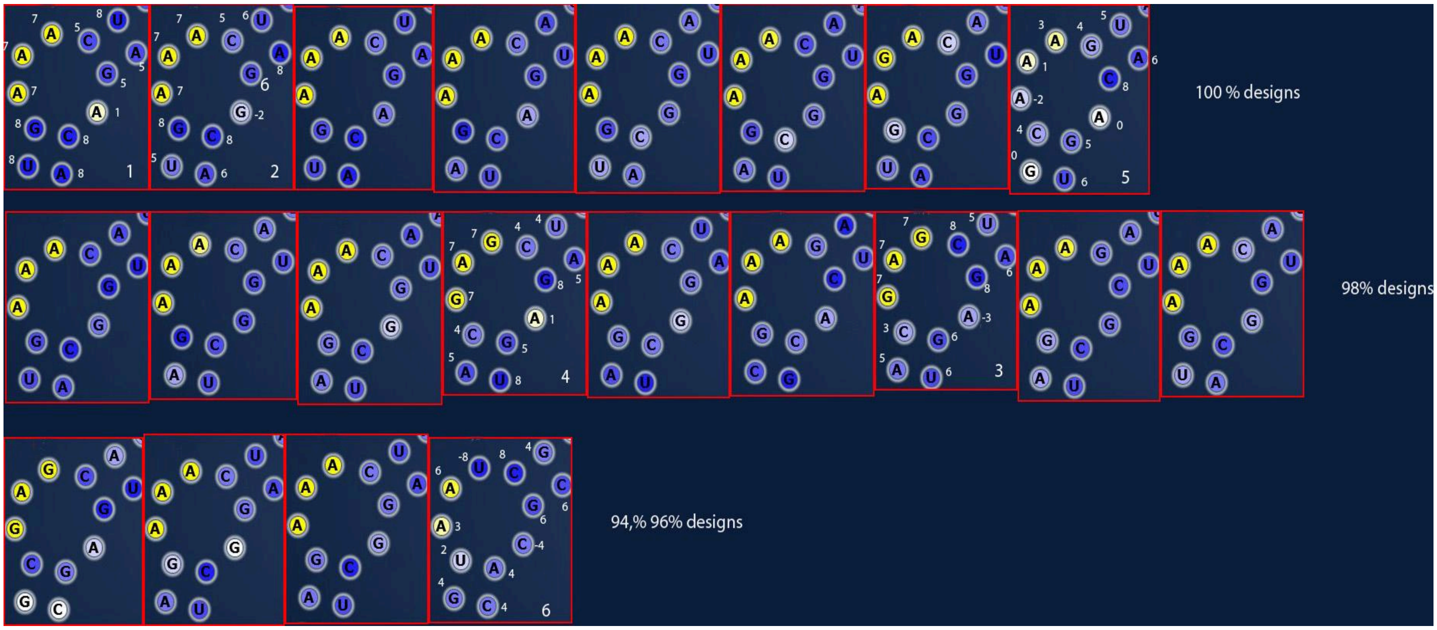
I have seen the other week in chat that the “strategies conflict with each other sometimes and it's inevitable that some strategies get violated in some cases”, and before Jee told me that Dev was not trying to mimic a player's style. So I thought: What was my playing style/strategy? At the time I was looking at the past updated lab results and doing screenshots. I have been comparing and ranking patterns of bases used on elements for my strategy, and thought: Could the ranking be done by computer with the right formula?

I’m still working on the idea of computationally selected elements, strictly because it is time consuming to visually score each base/element. But as the colours are generated computationally, it would be easy to display score as a number with or without the colours.

I have visually/manually scored some bases/elements as a quick check of the idea. I have been using the scale (Example below). But this can be subject to human error.



Below - The numbers in the corner are the order I think the elements would be ranked in visually.

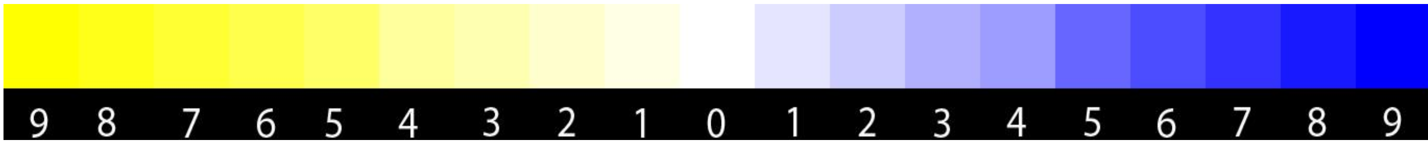


Here is the design format I ended up with.

