Разработка и оптимизация тестирующей системы и задач по биоинформатике в рамках платформы Розалинд

Алексей Кладов, 445 группа
Руководитель: Николай Вяххи

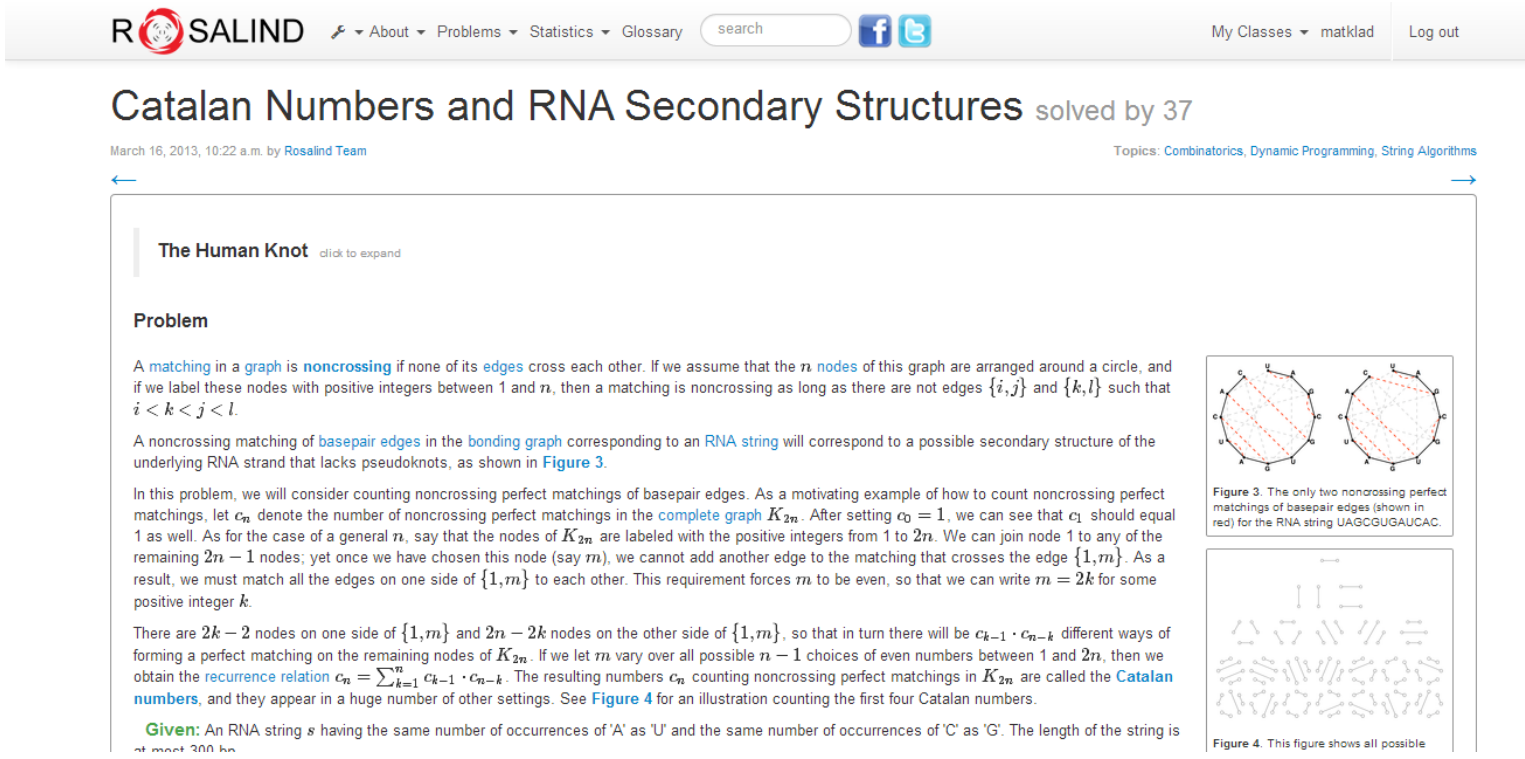
СПбГУ

aleksey.kladov@gmail.com

30 мая 2013 г.

