UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEINFO – DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA

GRADUAÇÃO EM BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

INTERAÇÃO ENTRE PATÓGENOS: ABORDAGENS COMPUTACIONAIS NA BUSCA POR PADRÕES EM GENOMAS FILOGENETICAMENTE DISTANTES

LEONARDO FIGUEIRÔA E SILVA

Trabalho de Graduação

Recife
07 de fevereiro de 2018

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEINFO – DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA

LEONARDO FIGUEIRÔA E SILVA

INTERAÇÃO ENTRE PATÓGENOS: ABORDAGENS COMPUTACIONAIS NA BUSCA POR PADRÕES EM GENOMAS FILOGENETICAMENTE DISTANTES

Trabalho apresentado ao Programa de GRADUAÇÃO EM BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO do DEINFO – DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA da UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO como requisito parcial para obtenção do grau de Bacharel em CIÊNCIA DA COMPUTAÇÃO.

Orientadora: *Profa. Jeane Cecília Bezerra de Melo*
Co-orientadora: *Profa. Nara Suzy Aguiar de Freitas*

Recife
07 de fevereiro de 2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S586i Silva, Leonardo Figueirôa e
Interação Entre Patógenos: Abordagens Computacionais na Busca por Padrões em Genomas Filogeneticamente Distantes / Leonardo Figueirôa e Silva. - 2018.
53 f. : il.

Orientadora: Jeane Cecilia Bezerra de Melo.
Coorientadora: Nara Suzy Aguiar de Freitas.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2020.

1. Biologia computacional. 2. Ecologia. 3. Reconhecimento de padrões. 4. Interação entre patógenos. 5. Descoberta de motifs. I. Melo, Jeane Cecilia Bezerra de, orient. II. Freitas, Nara Suzy Aguiar de, coorient. III. Título

CDD 004

In memoriam:
Adalgisa Soares Figueirôa

Agradecimentos

Agradeço a Deus por minha vida e por permitir a conclusão deste trabalho.

Agradeço a minha família que sempre esteve presente e procurou me apoiar de diversas maneiras durante minha graduação.

Agradeço a minha orientadora Jeane Melo por sua orientação e perseverança, estando comigo até a conclusão deste trabalho. Agradeço também a minha co-orientadora, Nara Freitas, pelas inúmeras reuniões de apoio para que o trabalho não fugisse do escopo, e também a Felipe Pessoa, amigo biólogo que juntamente com Nara me ajudou na elaboração do trabalho e de quem partiu a ideia aqui proposta neste documento.

Agradeço aos professores George Valença, Érica Souza e Jeísa Domingues pelo apoio, dicas e conversas que ajudaram bastante na elaboração desse texto técnico.

Agradeço aos professores constituintes da banca de defesa, Rodrigo Nonamor e Paulo Souza, por suas contribuições através de comentários voltados para a melhoria deste documento.

Agradeço a todos os amigos que fiz durante o período de graduação, em especial a Rodrigo Cunha, Henrique Duarte, Lucas de Holanda, Daniel Vilas-Boas, Thomás Leal, Pedro Pires, Thiago Duarte, Italo Lemos, Victor Sales, Suzana Saraiva, Daniel Nogueira e Dennys Barros com os quais compartilhei momentos felizes e difíceis nesta jornada e que levo para a vida.

Agradeço também aos amigos de fora da universidade, em especial a Ester Lim, Rebecka Borges e Egon Brandão, que acompanharam a elaboração deste documento, compartilhando ideias e me apoiando.

Por fim, a todos os professores de dentro e de fora desta instituição que contribuíram para a minha formação e a todas as pessoas que direta ou indiretamente contribuíram para a conclusão deste trabalho, o meu sincero muito obrigado!

"Tantas perguntas permanecem sem respostas. Talvez sejamos pobres por termos perdido uma possível explicação ou ricos por termos ganho um mistério. De qualquer forma, não são ambas as possibilidades igualmente intrigantes?"

—PETER WOHLLEBEN

Resumo

Considerada uma área emergente, o estudo da Interação Entre Patógenos (*PPI — Pathogen-Pathogen Interaction*, em inglês) tem recebido considerável atenção devido às implicações de saúde que ela representa para a população humana. No início do desenvolvimento desta pesquisa, biólogos do Departamento de Biologia da Universidade Federal Rural de Pernambuco realizaram análises nos genes e proteínas do *Papilomavírus humano* tipo 16 (HPV 16) contidos em bancos de dados de sequências do NCBI — *National Center for Biotechnology Information*. Essas análises iniciais resultaram em alinhamentos similares e em sintenia com o genoma da *Chlamydia trachomatis*. Como esses patógenos estão distantes filogeneticamente, pouco se sabe sobre seu histórico de interação e evolução a nível genético. Portanto, uma pesquisa que avalie as similaridades entre os genomas desses organismos poderia contribuir para uma melhor compreensão do processo de interação entre eles, estabelecendo uma relação ecológica e de padrões evolutivos que podem contribuir para a magnitude da infeção causada por esses agentes.

Analisar eventos evolutivos entre genomas filogeneticamente distantes envolve procurar por padrões que a *priori* não são conhecidos em regiões conservadas dos genomas, levando em consideração suas características específicas. Tendo em vista a não disponibilidade de métodos computacionais para tratar deste problema e suas especificidades, o presente trabalho se propôs realizar estudo sobre abordagens atuais para problemas deste tipo e a implementar uma heurística, utilizando métodos computacionais clássicos de busca por padrões em sequências e conhecimentos biológicos especificos, afim de investigar possíveis relações evolutivas e interações entre as espécies *Alphapapilomavirus* 9 e *Chlamydia trachomatis* através da aplicação de técnicas computacionais e genômica comparativa.

A implementação da heurística envolveu gerar informações sobre homogeneização dos genomas, uso de códon, propriedades físico-químicas dos aminoácidos e descoberta de *motifs* comuns às sequências, através da busca exaustiva. Como os resultados resultados obtidos foram volumosos, eles foram agrupados utilizando o método estatístico de análise de correspondência para fins de uma melhor visualização das relações entre as diferentes variáveis de análise e os resultados. O agrupamento final trouxe indícios que suportam a hipótese inicialmente levantada pelos biólogos, dando margens para novas interpretações sobre como esses organismos se relacionam.

Palavras-chave: biologia computacional, interação entre patógenos, descoberta de *motifs*.

Abstract

Considered an emerging area, the study of Pathogen-Pathogen Interaction has received considerable attention over the recent years because of the health implications it poses to the human population. At the beginning of this research project, biologists from the Department of Biology from the Federal Rural University of Pernambuco conducted an analysis on the genes and proteins of *Human papillomavirus* type 16 (HPV 16) contained in the National Center for Biotechnology Information (NCBI) sequences databases. The initial analysis resulted in similar alignments and in synteny with the genome of *Chlamydia trachomatis*. As these pathogens are phylogenetically distant, little is known about their history of interaction and evolution at the genetic level.

The analysis of evolutionary events between phylogenetically distant genomes involves looking for patterns that are not previously known in conserved regions of the genomes, taking into account their specific characteristics. Considering the non availability of computational methods to deal with this problem and its specificities, the present research project intends to study current approaches to similar problems and to implement a heuristic using classical computational methods for motif fiding and specific biological knowledge in order to investigate possible evolutionary relationships and interactions between the species *Alphapapilomavirus* 9 and *Chlamydia trachomatis* through the application of computational techniques and comparative genomics.

The implementation of the heuristics involved gathering information about genome homogenization, codon usage, physiochemical properties of amino acids, and finding motifs common to the sequences through exhaustive searching. Since the results obtained from the implementation of the heuristic were bulky, it was necessary to cluster them through a statistical method. The method chosen was correspondence analysis, which helps with data visualization and allows the view of relationships between the variables of the analysis and the results obtained. This clustering of the data gathered in the process provided clues that support the hypothesis initially raised by the biologists, allowing for the formulation of new interpretations as of how these organisms interact.

Keywords: computational biology, pathogen-pathogen interaction, motif finding, motifs.

Lista de Figuras

Lista de Tabelas

Sumário

PARTE I

Introdução

CAPÍTULO 1

Apresentação

"Na tríade epidemiológica clássica, a expressão clínica das doenças infecciosas é interpretada como um produto de uma relação intrínseca envolvendo um agente infeccioso, a resposta imune do hospedeiro e fatores ambientais. [...] existe um interesse crescente no fato de que agentes infecciosos frequentemente não agem de forma independente; mas seu potencial de virulência é mediado de diversas formas através de seus relacionamentos com outros patógenos."

—SINGER, 2010

A Biologia Computacional é definida pelo *National Institute of Health – NIH* (Instituto Nacional de Saúde, em português) como "O desenvolvimento e aplicação de métodos teóricos e de análise de dados, modelagem matemática e técnicas de simulação computacional para o estudo de sistemas biológicos, comportamentais e sociais" (M. HASELTINE F., 2000). Neste sentido, a biologia computacional é intrinsecamente multidisciplinar, abrangendo diversos campos das ciências da vida, desde moléculas a ecossistemas, tornando-se indispensável no avanço dessas ciências (BOURNE; BRENNER; EISEN, 2015). Uma das principais contribuições da Biologia Computacional é a recuperação e descoberta de novas informações a partir daquelas armazenadas em diferentes níveis de modelagem, tais como sequências e estruturas (NUSSINOV et al., 2015). Através da Computação, os processos de recuperação, extração e análise de dados, antes muito custosos, tornaram-se baratos e acessíveis, e o desenvolvimento das técnicas de simulação computacional possibilitou encontrar novas interpretações, como novas funções de proteínas (SCIACCA, 2009).

Considerada uma área ainda emergente, o estudo da Interação Entre Patógenos (PPI — *Pathogen-Pathogen Interaction*, em inglês) tem recebido considerável atenção devido às implicações de saúde que ela representa para a população humana, sendo assim uma forma de epidemiologia. A interação entre patógenos tem impacto em muitos fatores, mas principalmente nos fatores de virulência[1], o que pode potencializar as patologias causadas pelos agentes infecciosos no hospedeiro. Estudar a PPI contribui de maneira significativa para a pesquisa, tratamento e prevenção das implicações resultantes da interação sindêmica entre patógenos (CATTADORI; BOAG; HUDSON, 2008; SINGER, 2010).

[1] **Virulência:** capacidade de infecção de um agente, mas não necessariamente infecção patógena.

1.1 Motivação

No início do desenvolvimento desta pesquisa, biólogos do Departamento de Biologia da Universidade Federal Rural de Pernambuco realizaram análises nos genes e proteínas do *Papilomavírus humano* tipo 16 (HPV 16) contidos em bancos de dados de sequências do NCBI — *National Center for Biotechnology Information*. Essas análises resultaram em alinhamentos similares e em sintenia[2] com o genoma da *Chlamydia trachomatis*. Posteriormente, foi emitida uma nota no banco de dados indicando que o genoma continha um contaminante. Essa nota, por sua vez, instigou os biólogos a realizarem uma investigação da coexistência evolutiva entre ambos os patógenos.

O *Papilomavírus humano* (HPV) é um vírus pertencente à família *Papillomaviridae*, que apresenta um histórico íntimo de coevolução junto ao seu hospedeiro. O vírus é um dos principais agentes causadores de câncer cervical no mundo (CLIFFORD et al., 2003), e atualmente é foco de uma campanha de vacinação mundial, principalmente entre crianças e adolescentes. A *Chlamydia trachomatis* (CT) é uma bactéria gram-negativa, incapaz de sintetizar ATP (adenosina trifosfato — um nucleotídeo que é subproduto da respiração celular e é responsável por armazenar energia proveniente da respiração celular e fotossíntese para consumo imediato) e necessita viver, obrigatoriamente, no interior da célula do hospedeiro. Esta bactéria é responsável pelo desenvolvimento do tracoma e também é o principal responsável por doenças sexualmente transmissíveis, como a clamidíase (GAUNT et al., 2003; CHOROSZY-KRÓL et al., 2012).

Recentemente, dados sobre a infecção causada por esses dois organismos apontam para uma relação benéfica mútua entre ambos os patógenos (SIMONETTI et al., 2009; TAVARES et al., 2014; WOHLMEISTER et al., 2016). Por compartilharem o mesmo ambiente, o interior de células do colo do útero, permitiu-se levantar a hipótese de que ambos genomas estariam sujeitos a eventos de transferência lateral de genes comuns. Todavia, não existe informação sobre a ocorrência destes eventos entre ambos patógenos. Desta forma, uma pesquisa que avalie as similaridades entre os genomas de ambos os patógenos poderia contribuir para uma melhor compreensão do processo de interação entre os mesmos e estabelecer além da relação ecológica, a relação de padrões evolutivos que podem contribuir para a magnitude da infecção causada por esses agentes.

1.2 Problema de Pesquisa

Analisar eventos evolutivos entre genomas filogeneticamente distantes envolve procurar por padrões que a *priori* não são conhecidos em regiões conservadas dos genomas, levando em consideração suas características específicas. Atualmente essa análise é feita utilizando-se algumas ferramentas como o *Mauve*[3], o *MUSCLE*[4] e o *COPid*[5], as quais não são adequadas para

[2]Sintenia: estado no qual dois ou mais genes estão presentes em um mesmo cromossomo.

[3]Sistema que constrói alinhamentos múltiplos considerando processos de rearranjo e inversão gênica.

[4]Ferramenta utilizada para comparar alinhamentos múltiplos entre sequências de proteínas.

[5]Servidor *web* que auxilia na anotação das funções de proteínas, considerando sua composição, utilizando parte ou a proteína completa.

o tipo de análise que se deseja fazer nos organismos aqui estudados. Os resultados buscados diferem daqueles que se desejam analisar e apresentam baixa precisão quando aplicadas ao conjunto de dados deste trabalho. Um outro problema associado é o grande volume de dados gerados neste tipo de análise, visto que, como não conhecemos os padrões, todas as possibilidades devem ser enumeradas e o cruzamento dos dados, automatizado.

