CENTRO UNIVERSITÁRIO SENAI CIMATEC

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

Doutorado em Modelagem Computacional e Tecnologia Industrial

Tese de doutorado

Modelagem computacional aplicada a análise de similaridade e filogenia das proteínas.

Apresentada por: Luryane Ferreira de Souza
Orientador: Marcelo Albano Moret Simões Gonçalves
Coorientador: Bruna Aparecida Souza Machado

Salvador
2023

Luryane Ferreira de Souza

Modelagem computacional aplicada a análise de similaridade e filogenia das proteínas.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial, Curso de Doutorado em Modelagem Computacional e Tecnologia Industrial do Centro Universitário SENAI CIMATEC, como requisito parcial para a obtenção do título de **Doutor em Modelagem Computacional e Tecnologia Industrial**.

Orientador: Marcelo Albano Moret Simões Gonçalves

Coorientador: Bruna Aparecida Souza Machado

Salvador
2023

Ficha catalográfica elaborada pela Biblioteca do Centro Universitário SENAI CIMATEC

S719m Souza, Luryane Ferreira de

Modelagem computacional aplicada a análise de similaridade e filogenia das proteínas. / Luryane Ferreira de Souza – Salvador, 2023.

86 f. : il. color.

Orientador: Prof.. Dr. Marcelo Albano Moret Simões Gonçalves.
Coorientadora: Prof.ª. Dr.ª. Bruna Aparecida Souza Machado.

Tese (Doutorado em Modelagem Computacional e Tecnologia Industrial) – Programa de Pós-Graduação, Centro Universitário SENAI CIMATEC, Salvador, 2023.
Inclui referências.

1. Similaridade. 2. Proteína. 3. Autômato celular. 4. SARS-CoV-2. 5. Perfil de hidropatia. I. Centro Universitário SENAI CIMATEC. II. Gonçalves, Marcelo Albano Moret Simões. III. Machado, Bruna Aparecida Souza. IV. Título.

CDD 620.00113

NDI - 08

Nota sobre o estilo do PPGMCTI

Esta tese de doutorado foi elaborada considerando as normas de estilo (i.e. estéticas e estruturais) propostas aprovadas pelo colegiado do Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial e estão disponíveis em formato eletrônico (por solicitação via e-mail à secretaria do programa) e em formato impresso somente para consulta.

Ressalta-se que o formato proposto considera diversos itens das normas da Associação Brasileira de Normas Técnicas (ABNT), entretanto opta-se, em alguns aspectos, seguir um estilo próprio elaborado e amadurecido pelos professores do programa de pós-graduação supracitado.

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

Centro Universitário SENAI CIMATEC

Doutorado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leu e aprovou a Tese de doutorado, intitulada "**Modelagem Computacional Aplicada a Análise de Similaridade e Filogenia das Proteínas"**, apresentada no dia 08 de março de 2023, como parte dos requisitos necessários para a obtenção do Título de Doutora em Modelagem Computacional e Tecnologia Industrial.

Orientador:	**Prof. Dr. Marcelo Albano Moret Simões Gonçalves** SENAI CIMATEC
Coorientadora:	**Prof.ª Dr.ª Bruna Aparecida Souza Machado** SENAI CIMATEC
Membro Interno:	**Prof. Dr. Hernane Borges de Barros Pereira** SENAI CIMATEC
Membro Interno:	**Prof. Dr. Roberto Luiz Souza Monteiro** SENAI CIMATEC
Membro Externo:	**Prof. Dr. Thadeu Josino Pereira Penna** UFF
Membro Externo:	**Prof. Dr. Antônio-Carlos G. de Almeida** UFSJ

Dedico este trabalho à minha família pelo apoio incondicional. Em especial ao meu marido César pelo companheirismo e cuidado nesse período.

Agradecimentos

A realização dessa tese só foi possível graças ao apoio e incentivo de diferentes pessoas que passaram na minha vida ao longo desses 4 anos.

Primeiramente agradeço ao professor, orientador e amigo Marcelo A. Moret, pela sua orientação sempre atenta e presente, pelas conversas de incentivo e ajuda nos momentos de dúvidas.

A professora e coorientadora Bruna, pelas dicas e ajudas nessa reta final de construção da tese.

Em especial dedico essa conquista ao meu pai, Nélio, por todo carinho, amizade, amor, e que entendeu a minha ausência em algumas datas importantes. Você é o pilar da nossa família, obrigada por ser meu porto seguro. A minha mãe, Rita, que batalhou muito junto com meu pai para dar a melhor educação e as melhores oportunidades para mim e minhas irmãs. E que infelizmente não viveu para ver essa conquista.

Às minhas irmãs Lydiane e Yanara pela amizade, dicas e conversas que foram fundamentais para eu não desitir.

Ao meu marido César, que me incentivou a começar esse doutorado e que esteve comigo em todos os momentos, que sempre acreditou que eu conseguiria mesmo quando nem eu acreditava.

Aos professores da banca Hernane, Roberto, Thadeu e Antônio Carlos por terem aceito o convite e pelas correções e sugestões.

Aos professores e funcionários da pós pela ajuda e dedicação.

Ao Centro de Ciência Exatas da UFOB que possibilitaram meu afastamento nesse último ano de doutorado.

Salvador, Brasil
17 de Salvador
2023

Luryane Ferreira de Souza

Resumo

Com o crescente aumento de dados de sequenciamento de proteínas, um primeiro passo para identificação dessa macromolécula é feita através de comparação das proteínas para identificar suas similaridades com proteínas já conhecidas. Essa análise de similaridades pode identificar relações evolutivas entre as espécies comparadas. Sequências similares implicam espécies que compartilham um ancestral comum recente. Mas nem sempre as sequências são conservadas na evolução, então nesse caso uma comparação com a estrutura que é mais conservada pode ser uma alternativa para busca de relações evolutivas entre proteínas de mesma função. Este trabalho analisa as similaridades das sequências proteínas de diferentes espécies para classificar quanto às suas características evolutivas usando autômatos celulares unidimensionais para representar cada proteína como uma imagem. Calculamos as distâncias entre as imagens do autômatos usando a distância de Hamming. Essa distância mede as similaridades entre as imagens de autômato celular e a usamos para analisar relações evolutivas das espécies. Também propomos uma modelagem usando a diferença de perfil de hidropatia para analisar as diferenças ocorridas nas estruturas das variantes da SARS-CoV-2 ao longo da pandemia de COVID-19. Nosso método foi eficiente em aproximar espécies de mesmas classe animal e aproximou variantes do SARS-CoV-2 que compartilham da mutação N501Y. Além disso, usando a diferença no perfil de hidropatia podemos notar que a variante Omicron sofreu uma mudança significativa na região do RBD que pode está relacionada com os casos de reinfecção para essa variante.

Palavras-chave: Similaridade, Proteína, Autômato celular, SARS-CoV-2, Perfil de hidropatia.

Abstract

With the increasing number of protein sequencing data, the first step in identifying this macromolecule is to compare proteins to identify their similarities with known proteins. This similarity analysis can identify evolutionary relationships between the compared species. Similar sequences imply species that share a recent common ancestor. However, the sequences are not always conserved in evolution, so in this case, a comparison with the more conserved structure can be an alternative to searching for evolutionary relationships between proteins with the same function. This work analyzes the similarities of protein sequences from different species to classify them according to their evolutionary characteristics using one-dimensional cellular automata to represent each protein as an image. We calculated the distances between the automata images using the Hamming distance. This distance measures the similarities between the cellular automata images, and we use it to analyze species evolutionary relationships. We also propose modeling using the hydropathy profile difference to analyze the differences that occurred in the structures of the SARS-CoV-2 variants throughout the COVID-19 pandemic. Our method efficiently approached species of the same animal class and variants of SARS-CoV-2 that share the N501Y mutation. Furthermore, using the difference in the hydropathy profile, we can see that the Omicron variant underwent a significant change in the RBD region that may be related to the cases of reinfection for this variant

Keywords: Similarity, Protein, Cellular automaton, SARS-CoV-2, Hydropathy profile.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

AC	Autômato celular
SHD	Distância de Hamming estacionária.
RBD	Domínio de ligação ao receptor

Lista de Simbolos

$\psi(R,3)$	Hidrofobicidade local
$\psi(R,39)$	Hidrofobicidade conformacional

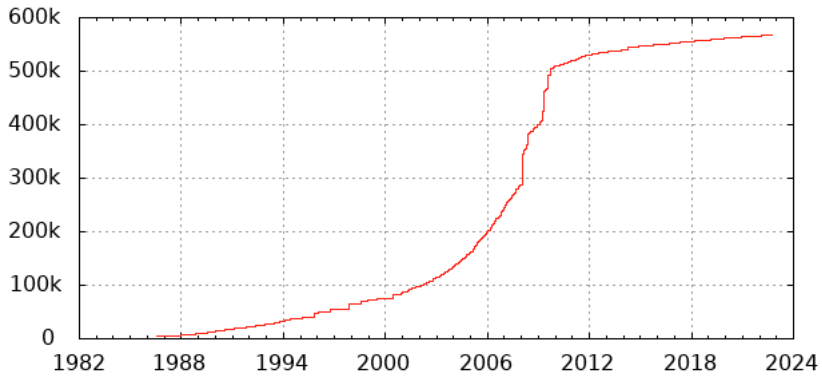
Capítulo Um

Introdução

As proteínas foram inicialmente sequenciadas por Frederick Sanger em 1955 (SANGER; THOMPSON; KITAI, 1955), correspondendo a macromoléculas compostas por variadas combinações de 20 tipos de diferentes aminoácidos, seus constituintes basilares. Tais combinações são responsáveis pela existência de uma enorme diversidade de proteínas desempenhando funções fundamentais em sistemas orgânicos, tais como: a catálise realizada pelas enzimas; o transporte de oxigênio função desempenhada pela hemoglobina; e as funções de regulação feita pelos hormonios de natureza proteica como a insulina. Além disso, considerando que cerca de 50% da massa seca de células é composta por proteínas, tanto em seres unicelulares como em multicelulares, é notável a relevância dessas macromoléculas para a existência da vida na Terra. Portanto, a evolução das espécies em nosso planeta está intrisecamente relacionada ao processo evolutivo das proteínas, sendo este um importante tema de estudo (KESSEL; BEN-TAL, 2018).

Algumas áreas se dedicam a estudar as proteínas, como a biologia molecular, bioinformática, biofísica, etc. Os últimos 20 anos têm sido de fundamental importância para essas áreas. O UniProtKB/Swiss-Prot (UNIPROT, 2022), um banco de dados específico para sequências de proteínas revisadas e suas funções cresceu mais de 4 vezes nesse período, como podemos observar na Figura 1.1. Esse crescimento se deve ao aumento do número de técnicas, e rapidez de sequenciamento. Em junho de 2022, o GenBank, um banco de dados de sequências genéticas, tinha 239.017.893 registros de sequências (GENBANK, 2022).

Figura 1.1: Número de entradas UniProtKB/Swiss-Prot por ano.



Fonte: (UNIPROT, 2022)

Com o crescimento desses bancos de dados surge a necessidade de tratar os dados da sequência de proteína. O principal objetivo do tratamento de dados da sequência de

proteína é analisar informações relevantes como estrutura, localização, função e relações evolutivas. Essas informações podem ser encontradas apenas fazendo a análise das sequências da proteína. E serão utilizadas para identificar a proteína e seu papel biológico, sendo de fundamental importância na produção de vacinas, de novos fármacos, tratamento de doenças, e para fins industriais.

Uma das técnicas mais utilizadas para caracterizar sequências de proteínas é a análise de similaridade. Sequências que apresentam um excesso de similaridades compartilham um ancestral comum recente (PEARSON, 2013). As sequências de proteínas guardam as informações evolutivas das espécies. Os métodos de similaridades de sequência de proteína são diversos, desde métodos que usam programação dinâmica para alinhar as sequências e encontrar regiões de similaridades (NEEDLEMAN; WUNSCH, 1970; LIPMAN; PEARSON, 1985; CAMPANELLA; BITINCKA; SMALLEY, 2003; HU; KURGAN, 2018), até métodos livres de alinhamento que analisam as similaridades das sequências das proteínas sem fazer essa correspondência de aminoácidos (MU; WU; ZHANG, 2013; MU et al., 2021; WU; XIAO; CHOU, 2010; YAO et al., 2008).

Para comparar sequências na busca por similaridades, algumas pesquisas desenvolveram uma representação gráfica da sequência de proteína como (XIAO et al., 2004; XIAO et al., 2005; XIAO; WANG; CHOU, 2008; YAO et al., 2008; LIAO et al., 2010; WU; XIAO; CHOU, 2010; MU; WU; ZHANG, 2013; CHAUDHURI et al., 2018; RAHMAN; BISWAS; BHUIYAN, 2019; MU et al., 2021).Todos esses trabalhos propõem uma representação 2D da sequência de proteína. Essas representações consideram características físico químicas dos aminoácidos (MU et al., 2021), estrutura molecular (CHAUDHURI et al., 2018) ou em alguns casos usam teoria da informação e teoria do reconhecimento molecular (XIAO et al., 2004). Alguns desses trabalhos utilizam autômatos celulares para representar a sequência da proteína como uma imagem (XIAO et al., 2004; XIAO et al., 2005; XIAO; WANG; CHOU, 2008). Essa técnica facilita comparar sequências sem necessariamente fazer a comparação para cada aminoácido, pois imagens muito distintas correspondem a proteínas diferentes ou de espécies diferentes.

Considerando que as sequências primárias de proteínas são menos conservadas que a estruturas da proteínas, uma análise dessa estrutura pode indicar semelhenças funcionais ou evolutivas entre as proteínas. Técnicas de perfil de hidropatia são uma alternativa a técnicas tradicionais de reconstrução da proteína para analisar semelhanças estruturais a partir da análise de sua sequência (LOLKEMA; SLOTBOOM, 1998). O perfil de hidropatia proposto por Kyte e Doolittle (KYTE; DOOLITTLE, 1982), foi o primeiro perfil de hidropatia a representar graficamente o caráter hidrofóbico da cadeia da proteína rastreando as regiões hidrofóbicas e hidrofílicas e estudando o melhor tamanho da janela de varredura que localiza essas regiões na sequência de proteína. Diversos trabalhos utilizam esse perfil para localizar regiões hidrofóbicas no interior da proteína enovelada, e regiões hidrofílicas

na superfície da proteína em contato com a água, são eles: (ESPOSTI; CRIMI; VENTUROLI, 1990; KRYSTEK; METZLER; NOVOTNY, 1995; DAMODHARAN; PATTABHI, 2004; PHILLIPS, 2009a; PHILLIPS, 2009b; PHILLIPS, 2021; PHILLIPS et al., 2022). Apesar do perfil de hidropatia localizar regiões hidrofóbicas e hidrofílicas da sequência da proteína, uma comparação desses perfis entre diferentes sequências mostra quais são mais semelhantes estruturalmente e relacionam evolutivamente as proteínas (PHILLIPS, 2021; PHILLIPS et al., 2022).

Motivados por esses trabalhos, propomos um método de comparação de sequências de proteínas usando as imagens de autômatos celulares. Utilizamos a métrica de distância de Hamming para comparar as imagens de autômatos celulares de diferentes proteínas, com estes valores de distância podemos comparar pares de espécies e construir matrizes de distâncias que serão úteis na construção de árvores filogenéticas. Quando aplicamos essa distância na análise filogenética de variantes de SARS-CoV-2, e em espécies animais podemos agrupar espécies de mesma classes, famílias ou que compartilham de mesmas mutações. E usando uma escala hidrofóbica baseada em fractalidade, considerando a hidrofobicidade local dos aminoácidos estudamos o perfil hidrofóbico das sequências de proteína Spike na região do RBD das variantes de SARS-CoV-2. Essa análise permite explicar uma diminuição na eficiência da vacina para a variante Omicron. Apesar das diferentes análises propostas nesse trabalho, elas concordam em comparar sequências de proteínas e na necessidade do estudo de relações evolutivas para entender como esse processo pode influenciar nossa forma de vida atual e futura.

1.1 Definição do problema

Como analisar similaridades/dissimilaridades entre sequências e estruturas de proteínas?

1.2 Objetivo

Avaliar modelos que analisam similaridade e relações evolutivas entre proteínas que nos permitam conhecer a origem das sequências e extrair informações biológicas de diferentes organismos.

1.2.1 Objetivos Específicos

- Identificar as proteínas para análise de similaridade para definir conjuntos de dados que apresentam um ancestral comum recente.

- Medir usando a distância de Hamming as similaridades/dissimilaridades das sequências de proteínas para construir a matriz de distância que compara pares de sequências.
- Construir árvore filogenética com a matriz distância para conhecer a origem dos diferentes seres estudados.
- Comparar nossos resultados com outros métodos usuais de distâncias entre sequências para avaliar a qualidade do método proposto.
- Comparar os perfis hidrofóbicos das sequências das proteínas Spike das variantes coronavírus SARS-CoV-2, em busca de similaridades/dissimilaridades de estruturas.

1.3 Importância da pesquisa

Toda a informação que a proteína precisa para se enovelar e assim atingir seu estado funcional esta na sequência da proteína, além disso as mutações ocorridas por essas sequências não são ao acaso e carregam uma mémoria do que ocorreu com a proteína no processo evolutivo. Uma das dificuldades da análise de similaridade de sequências é construir métodos que conseguem extrair as informações das proteínas a partir de suas sequências. A modelagem computacional das proteínas é uma ferramenta capaz de analisar e interpretar os dados dessa macromolécula,para uma melhor compreensão dos processos evolutivos de diferentes espécies.

1.4 Limites e limitações

Esta pesquisa propõem três métodos para análise das sequências de proteínas, todas as proteínas aqui estudadas passam pelo processo inicial de alinhamento das sequências comparadas. Esse passo incial é importante pois todos os métodos aqui propostos comparam sequências de mesmo tamanho. Sendo assim, ao selecionar os dados para avaliação do método, alinhamos as sequências de um mesmo tipo de proteína de espécies diferentes usando a ferramenta ClustalW no MEGA (THOMPSON; GIBSON; HIGGINS, 2003).

1.5 Questões e hipóteses

Os autômatos celulares são uma ferramenta gráfica para representar a sequência e podem ser utilizados na comparação dessas macromoléculas. Entretanto, as técnicas que utilizam essa ferramenta para comparar as sequências optam por retirar dados de texturas e

processamento da imagem do autômato celular para comparar as sequências. Diante do exposto a primeira questão que surge é como podemos medir as dissimilaridades a partir do autômato celular? Outra questão observada nesta pesquisa é que a variante Omicron do coronavírus SARS-CoV-2 sofreu muitas mutações na proteína Spike que é responsável pela ligação do vírus ao receptor do hospedeiro, essas mutações podem provocar mudanças estruturais nessa proteína. Sendo assim, buscamos responder a segunda questão problema: como a variante Omicron muda seu perfil hidrofóbico na região de ligação ao receptor na proteína Spike do coronavírus SARS-CoV-2? Diante dessas questões, nossas hipóteses são:

- H_1 A distância de Hamming pode medir essas dissimilaridades ao comparar os autômatos celulares.
- H_2 A variante Omicron tem seu perfil de hidropatia na região de ligação da proteína Spike bem diferente das outras variantes e do vírus inicialmente encontrado em Wuhan, sendo essa uma explicação por um aumento nos casos de reinfecção para essa variante e uma possível diminuição na eficácia da vacina.

1.6 Aspectos metodológicos

Este trabalho foi desenvolvido como uma pesquisa aplicada, conceituada em uma pesquisa exploratória bibliográfica. Isto porque, essa pesquisa bibliográfica inicial deve identificar o estado da arte em relação aos modelos usando autômatos celulares para proteínas e usando perfil de hidropatia. Nele, a autora propôs modelos de comparação de sequência mais específicos relacionados ao SARS-CoV-2 e depois propôs um modelo mais geral que engloba outras proteínas de espécies animais.

1.7 Organização da Tese de doutorado

Este trabalho foi divido em 6 capítulos de forma que cada capítulo corresponde a um tópico da tese:

- **Capítulo 1 - Introdução**: Apresentamos brevemente alguns dos métodos de similaridades, a importância desses métodos e do estudo de similaridades de proteínas, juntamente com os objetivos desse trabalho.
- **Capítulo 2 - Revisão da Literatura**: A revisão da literatura é apresentada. Nela a primeira seção se dedica aos conceitos básicos de proteínas que serão utilizados ao

longo do texto. Na segunda seção foi feita uma revisão dos métodos de similaridades de sequências de proteínas mostrando os tipos de métodos utilizados na comparação de proteína e abordando suas principais componentes. Em seguida, são apresentados os autômatos celulares e algumas aplicações desse método no estudo de proteínas. A última seção desse capítulo faz uma revisão pelos principais trabalhos de perfil de hidropatia e uso desse método no estudo do coronavírus SARS-CoV-2.

- **Capítulo 3 - Artigo 1**: Apresentamos o primeiro artigo publicado intitulado "Relating SARS-CoV-2 variants using cellular automata imaging". Nesse artigo usamos o gráfico de distância de Hamming entre autômatos celulares, para analisar as relações evolutivas entre as variantes de SARS-CoV-2 e comparar com um método tradicional de construção de árvore filogenética. Nesse artigo a ideia de distância de Hamming estacionária para medir a distância evolutiva foi introduzida.

- **Capítulo 4 - Artigo 2**: Apresentamos o segundo artigo que será submetido para publicação intitulado "New distance measure for comparing protein cellular automata image". Nesse trabalho usamos a mesma ideia da distância de Hamming estacionária para comparar autômatos celulares e construir matriz distância de similaridades que foram usadas nas construções de árvore filogenéticas. Comparamos nossa árvore resultante, com árvores feita com uma distância entre sequências de proteínas e verificamos que nossa medida é eficiente em agrupar espécies de mesma classe.

- **Capítulo 5 - Artigo 3**: Apresentamos o terceiro artigo que será submetido para publicação intitulado "Hydrophobic analysis of the SARS-CoV-2 Spike Protein". Nesse trabalho calculamos as diferenças entre o perfil de hidropatia das variantes de SARS-CoV-2 e do vírus inicialmente encontrado em Wuhan, na região do RBD da proteína Spike. Observamos que a diferença do perfil de hidropatia conformacional das variantes Alpha, Beta e Gamma são semelhantes sendo uma explicação para essas variantes não competirem nas regiões em que foram encontradas. Além disso observamos que a variante Omicron tem suas diferenças do perfil de hidropatia conformacional e local muito diferente das outras variantes, essa mudança pode está associada a uma diminuição na eficácia da vacina e o aumento de casos de reinfecção.

- **Capítulo 6 - Considerações Finais**: Apresentamos as principais conclusões e algumas ideias de trabalhos futuros.

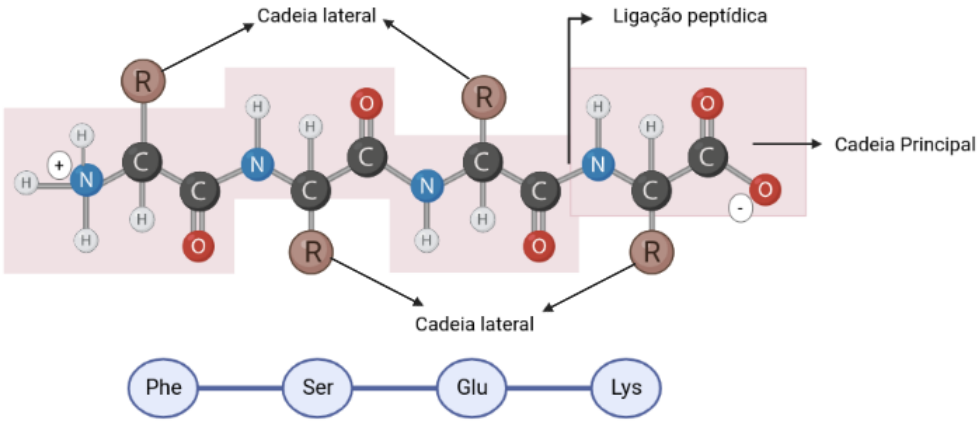
Capítulo Dois

Revisão da Literatura

2.1 *Proteínas e sistemas Complexos*

As proteínas são macromoléculas constituídas por 20 tipos de aminoácidos, que são estruturas mais simples. Esses aminoácidos, estão ligados por ligações covalentes formando uma sequência de aminoácidos denominada estrutura primária. A proteína é composta por uma cadeia principal e ligadas a ela estão as cadeias laterais correspondentes aos aminoácidos (Figura 2.1). Cada proteína tem sua sequência de aminoácidos, e esta sequência é responsável pelas diferentes funções das proteínas.

