Amanda Tábita da Silva Albanaz

Entendendo os Mecanismos Moleculares de Mutações que causam Esclerose Lateral Amiotrófica

Universidade Federal de Minas Gerais
Belo Horizonte
Fevereiro de 2019

Amanda Tábita da Silva Albanaz

Entendendo os Mecanismos Moleculares de Mutações que causam Esclerose Lateral Amiotrófica

Dissertação apresentada ao Programa Interunidades de Pós-graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para obtenção do título de Mestre em Bioinformática.

Orientador: Douglas Eduardo Valente Pires
Co-orientador: David Benjamin Ascher

Programa Interunidades de Pós-Graduação em Bioinformática
Universidade Federal de Minas Gerais - UFMG
Instituto de Ciências Biológicas
Belo Horizonte, Fevereiros de 2019

Agradecimentos

Gostaria de expressar minha gratidão aos meus orientadores. Ao Dr. Douglas Pires, por todo o apoio, incentivo e imensa compreensão, durante cada etapa do meu desenvolvimento acadêmico sob sua orientação. Ao Dr. David Ascher, que apesar da grande distância tem sido indispensável ao meu desenvolvimento e sempre se lembra de detalhes e materiais importantes. Obrigada à ambos pela oportunidade de aprender com vocês.

Agradeço também ao Instituto de Ciências Biológicas da UFMG, à todos os professores e equipes de suporte acadêmico, à pós-graduação em Bioinformática, imprescindíveis à formação acadêmica.

Ao Instituto René Rachou – Fiocruz Minas, obrigada pela oportunidade de aprendizado e crescimento. Sou grata aos colegas da Plataforma de Bioinformática, sem exceção àqueles que buscaram novos desafios em outras instituições e países. Agradeço à todos e em especial à Joicy, por todas as contribuições e suporte durante essa jornada.

Minha eterna gratidão aos meus pais, Vander e Marlene, à minha irmã Júlia e ao meu companheiro, João. Agradeço a imensa compreensão em todos os momentos, o suporte e incentivo imensuráveis. Vocês são parte do que sou. Agradeço também à minha bicharada que me proporcionou momentos insubstituíveis de desestresse e carinho.

Meus agradecimentos às agências de fomento, Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo suporte financeiro ao desenvolvimento deste trabalho.

"Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the Universe exist. Be curious. And however difficult life may seem, there is always something you can do and succeed at. It matters that you don't just give up."

Stephen Hawking.

RESUMO

A Esclerose lateral amiotrófica (do inglês, Amyotrophic Lateral Sclerosis - ALS) é uma doença rara caracterizada pela forma idade-dependente e rápida progressão degenerativa de neurônios motores superiores e inferiores, causando paralisia e óbito dentro de 2 a 4 anos. É causada principalmente por mutações *missense* que afetam a estrutura e funcionalidade protéicas. Neste contexto as proteínas SOD1, TDP-43 e FUS/TLS possuem particular importância para o mecanismo molecular da doença. Mutações nessas proteínas podem afetar a estrutura e função dessas proteínas e causar graves fenótipos da doença, muito embora seus mecanismos moleculares ainda não sejam bem compreendidos. Já foi reportado na literatura que a estabilidade estrutural é um fator importante para o mecanismo molecular da ALS, principalmente para a SOD1, entretanto foi hipotetizado que este apenas descreve parcialmente o repertório de efeitos moleculares ligados a esses fenótipos. Nesse sentido, no presente trabalho investigou-se propriedades moleculares que sejam relevantes para melhor entender a relação genótipo-fenótipo e os mecanismos moleculares ligados à ALS. Para tal, foi desenvolvida uma nova e ampliada base de dados relacional descrevendo mutações em ALS e dados clínicos de pacientes, a DynAMISM, que conta com dados estruturais e de sequência, predições do efeito de mutações e, no futuro, mecanismos moleculares putativos, à luz das estruturas das proteínas ou de seus modelos computacionais. A base inclui um aumento de 56,6% em casos clínicos em comparação à uma base de dados já estabelecida. A partir da análise de propriedades de mutações em SOD1, identificou-se uma forte correlação entre propriedades como flexibilidade e efeitos em interações proteína-proteína e dados clínicos de pacientes, o que pode indicar novos potenciais mecanismos moleculares relacionados a diferentes fenótipos. O poder preditivo desses dados foi, então, avaliado através se sua utilização em uma árvore de regressão e uma correlação de Pearson de até 0,7 foi obtida. Em trabalhos futuros é pretendida a disponibilização da base de dados em uma interface *web* amigável, validar os modelos preditivos por meios de testes cegos, refiná-los e estendê-los à outras proteínas. Entender o mecanismo molecular da ALS é um passo importante para auxiliar o manejo de pacientes bem como o desenvolvimento de tratamentos mais personalizados e mais eficazes e possívelmente este trabalho inicia um caminho para tal.

Palavras-chave: ALS, mutações *missense*, mecanismo molecular, fenótipo.

ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is an age-dependent rare disease, characterized by neurodegenerative effects on motor activity, causing paralysis and death within two to four years. It is mainly caused by *missense* mutation affecting protein structure and function and SOD1, TDP-43 and FUS/TLS have particular importance for the disease molecular mechanism. Mutations in these proteins can disturb protein structure and function, leading to severe phenotypes, even though their main molecular mechanisms remain unclear. It has been previously reported that protein stability plays an important role in the pathogenicity mechanism of the disease, especially for SOD1, however we hypothesize this effect only partially describes the repertoire of molecular effects leading to different phenotypes. In this work, we investigate mutation properties and their predicted effects in an effort to better understand molecular mechanisms of mutations in ALS. To achieve this, we developed a new and expanded relational database describing ALS mutations and patient clinical data, DynAMISM, which also encompasses structural and sequence features describing these mutations and, in the future, assignments of putative structure-based molecular mechanisms. The database represents an increase in 56.6% of total clinical cases in comparison with a well established database. By analyzing SOD1 mutations, we have identified a strong correlation between molecular properties such as flexibility and effects in protein-protein affinity with clinical outcomes, which might indicate new potential mutation molecular mechanisms. The predictive power of these findings were, then, assessed using a regression tree and a Pearson's correlation of up to 0.7 was achieved. As future work, we intend to made the database available as a user-friendly web interface, validate the predictive models using blind tests and extend them to other proteins. Understanding the molecular mechanisms of pathogenicity in ALS is an important step to guide better patient management and the development of more effective and personalized treatments, which we believe this work contributes to.

Key-words: ALS, *missense* mutation, molecular mechanisms, phenotype

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

3D – Tridimensional

ALS – *Amyotrophic Lateral Sclerosis*

ALSoD – *Amyotrophic Lateral Sclerosis Online Genetics Database*

CryoEM – Crio-Microscopia Eletrônica (do inglês, *Cryogenic electron microscopy*)

DNA – *Dexoxyribonucleic Acid*

DOPE - *Discrete Optimized Protein Energy*

DynAMISM – *Database of ALS Missense Mutations*

FALS – *Familial Amyotrophic Lateral Sclerosis*

FTD – Demência frontotemporal

FUS/TLS – *Fused in sarcoma/translocated in liposarcoma*

GnomAD – *Genome Aggregation Database*

HMMER – *Hidden Markov Models*

ID – Identificador

IUPAC – *International Union of Pure and Applied Chemistry*

KIF5A – *Kinesin Family Member* 5A

mCSM – *Mutation Cut-off Score Matrix*

mRNA – RNA mensageiro

MTR – *Missense Tolerance Ratio*

NMR – Ressonância Magnética Nuclear (do inglês, *Nuclear magnetic resonance*)

NLS – *Nuclear Localization Signal*

PDB – *Protein Data Bank*

pph2 – Polyphen 2

PPI – Interação proteína-proteína (do inglês, *Protein-protein interaction*)

PROVEAN – *Protein Variation Effect Analyzer*

Q-Q plot – gráfico quantil-quantil

RNA – *Ribonucleic Acid*

RRM – *RNA Recognition Motif*

RSA - Acessibilidade Relativa ao Solvente (do inglês, *Relative Solvent Accessibility*)

SALS – *Sporadic Amyotrophic Lateral Sclerosis*

SIFT – *Sorting Intolerant from Tolerant*

SOD1 – Superoxide dismutase 1

TDP-43 – *Transactive DNA-binding protein 43*

UniProt – *Universal Protein Resource*

WT – Selvagem (do inglês, *Wild Type*)

SUMÁRIO

1 INTRODUÇÃO

Apesar do avanço das tecnologias de estudo genético, os mecanismos moleculares relacionados às doenças não podem ser explicados apenas pela simples relação genótipo-fenótipo. Os mecanismos causadores das doenças resultam de perturbações mais complexas, envolvendo sistemas celulares e redes moleculares. Sendo assim, a cada dia tem se notado o desenvolvimento de diversas abordagens computacionais que podem ser utilizadas em conjunto para melhor compreender a ampla rede de relações por trás da associação genótipo-fenótipo. Muitas dessas abordagens computacionais se baseiam variadas metodologias de análises de proteínas, utilizando ou possibilitando também a integração de dados genômicos em larga escala e fenótipos de doenças. Assim se torna cada vez mais viável e acurada a predição de fenótipos de doenças e sua relação com os diversos mecanismos moleculares (WANG; GULBAHCE; YU, 2011).

1.1 A Esclerose Lateral Amiotrófica (ALS)

Neste contexto, a ALS é a doença humana, uma desordem neurodegenerativa progressiva, utilizada como estudo neste presente trabalho. Buscou-se compreender os mecanismos moleculares relacionados à fenótipos da doença, utilizando diferentes abordagens computacionais e metodologia de aprendizado de máquina para a predição dos fenótipos. Como será visto na seção de metodologia.

1.1.1 Contexto Histórico

Descoberta pela primeira vez em 1848 por Aran e melhor descrita mais tarde, em 1869, pelo neurologista francês Jean-Martin Charcot, a esclerose lateral amiotrófica (do inglês, *Amyotrophic Lateral Sclerosis*, ALS) é uma desordem neurodegenerativa progressiva e fatal, surgindo geralmente no início da vida adulta e com uma complexa base genética (FOROSTYAK; SYKOVA, 2017); MCLAUGHLIN et al., 2017; ROWLAND; SHNEIDER, 2001). A ALS é considerada uma doença rara, caracterizada principalmente pela rápida progressão degenerativa de neurônios motores superiores e inferiores, levando à paralisia e óbito comumente por falha respiratória, dentro de 2 a 4 anos desde o aparecimento dos sintomas (CHIÒ et al., 2013; DEL AGUILA et al., 2003; ROWLAND; SHNEIDER, 2001).