1.2.1 Pergunta de Pesquisa

Este trabalho procura responder a seguinte pergunta de pesquisa:

"Em que medida técnicas de busca e estratégias de análise podem ser combinadas em uma heurística para a descoberta de padrões no problema da análise de eventos evolutivos entre genomas filogeneticamente distantes?"

1.3 Objetivos

1.3.1 Objetivo Geral

Tendo em vista a não disponibilidade de um método de análise específico e automatizado para tratar deste problema, o presente trabalho se propõe a implementar uma heurística que utiliza métodos computacionais clássicos de busca por padrões em sequências e conceitos da biologia molecular, como a genômica comparativa, afim de investigar possíveis relações evolutivas e interações entre as espécies *Alphapapilomavirus 9* e *Chlamydia trachomatis*.

1.3.2 Objetivos Específicos

Os objetivos específicos desta pesquisa se configuram em:

1. Estudar as propriedades e restrições biológicas das regiões analisadas;
2. Definir uma heurística de análise através da combinação entre técnicas de Computação e Biologia;
3. Implementar o algoritmo de Busca Exaustiva adaptado para o problema, como o objetivo de encontrar padrões curtos e repetitivos que são comuns às sequências analisadas;
4. Obter informações sobre a homogeinização dos genomas, uso de códons por cada sequência e suas propriedades físico-quimicas;
5. Agrupar os dados obtidos utilizando Estatística Multivariada.

1.4 Conteúdo do Documento

O presente documento encontra-se dividido em três partes: a Parte I engloba os capítulos I, II, III e IV que tratam dos conceitos introdutórios. A Parte II está distribuída da seguinte forma: o Capítulo II traz os conceitos necessários para entender os elementos buscados, sua importância, estruturas e funções, do ponto de vista biológico. No Capítulo III, conceitos da computação necessários para a resolução do problema são explicados brevemente. No Capítulo IV são

apresentados os métodos relevantes na busca por padrões curtos e repetidos nas sequências, que são utilizados, criados ou adaptados em trabalhos relacionados. O Capítulo V traz a metodologia utilizada na pesquisa tais como suas fases e abordagens utilizadas, e outros aspectos como a natureza e abordagem da pesquisa são definidos. Seguido da metodologia, no capítulo VI são apresentados os resultados obtidos e suas discussões. Por fim, o Capítulo VII traz a conclusão da pesquisa e comenta brevemente sobre os impactos da pesquisa e encerrando-se com os trabalhos futuros. Os Capítulos VI e VII compõem a Parte III deste trabalho, intitulada Conclusão.

CAPÍTULO 2

Fundamentos Biológicos

"Do ponto de vista histórico, e num contexto contemporâneo e técnico, o lema da genética molecular e da biotecnologia segue sendo 'biologia é informação'."

—THACKER, 2005

Considerações iniciais

Procurar e analisar padrões que indicam interações gênicas entre patógenos exigem o conhecimento de alguns conceitos das áreas de Biologia Molecular e Genética. Este capítulo traz, de forma breve, alguns dos conceitos dessas áreas que estão ligados ao problema de pesquisa aqui tratado. Se desejar, o leitor é direcionado a consultar (DOUDNA; COX, 2012) e (PIERCE, 2012) para mais detalhes.

2.1 Moléculas Orgânicas: estruturas, funções e processos

O início da vida, em termos de moléculas orgânicas, se deu durante o período conhecido como sopa primordial. Foi durante esse período que processos químicos aleatórios deram origem a diversas moléculas orgânicas complexas e grandes, denominadas macromoléculas ou polímeros. Essas macromoléculas, formadas por subunidades repetidas chamadas monômeros, podem ter denominações diferentes baseadas em sua estrutura química e função, tais como ácidos nucleicos, carboidratos, lipídios e proteínas. O foco desta pesquisa está nos três polímeros responsáveis pela manutenção da informação genética e processos metabólicos da célula que são essenciais para todas as formas de vida conhecidas: os ácidos nucleicos, DNA e RNA, e as proteínas (Figura 2.1) (CECH, 2012; ROBERTSON; JOYCE, 2012; EIDHAMMER; JONASSEN; TAYLOR, 2000).

2.1.1 Ácidos Nucléicos

Os ácidos nucléicos são polímeros essenciais para que os organismos vivos como conhecemos pudessem surgir e evoluir, desempenhando um papel crucial nos processos de manutenção e regulação das funções biológicas, e também de herança de características (ACHAR; SÆTROM, 2015; CECH, 2012).

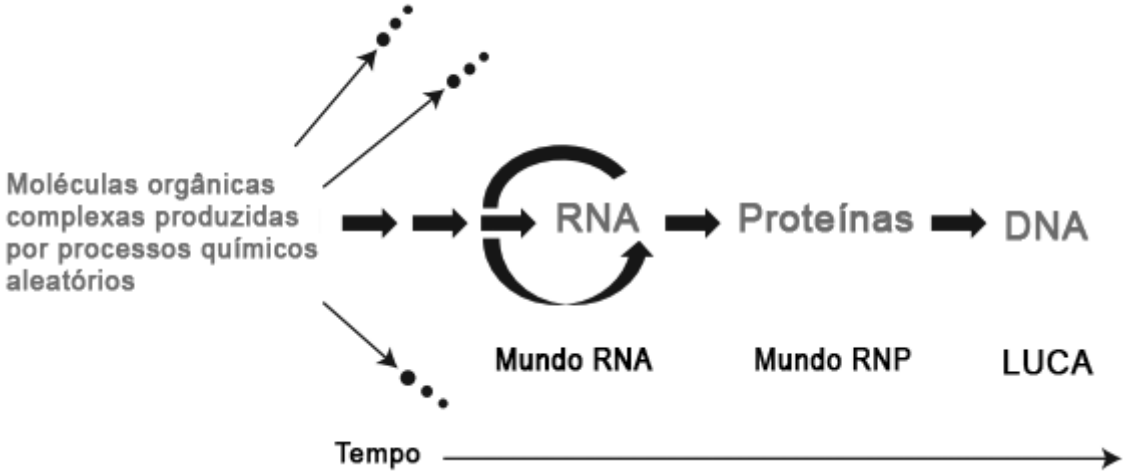


Figura 2.1: O RNA, precursor das proteínas e do DNA. Mundo RNP (ribonucleoproteína) se refere ao período em que as proteínas surgiram à partir dos processos de síntese de proteína iniciados graças ao surgimento do RNA. LUCA, do inglês *Last Universal Common Ancestor*, se refere ao último ancestral comum a todas as espécies conhecidas atualmente. Traduzido de: (ROBERTSON; JOYCE, 2012)

O ácido ribonucléico (RNA) é uma macromolécula que prosperou durante a fase do surgimento das moléculas orgânicas complexas há quase cinco milhões de anos. O seu aparecimento possibilitou que a expressão de hereditariedade e catalisação de reações químicas em células primitivas fosse possível. Posteriormente, o RNA deu origem a outras macromoléculas, o ácido desoxirribonucléico (DNA) e as proteínas, que assumiram os papeis de armazenar as informações genéticas, no caso do DNA, e de ser um catalisador e componente estrutural das células, no caso das proteínas. Apesar disso, o RNA permanece como mediador e catalisador em processos fundamentais nas células modernas. Sua função principal é participar da síntese de proteínas, mas ele também desempenha outras funções como corte e ligação de outras moléculas de RNA, e catálise na formação de ligações peptídicas nos ribossomos, chamados de ribozimas (ROBERTSON; JOYCE, 2012; ALBERTS et al., 2002).

O ácido desoxirribonucléico (DNA) é uma macromolécula que surgiu após o RNA, se especializando em armazenar informações genéticas que coordenam o desenvolvimento e funcionamento dos seres vivos, incluindo alguns vírus (JUNQUEIRA et al., 1998). Por muitos anos, os cientistas procuraram estabelecer a sua forma espacial sem sucesso. Apenas em 1952, a cientista britânica Rosalind Franklin, através de difração de raios X e cristalografia, obteve imagens que axuliaram na compreensão da estrutura desse polímero. Em 1953, Watson e Crick, baseados nos trabalhos de Rosalind e Wilkins, propuseram um modelo de fita dupla tridimensional para o DNA. O conjunto de sequências de DNA de um organismo é denominado gene, e, o conjunto de genes constitui um genoma (INSTITUTE, 2017; PRAY, 2008).

Em se tratando de sua composição química, ambos DNA e RNA são bastante semelhantes, sendo constituídos de monômeros chamados de nucleotídeos. Eles, por sua vez, são compos-

tos por um grupo fosfato, um açúcar (desoxirribose ou ribose) e uma base nitrogenada. Os nucleotídeos são divididos em purinas (Adenina e Guanina) e pirimidinas (Citosina, Timina e Uracila), compostos orgânicos formados por anéis aromáticos nitrogenados, sendo as purinas maiores do que as pirimidinas e possuindo um um anel de carbono-nitrogênio duplo. Um fato curioso, observado por Irwin Chargaff, é que as purinas se unem às pirimidinas através de pontes de hidrogênio buscando maximizar esse tipo de ligação para se tornarem estáveis, favorecendo o emparelhamento de base purina-pirimidina. A diferença de tamanho também tem um papel determinante para que esse emparelhamento ocorra sempre entre esses dois compostos. Esse fenômeno é conhecido como Regra de Chargaff e os emparelhamentos entre as bases são chamados de pareamentos complementares (DOUDNA; COX, 2012). Para o DNA, os nucleotídeos são a Citosina (C), Guanina (G), Adenina (A) e Timina (T). No RNA, a Timina (T) é substituída pela Uracila (U). A estrutura final dessas macromoléculas pode ser vista na Figura 2.2.

2.1.2 Proteínas

Após os surgimento dos procariotos e eucariotos, as propriedades evolutivas do código genético estabeleceram que os nucleotídeos dispostos em trincas dariam origem à estruturas orgânicas chamadas de aminoácidos. Os aminoácidos são os monômeros das proteínas, macromoléculas importantes na regulação das funções biológicas das células. Elas são responsáveis por assumir vários papeis na célula, dentre eles o de componente estrutural, onde elas provêm estrutura e suporte para as células; anticorpos, que se ligam à partículas estrangeiras, tais como vírus e bactérias; enzimas, que catalisam quase todas as reações químicas que ocorrem na célula; mensageiros, que transmitem sinais para coordenar processos biológicos entre células, tecidos e órgãos; e, transporte e armazenamento, onde elas carregam átomos e pequenas moléculas dentro e fora das células (DOUDNA; COX, 2012). Os quatro nucleotídeos podem estar dispostos em 64 combinações de trincas diferentes, formando 20 aminoácidos, dos quais nove são descritos por dois códons sinônimos[1], cinco são descritos por quatro códons diferentes, três são codificados por seis códons e dois aminoácidos são codificados por um códon. Apenas um aminoácido é codificado por três códons, mas o códon de terminação (*stop codon)* também é codificado por três trincas diferentes. Os códons sinônimos normalmente se diferem por um nucleotídeo na terceira posição, ou na segunda posição, em alguns aminoácidos (SUEOKA, 1961; GOUY; GAUTIER, 1982). Os códons sinônimos variam entre os genomas, possibilitando a verificação de relações evolutivas entre sequências diferentes. Um indicador de homogeneidade em trechos do genoma é a proporção de códons constituídos pelos nucleotídeos Guanina (G) e Citosina (C) (conteúdo-GC), já que sua presença em grandes taxas indicam que mutações deletérias ocorreram com menos frequências naquele código genético (LOBRY; CHESSEL, 2003; LI, 1987; ARCHETTI, 2004).

[1]Códons sinônimos: códons que possuem uma leve mudança em suas bases (geralmente na segunda ou terceira base) mas que representam o mesmo aminoácido. Um exemplo de códons sinônimos seriam os códons UUU e UUC, que representam o aminoácido Fenilalanina.

