Università degli Studi di Milano Bicocca

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

Integer Linear Programming approaches on the DNA recombination problem

Relatore: *Bonizzoni Paola*

Co-relatore: *Della Vedova Gianluca*

Relazione della prova finale di:

Antonio Vivace

Matricola 793509

Anno Accademico 2016-2017

Abstract

We introduce the *Computational Biology* field and familiarise with *Integer Linear Programmming*, defining its inception, uses and approach, and how ILP-based approaches have become a standard optimization technique in bioinformatics, reviewing some applications.

Then, we formalise the "DNA Recombination and Rearrangement" problem based on the what is observed in some species of ciliates, followed by an analysis and report of some of existent approaches and their central ideas, limitations and reductions applied.

Finally, an ILP formulation of the DNA Recombination problem is given, describing the implementative tools used and the main encountered difficulties.

Contents

1 — Introduction

1.1 Computational Biology

Computational Biology is defined as the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems[1].

Some of the most important challenges in the field includes[2]:

- Protein structure prediction;

- Homology searches;

- Multiple alignment and phylogeny construction;

- Genomic sequence analysis and gene-finding.

In particular, *Computational Molecular Biology* (bioinformatics) focuses on studying existing and emerging approaches, techniques and algorithms for string computation (sequences) providing a significant intersection between computer science and molecular biology.

Note that the term *bioinformatics* is used also as an umbrella term for the (wider) body of biological studies using computer programming as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the field of genomics.

2 — Integer Programming

2.1 Definition

Linear programming is a technique for the mathematical optimization of a linear objective function, subject to linear equality and linear inequality constraints.

Linear programs are problems that can be expressed in canonical form as:

$$
\begin{aligned}
\text{maximize} \quad & \mathbf{c}^\mathrm{T}\mathbf{x} \\
\text{subject to} \quad & A\mathbf{x} \leq \mathbf{b} \\
\text{and} \quad & \mathbf{x} \geq \mathbf{0} \\
& (\mathbf{x} \in \mathbb{Z}^n)
\end{aligned}
$$

If the variables are forcibly constrained to be integers, we call the program *Integer* or *Integer Linear*.

0-1 integer programming or binary integer programming (*BIP*) is the special case of integer programming where variables are required to be 0 or 1 ($\mathbf{x} \in \{0, 1\}$).

In contrast to linear programming, which can be solved efficiently in the worst case, integer programming problems are in many practical situations (bounded variables) NP-hard. BIP are classified as NP-hard too ("01 integer programming" is one of the *Karp's 21 NP-complete problems*).

2.2 In Computational Biology

2.3 Examples

2.3.1 Problem 1

3 — The Problem

3.1 Biological Background

3.2 Formalisation

3.3 Existent Approaches/Solutions

3.4 ILP formulation

Bibliography

[1] NIH Biomedical Information Science and Technology Initiative Consortium. NIH working definition of bioinformatics and computational biology. `https://web.archive.org/web/20120905155331/http://www.bisti.nih.gov/docs/CompuBioDef.pdf`, 2000.

[2] D. Searls. Computational Methods in Molecular Biology.