Stephen Hawking, indiscutivelmente um dos maiores cientistas de todos os tempos, autor de vários *best-sellers* e estudioso de buracos negros foi diagnosticado com ALS em 1963. A luta do cientista contra a doença foi árdua e peculiarmente longa. Somente 10% dos pacientes com

ALS sobrevivem mais de 10 anos. Hawking, entretanto, teve uma sobrevida de 55 anos, se tornando uma das maiores inspirações para muitas pessoas, em especial para os portadores de ALS.

Hawking e várias personalidades lutaram para promover a pesquisa e conscientização sobre ALS. Dentre vários movimentos nesse sentido, o Desafio do Balde de Gelo (do inglês, *Ice Bucket Challenge*), ganhou destaque em 2014. Nessa aparente brincadeira, que ficou popular nas mídias sociais, uma pessoa desafia outra a divulgar, em redes sociais, um vídeo em que o desafiado derrama sobre sua cabeça um balde de água com gelo. A pessoa nomeada pode aceitar o desafio ou fazer uma doação à alguma instituição de caridade à ALS, ou optar por ambos. Essa campanha alcançou tanto êxito que o dinheiro arrecadado possibilitou o avanço de alguns projetos[1] que buscam tratamento e cura para a ALS, entre eles estão: ALS *Accelerated Therapeutics*, *The New York Genome Center*, *Neuro Collaborative* e o *Project MinE*. O financiamento possibilitou também a descoberta de dois genes associados à ALS, o NEK1 (MAKAR et al., 1975) e KIF5A (NICOLAS et al., 2018)[2].

1.1.2 Incidência Global da ALS

Estima-se que globalmente cerca de 450.000 pessoas vivam com ALS, com incidência (taxa de manifestação de uma determinada doença) global de 1,9:100.000 habitantes (CHIÒ et al., 2013) e prevalência média (número de casos de uma doença em uma população, durante um período específico de tempo) de 5,2:100.000 (WIJESEKERA; LEIGH, 2009).

A incidência e prevalência globais desta doença possuem ampla variabilidade, tanto em termos de tempo de sobrevivência quanto das idades de início dos portadores da doença. Isso se deve, provavelmente à fatores como variabilidade no tamanho das populações pesquisadas, inclusão ou exclusão de variações fenotípicas, subestimação de casos devido à diferença de acesso aos sistemas de saúde e variação na qualidade de recursos médicos adequados entre os países. Nesse mesmo contexto é necessário considerar ainda o abismo socioeconômico e social ainda significativo, fatores genéticos, populações ancestrais, ambiente, estilo de vida e variações raciais (ARTHUR et al., 2016; CHIÒ et al., 2013).

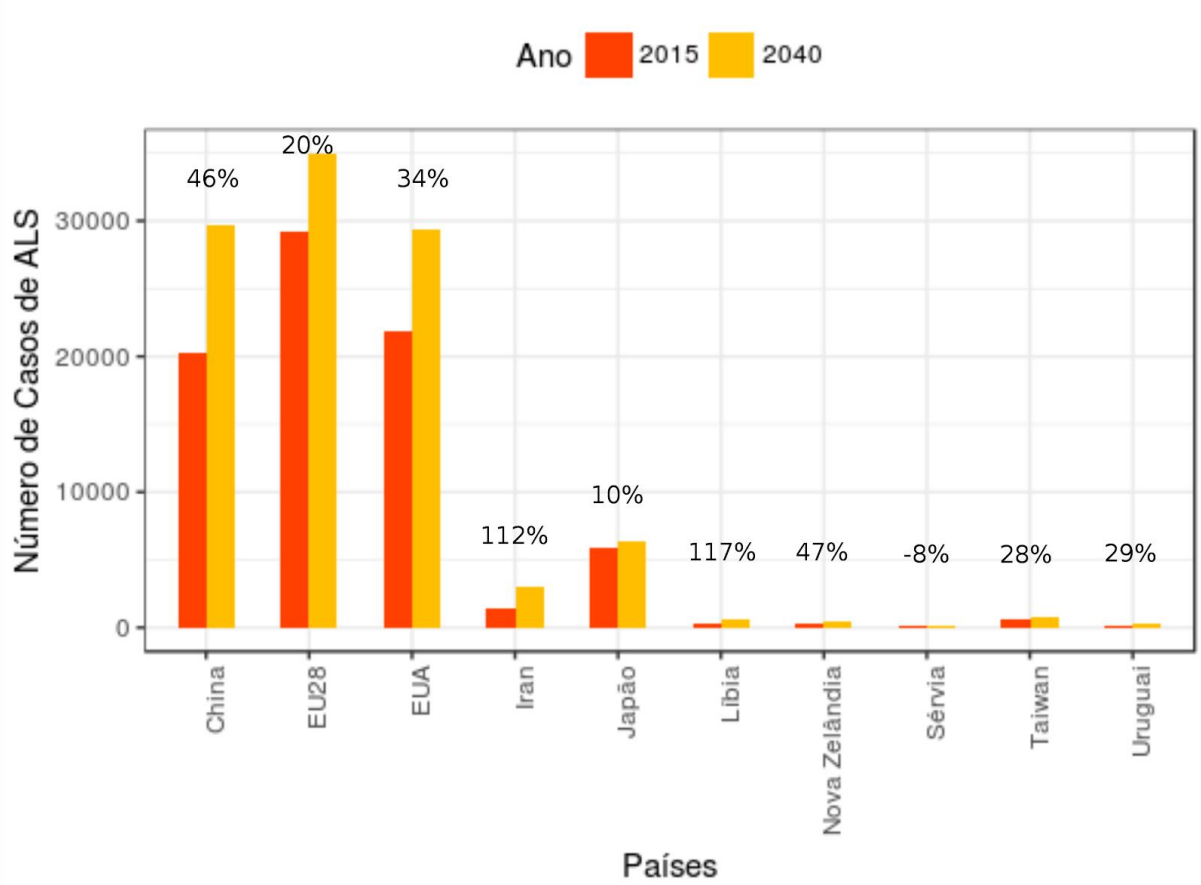
Apesar do mencionado anteriormente, não se considera que as variações na prevalência da ALS sejam devido à etnicidade e variáveis demográficas. Considerando relações com outros fatores