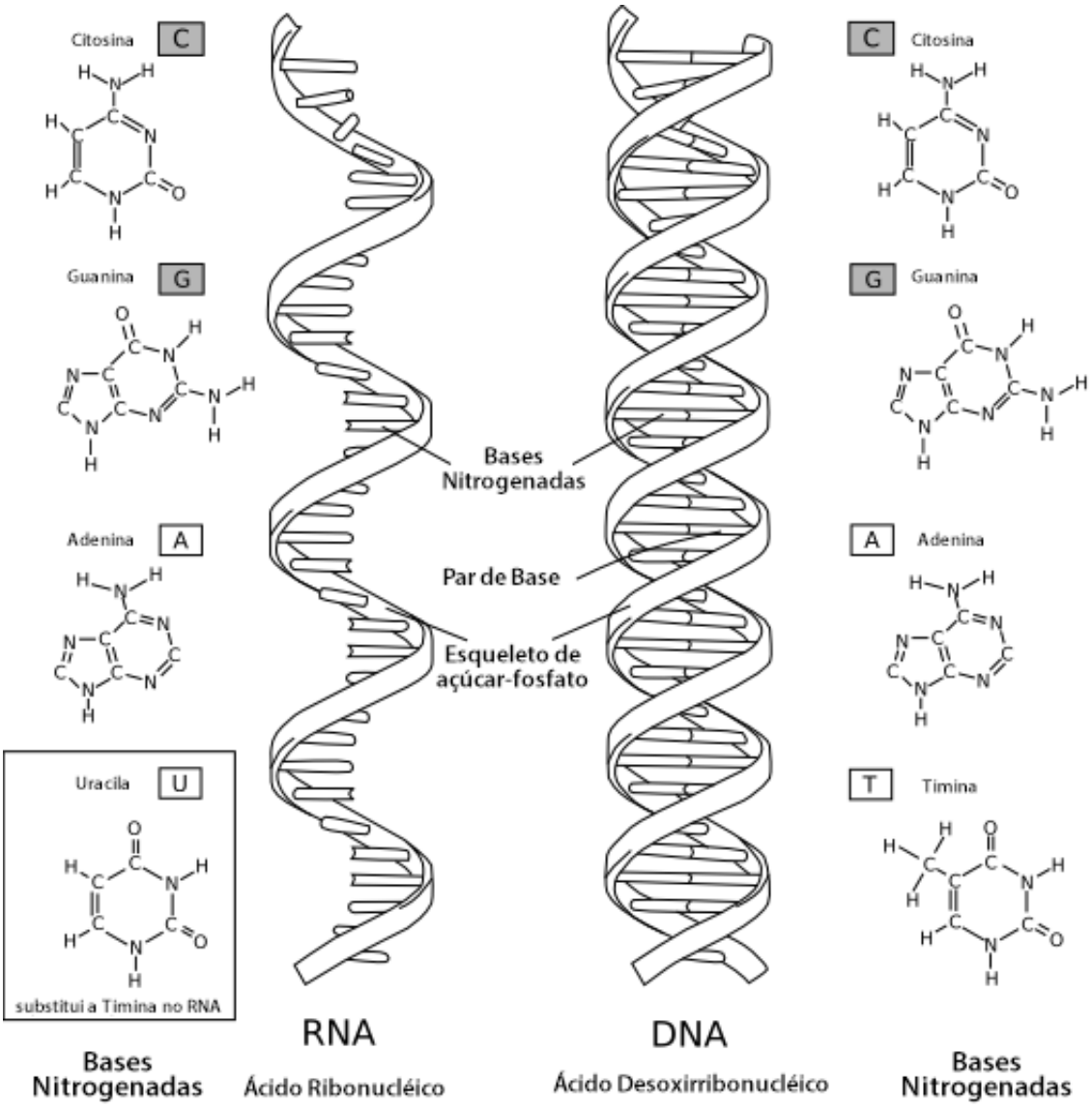


Figura 2.2: Estrutura dos ácidos nucléicos DNA e RNA. Traduzido de: <https://en.wikibooks.org/wiki/An_Introduction_to_Molecular_Biology/RNA:The_ribonucleic_acid>

2.1.3 *Motifs*

Durante a síntese de proteínas (tradução), na etapa de iniciação, regiões da sequência do RNA_m interagem com um complexo de inicialização que é formado por proteínas e enzimas que auxiliam no processo de ligação do ribossomo ao RNA_m. Essas regiões do RNA_m são um consenso de bases que indicam o local de ligação do ribossomo para a síntese da proteína. Esse consenso ou sequência de bases (aqui chamados de *motif*) é denominado de sequência reguladora, e está associada a expressão gênica. Tais *motifs* estão presentes em grande número nas sequências localizadas no início dos genes e se dividem em dois tipos: promotores e reforçadores.

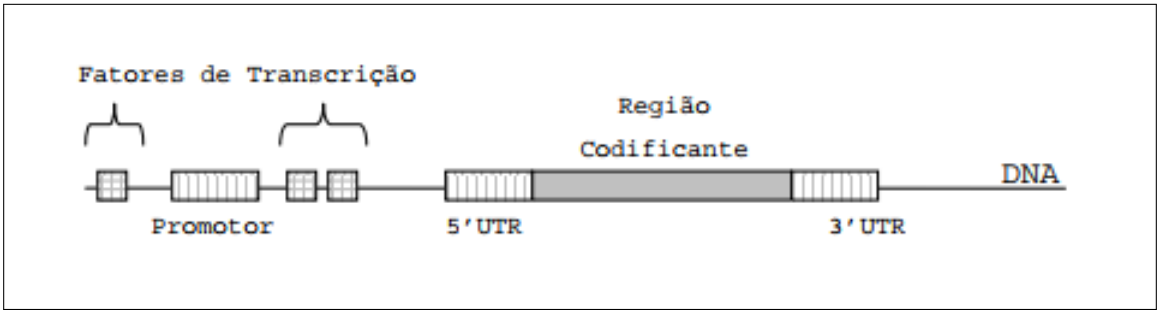


Figura 2.3: Motifs, promotores e fatores de transcrição. Fonte: (LEMOS; ARAGAO; CASANOVA, 2003)

Quando os fatores (proteínas e enzimas auxiliares) estão nas proximidades de um sítio de ligação (*motif*), o ribossomo irá iniciar o processo de síntese à partir daquele local e irá seguir a sequência até a fase de término. A presença do códon de término (*stop códon*) e de *motifs* reguladores irá determinar o fim da síntese da proteína. Sendo assim, os *motifs* possuem um papel importante na síntese de proteínas e expressão gênica, sendo possível que um mesmo trecho de uma sequência de RNA possa sintetizar mais de uma proteína, como mostra a Figura 2.4. Portanto, *motifs* de mesmo comprimento em locais e de sequência de bases semelhantes que estão presentes em trechos de genomas de dois organismos, podem indicar que esses organismos sintetizam as mesmas proteínas ou proteínas similares (PIERCE, 2012).

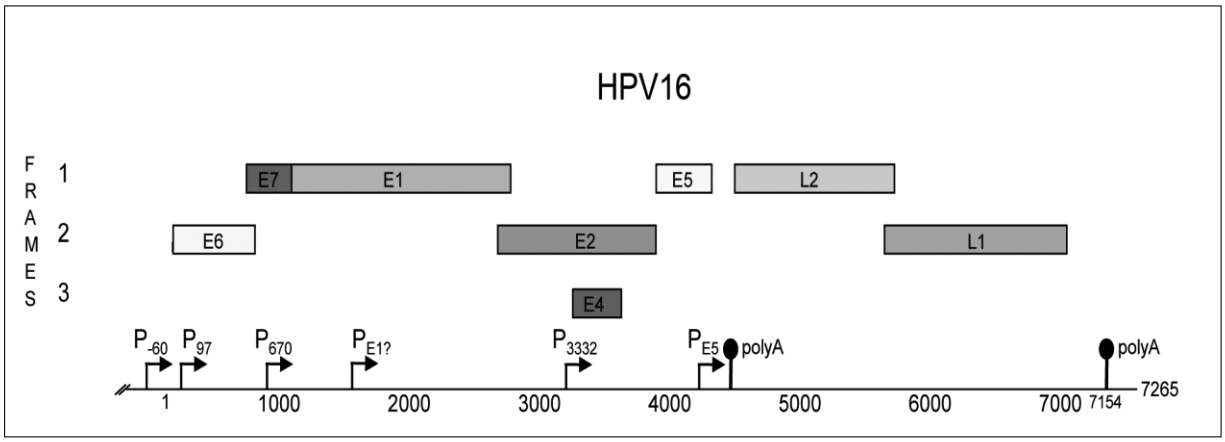


Figura 2.4: *Frames* de leitura (ORFs) e RNAm do genoma do HPV16. Observe que há uma sobreposição nas regiões E2 e E6, o que indica que um mesmo trecho de RNAm sintetiza proteínas diferentes. Fonte: (SCHWARTZ, 2013).

CAPÍTULO 3

Fundamentos Computacionais

Considerações iniciais

A palavra "algoritmo"é derivada do nome do árabe matemático al-Khwarizmi, que viveu durante o século IX. Um algoritmo é um conjunto de instruções bem definidas que frequentemente envolve a repetição de uma operação com o propósito de desempenhar uma tarefa específica, como calcular o mínimo divisor comum entre dois números inteiros ou descobrir padrões em sequências de DNA (CORMEN, 2009). Como muitos problemas da Biologia Computacional são NP-Completos, este capítulo fala brevemente sobre conceitos de complexidade e uma das formas de tratar do problema de busca por *motifs* regulatórios. Caso o leitos não esteja familiarizado com tais conceitos, pode-se consultar (CORMEN, 2009), (FEOFILOFF, 1999) e (STORMO, 2004) para um entendimento mais aprofundado.

3.1 Complexidade do Problema

Em Ciência da Computação costuma-se classificar problemas de acordo com sua complexidade, medida em função do tamanho da entrada, referênciado como n. Problemas para os quais conhecemos uma solução de complexidade polinomial, considerados *tratáveis* ou *fáceis*, são classificados como pertencentes à classe P, ou seja, o número de operações necessárias para resolvê-lo é descrita como um polinômio em função do tamanho de entrada n (CORMEN, 2009; FEOFILOFF, 1999).

No entanto, há uma classe de problemas para os quais não se conhece uma solução polinomial. Porém, não foi demonstrada que ela não existe. A solução deste conjunto de problemas, por sua vez, pode ser verificada em tempo polinomial. Estes problemas são classificados como pertencentes à classe NP, ou seja, para os quais existe um certificado que pode ser verificado em complexidade polinomial por um algoritmo determinístico. NP é uma abreviação para *Non-deterministic Polynomial*. Basicamente, pertencer à classe NP implica em poder veriirificar, em tempo polinomial, se uma suposta solução de uma instância do problema é de fato uma solução.

Uma máquina não determinística é uma abstração. Trata-se de um modelo computacional que são constituídos de uma fase onde uma função de Escolha E escolhe uma possível solução em um conjunto, e esta solução é verificada utilizando comandos de uma máquina determinística. Se o elemento informado pela função escolha for uma solução para o problema a máquina irá aceitar a solução e, caso contrário, a rejeitará. Quando esta verificação é feita em tempo polinomial, o problema é dito Não-deterministicamente Polinomial, pertencendo, portanto, à classe NP.

Um outro aspecto também considerado no estudo desta classe de problemas é o conceito de *redutibilidade*. Dizemos que um problema A pode ser reduzido a um problema B quando podemos transformar uma instância de A em uma instância de B e uma solução de B em uma solução de A. Desta forma, se conhecermos uma solução para B, no caso, um algoritmo polinomial para B, podemos resolver A através de B, utilizando as transformações. Se as transformações são feitas em tempo polinomial, dizemos que A é polinomialmente redutível a B. Um problema que pertence à classe NP e é polinomialmente redutível aos demais problemas desta classe é dito *NP-difícil* (CORMEN, 2009; FEOFILOFF, 1999).

3.2 Descobrindo *Motifs*

A identificação de padrões conservados entre sequências pode dar indícios de que elas estão relacionadas funcionalmente e estruturalmente. No entanto, procurar por *motifs* em biossequências envolve procurar por padrões onde há pouca variabilidade e bastante repetição, já que o código genético está descrito num alfabeto de quatro caracteres para o DNA e o RNA e 20 para as proteínas (STORMO, 2004).

3.2.1 Busca Exaustiva

Os problemas de descoberta de padrões são NP-difíceis e, portanto, por não se conhecer uma solução polinomial faz-se necessário abordá-lo de modo a encontrar uma melhor solução em tempo viável. Métodos comumente utilizados neste tipo de problema são baseados na busca exaustiva. A busca exaustiva, apesar de ter complexidade de tempo exponencial no pior caso, pode ser utilizada em conjunto com técnicas que diminuem o espaço de busca afim de reduzir a complexidade de tempo. Inicialmente, o método mais simples é enumerar todos os padrões que satisfazem as restrições das sequências analisadas, contar suas ocorrências e comparar se elas existem em ambas. Esse método é cabível para padrões pequenos e simples, como os *motifs*. Felizmente, muitos sítios de ligação são padrões sem interrupções e com pouca variação, sendo possível utilizar a busca exaustiva como abordagem para esse tipo de problema (BREJOVÁ et al., ; STORMO, 2004).

Informalmente, o problema da descoberta de *motifs* é descobrir padrões relacionados de um tamanho especificado em uma coleção de um fragmento de DNA ou RNA. Seja $D = (D_1, \cdots, D_t)$ uma coleção de trechos de DNA sobre um alfabeto de nucleotídeos $\Sigma = (A, C, T, G)$, e, seja $|D_i| = n$ e $l < n$ o tamanho do trecho que estamos procurando, o problema pode ser formalizado como:

> Dada uma coleção de trechos de DNA $D = (D_1, \cdots, D_t)$, cada trecho de tamanho n e um número inteiro $0 < l \leq n$, encontre a lista de posições iniciais (alinhamento) $s = (s_1, \cdots, s_t)$ que possuem os trechos mais similares possíveis.

É possível reformular o problema da busca por *motifs* em um problema mais simples. Considere o problema da Busca pela *String* Mediana:

Dados dois l-mers (trechos de tamanho l) v e w, podemos calcular a distância de Hamming entre eles como o número de posições que diferem entre v e w.

A distância de Hamming entre dois trechos v e s é dada por

$$d_H(v,s) = \sum_{j=1}^{t} d_H(v, D_j[s_j : s_{j+l}])$$

Portanto, a distância total entre um trecho v e a coleção D é

$$DistanciaTotal(v,D) = mind_H(v,s).$$

Logo, o problema da *string* mediana consiste em achar l-mers que satisfazem a restrição da distância de edição entre dois trechos, dadas posições iniciais nas sequências de entrada. Esse problema se classifica como um problema de minimização: estamos procurando ocorrências que minimizem $d_H(v,w)$ através de todas as posições iniciais de uma sequência $S = (s_1, s_2, \ldots, s_t)$ para $1 \leq s_1 \leq n - l + 1$ (STORMO, 2004).

CAPÍTULO 4

Trabalhos Relacionados

Considerações Iniciais

Neste capítulo são citados e discutidos alguns métodos que se propõem a resolver o problema da busca por padrões curtos (*motifs*) entre sequências genéticas, que constitui parte essencial do problema de pesquisa deste trabalho.

