UNIVERSIDAD
DE GRANADA

# DE NOVO GENOME ASSEMBLY USING QUANTUM ANNEALING

JOSÉ ANTONIO ÁLVAREZ OCETE

Trabajo Fin de Grado

Doble Grado en Ingeniería Informática y Matemáticas

**Tutores**
Carlos Cano
Antonio Lasanta

FACULTAD DE CIENCIAS
E.T.S. INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

*Granada, a 5 de abril de 2021*

# ÍNDICE GENERAL

## RESUMEN

TODO: resumen

# Parte I

# PARTE DE MATEMÁTICAS

TODO: cambiar la descripción de la parte (antes del titulo)

# 1

# THE GENOME ASSEMBLY PROBLEM

The genome of an organism is all its genetic material [Rot19]. The deoxyribonucleic acid (DNA) is the carrier of that genetic information. It consists of two long chains twisted to form a double helix [AJL+07]. Each of these chains is composed of a series of nucleotides or bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Since these bases are matched in pairs in the DNA double helix, those are called base pairs (bp).

A genome sequence is the complete list of nucleotides of every chromosome of an organism. With today's technology, automated sequence machines can read up to 1000 bp at a time [SKJ+11] while the human genome contains 3 Mbp, so we can't just read the whole genome. This is where genome assembly comes in.

Genome assembly refers to the process of, given a large number of short DNA reads, stitch them together to form a large representation of the original chromosome where the reads came from. The two main techniques used to reconstruct these sequences are the ab initio reference-free alignment and the de novo reference-based assembly.

## 1.1 AB INITIO REFERENCE-BASED ALIGNMENT

In this method, the DNA reads are matched against a known trusted reference of the same organism. This is essentially a pattern matching problem, where we find the index of a given sub-string in a larger string. However, after the reconstruction is complete the result is compared to the reference in order to identify implications; therefore introducing bias based on the reference [SAAB20].

In the naive approach, the short sub-string is compared to the reference starting at the first index. If the end of the sub-string is reached with a positive, a match is obtained. Otherwise, the sub-string is shifted a single position and we compare again. Heuristic methods that improve on this idea are based on shifting a greater number of spaces after a mismatch.

Different number of strategies have been developed in this direction. For instance, the classic Boyer-Moore and Knuth-Pratt-Morris algorithms [HG99]. However, these

in particular are not adequate for genome assembly since these are exact string matching algorithms and DNA reads usually need approximate matches due to reads errors. Other algorithms worth mentioning are the Needleman-Wunsch algorithm [NW70] and the Smith-Waterman algorithm [SW81], for global and local alignment respectively. These are dynamic programming algorithms designed specifically with DNA reads in mind.

State of the art algorithms trades off accuracy for speed and memory. The approximation and errors introduced prevents the application of this technique to critical areas such as personalized medicine. Given enough computational power, the de novo method yields better results.

## 1.2 DE NOVO REFERENCE-FREE ASSEMBLY

This method is reference-free, being based only on DNA reads. Thus, it has no reference bias but it is more computationally complex. It is usually used the first time a species DNA is read.

In this technique, multiple copies of the same DNA are made before slicing it. After chopping each copy at random places the data is redundant and the different reads overlap, making the assembly easier.

# BIBLIOGRAFÍA


[AJL+07] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P And Walter. Molecular Biology of the Cell - NCBI Bookshelf, 2007.

[HG99] Susan P. Holmes and Dan Gusfield. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Technical Report 447, 1999.

[NW70] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, mar 1970.

[Rot19] Stephanie Clare Roth. What is genomic medicine? *J. Med. Libr. Assoc.*, 107(3):442–448, jul 2019.

[SAAB20] Aritra Sarkar, Zaid Al-Ars, and Koen Bertels. QuASeR Quantum Accelerated De Novo DNA Sequence Reconstruction. *arXiv*, pages 1–24, 2020.

[SKJ+11] Barton E. Slatko, Jan Kieleczawa, Jingyue Ju, Andrew F. Gardner, Cynthia L. Hendrickson, and Frederick M. Ausubel. 'First generation' automated DNA sequencing technology. *Curr. Protoc. Mol. Biol.*, (SUPPL.96), oct 2011.

[SW81] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. Technical Report 1, 1981.