[1] http://www.alsa.org/news/media/press-releases/ibc-initial-commitment.html
[2] http://www.alsa.org/news/media/press-releases/significant-gene-discovery-072516.html

mencionados anteriormente, é possível citar diferenças de critério de diagnóstico e o momento na história da doença no paciente em que esse diagnóstico é feito, além da ausência de um padrão de avaliação (BEGHI et al., 2011; LOGROSCINO et al., 2008). Em contrapartida, há uma classificação específica à uma região, a ALS Guamaniana, que se deve à alta prevalência observada na isolada ilha de Guam e nos Territórios do Pacífico, na década de 1950. Esta classificação, geograficamente baseada, possui etiologia desconhecida e seu diagnóstico diferencial inclui tumores cervicais e/ou no forame magno (BEGHI et al., 2011; LOGROSCINO et al., 2008).

Estudos retrospectivos e prospectivos demonstram, em geral, aumento da incidência e prevalência da ALS, o que pode ser resultado da limitação de processos de coleta de dados adequados em tempos de pesquisa retrospectiva (BEGHI et al., 2006; LOGROSCINO et al., 2008). A tendência desse aumento se concentra principalmente na terceira idade e em nações em desenvolvimento (Figura 1), já que essas nações tendem a um aumento na expectativa de vida populacional e consequentemente, da população idosa (ARTHUR et al., 2016).

Figura 1 - Estimativa de Aumento de Casos de ALS do ano de 2015 a 2040.



As barras mostram dados referentes à indivíduos do gênero masculino e feminino portadores da ALS e a estimativa do aumento de casos reportados da doença entre os anos de 2015 e 2040. A barra de

cor laranja, referente ao ano de 2015, denota todos os casos reportados até o dado ano. Dados extraídos e modificados de ARTHUR et al., 2016.

A Figura 1 mostra no ano de 2015 aproximadamente 80.000 casos ao todo (46.000 homens e 34.000 mulheres). A projeção para casos da doença até 2040 é superior a 105.000 casos (60.000 homens e 45.000 mulheres), percentual de aumento superior a 31% em um intervalo de 25 anos (ARTHUR et al., 2016).

1.1.3 Incidência da ALS por Idade

A ALS possui uma curva de incidência de idade similar a outras doenças neurodegenerativas como Parkinson e Alzheimer (LOGROSCINO et al., 2015). Sua ocorrência é rara abaixo dos 30 anos de idade, aumenta aos 40 e apresenta um pico de incidência entre os 70 e 80 anos, e então reduz novamente, primeiramente no gênero masculino e depois no feminino (LOGROSCINO et al., 2015; MARIN et al., 2018).

O rápido declínio de incidência em pessoas com mais de 75 anos pode estar relacionado à dificuldade no diagnóstico de ALS nessa faixa etária devido à comorbidade, menor acesso a cuidados especializados ou ainda à quadros de ALS agressivos e rápidos que levam à morte antes do diagnóstico. Outra possível explicação para essa observação é que após os 75 anos os indivíduos suscetíveis ao desenvolvimento da ALS já teriam em sua maior parte sucumbido à doença, remanescendo os não suscetíveis (LOGROSCINO et al., 2008). A análise da correlação entre número de casos e idade é de suma importância, uma vez que pode dar suporte à descoberta das causas da patogenia da ALS (MARIN et al., 2018).

1.1.4 Anatomia da ALS

A ALS comumente afeta os neurônios motores superiores no córtex motor e os neurônios motores inferiores no tronco cerebral e medula espinhal. Adicionalmente, cerca de 50% dos casos de ALS caracterizados por variações no espectro fenotípico, podendo incluir outras síndromes, podem afetar sistemas extra motores, como os circuitos temporal, comportamental e frontal (considerado o centro executivo do cérebro). Dessa forma, a ALS pode ser considerada uma desordem multissistêmica, onde há uma conexão entre neurônios motores e outras células neurais, como astrócitos, oligodendrócitos e células da micróglia (MCLAUGHLIN et al., 2017).

A característica progressiva dessa doença neurodegenerativa leva os neurônios motores à morte. Consequentemente, o cérebro perde a capacidade de iniciar e controlar movimentos musculares,

levando à paralisia em pacientes em estágio avançado da doença[3]. Essa característica também acarreta consequências para as funções cognitivas e executivas, podendo levar à outras doenças e síndromes, como problemas cognitivos e demência frontotemporal (FTD)[4]. Esta última, de ocorrência estimada em 20% dos casos de ALS, afeta o prognóstico dos pacientes, podendo tornar mais difícil visualizar as possibilidades terapêuticas e a evolução da doença (PHUKAN et al., 2012). Outros espectros de fenótipos da ALS podem incluir fenótipos de psicose, degeneração corticobasal, comportamento suicida e parkinsonismo, sendo que em casos familiares pode ocorrer o raro complexo ALS-parkinsonismo-demência (SILVERMAN et al., 2016).