Figura 2.1: Partes de uma proteína. A cadeia principal está marcada de rosa. E a cadeia lateral compostas pelos aminoácidos foi representado por R (resíduos). Em sua forma mais simplificada representamos apenas a cadeia lateral.



Fonte: Criado com o BioRender.com

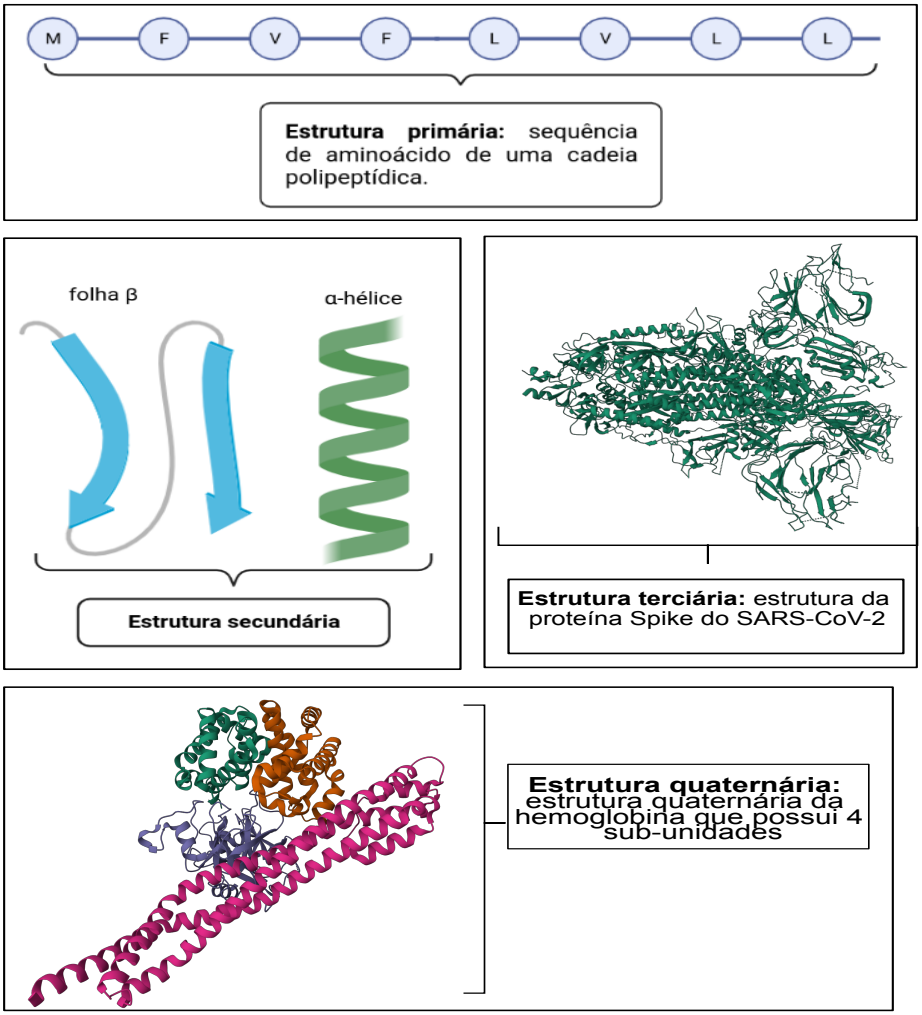
Para facilitar o uso de sequências de proteínas em software de bioinformáticas Margaret Dayhoff representou os aminoácidos por letras ao invés de siglas e essa representação é utilizada até hoje nos principais bancos de proteínas (GAUTHIER et al., 2019).

Quando estão ativas desenvolvendo suas funções, elas apresentam uma forma tridimensional denominada estrutura nativa. As funções da proteína são determinadas pelos tipos de aminoácidos presentes na cadeia e a ordem com que eles aparecem nas sequências. Anfinsen acreditava que toda informação necessária para a proteína enovelar-se está na sua sequência de aminoácidos. E que a estrutura nativa sempre busca um mínimo de energia livre (ANFINSEN, 1973). Para adquirir essa forma, a proteína passa pelo processo de enovelamento, que é o processo de formação das estruturas secundárias, terciárias e quaternárias (Figura 2.2).

As estruturas da proteína são:

- estrutura primária - composta por uma cadeia principal e uma cadeia lateral onde estão os aminoácidos ligados por ligações peptídicas
- estrutura secundária - arranjo espacial dos aminoácidos que estão próximos, as estruturas secundárias mais frequentes são as α-hélices e as folhas β.
- estruturas terciárias - interações entre aminoácidos distantes mas aproximados pelo enovelamento.
- estruturas quartenárias - é formada por um complexo de mais de uma cadeia polipeptídica (Capítulo XI, (NUSSENZVEIG, 2008)).

Figura 2.2: Estruturas da proteína.



Fonte: Criado com o BioRender.com

O enovelamento é estudado em seus diferentes aspectos: biológicos, químicos, físicos e computacionais. A principal motivação é entender quais fatores que influenciam no enovelamento e como ele ocorre. Entender esse processo é importante, pois as conformações

nativas das proteínas influenciam nas suas funções. Quando ocorrem falhas no enovelamento de proteínas ou mudanças na sua forma nativa, em alguns casos esse comportamento pode provocar o surgimento de doenças como Alzheimer, Parkinson (BRILKOVA et al., 2022; ELLIS; PINHEIRO, 2002), doença da vaca louca, etc (CHAUDHURI; PAUL, 2006). Estudar o processo de enovelamento em seus vários aspectos é necessário para entender como ocorre o aparecimento de tais doenças.

Outro fator que influencia no enovelamento de uma cadeia polipeptídica é a distribuição dos seus aminoácidos polares e apolares na cadeia. Em contato com a água, esses aminoácidos buscam uma conformação que favorece o contato com a água para os aminoácidos polares e diminui o contato com a água dos aminoácidos apolares. Essas interações entre os aminoácidos apolares e polares são conhecidas como as interações hidrofóbicas e tem papel fundamental no enovelamento de proteínas (MORET; ZEBENDE, 2007) e juntamente com as ligações de hidrogênio são as principais forças estabilizadoras (PACE et al., 1996) .

A estrutura da proteína pode ser determinada por experimentos em laboratórios mas como os número de sequências nos principais bancos de dados são elevados, outras técnicas foram utilizadas nas identificações estruturais das proteínas. Tendo em vista o custo elevado em desenvolver pesquisas em laboratórios foram desenvolvidos modelos teóricos, matemáticos e computacionais para descrever o enovelamento de proteínas e predição da estrutura.

A hidrofobicidade é um dos tipos de interações fracas que os aminoácidos podem apresentar quando a proteína está enovelada. As ligações fracas são ligações não covalentes entre os aminoácidos da cadeia. Interações eletrostáticas, pontes do tipo hidrogênio, Forças de Van der Waals são exemplos de ligações fracas que os aminoácidos podem realizar entre si. Esse tipo de ligação é mais importante na determinação de funções do que as ligações fortes (entre proteínas) (PHILLIPS, 2009a).

Quando a proteína está enovelada os aminoácidos podem participar de diversas ligações fracas com grupos de aminoácidos que foram aproximados pela configuração tridimensional obtida com o enovelamento, mas não é possível manter todas as interações ao mesmo tempo. Dessa forma, as interações fracas entre aminoácidos são alternadas, buscando sempre minimizar a energia livre. E essa é uma das características de sistemas complexos, a frustração. Quando estudamos proteínas, não podemos explicá-las como a soma das propriedades individuais de seus aminoácidos constituintes, essas moléculas se comportam como um sistema complexo. A criticalidade auto-organizada também é uma característica desse sistemas. Quando a quantidade de aminoácidos em uma cadeia simples atinge um número mínimo, essa cadeia pode enovelar formando estruturas secundárias estáveis (MORET, 2011), que se assemelha ao modelo de pilha de areia (Capítulo I, (NUSSENZVEIG, 2008)).

Outra característica importante é o efeito de memória presente no enovelamento. De acordo com o paradoxo de Levinthal durante o processo de enovelamento essa molécula não experimenta todas as configurações existentes para encontrar sua forma tridimensional, existe nesse processo um efeito de memória que favorece essa proteína encontrar sua configuração final. Supondo uma sequência com 100 aminoácidos (número médio de aminoácidos em uma proteína) e que cada aminoácido possui duas configurações espaciais, então o número de configurações possíveis para a proteína seria 2^{100} ($\approx 10^{30}$). Considerando que na natureza os movimentos mais rápidos ocorrem em 10^{-12} segundos, o tempo gasto para a proteína verificar todas as configurações possíveis seria 10^{18} (tempo maior que a idade do Universo), mas esse processo ocorre em menos de um segundo (CONTESSOTO et al., 2018). Consequentemente, a evolução atuou na seleção das funções das proteínas e influencia no processo de enovelamento.

2.2 Similaridades de Sequências de Proteínas

Após o sequenciamento da proteína o próximo passo para identificar características como função, localização, e relações evolutivas, é a busca por similaridades. Sequências que são similares podem ser inferidas como homólogas, ou seja, apresentam um ancestral comum recente. Existem dois tipos de similaridades, as similaridades das sequências das proteínas e a similaridade funcional. A similaridade funcional é mais difícil de ser quantificada e demanda conhecer as estruturas das proteínas (PEARSON, 2013). Para uma análise inicial, a similaridade da sequência é capaz de mostrar relações evolutivas entre as espécies analisadas mesmo sem conhecer sua estrutura. Aqui vamos nos restringir ao estudo de similaridades de sequências de proteínas. Fitch em 1970, separou os homólogos em dois tipos: os genes que divergem por duplicação e acabam mudando suas funções são denominados parálogos. Já os que divergem por especiação e cada gene mantém a função do gene de origem é denominado ortólogos (FITCH, 1970).

Muitos métodos de comparação de sequência para análise de similaridades foram desenvolvidos nos últimos anos. Eles se dividem em dois tipos básicos: os que usam ou produz alinhamento e os que não usam nem produzem alinhamento. O alinhamento ocorre quando fazemos correspondência entre os aminoácidos de duas ou mais sequências (ZIELEZINSKI et al., 2017).

Os métodos livres de alinhamentos são uma solução mais simples para os métodos com alinhamento. Como as sequências de uma mesma proteína podem mudar de tamanho de acordo com a espécie, a comparação em métodos livre de alinhamento não é feita diretamente com a sequência da proteína. Por exemplo, alguns métodos fazem uma representação gráfica da sequência e a partir dessa representação retiram vetores descritores da proteína e assim comparam usando alguma medida de distância (YAO et al., 2008;

MU; WU; ZHANG, 2013; YAO et al., 2014). Para comparar sequências usando métodos livres de alinhamentos alguns desses trabalhos propõem novas medidas de distâncias (OTU; SAYOOD, 2003) ou até mesmo representação gráfica da sequência de proteína para melhor compará-la (YAO et al., 2008; LIAO et al., 2010; WU; XIAO; CHOU, 2010; MU; WU; ZHANG, 2013; HUANG; HU, 2013; YAO et al., 2014; RAHMAN; BISWAS; BHUIYAN, 2019; MU et al., 2021).

As análises filogenéticas que usam o alinhamento são bem comuns na literatura e atualmente existem diferentes softwares que fazem o alinhamento da sequência da proteína em poucos minutos. Os principais softwares que fazem alinhamento são ClustalW, MAFFT, BLAST (THOMPSON; GIBSON; HIGGINS, 2003; KATOH; STANDLEY, 2013; ALTSCHUL et al., 1990). Mas para que serve o alinhamento? Quando as sequências estudadas são pesquisadas nos bancos de dados seus tamanhos são variados, dependendo das espécies consideradas. Assim, para que as regiões correspondentes da sequência sejam comparadas é necessário o alinhamento das sequências. Os softwares de alinhamento buscam alinhar os aminoácidos em regiões mais semelhantes possível, para isso ele insere lacunas para aproximar mais as sequências analisadas (CHAO; TANG; XU, 2022).

Os métodos de análise filogenéticos que usam o alinhamento são divididos em métodos de distância, métodos de máxima parcimônia, métodos de máxima verossimilhança e métodos Bayesianos. As etapas do método de distância são: primeiro o alinhamento múltiplo das sequências, segundo distâncias evolutivas e terceiro a construção de árvores (LIU; WANG, 2006). O método da distância calcula para cada par de sequências uma medida que corresponde ao comprimento do ramo que os separa, construindo assim uma matriz distância. A matriz distância é utilizada na construção da árvore filogenética, juntamente com modelos evolutivos para construção da árvore (FELSENSTEIN, 1996). Os métodos de máxima parcimônia e máxima verossimilhança possuem apenas duas etapas: o alinhamento múltiplo e as árvores. A máxima parcimônia é um método que não utiliza a matriz distância, constrói a árvore diretamente da análise dos caracteres das sequências, não utiliza modelos evolutivos para construção da árvore. A árvore mais simples é escolhidas entre todas as árvores possíveis. A máxima verossimilhança busca a árvore mais verossímil de acordo com os dados das sequências alinhadas e com base no modelo evolutivo (CALDART et al., 2016). Já os métodos Bayesianos são baseados em métodos de máxima verossimilhança mas incorporam a probabilidade (LIU; WANG, 2006). Neste trabalho vamos nos restringir aos métodos de distâncias.

Os métodos de distância usam uma matriz de distância para construir a árvore filogenética e diferentemente dos métodos de máxima parcimônia não percorre todas as topologias em busca de uma árvore de evolução mínima. Quando se compara um grande número de sequências o método de máxima parcimônia não é eficiente pois não consegue examinar todas as topologias, por esse motivo usam-se métodos de distância que constrói uma

árvore no final do processo. Vários métodos de distâncias foram desenvolvidos nos últimos anos e são utilizados principalmente quando a comparação é feita com muitas sequências (FARRIS, 1972; TATENO; NEI; TAJIMA, 1982; FAITH, 1985; SAITOU; NEI, 1987; KUHNER; FELSENSTEIN, 1994; ZHANG; SUN, 2008).

Na construção da árvore podemos usar métodos de agrupamento que são métodos que separam os dados analisados em subgrupos. Os agrupamentos hierárquicos utilizam uma ordem para agrupar os indivíduos. E representam as árvores filogenéticas por dendrogramas. Essa técnica pode ser dividida em dois métodos, os aglomerativos que separam inicialmente todos os N elementos em N clusters, e a medida que analisa as similaridades unem os clusters que são similares. E os métodos divisivos, que inicialmente mantém os N indivíduos em um único cluster e a medida que analisa as distâncias separam os indivíduos em agrupamentos menores. Vamos utilizar os métodos aglomerativos para construção das árvores em nossos artigos. Os métodos aglomerativos se diferenciam pela forma como calculam as distâncias entre dois grupos: single, complete, average (UPGMA), centroid, weighted average, median e Wards (EVERITT et al., 2011).

2.3 Autômatos Celulares

Um autômato celular (AC) é sistema discreto no tempo e espaço. Suas unidades mais básicas são as células, que possuem um estado atribuído. Os estados são finitos e vão depender do objeto de estudo. Para evolução do autômato é necessário definir uma regra de evolução que determina o estado futuro de cada célula dependendo de seu estado atual e dos seus vizinhos. A condição inicial mostra os estados iniciais da célula, de onde o autômato celular começa sua evolução. A partir dessa condição os outros passos do autômato serão determinados pela regra de evolução (SARKAR, 2000).

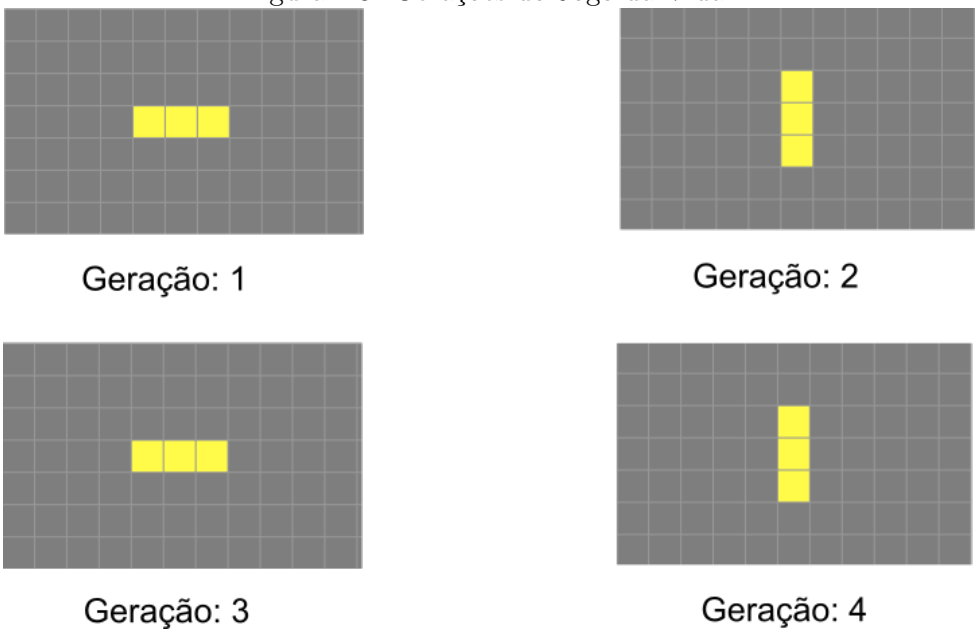
O primeiro estudo formal de autômatos celulares foi realizado por John von Neumann, ele conseguiu construir uma máquina que se auto reproduz e que portanto pode ser usado na aplicação de sistemas biológicos complexos (NEUMANN, 1963). Em 1970 John Conway criou o jogo da vida um jogo baseado nas ideias de Neumann de autômato celular. Conway construiu um jogo com regras baseadas em uma sociedade de organismos vivos (GARDNER, 1970). O autômato do jogo da vida era um autômato celular de duas dimensões onde as células tinham dois estados vivo ou morto. A vizinhança de Moore é composta por oito células as quatro adjacentes ortogonalmente e quatro adjacentes diagonalmente. E as regras são:

- Cada célula viva com dois ou três vizinhos vivos sobrevive para a próxima geração (evolução).

- Cada célula viva com quatro ou mais vizinhos vivos morre por superpopulação.
- Toda célula viva com um vizinho ou nenhum vizinho vivo morre por isolamento.
- Toda célula morta com três vizinhos vivos, torna-se viva na próxima geração é uma célula de nascimento (GARDNER, 1970).

Na Figura 2.3 mostramos a evolução do Jogo da vida para três células vivas esse padrão é do tipo oscilador que se repete a cada 2 passos de evolução. Conway construiu um dos autômatos mais conhecidos do mundo.

Figura 2.3: Gerações do Jogo da Vida



Fonte: Criado em <https://playgameoflife.com/>

Um dos principais nomes na área de autômato celular foi Wolfram, ele se dedicou especificamente aos autômatos celulares unidimensionais, aos estudos de suas regras e comportamentos (WOLFRAM, 1984; WOLFRAM, 2002). Os autômatos celulares são aplicados em diferentes áreas como a física, química, biologia, computação e matemática, sua principal vantagem é a simplicidade no funcionamento mas podem levar a comportamentos complexos (SARKAR, 2000).

A modelagem tradicional por meios de equações diferenciais modelam alguns sistemas biológicos mas para os casos de biomoléculas esse tipo de modelagem pode ser desafiador devido ao número de variáveis e suas interações, assim como alternativa os autômatos celulares podem ser utilizado como uma técnica mais eficiente para modelar esses sistemas (BONCHEV et al., 2010).

Os AC são usados em diversas pesquisas de proteínas, como na predição de suas estruturas

terciárias (SANTOS; VILLOT; DIéGUEZ, 2013; SANTOS; VILLOT; DIéGUEZ, 2013) na modelagem de enovelamento (VARELA; SANTOS, 2016; VARELA; SANTOS, 2022) e na classificação de suas propriedades através das imagens de autômatos celulares (XIAO et al., 2004; XIAO et al., 2005; KAVIANPOUR; VASIGHI, 2017; CHAUDHURI et al., 2018).

O autômato celular (AC) é composto por quatro elementos principais (L, S, N, f):

- L é a grade que pode ter uma, duas ou três dimensões, conforme Figura 2.4 . Podemos representá-las por diferentes figuras regulares: quadrados, triângulos e hexágonos;
- S são os estados das células;
- N é a vizinhança que consideramos para a célula analisada;
- f regra de transição dos autômatos.

Figura 2.4: Grades a) unidimensional, b) duas dimensões e c) três dimensões, respectivamente.

Fonte: a) Autoria Própria, b) (SAHELY, 2022), c) (WOLFRAM, 2022)

Estamos interessados em estudar sequências de proteínas desdobradas, portanto restringimos o estudo de autômatos celulares a grade unidimensional quadradas. As células x_i^t são as unidades básicas do autômato de tamanho N, onde i representa a posição no autômato $i = 1, 2, 3, ..., N$, e a variável t representa os passos da evolução do autômato. Consideramos dois estados para as células, $S = \{0, 1\}$ onde representaremos 0 pela cor branca e 1 pela cor preta. A vizinhança da célula x_i^t leva em consideração os estados das células $V(x_i^t) = \{x_{i-1}^t, x_i^t, x_{i+1}^t\}$. Trata-se de uma vizinhança de raio 1. A regra de transição define o estado da célula em cada instante, tendo como base os estados das células da vizinhança no tempo anterior, como é apresentado na Eq. 2.1:

$$x_i^{t+1} = f(x_{i-1}^t, x_i^t, x_{i+1}^t) \tag{2.1}$$

o estado da célula x_i^{t+1} , depende do estado anterior dessa célula e de suas vizinhas imediatas.

Além das componentes do autômato celular existem também outros elementos que precisam ser considerados para a evolução do autômato celular: condição de fronteira e a condição inicial. A condição de fronteira define a vizinhança das células nas extremidades x_1^t e x_N^t do autômato. Como a célula x_1^t não tem vizinhos à esquerda e a célula x_N^t não tem vizinhos à direita precisamos definir quem serão esses vizinhos. Definimos por exemplo, a condição fronteira periódica onde a célula x_1^t tem como vizinho à esquerda a última célula do autômato a célula x_N^t, como visto na Eq. 2.2. E a célula x_N^t tem como vizinho à direita a célula x_1^t, como visto na Eq. 2.3.

$$V(x_1^t) = \{x_N^t, x_1^t, x_2^t\} \tag{2.2}$$

$$V(x_N^t) = \{x_{N-1}^t, x_N^t, x_1^t\} \tag{2.3}$$

Antes de evoluir um autômato celular é necessário conhecer os estados de cada célula antes da primeira evolução, esse estado inicial é denominado condição inicial. Aqui essa condição inicial será definida pela sequência da proteína. A representação da sequência de proteína são sequências de caracteres que representam os 20 tipos de aminoácidos. Para transformar a sequência de caracteres na condição inicial do autômato um código que transforma cada aminoácido da sequência em um código binário é necessário, assim o autômato terá dois estados que serão representados por 0 e 1 (branco e preto). Trabalhos como (XIAO et al., 2004; XIAO; CHOU, 2007; KAVIANPOUR; VASIGHI, 2017; CHAUDHURI et al., 2018) propõem códigos para transformar a sequência de proteína em sequência binária. A grande preocupação nessa codificação é perder a menor quantidade de informações dos aminoácidos com a codificação (XIAO et al., 2004). Cada um dos trabalhos citados acima levam em consideração características dos aminoácidos. O código proposto por (XIAO et al., 2004) propõem um código de 5 dígitos baseado em regras de similaridades, complementaridade, teoria molecular, e teoria da informação. Já em um artigo posterior (XIAO; CHOU, 2007) propõem um código também de 5-dígitos baseado no índice de hidrofobicidade dos aminoácidos. Baseado nesse trabalho, (KAVIANPOUR; VASIGHI, 2017) propõem um código mais simples com 3-dígitos para estudar a classe estrutural das proteínas. Em (CHAUDHURI et al., 2018) propõem um código com 8-dígitos baseados nas análise da estrutura molecular dos aminoácidos. Não existe um código ideal, cada um dos trabalhos citados têm uma aplicação diferente para uso dos códigos propostos. Nosso método leva em consideração os códigos propostos por (XIAO; CHOU, 2007) e (CHAUDHURI et al., 2018) , dependendo do objeto de estudo.

