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.
- (advantage) The runtime of the algorithms is O(n3) where n is the length of sequence,
- (disadvantage) 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,
normally, if a RNA molecule don't have conformational switch, the energy difference between minimum free energy and energy of the first sub-optimal confirmation is enough big, so that the mfe-point can not go up and reach the point.
- 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 line in the energy landscape and calculate all suboptimal structure under the line.
- 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.
- one is the solution is we can further classify secondary structures space
- abstract shapes is one approach in this direction
- the initial idea is the user is only interesting in structures that show fundamental differences
- small change like bulge loop and internal loops are not every important.
- the idea is we can abstract from such unimportant elements, for example, bulge loop and internal loops
- last point is every shape has a representative, it is also called as shrep, it is minimum free energy within 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 text form.
- from the intuitive perspective, the abstract shape is similar than the 2 dimension version. for example in record 3, as seen from the 2 dimension version, it is a cloverleaf, correspondingly, the abstract shape is a pair brackets that nest 3 pairs inner brackets.

11.
- the abstract shape works pretty well, but one drawback is still there. namely 照读.
- 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.
- this is the reason why 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. To differ from multiloop and hairpin loop, we attach a letter 'm' to the helices indices of multiloop. for example, in the record, 30m means the position comes from a multiloop.
- the most interesting information behind the list are 3 records, namely this one, this one and this one
- if we compare the 3 records, we can find the helices indices of this record is a combination of the first 2 records.
- what does it mean? ths answer is in the next slide.

15.
- if we draw the 3 records in energy landscape, it must be like this.
- the first record is global minimum structure and the second record is a local minimum. They have almost the same free energy.
- furthermoare, the combination records is a saddle point
- so if we move from this point to this point, we have to pass the saddle point. Because the free energy of all 3 helices indices are known, therefore, we can calculate the difference easily. the difference is barrier energy.
- in our example, （返回去）, the mfe-point has energy 10.7, this point has energy 2.3, and the barrier energy is about 8.4

16.
- from this slide, I will talk about the first problem we encountered, this is a runtime problem.
- the figure shows 照读
  in the figure, x axis is sequence length, y axis is number of helices indices or abstract shapes
  the line means the helices indices space, the line means the abstract shapes space.
- we notice, 照读, therefore it caused a runtime problem. In practice, for a sequence with 72 nucleotides, we need about 32 hours to calculate the helices indices on our super computer, it is not satisfactory.
- how can we solve this problem?

17.
- one solution is that 照读,
- the figure shows 照读
- there are altogether 5 lines in the figure
  the first line means the helices indices without energy limitation
      second is with a energy limit of 20 kcal/mol, this is with 15, this is with 10, this is with 5.
- although all helices indices spaces grow exponentially with sequence length, but the degree of steepness are different,
- the tendency is the lower the energy range (is), the flater the line (is) and the better the runtime will be.
- so if we set a lower enery range 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.
- of course, 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.