1.1.5 Progressão, Sintomas e Diagnóstico da ALS

O início e severidade da ALS são dependentes da idade. A ALS, muito embora rara em âmbito geral, é considerada a doença neurodegenerativa mais comum em pessoas entre a idade adulta e a terceira idade, predominantemente entre 40 e 70 anos de idade, considerando os 60 anos a média de início da desordem[5].

A heterogeneidade clínica dificulta o diagnóstico (ANDERSEN, 2006), mas alguns fatores de risco para a doença incluem idade acima de 40 anos, histórico familiar de ALS e ser do gênero masculino (WROE et al., 2008).

A morte dos neurônios motores causa fraqueza e atrofia muscular observadas nos portadores da ALS. A fraqueza se inicia de forma focalizada e se espalha através do sistema nervoso, levando à paralisia e morte geralmente dentro de 2 a 3 anos nos casos da ALS de início bulbar e entre 3 a 5 anos quando tem início espinhal. Isso se deve principalmente ao fato de que a habilidade de controlar os movimentos é comprometida, em especial habilidades para comer e respirar (MCLAUGHLIN et al., 2017; PHUKAN; PENDER; HARDIMAN, WIJESEKERA; LEIGH, 2007).

Existem duas grandes classificações para a ALS, podendo ser de ocorrência esporádica (SALS - do inglês, *Sporadic* ALS) ou familiar (FALS - do inglês, *Familial* ALS). Apesar da SALS ser a forma mais comum da doença, representando 90% dos casos de ALS, seu mecanismo de patogênese e base genética são menos conhecidos em relação à FALS. Enquanto os casos familiares, caracterizados por histórico familiar com dois ou mais familiares com a doença, são

[3] http://www.alsa.org/about-als/what-is-als.html?referrer=http://www.alsa.org/about-als/
[4] https://ghr.nlm.nih.gov/condition/amyotrophic-lateral-sclerosis
[5] https://ghr.nlm.nih.gov/condition/amyotrophic-lateral-sclerosis

mais raros, representando cerca de 10% dos casos. Na FALS existe uma chance de 50% de cada filho herdar uma mutação genética e desenvolver a doença, onde aproximadamente 60% dos indivíduos com FALS possuem pelo menos uma mutação genética (GOMES et al., 2008).

SALS e FALS apresentam semelhanças nos aspectos clínicos, o que pode significar que ambos convergem para o mesmo caminho e envolvem fatores comuns (ROTUNNO; BOSCO, 2013). Casos esporádicos podem surgir de fatores genéticos, bem como de fatores ambientais e comportamentais, como tabagismo, exercícios físicos em excesso, lesões e exposição a toxinas ambientais capazes de comprometer neurônios motores de vias glutamatérgicas (D'AMICO et al., 2013; MORAHAN; PAMPHLETT, 2006; ROTUNNO; BOSCO, 2013). Em uma ocorrência menor, a ALS pode ser herdada em um padrão autossômico recessivo, uma vez que os pais de um indivíduo com FALS podem não ser afetados, sendo frequentemente confundida com a forma esporádica[6].

Apesar de existirem semelhanças, os sintomas entre SALS e FALS se diferenciam: na condição esporádica os pacientes geralmente desenvolvem as características em seus primeiros 50 ou 60 anos, enquanto em pacientes com FALS, os sinais aparecem no início dos 40 ou 50 anos, raramente essas pessoas desenvolvem os sinais na infância ou adolescência, condição chamada de ALS juvenil.

1.1.6 Patogenia - O Papel de Disfunções Proteicas

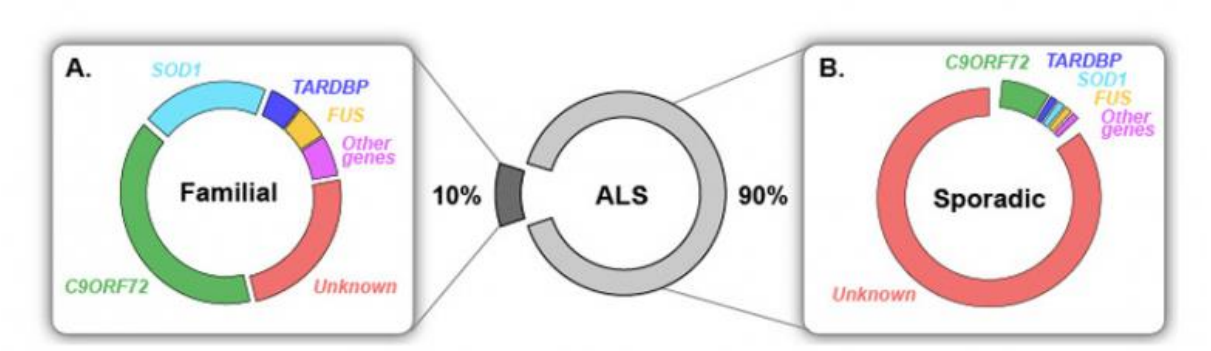
A patogênese da ALS se caracteriza principalmente pela associação de proteínas específicas disfuncionais. Tais disfunções são consequência de mutações genéticas, mais especificamente mutações *missense*. Essas mutações podem alterar a estrutura e função proteicas (ZEINEDDINE et al., 2017). Pesquisas revelam que mais de 50 genes podem estar associados à patogênese da ALS, levando à um amplo espectro de fenótipos (Tabela 6 - Apêndice) (TAYLOR; BROWN; CLEVELAND, 2016; ZUFIRÍA et al., 2016).