Busca por *motifs* Regulatórios

No livro *An Introduction to Bioinformatics Algorithms*, Stormo (STORMO, 2004) discute o método de busca exaustiva como uma abordagem inicial para tratar a busca por *motifs* regulatórios (*regulatory motif finding*). Uma outra estratégia proposta por Stormo apresenta um problema semelhante: a utilização da *string* mediana para demonstrar a equivalência computacional entre os dois problemas. O problema da *string* mediana é utilizado para solucionar a busca por *motifs* juntamente com a técnica de poda das Árvores de Busca (*search trees*), constituindo uma abordagem por força bruta.

Montanari (MONTANARI et al., 2016) traz uma solução implementada que está alinhada com o problema em questão através de uma implementação do algoritmo R-MBP (*Root-element Best-matching Problem*) utilizando a técnica de Programação Dinâmica (*DP - Dynamic Programming*, em inglês). No trabalho, o problema de busca pelos padrões é definido como pontos de interesse em uma trilha (*track*) no qual as regiões do genoma são reduzidas aos pontos. O casamento (*match*) entre a *query* e uma trilha é definido através de uma função injetora, onde o custo é definido em função dos *matches* iniciais. Sendo assim, o problema tratado pelo R-MBP consiste em encontrar e determinar uma função $f*$ com o menor custo por *match* para duas trilhas de entrada.

Fan (FAN et al., 2015) utilizaram uma abordagem promissora adotando uma estratégia de variação no comprimento dos *motifs*. Através do uso de um *framework* de Algoritmos Genéticos (GA), os autores desenvolveram um algoritmo chamado de ALDILM. O algoritmo lida com a busca de *motifs* ótimos em sequências de DNA comparando com um *motif* ótimo pré-determinado. Ele considera um *motif* inicial de tamanho três e uma população de 64 possíveis indivíduos (pois cada sítio do *motif* apenas assume uma das quatro bases nitrogenadas). A partir daí, o tamanho dos indivíduos irá crescer em uma unidade a cada época até que ele atinja um tamanho máximo determinado. Para isso, uma função de escore (*scoring*) é utilizada para encontrar o tamanho e o *motif* ótimos que serão utilizados na análise. Pela natureza genética do algoritmo, três operações são utilizadas: mutação, adição e deleção. A mutação é uma operação

importante pois impede a repetição de *motifs* durante as iterações do algoritmo (no experimento ela foi fixada em 0.2% para garantir a individualidade de cada um dos 64 tipos iniciais destacados pelo autor). A operação de adição confere uma nova análise de escore para os motifs através da adição aleatória de uma base ao final de ambos os *motifs* comparados. Por último, a deleção apenas garante que os novos *motifs* obtidos através da operação de adição possam ser restaurados a sua forma original. O algoritmo foi testado com dados simulados e dados reais. Segundo os autores, os resultados são consistentes com a realidade e similares com o de três métodos conhecidos: *Gibbs Sampler*, *MEME* e *Weeder*. No que diz respeito ao desempenho, ele possui desempenho similar ou até melhor que os métodos mais utilizados, porém, dada sua natureza estocástica, não há garantias de que ele consiga achar o tamanho e *motif* ótimos em todas as ocasiões.

Al-Ouran (AL-OURAN et al., 2015) trata o problema com métodos combinatórios como o RILP (*Relaxed Integer Linear Programming*). Nele o problema é abordado como o um outro problema (*Set Cover Problem*) onde o problema de seleção de *motif* é tratado como um um programa linear inteiro 0-1 e o objetivo é encontrar um vetor 0-1 com tamanho que satisfaça a restrição de que, o subconjunto encontrado seja o menor possível que case com o conjunto inicial (chamado de conjunto universo). Esse algoritmo utiliza aproximação em tempo linear, ou seja, é possível obter uma solução em tempo polinomial no pior caso.

Brown (BROWN, 2012) adotou uma técnica evolucionária conhecida como mundos múltiplos (*Multiple Worlds*), onde agentes da população estudada devem evoluir para se especializarem um papel. Ela utiliza uma função de *fitness* para avaliar a adaptação da população.

Maiti e Mukherjee (MAITI; MUKHERJEE, 2015) utilizam o método de Monte-Carlo para maximizar a acurácia de descoberta de padrões nas sequências, onde há a seleção da cadeia de Markov mais promissora e é introduzido um fator de aleatoriedade. Após isso, é feita uma simulação para atualizar uma variável de entrada θ. Isso ajuda o algoritmo a evitar um máximo local e o torna mais efetivo na busca pelos padrões. Como conclusão, é feita uma comparação entre a abordagem tradicional (sem o fator de aleatoriedade) e a nova abordagem descrita pelos pesquisadores.

Falah, Maroua e Mourad (FALAH; GHNIMI; ELLOUMI, 2014) discutem um algoritmo novo, criado pelos próprios autores, chamado *SMS_H_CCA*. Ele recebe um conjunto de *strings*, um limiar e dois *quorums* e devolve um conjunto contendo o os padrões mais específicos com, no máximo, tamanho l. Os resultados dos experimentos foram conduzidos em dados pseudo-aleatórios gerados usando um algoritmo chamado de *KISS* em dois alfabetos de tamanho quatro (para simular sequências de DNA) e 20 (para simular sequências de proteínas). Por último, foi mensurado o tempo de processamento para os experimentos com DNA e proteínas. Eles concluem que, embora o *SM_H_CCA* não consiga lidar com todas as variações de padrões como o *SMS-H-Forbid*, o novo algoritmo consegue achar rapidamente os padrões mais específicos e, portanto, os mais relevantes.

Autor	Ano	Abordagem	Método
Montanari et al.	2016	Programação Dinâmica	RMBP
Fan et al.	2015	Algoritmos Genéticos	Método Próprio
Al-Ouran et al.	2015	Análise Combinatória	RILP
Maiti e Mukherjee	2015	Algoritmos Probabilísticos	Monte-Carlo
Falah, Maroua e Mourad	2014	-	Método Próprio
Brown	2012	Computação Evolucionária	Mundos Múltiplos
Stormo	2004	Força Bruta	Busca Exaustiva

Tabela 4.1: Tabela de Resumo de Trabalhos Relacionados. Fonte: o autor.

Considerações Finais

Os trabalhos relacionados aqui mencionados propõem adaptações de métodos e estratégias clássicas para melhor solucionar o problema de busca de *motifs* entre sequências de DNA e RNA. O problema de pesquisa aqui proposto envolve, também, a descoberta de *motifs* entre duas sequências de entrada, que pode ser realizada através da busca exaustiva ou por uma abordagem semelhante àquela proposta no ALDILM, e a subsequente busca por ocorrências de tais *motifs*, que pode ser realizada através dos métodos apresentados nos demais trabalhos mencionados. Tais abordagens podem aplicadas ao um conjunto de dados aqui utilizado. No entanto, por se tratar de dados inéditos e, buscando considerar as especificidades do mesmo, optamos por utilizar a busca exaustiva com o objetivo de alinhar com a técnica computacional, estratégias habitualmente utilizadas na biologia molecular para esses dados em específico.

PARTE II

Materiais e Métodos

CAPÍTULO 5

Metodologia

"Método científico é o conjunto de processos ou operações mentais que se devem empregar na investigação. É a linha de raciocínio adotada no processo de pesquisa."

—GIL, 1999; LAKATOS;MARCONI, 1993

5.1 Da Natureza da Pesquisa

Este trabalho configura-se como uma pesquisa empírica de natureza quantitativa. A abordagem é dedutiva e de caráter exploratório e descritivo, tendo como principal objetivo descobrir e interpretar dados obtidos à partir de experimentos e observação (GERHARDT; SILVEIRA, 2009).

5.2 Fases da Pesquisa

Nesta seção são apresentadas as fases da pesquisa, desde a obtenção do banco de dados até a análise final dos dados.

Fase I – Montagem do Banco de Dados de Sequências:
A fase inicial consistiu na obtenção das sequências de entrada de repositórios *online*. Esta fase está detalhada na Seção 5.4.

Fase II – Estudo das Características das Sequências:
Nesta fase foram realizados estudos e reuniões com os biólogos, afim de entender a biologia por trás dos patógenos e propriedades gerais dos genomas e como os mecanismos de evolução e transferência lateral de genes determinariam as mudanças nas sequências analisadas.

Fase III – Definição da Heurística de Análise
Durante essa fase foram elaboradas estratégias voltadas para a obtenção das informações que se desejavam analisar, através da união de conceitos computacionais e biológicos definidos numa heurística de busca.

Fase IV – Implementação da Heuristica
Nesta fase foi realizada a implementação dos procedimentos de busca utilizando abordagens computacionais e os materiais definidos nesta metodologia. Esta etapa está detalhada na Seção 5.5.

Fase V – Análise Estatística dos Resultados
Finalmente, nesta fase os dados obtidos foram analisados utilizando métodos e ferramentas

estatísticas. Os resultados obtidos serão apresentados nos próximos capítulos. Esta fase está detalhada na Seção 5.6.

5.3 Materiais

Nesta seção os materiais utilizados na obtenção e análise dos dados, tais como Python e R, são apresentados de forma breve.

5.3.1 Python

Python é uma linguagem de alto-nível, interpretada e multi-paradigma. Ela foi pensada com foco na produtividade e legibilidade, sendo possível expressar procedimentos de maneira natural e próximo de notações matemáticas. Sua tipagem é dinâmica, forte e *duck typing*, o que a torna uma linguagem interessante para o desenvolvimento de aplicações rápidas. Por ser uma linguagem interpretada, os códigos escritos podem ser diretamente executados no interpretador, dando margem para testes rápidos. Python conta com uma vasta biblioteca padrão e suporta módulos e pacotes, o que encoraja modularidade e reuso de código. Além disso, Python também conta com diversos pacotes mantidos por sua comunidade de usuários, e, dentre eles, pacotes voltados à biologia computacional, como o Biopython, que possui diversas funções e módulos prontos para uso, agilizando a implementação de soluções dentro do contexto biológico, como a função de leitura de cabeçalhos de arquivos FASTA (FOUNDATION, 2018a).

5.3.2 R

R é uma linguagem funcional. Ela é um dialeto da linguagem S, criada por John Chambers, que permite o usuário realizar análises estatísticas afim de visualizar, transformar e modelar dados de modo interativo. R pode ser utilizado para plotar curvas, fazer análise de agrupamentos (*clusters*), análise de *microarrays*, classificações, genômica comparativa, e até mesmo aprendizado de máquina, modelagem e simulação (FOUNDATION, 2018b).

5.3.3 Pacote Vegan

Vegan é um pacote de funções e métodos estatísticos voltados principalmente para ecologistas. Seus métodos incluem todos os métodos de ordenação mais comuns: análise de componentes principais, análise de correspondência, análise de correspondência destendenciada e escalonamento multidimensional não-métrico (OKSANEN, 2015).

5.4 Montagem do Banco de Dados de Sequências

A montagem do banco de dados de sequências se deu através da obtenção de arquivos FASTA referentes as sequências de nucleotídeos e aminoácidos dos genes do HPV, disponíveis na plata-

forma *Papillomavirus Episteme* (<https://pave.niaid.nih.gov/>). Essa plataforma disponibiliza informações genéticas curadas sobre os diferentes tipos de papilomavirus de forma sistemática. As sequências de DNA dos plasmídeos da clamídia foram obtidos no banco de dados *Genome*, disponibilizado pelo NCBI (<https://www.ncbi.nlm.nih.gov/>) também em formato FASTA. Posteriormente, os genomas obtidos foram processados em *softwares* como o *MAUVE* 2.4.0 e *Islandviewer* 4.0, afim de identificar ilhas cromossômicas e regiões similares relacionados à virulência e patogenicidade (SILVA, 2017).

5.5 Análise Comparativa das Regiões

Para realizar a comparação entre as regiões homólogas dos patógenos foram obtidos dados referentes ao conteúdo-GC (homogeinização), propriedades físico-quimicas dos aminoácidos, uso de códon (taxa de códons dentro da sequência) e *motifs* comuns às duas sequências.

5.5.1 Tradução de Códons e Síntese de Proteínas

Realizar a síntese de aminoácidos é um procedimento simples. Os aminoácidos que irão compor uma proteína são determinados pela sequência de bases presentes no DNA que são transcritas para o RNA_m, que será lido num passo de três em três bases pelo ribossomo para realizar a tradução na sequência de proteínas. Considere que exista uma estrutura semelhante a um dicionário que possui dois atributos: uma chave única, que identifica um índice no dicionário e um valor, podendo ser o valor um conjunto de valores, como visto abaixo:

```
dicionário -> chave:{valor}
```

Como um conjunto de códons podem ser traduzidos em um mesmo aminoácido (códons sinônimos), e para cada códon possível existe apenas uma correspondência de aminoácido, na estrutura de dicionário cada aminoácido é uma chave que possui um conjunto de códons que traduzem neste aminoácido. Desta forma, o dicionário terá a seguinte estrutura:

```
dicionário -> Aminoácido:{codon}
```

As correspondências de tradução foram determinadas conforme o círculo de tradução encontrado no Apêndice A. O procedimento de definição do dicionário de aminoácidos utilizou a técnica de compreensão de listas, suportada pelo Python. A sintaxe de compreensão de lista foi influenciada pela notação matemática dos conjuntos, onde, matematicamente:

$$S = \{x^2 : x \in \{0 \ldots n\}\}$$

o que se traduz como uma lista L e $L =$ [expresão é executada [se `condição`]], em pseudo-código. Sendo assim, considere um alfabeto de nucleotídeos $\Sigma = \{A, C, G, T\}$ e uma biossequência de DNA S tal que $S = (s_1, s_2, ..., s_n) : s_i \in \Sigma$. Um dicionário de aminoácidos pode ser construído à partir de σ ao percorrer uma lista de triplas de nucleotídeos e associar cada tripla a uma chave (aminoácido) do dicionário.

