1.
- Today I will talk about my project.
- The topic of my project is about development of a new structure abstraction, we call it as "helices indices of RNAs"
- If you have some questions during my presentation, feel free to interrupt me.

2.
I've divided my presentation into 4 parts
- in the first part, I will talk about the basic secondary structure elements of RNA and introduce a classical RNA secondary structure prediction algorithm
- because the concept of "helices indices of RNAs" is based on the concept of "abstract shapes", that was developed by Bjoern, so I will introduce
the concept of "abstract shapes" in the 2nd part.
- give the basic idea about the helices indices.
- in the last part, I will give an outlook of the project

3.
- In this slide, I will talk about the basic secondary structural elements of RNA,
- As the figure showed, RNA is composed of 2 kinds of elements, like 6 种.
  the first kind is energy favorable part, it includes dangling end and base stacking
- such energy favorable elements lead to formation of energy unfavorable elments, they are 4 种.
- all double stranded regions are also called as as 'helices', it is always related with a loop type.
  for example we call this part as helics of hairpin loop, this part as helices of multiloop, this part as helices of bulge loop, this part as helices of internal loop.
- the concept is very important, because we will develop a new methods
  that based on the helices. I will explain it later in this presentation in more detail.

4.
- In this slide, I will introduce the classic algorithm for RNA secondary structure prediction.
  namely Zuker algorithm. The algorithm was first described by Zuker and Stiegler in 1981.
- the basic idea is sequence can be folded into many different secondary structures.
  furthermore, we can calculate for every secondary structure a unambiguous free energy value by adding the free energies of all elements of the secondary structure.
  after that, the algorithm choose the structure with the minimum free energy.
- The runtime of the algorithms is O(n3) where n is the length of sequence,
- From this argorithm, we can get only one solution.

5.
- To better understand it, I prepared a small example,
- as I just said, in Zuker algorithm, for every secondary structure we calculate a free energy value by ...
- as the figure shows, the secondary structure consists of a hairpin loop, dangling end, 2 base stacking, a bulge loop,
- among them, the hairpin loop and dangling end are energy favorable, the have the negative free energy, and in contrast, hairpin loop and bulge loop are energy unfavorable, they have positive free energy.
- after addition of all elements, we get a free energy for the whole secondary structure, namely -4.6 kcal/mol.

6.
- if we calculate free energy for every secondary structure and draw it on a graph, we get a landscape.

- as the figure showed, the x-axis means conformations,
the y-axis means free energy,
if a RNA molecule don't have conformational switch, typically, if we see it from the landscape, the energy difference between minimum free energy and energy of the first sub-optimal confirmation is enough big.
- In Zuker algorithm, we calculate only this point.

7.
- aber in practice, the native structure is not always the one 照 读...
- what can we do to find the native structure? One of solutions is that we define a range and enumerate all suboptimal 照 读...

8.
- the figure shows the idea I just explained last slide, firstly, we define a linie in the energy landscape and calculate all suboptimal structure under the linie.
- but it caused another problem, namely the number of suboptimal structures grows exponentially with the size of the energy range.
- How can we solve the problem?

9.
- 照读: One of the solution is further classify ...
abstract shape is one solution in the direction.
it was developed by ... initial ideas ...
- central to this approach is ...
for example in abstract shape type 5, all loops excepts for internal loop and multiloop are abstracted.
- each shape has a representative structure called shrep. It is the structrue with the minimum free energy within the shape class.

10.
- this figure shows an output example from abstract shapes by setting an envergy range of 5 kcal/mol above the mfe.
- the picture above is 2 dimension representation, the picture below is the textual output.
- we notice the 2 dimension representation and abstract shapes are very similar, if we image the first opening bracket and last closing bracket as a stem of tree. and all inner brackets pairs are branches of the trees.
- the abstract shapes are identical as the representation above.
- in the middle is the secondary structure of shrep within the shape class
on the right side is the free energy of shrep

11.
- the abstract shape works pretty well, but one drawback is still there. namely 照读.
- in other words, it doesn't contain any information about the position in abstract shapes.
- What is consequence of the drawback?
- To explain the drawback, I prepared an example, as the figure shows, although the abstract shape above and below are identical, but position of the hairpin loop are totally different.
- the drawback make 照读...