The amount of elements I have checked so far is very small and would need to be increased to refine the idea and average out the -/+ 4% error rate in the lab results. Below I have put some ideas about how it could be done and what I look for in the results.

Idea for Computationally Selected Elements

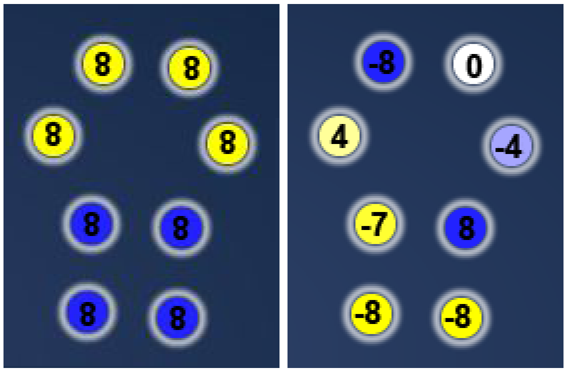
I converted the colours to numbers with the aid of a colour scale. The formula to reproduce the colors was sent to me by Jee. The results are showed in 9 different shades of yellow, and 9 different shades of blue and plus one white for each nucleotide.



-- For paired nucleotides (like in a stem), Dark Blue is good and yellow is bad.
So a dark blue (which is best) would be a 9 on the scoring scale.
The lighter the blue the lower the score and a yellow would be a minus.

-- For non paired nucleotides (like in a loop), Dark yellow is best and blue is bad.
So a dark yellow (which is best) would be a 9.
The lighter the yellow the lower the score and a blue would be a minus.

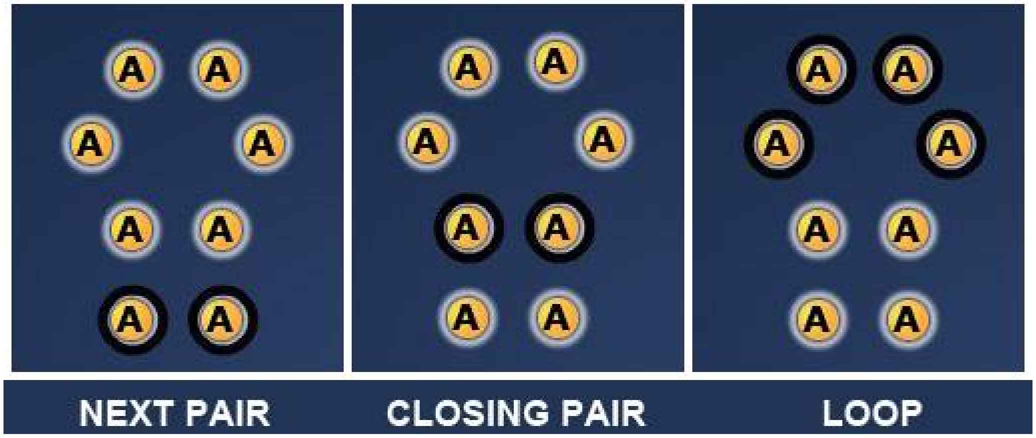
(Note: All designs results have been converted to numbers using a reduced scoring range of 8 yellow to 8 blue. Due to yellow 8 & 9 being very similar in colour, I wanted to reduce my error rate of the conversion.)



The picture to the left shows a perfectly working element. The Stem is blue and paired, the Loop is yellow and unpaired. So all nucleotides receive the top score of +8 pr. nucleotide x 8 = 64. The element to the right shows a hairpin, where the stem is not binding but the loop has paired up. Thus the overall negative score -23.

Here are my strategies to show what I call ranking of elements.

