

MAC0499 - Trabalho de Formatura Supervisionado  
Proposta para Monografia

**Subseqüência comum máxima sem repetições**

Christian Tjandraatmadja - Número USP: 5122760  
Supervisora: Profa. Dra. Cristina Gomes Fernandes

## 1 Resumo

O problema da subseqüência comum máxima é um problema clássico em computação. Uma de suas variantes é o problema da subseqüência comum máxima sem repetições, abreviado RFLCS, do inglês *repetition free longest common subsequence*. O RFLCS é definido da seguinte forma: dadas duas seqüências  $s$  e  $t$ , encontrar uma subseqüência comum de  $s$  e  $t$  tal que seus símbolos sejam dois a dois distintos e seu comprimento seja máximo. Neste projeto de iniciação científica, esse problema será abordado do ponto de vista de algoritmos de aproximação e de programação inteira.

A motivação para o estudo do RFLCS tem base em Biologia Computacional. Podemos considerar seqüências como genomas, compostas de genes. Subseqüências comuns de genomas modernos podem apontar para uma evolução independente destes a partir de um mesmo genoma ancestral. A primeira abordagem que considera subseqüências sem repetições é a de Sankoff [2] (veja também Bonizzoni [1]), que propôs um modelo de ordenação de um genoma por reversões no qual para cada família de genes queremos encontrar no máximo um representante. Para esta aplicação, é interessante a variante do RFLCS em que permitimos reversões de trechos das seqüências.

## 2 Objetivos

O objetivo desse projeto de iniciação científica é estudar o problema do RFLCS. Para isso, tanto algoritmos de aproximação como algoritmos baseados em técnicas de combinatória poliédrica serão estudados. Um estudo do politopo associado ao RFLCS será feito para conhecer melhor o problema e aplicar esse conhecimento nos algoritmos. Em particular, um algoritmo exato baseado na técnica de programação inteira *branch and cut* será estudado. Esses algoritmos serão implementados e analisados experimentalmente. Além disso, serão estudados casos do RFLCS em que existem algoritmos melhores. Uma variante do problema em que se permite reversões também será estudado se o tempo permitir.

### 3 Atividades já realizadas

Para entender melhor o problema, foram observadas algumas abordagens para o problema do RFLCS. Isso serviu de base ao estudo dos algoritmos.

Alguns algoritmos de aproximação para o problema do RFLCS foram estudados e implementados. Esses algoritmos foram testados com seqüências aleatórias de diversas classes e foi feito uma breve observação dos comportamentos desses algoritmos.

Implementou-se também uma modelagem de programação inteira do RFLCS usando uma biblioteca para solução de problemas de programação linear (**glpk**) para resolvê-la, mas no momento isso é feito sem nenhum conhecimento do problema além da modelagem.

### 4 Cronograma de atividades

Ativ/Mês	JUL	AGO	SET	OUT	NOV
1	✓	✓			
2		✓	✓	✓	
3			✓	✓	
4			✓	✓	✓
5				✓	✓
6				✓	✓
7					✓
8	✓	✓	✓	✓	✓

#### Legenda:

1. Estudo e implementação de algoritmos de aproximação para o RFLCS.
2. Estudo de um politopo associado ao RFLCS.
3. Busca de algoritmos melhores para casos particulares do RFLCS.
4. Implementação do algoritmo exato baseado no método *branch and cut*.
5. Realização de testes experimentais.
6. Extensão dos resultados para incorporar uma reversão.
7. Preparação do poster e apresentação do trabalho.
8. Preparação do texto da monografia.

## 5 Estrutura esperada da monografia

A monografia terá uma parte técnica e uma parte subjetiva. Ela será dividida nos tópicos listados abaixo, mas ela pode ser alterada conforme se prossegue a disciplina.

### 5.1 Parte técnica

- **Introdução:** motivação ao estudo do RFLCS, descrição do problema e definições.
- **Abordagens:** algumas abordagens observadas sobre o RFLCS.
- **Algoritmos de aproximação:** um estudo sobre alguns algoritmos de aproximação para o RFLCS, detalhes sobre suas implementações e resultados experimentais.
- **Politopo do RFLCS:** um estudo sobre o politopo do RFLCS.
- **Algoritmo exato:** um estudo sobre um algoritmo exato para o RFLCS baseado na técnica *branch and cut*, e como os algoritmos de aproximação e o estudo do politopo foram usados para o desenvolvimento desse algoritmo. Detalhes sobre sua implementação e resultados experimentais.
- **Casos particulares:** um estudo sobre casos particulares do RFLCS em que haja algoritmos melhores.
- **RFLCS com reversão:** um estudo de como estender o projeto para que se inclua uma ou mais reversões. Essa seção depende da disponibilidade de tempo durante a disciplina.
- **Conclusão:** conclusão sobre a parte técnica.
- **Bibliografia**

### 5.2 Parte subjetiva

- **Experiência pessoal:** desafios, frustrações e outras experiências sobre o projeto.
- **Conceitos aprendidos no curso:** uma lista de disciplinas cursadas relevantes para o projeto e observações sobre as aplicações dos conceitos estudados em cada uma delas no projeto.
- **Conclusão:** passos para aprimorar os conhecimentos da área se continuar atuando nela e palavras finais.

## Referências

- [1] P. Bonizzoni, G. Della Vedova, R. Dondi, G. Fertin, and S. Vialette. Exemplar longest common subsequence. In 3992, editor, *Proceedings of IWBRA*, Lecture Notes in Computer Science, pages 622–629. Springer, 2006.
- [2] D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.