5.5.2 Análise de Propriedades Físico-químicas

Para analisar as propriedades físico-quimicas dos aminoácidos dos peptídeos resultante da tradução das sequências, foi utilizado o pacote *pepdata 0.7.0* (<https://pypi.python.org/pypi/pepdata/0.7.0>). Nele, o módulo `amino_acid` conta com uma variedade de funções incluindo a análise de propriedades físicas e químicas para ambos resíduos de aminoácidos e interações entre pares de resíduos. As propriedades físico-químicas analisadas são mostradas na Figura 5.1.

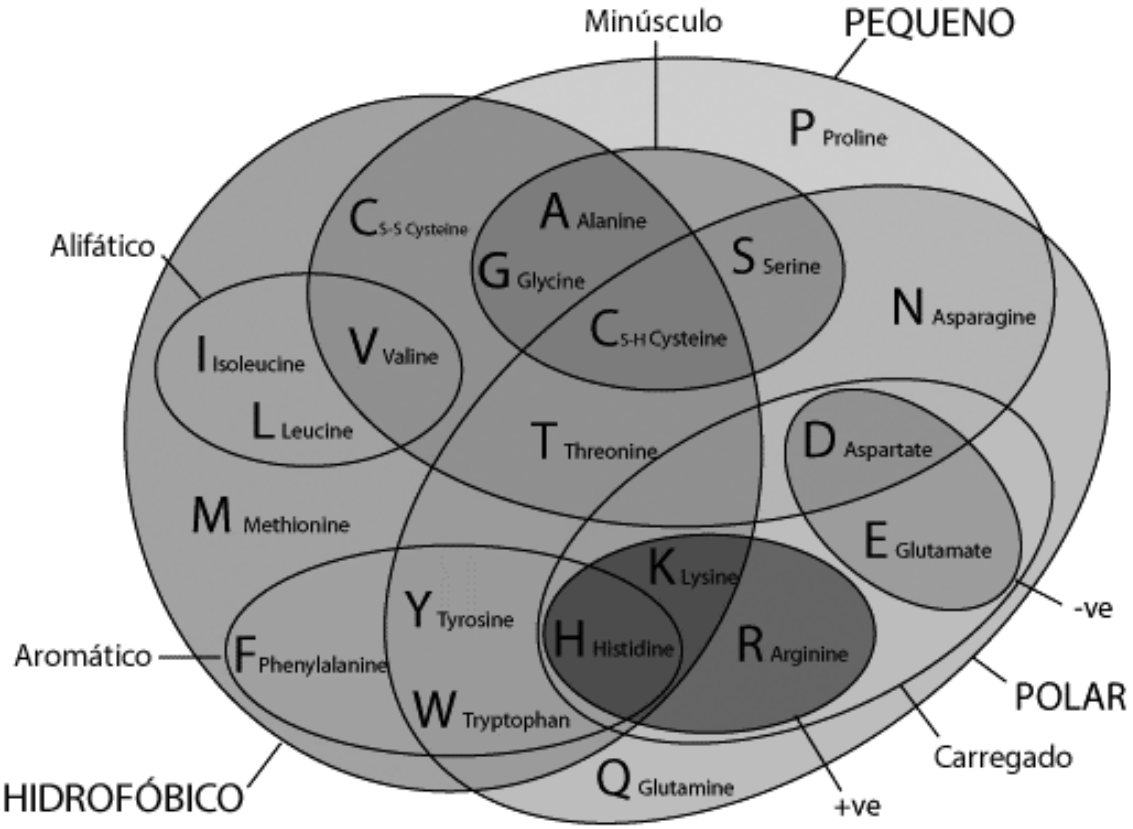


Figura 5.1: Propriedades físico-químicas dos aminoácidos. Cada propriedade está assicada a uma função da proteína resultante. Se, por exemplo, a proteína for formada em sua maioria por códons de propriedade hidrofóbica, isso pode significar que esta proteína esta associada com funções estruturais, como a formação da membrana plasmática. Fonte: (LIVINGSTONE; BARTON, 1993)

5.5.3 Homogeinização dos Genomas — Determinando o Conteúdo-GC

Determinar o conteúdo-GC é trivial, pois envolve apenas contar as ocorrências das bases na biosequência e, através de uma razão simples, determinar o percentual das ocorrências das bases desejadas.
Seja um alfabeto $\Sigma = \{A, C, G, T\}$. Seja S uma palavra sobre Σ, tal que $S = (s_1, s_2, ..., s_n) : s_i \in \Sigma_S$.

$$\text{Seja uma função } Q(x) = |\{i \in \mathbb{N} : s_i = x\}|$$

O percentual de conteúdo-GC é dado por:

$$P_{GC} = \frac{Q(C)+Q(G)}{Q(A)+Q(C)+Q(G)+Q(T)} \times 100$$

No entanto, pela propriedade de pareamento complementar, podemos determinar o conteúdo-GC por:

$$P_{GC} = \frac{Q(G)\times 2}{Tamanho(S)} \times 100$$

Então, o conteúdo-GC é definido como o percentual de ocorrências dos símbolos C ou G na cadeia. Isso pode ser calculado com um percurso na cadeia, em complexidade proporcional ao comprimento da cadeia.

5.5.4 Encontrando *Motifs*

Por fim, para descobrir *motifs* nas sequências de DNA e RNA foi aplicado o método de força-bruta, o qual consiste em gerar e buscar por todas as possibilidades. Como estamos tratando do problema de busca por *motifs* como o Problema de Busca da *String* Mediana, chegamos ao seguinte algoritmo descrito em (STORMO, 2004):

```
   Entrada: biossequências S de DNA, t, n, l
   Saída: melhor motif encontrado
 1 início
 2     melhorMotif <- AAA...AA
 3     melhorDistancia <- ∞
 4     para cada l-mer de AAA...A até TTT...T faça
 5         se DistanciaTotal(l-mer, S) < melhorDistancia então
 6             melhorDistancia <- DistanciaTotal(l-mer, S)
 7             melhorMotif <- l-mer
 8         fim
 9     fim
10 fim
11 retorna melhorMotif
```

Algoritmo 1: Procedimento para determinar o melhor conjunto de motifs entre um par de sequências.

Na heurística definida, decidimos utilizar uma estratégia de *frames* de leitura, que consistem em analisar as sequências de seis modos distintos, como mostra a Figura 5.2. Se existe uma sequência S, onde $S = (s_1, s_2, ..., s_i)$ é o conjunto de posições iniciais dessa sequência, sendo o primeiro *frame* iniciado na posição $i = 0$ da sequência. Os dois *frames* subsequentes iniciam na posição $i+1$ e $i+2$ respectivamente. Os três *frames* restantes seguem a mesma estratégia, porém com a sequência ao inverso. A razão pela qual os cinco *frames* estão definidos desta forma é para garantir que *motifs* que tenham sofrido mutação em sua segunda ou terceira base (*frames* dois e três) sejam considerados na busca, e para os frames restantes, a leitura reversa certifica que os *motifs* na leitura 3' -> 5' foram cobertos (STORMO, 2004). Portanto, o procedimento acima foi executado seis vezes para cada tamanho l de *motif*, sendo $3 \leq l \leq 10$.

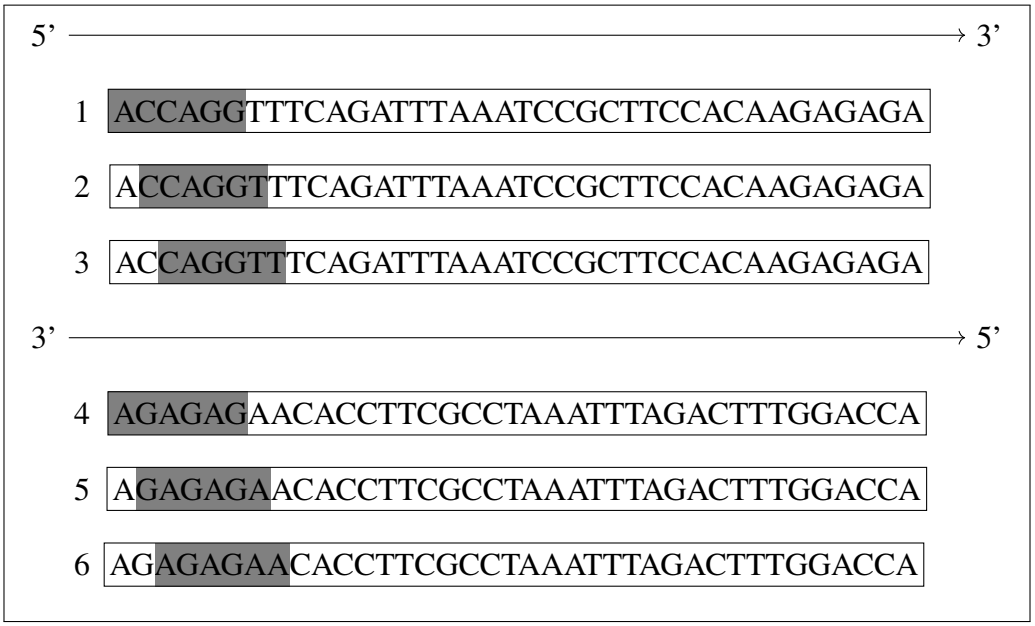


Figura 5.2: Os seis *frames* de leitura. No exemplo, o tamanho do *motif* procurado foi fixado em seis bases.

Após identificar todos os motifs das sequências de entrada, eles são comparados nas sequências como mostra a Figura 5.3 e armazenados em uma lista, onde suas posições e quantidades de repetições na sequências são também armazenadas. Por fim, a lista é transformada numa tabela e salva numa planilha com as outras informações levantadas previamente.

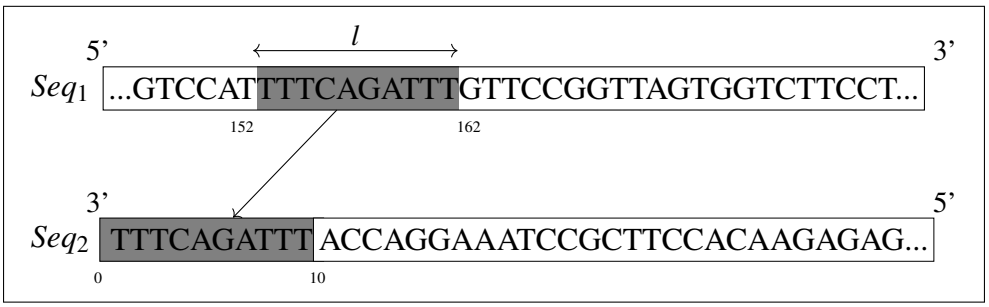


Figura 5.3: Identificando ocorrências de *motifs* comuns as sequências.

5.6 Análise de Resultados

Estatística Multivariada — Análise de Correspondência

Os resultados obtidos foram armazenados em planilhas para cada sequência que armazenava as informações obtidas pelos procedimentos descritos anteriormente. No entanto, devido ao grande volume de dados levantados, tornou-se inviável realizar uma análise univariada e, portanto, decidimos utilizar uma técnica de análise multivariada. A análise multivariada é indicada para quando existe um maior número de variáveis, pois ela consiste em condensar os dados em componentes principais. Segundo Mingoti (2005), os métodos de estatística multivariada tem como propósito simplificar ou facilitar a interpretação dos dados concentrando em um mesmo plano as informações contidas em um universo multidimensional. Para poder fazer esta análise, no entanto, foi necessário mudar a disposição dos dados nas planilhas e condensar as 276 planilhas obtidas em seis, uma planilha para cada *frame* de leitura. O modelo final de planilha obtido pode ser visto na Tabela 5.1.

Sequências	%GC	Códons Presentes				Nº de motifs Econtrados			
		UGU	UCU	AUA	...	Tam. 3	Tam. 4	Tam. 5	
Sequência 1									...
Sequência 2									
Sequência 3									
⋮									

Tabela 5.1: Modelo reorganizado de planilha utilizada para fazer a análise multivariada.

A análise de correspondência canônica é uma técnica de ordenação multivariada que consiste na análise exploratória de dados categorizados, sendo um método de associação entre os elementos de dois ou mais conjuntos de dados, buscando estabelecer uma estrutura de associação dos fatores em questão. Ela é comumente utilizada por ecologistas afim de identificar agrupamentos e relações entre espécies e variáveis ambientais, através de gráficos que permitem a visualização da relação entre os conjuntos, revelando relações que não teriam sido percebidas se a análise fosse feita aos pares de variáveis (CZERMAINSKI, 2004; LUCIO; TOSCANO; ABREU, 1999).

A análise de correspondência foi realizada no R, utilizando o pacote *vegan* que disponibiliza a função `cca()`. Após importar os dados para o ambiente, os nomes das sequências foram anexados às amostras em uma lista. Optamos por desconsiderar os dados referentes aos *motifs* de tamanho três e quatro, pois observamos que haviam muitas ocorrências de repetições do mesmos devido ao seu tamanho, gerando ruído na análise. Após organizar as amostras no R, foi possível executar a análise de correspondência e representar no gráfico a tendência das amostras, considerando as propriedades de homogeinização, uso de códon e quantidades de *motifs* em comum entre as sequências. O gráfico de resultados foi desenhado com a opção `scaling=3`, configurada para permitir uma melhor visualização das amostras agrupadas (menos espalhamento).