Tabela 2.1: Codificação de aminoácidos (CHAUDHURI et al., 2018), deleções e dados de sequência de proteínas ausentes após o alinhamento.

Aminoácidos	Código	Aminoácidos	Código
Glicina (G)	00000000	Cisteina (C)	01000100
Alanina (A)	00000100	Treonina (T)	00110100
Prolina (P)	00100110	Asparagina (N)	00101110
Valina (V)	00010110	Glutamina (Q)	00101111
Metionina (M)	00110110	Tirosina (Y)	10100100
Triptofano (W)	10110110	Histidina (H)	01111110
Fenilalanina (F)	10000100	Lisina (K)	00110111
Isoleucina (I)	00011110	Arginina (R)	01111111
Leucina (L)	00010111	Ácido aspártico (D)	01110100
Serina (S)	00100100	Ácido glutâmico (E)	01110110
Deleção (-)	11111111	Informação Faltante (?)	11111110

Usando o código da Tabela 2.1 podemos definir a condição inicial do autômato celular. Por exemplo, a sequência da proteína beta-globina de humanos que tem como sequência de aminoácidos alinhado o código:

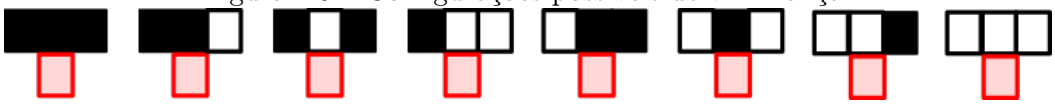
$$MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD$$
$$LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDK$$
$$LHVDPENFRLLGNVLVCVLAHHFGK-EFTPPVQAAYQKVVAGVANALA$$
$$HKYH$$

Terá como condição inicial o código binário:

00110110000101100111111000010111001101000010011001110110011101
00011011100100100000001000001011000110100000001000001011110110
10000000000011011100010110001011100001011001110100011101100001
11000000000000000000111011000000100000101110000000001111111000
01110001011100010110000101101010010000100110101101100011010000
01111011111111000010010000100011101100010010010000100000000001
11010000010111001001000011010000100110011101000000010000010110
01101100000000000101110001001100011011100010110001101110000100
01111110000000000011011100110111000101100001011100000000000010
01000010000100100011101000000000000010111000001000111111000010
11011101000010111000010111001101110000000000110100100001000000
10000110100000101110010010001110110000101110111111001110100011
01000011011100010111011111100001011001110100001001100111011000
01110100001000111111100010111000101110000000000101110000101100
01011100010110011101000001011000010111000001000111111001111110
00001000000000000110111111111110111011010000100001101000010011
00100110000101100010111100000100000001001010010000101111001101
10001011000010110000001000000000000001011000000100001011100000
00000101110000010001111110001101111010010001111110

Quando usamos autômatos celulares unidimensionais com dois estados e três células na vizinhança as regras de evolução local utilizadas são denominadas regras de autômatos celulares elementares de Wolfram (WOLFRAM, 2002). Essas regras são capazes de reproduzir comportamentos bem variados podendo modelar diferentes sistemas. Como o autômato tem dois estados e a vizinhança leva em consideração o estado de três células, teremos ao todo $2^3 = 8$ configurações de vizinhança possíveis, ver Figura 2.5.

Figura 2.5: Configurações possíveis de vizinhança

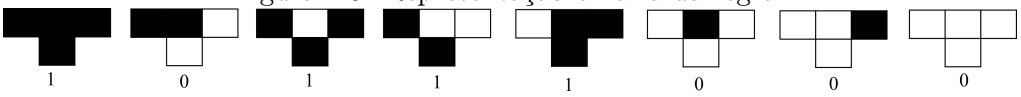


Fonte: Autoria Própria

Para cada configuração da vizinhança da Figura 2.5 a célula do meio pode assumir 2 estados possíveis na evolução do autômato então temos ao todo $2^8 = 256$ regras possíveis para o autômato celular unidimensional considerando a vizinhança de raio 1. Cada uma

das 256 regras de Wolfram pode ser representada por um número de 0 a 255. Por exemplo a regra da Figura 2.6:

Figura 2.6: Representação binária da regra



Fonte: Autoria Própria

Para encontrar o número da regra correspondente, basta escrever o número binário como um número na base decimal:

$$1 \times 2^7 + 0 \times 2^6 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0 = 184 \qquad (2.4)$$

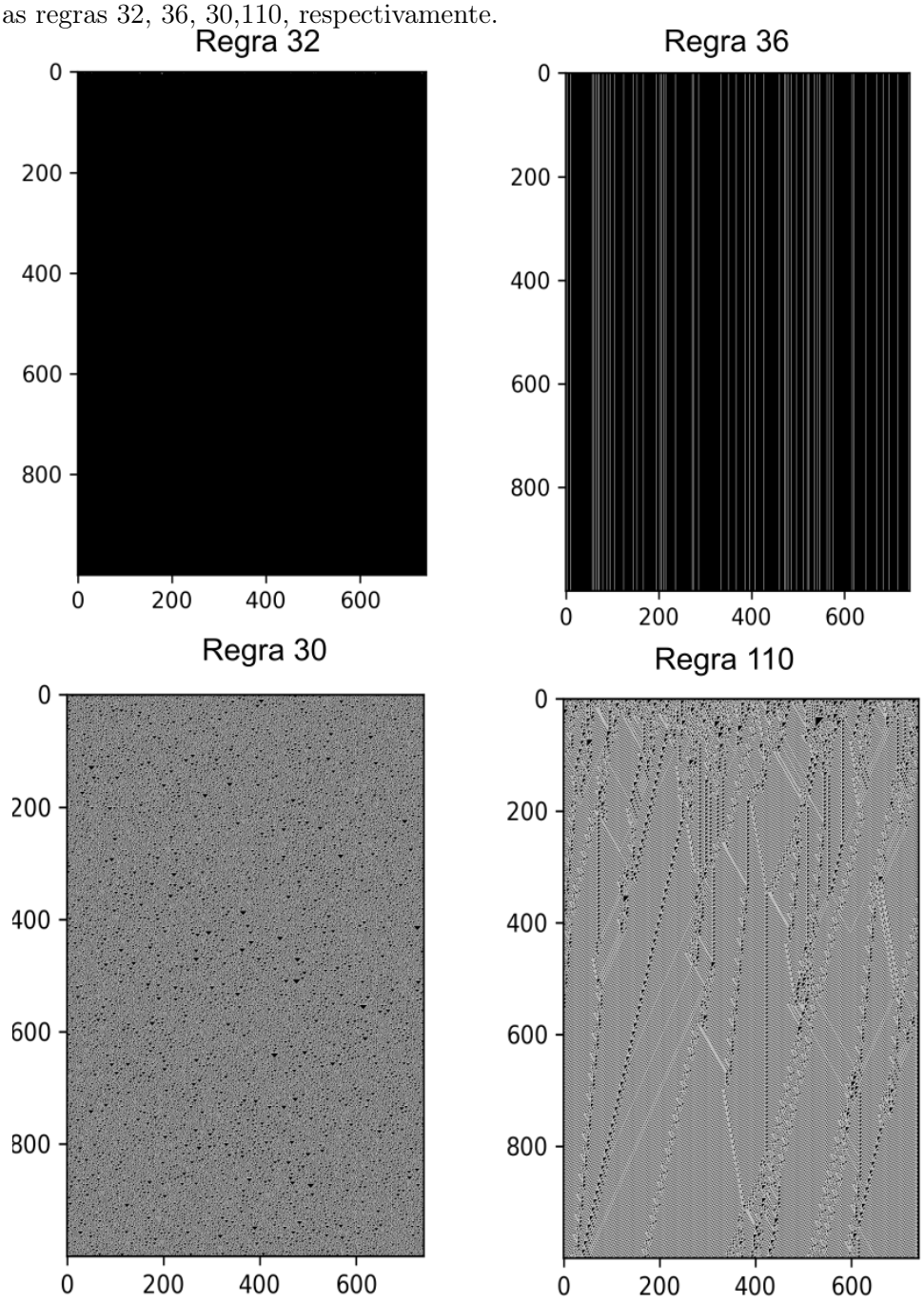
A Eq. 2.4 mostra que a regra da Figura 2.6 é a regra 184 de Wolfram. Caso tenhamos o número da regra também é possível saber o que ela faz com cada uma das configurações de vizinhanças fazendo a conversão de número decimal para binário e colocando na ordem que as vizinhanças seguem na Figura 2.5, colocando o binário que acompanha as maiores potências da esquerda para direita.

As regras do autômato celular podem gerar comportamentos interessantes. Na Figura 2.7, a condição inicial do autômato celular é a sequência da proteína beta-globina do pombo doméstico (P11342.1, NCBI) usando o código binário proposto por (XIAO; CHOU, 2007) para os aminoácidos, quatro regras diferentes foram consideradas e para a mesma condição inicial temos quatro imagens diferentes.

A escolha de tal regra pode ser relacionada ao tipo de molécula estudada (proteína, DNA, RNA), localização celular ou comportamento do sistema analisado (XIAO et al., 2005; WANG et al., 2005; RAHMAN; BISWAS; BHUIYAN, 2019), cada problema passará por uma etapa de escolha dessa regra para que se obtenha o melhor resultado. Trabalhos anteriores (XIAO et al., 2005; WANG et al., 2005; RAHMAN; BISWAS; BHUIYAN, 2019) já definiram algumas regras para o estudo de proteínas e DNA. Escolhida a regra o autômato é evoluído e obtêm-se a imagem do autômato celular (IAC).

As imagens dos autômatos celulares são importantes pois muitas características que não podem ser observadas diretamente nas sequências das moléculas podem ser observadas na IAC. Para melhor analisar essas imagens são usados métodos estatísticos de textura como o usado por (KAVIANPOUR; VASIGHI, 2017) que calcula contraste, correlação, homogeneidade a partir de recursos de Haralick e com esses dados usam redes neurais artificiais para prever, ou classificar estruturas. A comparação visual das imagens também é possível (XIAO et al., 2004; XIAO et al., 2005) mas para uma comparação de imagens

Figura 2.7: Imagem dos autômatos celulares da proteína beta-globina em pombos domésticos usando as regras 32, 36, 30,110, respectivamente.



Fonte: Autoria própria

mais rigorosa, métodos de comparação entre a imagem são necessários. Buscamos com essa técnica analisar similaridades entre sequência de proteína comparando suas IAC usando a métrica de distância de Hamming.

2.4 Perfil de Hidropatia

As forças hidrofóbicas são os fatores mais importante para manter a proteína enovelada (KAUZMANN, 1959). Além disso, toda informação para a proteína se enovelar está na sequência, sua estrutura e consequentemente sua função são determinadas pela sequência de aminoácido (ANFINSEN, 1973). Como os 20 aminoácidos possuem características hidrofóbicas e hidrofílicas, isso deve ser levado em consideração quando a proteína busca sua estrutura estável. Mas além da característica hidrofóbica/hidrofílica individual do aminoácido os seus vizinhos contribuem para a formação da estrutura nativa da proteína. Assim as regiões hidrofóbicas (interior) e regiões hidrofílicas (na superfície da proteína) são determinadas pela hidrofobicidade dos aminoácidos e o efeito dos seus vizinhos (PHILLIPS, 2021). O perfil de hidropatia mostra graficamente as regiões das sequências que são mais hidrofóbicas e hidrofílica levando em consideração as características hidrofóbicas dos aminoácidos com o efeito dos seus vizinhos.

O perfil de hidropatia foi utilizado em diversas pesquisas, suas principais aplicações são prever regiões com estrutura secundária, prever regiões transmembranares de proteínas, procurar novas proteínas de membrana (KRYSTEK; METZLER; NOVOTNY, 1995; ESPOSTI; CRIMI; VENTUROLI, 1990; CLEMENTS; MARTIN, 2002). Como as sequências da proteína estão mais sujeitas a mutações que suas estruturas, mudanças no perfil de hidropatia podem sinalizar mudanças nas estruturas, indicando uma divergência evolutiva nas proteínas que não pode ser verificada facilmente analisando apenas a sequência.

O perfil de hidropatia é a média dos índices de hidrofobicidades de um aminoácido e de seus vizinhos a medida que se percorre a sequência da proteína, dando uma visualização gráfica do caráter hidrofóbico/hidrofílico da cadeia de uma extremidade à outra, em relação a uma linha média universal (KYTE; DOOLITTLE, 1982). O perfil de hidropatia destaca-se por ser um método de fácil execução e que consegue obter informações importantes das estruturas das proteínas através de sua sequência. Para determinar o perfil de hidropatia é necessário escolher os índices de hidrofobicidade dos aminoácidos e a janela de varredura para o cálculo da média. Os índices de hidrofobicidade dos aminoácidos podem assumir diferentes valores dependendo da escala considerada, mas suas características hidrofóbicas e hidrofílicas não mudam, assim independente da escala considerada o gráfico do perfil de hidropatia mantém um comportamento similar para as diferentes escalas (KRYSTEK; METZLER; NOVOTNY, 1995).

Um dos principais trabalhos de perfil de hidropatia foi proposto por Kyte e Doolittle (KYTE; DOOLITTLE, 1982), nesse trabalho os índices de hidrofobicidade de cada aminoácido são derivadas das energias livres de transferência de vapor de água e da distribuição interior-exterior das cadeias laterais de resíduos. Para proteínas globulares, ocorreu uma correspondência entre as regiões internas das proteínas com regiões do perfil de hidro-

patia que são hidrofóbicas , assim como as regiões externas da proteína e com regiões do perfil de hidropatia que são hidrofílicas. Esse método também foi capaz de identificar em proteínas de membranas as porções dessa proteína que estão dentro da bicamada lipídica. Kyte e Doolittle também analisaram o tamanho da janela de varredura e concluíram que para encontrar regiões com estruturas secundárias a janela de 7 a 9 aminoácidos é a mais adequada.

Esposti, Crimi e Venturoli (ESPOSTI; CRIMI; VENTUROLI, 1990), propôs uma análise estatística entre diferentes escalas para verificar a predição de segmentos transmembranares de proteínas e selecionou critérios de escolha do melhor índice de hidrofobicidade utilizado nos perfis de hidropatia para proteínas de membrana. Outros trabalhos também usam perfil de hidropatia para detectar relações evolutivas distantes entre proteínas de membrana (LOLKEMA; SLOTBOOM, 1998) e encontrar novas proteínas de membrana (CLEMENTS; MARTIN, 2002).

O perfil de hidropatia também foi usado em (PHILLIPS, 2021) para analisar as mudanças no perfil hidropático da proteína Spike do coronavírus SARS-CoV-2. Phillips mostra que o vírus SARS-CoV-2 consegue nivelar os mínimos no perfil de hidropatia na região do domínio N terminal S1 responsável pela ligação celular. Os vírus podem nivelar seus mínimos hidrofílicos para prolongar a vida útil da proteína. Essa capacidade de nivelar seus mínimos não está presente no SARS-CoV-1. Assim o SARS-CoV-2 promove uma ligação viral mais forte e aumenta a contagiosidade. O nivelamento dos mínimos no perfil hidropático ocorre apenas quando o índice de hidrofobicidade proposto por Moret e Zebende (MORET; ZEBENDE, 2007) foi utilizado. Esse índice caracteriza o processo de enovelamento como uma transição de fase de segunda ordem. Outros índices de hidrofobicidades não exibem esse nivelamento presente na escala Moret e Zebende. Essa escala leva em consideração a perda de área de superfície acessível ao solvente de cada aminoácido. A perda de área se comporta como uma lei de potência e os expoentes podem ser caracterizados como os índices de hidrofobicidade dos aminoácidos.

A área de superfície acessível ao solvente é calculada considerando cada átomo mais externo à molécula como uma esfera com raio de Van der Waals. A área da molécula acessível ao solvente (ASA) é definida passando uma esfera de raio menor que representa o solvente pela superfície da molécula. Moret e Zebende (MORET; ZEBENDE, 2007) consideram um conjunto de 5526 cadeias de proteínas para medir a área de superfície acessível ao solvente dos aminoácidos e tomando vários fragmentos de proteínas de tamanhos variados entre 3 a 45 aminoácidos. Moret e Zebende mostraram que quando o número de aminoácidos é maior que 8 a área de superfície acessível ao solvente dos aminoácidos se comporta como uma lei de potência, Eq. 2.5:

$$ASA \propto N^{\gamma} \tag{2.5}$$

Na Eq. 2.5, o N é o número de aminoácidos vizinhos no segmento de proteínas extraídos das 5526 proteínas do banco de dados. E γ determina a hidrofobicidade dos aminoácidos.

Em um trabalho mais recente (PHILLIPS et al., 2022), estabelece qual o valor ideal da janela de varredura W, para se calcular o perfil de hidropatia para o SARS-CoV-2 e obter características estruturais da proteína Spike do SARS-CoV-2. O valor ideal de W é próximo a 40. E nesse trabalho também foi utilizada a escala (MORET; ZEBENDE, 2007). Vamos usar a escala Moret e Zebende juntamente com o valor de janela de varredura para analisar as mudanças ocorridas nas variantes de SARS-CoV-2 na região do RBD da proteína Spike.

Capítulo Três

Artigo 1

3.1 Relating SARS-CoV-2 variants using cellular automata imaging.

Artigo publicado na Revista Scientific Reports v. 12, n. 10297 (2022)

Autores: Luryane Ferreira de Souza, Tarcísio Marciano da Rocha Filho, Marcelo Albano Moret Simões Gonçalves.

DOI: <https://doi.org/10.1038/s41598-022-14404-6>

scientific reports

Check for updates

OPEN

Relating SARS-CoV-2 variants using cellular automata imaging

Luryane F. Souza[1,2]✉, Tarcísio M. Rocha Filho[3] & Marcelo A. Moret[2,4]

We classify the main variants of the SARS-CoV-2 virus representing a given biological sequence coded as a symbolic digital sequence and by its evolution by a cellular automata with a properly chosen rule. The spike protein, common to all variants of the SARS-CoV-2 virus, is then by the picture of the cellular automaton evolution yielding a visible representation of important features of the protein. We use information theory Hamming distance between different stages of the evolution of the cellular automaton for seven variants relative to the original Wuhan/China virus. We show that our approach allows to classify and group variants with common ancestors and same mutations. Although being a simpler method, it can be used as an alternative for building phylogenetic trees.

The disruption caused during the last two years by the COVID-19 pandemic is hard to be underestimated, from more than five million deaths and 270 million cases world-wide, according to official sources [1], to economic disruption in most countries[2]. In December 31, 2019 the first case was reported in the city of Wuhan, China, and in January 9, 2020, the World Health Organization (WHO) informed that Chinese scientists reported that the disease was caused by a new coronavirus. In February 11, 2020, in order to not associate the disease with any locality or groups of people the new coronavirus was named SARS-CoV-2 and the disease it caused COVID-19. In March 11 of that same year the WHO declared the outbreak a pandemic[3].

The SARS-CoV-2 virus is part of the same virus family as the SARS-CoV and MERS-CoV viruses, the Sarbecovirus subgroup of the subdivision of the Betacoronavirus genera, which were responsible for epidemics in China (2003) and Saudi Arabia (2012)[4]. The last decade or so witnessed important developments in genome sequencing techniques, with a major increase in information gathering (data) on DNA, RNA, and protein sequences, as exemplified by the amount of data in databases such as GenBank[5] and UniProt[6] (for a more thorough account on genomic databases see[7,8]). Genomic information on animals, plants, and significant disease-causing viruses and bacteria are now easily available to researchers worldwide. Even before COVID-19 was declared a pandemic researchers in China determined the genomic sequencing of the virus[9]. Genomic sequencing is a crucial for designing vaccines, identify variants, determine the virus family and to drugs development[10–12]. The SARS-CoV-2 is a single-stranded RNA virus, with a genome size of 30 Kb, and four structural proteins: Nucleocapsid (N), Matrix (M), Envelope (E) and the Spike (S)[4,10]. The latter is responsible for recognizing and allowing the virus to enter the cell, possibly the main reason why this protein has been widely studied. Mutations in the SARS-CoV-2 viruses result in new variants with mutations in the spike protein increasing replication within cells, and an increased transmissibility[8].

A protein can be depicted as a primary structure formed by a sequence of long strings of characters containing all information: structure, function, hydrophobicity and different motifs. Several researchers have studied how to extract different properties, e. g. hydrophobicity[13–16], fractality[17,18], geometric and thermodynamic aspects[19–21]. Cellular Automata have been widely used to model complex systems with simple, easy-to-understand rules[22], and in recent years many papers were devoted study protein related problems using this approach. Sleit and Mdain[23] proposed a protein folding model based on cellular automata, with straightforward evolutionary rules based on the hydrophobicity of amino acids. Other works dedicated to the same problem include[24–26]. Cellular Automata Image (CAI) analysis[27] is a powerful tool to classify protein structure[28–30] and virus taxonomy[31]. These images can contain important information on the modeled system, for example, CAI allows to differentiate similar systems with respect to those significantly different. The identification of functions, structures, location, and common ancestry of a protein sequence can be performed by a comparison with other know proteins in databases, using alignment, similarity, and homology techniques[32]. In the present paper we propose a protein comparison approach using a cellular automaton image and the information theoretic Hamming metric for the distance between such images, as a measure of similarity and difference, applied to the spike protein. The distance is measured with respect to the S protein in the initial virus strain as first detected in Wuhan, and for the following

[1]CCET, Universidade Federal do Oeste da Bahia, Barreiras 47808-021, Brazil. [2]SENAI-CIMATEC, Salvador 41650 -010, Brazil. [3]ICCMP & IF, Universidade de Brasília, Brasília 70910-900, Brazil. [4]DCET, UNEB, Salvador, Brazil. ✉email: luryane.souza@ufob.edu.br

Amino acids	K	N	D	E	P	Q	R
Decimal code	6	8	9	10	11	12	13
Binary code	00110	01000	01001	01010	01011	01100	01101
Amino acids	**S**	**T**	**G**	**A**	**H**	**W**	**Y**
Decimal code	14	15	16	17	18	20	21
Binary code	01110	01111	10000	10001	10010	10100	10101
Amino acids	**F**	**L**	**M**	**I**	**V**	**C**	
Decimal code	23	24	26	27	28	30	
Binary code	10111	11000	11010	11011	11100	11110	

Table 1. Coding for each of the 20 possible amino acids[28].

Figure 1. Rule 184 from[36] for an elementary cell automaton with three neighbors. The state 0 is represented in white and 1 in black.

Variants Of Concern (VOCs) with mutations of the Spike protein: Alpha (first identified in the United Kingdom), Beta (South Africa), Gamma (Brazil), Delta (India), and the more recent Omicron (South Africa), B.1.1.28, and P2 (Brazil). Our goal is to explicitly obtain the evolutionary relationships between these SARS-CoV-2 variants.

The cellular automata image approach for protein classification and the Hamming distance are presented in "Methods" section. Our results are presented and discussed in "Results and discussion" section, and we close the paper with some concluding remarks in "Concluding remarks" section.

Methods

Cellular automata are discrete dynamical systems with simple local evolution rules and, despite this, can show complex behavior[22]. The rules take into account the state of neighboring cells, analogous to protein structure since physicochemical characteristics of neighboring amino acids influence the folding or function of the protein. The cellular automata considered here has four components: a grid, the set of states, the neighborhood of each state, and the local transition rule. Several possibilities were proposed for encoding the sequence of the 20 types of amino acids in a protein: an 8-digit code for each amino acid[33], or codes reflecting physicochemical characteristics and degeneracy, based on rules of similarity and complementarity: based on molecule recognition and information theory, with a 5-digit code for each amino acid[34], or by representing the amino acid sequences using the hydrophobicity index of each amino acid[28]. The latter in the present work as it allows to better describe the evolutionary relationships between SARS-CoV-2 variants, resulting in smaller distantes for variants with the same mutations and those that emerged in the same period throughout the pandemic. It also groups together variants that share a mutation in the amino acid N501Y. Coronaviruses that cause MERS, SARS and COVID-19 diseases are all closely related, and it is natural to expect that the same coding scheme will be a good representation of the SARS-CoV-2 proteins based in the same molecules. This is reinforced by the discussion in[35] (see particularly Figure 3 of this paper) that shows that the Spike proteins of these viruses have very similar characteristics. Different binary codes were used to distinguish SARS-CoV viruses from other coronaviruses, such as the one used by Xiao et al.[34], which is a simpler code and does not take into account physicochemical amino acids.