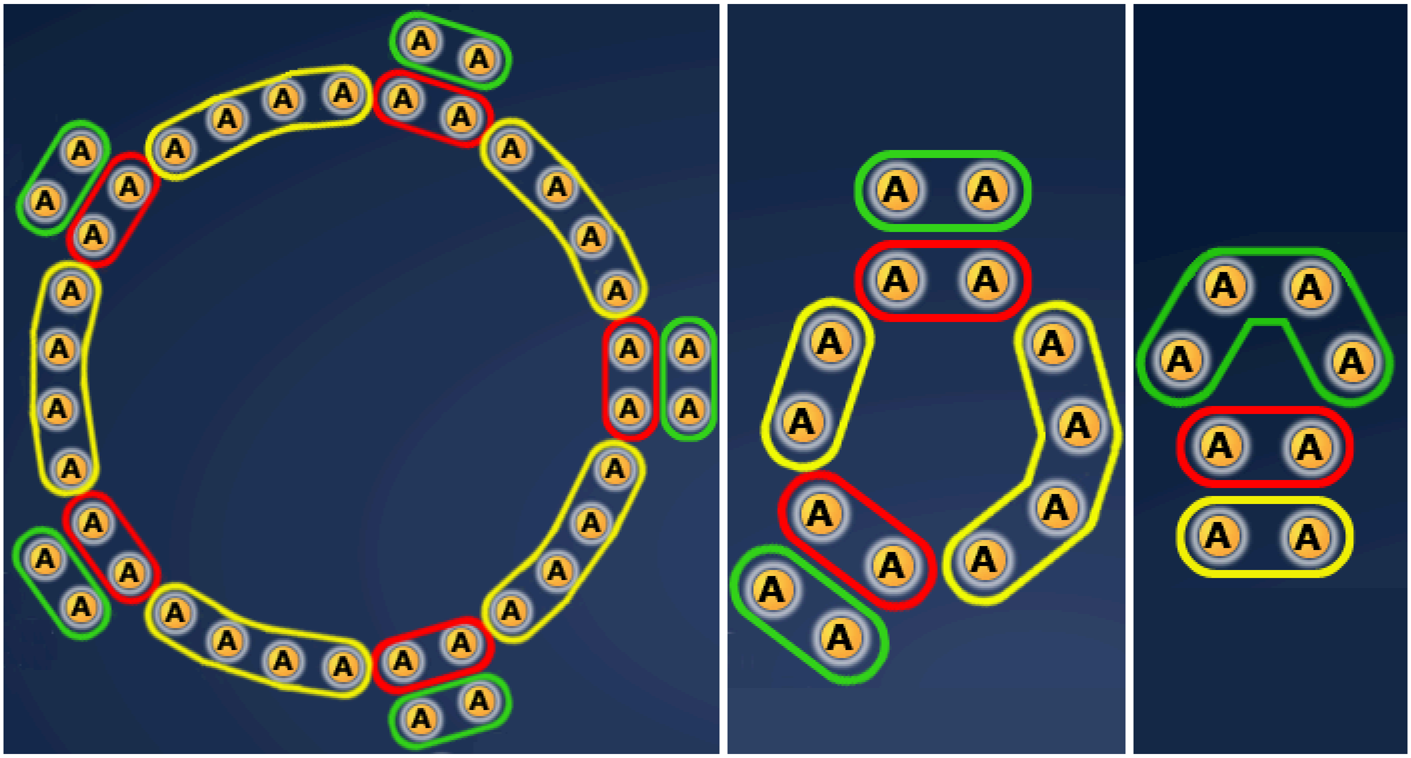
Terminology explanation



1) Total score
Add the total score of all the bases in the element.

2) Closing pairs & loop
Add the total score of the Closing pairs & loop bases.

3) Ranking of sections of an element.
Each of the sections would have a different percentage that would count to the overall total.



Possible score totals

	Multiloops	2-4 Loop	Tetraloop
Possible score			
Total:		80	80
Loop:		6x8 = 48	6x8 = 48
Closing pair:		2x8 = 16	2x8 = 16
Next pair:		2x8 = 16	2x8 = 16

Percentage of total possible score

	Multiloops	2-4 Loop	Tetraloop
Loop:	= 50%	= 60%	= 50%
Closing pair:	= 25%	= 20%	= 25%
Next pair:	= 25%	= 20%	= 25%

Possible score totals is the total score of all the bases in the element counted together. Percentage of total possible score is the distribution of score to the sections, loop, closing basepair and next pair.

Different strategies for the scoring:

A – 3 sections that would count to the final score.
Possible percentages could be
Closing pairs : 60%
loop bases : 30%
Next pairs : 10%

B – 2 sections that would count to the final score.
Percentage of could be
Closing pairs : 70%
loop bases : 30%

Why did I choose these score percentages?

It is based on the intuition I gained from my previous lab experience, lab by lab, round by round. I would analyze and design my lab submission in my head mostly; sometimes I looked at the dot plot. I would design while I was out walking; I can picture both an element and a whole design in my head. Those are the approximate percentages I used to analyze a element based on my intuition.

As an example, I give closing pairs more importance because I want elements to form correctly. Closing basepairs are more important for correct loop formation than the next pair. If the element did not form correctly, it would not be in the correct shape.

Later I converted nine lab designs into numbers to explain my system to Jee. I picked the highest scoring designs. I screenshot the entire design and picked elements from those designs. To save time, I screenshot designs from an entire lab to pick elements from those designs. I later did screenshots of the elements I wanted to use for my reference sheets.