PARTE III

Conclusão

CAPÍTULO 6

Resultados e Discussões

Após realizar a análise de correspondência no R, os gráficos gerados mostraram que houve um agrupamento das amostras em três grupos. O posicionamento das amostras (e dos agrupamentos) se dá pela associação das amostras com as variáveis através dos cálculos de inércia e da qualidade. A inércia é uma medida de dispersão entre as variáveis da tabela dada pelo Qui-quadrado de Pearson, dividido pelo total das frequências. A qualidade diz respeito a confiabilidade da representação dos pontos no sistema de coordenadas definido pelo número de dimensões escolhido para a análise. Ela é definida pela razão entre o quadrado da distância do espaço definido pelo número de dimensões escolhidas. Portanto, quanto mais próximo de um, melhor será a representação do ponto no espaço escolhido. Os gráficos completos obtidos para os seis *frames* podem ser vistos no Apêndice B e também estão disponíveis em <https://drive.google.com/drive/folders/1mplUO87kGWVhTA2AZlrj5GPfFVCzcc-C?usp=sharing>. Abaixo, um resumo com cortes dos gráficos é apresentado na Figura 6.1.

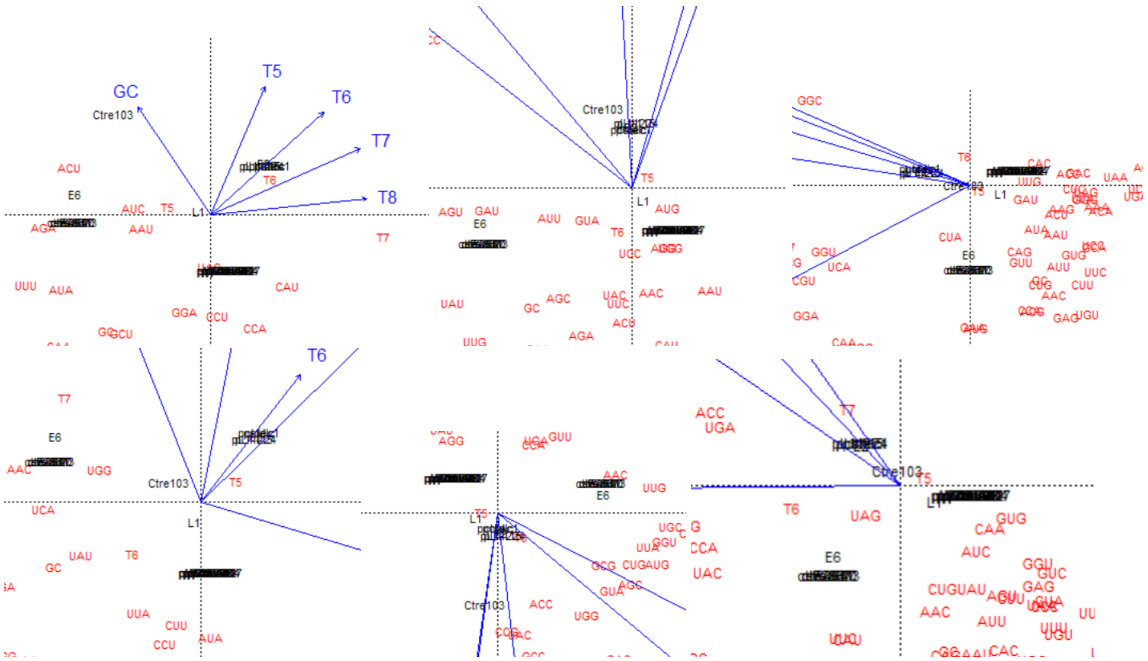


Figura 6.1: Gráficos resultantes da análise de correspondência para os seis *frames* de leitura. Fonte: o autor.

Em se tratando de organismos filogeneticamente distantes, espera-se observar um distanciamento natural entre as suas propriedades genéticas, o que diz respeito às particularidades

de cada espécie. No entanto, os gráficos resultantes mostraram que, em todos os *frames* de leitura, as regiões tendem a se agrupar, com exceção da CTre103, que se manteve a uma distância razoável de todos os agrupamentos em todos os gráficos. O fato de que estes organimos coexistem no mesmo ambiente, juntamente com a análise realizada neste trabalho revela um forte indício da existência de interações a níveis moleculares, como a transferência lateral de genes, entre estes organismos. Os resultados aqui apresentados estão próximos aos encontrados por (SILVA, 2017), conforme mostra a Figura 6.2, onde, através de investigação manual, ele chegou em um agrupamento que segue os sorotipos E, L2b e L1 do HPV, que também se repetiram nos agrupamentos da análise de correspondência.

Plasmídeos	Gene CT	Região (pb)	Gene HPV	Região (pb)	Sorotipo
ctrE-103	*pGP3*	*269 - 374*	*L1*	593 - 698	E
CtrE-547	*pGP6*	*128 - 303*	*E6*	263 - 445	E
ctre8873	*pGP6*	*128 - 303*	*E6*	263 - 445	E
ctredk20	*pGP6*	*128 - 303*	*E6*	263 - 445	E
e32931	*pGP6*	*128 - 303*	*E6*	263 - 445	E
f6068	*pGP6*	*128 - 303*	*E6*	263 - 445	F
pAms1	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pAms2	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pAms3	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pAms4	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pAms5	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pctdec1	*pGP2*	*492 - 629*	*E2*	320 - 458	D-EC
pctdlc1	*pGP2*	*492 - 629*	*E2*	320 - 458	D-LC
pL2b795	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2b820007	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2bCan1	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2bCan2	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2bCV204	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2bLST	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL2bUCH2	*pGP4*	*79 - 263*	*L1*	791 - 972	L2b
pL1115	*pGP2*	*460 - 740*	*E2*	784 - 1079	L1
pL1224	*pGP2*	*460 - 740*	*E2*	784 - 1079	L1
pL11322	*pGP4*	*79 - 263*	*L1*	791 - 972	L1

Figura 6.2: Resultados encontrados por análise manual. Fonte: (SILVA, 2017).

CAPÍTULO 7

Conclusão

O trabalho aqui apresentado trouxe a aplicação de um método computacional (busca exaustiva) em um problema real da biologia, constituindo um trabalho multidisciplinar. Uma heurística de busca foi definida, observando as restrições dsa características do problemas e dos próprios biólogos. Após a implementação da heurística de busca, os dados obtidos foram analisados através do método estatístico de análise de correspondência e mostraram que houve um agrupamento das sequências que também foi observado em (SILVA, 2017).

7.1 Impacto da Pesquisa

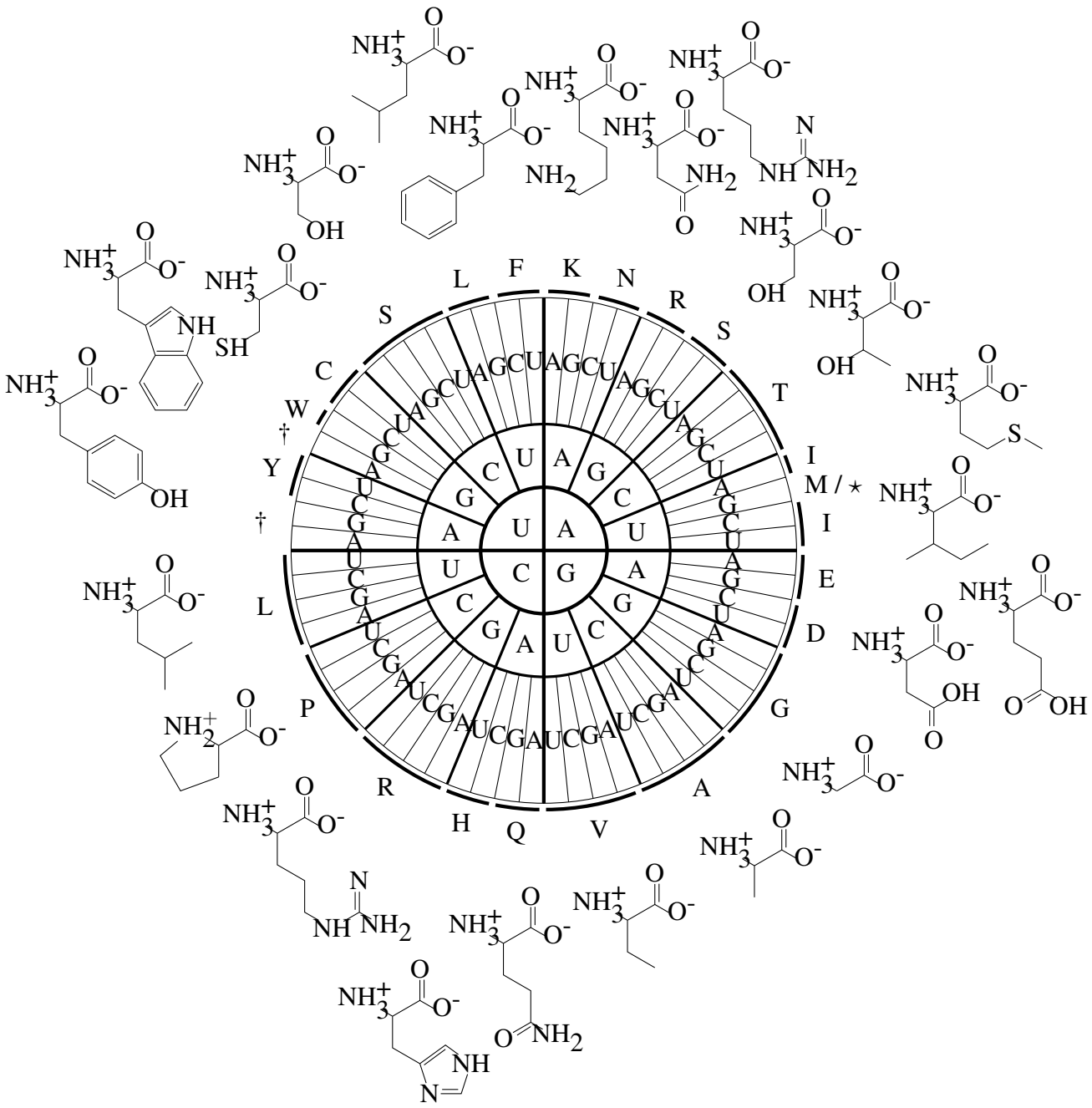
Os resultados obtidos por meio da heurística implementada podem auxiliar no entendimento e na validação e e formação de novas hipóteses sobre sobre a natureza, caminhos e implicações de saúde advindos da interação entre os patógenos estudados.

7.2 Trabalhos Futuros

Como trabalho futuro, pretendemos fazer análises mais detalhadas dos resultados através de implementações de soluções mais sofisticadas, o que nos fornecerá dados mais precisos para uma análise estatística mais robusta. Uma análise biológica dos dados obtidos também está prevista, à partir da qual podemos acrescentar especificidades e explorar as características dos elementos detectados. Por exemplo, após descobrir todos os *motifs* de um determinado comprimento l, é possível chegar a um *motif* consenso para aquele tamanho e usar essa informação na análise multivariada. Isso poderia refinar o agrupamento ou até mesmo modificá-lo, oportunizando novas interpretações. Além da busca exaustiva, outras técnicas computacionais podem ser aplicadas no problema, como o ALDILM, visto em (FAN et al., 2015), que aplica a abordagem de algoritmos genéticos para tratar do problema, demostrando boa acurácia na descoberta de *motifs* entre sequências conforme demonstrado pelos autores.

APÊNDICE A

Círculo de Tradução de Códons



O círculo de tradução é uma forma de se construir os aminoácidos à partir de trincas de bases nitrogenadas. Para ler o círculo, comece pelo centro e escolha uma base nos próximos dois níveis. O aminoácido resultante está representado por uma letra.

APÊNDICE B

Resultados da Análise de Correspondência

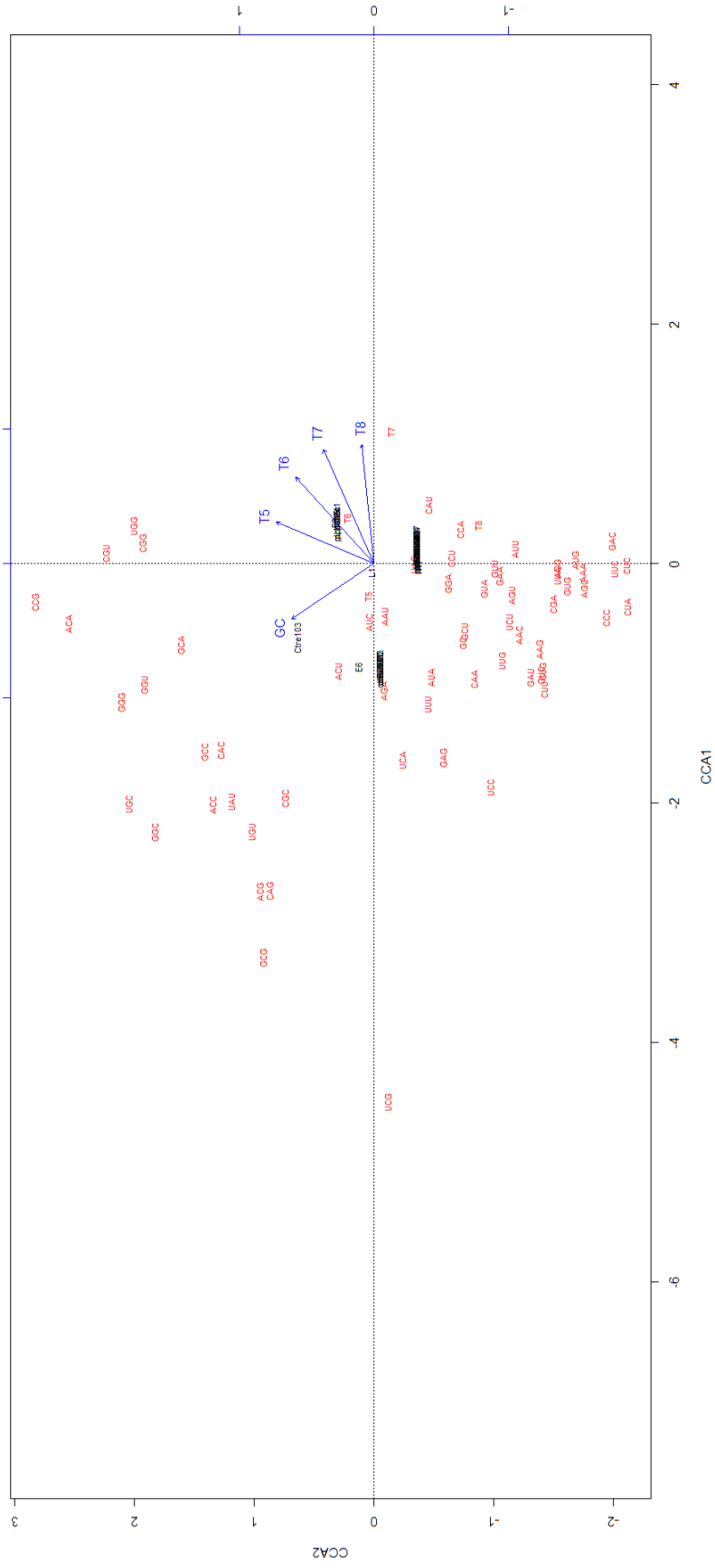


Figura B.1: Frame 1. Fonte: elaborado pelo autor.

