Universidade de Aveiro
2017

Departamento de
Electrónica, Telecomunicações e Informática,

**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

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

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 [6] 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 the DNA single-lettered code and sequences of musical notes. As Clare Sansom exposed in *The Biochemist* [10], 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 rhythm 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 hemoglobin and the globin from the tuatara than it was with conventional sequence alignment methods.

This dissertation aims to retrieve conclusions in the same direction as studies like the one mentioned before. The main technique for finding similarities in different biological sequences (known as homologous regions) is sequence alignment (these concepts will be further explored in chapter 1.3.). By representing successfully these similarities with music, we are obtaining an alternative and possibly more familiar way to identify such regions and by doing so, we can even speculate on the evolutionary relationship between the species associated with the sequences.
The decision to focus on mitochondrial DNA (mtDNA) sequences is mainly justified by the fact that mitochondria are mainly passed between generationsby maternal inheritance, as usually the mitochondria in the sperm are destroyed after fertilization (as it is explained in [4]). This pattern of inheritance suggests that mtDNA sequences may the most suitable for analysis in this work, since they provide stronger evidences of proximity between a group of species: if we find homologous regions between different mtDNA sequences, we can speculate that there's a higher probability of a stronger evolutionary relationship between such species than in the opposite way. We also must consider that the volume of mtDNA is considerably small when

compared to nuclear DNA, which would demand high computation time and elevated memory consumption. It would also be far more complex to produce listenable tracks of music from such large sequence patterns.

This work should be regarded as the follow-up to *Música Genómica* [1], a dissertation written by Rui Antunes in 2015. The core of this work was based on representing genetic sequences in musical sound. In order to map DNA sequences to musical notes, several approaches were explored and will serve as a basis for building a virtual orchestra of instruments mapped to DNA sequences (as it is exposed in chapter 3, *From code to music*). By listening to the sound of the orchestra, we should be able to identify which species are closer in terms of phylogenetics. There will be two main convergence points for observing similarities between species:

1. the families of instruments. Sequences with the greatest amount of homology should be mapped to instruments of the same family (for example, brass or percussion);

2. dynamics and harmony in the composition, which should directly proportional with the similarities observed in sequence alignment.

Finally, this dissertation will require a strong focus on data science techniques to statistically compare the alignments we produce, as well the evolutionary proximity between species in each of these alignments. Pattern discovery, clustering and classification methods will be applied, [referencias???; investigar mais métodos], in order to obtain the most suitable groups of sequences and their homologous regions in the virtual orchestra.

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 both structural and evolutionary resemblance, as Durbin and al.

designate [9]. Such resemblances can be regarded a shared ancestry. Such techniques for estimating homologous regions are called alignments. Alignment methods usually evaluate homology between biological sequences using scoring algorithms. If we are examining a pair of sequences, we use pairwise alignments. But in this work we intent to find similarities between variable sets of mitochondrial DNA sequences. In this case, our goal is to perform multiple sequence alignment methods (they will be referred as MSA, 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 (R. Durbin et al.). This dependency is represented by a phylogenetic tree (we will explore this more ahead). 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 alignment 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 generalization of the dynamic programming alignment approach that exists for pairwise alignments.

   This process can have several performance issues: it demands 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.

   [PLACEHOLDER FOR PICTURE]

2. Progressive Alignment: the most common approach. A progressive alignment is obtained with successive pairwise alignments, aligning each sequence with the previously retrieved alignment. The main differences between progressive alignments are in their scoring algorithms, the methods that are used to choose the order of the alignment and whether the progression is based in a unique alignment that is incrementally built or in a tree structure with alignments aligned with other alignments.

3. Iterative Methods: these are more of a refinement to progressive alignments. In the of the 1980's and the beggining of the 1990's, several

articles were published [8] [7] suggesting different approaches to an ongoing issue with progressive alignments: once you align a subset of sequences, the alignment is “frozen”(denomination by Durbin et al.), meaning you can no longer change a previously aligned group. The group’s relatedness and homology can be inferred differently as further sequences are aligned and by being so, “pure” progressive alignments may prove themselves considerably inaccurate. Iterative refinements are algorithms in which an initially obtained alignment is obtained and then one or more sequences are taken out and realigned with a given profile of the remaining sequences. This is iteratively done until no changes are verified in the alignment.

4. Hidden Markov Models [3] (HMM). In a brief overview, HHMs are statistical models, in which each nucleotide is assigned with three labels for three transition states and their associated probabilities. By analysing the path that it is obtained between state transitions, we can obtain a Markov chain. Each path is therefore a hidden Markov chain. HMM can be trained with several data sources (in this case, biological sequences) to improve their alignment accuracy. In the context of MSA, HMMs can be used map probabilities of gaps, matches or mismatches in alignments between sequences. MSA can be performed with an already known model or we can conceive our These models are also highly used in several data modelling problems, such as speech recognition systems, pattern recognition, data compression, among others [5].

These methods are further detailed in the above mentioned book by Durbin et al. Some MSA implementations were chosen to find similarities between biological sequences in order to cluster all homologous regions between sequences, which will serve as a basis to the assigning of each family of instruments in the virtual orchestra. They are detailed in the *Multiple Sequence Alignment* of chapter 3, *The virtual orchestra.*

In this section, we observed the relevancy of sequence alignment, as it is fundamental to find relationships between mtDNA sequences. We will also cluster different sequences by obtaining their phylogenetic trees.

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

3.1 Multiple Sequence Alignment

3.1.1 Comparison between methods

3.2 Assigning species to instruments

3.3 Building the orchestra

3.4 Data relations: what species exist in this 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] Rui Antunes. Música genómica. Master's thesis, Universidade de Aveiro, 2015.

[2] John Dunn. Algorithmic arts. `http://algoart.com/music.htm`. Accessed: 2016-12-05.

[3] Sean R Eddy. What is a hidden markov model? *Nature Biotechnology*, 22(10):1315–1316, 2004.

[4] Centre for Genetics Education. Mitochondrial inheritance. `http://www.genetics.edu.au/Publications-and-Resources/Genetics-Fact-Sheets/FactSheetMitochondria`. Accessed: 2016-12-08.

[5] Z. Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

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

[7] Osamu Gotoh. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *CABIOS*, 9(3):361–370, 1993.

[8] Geoffrey J.Barton and Michael J.E.Sternberg. Evaluation and improvements in the automatic alignment of protein sequence. *Protein engineering*, 1(2):89–94, 1987.

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

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