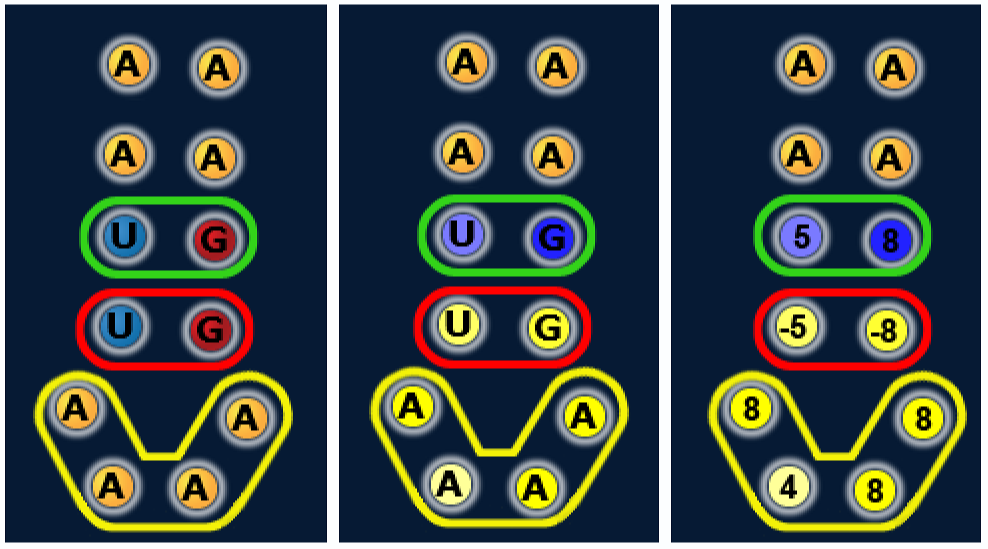
I made them so I could do comparison/split screen analysis of designs and elements and later use those as a for reference for designing. I could also watch the tendencies of a sequence over many different design/labs. I had more of a focus on elements, the more labs I did, because most of the labs (later) had stems of a similar length.

After that, I set up my system for the following elements: tetraloops, 1-3 loops and 2-4 loops.

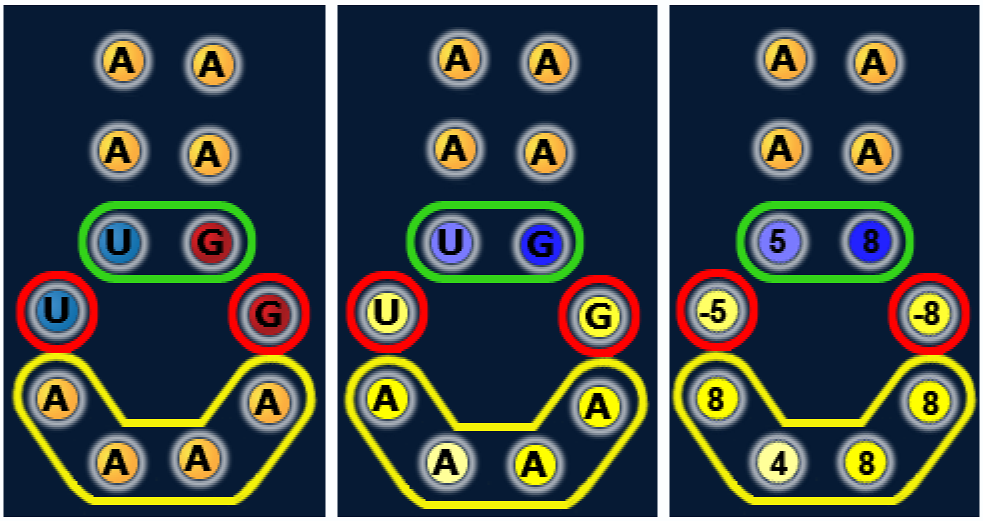
Extra Penalties

Closing pairs seem to have a stronger effect on the design than pairs buried in the middle of a stack. I want to penalize closing pairs more based on looking at the spreadsheet numbers on multiloops, and on my intuition. The numbers were higher than I would like for a few designs with mispaired closing pairs. I would want to penalize closing pairs with minus scoring bases more than other bases placed at less important spots. Longer strings are less prone to a few negatively scored bases.

Here is an example with a tetraloop in the shape it is supposed to fold up in. Notice the two yellow nucleotides marked by red in the shape data picture in the center. This closing basepair is not binding.



Below is how the element is actually folding when the data got back from lab. The broken tetraloop is forming a hexaloop instead.



Those nucleotides that are breaking apart in the red circles would need to get penalized harder, to reflect what breaking of the loop does to the lab score of the overall structure.