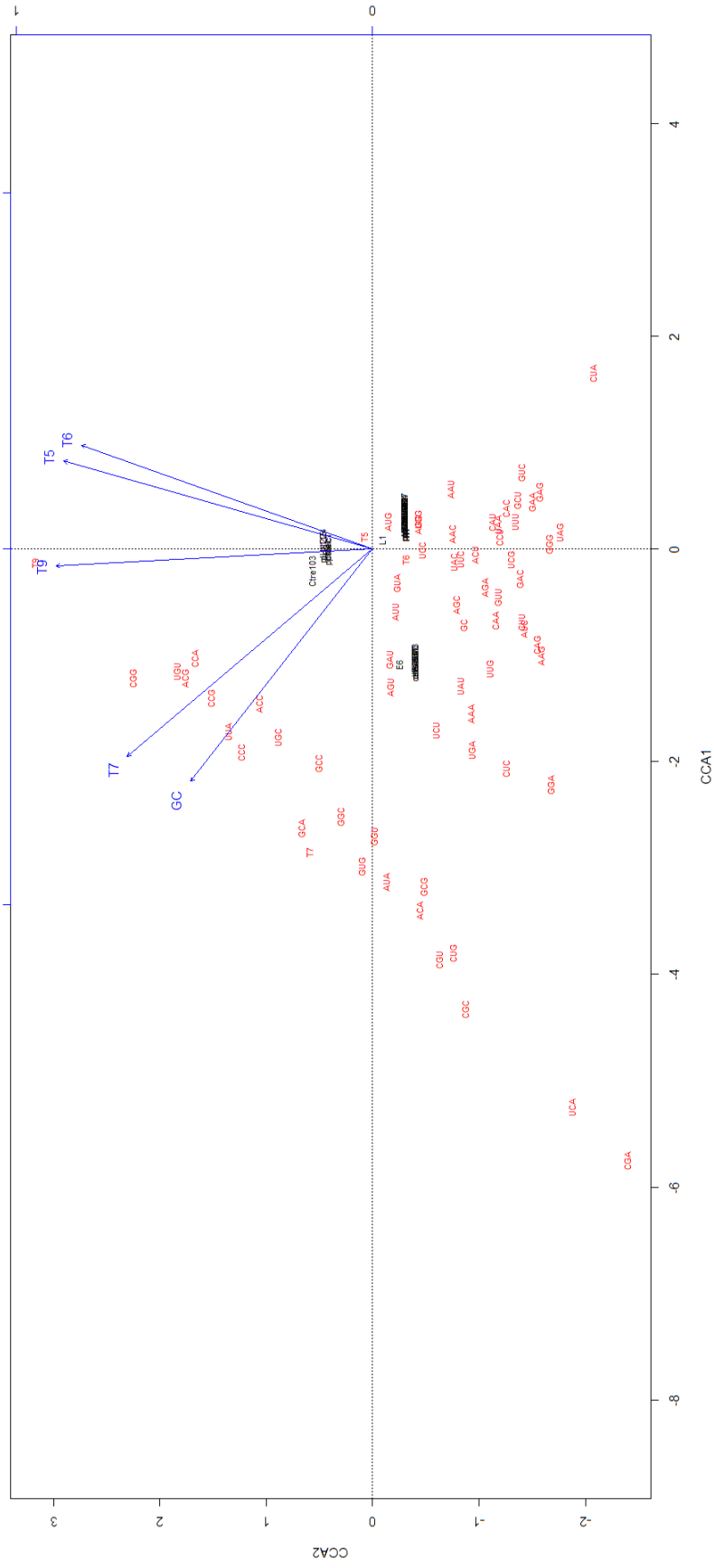


Figura B.2: Frame 2. Fonte: elaborado pelo autor.

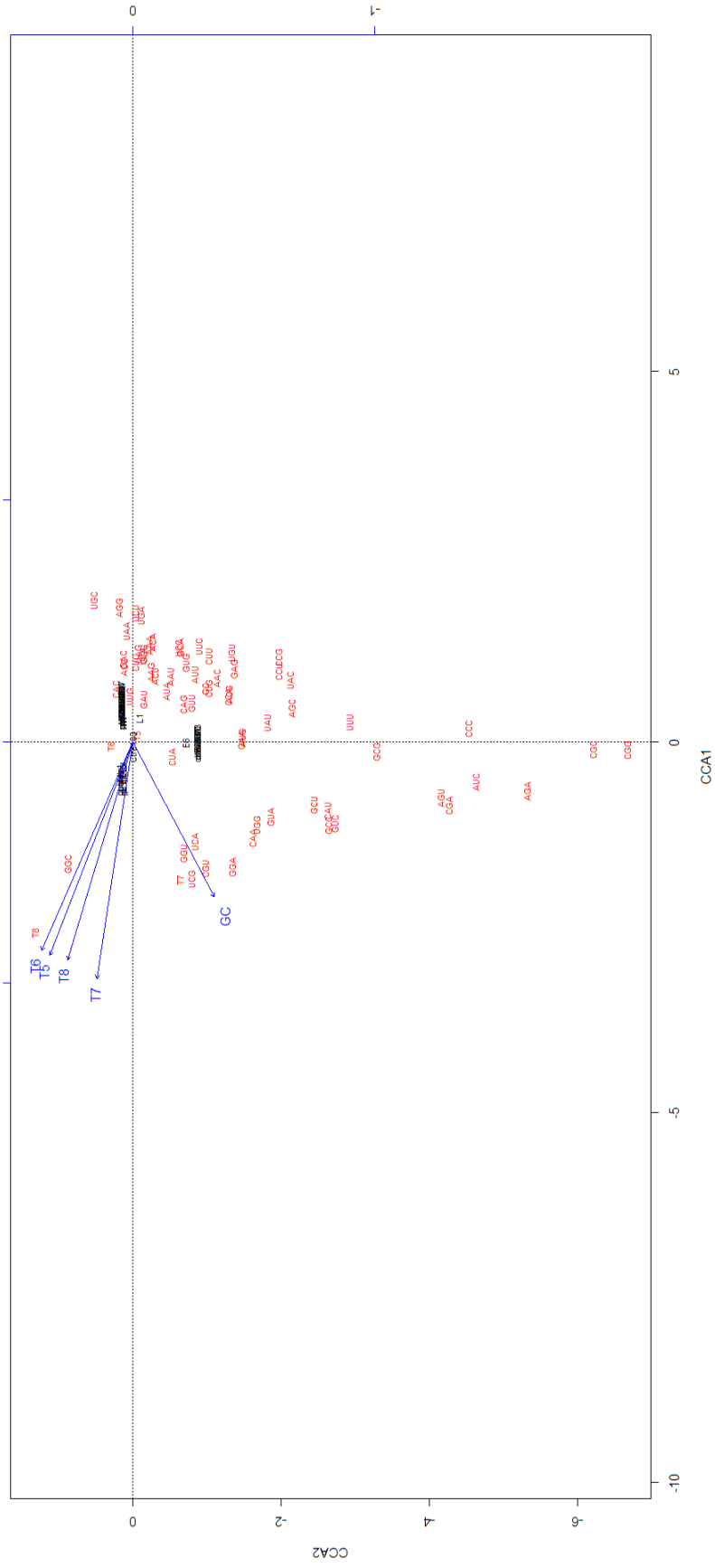


Figura B.3: Frame 3. Fonte: elaborado pelo autor.

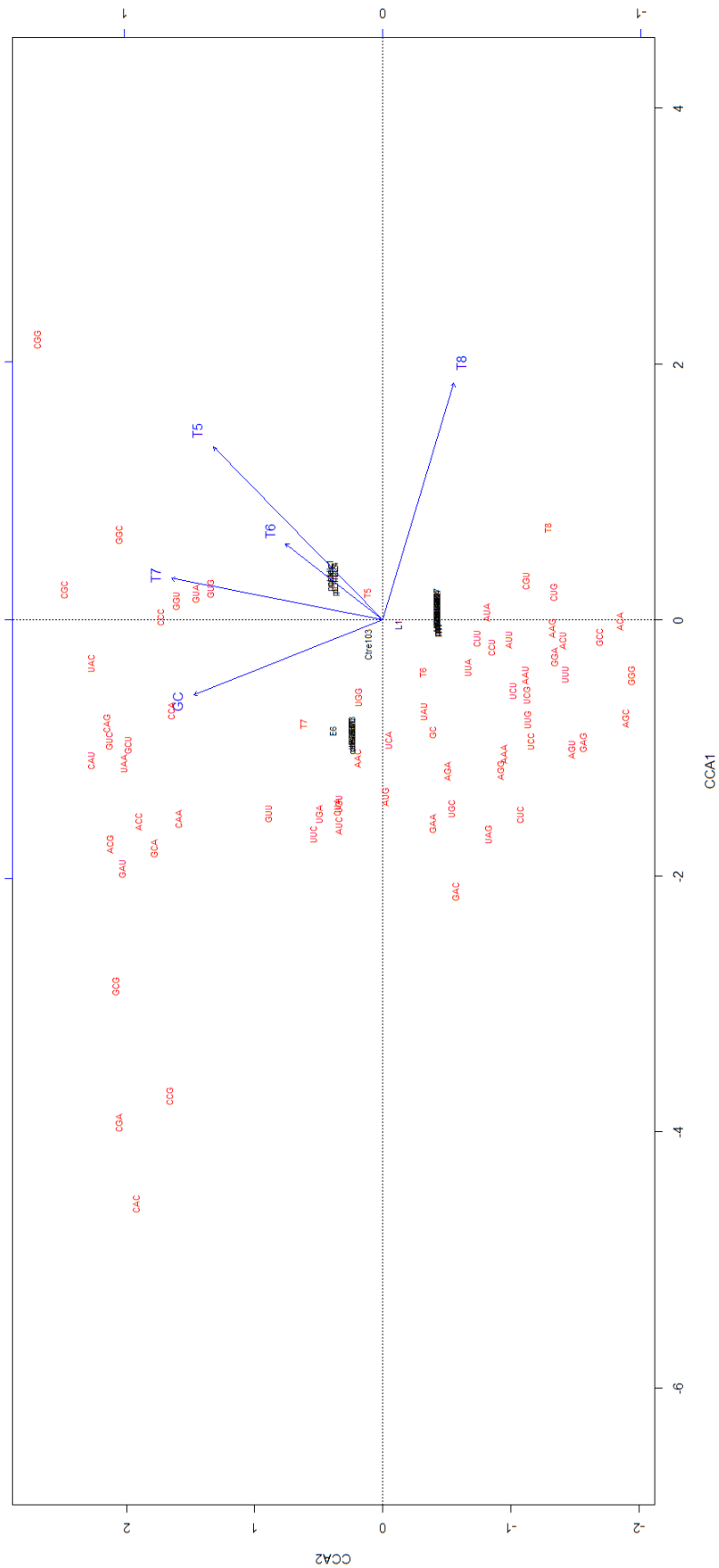


Figura B.4: Frame 4. Fonte: elaborado pelo autor.