Обзор Rosalind

- e-learning
- Биоинформатика
- Решение задач
- Геймификация



ROSALIND About Problems Statistics Glossary search My Classes matklad Log out

Catalan Numbers and RNA Secondary Structures solved by 37

March 16, 2013, 10:22 a.m. by Rosalind Team

Topics: Combinatorics, Dynamic Programming, String Algorithms

The Human Knot click to expand

Problem

A matching in a graph is **noncrossing** if none of its edges cross each other. If we assume that the n nodes of this graph are arranged around a circle, and if we label these nodes with positive integers between 1 and n, then a matching is noncrossing as long as there are not edges $\{i,j\}$ and $\{k,l\}$ such that $i < k < j < l$.

A noncrossing matching of basepair edges in the bonding graph corresponding to an RNA string will correspond to a possible secondary structure of the underlying RNA strand that lacks pseudoknots, as shown in Figure 3.

In this problem, we will consider counting noncrossing perfect matchings of basepair edges. As a motivating example of how to count noncrossing perfect matchings, let c_n denote the number of noncrossing perfect matchings in the complete graph K_{2n}. After setting $c_0 = 1$, we can see that c_1 should equal 1 as well. As for the case of a general n, say that the nodes of K_{2n} are labeled with the positive integers from 1 to $2n$. We can join node 1 to any of the remaining $2n-1$ nodes; yet once we have chosen this node (say m), we cannot add another edge to the matching that crosses the edge $\{1,m\}$. As a result, we must match all the edges on one side of $\{1,m\}$ to each other. This requirement forces m to be even, so that we can write $m = 2k$ for some positive integer k.

There are $2k-2$ nodes on one side of $\{1,m\}$ and $2n-2k$ nodes on the other side of $\{1,m\}$, so that in turn there will be $c_{k-1} \cdot c_{n-k}$ different ways of forming a perfect matching on the remaining nodes of K_{2n}. If we let m vary over all possible $n-1$ choices of even numbers between 1 and $2n$, then we obtain the recurrence relation $c_n = \sum_{k=1}^{n} c_{k-1} \cdot c_{n-k}$. The resulting numbers c_n counting noncrossing perfect matchings in K_{2n} are called the **Catalan numbers**, and they appear in a huge number of other settings. See Figure 4 for an illustration counting the first four Catalan numbers.

Given: An RNA string s having the same number of occurrences of 'A' as 'U' and the same number of occurrences of 'C' as 'G'. The length of the string is at most 300 bp.

Figure 3. The only two noncrossing perfect matchings of basepair edges (shown in red) for the RNA string UAGCGUGAUCAC.

Figure 4. This figure shows all possible

Rosalind achievements

- Грант Start Fellows
- 6000 пользователей
- 120(80 сделал я) задач
- 20 классов (один из них мой)

Проблемы и их решения

- Имплементация задач и улучшение тестирующей системы
- Всё медлiенно – распараллелил тесты
- Всё медленно – добавил возможность писать задачи не только на Python, но и на C++
- Не ортагональное API для задачи – исправление в тестирующей системе + планы на будущее
- Неправильные ссылки – валидация условий
- Некому тестировать задчи – бета-тестирование
- Много задач и непонятно, что на какой стадии и как должно работать – добавленны вкусности в тестирующую систему
- Биоинформатические бибилиотеки такие биоинформатические – баг репорты, работа с напильником

Некоторые цифры

	Задача	Python	C++
1	GAFF	20 с	0.01 с
2	QRTD	163 с	0.03 с
...	...	...	...
30	...	много	мало

- Параллелизм — ускорение тестов с 8 минут до 2.
- Проверки — обнаружение проблем в 4 уже опубликованных задачах.

Что дальше?

- Новый проект
- Больше областей
- Больше геймификации
- Больше википедии
- Больше типов задач, хороших и разных
- Больше датамайнинга

Результаты

- Решены проблемы с производительностью
- Улучшенна 'диагностика' задач
- Разрабатывается структура textbook для нового проекта
- Прототипируются отдельные части новой системы