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 Programming*, 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* (ILP) 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:

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

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

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

At its inception, the focus of Computational Biology was on the development of efficient algorithms and data structures that were able to deal with the data being introduced in life science applications. Lately, the introduction of high throughput methods for biomedical data analysis and the rise of Systems Biology (the study of systems of biological components) made Statistical Learning approaches a standard.

Furthermore, new and accessible sequencing methods caused the quantity of the data produced to grew exponentially.

This element and the fact that biological processes are usually reduced and studied as simulations (because the actual nature of them is still being investigated, as in the case of our problem) lead to the introduction of a lot new optimization problems in the field.

In most cases, these optimisazion problems are discrete ones: hence the success of ILP-based approaches.

### 2.2.1 Advantages

There are a number of additional reasons why ILP should be taken into consideration, even when the problems seems to not require it or the advantage of introducing an ILP formulation isn't initally clear:

- Commercial ILP *solvers* are available;
- The progress of those solvers has been spectacular: benchmark ILP problems can be solved *200-bilion* times faster than twenty-years ago;
- Even for a problem where a worst-case efficient general algorithm might be possible, the time and effort needed to find it, implement it as a computer program, is typically much greater than the time and effort needed to formulate and implement an ILP solution to the problem.
- Some problems can be modeled in a much more efficient with ILP.

## 2.3 An real-world case: Multiple Sequence Alignment

This is a classical situation where dynamic programming can be exploited: the problem was initially treated exposing the similarities and differences of the given set of sequences by calculating a two-dimensional matrix where each row represents a sequence and the columns exhibit their common patterns and their differences.

To evaluate the quality of alignments, a large number of scoring functions has been suggested, leading to this definition:

**Problem (MSA):** Given a set of sequences and a scoring function, calculate an alignment of the sequences that is optimal with respect to the scoring function.

The runtime and storage requirements of such approach are very limiting and exponentially growing when raising the number of the sequences: an instance of the problem with 10 of them was, and still is, a real challenge, while in many realistic applications users would like to compare dozens of sequences[citation].

A variant of this problem is the Maximum Weight Trace problem (MWT), introduced by John Kececioglu [quotation needed].

**Problem (MWT):** [MWT brief description]

Every Multiple Sequence Alignment problem can be cast using this formulation.

In 1997, Reinert et al. [quotation needed] proposed an ILP formulation for MWT

## 2.4 Design of an ILP formulation

### 2.4.1 Idioms

Here's how many logic expressions can be expressed as linear disequalities without side effects or uncovered cases:

**If Then** todo

**Only If** (binary variables)

$$z \leftrightarrow (L \geq b) \qquad (z = 1 \text{ only if } L \geq b)$$

Let $s$ be the smallest value that L can achieve and set $m = s - b$

$$L + m \times z \geq m + b$$

# 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.