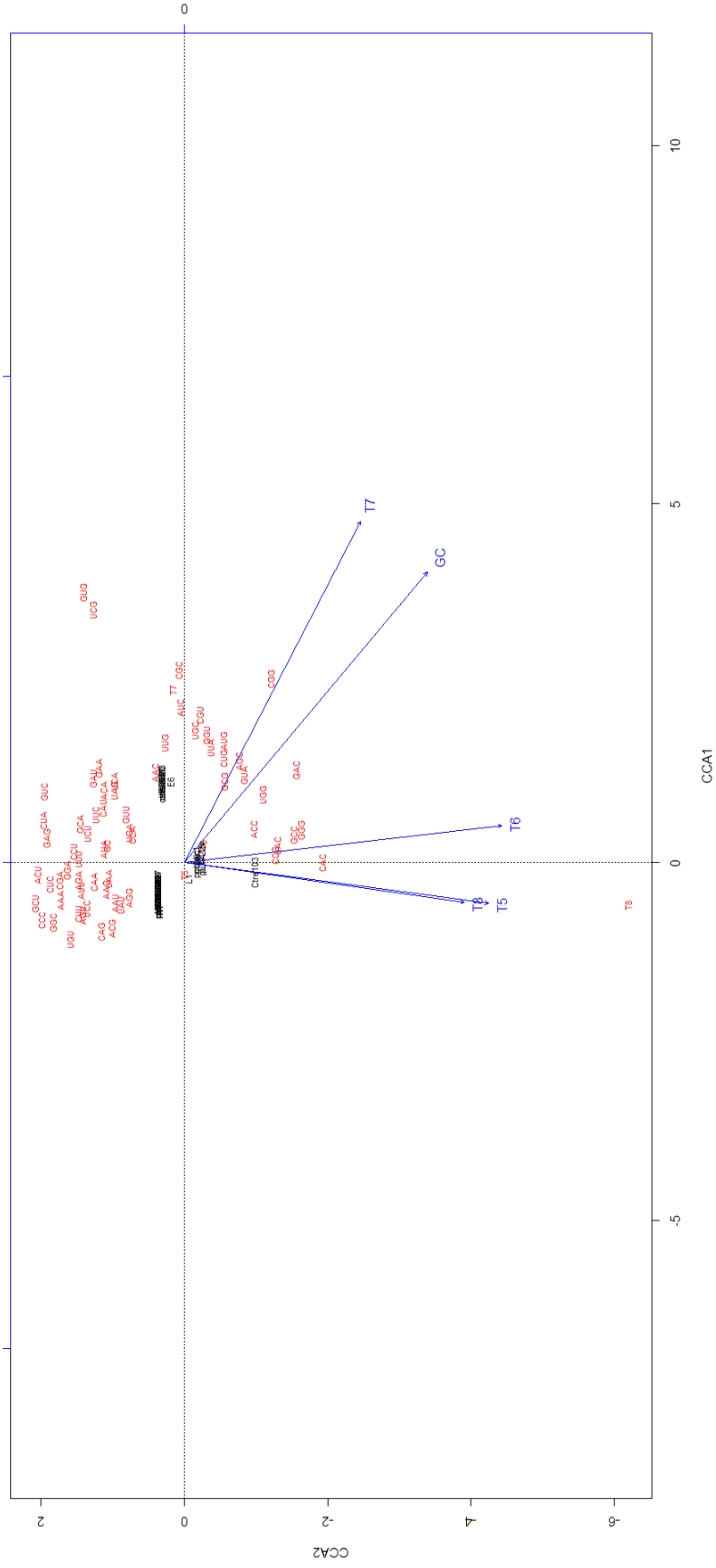


Figura B.5: Frame 5. Fonte: elaborado pelo autor.

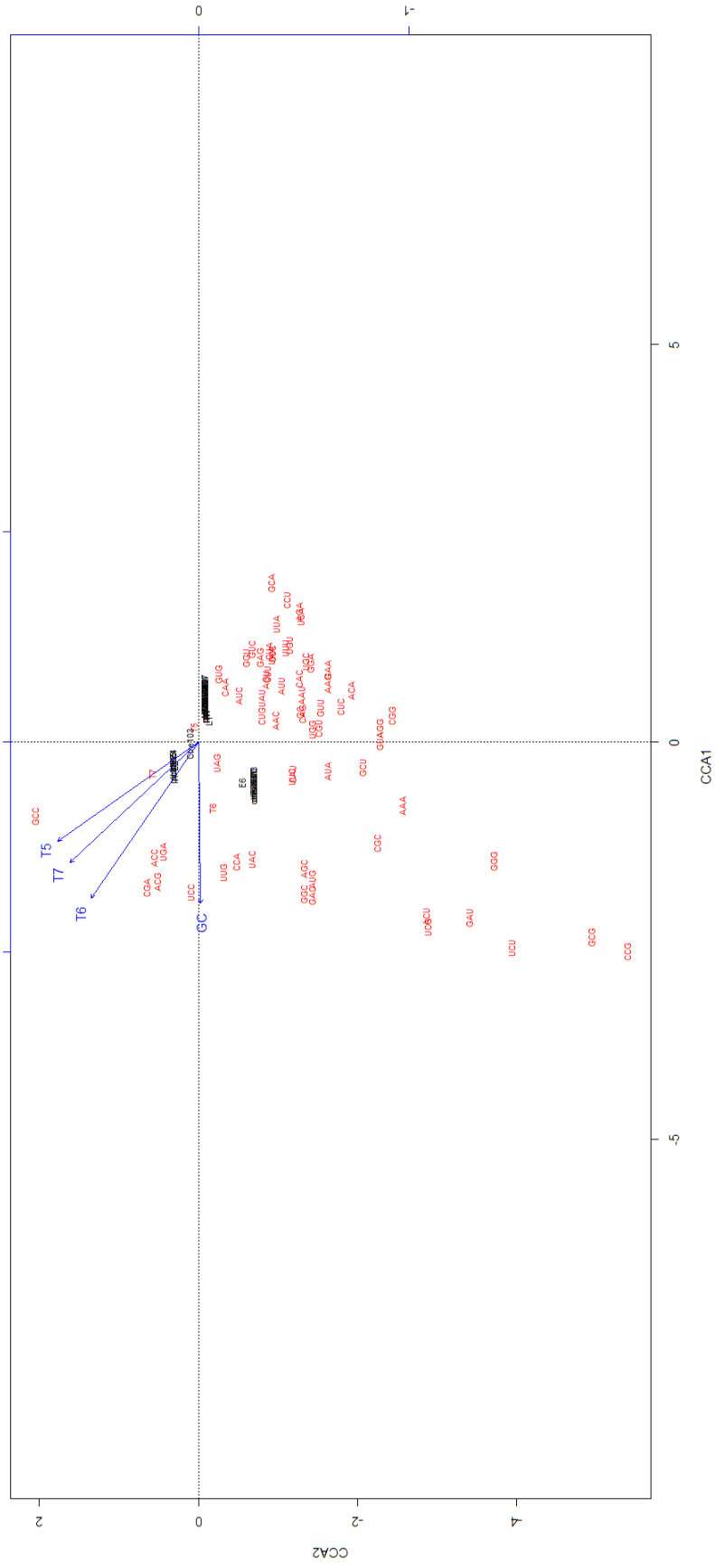


Figura B.6: Frame 6. Fonte: elaborado pelo autor.

Referências Bibliográficas

ACHAR, A.; SÆTROM, P. Rna motif discovery: a computational overview. **Biology direct**, BioMed Central, v. 10, n. 1, p. 61, 2015.

AL-OURAN, R. et al. Discovering gene regulatory elements using coverage-based heuristics. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, 2015.

ALBERTS, B. et al. Molecular biology of the cell, (garland science, new york, 2008). **Google Scholar**, p. 652, 2002.

ARCHETTI, M. Selection on codon usage for error minimization at the protein level. **Journal of molecular evolution**, Springer, v. 59, n. 3, p. 400–415, 2004.

BOURNE, P. E.; BRENNER, S. E.; EISEN, M. B. Ten years of plos‡ computational biology: A decade of appreciation and innovation. **PLoS computational biology**, Public Library of Science, v. 11, n. 6, p. e1004317, 2015.

BREJOVÁ, B. et al. **Project Report for CS798g, University of Waterloo, 2000**.

BROWN, J. A. Multiple worlds model for motif discovery. In: IEEE. **Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on**. [S.l.], 2012. p. 92–99.

CATTADORI, I.; BOAG, B.; HUDSON, P. Parasite co-infection and interaction as drivers of host heterogeneity. **International journal for parasitology**, Elsevier, v. 38, n. 3-4, p. 371–380, 2008.

CECH, T. R. The rna worlds in context. **Cold Spring Harbor perspectives in biology**, Cold Spring Harbor Lab, v. 4, n. 7, p. a006742, 2012.

CHOROSZY-KRÓL, I. et al. Characteristics of the chlamydia trachomatis species-immunopathology and infections. **Adv Clin Exp Med**, v. 21, n. 6, p. 799–808, 2012.

CLIFFORD, G. et al. Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. **British journal of cancer**, Nature Publishing Group, v. 88, n. 1, p. 63, 2003.

CORMEN, T. H. **Introduction to algorithms**. [S.l.]: MIT press, 2009.

CZERMAINSKI, A. B. **Análise de correspondência**. Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo. Piracicaba, 2004.

DOUDNA, J.; COX, M. Biologia molecular-princípios e técnicas. 2012.

EIDHAMMER, I.; JONASSEN, I.; TAYLOR, W. R. Structure comparison and structure patterns. **Journal of Computational Biology**, Mary Ann Liebert, Inc., v. 7, n. 5, p. 685–716, 2000.

FALAH, T. E.; GHNIMI, M.; ELLOUMI, M. A consensus algorithm for simple motifs finding. In: IEEE. **Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on**. [S.l.], 2014. p. 33–37.

FAN, Y. et al. An algorithm for motif discovery with iteration on lengths of motifs. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)**, IEEE Computer Society Press, v. 12, n. 1, p. 136–141, 2015.

FEOFILOFF, P. Análise de algoritmos. **Internet: http://www. ime. usp. br/˜ pf/analise_de_algoritmos**, v. 2009, 1999.

FOUNDATION, P. S. **What is Python? Executive Summary**. 2018. Disponível em: <https://www.python.org/doc/essays/blurb/>.

FOUNDATION, T. R. **What is R?** 2018. Disponível em: <https://www.r-project.org/about.html>.

GAUNT, M. W. et al. Mechanism of genetic exchange in american trypanosomes. **Nature**, Nature Publishing Group, v. 421, n. 6926, p. 936, 2003.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. [S.l.]: Plageder, 2009.

GOUY, M.; GAUTIER, C. Codon usage in bacteria: correlation with gene expressivity. **Nucleic acids research**, Oxford University Press, v. 10, n. 22, p. 7055–7074, 1982.

INSTITUTE, S. H. **James Watson, Francis Crick, Maurice Wilkins, and Rosalind Franklin**. 2017. Disponível em: <https://www.sciencehistory.org/historical-profile/james-watson-francis-crick-maurice-wilkins-and-rosalind-franklin>.

JUNQUEIRA, L. C. U. et al. **Biología celular y molecular**. [S.l.]: McGraw-Hill Interamericana, 1998.

LEMOS, M.; ARAGAO, M. V. S. P.; CASANOVA, M. A. **Padrões em Biossequências**. [S.l.]: PUC, 2003.

LI, W.-H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. **Journal of molecular evolution**, Springer, v. 24, n. 4, p. 337–345, 1987.

LIVINGSTONE, C. D.; BARTON, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. **Bioinformatics**, Oxford University Press, v. 9, n. 6, p. 745–756, 1993.

LOBRY, J. R.; CHESSEL, D. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. **Journal of applied genetics**, INSTITUTE OF PLANT GENETICS, v. 44, n. 2, p. 235–262, 2003.

LUCIO, P.; TOSCANO, E. de; ABREU, M. de. Caracterização de séries climatológicas pontuais via análise canônica de correspondência. estudo de caso: Belo horizonte–mg (brasil). **Brazilian Journal of Geophysics**, v. 17, n. 2-3, p. 193–207, 1999.

M. HASELTINE F., L. Y. H. **NIH Working Definition of Bioinformatics and Computational Biology**. 2000. Disponível em: <http://www.binf.gmu.edu/jafri/math6390-bioinformatics/workingdef.pdf>.

MAITI, A.; MUKHERJEE, A. On the monte-carlo expectation maximization for finding motifs in dna sequences. **IEEE journal of biomedical and health informatics**, IEEE, v. 19, n. 2, p. 677–686, 2015.

MONTANARI, P. et al. Pattern similarity search in genomic sequences. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 28, n. 11, p. 3053–3067, 2016.

NUSSINOV, R. et al. From "what is?" to "what isn't?" computational biology. **PLoS computational biology**, Public Library of Science, v. 11, n. 7, p. e1004318, 2015.

OKSANEN, J. Vegan: an introduction to ordination. **URL http://cran. r-project. org/web/packages/vegan/vignettes/introvegan. pdf**, 2015.

PIERCE, B. A. **Genetics: A conceptual approach**. [S.l.]: Macmillan, 2012.

PRAY, L. Discovery of dna structure and function: Watson and crick. **Nature Education**, v. 1, n. 1, p. 100, 2008.

ROBERTSON, M. P.; JOYCE, G. F. The origins of the rna world. **Cold Spring Harbor perspectives in biology**, Cold Spring Harbor Lab, v. 4, n. 5, p. a003608, 2012.

SCHWARTZ, S. Papillomavirus transcripts and posttranscriptional regulation. **Virology**, Elsevier, v. 445, n. 1-2, p. 187–196, 2013.

SCIACCA, E. Contributions in computational biology. 2009.

SILVA, F. Alphapapillomavirus 9 vs chlamydia trachomatis: Análise genômica e tendências evolutivas. **Universidade Federal Rural de Pernambuco**, 2017.

SIMONETTI, A. C. et al. Immunological's host profile for hpv and chlamydia trachomatis, a cervical cancer cofactor. **Microbes and infection**, Elsevier, v. 11, n. 4, p. 435–442, 2009.

SINGER, M. Pathogen-pathogen interaction: a syndemic model of complex biosocial processes in disease. **Virulence**, Taylor & Francis, v. 1, n. 1, p. 10–18, 2010.

STORMO, G. Exhaustive search. In: **An Introduction to Bioinformatics Algorithms**. [S.l.]: MIT Press, 2004.

SUEOKA, N. Compositional correlation between deoxyribonucleic acid and protein. In: COLD SPRING HARBOR LABORATORY PRESS. **Cold Spring Harbor symposia on quantitative biology**. [S.l.], 1961. v. 26, p. 35–43.

TAVARES, M. C. M. et al. Chlamydia trachomatis infection and human papillomavirus in women with cervical neoplasia in pernambuco-brazil. **Molecular biology reports**, Springer, v. 41, n. 2, p. 865–874, 2014.

WOHLMEISTER, D. et al. Association of human papillomavirus and chlamydia trachomatis with intraepithelial alterations in cervix samples. **Memórias do Instituto Oswaldo Cruz**, SciELO Brasil, v. 111, n. 2, p. 106–113, 2016.