Mutações que ocorrem no gene da proteína *superoxide dismutase 1* (SOD1) e das proteínas de ligação ao DNA e RNA, TDP-43 e FUS/TLS, possuem particular importância por estarem associadas a um maior número de casos de ALS (Figura 2) (HAYDEN; CONE; JU, 2017; KABASHI et al., 2008; ROSEN et al., 1993; SCOTTER; CHEN; SHAW, 2015; TAYLOR; BROWN; CLEVELAND, 2016; VANCE et al., 2009). Cerca de 20% dos casos de FALS e 2%

[6] https://ghr.nlm.nih.gov/condition/amyotrophic-lateral-sclerosis#inheritance

de casos de SALS, possuem ao menos uma mutação na SOD1 e, aproximadamente 2% a 5% dos casos de SALS possuem mutações na TDP-43 (WIJESEKERA; LEIGH, 2009).

Figura 2 - Participação dos Principais Genes na ALS Hereditária e Esporádica



A figura representa o percentual de casos de FALS (A) e SALS (B) entre os casos de ALS e ilustra a participação relativa dos genes C9ORF72, SOD1, TARDP e FUS. Retirada de (LAFERRIERE; POLYMENIDOU, 2015).

Devido às mutações genéticas, proteínas são expressas em formas patogênicas associadas à ALS. Essas formas apresentam instabilidade estrutural, com alta chance de problemas relacionados ao enovelamento, além de carga e solubilidade anormais. Nos casos de SALS é comumente observado o acúmulo de proteínas erroneamente enoveladas (do inglês, *misfolded*) e agregados insolúveis no citoplasma de neurônios motores. A SOD1 e TDP-43 são as principais proteínas observadas nessas formas patogênicas, tanto em sua forma mutante quanto selvagem (ZEINEDDINE et al., 2017).

As proteínas SOD1, TDP-43 e FUS/TLS são caracterizados em diferentes níveis de desordem intrínseca, o que também tem importante participação em suas funções, mas também estão relacionados à patogenicidade (SANTAMARIA et al., 2017). Essas três proteínas desempenham papel chave para o funcionamento normal dos neurônios motores e também de outras células[7].

A patologia da ALS possui um padrão característico de acúmulo e deposição das proteínas SOD1, TDP-43 e FUS/TLS nos neurônios motores, provavelmente em razão da propagação da forma patogênica dessas três proteínas (ZEINEDDINE et al., 2017). Sabe-se que o enovelamento incorreto da proteína, o qual provavelmente ocorre devido à interação

[7] https://ghr.nlm.nih.gov/condition/amyotrophic-lateral-sclerosis#inheritance

inapropriada das superfícies hidrofóbicas expostas com os componentes celulares, pode levar à disfunção celular (BOLOGNESI et al., 2010).

Mutações em proteínas (como a TDP-43 e FUS/TLS) que interagem com moléculas de RNA, impactam no metabolismo deste mesmo material, o que possui importante papel na degeneração dos neurônios motores (RENTON; CHIÒ; TRAYNOR, 2014). Algumas proteínas, assim como a TDP-43, FUS/TLS e SOD1, podem se tornar patogênicas via mutações germinativas ou através de modificações pós-traducionais não-Mendelianas, especialmente nas doenças neurodegenerativas (BEYER; ARIZA, 2013).

A ALS é uma desordem complexa e multifatorial e alguns casos clínicos podem estar relacionados a mais de uma proteína. A SOD1, por exemplo, pode desempenhar papel importante no desenvolvimento da ALS, seja juntamente com outras proteínas, ou não, demonstrando como a doença pode ser complexa (FORSBERG et al., 2010; POKRISHEVSKY et al., 2012). Um experimento in vitro mostrou que a presença de agregados de SOD1 permite a observação de fragmentos de TDP-43 no citoplasma das células, sugerindo que a adição de agregados de SOD1 ao ambiente extracelular de células neurais resulta na mudança de localização, agregação e fragmentação da TDP-43 (CIRYAM et al., 2017; ZEINEDDINE et al., 2017) .

Pesquisas têm mostrado que fenótipos de ALS são em sua maioria causados por proteínas que interferem na maquinaria autofágica e/ou estão envolvidas no metabolismo de DNA e RNA (RATTI; BURATTI, 2016).

1.1.6.1 Conhecendo a SOD1: funções e relação com a ALS

A SOD1 é uma metaloenzima ubíqua, que se liga ao cobre e zinco, homodimérica, composta por 154 aminoácidos (Apêndice 10.6) (NARUSE et al., 2013; ROTUNNO; BOSCO, 2013). Cada subunidade da SOD1 é formada por oito folhas-beta antiparalelas, contendo uma ponte dissulfeto intramolecular e uma interação com um átomo de zinco e um átomo de cobre (Figura 31), os quais são componentes importantes do sítio catalítico. Essa formação também confere estabilidade estrutural, também promovida pela ponte dissulfeto e modificações pós-traducionais (KAWAMATA; MANFREDI, 2010; KUMAR et al., 2017; NARUSE et al., 2013; ROTUNNO; BOSCO, 2013a, 2013b; SRINIVASAN; RAJASEKARAN, 2017a, 2017b).