Table 1 shows the coding of Ref.[28] that will be used throughout the rest of this work. The Spike protein sequence has 1273 amino acids, and each one is coded as a 5 digit sequence, and thence $N = 6365$ cells with 0 and 1 as possible state, and composing the first line of the cellular automata (initial condition). The state of the i-th cell at step t is notated as $x_i^t = 0, 1, i, i = 1, \ldots, N$. The neighborhood of the cell at position i is composed by the three cells at positions $i-1$, i and $i+1$, resulting in $2^3 = 8$ different states for the neighborhood. We also use periodic boundary conditions. For each possible configuration of the neighborhood, the middle cell can assume two possible states, and thus the number of possible evolution rules is given by $2^8 = 256$[36]. As discussed in[36] and [31], the most appropriate evolution for the celular atomaton rule for SARS-CoV virus classification and for distinguishing them from other viruses, is Wolfram's 184 and depicted in Fig. 1. This rule yield as a typical feature of SARS-CoV viruses a V pattern pattern in the cellular automaton image (see below).

In order to implement a numeric metric to distinguish different images, we consider here the information theoretic Hamming distance D_H, which is commonly used for the distance between sequences of same length and is a simple measure the number of different positions/errors with all required mathematical properties[37]. Here the sequences considered are the states of the automata at the same step$t$. In this case the distance can be written as:

$$D_H(t) = \frac{1}{N}\sum_{i=1}^{N} \|x_i^t - \overline{x}_i^t\|, \tag{1}$$

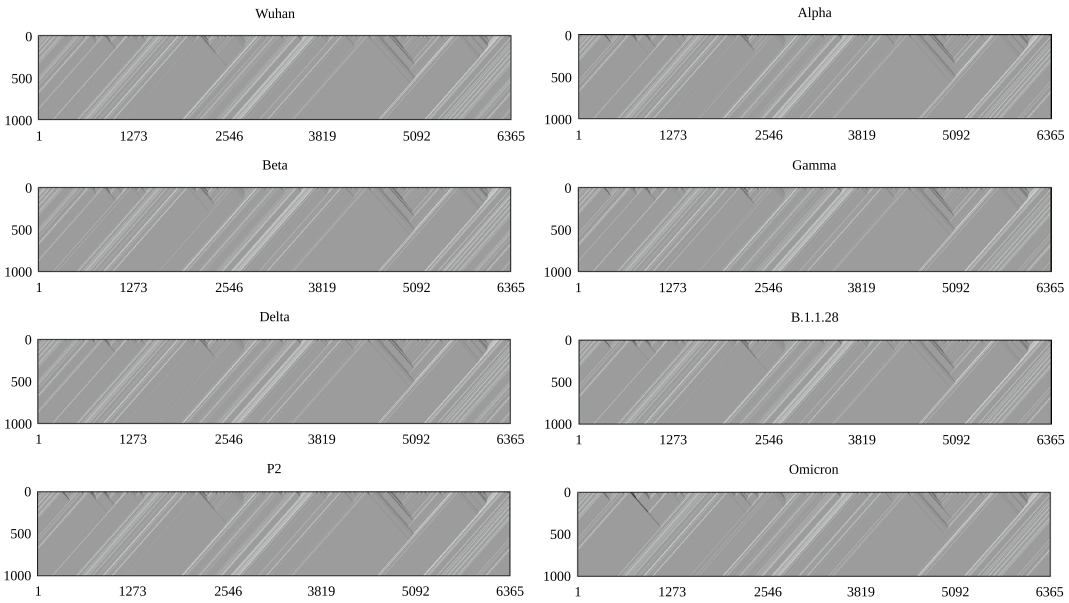


Figure 2. Evolution of protein cellular automata from coding in from Table 1 and the Wolfram's rule in Fig. 1, for the different variants. The horizontyal and vertical axes are the cell number i and the evolution step t, respectively.

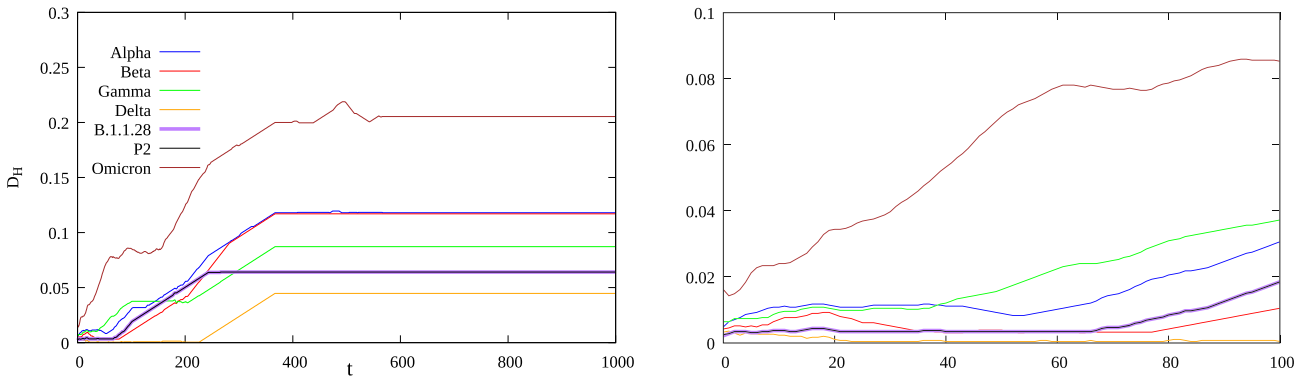


Figure 3. Left: Hamming distance as a function of step t for the time evolution of the cellular automata associated to the spike protein between each variant and the initial Wuhan strain. Right: Zoom over the initial values of t.

with N the size of the grid, x_i^t the state of the cell at step t for the S protein in the initial Wuhan strain and $\bar{x}_i^t$ the Spike protein of the given variant.

Ethics declarations. No human samples/human data were used in the present work.

Results and discussion

The cellular automaton for the SARS-CoV-2 spike protein using available genomic data data for Alpha, Beta, Gamma, Delta, B.1.1.28, P2 variant and the original strain are available at[5] and[38] for Omicron, represented with the coding in Table 1, and evolved according to the rule in Fig. 1 over 1000 time steps. Deletions in the protein sequence were represented by the code 00000 and insertions by introducing the deletion code in the other proteins at the corresponding position. Figure 2 shows the resulting image representing the evolution of the automaton for each considered variant, where the V shaped patterns characteristic of SARS-CoV viruses[31] are clearly visible. Figure 3 shows the time evolution of the Hamming distance D_H for each variant with respect to the original Wuhan strain. For the initial steps the distance has small values, as expect for variants of the same virus, and increases with t up to an asymptotic constant value after approximately $t = 400$ steps. The small number of mutations, if compared to the number of amino-acids in the protein and measured by the small Hamming distance at $t = 0$, is amplified by the evolution of the cellular automata and results in quite different asymptotic values of D_H, after an irregular transient of roughly 200 time steps. This allows us to classify the cellular automata as Wolfram Class IV, with an intermediate behavior between Classes II (periodical) and III (chaotic). Although

Variant	Mutations
Alpha	HV69-70del, Y145del, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H
Beta	L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, A701V
Gamma	L18F, T20N, P26S, HV69-70del, D138Y, Y145H, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F
Delta	T95I, G142D, E154K, L452R, E484Q, D614G, P681R, Q1071H
B.1.1.28	HV69-70del, Y145del, D614G, V1176F
P2	HV69-70del, Y145del, E484K,D614G,V1176F
Omicron	A67V, HV69-70del, T95I, G142D, VYY143-145del, N211I, L212del, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F

Table 2. Mutations the Spike protein of the SARS-CoV-2 variants from[39].

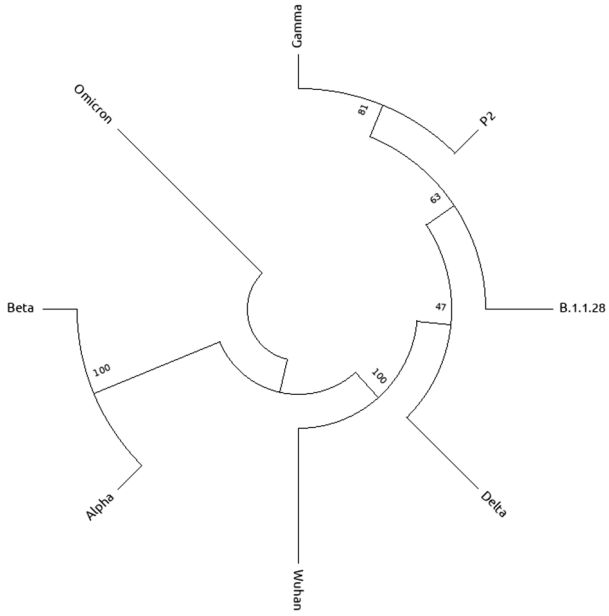


Figure 4. Phylogenetic tree of SARS-CoV-2 variants and the Wuhan strain sequences from the neighbor-joining method[43].

the Omicron variant presents more mutations (and therefore a higher value of D_H) than other known VOCs, with 33 amino acid changes in the spike protein[39], its distance plot remains close to the variants sharing the N501Y mutation (see Table 2 for the characteristic mutations of each variant). This large number of modifications seems to be linked to an increased transmissibility and possibly smaller efficiency of curent vaccines[40].

Table 2 shows the different mutations present in each main variant of the SARS-CoV-2 virus. We then see from Fig. 3 that the present approach groups the variants carrying the N501Y mutation, the sense that final stationary Hamming distance between these variants and the original are more closer and with higher values. The Gamma and P2 variants are also closer as they have the same clade B.1.1.28 (note that the distance for P2 and B.1.1.28 are practically the same in the Figure), while the Delta variant, which carries the P681R mutation unfamiliar to the other variants, is the one with smallest distance. We believe that the present approach is a straightforward way to measure evolutionary distances between SARS-CoV-2 variants, much simpler that other techniques as in[41,42] were a normalized Laplacian pyramid is employed to measure pairwise similarities in cellular automata image wavelet images in order to build phylogenetic trees.

In order to show that the present approach properly relates the variants, we computed the phylogenetic tree from the the neighbor-joining method with alignment[43], which calculates evolutionary distances between species. Figure 4 shows the results for the main known variants of SARS-CoV-2. We see that the variants Gamma, P2, and B.1.1.28 are in the same clade in the tree, while in Fig. 3 these same variants have closer stationary distances. Our results for the Hamming distance fo Delta, Gamma, P2, and B.1.1.28 variants shows that they are closer to the protein initially found in the Wuhan strain, as expected as they are in the same clade in Fig. 4. The same occurs for Alpha and Beta variants, which are in the same clade and have close stationary Hamming distances, while in both approaches the Omicron variant is clearly separated from the other variants. On the other hand, variants B.1.1.28 and P2 have the same stationary Hamming distance, as they have very similar mutations (see Table 2) while P2 is more close to Gamma in the phylogenetic tree.

The variants with smallest values of D_H are those with the smallest number of mutations in Table 2: Delta, B.1.1.28 and P2, which are also the variants without the N501Y mutation. Despite the differences in the images of each variant, resulting from different mutations, the celular automaton rule also results in the V-shaped pattern for SARS-CoV-2 type coronaviruses. This V pattern is characteristic of SARS-CoV-like coronaviruses as discussed in length in Refs.[31,36]. Despite the fact that the SARS-CoV-2 virus is different from SARS-CoV, they share this pattern from their common ancestors. During the COVID-19 pandemic many mutations occurred in the virus sequence, but without a functional change in the Spike protein, although some of these mutations may bring some advantages. However, since different sequencesi perform the same function, mutations in proteins are degenerate, a behavior fundamental for natural selection to occur. Without degeneracy, there is no genetic variability, and this hinders natural selection from acting[44].

Concluding remarks

The approach presented here allows to cluster variants with common ancestors by using a cellular automaton and the asymptotic Hamming distance for the resulting images for each variant, as shown in Fig. 2, and is a more straightforward and simpler evolutionary classification of those variants, than other approaches such as alignment technique, similarity analysis and image processing. Iti particularly discerns the deviation of Omicron with respect to other variants, preserving the V shaped pattern characteristic of the SARS-CoV viruses, despite having the largest number of mutations among known variants, and grouping variants with the N501Y mutation. Furthermore, after just three iterations of the automaton for the protein in the Wuhan strain, the amino acid at position 501 changed from N to Y. This rapid convergence suggest an alternative explanation for the emergence of Alpha, Beta, and Gamma on three continents simultaneously, an evolutionary convergence. We also note that without degeneration, mutations could lead to unfavorable structures for the virus, making it easier to control its spread[44]. Cellular automata are a simple tool to extract meaningful information from proteins sequences, with a very low computational cost. We hope that the present work will contribute as an useful tool to build protein phylogenetic trees.

Data availability

Datasets used during the current study are the sequences of the Spike proteins of the virus initially found in Wuhan [YP_009724390.1] and its variants Alpha [QWP89177.1], Beta [UAL50115.1], Gamma [QXF22923.1], Delta [QXP08802.1], Omicron [UGO97992.1], B.1.1.28 [QQK84800.1] and P2 [QXF22396.1] which are available at https://www.ncbi.nlm.nih.gov/genbank/.

Received: 2 February 2022; Accepted: 7 June 2022
Published online: 18 June 2022

References

1. John Hopkins University. John Hopkins Coronavirus Resource Center (2021). Available online: https://coronavirus.jhu.edu/map.html. Accessed 4 Sept 2021.
2. Tooze, A. *Shutdown—How Covid Shook the World's Economy* (Penguin Random House, 2021).
3. World Health Organization. WHO timeline-COVID-19. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=CjwKCAiA7dKMBhBCEiwAO_crFAhknuq4kc_PZRW1qx3v_bMHTvAmmEewQ2vyKtZ47HyUy7DLGlZxoCkC4QAvD_BwE#event-115 (2020). Accessed 17 Nov 2021.
4. Machhi, J. *et al.* The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections. *J. Neuroimmune Pharmacol.* **15**, 359–386. https://doi.org/10.1007/s11481-020-09944-5 (2020).
5. GenBank. National Center for Biotechnology Information (2021).
6. UniProt. The Universal Protein Resource (2021).
7. Chen, C., Huang, H. & Wu, C. H. Protein bioinformatics databases and resources. *Methods Mol. Biol.* **1558**, 3–39. https://doi.org/10.1007/978-1-4939-6783-4_1 (2017).
8. NIH—National Library of Medicine. NCBI SARs-CoV-2 Resources (2021).
9. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 1–8. https://doi.org/10.1038/s41586-020-2008-3 (2020).
10. Khan, M. T. *et al.* Structures of SARS-CoV-2 RNA-binding proteins and therapeutic targets. *Intervirology* **64**, 1–14. https://doi.org/10.1159/000513686 (2021).
11. Chou, K. C., Wei, D. Q. & Zhong, W. Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun.* **308**, 148–151. https://doi.org/10.1016/S0006-291X(03)01342-1 (2003).
12. Chou, K. C., Wei, D. Q., Du, Q. S., Sirois, S. & Zhong, W. Z. Progress in computational approach to drug development against SARS. *Curr. Med. Chem.* **13**, 3263–3670. https://doi.org/10.2174/092986706778773077 (2006).
13. Moret, M. A. & Zebende, G. F. Amino acid hydrophobicity and accessible surface area. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **75**, 011920. https://doi.org/10.1103/PhysRevE.75.011920 (2007).
14. Phillips, J. C. Scaling and self-organized criticality in proteins I. *Proc. Natl. Acad. Sci.* **106**, 3107–3112. https://doi.org/10.1073/pnas.0811262106 (2009).
15. Phillips, J. C. Synchronized attachment and the Darwinian evolution of coronaviruses CoV-1 and CoV-2. *Physica A Stat. Mech. Appl.* **581**, 126202. https://doi.org/10.1016/j.physa.2021.126202 (2021).
16. Li, S., Cai, C., Gong, J., Liu, X. & Li, H. A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing. *Proteins Struct. Funct. Bioinform.* **89**, 1541–1556. https://doi.org/10.1002/prot.26176 (2021).
17. Moret, M. A., Miranda, J. G. V., Nogueira, E., Santana, M. C. & Zebende, G. F. Self-similarity and protein chains. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **71**, 012901. https://doi.org/10.1103/PhysRevE.71.012901 (2005).
18. Moret, M. A., Santana, M. C., Nogueira, E. & Zebende, G. F. Protein chain packing and percolation threshold. *Physica A Stat. Mech. Appl.* **361**, 250–254 (2006).
19. Moret, M. A. Self-organized critical model for protein folding. *Physica A Stat. Mech. Appl.* **390**, 3055–3059. https://doi.org/10.1016/j.physa.2011.04.008 (2011).
20. Xu, X. L., Shi, J. X., Wang, J. & Li, W. Long-range correlation and critical fluctuations in coevolution networks of protein sequences. *Physica A Stat. Mech. Appl.* **562**, 125339. https://doi.org/10.1016/j.physa.2020.125339 (2021).

21. Nelson, E. D. & Onuchic, J. N. Proposed mechanism for stability of proteins to evolutionary mutations. *Proc. Natl. Acad. Sci.* **95**, 10682–10686. https://doi.org/10.1073/pnas.95.18.10682 (1998).
22. Toffoli, T. & Margolus, N. *Cellular Automata Machines: A New Environment for Modeling* (MIT Press in Scientific Computation, 1987).
23. Sleit, A. & Madain, A. Protein folding in the two-dimensional hydrophobic polar model based on cellular automata and local rules. *Int. J. Comput. Netw. Inf. Secur.* **16**, 48 (2016).
24. Varela, D. & Santos, J. Protein folding modeling with neural cellular automata using Rosetta. In GECCO '16 Companion: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion, GECCO '16 Companion, 1307–1312 (Association for Computing Machinery, 2016).
25. Varela, D. & Santos, J. Protein folding modeling with neural cellular automata using the Face-Centered Cubic model (2017). Published in IWINAC 19 June 2017.
26. Varela, D. & Santos, J. Automatically obtaining a cellular automaton scheme for modeling protein folding using the FCC model. *Nat. Comput.*https://doi.org/10.1007/s11047-018-9705-y *(2019).*
27. Wolfram, S. Cellular automata as models of complexity. *Nature* **311**, 419–424 (1984).
28. Xiao, X. & Chou, K. Digital coding of amino acids based on hydrophobic index. *Protein Pept. Lett.* **14**, 871–5 (2007).
29. Xiao, X., Wang, P. & Chou, K. C. Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image. *J. Theor. Biol.* **254**, 691–6. https://doi.org/10.1016/j.jtbi.2008.06.016 (2008).
30. Kavianpour, H. & Vasighi, M. Structural classification of proteins using texture descriptors extracted from the cellular automata image. *Amino Acids* **49**, 261–271. https://doi.org/10.1007/s00726-016-2354-5 (2017).
31. Wang, M. *et al.* A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. *Med. Chem.*https://doi.org/10.2174/1573406053402505 *(2005).*
32. Gabler, F. *et al.* Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinform.* **72**, e108. https://doi.org/10.1002/cpbi.108 (2020).
33. Ghosh, S. & Chaudhuri, P. P. Cellular automata model for proteomics and its application in cancer immunotherapy. In *Cellular Automata. ACRI 2018. Lecture Notes in Computer Science*, 3–15 (Springer International Publishing, 2018).
34. Xiao, X., Shao, S., Ding, Y. & Chen, X. Digital coding for amino acid based on cellular automata. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 5, 4593–4598, https://doi.org/10.1109/ICSMC.2004.1401256 (2004).
35. Phillips, J. C., Moret, M. A., Zebende, G. F. & Chow, C. C. Phase transitions may explain why SARS-CoV-2 spreads so fast and why new variants are spreading faster. *Physica A* **598**, 127318. https://doi.org/10.1016/j.physa.2022.127318 (2022).
36. Xiao, X. *et al.* Using cellular automata to generate image representation for biological sequences. *Amino Acids* **28**, 29–35. https://doi.org/10.1007/s00726-004-0154-9 (2005).
37. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x (1950).
38. Mullen, J. L. *et al.* Outbreak. Info (2021). Accessed 17 Dec 2021.
39. European Centre for Disease Prevention and Control. Implications of the emergence and spread of the SARS-CoV-2 b.1.1. 529 variant of concern (Omicron) for the EU/EEA. https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-emergence-sars-cov-2-variant-b.1.1.529 (2021). Accessed 17 Dec 2021.
40. World Health Organization. Enhancing Readiness for Omicron (b.1.1.529): Technical brief and priority actions for member states. https://www.who.int/publications/m/item/enhancing-readiness-for-omicron-(b.1.1.529)-technical-brief-and-priority-actions-for-member-states (2021). Accessed 17 Dec 2021.
41. Wu, Z. C., Xiao, X. & Chou, K. C. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J. Theor. Biol.* **267**, 29–34. https://doi.org/10.1016/j.jtbi.2010.08.007 (2010).
42. Rahman, M. M., Biswas, B. A. & Bhuiyan, M. I. H. Protein similarity analysis by wavelet decomposition of cellular automata images. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6 (IEEE, 2019).
43. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–442. https://doi.org/10.1093/oxfordjournals.molbev.a040454 (1987).
44. Edelman, G. M. & Gally, J. A. Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci.* **98**, 13763–13768. https://doi.org/10.1073/pnas.231499798 (2001).

Author contributions

L.F.S. conducted the computer modelling, M.A.M. separated the protein sequence data from the variants,T.M.R.F. made the images of the automata. All analyzed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.F.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Capítulo Quatro

Artigo 2

4.1 New distance measure for comparing protein cellular automata image.

Artigo enviado para Royal Society Open Science.

Autores: Luryane Ferreira de Souza, Hernane Borges de Barros Pereira, Tarcísio Marciano da Rocha Filho, Bruna Aparecida Souza Machado, Marcelo Albano Moret Simões Gonçalves.

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Article submitted to journal

Subject Areas:
xxxxx, xxxxx, xxxx

Keywords:
Cellular Automata, Hamming Distance, Phylogeny, Proteins

Author for correspondence:
Luryane Ferreira de Souza
e-mail: luryane.souza@ufob.edu.br

New distance measure for comparing protein cellular automata image

Luryane F. Souza[1,2], Hernane B. de B. Pereira[2,3], Tarcisio M. da Rocha Filho[4], Bruna A. S. Machado[2] and Marcelo A. Moret[2,3]

[1]Universidade Federal do Oeste da Bahia, CCET, Barreiras, Brazil
[2]SENAI-CIMATEC, Salvador, Brazil
[3]Universidade do Estado da Bahia, UNEB, Salvador, Brazil
[4]ICP and Instituto de Física, Universidade de Brasília, Brasília, Brazil

One of the first steps in protein sequence analysis is comparing sequences to look for similarities. We propose an information theoretical distance to compare cellular automata representing protein sequences, and determine similarities. Our approach relies in a stationary Hamming distance for the evolution of the automata according to a properly chosen rule, and to build a pairwise similarity matrix and determine common ancestors among different species in a simpler and less computationally demanding computer codes when compared to other methods.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

Bioinformatics has been growing recently and consolidating itself as a research area, bringing together researchers from different areas such as molecular biology, physics, mathematics, and computer science, among others. Its origins date back to the first protein sequencing studies in the 1950s when insulin was sequenced well before the first microcomputers [1–3]. More recently, there has been a significant increase in sequencing tools, with sequence analysis software such as ClustalW [4] and protein databases as UniProt [5], GenBank [6], PDB [7], and PDB2 [8]. Alongside bioinformatics, molecular biology has grown in recent years with the emergence of different computational approaches designed to study protein folding and sorting [9–11]. A protein is a macromolecule made up of 20 types of amino acids that can represented in its primary form as a string of characters for each amino acid. Different approaches for graphical representations of protein sequences were proposed in Refs. [11–21]. The approach we use in the present work is based on cellular automata (CA) images generated using the sequence of a given protein as initial state. Amino acids are encoded into valid entries of a cellular automaton (see for instance Xiao et al. [17]), using a digital code based on the rules of similarity, complementarity, molecular recognition theory and information theory [16]. A coding based on hydrophobicity indices of amino acids was used in a reduced form by Kavianpour and Vasighi [11] to encode protein sequences to extract features from the images and determine the structural class of the protein. Chaudhuri et al. [20] used a coding with an eight-digit binary code based on the analysis of the molecular structure of each amino acid.