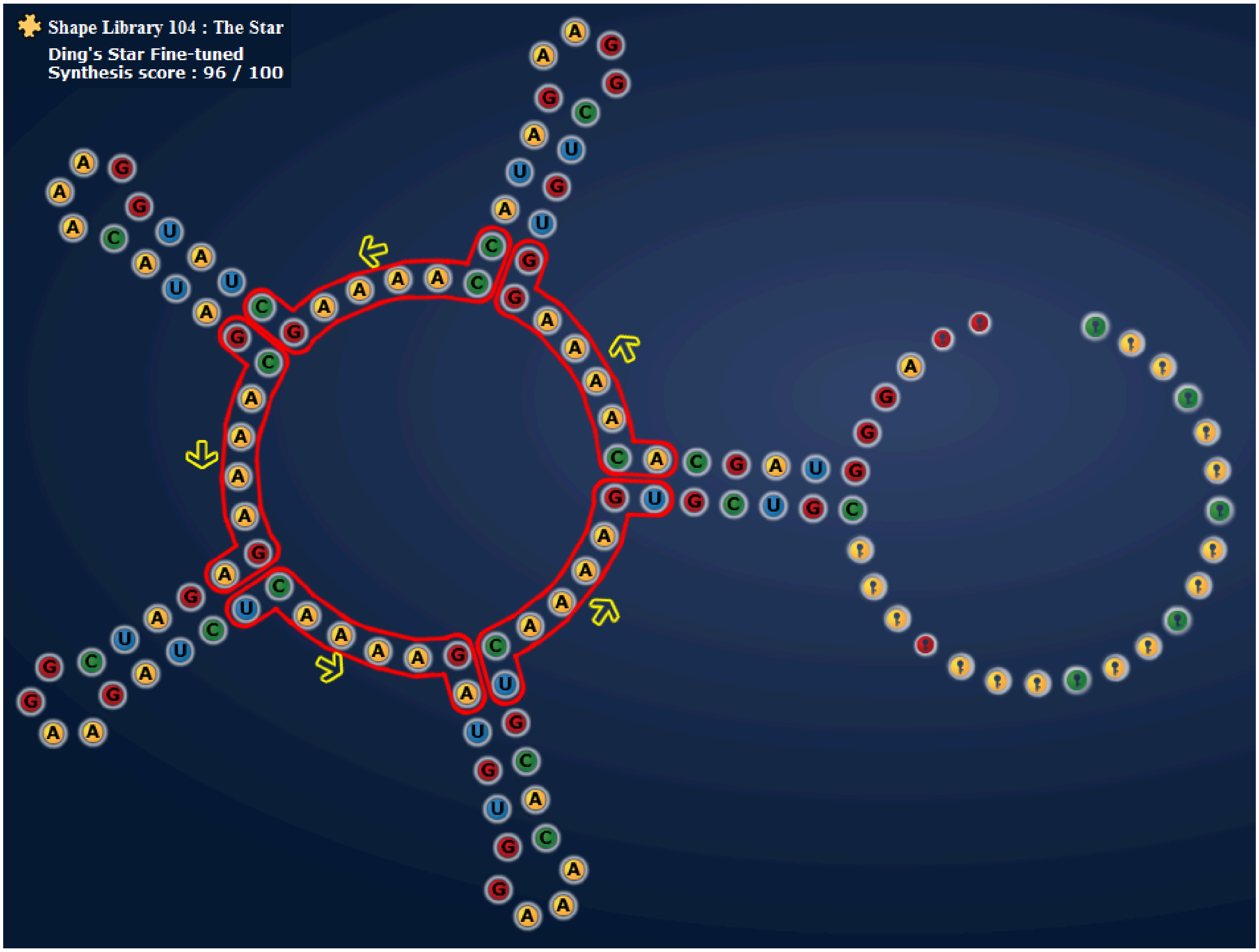
I also take into consideration how the design scored in lab and how the same sequence did on other designs, if available.

Finding Patterns with the Scoring System (this section written by Eli Fisker)

Mat asked me (Eli) to take a look at the spreadsheet where his strategy scored the multiloops in The Star lab. This spreadsheet contains pictures with the shape data of all the multiloops from the Star lab that scored 90% and above. The sequences were arranged in order of how they were scored as being the most stable and perfectly folding multiloops, according to the shape data.

What I saw was that Mat's system has picked up the pattern of Right orientation of GC-pairs in multiloops. The top scoring half follow the pattern perfectly. (See picture with sequences, two pictures down.) The green and red columns mark where the closing bases of the multiloop will appear, and their color if they follow the right direction pattern.