A SOD1 é secretada pelas células neurais, sendo encontrada no fluido cerebrospinal e em todas as células, por isso é considerada uma proteína ubíqua. Representa cerca de 0,5% de todo o

conteúdo proteico do cérebro humano e aproximadamente 90% do total de proteínas do tipo SOD (ANDERSEN, 2006; SUNDARAMOORTHY et al., 2013; WANG et al., 2008). Esta proteína está majoritariamente localizada no citoplasma e, até certo nível, no espaço intermembrana da mitocôndria e no núcleo celular, mas também pode ser encontrada em muitos outros compartimentos celulares em quantidades menores (KAWAMATA; MANFREDI, 2010; PAPA; MANFREDI; GERMAIN, 2014).

A SOD1 é responsável por catalisar a dismutação, dependente de metal, de ânions superóxidos tóxicos em peróxido de hidrogênio e oxigênio, através de reduções e reoxidação de íons de cobre, sendo assim a coordenação do cobre dentro dos monômeros da SOD1 é essencial à sua funcionalidade (ROTUNNO; BOSCO, 2013).

Considerando a catálise descrita, a SOD1 possui função protetora contra o estresse oxidativo (ROTUNNO; BOSCO, 2013). E associada à essa função, a SOD1 também possui papel importante na homeostase celular, destruindo os radicais que são normalmente produzidos dentro das células (ANDERSEN, 2006).

A catálise realizada pela SOD1 desempenha uma importante cascata liberando o peróxido de hidrogênio, implicando em outras modulações da via, como expressão gênica, proliferação, diferenciação e morte celular e, transdução de sinal. O papel da SOD1 na transdução de sinal é modulado pelo peróxido de hidrogênio liberado e pela SOD1 extracelular secretada pelas células, como microglias, podendo aumentar os níveis de cálcio intracelular, o qual tem efeito neuroprotetor nos neurônios granulares cerebelares (POLAZZI et al., 2013; ROTUNNO; BOSCO, 2013).

A associação da SOD1 com a ALS foi descoberta em 1993 (ROSEN et al., 1993), sendo o primeiro gene onde foram encontradas mutações causadoras da ALS, responsáveis por 6% de todo os casos e 15-20% dos casos de FALS (BUNTON-STASYSHYN et al., 2015; PASINELLI et al., 2004; ROTUNNO; BOSCO, 2013; SILVERMAN et al., 2016). Essa associação SOD1-ALS é complexa e pode envolver estresse oxidativo, desmetilação, perda de função de protease e autofagossoma, disfunção bioenergética mitocondrial, transporte axonal prejudicado, excitotoxicidade do glutamato, *mRNA splicing* e outras modificações pós-traducionais (ROTHSTEIN, 2009; VAN ZUNDERT; BROWN, 2017).

A SOD1 possui tendência a formar agregados fibrilares citotóxicos na ausência de pontes dissulfeto, formadas em cada monômero pelas cisteínas 57 e 146, e também pela ausência de

íons de zinco, já que a interação com o zinco promove a dimerização e estabilização da estrutura nativa e a ponte dissulfeto é muito importante para a estabilidade, influenciando na temperatura de *melting* (ROTHSTEIN, 2009; VAN ZUNDERT; BROWN, 2017). A estabilidade estrutural também é influenciada por íons de cobre (DAS; PLOTKIN, 2013).

A formação de agregados pode envolver também um *cross-linking* dissulfeto, envolvendo a cisteína 111 (COZZOLINO et al., 2008). Além disso, a reação de desmetilação e/ou redução da interação dissulfeto reduz a afinidade pelos metais cobre e zinco. As proteínas envolvidas na ALS têm alta tendência em perder seus metais e de tornar monômeros seus homodímeros (DAS; PLOTKIN, 2013).

As mutações podem se estender por toda a estrutura, sendo que cerca de 80% são substituições do tipo *missense* (troca de um resíduo de aminoácido por outro). De modo geral, mutação em regiões conservadas são mais propensas a causar ALS, visto a importância evolutiva e funcional para a proteína (ROTUNNO; BOSCO, 2013).

As formas mutantes da SOD1 são mais propensas a erros de enovelamento e prejuízo da formação da ponte dissulfeto e da ligação com os metais, fatores associados à toxicidade (SRINIVASAN; RAJASEKARAN, 2017). As formas selvagens da proteína também podem se enovelar erroneamente sob perturbações não genéticas, como por exemplo depleção dos metais, ruptura da estrutura quaternária, oxidação e super expressão, o que pode fazer com que a SOD1 assuma uma conformação tóxica, similar às suas variantes mutantes (ROTUNNO; BOSCO, 2013; SILVERMAN et al., 2016; SUNDARAMOORTHY et al., 2013).

Formas mal enoveladas podem funcionar como “sementes”, espalhando o erro entre SOD1 endógenas não mutantes, célula a célula, em uma forma de propagação *prion-like*, modelando a transmissão intracelular da doença pelo centro neural. Essa transmissão pode ocorrer como agregados de proteína que são liberados de células mortas e absorvidas por células vizinhas via macropinocitose ou por uma via dependente de exoma, com a associação de vesículas que são liberadas no ambiente extracelular e absorvidas via endocitose (GRAD et al., 2014a, 2014b; MÜNCH; O’BRIEN; BERTOLOTTI, 2011; SILVERMAN et al., 2016).