Similarities among sequences, i. e. homologous sequences, hold important information, such as similar functions or as an indication of a recent common ancestor [22]. Indeed, evolutionary relationships between protein sequences can be determined from sequence comparison methods [23–28]. Rahman et al. [21] proposed a method for decomposing CA images using wavelet decomposition and used the horizontal image of this decomposition for protein comparison from an image quality metric. The Hamming distance have been successfully used to evaluate the stability of the concentration of soot during controlled combustion of acetylene and natural gas, within the spatiotemporal standards generated by the evolution of the CA-based system [29]. In previous work by some of the authors [28] we used cellular automata imaging of the Spike proteins to compare variants of the SARS-CoV-2 virus using the stationary Hamming distance, and determined the variants that shared recent common ancestors by looking only at the evolution of the distance between the variant CA and the one for the reference protein initially found in Wuhan-China. As a continuation, we propose a method to build the distance matrix between species pairs using the stationary Hamming distance measuring the dissimilarity between different species. We apply this method for different proteins: ND5, ND6, transferrin, and beta-globin. The proposed approach is effective for grouping similar species and, using cophenetic correlation coefficients, building dendrograms similar to those obtained using the p-distance from the package MEGA [30].

Considering that the p-distance measures the differences between two sequences and that information loss may occur when transforming a protein sequence into a cellular automaton, our results confirm that this loss is minimal, and that our methodology can be used in the analysis of similar proteins. Another advantage is that we use a simple comparison metric not requiring more elaborate processing methods or image textures.

2. Methods

A Cellular automaton (CA) is a discrete dynamical system in both space and time evolving under a given spatially local rule. Despite their simplicity they often model complex systems [31]. It is defined from five components: L, S, N, f and B, with L a n-dimensional spatial lattice of cells with values c_i^t, $i = 1, 2, 3, \ldots, M$ at time t. Each cell assumes values in the set of possible states S.

The neighborhood N of a given cell i is the set of cells considered in the transition rule f. Finally, the boundary conditions of the automaton is represented by B.

Here we consider in the presen work one-dimensional cellular automata with a neighborhood of cell i it given by the cells $i-1$, i and $i+1$:

$$N(c_i^t) = \{c_{i-1}^t, c_i^t, c_{i+1}^t\} \;\; i = 1, 2, \ldots, M.$$

and a set of two possible states $S = \{0, 1\}$. Therefore, we have a total of $2^3 = 8$ different possible neighborhoods. The transition function f expresses the state assumed by each cell in the next time step according to its neighborhood as a query list for each possible neighborhood state. We thus have a total of $2^8 = 256$ possible evolution rules for the cellular automaton, each rule enumerated 0 to 255 given by the decimal form of it binary representation, as exemplified in Figure 1. The boundary condition B determining the neighborhood of cells at the extremities of the cellular automaton can be of four types: fixed contour, random, periodic and reflecting [32]. We consider here periodic boundary conditions such that that $c_{M+1} = c_1$ and $c_0 = c_M$. Figure 2 illustrates the procedure of forming an image of the cellular automaton composed by the lines for each discrete time value t according to the evolution rule.

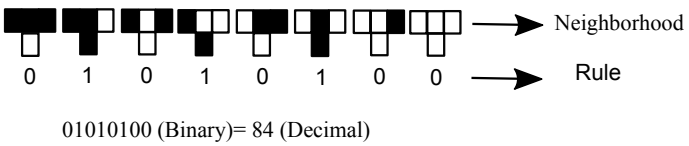


Figure 1: Cellular automaton rule n^o 84 with for the eight types of neighborhoods in the cellular automaton.

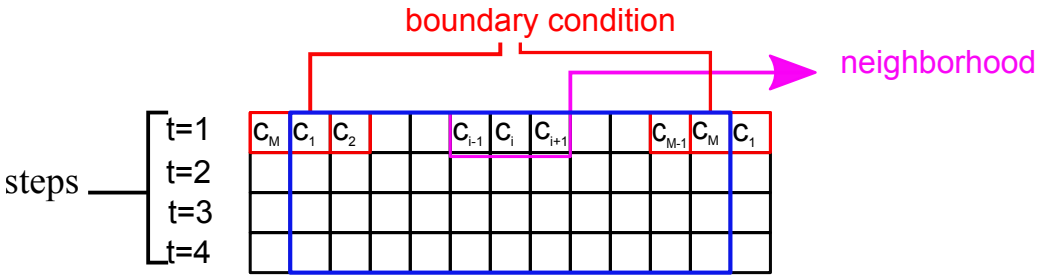


Figure 2: Cellular automaton image

Since we use a two state cellular automaton, each amino acid will be encoded in a binary code, with different possibilities studied in the literature [11,16,20,33]. For the present study we use the code proposed by Chaudhuri et al. [20] and shown in Table 1, which encodes each amino acid with an 8-digit code based on the molecular structures of amino acids. Among other possibilities, this choice is justified by the fact that it yields a better grouping of close species.

The first step is to align the different protein sequences considered such that each associated automata are of the same size. Deletions and losses are identified and properly represented in each automaton by the respective codes in Table 1. Considering a protein of size P, the initial condition will have a size of $M = 8 \times P$. As a first illustration of our approach we show in Fig. 3 the cellular automaton image of Beta-Globin protein for six different animal species, for a total evolution of $t = 500$ steps (this value will be used for the remaining of the present paper).

The images of the associated cellular automata provide signatures for different proteins and are used to determine similarities/differences between species. A comparison between those images

Table 1: Encoding of amino acids, deletions and missing protein sequence data after alignment.

Aminoacid	Code	Aminoacid	Code
Glycine (G)	00000000	Cysteine (C)	01000100
Alanine (A)	00000100	Threonine (T)	00110100
Proline (P)	00100110	Asparagine (N)	00101110
Valine (V)	00010110	Glutamine (Q)	00101111
Methionine (M)	00110110	Tyrosine (Y)	10100100
Tryptophan (W)	10110110	Histidine (H)	01111110
Phenylalanine (F)	10000100	Lysine (K)	00110111
Isoleucine (I)	00011110	Arginine (R)	01111111
Leucine (L)	00010111	Aspartic Acid (D)	01110100
Serine (S)	00100100	Glutamic Acid (E)	01110110
Deletion (-)	11111111	Missing information (?)	11111110

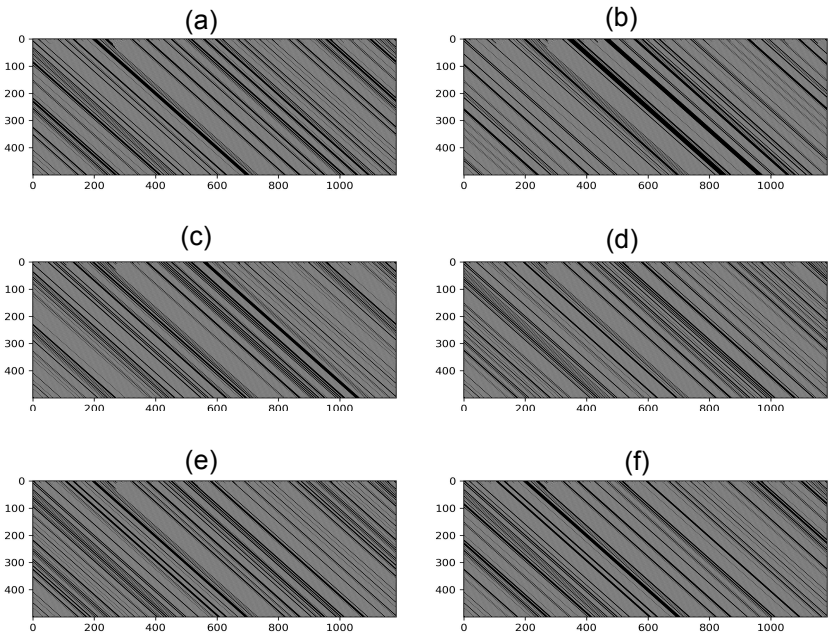


Figure 3: Cellular automaton image of the Beta-Globin protein for A) Human, B) Shark, C) Catfish, D) Turtle, E) Swift and F) Coyote species.

can be performed with a low computational cost using the Hamming distance from information theory [34] given by the number of changes needed to transform one sequence into another, and implement for the generated images of two cellular automata CA_A and CA_B as:

$$D_H(t) = \frac{1}{M}\sum_{i=1}^{M} |a_i^t - b_i^t|, \tag{2.1}$$

with a_i^t and b_i^t the values of the ith cells of automata CA_A and CA_B at step t, respectively. The size of the automaton is $M = 8 \times P$, where P is the size of the aligned protein sequence. As shown in the next section the Hamming distance saturates after a relatively small number of steps. We denominate this saturated value the Stationary Hamming Distance (SHD), which is then used to build the similarity matrix.

3. Results

We apply our approach to the following four protein sequences: beta-globin, NADH Dehydrogenase 5 (ND 5), NADH Dehydrogenase 6 (ND 6) and transferrin. This choice was motivated by the requirement to be able to perform comparisons with previous results in the literature [13,35]. All sequences were aligned using the ClustalW system [30,36]. Our results are then compared to those obtained from pairwise p-distance from ClustalW.

3.1 NADH Dehydrogenase 5 (ND 5)

Protein ND 5 is a sub-unit of the mitochondrial respiratory enzyme complex I (NADH: ubiquinone oxidoreductase) [37], and is responsible for mitochondrial electron transport. Mutations and defects in this enzyme can cause Leigh's disease and MELAS syndrome. Being highly conserved in eukaryotes, we use data from these sequences to analyze similarities between mammalian species. We consider here the following nine species: Human, Gorilla, Pigmy Chimpanzee, Common Chimpanzee, Fin Whale, Blue Whale, Rat, Mouse, and Opossum. All sequences were taken from the NCBI protein database [6], and their identifications are given in Table S1 of the Supplementary Information. The aligned sequences have 613 entries each, and are represented using the coding in Table 1, resulting into a binary sequence of length 4904 for the initial condition of the cellular automaton. The cellular automata image is the generated from the prescription in the previous section and the Hamming distance in Eq. (2.1) between two species as a function of the number of steps. The results for the distance between each of the nine species and Humans are shown in Fig. 4.

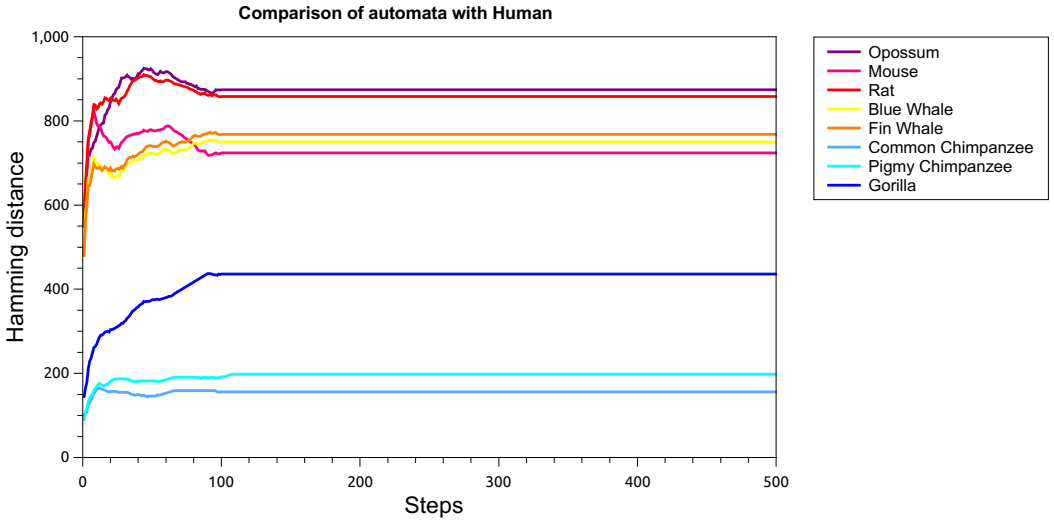


Figure 4: Hamming distance between the cellular automata image for some different mammalian species and Human, as a function of the number of steps.

We then build distance matrices from the SHD and the p-distance, and obtain corresponding dendrograms using the average method from R studio hierarchical grouping [38,39], and shown in Fig. 5. Both dendrograms are identical, with the exception of a small difference of the closest relative of human. Nevertheless both methods correctly group families: Hominidae, Balaenopteridae, Muridae, and Didelphidae. Other methods such as [35] yield dendrograms identical to the one obtained from our method. The cophenetic correlation coefficient [40] between the two dendrogram in Fig. 5 is 0.9940, indicating a very close similarity.

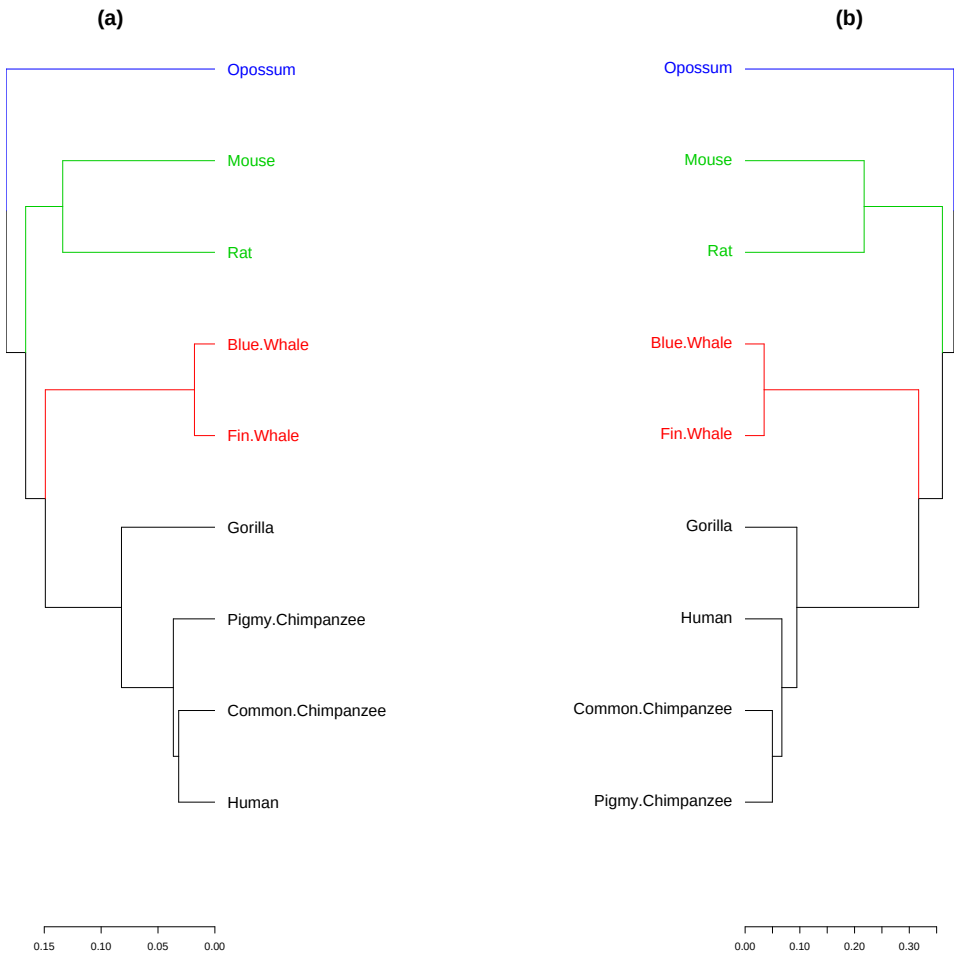


Figure 5: Dendrograms from the ND 5 protein from (a) SHD and (b) p-distance. The four families are well grouped in both dendrograms: Didelphidae (blue), Muridae (green), Balaenopteridae (red), and Hominidae (black).

3.2 NADH Dehydrogenase 6 (ND 6)

The NADH Dehydrogenase 6 (ND 6) protein is a sub-unit of the NADH dehydrogenase (ubiquinone) enzyme, located in the mitochondrial inner membrane. Mutations or errors in their sequences can cause Leigh's disease and spinal muscular atrophy [41]. Protein sequences were obtained from NCBI [6], and their identifications are given in the Table S2. The aligned sequences have 176 positions, and thence the initial condition for this protein has $8 \times 176 = 1408$ cells. We follow the same procedure as for the previous case: the SHD between each pair among the

following species: Human, Gorilla, Common Chimpanzee, Gray Seal, Harbor Seal, Rat, Mouse, and Wallaroo. Dendrograms are then obtained from the distance matrices using the SHD and the p-distance, and shown in Fig. 6. The two dendrograms are identical and both methods group families correctly: Macropodidae, Muridae, Phocidae, and Hominidae. We note that other methods as alignment-free similarity analysis [35] cannot separate the Macropodidae family from the Muridae. The cophenetic correlation coefficient between the two dendrograms in Fig. 6 is 0.9797, indicating again a very close similarity.

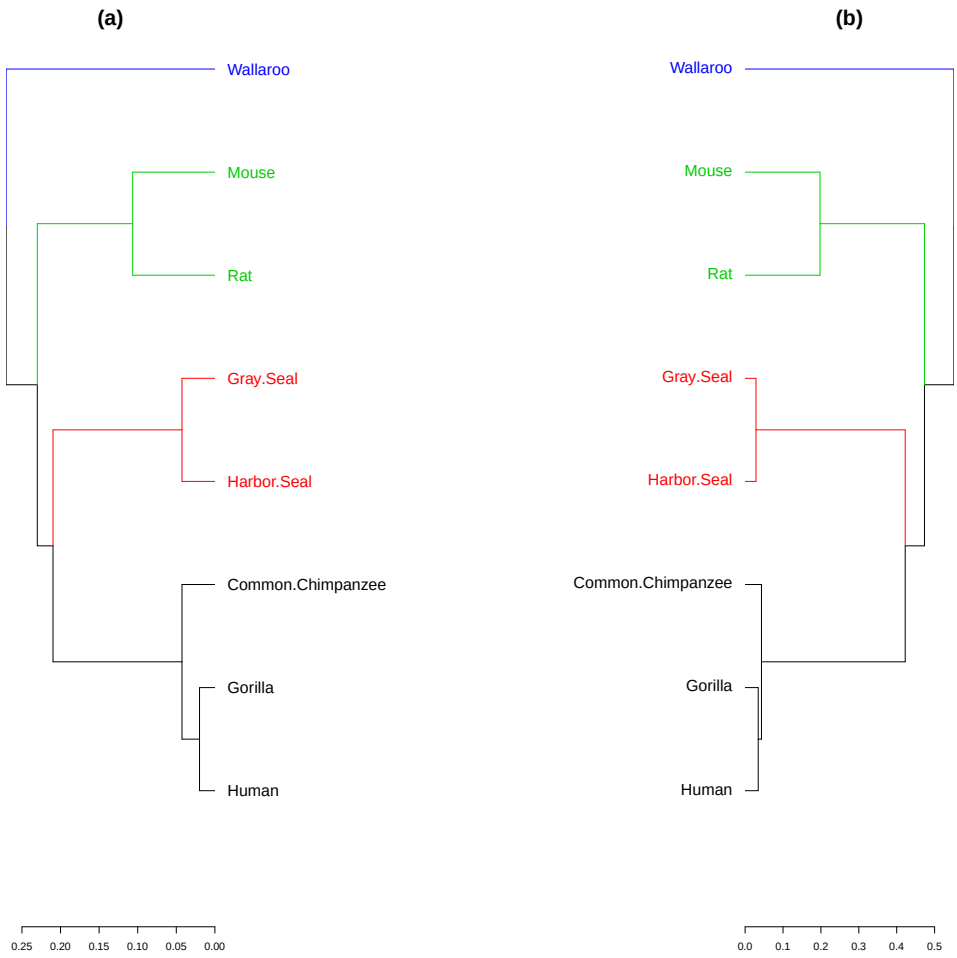


Figure 6: Dendrograms from the ND 6 protein obtained from distance matrices using (a) SHD and (b) p-distance, with the indication of family groupings: Macropodidae (blue), Muridae (green), Phocidae (red), and Hominidae(black).

3.3 Transferrin

Transferrin is an iron-binding protein keeping iron at a low concentration in biological fluids. Serum transferrin (TF) is present in mammals, amphibians, and fish [42], and plays an essential role in fighting bacterial infections in fish [43]. Blood iron overload is a rare condition that characterizes hereditary atransferrinemia [44]. We consider a set of 24 transferrin protein sequences across Mammalia, Amphibian, and Actinopterygii species from the NCBI database [6] with the respective NCBI identification given in the Table S3. The aligned sequences have 750

positions. Each one is then encoded into a binary code of size $8 \times 750 = 6000$. The dendrogram obtained from the SHD and p-distance are shown in Fig. 7. Our approach correctly classifies all species into their respective groups: Mammalia, Amphibian and Actinopterygii and separately grouping mammals' serum transferrin (TF) and lactotransferrin (LF). It also group correctly species from the genus Salmo (Brown Trout, Atlantic Salmon), Salvelinus (Japanese Char, Brook Trout, Lake Trout), and Oncorhynchus (Amago Salmon, Sockeye Salmon, Rainbow Trout, Coho Salmon, Chinook Salmon) in the Actinopterygii class. Only Amago Salmon (TF) and Sockeye Salmon (TF) were grouped differently, but the same problem has already been reported in previous works [43]. The cophenetic correlation coefficient between the two dendrograms is 0.9671, again indicating a very good similarity between the clusters.

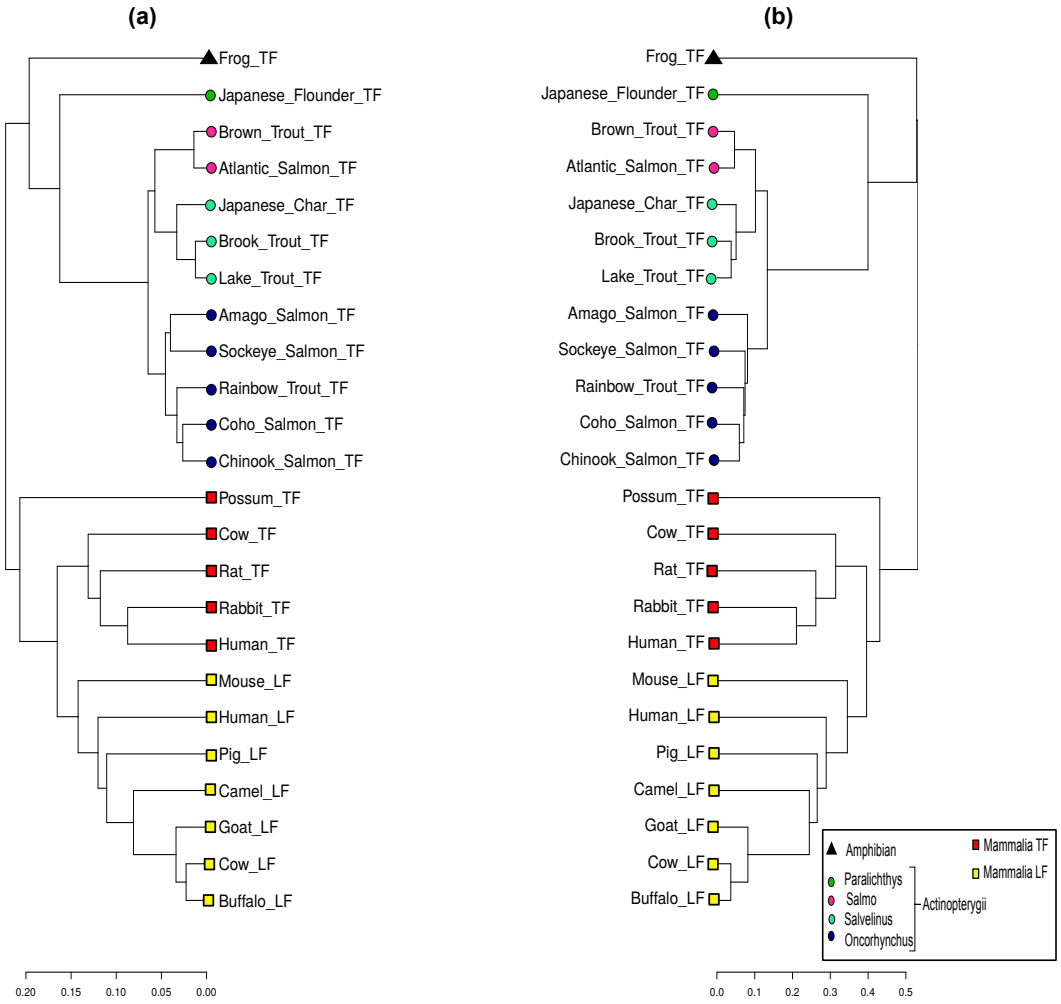


Figure 7: Transferrin protein dendrograms from (a) SHD and (b) p-distance distance matrices.