12.
- for the reason we will develop a new structure abstraction namely helices indices. It includes the information of positions of helices.
- to develop it, 2 things have to been decided,
- the first thing is Which secondary structure element should be recorded?
- we have 4 candidate ...
- the second point is which position of this element should be recorded?
- we have also 4 candidates,
  i is first position of helix, j is last position of helix,
  we can also record the both positions at the same time or the central position of them.

13.
- To better understand it, I prepared a small example,
- In this example, We record only hairpin loop and use the central position
- So as the figure showed, the structure is composed of 2 helices, namely this one and this one.
  because we abstract from the bulge loop, so we don't consider this part.
- the position of i is 8, j is 13, i+j/2 is 10.5, similarly, k is 35 and l is 41, and k+l/2 is 38. Therefore, this structure would be abstracted to [10.5, 38].

14.
- the slide showed 照读
- in this example, we record multiloop and hairpin loop and use the central position
- the most interesting information from it is 3 records that are marked with asterisk
- on the left side, it is the free energy of shrep, namely this record, this record and this record
  in the middle, it is the secondary structure of the shrep
  on the right side, it is helices indices
- if we compare the 3 records, we observe the helices indices of this record is a combination of the helices indices of the first 2 records.
- what does it mean? ths answer is in the next slide.

15.
- we can observe the first record as the global minimum structure and the second record as a suboptimal structure. They have almost the same free energy.
- furthermoare, we can observe the combination record as a saddle point in energy landscape.
- If we will move from this point to this point, we have to pass the saddle point. Because the free energy of all the helices indices are known, therefore, we can calculate barrier energy between saddle point and mfe point easily.
- in our example, （返回去）, the minimum is -10.7, saddle point is -2.3, and the barrier energy is therefore 8.4 kcal/mol

16.
- from this slide, I will point out one serious problem we might encountered, namely the runtime problem.
- the figure shows 照读
  x axis is sequence length, y axis is number of helices indices or abstract shapes
  the linie means the helices indices space, the linie means the abstract shapes space.
- we notice, 照读 it will cause serious calculation runtime problem. For example, it needs about 30 hours to calculate the helices indices for a 72 nt long sequence on our super computer.
- how can we solve this problem?

17.
- one solution is that we can set an energy range to limit the helices indices space.
- the figure shows the evaluation result by setting different energy ranges.
  x axis is sequence length, y axis is number of helices indices
  the linie means the helices indices without energy limitation
  ... with a energy limit of 20 kcal/mol
  .. 15, 10, 5
- although all helices indices space grow exponentially with sequence length, but the degree of steepness are different, the stricter the limitation, the flater the linie and the better the runtime will be.
- so if we set a strict limitation on the calculation, we will get a better runtime and solve the problem.

18.
- in last slide, I will give an outlook of the project.
- at first, we will develop a new structure abstraction, namely the helices indices
- after that, the idea will be implemented based on the idea.
- afterwards, the software will be evaluated by benchmark program
- lastly, we will design a RNA class predictor

19.
Thank you for your attention!
Are there any questions?

questions:

1. How was implemented the idea in detail?

dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems.

The key idea behind dynamic programming is quite simple. In general, to solve a given problem, we need to

solve different parts of the problem (subproblems), then combine the solutions of the subproblems to reach an overall solution.

Often, many of these subproblems are really the same. The dynamic programming approach seeks to solve each subproblem only once,

thus reducing the number of computations. This is especially useful when the number of repeating subproblems is exponentially large.

(Page 40 from Diss)

2. Why is mfe structures not the native one?

This can be explained by the existence of modified bases in native tRNAs, which leads to the formation of a structure that is not

the optimal under the energy model used.

3.

It exists a software, it can analyse the folding space completely. Only the runtime of the program is very bad.

This is also the reason why we will develop a new program because we will improve the runtime.

4.

- In this algorithm, we have to assume one condition, namely we get away the knots from the calculation.

  The reason why we can get away the knots from the calculation is the calculation result of secondary structure prediction are used to infer

  3-dimensional structures and knots can be inferred at this stage as well.