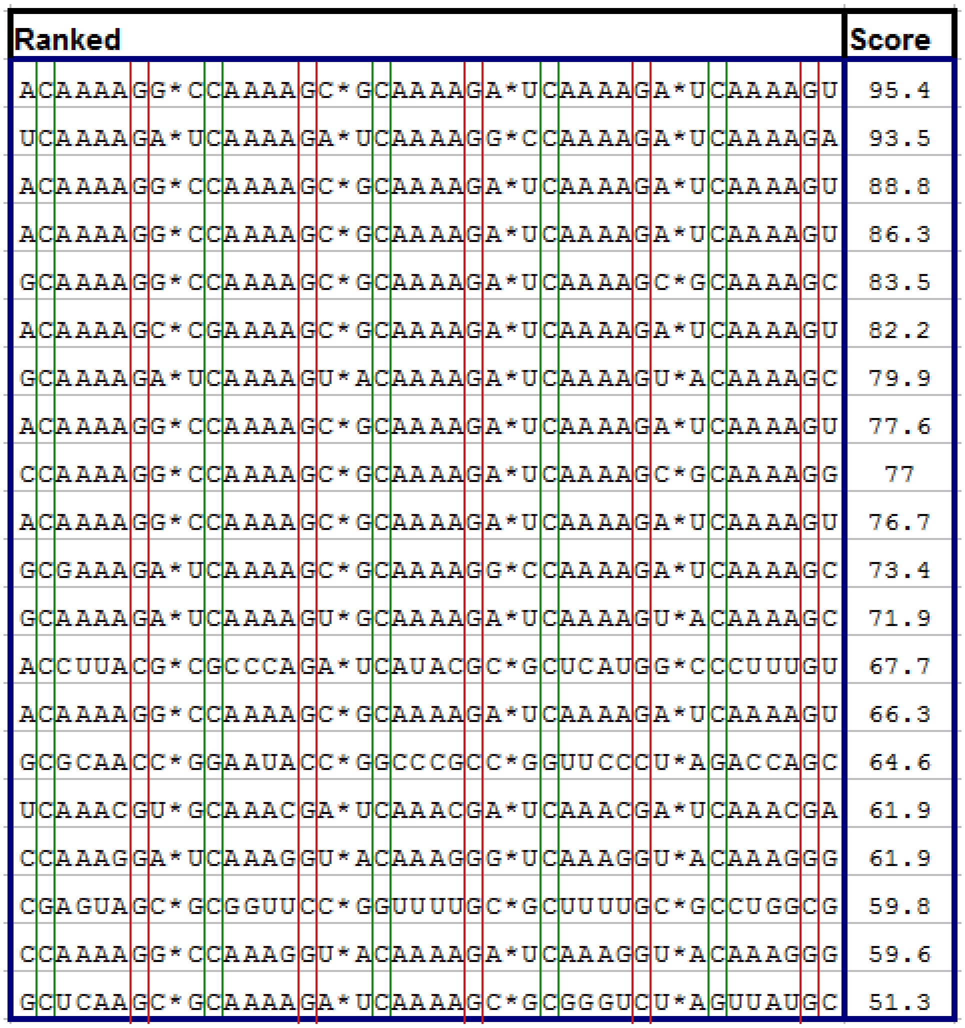
Below is the explanation for reading the coming list of sequences. The section breaks are marked with * in the sequence list. And the arrows point out the reading direction.