3.4 Beta-Globin

Hemoglobin comprises two chain pairs α and β, which have distinct chains of amino acids, two dimers of $\alpha - \beta$ form hemoglobin. Its principal function is to carry oxygen from blood to tissues.

Mutations in the beta-globin chain can cause sickle cell anemia [45]. We consider here 50 beta-globin sequences from different species taken from the NCBI database [6], with identifications given in the Table S4. The aligned sequences have 148 positions. The initial condition is then coded in $8 \times 148 = 1184$ cell digits. The corresponding distance matrices for SHD and p-distance then have $50 \times 50 = 2500$ entries, and the corresponding dendrograms are shown in Fig. 8.

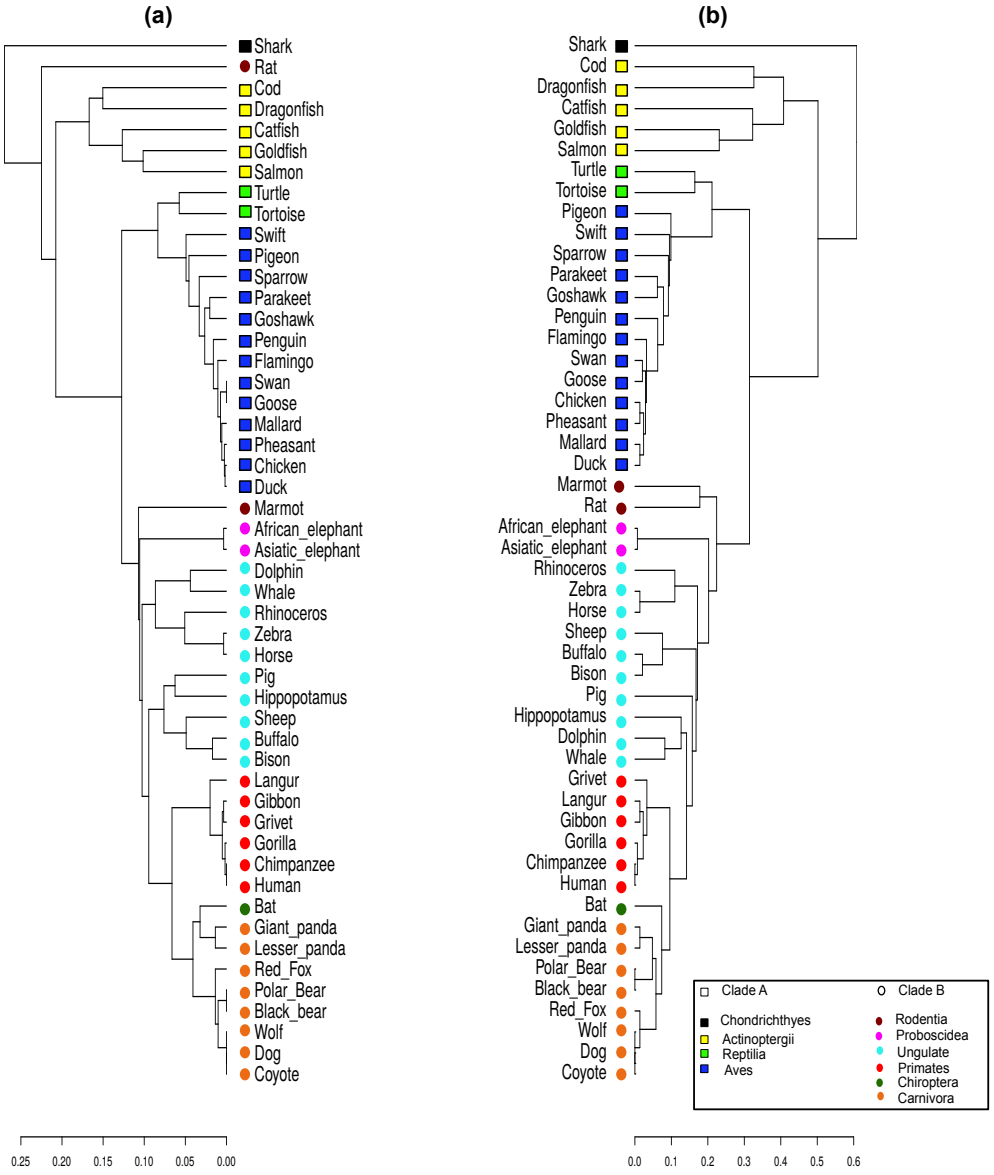


Figure 8: Beta-globin protein dendrograms from SHD (a) and p-distance (b) values. Clade A is represented by squares and Clade B by circles. Different animal groups are represented by colors.

Our approach yields a consistent classification of the identified clusters. At variance the results in [13] and the dendrogram obtained using the p-distance, that separates mammals from non-mammals, our approach failed in this point, with rat classified in Clade A. This same inconsistency was also observed in a previous work [35]. Other divergences of the method involve some more recent families but keep the tree similar to the one obtained using the p-distance. These

discrepancies observed for Beta-Globin may be due to the fact that we considered a significant and more diverse number of species, and is reflected in the value of 0.8790 for the cophenetic correlation coefficient between the two dendrograms, clearly below the other proteins considered here, but still acceptable.

4. Concluding remarks

We discussed and showed that Cellular automata are a tool for visual comparing of protein sequences and for determining their similarity. We expanded the use of this tool by introducing the use of the Hamming distance from information theory, in order to compare the cellular automata images obtained. Our approach allows to determine phylogenetic relations among species with a good accuracy if one considers that one protein was used in each of the dendrogram presented, but nevertheless has some limitations. We applied it to lysozyme protein sequences (not shown here), with inconclusive results, with the possible explanation that the sequences fir this cases not homologous but are the result of convergent evolution. In this case, the resulting dendrograms from both our method and by using p-distance cannot approximate species with recent common ancestors.

The main advantage of our method compared to using the p-distance to build the distance matrix is that it codes each amino acid according to its structure, such that similar amino acids have closer codings, such that the distance measure we use will give different weights for different mutations, as observed in [28], which used a code based on the hydrophobicity of amino acids such that the sequences that underwent mutations with a change in hydrophobicity had a greater distance and also a different weight for each type of mutation.

Other approaches use textures from images [21] to compare cellular automata. Ours requires a low computational cost and no processing methods or image textures, with an efficient protein comparison. As a first work we used an evolution rule previously proposed in the literature, but in forthcoming research we will consider other possibilities, and are currently investigating the possibility of coding proteins using the hydrophobicity scale proposed by Moret and Zebende [46].

Ethics. No human samples/human data were used in the present work.

Data Accessibility. The accession numbers of the sequences used are shown in the supplementary material.

Authors' Contributions. L.F.S. and M.A.M. conceptualization, data curation, investigation; M.A.M Project administration, supervision; L.F.S formal analysis, methodology, validation, writing − original draft; T.M.R.F visualization; T.M.R.F, B.A.S.M and M.A.M funding acquisition, resources; H.B.B.P. software, validation. All authors writing review and editing.

Competing Interests. The authors declare no competing interests.

Funding. This work was partially funded by National Council of Technological and Scientific Development – CNPq (Brazil), grant numbers 312857/2021-7 TMRF and 305096/2022-2 MAM.

References

1. Sanger F, Thompson EOP. 1953a The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **53**, 353–366.
2. Sanger F, Thompson EOP. 1953b The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates.. *The Biochemical journal* **53**, 366–374.

3. Gauthier J, Vincent AT, Charette SJ, Derome N. 2018 A brief history of bioinformatics. *Briefings in Bioinformatics* **20**, 1981–1996.
4. Thompson JD, Gibson TJ, Higgins DG. 2003 Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics* **00**, 2.3.1–2.3.22.
5. UniProt. 2021 The Universal Protein Resource. .
6. GenBank. 2021 National Center for Biotechnology Information. .
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000 The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242.
8. PDB. 2023 Protein Data Bank. .
9. Moret MA, Santana MC, Nogueira E, Zebende GF. 2006 Protein chain packing and percolation threshold. *Physica A* **361**, 250–254.
10. Moret M. 2011 Self-organized critical model for protein folding. *Physica A* **390**, 3055–3059.
11. Kavianpour H, Vasighi M. 2017 Structural classification of proteins using texture descriptors extracted from the cellular automata image. *Amino Acids* **49**, 261–271.
12. Mu Z, Wu J, Zhang Y. 2013 A novel method for similarity/dissimilarity analysis of protein sequences. *Physica A: Statistical Mechanics and its Applications* **392**, 6361–6366.
13. Mu Z, Yu T, Liu X, Zheng H, Wei L, Liu J. 2021 FEGS: a novel feature extraction model for protein sequences and its applications. *BMC Bioinformatics* **22**.
14. Liao B, Liao B, Sun X, Zeng Q. 2010 A Novel method for similarity analysis and protein subcellular localization prediction. *Bioinformatics* **26**, 2678–2683.
15. Wu Z, Xiao X, K.C. C. 2010 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J Theor Biol.* **267(1)**, 29–34.
16. Xiao X, Shao S, Ding Y, Chen X. 2004 Digital coding for amino acid based on cellular automata. In *004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* vol. 5 pp. 4593–4598.
17. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou K. 2005 Using cellular automata to generate image representation for biological sequences. *Amino Acids* **28**, 29–35.
18. Xiao X, Wang P, Chou K. 2011 Cellular automata and its applications in protein bioinformatics.. *Curr Protein Pept Sci.* **12(6)**, 508–519.
19. Wang M, Yao JS, Huang ZD, Xu ZJ, Liu GP, Zhao HY, Wang XY, Yang J, Zhu YS, Chou KC. 2005 A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. *Medicinal Chemistry* pp. 39–47.
20. Chaudhuri PP, Ghosh S, Dutta A, Choudhury SP. 2018 pp. 291–325. In *Cellular Automata (CA) Model for Protein*, pp. 291–325. Singapore: Springer Singapore.
21. Rahman MM, Biswas BA, Bhuiyan MIH. 2019 Protein Similarity Analysis by Wavelet Decomposition of Cellular Automata Images. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* pp. 1–6.
22. Pearson WR. 2013 An Introduction to Sequence Similarity (âĂIJHomologyâĂİ) Searching. *Current Protocols in Bioinformatics* **42**, 3.1.1–3.1.8.
23. Lipman DJ, Pearson WR. 1985 Rapid and Sensitive Protein Similarity Searches. *Science* **227**, 1435–1441.
24. Campanella JJ, Bitincka L, Smalley J. 2003 MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences.. *BMC Bioinformatics* **4**.
25. Prakash A, Jeffryes M, Bateman A, Finn RD. 2017 The HMMER Web Server for Protein Sequence Similarity Search. *Current Protocols in Bioinformatics* **60**, 3.15.1–3.15.23.
26. Hu G, Kurgan L. 2019 Sequence Similarity Searching. *Current Protocols in Protein Science* **95**, e71.
27. Moret MA, Miranda JGV, Nogueira E, Santana MC, Zebende GF. 2005 Self-similarity and protein chains. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **71**, 012901.
28. Souza LF, Rocha Filho TM, Moret MA. 2022 Relating SARS-CoV-2 variants using cellular automata imaging. *Scientific Reports* **12**.
29. Souza J, Pereira H, Santos A, Senna V, Moret M. 2014 A new proposal for analyzing combustion process stability based on the Hamming distance. *Physica A* **413**, 301–306.

30. Tamura K, Stecher G, Kumar S. 2021 MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* **38**, 3022–3027.
31. Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M. 2008 Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids* **34**, 111–117.
32. Pereira HBB, Zebende GF, Moret MA. 2010 Learning computer programming: Implementing a fractal in a Turing Machine. *Computers & Education* **55**, 767–776.
33. Xiao X, Chou K. 2007 Digital Coding of Amino acids based on hydrophobic index.. *Protein and Peptide Letters* **14**, 871–875.
34. Hamming RW. 1950 Error detecting and error correcting codes. *The Bell System Technical Journal* **29**, 147–160.
35. Saw AK, Tripathy BC, Nandi S. 2019 Alignment-free similarity analysis for protein sequences based on fuzzy integral. *Scientific Reports* **9**.
36. Thompson JD, Higgins DG, Gibson TJ. 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680.
37. Cardol P. 2011 Mitochondrial NADH:ubiquinone oxidoreductase (complex I) in eukaryotes: A highly conserved subunit composition highlighted by mining of protein databases. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807**, 1390–1397.
38. Saraçli S, Doğan N, Doğan İ. 2013 Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications* **203**, 1–8.
39. R Core Team. 2018 R: A Language and Environment for Statistical Computing. .
40. Sokal RR, Rohlf FJ. 1962 The Comparison of Dendrograms by Objective Methods. *Taxon* **11**, 33–40.
41. NCBI gene. 2022 MT-ND6 mitochondrially encoded NADH dehydrogenase 6 [homo sapiens (human)] - gene - NCBI. .
42. Lambert LA, Perri H, Meehan T. 2005 Evolution of duplications in the transferrin family of proteins. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **140**, 11–25.
43. Ford MJ. 2001 Molecular Evolution of Transferrin: Evidence for Positive Selection in Salmonids. *Molecular Biology and Evolution* **18**, 639–647.
44. Aslan D, Crain K, Beutler E. 2007 A New Case of Human Atransferrinemia with a Previously Undescribed Mutation in the Transferrin Gene. *Acta Haematologica* **118**, 244–247.
45. Hsia CC. 1998 Respiratory Function of Hemoglobin. *New England Journal of Medicine* **338**, 239–248.
46. Moret MA, Zebende GF. 2007 Amino acid hydrophobicity and accessible surface area. *Phys. Rev. E* **75**, 011920.

Supplementary Information

All sequence data were taken from the NCBI genome database https://www.ncbi.nlm.nih.gov/protein. Below are the identifications of each species and protein used.

Table S1 . Identification of ND5 protein sequences at NCBI.

Sequence name	Species	ID (NCBI)
Human	Homo sapiens	AP 000649.1
Gorilla	Gorilla gorilla	NP 008222.1
Common chimpanzee	Pan troglodytes	NP 008196.1
Pigmy chimpanzee	Pan paniscus	NP 008209.1
Fin whale	Balenoptera physalus	NP 006899.1
Blue whale	Balenoptera musculus	NP 007066.1
Rat	Rattus norvegicus	AP 004902.1
Mouse	Mus musculus	NP 904338.1
Opossum	Didelphis virginiana	NP 007105.1

Table S2 . Identification of ND6 protein sequences at NCBI.

Sequence name	Species	ID (NCBI)
Human	Homo sapiens	YP 003024037.1
Gorilla	Gorilla gorilla	NP 008223.1
Common chimpanzee	Pan troglodytes	NP 008197.1
Harbor seal	Phoca vitulina	NP 006939.1
Gray seal	Halichoerus grypus	NP 007080.1
Rat	Rattus norvegicus	AP 004903.1
Mouse	Mus musculus	NP 904339.1
Wallaroo	Osphranter robustus	NP 007405.1

Table S3 . Identification of Transferrin protein sequences at NCBI.

Sequence name	Species	ID (NCBI)
Human TF	Homo sapiens	S95936.1
Rabbit TF	Oryctolagus cuniculus	P19134.4
Rat TF	Rattus norvegicus	D38380.1
Cow TF	Bos Taurus	U02564.1
Buffalo LF	Bubalus bubalis	AJ005203.1
Cow LF	Bos Taurus	X57084.1
Goat LF	Capra hircus	X78902.1
Camel LF	Camelus dromedarius	AJ131674.1
Pig LF	Sus scrofa	AAA31102.1
Human LF	Homo sapiens	NM 002343.6
Mouse LF	Mus musculus	NM 008522.3
Possum TF	Trichosurus vulpecula	AF092510.1
Frog TF	Xenopus laevis	X54530.1
Japanese flounder TF	Paralichthys olivaceus	D88801.1
Atlantic salmon TF	Salmo salar	L20313.1
Brown trout TF	Salmo trutta	D89091.1
Lake trout TF	Salvelinus namaycush	D89090.1
Brook trout TF	Salvelinus fontinalis	D89089.1
Japanese char TF	Salvelinus leucomaenis pluvius	D89088.1
Chinook salmon TF	Oncorhynchus tshawytscha	AH008271.2
Coho salmon TF	Oncorhynchus kisutch	D89084.1
Sockeye salmon TF	Oncorhynchus nerka	D89085.1
Rainbow trout TF	Oncorhynchus mykiss	D89083.1
Amago salmon TF	Oncorhynchus masou	D89086.2

Table S4 . Identification of Beta-Globin protein sequences at NCBI.

Sequence name	Species	ID (NCBI)
Human	Homo sapiens	AAA16334.1
Pigeon	Columba Livia	P11342.1
Goshawk	Accipiter gentilis	P08851.1
Black Bear	Ursus thibetanus	P68012.1
Lesser Panda	Ailurus fulgens	P18982.1
Asiatic Elephant	Elephas maximus	P02084.1
Giant Panda	Ailuropoda melanoleuca	P18983.2
African Elephant	Loxodonta africana	P02085.1
Sheep	Ovis aries	P02075.2
Tortoise	Chelonoidis niger	P83123.3
Duck	Anas platyrhynchos	P02114.2
Grivet	Chlorocebus aethiops	P02028.1
Mallard	Anas platyrhynchos platyrhynchos	P02115.1
Gorilla	Gorilla gorilla gorilla	P02024.2
Goose	Anser anser anser	P02117.1
Shark	Heterodontus portusjacksoni	P02143.1
Rat	Rattus norvegicus	CAA33114.1
Hippopotamus	Hippopotamus amphibius	P19016.1
Penguin	Aptenodytes forsteri	P80216.1
Horse	Equus caballus	P02062.1
Swift	Apus apus	P15165.1
Gibbon	Hylobates lar	P02025.1
Coyote	Canis latrans	P60525.1
Whale	Balaenoptera acutorostrata	P18984.1
Catfish	Silurus asotus	O13163.2
Bat	Macroderma gigas	P24660.1
Bison	Bison bonasus	P09422.1
Red Fox	Vulpes vulpes	P21201.1
Swan	Cygnus olor	P68945.1
Marmot	Marmota marmota	P08853.1
Buffalo	Bubalus bubalis	P67820.1
Salmon	Salmo salar	Q91473.3
Dog	Canis lupus familiaris	P60524.1
Sparrow	Passer montanus	P07406.1
Chimpanzee	Pan troglodytes	P68873.2
Pheasant	Phasianus colchicus colchicus	P02113.1
Dolphin	Tursiops truncatus	P18990.1
Flamingo	Phoenicopterus ruber	P02121.1
Goldfish	Carassius auratus	P02140.1
Pig	Sus scrofa	P02067.3
Polar bear	Ursus maritimus	P68011.1
Dragonfish	Cygnodraco mawsoni	ADD73488.1
Rhinoceros	Rhinoceros unicornis	P09907.1
Parakeet	Psittacula krameri	P21668.1
Chicken	Gallus gallus	P02112.2
Zebra	Equus zebra	P67824.1
Wolf	Chrysocyon brachyurus	P60526.1
Cod	Gadus morhua	O13077.2
Turtle	Chrysemys picta bellii	P13274.1
Langur	Semnopithecus entellus	P02032.1

Capítulo Cinco

Artigo 3

5.1 Hydrophobic analysis of the SARS-CoV-2 Spike Protein.

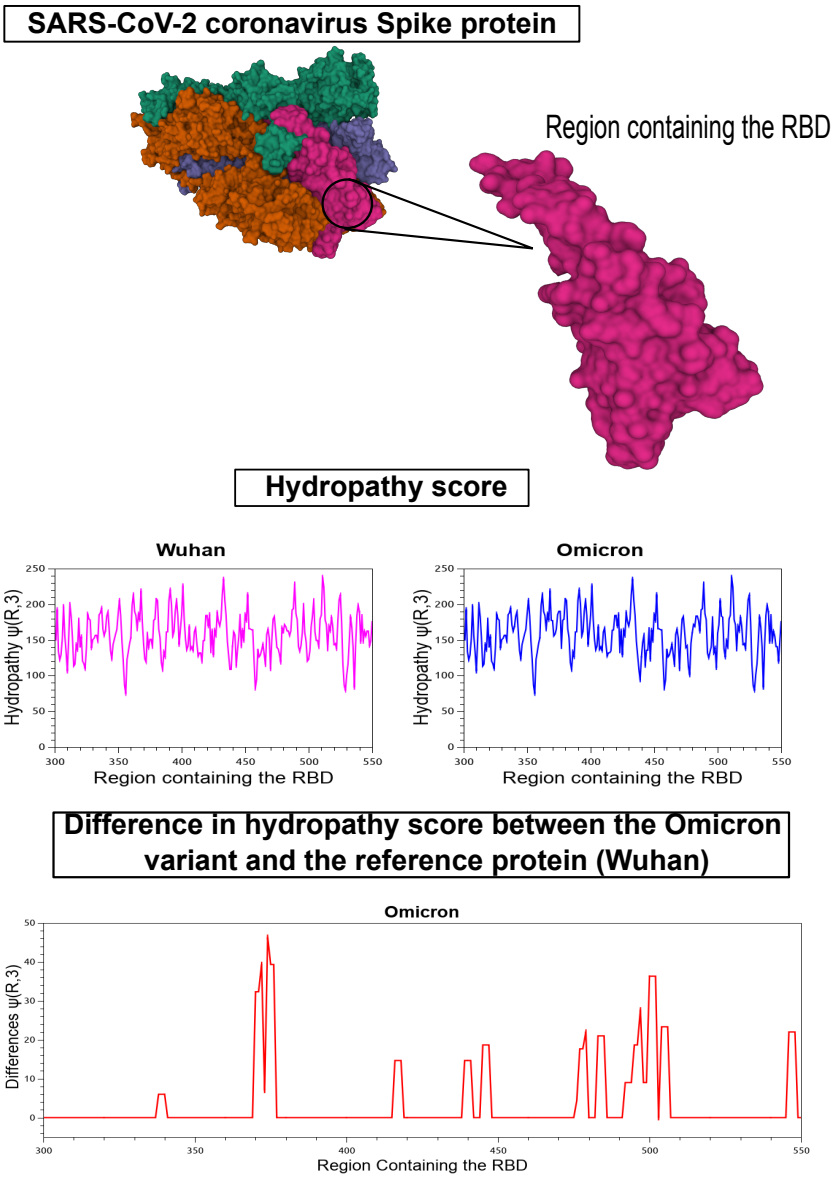
Artigo enviado para Journal of Molecular Biology (JMB).

Autores: Luryane Ferreira de Souza, Thiago Barros Murari, James C. Phillips, Hernane Borges de Barros Pereira, Tarcísio Marciano da Rocha Filho, Bruna Aparecida Souza Machado, Marcelo Albano Moret Simões Gonçalves.

Graphical Abstract

Hydrophobic analysis of the SARS-CoV-2 Spike Protein

Luryane F. Souza, Thiago B. Murari, James C. Phillips, Hernane B. B. Pereira, Tarcisio M. da Rocha Filho, Bruna A. S. Machado, Marcelo A. Moret



Highlights

Hydrophobic analysis of the SARS-CoV-2 Spike Protein

Luryane F. Souza, Thiago B. Murari, James C. Phillips, Hernane B. B. Pereira, Tarcisio M. da Rocha Filho, Bruna A. S. Machado, Marcelo A. Moret

- The Spike (S) protein from the SARS-CoV-2 coronavirus is the target of effective vaccines and treatments for COVID-19. The receptor-binding domain (RBD) region in the S1 region binds to the host receptor and is responsible for the first interaction between the virus and the host. Therefore, mutations in this region may affect the efficiency of vaccines and treatments. This paper investigates changes in the hydropathic profile that occurred in the RBD of variants of concern (VOCs).
- The hydrophobic shapes of the alpha, beta, and gamma variants are similar, which may be one of the factors for these variants not competing in the case of numbers in 2020 when they appeared.
- We showed that the Omicron variant was the variant that most changed the hydropathic profile in the RBD region, and these changes may be related to the prevalence of this variant in all regions of the planet, and greater resistance of this variant to treatments and vaccines compared to the other VOCs studied here.

