O Problema da Subseqüência Comum Máxima sem Repetições

Christian Tjandraatmadja (christj@ime.usp.br)
Supervisora: Cristina Gomes Fernandes (cris@ime.usp.br)
Orientador: Carlos Eduardo Ferreira (cef@ime.usp.br)

Instituto de Matemática e Estatística
Universidade de São Paulo
Apoio financeiro: FAPESP (processo 07/54282-6)

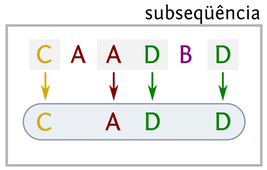
Introdução

Objetivo: Estudar o problema do RFLCS, definido a seguir, e implementar um algoritmo para a sua resolução.

Considere seqüências s e t sobre um alfabeto.

Uma subseqüência de s é obtida escolhendo elementos de s de forma a manter a ordem relativa entre eles.

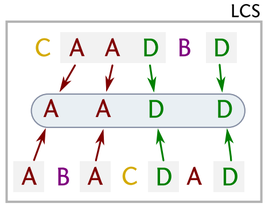
subseqüência



Uma subseqüência comum de s e t é uma subseqüência tanto de s e como de t.

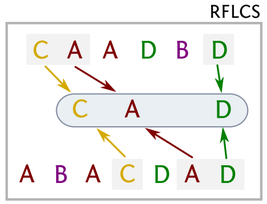
Dizemos que tal subseqüência é um LCS (*longest common subsequence*, ou subseqüência comum máxima) se ela tem comprimento máximo.

LCS



Dizemos que ela é um RFLCS (*repetition free longest common subsequence*, ou subseqüência comum máxima sem repetições) se ela tem comprimento máximo entre as que não têm mais de uma ocorrência de um mesmo símbolo.

RFLCS



O problema em que estamos interessados é, dadas duas seqüências, encontrar um RFLCS delas.

Abordagens principais:
- Programação inteira
- Combinatória poliédrica

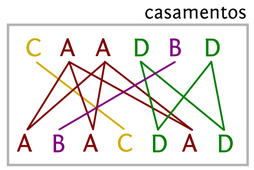
Aplicações:
- LCS: Diferenciação de arquivos, compressão de dados
- LCS e RFLCS: Biologia Computacional (similaridade entre genomas)

Complexidade:
- LCS: Existe algoritmo polinomial (para número fixo de seqüências)
- RFLCS: NP-difícil

LCS

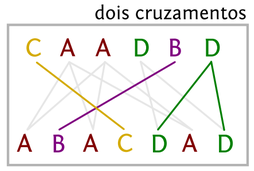
Definimos casamento como um par (i,j) que liga o i-ésimo elemento de s ao j-ésimo elemento de t se eles são do mesmo símbolo.

casamentos



Dizemos que dois casamentos (i,j) e (k,ℓ) se cruzam se (i ≤ k e j ≥ ℓ) ou (k ≤ i e ℓ ≥ j).

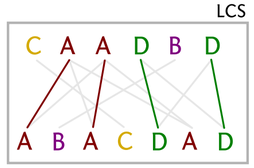
dois cruzamentos



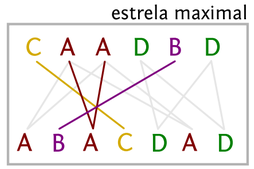
Um LCS pode ser visto como uma escolha do maior número de casamentos que não se cruzam, pois um cruzamento troca a ordem relativa entre os elementos.

Definimos estrela como um conjunto de casamentos que se cruzam dois a dois. Dizemos que ela é maximal se não há outro casamento fora do conjunto que cruza com todos do conjunto. Observe que o LCS pode ser visto como escolher o maior número de casamentos tal que não haja dois em uma mesma estrela maximal.

LCS



estrela maximal



Assim, podemos formular o problema do LCS como um problema de programação inteira da seguinte forma:

Sejam C o conjunto de casamentos e $z \in \{0,1\}^C$ o vetor em que $z_{ij} = 1$ se e somente se escolhemos o casamento (i,j).

$$\max \sum_{(i,j) \in C} z_{ij}$$ → maximizar a escolha de casamentos de forma que

$$\text{sujeito a} \sum_{(i,j) \in S} z_{ij} \leq 1 \quad \text{para toda estrela maximal } S$$ → para cada estrela maximal, podemos selecionar no máximo 1 casamento,

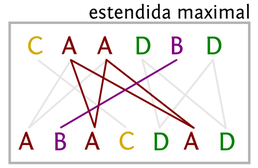
$$z_{ij} \in \{0,1\} \quad \text{para todo casamento } (i,j)$$ → cada casamento é escolhido 0 ou 1 vezes.

Podemos provar que, do ponto de vista de combinatória poliédrica, não existe formulação melhor.

RFLCS

No RFLCS, surge a restrição de que existe no máximo uma ocorrência de cada símbolo na solução. Isto é, não podemos escolher mais de um casamento associado ao mesmo símbolo.

estrela estendida maximal



Definimos então estrela estendida como um conjunto de casamentos que, dois a dois, ou se cruzam ou estão associados ao mesmo símbolo. Note que, em particular, os conjuntos de todos os casamentos de mesmo símbolo são estrelas estendidas.

Podemos obter, então, a seguinte formulação:

Sejam C o conjunto de casamentos e $z \in \{0,1\}^C$ o vetor em que $z_{ij} = 1$ se e somente se escolhemos o casamento (i,j).

$$\max \sum_{(i,j) \in C} z_{ij}$$ → maximizar a escolha de casamentos de forma que

$$\text{sujeito a} \sum_{(i,j) \in S} z_{ij} \leq 1 \quad \text{para toda estrela estendida maximal } S$$ → para cada estrela estendida maximal, podemos selecionar no máximo 1 casamento,

$$z_{ij} \in \{0,1\} \quad \text{para todo casamento } (i,j)$$ → cada casamento é escolhido 0 ou 1 vezes.

O algoritmo implementado para o projeto se baseia na formulação acima. Como o número de estrelas estendidas maximais é exponencial no tamanho do problema, as restrições de estrelas estendidas são adicionadas conforme elas são violadas durante a execução do algoritmo.

Resultados

- Conhecimento da estrutura dos problemas do LCS e do RFLCS e das suas caracterizações poliédricas;
- Implementação de um algoritmo para a resolução do problema do RFLCS baseado na técnica *branch and cut* com o uso do pacote glpk.