Shape Library 104 : The Star

Ding's Star Fine-tuned

Synthesis score : 96 / 100

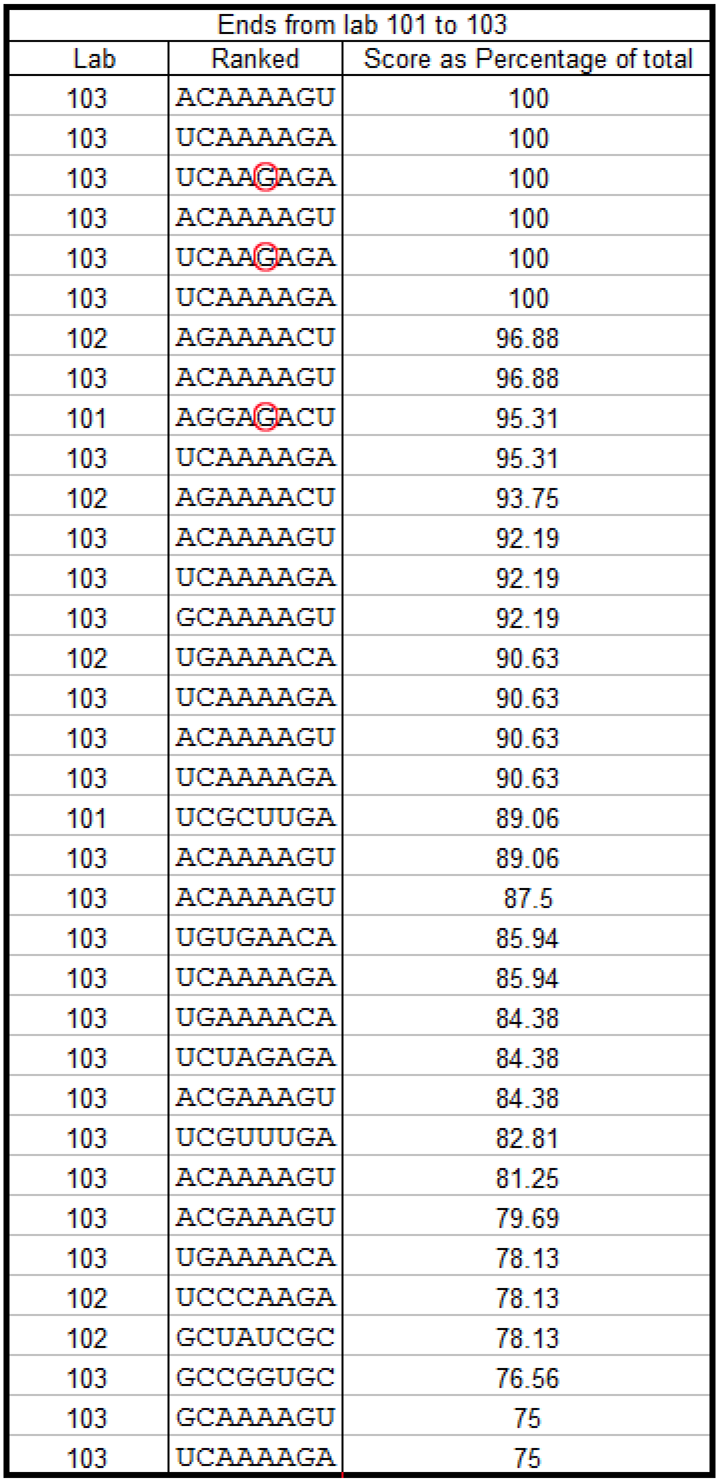


Ranked	Score
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	95.4
UCAAAAGA*UCAAAAGA*UCAAAAGG*CCAAAAGA*UCAAAAGA	93.5
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	88.8
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	86.3
GCAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGC*GCAAAAGC	83.5
ACAAAAGC*CGAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	82.2
GCAAAAGA*UCAAAAGU*ACAAAAGA*UCAAAAGU*ACAAAAGC	79.9
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	77.6
CCAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGC*GCAAAAGG	77
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	76.7
GCGAAAGA*UCAAAAGC*GCAAAAGG*CCAAAAGA*UCAAAAGC	73.4
GCAAAAGA*UCAAAAGU*GCAAAAGA*UCAAAAGU*ACAAAAGC	71.9
ACCUUACG*CGCCCAGA*UCAUACGC*GCUCAUGG*CCCUUUGU	67.7
ACAAAAGG*CCAAAAGC*GCAAAAGA*UCAAAAGA*UCAAAAGU	66.3
GCGCAACC*GGAAUACC*GGCCCGCC*GGUUCCCU*AGACCAGC	64.6
UCAAACGU*GCAAACGA*UCAAACGA*UCAAACGA*UCAAACGA	61.9
CCAAAGGA*UCAAAGGU*ACAAAGGG*UCAAAGGU*ACAAAGGG	61.9
CGAGUAGC*GCGGUUCC*GGUUUUGC*GCUUUUGC*GCCUGGCG	59.8
CCAAAAGG*CCAAAGGU*ACAAAAGA*UCAAAGGU*ACAAAGGG	59.6
GCUCAAGC*GCAAAAGA*UCAAAAGC*GCGGGUCU*AGUUAUGC	51.3

The sequences are from this table of results, located at B 42.

Another pattern stood out strongly too. Watch the picture above again. This time focus on the middle row between the red and green columns. Notice the all A pattern that stands out in the stablest and highest scoring multiloop elements. The most successful version of the multiloop ring is clearly yellow.

In the spreadsheet on Lab 101 and 103 Tetraloops, another pattern stands out clearly. What I see when looking from the top and going down, is the overall appearance of mainly A's in the middle section. That is the all A pattern that Mat successfully has been using for solving tetraloops in the single shape labs. See Pictures on successful loop sequences. Most of the highest scoring loops are the all A ones.