Hydrophobic analysis of the SARS-CoV-2 Spike Protein

Luryane F. Souza[a,b,*], Thiago B. Murari[b], James C. Phillips[c], Hernane B. B. Pereira[b,d], Tarcisio M. da Rocha Filho[e], Bruna A. S. Machado[b], Marcelo A. Moret[b,f]

[a] *CCET, UFOB, Rua Professor José Seabra de Lemos, 316, Recanto dos Pássaros, Barreiras, 47808-021, BA, Brazil*
[b] *SENAI-CIMATEC, Av. Orlando Gomes, 1845, Piatã, Salvador, 41650-010, BA, Brazil*
[c] *Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, United States of America*
[d] *DCH, UNEB, Rua Silveira Martins, 2555,Cabula, Salvador, 41150-000, BA, Brazil*
[e] *Instituto de Física, Universidade de Brasília, Campus Universitário Darcy Ribeiro, Brasília, 70910-900, DF, Brazil*
[f] *DCET, UNEB, Rua Silveira Martins, 2555,Cabula, Salvador, 41150-000, BA, Brazil*

Abstract

The hydrophobic properties of the amino acids present in the primary structure of the SARS-CoV-2 spike proteins are investigated through the thermodynamic amino acid scale. Estimates for the hydrophobic shape are obtained for the receptor binding domain in several Spike proteins of SARS-CoV-2 variants. The findings show a possible explanation for the decrease in immune antibodies and effectiveness of vaccines for the Omicron variant and the consequent increase in the rate of reinfection. In addition, the hydrophobic form of the alpha, beta, and gamma variants is one of the factors that contributed to these variants not competing during their outbreaks.

Keywords: COVID-19, Hydrophobic scale, receptor binding domain
PACS: 87.15.kp, 82.70.Uv, 05.45.Df, 87.15.Qt

*Corresponding author
Email addresses: luryane.souza@ufob.edu.br (Luryane F. Souza), mamoret@gmail.com (Marcelo A. Moret)

1. Introduction

The emergence of new variants of SARS-CoV-2 was the reason for several outbreaks of the COVID-19 pandemic. Through 2022, the most prevalent variant of concern worldwide is the Omicron [1] variant and its subvariants. The first cases of the Omicron variant occurred on 11 November 2021 in patients from Botswana. Since it was detected, this variant has aroused the interest of scientists, as it has many mutations, mainly in the receptor binding domain (RBD) [2]. The RBD subunit is a region of the Spike protein responsible for binding to receptor angiotensin-converting enzyme 2 ACE2 to begin replication in human cells [3]. As the target of most vaccines is to neutralize the RBD, this number of mutations in this region may be favorable to the virus and decrease the vaccine's effectiveness for this new variant [4].

It is well known that proteins are part of almost all biological processes. Hydrophobic forces and the evolution of proteins are the main drivers of the packing of proteins toward their native structures. The phase transition between the unfolded state and the native structure is not necessarily a first-order transition. This type of phase transition in proteins generates fractality [5, 6, 7, 8], and self-organized criticality [9, 10, 11, 12, 13], among other typical characteristics of complex systems [14, 15].

Furthermore, this phase transition gives rise to a thermodynamic scale based on the power law of the loss of solvent-accessible surface area (ASA) around a specific amino acid, which characterizes the index of hydrophobicity of the amino acid. Moret and Zebende [16] proposed this index. Different contexts have applied this index, such as protein evolution [17, 18], protein networks [19], critical fluctuations in nature [12, 13], estimation of the free energy contribution of each amino acid [20], and long-range correlation in protein dynamics [21]. Additionally, to build the matrix of hydropathy scores ($\Psi(R, W)$) [18, 22] among other applications.

The hydropathy profile of the protein sequence is used to analyze its internal hydrophobic, and outer hydrophilic regions, such as [23, 24, 25, 26]. In addition, they are also used in the analysis of secondary structures and transmembrane regions. This profile will be the average between the hydrophobicity indexes of the amino acids in the considered neighborhood [27].

Proteins are synthesized in messenger RNA. Coronaviruses are large RNA viruses, with the largest among them being characterized as non-segmented and single-stranded positive-sense RNA [28]. SARS-CoV-2 has a 29.9 kb genome capable of encoding several proteins [29]. Among them, the spike

protein of SARS-CoV-2 plays a critical role in the virus, as it is responsible for binding to the receptor (ACE2) on the host cell and determining host tropism [30]. A small number of amino acids appears to be important for the binding of ACE2 and S proteins [31]. The S protein is responsible for binding to the receptor on the host cell and for fusing the viral membrane to that of the host cell [32, 33].

In this context, we will analyze the difference in the hydropathy profile between the five sequences of the Spike protein of the worrying variants of SARS-CoV-2, Alpha, Beta, Gamma, Delta, and Omicron and the reference sequence (Wuhan) in the RBD region.

2. Results

The local hydrophobic profile can be estimated by the average hydrophobic behavior of three neighboring amino acids ($\Psi(R, 3)$). Mutations cause hydrophobic changes in SARS-CoV-2 RNA. Figure 1 depicts these changes in the Spike protein RDB of SARS-CoV-2 VOCs compared to the first measured Spike protein (Wuhan).

In Figure 1, the Alpha, Beta, Gamma, and Omicron variants present the N501Y mutation, which changes the local hydrophobic profile of these VOCs. This mutation increases the binding of these VOCs to human ACE2 [34]. Furthermore, what draws attention in Figure 1 is the difference in the local hydrophobic profile of the Omicron variant compared to the local profile of the virus initially found in Wuhan, in the RBD region. This change in the profile of Omicron's local hydropathy can be hydrophobic and hydrophilic. Therefore, Figure 1 cannot be used to find the internal and external regions of the folded protein. However, we can easily visualize with this figure that of all the VOCs studied here, Omicron is the one that most changes its local hydrophobic profile.

For a better understanding of the weights of the mutations in the protein conformation, we calculated the difference in the $\Psi(R, 39)$ hydrophobicity among the first Spike protein observed (Wuhan) and their VOCs in the RBD region (Figure 2). Again, we observed a remarkable change in the Omicron variant.

Figure 2 shows $|\Psi(R, 39)_{Wuhan} - \Psi(R, 39)_{VOC}|$ which could be similar to the hydrophobic effect when measured far from the spike protein. These mutations in the Omicron variant result in an attack rate, and herd immunity has lost its importance in the fight against COVID-19 [35]. Furthermore,

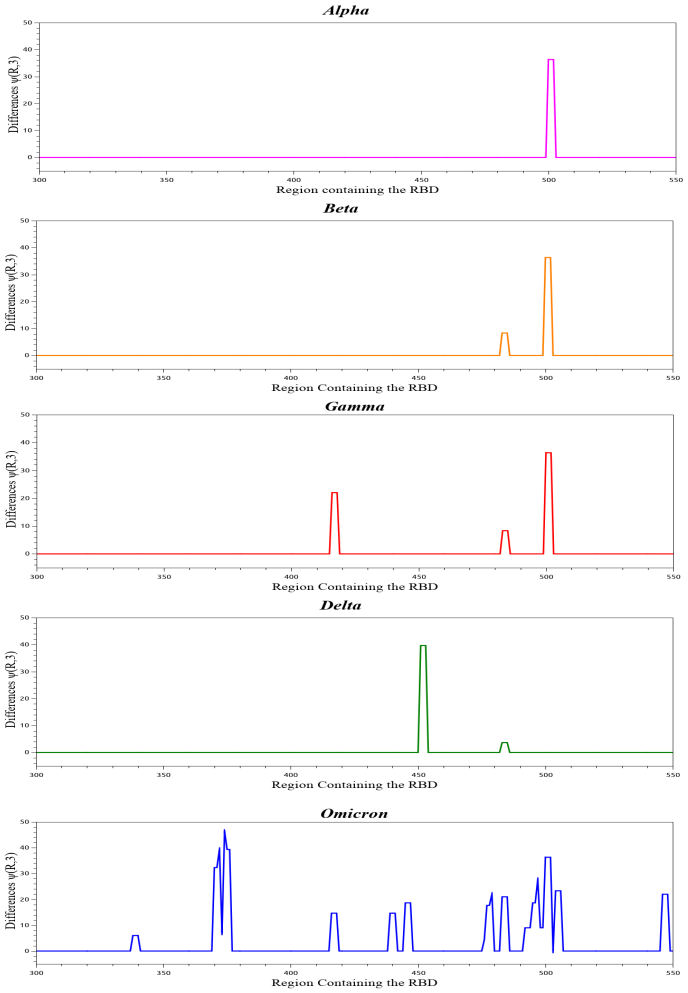


Figure 1: The difference in the hydrophobic behavior to the local hydrophobicity ($\Psi(R, 3)$.

previous work concluded that this variant is resistant to therapeutic monoclonal antibodies. The factor that most contributes to this resistance is not the weight of the mutations separately, but the combination of these mutations that makes this variant especially worrying [36].

Alpha, Beta, and Gamma variants cocirculated worldwide in the same period, unlike Delta and Omicron. Due to the N501Y mutation, these three variants caused waves of COVID-19 in three different regions of the world: South Africa, the United Kingdom, and Brazil. This mutation was responsi-

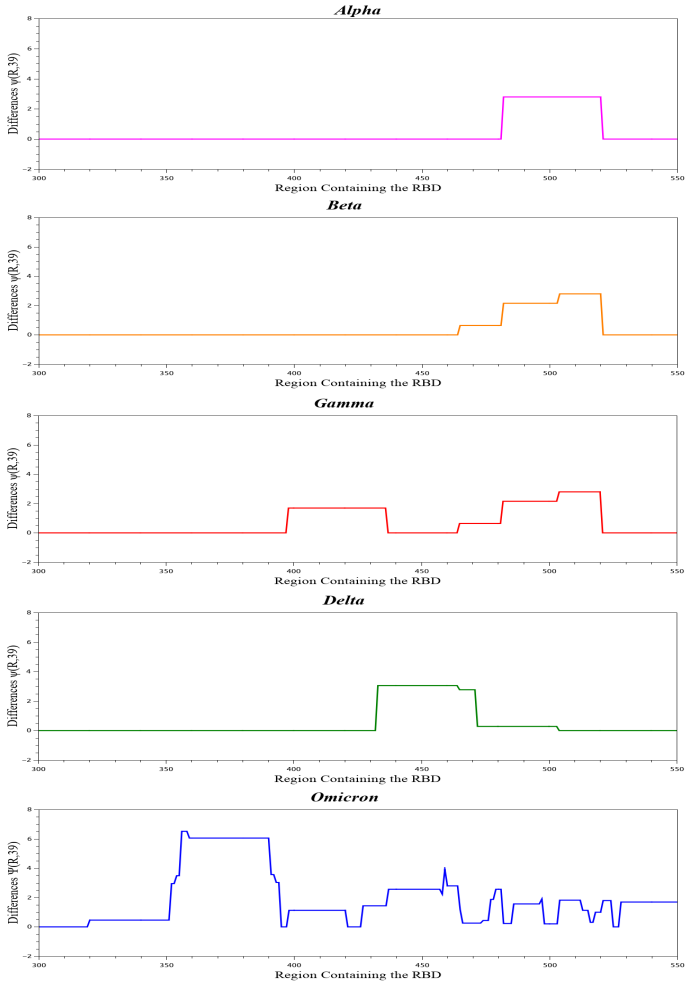


Figure 2: The matrix of hydropathy scores ($\Psi(R, 39)$) presents a similar behavior to alpha, beta and gamma variants. Furthermore, this behavior should be similar to the hydrophobic profile for objects far from the surface of the protein.

ble for increased transmission rates and the ability to evade immunity [37]. The hydrophobic profiles presented in Figures 1 and 2 agree in the same sense, showing that these variants have similar hydrophobic behavior in many aspects.

3. Discussion

It was observed that the N501Y mutation occurs after three iterations in cellular automata [38]. This fact may explain the emergence of Alpha, Beta, and Gamma. Notably, the Beta or Gamma variants do not show significant contamination in countries where Alpha is the dominant variant in the COVID-19 outbreak. Similar behavior occurs when Beta is the dominant alpha variant, or Gamma does not have a significant percentage of contamination. This behavior also occurs for the Gamma variant. This behavior of the dominant variant only changed when the Delta variant emerged and spread worldwide.

The self-similar behavior of the amino acid hydrophobicity is a key to knowledge about the virus and cell binding. Similar hydrophobic behavior to Alpha, Beta, and Gamma variants and dissimilar behavior to Delta and Omicron variants were observed in $\Psi(R, 3)$ and $\Psi(R, 39)$ (Figure 1 and 2).

The hydrophobic shape in the RBD region changes significantly for the Omicron variant. Therefore, $\Psi(R, 3)$ provides the hydrophobic landscape that may explain the decreased vaccine efficacy for the Omicron variant [39]. Otherwise, $\Psi(R, 39)$ explains why Alpha, Beta, and Gamma do not compete with each other.

Finally, both $\Psi(R, 3)$ and $\Psi(R, 39)$ significantly change the hydrophobic shape to the Omicron variant. These amino acid mutations are key to the reinfection observed from Omicron variants.

4. Methods

The analysis method uses a thermodynamic amino acid scale [16] that considers the loss of solvent-accessible surface area (ASA) around a central residue. The ASA for a specified amino acid residue in the center of a protein fragment scales as a power law of fragment length with a well-ordered negative exponent for all 20 amino acids. Comparing the loss of ASA and power-law gives us a measure of whether an amino acid has a nonpolar or polar side chain. Therefore, it is possible to infer the hydrophobicity of the amino acid, that is, whether an amino acid has a hydrophobic side chain or if it has a hydrophilic chain. In addition, the power law scale is the hallmark for the critical point of a second-order phase transition and the existence of universal parameters of amino acid-water interactions is evidence of a thermodynamic phase transition model for proteins [40, 41, 17]. To the best of

our knowledge, proteins are the only large-scale networks that exhibit both first order unfolding phase transitions and second-order conformational phase transitions described by fractals.

Specifically, the 20 exponents provide a measure of the average hydropathy of each residue at the center of an arbitrary background neighborhood. The smaller the magnitude of the exponent (see Table 1), the slower the loss of ASA and therefore the more hydrophilic the amino acid will be on average. This differs from many attempts to attribute hydropathy based on the chemical properties of an amino acid alone [16].

Amino Acid	A	C	D	E	F	G	H	I	K	L
Hydropathy Index	157	246	87	94	218	156	152	222	69	197
Amino Acid	M	N	P	Q	R	S	T	V	W	Y
Hydropathy Index	221	113	121	105	78	100	135	238	174	222

Table 1: Thermodynamic amino acid scale [16]

For a globular protein, hydrophilic regions are more likely to reside near the outside of the protein and vice versa for more hydrophobic regions. Then the effect of neighboring residues must be considered to be a central amino acid. This requires deducing an effective domain length that dominates the conformation of the protein, the details of which at the molecular level are unknown. This length may differ from protein to protein, although there may be preferred lengths. Specifically, for Spike proteins this length is close to 40 amino acids [22, 42].

We analyzed the hydropathic profile of a region of the Spike protein sequence, the RBD. The RBD of the SARS-CoV-2 Spike protein is identified in the region between residues 331 and 524 [43]. We will consider regions 300 to 550 in the sequence for the hydropathy profile. We consider two different hydropathy profiles. The first one analyzes the weight of mutations in the local profile that will be calculated by:

$$|\Psi(R,3)_{Wuhan} - \Psi(R,3)_{VOC}|, \tag{1}$$

where $\Psi(R,3)$ is the hydropathy of residue R considering the average hydropathy between this residue and its immediate neighbors in the sequence. The second hydropathy profile measures the weight of mutations in the protein conformation and will be calculated by:

$$|\Psi(R,39)_{Wuhan} - \Psi(R,39)_{VOC}|, \quad (2)$$

where $\Psi(R,39)$ is the hydropathy of residue R considering the average hydropathy between this residue and its nineteen neighbors on the right and nineteen neighbors on the left. We calculated local and conformational hydropathic profile differences for the following variants: Alpha, Beta, Gamma, Delta, and Omicron. The results are shown in Figures 1 and 2.

Acession Numbers

In the current study, we used the following data taken from the NCBI website [44], for the Spike protein sequence of the SARS-CoV-2 virus and its variants: Wuhan (NCBI: **YP_009724390.1**), Alpha (GenBank: **QWP89177.1**), Beta (GenBank: **UAL50115.1**), Gamma (GenBank: **QXF22923.1**), Delta (GenBank: **QXP08802.1**) and Omicron (GenBank: **UGO97992.1**)

References

[1] W. Ou, J.and Lan, X. Wu, T. Zhao, B. Duan, P. Yang, Y. Ren, L. Quan, W. Zhao, J. Seto, D.and Chodosh, Z. Luo, Q. Wu, J.and Zhang, Tracking SARS-CoV-2 omicron diverse spike gene mutations identifies multiple inter-variant recombination events, Signal Transduction and Targeted Therapy 7 (2022) 138. doi:10.1038/s41392-022-00992-2.

[2] X. He, W. Hong, X. Pan, G. Lu, X. Wei, SARS-CoV-2 omicron variant: Characteristics and prevention, MedComm 2 (4) (2021) 838–845. doi:https://doi.org/10.1002/mco2.110.

[3] N. K. Routhu, N. Cheedarla, V. S. Bollimpelli, S. Gangadhara, V. V. Edara, L. Lai, A. Sahoo, A. Shiferaw, T. M. Styles, K. Floyd, S. Fischinger, C. Atyeo, S. A. Shin, S. Gumber, S. Kirejczyk, K. H. Dinnon, P.-Y. Shi, V. D. Menachery, M. Tomai, C. B. Fox, G. Alter, T. H. Vanderford, L. Gralinski, M. S. Suthar, R. R. Amara, SARS-CoV-2 RBD trimer protein adjuvanted with Alum-3M-052 protects from SARS-CoV-2 infection and immune pathology in the lung, Nature Communications 12 (2021) 3587. doi:10.1038/s41467-021-23942-y.

[4] T. B. Murari, L. M. S. Fonseca, H. B. B. Pereira, A. S. Nascimento Filho, H. Saba, F. A. Scorza, A. C. G. Almeida, E. L. N. Maciel, J. F. F.

Mendes, T. M. Rocha Filho, D. J. R., R. Badaró, B. A. S. Machado, M. A. Moret, Retrospective cohort study of COVID-19 in patients of the brazilian public health system with SARS-COV-2 omicron variant infection, Vaccines 10 (9) (2022) 1504. doi:10.3390/vaccines10091504.

[5] J. Liang, K. A. Dill, Are proteins well-packed?, Biophysical Journal 81 (2) (2001) 751–766. doi:https://doi.org/10.1016/S0006-3495(01)75739-6.

[6] M. A. Moret, J. G. V. Miranda, E. Nogueira, M. C. Santana, G. F. Zebende, Self-similarity and protein chains, Phys. Rev. E 71 (2005) 012901. doi:10.1103/PhysRevE.71.012901.

[7] M. A. Moret, M. C. Santana, E. Nogueira, G. F. Zebende, Protein chain packing and percolation threshold, Physica A: Statistical Mechanics and its Applications 361 (1) (2006) 250–254. doi:https://doi.org/10.1016/j.physa.2005.08.001.

[8] M. A. Moret, M. C. Santana, G. F. Zebende, P. G. Pascutti, Self-similarity and protein compactness, Physical Review E 80 (4) (2009) 041908. doi:https://doi.org/10.1103/PhysRevE.80.041908.

[9] J. C. Phillips, Scaling and self-organized criticality in proteins I, Proceedings of the National Academy of Sciences of the United States of America 106 (2009) 3107–3112. doi:10.1073/pnas.0811262106.

[10] J. C. Phillips, Scaling and self-organized criticality in proteins II, Proceedings of the National Academy of Sciences 106 (9) (2009) 3113–3118. doi:10.1073/pnas.0811308105.

[11] M. A. Moret, Self-organized critical model for protein folding, Physica A: Statistical Mechanics and its Applications 390 (17) (2011) 3055–3059. doi:https://doi.org/10.1016/j.physa.2011.04.008.

[12] Q. Y. Tang, Y. Y. Zhang, J. Wang, W. Wang, D. Chialvo, Critical fluctuations in the native state of proteins, Phys. Rev. Lett. 118 (2017) 088102. doi:10.1103/PhysRevLett.118.088102.

[13] D. R. Chialvo, Life at the edge: complexity and criticality in biological function, Acta Physica Polonica B 49 (12) (2018) 1001–1025.

[14] R. D. Young, R. Scholl, Proteins as complex systems, Journal of Non-Crystalline Solids 131-133 (1991) 302–309, Proceedings of the International Discussion Meeting on Relaxations in Complex Systems. doi:https://doi.org/10.1016/0022-3093(91)90320-6.

[15] H. Frauenfelder, Proteins: Paradigms of complexity, Proceedings of the National Academy of Sciences 99 (suppl_1) (2002) 2479–2480. doi:https://doi.org/10.1073/pnas.01257999.

[16] M. A. Moret, G. F. Zebende, Amino acid hydrophobicity and accessible surface area, Phys. Rev. E 75 (2007) 011920. doi:10.1103/PhysRevE.75.011920.

[17] J. C. Phillips, Thermodynamic scaling of interfering hemoglobin strain field waves, The Journal of Physical Chemistry B 122 (40) (2018) 9324–9330. doi:10.1021/acs.jpcb.8b07550.

[18] J. C. Phillips, Scaling and self-organized criticality in proteins: Lysozyme c, Physical review. E, Statistical, nonlinear, and soft matter physics 80 (5) (2009) 051916. doi:10.1103/PhysRevE.80.051916.

[19] E. Faraggi, Y. Zhou, A. Kloczkowski, Accurate single-sequence prediction of solvent accessible surface area using local and global features, Proteins 82 (11) (2014) 3170–3176. doi:10.1002/prot.24682.

[20] L. J. Williams, B. J. Schendt, Z. R. Fritz, Y. Attali, R. H. Lavroff, M. L. Yarmush, A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme, Technology (Singap World Sci) 7 ((1-2)) (2019) 12–39. doi:10.1142/s233954781950002x.

[21] Q. Y. Tang, K. Kaneko, Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis, PLoS Comput Biol 16 (2) (2020) e1007670. doi:10.1371/journal.pcbi.1007670.

[22] J. C. Phillips, Synchronized attachment and the darwinian evolution of coronaviruses CoV-1 and CoV-2, Physica A: Statistical Mechanics and its Applications 581 (2021) 126202. doi:10.1016/j.physa.2021.126202.

[23] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1) (1982) 105–132. doi:https://doi.org/10.1016/0022-2836(82)90515-0.

[24] M. D. Esposti, M. Crimi, G. Venturoli, A critical evaluation of the hydropathy profile of membrane proteins, European journal of biochemistry 190 1 (1990) 207–219. doi:10.1111/J.1432-1033.1990.TB15566.X.

[25] J. S. Lolkema, D. J. Slotboom, Estimation of structural similarity of membrane proteins by hydropathy profile alignment, Molecular membrane biology 15 1 (1998) 33–42.

[26] L. Damodharan, V. Pattabhi, Hydropathy analysis to correlate structure and function of proteins, Biochemical and Biophysical Research Communications 323 (3) (2004) 996–1002. doi:https://doi.org/10.1016/j.bbrc.2004.08.186.

[27] S. R. Krystek Jr., W. J. Metzler, J. Novotny, Hydrophobicity profiles for protein sequence analysis, Current Protocols in Protein Science 00 (1) (1995) 2.2.1–2.2.13. doi:https://doi.org/10.1002/0471140864.ps0202s00.

[28] B. Hu, H. Guo, P. Zhou, Z. L. Shi, Characteristics of SARS-CoV-2 and COVID-19, Nature Reviews Microbiology 19 (2021) 141–154. doi:10.1038/s41579-020-00459-7.

[29] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. Liu, D. Wang, W. Xu, E. Holmes, G. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding., Lancet 395 (2020) 565–574. doi:10.1016/S0140-6736(20)30251-8.

[30] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, J. Sheng, L. Quan, Z. Xia, W. Tan, G. Cheng, T. Jiang, Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China, Cell Host & Microbe 27 (3) (2020) 325–328. doi:https://doi.org/10.1016/j.chom.2020.02.001.

