Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2017

Rui Espinha Ribeiro

Multiple Sequence Alignment for Mitochondrial Orchestra

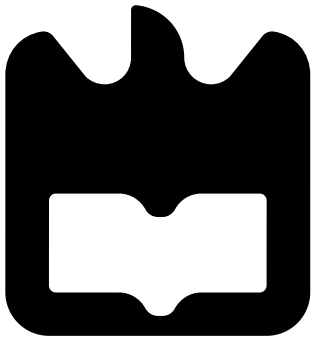
DOCUMENTO PROVISÓRIO

Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2017

Rui Espinha Ribeiro

Alinhamento Sequencial Múltiplo para Orquestra Mitocondrial

DOCUMENTO PROVISÓRIO



Rui Espinha Ribeiro

Multiple Sequence Alignment for Mitochondrial Orchestra

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Doutor Carlos Alberto da Costa Bastos, Professor do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e da Doutora Vera Mónica Almeida Afreixo, Professora do Departamento de Matemática da Universidade de Aveiro.

DOCUMENTO PROVISÓRIO

o júri / the jury

presidente / president

ABC
Professor Catedrático da Universidade de Aveiro (por delegação da Reitora da Universidade de Aveiro)

vogais / examiners committee

DEF
Professor Catedrático da Universidade de Aveiro (orientador)

GHI
Professor associado da Universidade J (co-orientador)

KLM
Professor Catedrático da Universidade N

agradecimentos / acknowledgements

NO ONE AT ALL

Resumo Um bom resumo da minha tese.

Abstract An abstract for the ages

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last two decades, there has been a growing focus on an atipical way of analysing genetic information to establish any relation between two species, the so-called "genomic music". Although we can speculate about the scientific and artistic interest of such studies, only the first one is actually explored in this work's scope as the second one and its subsequent discussions are a consequence of the musical elements we introduce to the game (in fact, composers like Michael Zev Gordon [1] have explored this crossing of different universes).
The proposal for this thesis is to build a bridge between phylogenetic studies and music through data science algorithms, as the main intent is to explore associations between different genetic codes which enable us to assume a relation between two different species. In practice, the final work is an application that creates a virtual orchestra with the mitochondrial genetic code. The instruments and the music's dynamics should be binded with the subspecies found in a specific genetic code, allowing us to give a specie a certain musical identity.
This will allow us to associate separate species by simply listening them, which enables any non-science related individual to take some sort of conclusion by listening to "the sound of a specie". It can also serve as a contribution to the long-term scientific battle that is the study of data science techniques that find evidence of relations in different genetic codes, which obviously can bring benefits in future research works.

1.1 Motivation

We can establish a strong resemblance between the *modus operandi* of between the DNA single-lettered code and sequences of musical notes. As Clare Sansom explained in *The Biochemist* [4], we can express the gene with any of the four letters A,C,G and T (we will deepen these molecular biology concepts in 1.3.), which can code each protein assembled by genetic

information. However, if we only know the letters of the sequence we don't really have relevant information to infer the structure of a protein. Very much like in the context of a musical tune, we can know the **sequences** of notes but that's not the only information required to infer the rythm nor the dynamics (we will also explore these musical concepts on 1.4.).
As stated before, scientists and artists have explored this curious analogy between this seemingly distant but yet close worlds. A good example of the interception between biological researchers and musicians is the case of the colaboration between Texas University biologist Mary Ann Clark and artist John Dunn, owner of the website Algorithmic Arts [2]. In this cooperative project, both entities aimed at converting protein structures to musical compositions. We can even listen to some results of their work, like the sound produced by a spidroin. In the same *The Biochemist* article mentioned earlier we can find an inquiring conclusion of Clark and Dunn's work: assigning music pitch and instruments to changes verified in helix and strands (REVER??), they were able to actually build *motifs* that repeated themselves in the global context of a song, very much like in a "regular" completely non-artificial composition. The main scientific interest in these observations came with the fact that it was easier to notice the similarity between the sound of human haemoglobin and the globin from the tuatara than it was with conventional sequence alignment methods.
CONTINUA

1.2 Genomic music

1.2.1 Scientific background

1.2.2 Existing projects

1.2.3 Music from the genetic code: what does it mean in 2016?

1.3 Interdisciplinary concepts: Biology

1.3.1 Basic notions

1.3.2 Multiple Sequence Alignment

One of the major tasks in this work is to find homologous sequences so we can use them to map music in a coherent way. By homologous, we mean as having structural and evolutionary resemblance - a shared ancestry between them. [3, chapter, p. 215] Such techniques for estimating homologous regions are called alignments. More specifically, the global alignment of such two sequences is called a pairwise alignment. But in this work we intent to find homologous regions between large sets of mitochondrial DNA

sequences. In this case, our goal is to perform multiple sequence alignment methods (MSA, as we will refer to them in descriptions, from this chapter on).
The analysis of these methods is highly complex as their correctness depends on the relatedness of our set of sequences. Alignment techniques must always keep in mind two key features: the fact that some positions (in the alignment) are more conserved than others, e.g. position-specific scoring; and the fact that the sequences are not independent, but instead are related by a phylogenetic tree (which we will explore more ahead) (R. Durbin et al.). MSA algorithms usually include indels (or "gaps", as they are commonly known). Such gaps are relative to insertion or deletion events that are represented between characters in an alignment column.
There are 4 main types of MSA methods:

1. Dynamic programming: this is the most direct approach to MSA. As the name suggests, it uses dynamic programming techniques (mainly, storing computationally complex results in a data structure): we usually have a gap penalty and a substitution matrix (the storage structure). The score (probability) of an aligment between each pair of amino acids are mapped in the matrix based on the similarity of their chemical properties and evolutionary probability of the mutation. This method is the result of the generalisation of the dynamic programming alignment approach that exists for pairwise alignments.
   This process can have several performance issues: it demans the computation of the whole substitution matrix, which implies that for a set of N sequences, we always estimate $2^n - 1$ pairwise alignments in a $L.N$ matrix, where L is the length of the sequences.

   [A PICTURE THAT SHOWS WHAT'S HAPPENING]

2. Progressive Alignment

3. Iterative Methods

4. Hidden Markov Models (HMM)

Furthermore: Consensus Methods

1.3.3 Phylogenetics

Distance matrices

Parsimony

Bayesian inference??

1.4 Interdisciplinary concepts: Music

1.4.1 Basic notions

1.4.2 Dynamics

1.4.3 Families of instruments in the orchestra

1.5 General Dissertation Structure

Chapter 2

From code to music

2.1 Musical synthesis

2.2 Classifying music: Zipf's law

2.3 Mapping code into music

2.3.1 Frequency of occurrence

2.3.2 Distances between words

2.3.3 Dynamics

Chapter 3

The virtual orchestra

Chapter 4

Programming environment

4.1 Python as a data science language

4.2 BioPython

4.3 Data analysis methods

Chapter 5

Final Application

5.1 Interface

5.2 Tests

5.2.1 Accuracy in building phylogenetic tree

5.2.2 Comparison between different species

5.2.3 Comparison between music mappings

5.2.4 Auditive comparison

Chapter 6

Results

Chapter 7

Conclusions

Bibliography

[1] Michael Zev Gordon. Symphony of life: making music out of the human genome. *The Guardian*, 2010.

[2] Algorithmic arts.

[3] A. Krogh R. Durbin, S. Eddy and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[4] Clare Sansom. Dna makes protein - makes music? *Cyberbiochemist*, 2002.