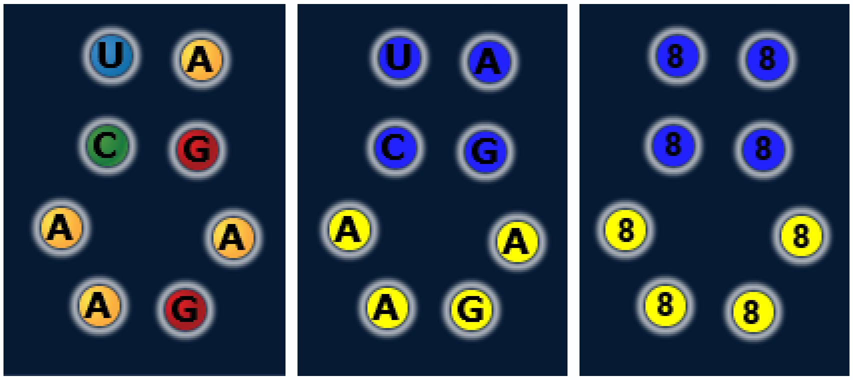
Ends from lab 101 to 103

Lab	Ranked	Score as Percentage of total
103	ACAAAAGU	100
103	UCAAAAGA	100
103	UCAAGAGA	100
103	ACAAAAGU	100
103	UCAAGAGA	100
103	UCAAAAGA	100
102	AGAAAACU	96.88
103	ACAAAAGU	96.88
101	AGGAGACU	95.31
103	UCAAAAGA	95.31
102	AGAAAACU	93.75
103	ACAAAAGU	92.19
103	UCAAAAGA	92.19
103	GCAAAAGU	92.19
102	UGAAAACA	90.63
103	UCAAAAGA	90.63
103	ACAAAAGU	90.63
103	UCAAAAGA	90.63
101	UCGCUUGA	89.06
103	ACAAAAGU	89.06
103	ACAAAAGU	87.5
103	UGUGAACA	85.94
103	UCAAAAGA	85.94
103	UGAAAACA	84.38
103	UCUAGAGA	84.38
103	ACGAAAGU	84.38
103	UCGUUUGA	82.81
103	ACAAAAGU	81.25
103	ACGAAAGU	79.69
103	UGAAAACA	78.13
102	UCCCAAGA	78.13
102	GCUAUCGC	78.13
103	GCCGGUGC	76.56
103	GCAAAAGU	75
103	UCAAAAGA	75

Another type of loop stands out among the top scoring loop patterns, with G's differently positioned compared to the more typical player loop GAAA. The tetraloop sequence AAGA stands out (marked by red circles) .

Here is an example on this new AAGA loop, *Dimension9's* design *JewelOfTheMoon* (97%) from lab 103. It is one of the loops with a perfect score of as percentage of total.

Sequence UCAAGAGA, score 100 as percentage of total.



Purpose of Computationally Selected Elements

Basically this Computationally Selected Elements tool can be used to pick up patterns for what works and makes an element stable. This can aid the designing process for us and have the potential to save computation time for our robot, because this method can pick up some most stable elements from a pattern library.

So far I have only converted elements with the first two pairs of bases on the stem into numbers, but it would be interesting to look at good loop patterns, depending on the length of the stems.

The amount of data coming our way will continue to grow. One purpose for this tool was to make it possible to search through and analyze large amounts of data in ways that could mimic a player's analysis style, as this data filter can rank elements in order of how they have done according to the shape data.

I also have a simple strategy that I was using for my later lab single state designs, made up of strategies plus overall rules, on how to use an overall designing strategy together with this analysis tool.

I had the idea for the project and asked Eli for help with the presentation. We discussed the system and Eli has added a section of analysis. Mike and Starryjess helped with the grammar errors and here is the result.

Datasheets

Scored Multiloop

Lab 104 The Star - Scored Multiloop

Scored Tetraloops

Lab 103 - One Bulge Cross - Scored Tetraloops

Lab 102 - The Cross - Scored Tetraloops

Lab 101 The Finger - Scored Tetraloops

Scored Elements

Lab 104 The Star - Scored Elements

Lab 103 One Bulge Cross - Scored Elements

Lab 102 The Cross - Scored Elements

Lab 101 The Finger - Scored Elements

Element by score

Lab 101 and 103 Tetraloops

Reference sheets

Reference sheets 104 - The Star

Reference sheets 103 - One Bulge Cross

Reference sheets 102 - The Cross

Reference sheets 101 - The Finger