[31] J. Damas, G. M. Hughes, K. C. Keough, C. A. Painter, N. S. Persky, M. Corbo, M. Hiller, K. P. Koepfli, A. R. Pfenning, H. Zhao, D. P. Genereux, R. Swofford, K. S. Pollard, O. A. Ryder, M. T. Nweeia, K. Lindblad-Toh, E. C. Teeling, E. K. Karlsson, H. A. Lewin, Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates, Proc Natl Acad Sci USA 117 (36) (2020) 22311–22322. doi:10.1073/pnas.2010146117.

[32] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, J. S. McLellan, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation, Science 367 (6483) (2020) 1260–1263. doi:10.1126/science.abb2507.

[33] Q. Wang, Y. Zhang, L. Wu, S. Niu, C. Song, Z. Zhang, G. Lu, C. Qiao, Y. Hu, K. Y. Yuen, Q. Wang, H. Zhou, J. Yan, J. Qi, Structural and functional basis of SARS-CoV-2 entry by using human ACE2., Cell 181 (4) (2020) 894–904. doi:10.1016/j.cell.2020.03.045.

[34] R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, D. Hardie, V. Hill, N. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph, L. Kooppile, S. L. K. Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj, B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula, K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. Mbulawa, A. Mendes, K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi, M. Motswaledi, T. Mphoyakgosi, N. Msomi, P. Mwangi, Y. Naidoo, N. Ntuli, M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. San, L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Subramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. V. Wyk, S. Weaver, C. Wibmer, E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zuze, D. Goedhals, W. Preiser, F. Treurnicht, M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 omicron variant in southern Africa., Nature 603 (7902) (2022) 679–686. doi:https://doi.org/10.1038/s41586-022-04411-y.

[35] T. M. Rocha Filho, M. A. Moret, C. C. Chow, J. C. Phillips, A. J. A. Cordeiro, F. A. Scorza, A. G. Almeida, J. F. F. Mendes, A data-driven model for COVID-19 pandemic - evolution of the attack rate and prognosis for Brazil, Chaos Solitons Fractals 152 (2021) 111359. doi:10.1016/j.chaos.2021.111359.

[36] T. Tada, H. Zhou, B. M. Dcosta, M. I. Samanovic, V. Chivukula, R. S. Herati, S. R. Hubbard, M. Mulligan, N. R. Landau, Increased resistance of sars-cov-2 omicron variant to neutralization by vaccine-elicited and therapeutic antibodies., EBioMedicine. 78 (103944) (2022) 1–11. doi:10.1016/j.ebiom.2022.103944.

[37] Y. Liu, J. Liu, K. S. Plante, J. A. Plante, X. Xie, X. Zhang, Z. Ku, Z. An, D. Scharton, C. Schindewolf, S. G. Widen, V. D. Menachery, P.-Y. Shi, S. C. Weaver, The N501Y spike substitution enhances SARS-CoV-2 infection and transmission, Nature 602 (2022) 294–299. doi:https://doi.org/10.1038/s41586-021-04245-0.

[38] L. F. Souza, T. M. Rocha Filho, M. A. Moret, Relating SARS-CoV-2 variants using cellular automata imaging, Scientific Reports 12 (2022) 10297. doi:10.1038/s41598-022-14404-6.

[39] E. Eythorsson, H. L. Runolfsdottir, R. Ingvarsson, M. I. Sigurdsson, R. Palsson, Rate of SARS-CoV-2 reinfection during an omicron wave in iceland, JAMA Netw Open. 5 (8) (2022) e2225320. doi::10.1001/jamanetworkopen.2022.25320.

[40] M. A. Muñoz, Colloquium: Criticality and dynamical scaling in living systems, Rev. Mod. Phys. 90 (2018) 031001. doi:10.1103/RevModPhys.90.031001.

[41] J. C. Phillips, Fractals and self-organized criticality in proteins, Physica A: Statistical Mechanics and its Applications 415 (2014) 440–448. doi:10.1016/j.physa.2014.08.034.

[42] J. C. Phillips, M. A. Moret, G. F. Zebende, C. C. Chow, Phase transitions may explain why SARS-CoV-2 spreads so fast and why new variants are spreading faster, Physica A 598 (2022) 127318. doi:10.1016/j.physa.2022.127318.

[43] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, L. Du, Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine, Cellular and Molecular Immunology 17 (2020) 613 – 620.

[44] NCBI, National Center for Biotechnology Information (2022).
URL `https://www.ncbi.nlm.nih.gov/genbank`

Capítulo Seis

Conclusões

Nesta tese mostrou-se que a modelagem usando autômatos celulares é uma maneira eficiente para analisar similaridades entres sequências de proteínas, pois essa modelagem representa a sequência da proteína por uma imagem que em uma primeira análise já pode ser usada para comparar as proteínas. E em uma segunda análise usando a distância de Hamming estacionária pode-se concluir se de fato as sequências comparadas são similares. Nossa metodologia avalia as similaridades medindo as dissimilaridades do autômato, ao contrário de outras técnicas que fazem uma comparação apenas visual das imagens de autômato celular.

Além disso, notamos que os autômatos celulares da proteína Spike do coronavírus SARS-CoV-2 apresentam o mesmo padrão em forma de V que as imagens do coronavírus SARS-CoV-1 (WANG et al., 2005). Esse padrão pode ser usado como assinatura para coronavírus do tipo SARS, visto que outros coronavírus não apresentam esse comportamento. Vale ressaltar que essa assinatura está relaciona a escolha da regra de evolução 184, que consegue diferenciar esses coronavírus como visto em (WANG et al., 2005). Essa regra guarda a informação evolutiva da sequência da proteína, nas evoluções dos autômato sendo possível utilizar a distância de Hamming estacionária como uma medida de similaridades entre as proteínas. Podemos notar que essa é uma características de algumas regras de autômatos celulares elementares.

A distância de Hamming estacionária como medida de similaridade é uma métrica eficiente para classificar os animais quanto a classe como visto na análise da beta-globina e transferrina. Podemos notar que essa metodologia agrupa proteínas homólogas diferentes como a tranferrina e a lactotransferrina, essas proteínas são membros de uma mesma família de proteínas mas as lactotransferrina estão presentes em mamíferos. Nosso método separou a transferrina dos peixes, anfíbios e mamíferos da lactotransferrina dos mamíferos indicando ser uma medida eficiente para análise filogenéticas de proteínas.

O outro método proposto nesta tese, a diferença do perfil de hidropatia mostra as regiões onde ocorreram mudanças hidrofóbicas ou hidrofílicas nas variantes dos SARS-CoV-2 em comparação com o vírus inicialmente encontrado em Wuhan. Esse método não é capaz de indicar quais regiões são internas ou externas à proteína enovelada. Mas as mudanças ocorridas no RBD da proteína Spike das variantes do coronavírus SARS-CoV-2, podem indicar uma menor eficiência da vacina para as variantes Omicron e explicar um aumento nos casos de reinfecção para essa variante. A semelhança entre os perfis de hidropatia de $\psi(R, 39)$ das variantes Alpha, Beta e Gamma, indica uma semelhança estrutural dessas

variantes no RBD. Sendo uma possível explicação para a dominação dessas variantes em seus locais de origem.

6.1 Conclusões

Ao longo desse trabalho usamos direta ou indiretamente várias técnicas de modelagem de sistemas complexos, por exemplo, a fractalidade na escala Moret-Zebende usada para construir o perfil de hidropatia, a criticalidade auto-organizada considerada para estudar a evolução darwiniana do SARS-CoV. Os efeitos de memória presentes nas sequências na análise de similaridade, muitas dessas sequências sofreram mutações que as diferenciam dos seus ancestrais, mas a informação da evolução não é perdida por completo. O uso do autômato celular para modelar a sequência, pois esse método de modelagem surgiu como uma alternativa a métodos tradicionais para modelar sistemas complexos. A soma das propriedades individuais dos aminoácidos que não podem descrever as estruturas, evolução, enovelamento, ou outras propriedades da proteína. Assim, todas essas características corroboram para concluirmos que as proteínas são sistemas complexos.

6.2 Contribuições

Esses resultados contribuem para que o autômatos celulares sejam usados em análises de similaridades de sequencia de proteínas contribuindo assim na análise filogenética de proteínas. Além disso, o método permite analisar as mudanças estruturais ocorridas entre proteínas Spike do coronavírus SARS-CoV-2 sem precisar de ferramentas que modelem estruturalmente a proteína, analisando apenas a diferença no perfil de hidropatia que foi calculada levando em consideração a sequência da proteína. Essas técnicas são alternativas simples e muito eficiente no estudo de proteínas

6.3 Limitações

O método de comparação de autômatos celulares usando distância de Hamming estacionária não pode ser utilizado em proteínas não homólogas, assim como em outros métodos tradicionais de distância filogenética. A análise de similaridade de proteínas que passaram por evolução convergente como por exemplo a lisozima, mostrou-se inconclusiva pois aproximavam espécies que não são agrupadas em métodos usuais de análise de similaridade de sequência.

6.4 Atividades Futuras de Pesquisa

Nos métodos propostos usamos regras de autômatos celulares já consolidadas e mostramos que essas regras guardam informações das sequências iniciais como um efeito de memória, em trabalhos futuros queremos identificar quais regras tem essa mesma característica. Além disso, vamos propor um código binário do aminoácidos com base na escala de (MORET; ZEBENDE, 2007). Queremos aplicar a distância de Hamming estacionária no estudo filogenético do vírus monkeypox.

Apêndice A

Produção Técnica e Científica

Total de Publicação: 1 artigo em revista (Quali Capes A1) / 1 trabalho apresentado em congresso

A.1 Artigo publicado em periódico

SOUZA, L. F.; ROCHA FILHO, T. M.; MORET, M. A. Relating SARS-CoV-2 variants using cellular automata imaging. Scientific Reports, v. 12, n. 10297, 2022.<https://doi.org/10.1038/s41598-022-14404-6>

A.2 Trabalho publicado em congresso

SOUZA, L. F.; GONÇALVES, M. M. Cellular automata and their applications in proteins. In: Anais do Encontro Nacional de Modelagem Computacional, Encontro de Ciência e Tecnologia de Materiais, Conferência Sul em Modelagem Computacional e Seminário e Workshop em Engenharia Oceânica. Anais.Pelotas(RS) UFPel / FURG / UNIPAMPA, 2022. Disponível em:

<https//www.even3.com.br/anais/enmcmcsulsemengo2022/533942-CELLULAR-AUTOMATA-AND-THEIR-APPLICATIONS-IN-PROTEINS>.

Referências Bibliográficas

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, v. 215, n. 3, p. 403–410, 1990. 2.2

ANFINSEN, C. B. Principles that govern the folding of protein chains. *Science*, v. 181, n. 4096, p. 223–230, 1973. 2.1, 2.4

BONCHEV, D.; THOMAS, S.; APTE, A.; KIER, L. Cellular automata modelling of biomolecular networks dynamics. *SAR and QSAR in environmental research*, v. 21, p. 77–102, 2010. 2.3

BRILKOVA, M. et al. Error-prone protein synthesis recapitulates early symptoms of alzheimer disease in aging mice. *Cell Reports*, v. 40, n. 13, p. 111433, 2022. 2.1

CALDART, E. T.; MATA, H.; CANAL, C. W.; P, R. A. Análise filogenética: conceitos básicos e suas utilizações como ferramenta para virologia e epidemiologia molecular. *Acta Scientiae Veterinariae*, v. 44, n. 1392, p. 1–20, 2016. 2.2

CAMPANELLA, J. J.; BITINCKA, L.; SMALLEY, J. Matgat: An application that generates similarity/identity matrices using protein or dna sequences. *BMC Bioinformatics*, v. 4, n. 29, 2003. 1

CHAO, J.; TANG, F.; XU, L. Developments in algorithms for sequence alignment: A review. *Biomolecules*, v. 12, n. 4, p. 546, 2022. 2.2

CHAUDHURI, P. P.; GHOSH, S.; DUTTA, A.; CHOUDHURY, S. P. *A New Kind of Computational Biology*. Singapore: Springer, 2018. (document), 1, 2.3, 2.3, 2.1

CHAUDHURI, T. K.; PAUL, S. Protein-misfolding diseases and chaperone-based therapeutic approaches. *The FEBS Journal*, v. 273, n. 7, p. 1331–1349, 2006. 2.1

CLEMENTS, J. D.; MARTIN, R. E. Identification of novel membrane proteins by searching for patterns in hydropathy profiles. *European Journal of Biochemistry*, v. 269, n. 8, p. 2101–2107, 2002. 2.4

CONTESSOTO, V. G.; JUNIOR, A. B. O.; J., C.; OLIVEIRA, R. J.; LEITE, V. B. P. Introdução ao problema de enovelamento de proteínas: uma abordagem utilizando modelos computacionais simplificados. *Rev Bras Ensino Física [Internet]*, v. 40, n. 4, 2018. 2.1

DAMODHARAN, L.; PATTABHI, V. Hydropathy analysis to correlate structure and function of proteins. *Biochemical and Biophysical Research Communications*, v. 323, n. 3, p. 996–1002, 2004. 1

ELLIS, R. J.; PINHEIRO, T. J. T. Medicine:danger-misfolding proteins. *Nature*, v. 416, n. 6880, p. 483–484, 2002. 2.1

ESPOSTI, M. D.; CRIMI, M.; VENTUROLI, G. A critical evaluation of the hydropathy profile of membrane proteins. *European journal of biochemistry*, v. 190, n. 1, p. 207–219, 1990. 1, 2.4

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. *Cluster Analysis*. London, 5th Edition: John Wiley & Sons, 2011. 2.2

FAITH, D. P. Distance methods and the approximation of most-parsimonious trees. *Systematic Biology*, v. 34, n. 3, p. 312–325, 1985. 2.2

FARRIS, J. S. Estimating phylogenetic trees from distance matrices. *The American Naturalist*, v. 106, n. 951, p. 645–668, 1972. 2.2

FELSENSTEIN, J. [24] inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. In: *Computer Methods for Macromolecular Sequence Analysis*. [S.l.]: Academic Press, 1996. v. 266, p. 418–427. 2.2

FITCH, W. M. Distinguishing homologous from analogous proteins. *Systematic Zoology*, v. 19, n. 2, p. 99–113, 1970. 2.2

GARDNER, M. Mathematical games: The fantastic combinations of john conway's new solitaire game "life". *Scientific American*, v. 223, n. 4, p. 120–123, 1970. 2.3

GAUTHIER, J.; VINCENT, A. T.; CHARETTE, S. J.; DEROME, N. A brief history of bioinformatics. *Briefings in bioinformatics*, v. 20, n. 6, p. 1981–1996, 2019. 2.1

GENBANK. *National Center for Biotechnology Information*. 2022. Https://www.ncbi.nlm.nih.gov/genbank. 1

HU, G.; KURGAN, L. Sequence similarity searching. *Current Protocols in Protein Science*, v. 95, n. 1, p. e71, 2018. 1

HUANG, G.; HU, J. Similarity/dissimilarity analysis of protein sequences by a new graphical representation. *Current Bioinformatics*, v. 8, n. 5, p. 539–544, 2013. 2.2

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, v. 30, n. 4, p. 772–780, 2013. 2.2

KAUZMANN, W. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, v. 14, p. 1–63, 1959. 2.4

KAVIANPOUR, H.; VASIGHI, M. Structural classification of proteins using texture descriptors extracted from the cellular automata image. *Amino Acids*, v. 49, n. 2, p. 261–271, 2017. 2.3, 2.3, 2.3

KESSEL, A.; BEN-TAL, N. *Introduction to Proteins: Structure, Function, and Motion*. New York, 2 edition: Chapman & Hall/CRC Mathematical and Computational Biology, 2018. 1

KRYSTEK, S. R.; METZLER, W. J.; NOVOTNY, J. Hydrophobicity profiles for protein sequence analysis. *Current Protocols in Protein Science*, v. 00, n. 1, p. 2.2.1–2.2.13, 1995. 1, 2.4

KUHNER, M. K.; FELSENSTEIN, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, v. 11, n. 3, p. 459–468, 1994. 2.2

KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, v. 157, n. 1, p. 105–132, 1982. 1, 2.4

LIAO, B.; LIAO, B.; SUN, X.; ZENG, Q. A novel method for similarity analysis and protein sub-cellular localization prediction. *Bioinformatics*, v. 26, n. 21, p. 2678–2683, 2010. 1, 2.2

LIPMAN, D. J.; PEARSON, W. R. Rapid and sensitive protein similarity searches. *Science*, v. 227, n. 4693, p. 1435–1441, 1985. 1

LIU, N.; WANG, T. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. *FEBS Letters*, v. 580, n. 22, p. 5321–5327, 2006. 2.2

LOLKEMA, J. S.; SLOTBOOM, D.-J. Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins. *FEMS Microbiology Reviews*, v. 22, n. 4, p. 305–322, 1998. 1, 2.4

MORET, M. Self-organized critical model for protein folding. *Physica A: Statistical Mechanics and its Applications*, v. 390, n. 17, p. 3055–3059, 2011. 2.1

MORET, M. A.; ZEBENDE, G. F. Amino acid hydrophobicity and accessible surface area. *Phys. Rev. E*, v. 75, n. 1, p. 011920, 2007. 2.1, 2.4, 2.4, 6.4

MU, Z.; WU, J.; ZHANG, Y. A novel method for similarity/dissimilarity analysis of protein sequences. *Physica A: Statistical Mechanics and its Applications*, v. 392, n. 24, p. 6361–6366, 2013. 1, 2.2

MU, Z. et al. FEGS: a novel feature extraction model for protein sequences and its applications. *BMC Bioinformatics*, v. 22, n. 297, 2021. 1, 2.2

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, v. 48, n. 3, p. 443–453, 1970. 1

NEUMANN, J. V. The general and logical theory of automata. In: *In J. von Neumann Collected Works*. [S.l.]: Q. H. Taub, 1963. v. 5, p. 288–328. 2.3

NUSSENZVEIG, H. M. *Complexidade E Caos*. Rio de Janeiro, 3 ed.: Ed. UFRJ/Copea, 2008. 2.1, 2.1

OTU, H. H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, v. 19, n. 16, p. 2122–2130, 2003. 2.2

PACE, C. N.; SHIRLEY, B. A.; MCNUTT, M.; GAJIWALA, K. Forces contributing to the conformational stability of proteins. *The FASEB Journal*, v. 10, n. 1, p. 75–83, 1996. 2.1

PEARSON, W. R. An introduction to sequence similarity (homology) searching. *Current Protocols in Bioinformatics*, v. 42, n. 1, p. 3.1.1–3.1.8, 2013. 1, 2.2

PHILLIPS, J. C. Scaling and self-organized criticality in proteins i. *Proceedings of the National Academy of Sciences*, v. 106, n. 9, p. 3107–3112, 2009. 1, 2.1

PHILLIPS, J. C. Scaling and self-organized criticality in proteins ii. *Proceedings of the National Academy of Sciences*, v. 106, n. 9, p. 3113–3118, 2009. 1

PHILLIPS, J. C. Synchronized attachment and the darwinian evolution of coronaviruses cov-1 and cov-2. *Physica A: Statistical Mechanics and its Applications*, v. 581, n. 1, p. 126202, 2021. 1, 2.4

PHILLIPS, J. C.; MORET, M. A.; ZEBENDE, G. F.; CHOW, C. C. Phase transitions may explain why sars-cov-2 spreads so fast and why new variants are spreading faster. *Physica A*, v. 598, n. 1, p. 127318, 2022. 1, 2.4

RAHMAN, M. M.; BISWAS, B. A.; BHUIYAN, M. I. H. Protein similarity analysis by wavelet decomposition of cellular automata images. In: *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. Bangladesh: IEEE, 2019. p. 1–6. 1, 2.2, 2.3

SAHELY, B. *Visualizing Conway's Game of Life.* 2022. Http://demonstrations.wolfram.com/VisualizingConwaysGameOfLife/. 2.4

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, v. 4, n. 4, p. 406–425, 1987. 2.2

SANGER, F.; THOMPSON, E. O.; KITAI, R. The amide groups of insulin. *The Biochemical journal*, v. 3, n. 59, p. 509–518, 1955. 1

SANTOS, J.; VILLOT, P.; DIéGUEZ, M. Protein folding with cellular automata in the 3d hp model. In: *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*. New York, NY, USA: Association for Computing Machinery, 2013. p. 1595–1602. 2.3

SANTOS, J.; VILLOT, P.; DIéGUEZ, M. Cellular automata for modeling protein folding using the hp model. In: *2013 IEEE Congress on Evolutionary Computation*. Cancun, Mexico: IEEE, 2013. p. 1586–1593. 2.3

SARKAR, P. A brief history of cellular automata. Association for Computing Machinery, New York, NY, USA, v. 32, n. 1, p. 80–107, March 2000. 2.3, 2.3

TATENO, Y.; NEI, M.; TAJIMA, F. Accuracy of estimated phylogenetic trees from molecular data. i. distantly related species. *J Mol Evol.*, v. 18, n. 6, p. 387–404, 1982. 2.2

THOMPSON, J. D.; GIBSON, T. J.; HIGGINS, D. G. Multiple sequence alignment using clustalw and clustalx. *Current Protocols in Bioinformatics*, v. 00, n. 1, p. 2.3.1–2.3.22, 2003. 1.4, 2.2

UNIPROT. *The Universal Protein Resource*. 2022. Https://www.uniprot.org. 1, 1.1

VARELA, D.; SANTOS, J. Protein folding modeling with neural cellular automata using rosetta. In: *GECCO '16 Companion: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. New York, NY, USA: Association for Computing Machinery, 2016. p. 1307–1312. 2.3

VARELA, D.; SANTOS, J. Evolving cellular automata schemes for protein folding modeling using the rosetta atomic representation. *Genetic Programming and Evolvable Machines*, v. 23, n. 2, p. 225–252, 2022. 2.3

WANG, M. et al. A new nucleotide-composition based fingerprint of sars-cov with visualization analysis. *Medicinal Chemistry*, v. 1, n. 1, p. 39–47, 2005. 2.3, 6

WOLFRAM, S. Cellular automata as models of complexity. *Nature*, v. 311, n. 0, p. 419–424, 1984. 2.3

WOLFRAM, S. *A New Kind of Science.* [S.l.]: Wolfram Media, 2002. 2.3, 2.3

WOLFRAM, S. *3D Totalistic Cellular Automata.* 2022. Http://demonstrations.wolfram.com/3DTotalisticCellularAutomata/. 2.4

WU, Z.; XIAO, X.; CHOU, K. 2d-mh: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J Theor Biol.*, v. 267, n. 1, p. 29–34, 2010. 1, 2.2

XIAO, X.; CHOU, K. Digital coding of amino acids based on hydrophobic index. *Protein and Peptide Letters*, v. 14, n. 9, p. 871–875, 2007. 2.3, 2.3

XIAO, X.; SHAO, S.; DING, Y.; CHEN, X. Digital coding for amino acid based on cellular automata. In: *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. The Hague: IEEE, 2004. v. 5, p. 4593–4598. 1, 2.3, 2.3, 2.3

XIAO, X. et al. Using cellular automata to generate image representation for biological sequences. *Amino Acids*, v. 28, n. 1, p. 29–35, 2005. 1, 2.3, 2.3

XIAO, X.; WANG, P.; CHOU, K. C. Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image. *Journal of Theoretical Biology*, v. 254, n. 3, p. 691–696, 2008. 1

YAO, Y. et al. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. *Evolutionary bioinformatics online*, v. 10, n. 0, p. 87–96, 2014. 2.2

YAO, Y.-H. et al. Analysis of similarity/dissimilarity of protein sequences. *Proteins: Structure, Function, and Bioinformatics*, v. 73, n. 4, p. 864–871, 2008. 1, 2.2

ZHANG, W.; SUN, Z. Random local neighbor joining: A new method for reconstructing phylogenetic trees. *Molecular Phylogenetics and Evolution*, v. 47, n. 1, p. 117–128, 2008. 2.2

ZIELEZINSKI, A.; VINGA, S.; ALMEIDA, J.; KARLOWSKI, W. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, v. 18, n. 1, p. 186, 2017. 2.2

Modelagem computacional aplicada a análise de similaridade e filogenia das proteínas.

Luryane Ferreira de Souza

Salvador, Salvador
2023.