A propagação independente de contato célula-célula, dependendo apenas da liberação extracelular de agregados, pode explicar por que a forma predominante de ALS, SALS, pode perpetuar e se disseminar sistemática e progressivamente dentro do corpo. As formas mal enoveladas da SOD1 podem representar cerca de 4% do total das proteínas SOD1 presente na

medula espinhal de pacientes com SALS (GRAD et al., 2014; MÜNCH; O'BRIEN; BERTOLOTTI, 2011).

Essas formas da SOD1 podem causar uma cascata de desregulações dentro do sistema celular. O estresse e apoptose do retículo endoplasmático, fragmentação do Golgi e problemas mitocondriais, têm sido relatados como parte da patogênese estudada em SALS e FALS (SUNDARAMOORTHY et al., 2013). Essas formas podem ativar células da micróglia, elevando a síntese de óxido nítrico e a secreção de superóxidos e citocinas pró-inflamatórias (ROTUNNO; BOSCO, 2013). Além disso, as formas mal enoveladas presentes nos axônios podem inibir o transporte axonal através de mecanismos envolvendo fosforilação de proteínas, comprometendo a integridade de neurônios motores (ROTUNNO; BOSCO, 2013).

1.1.6.2 Conhecendo a TDP-43: funções e relação com a ALS

A *transactive response* (TAR) *DNA binding protein* 43 kDa (TDP-43) é uma ribonucleoproteína de 414 aminoácidos codificada pelo gene TARDBP (MACKENZIE; RADEMAKERS, 2008). Essa proteína possui dois motivos de reconhecimento de RNA e uma região C-terminal rica em glicina que garante que essa proteína se ligue a uma fita única de DNA, RNA e a outras proteínas (BURATTI et al., 2001; KUO et al., 2014; WANG et al., 2004).

A TDP-43 é, assim como a SOD1, ubíqua, sendo expressa em muitos tecidos, incluindo o cérebro, onde se localiza dentro do núcleo de neurônios e algumas células da glia (BURATTI et al., 2001; NEUMANN et al., 2006).

Entre as funções da TDP-43 é possível citar sua participação na repressão da transcrição do DNA (OU et al., 1995), estabilidade de mRNA, biogênese de microRNA, apoptose e divisão celular (BURATTI; BARALLE, 2008). Possui papel na regulação da plasticidade neuronal, atuando como fator de resposta à atividade neuronal (WANG et al., 2008), sendo responsável também por regular muitos genes (incluindo seu próprio gene) que possuem papel na manutenção de diversas células, incluindo os neurônios (RATTI; BURATTI, 2016).

A TDP-43 possui domínios clássicos de ribonucleoproteínas heterogêneas (hnRNP), N-terminal e C-terminal, regiões usualmente designadas à interações protéicas, sendo que a região C-terminal envolvida nas interações de ribonucleoproteínas e *splicing*; além de dois motivos de reconhecimento de RNA, para ligação ao RNA alvo (LUKAVSKY et al., 2013; RENTON; CHIÒ; TRAYNOR, 2014). Essa proteína também se liga à pre-mRNA nascentes para regular

seu processamento, participa da regulação de RNAs não codificantes, processamento de miRNA, estabilidade de mRNA, transporte e tradução (RATTI; BURATTI, 2016).

No contexto patológico, a TDP-43 pode formar grânulos de agregados e espalhar essa tendência via processo de macropinocitose (ZEINEDDINE et al., 2015; RATTI; BURATTI, 2016). É um fator em comum entre FALS e SALS e para a ALS-FTD com inclusões ubiquitinadas (ARAI et al., 2006; MACKENZIE; RADEMAKERS, 2008; NEUMANN et al., 2006). A associação da TDP-43 com mecanismos que causam ALS foi descrita pela primeira vez em 2006, sendo reportada como principal componente dos agregados observados em neurônios na ALS (ARAI et al., 2006).

As inclusões imunorreativas, pré-inclusões citoplasmáticas granulares e neurites distróficas são características comuns observadas em vários espectros fenotípicos da ALS e devido à toxicidade causada (CAIRNS et al., 2007; MACKENZIE; RADEMAKERS, 2008).

Grande parte das formas patogênicas observadas se devem a ocorrência de mutações *missense* na TDP-43, as quais participam de cerca de 5% dos casos de FALS (BURATTI, 2015). Essas mutações localizam-se predominantemente na região C-terminal da proteína e podem estar relacionadas à propensidade de formar agregados, ao tempo de meia-vida da proteína, às interações proteicas e à alteração na localização subcelular (BURATTI, 2015). A localização subcelular anormal da TDP-43 sugere um mecanismo patogênico que envolve a perda da função nuclear normal de regulação da transcrição, alterações em mecanismos de *splicing* e estabilidade de mRNA. O prejuízo no processamento de RNAs e a presença de agregados no citoplasma podem causar o sequestro de proteínas e RNAs, causando toxicidade de neurônios motores e corticais (BURATTI; BARALLE, 2008; VANCE et al., 2009).

A região C-terminal é altamente conservada. As mutações observadas, comumente para serina e treonina, podem aumentar a fosforilação dessa região da proteína e da formação de agregados, formação de inclusões, mudança de localização subcelular e causar ALS (KABASHI et al., 2008).

A progressão da ALS associada a TDP-43 está associada à erros de enovelamento e sua automontagem amilóide, à transmissão das formas patogênicas célula a célula em via *prion-like*, como anteriormente descrito para a proteína SOD1 (KANOUCHI; OHKUBO; YOKOTA, 2012; NONAKA et al., 2013; PELED et al., 2017). Mas este não é o único mecanismo de propagação, existindo ainda a disseminação de TDP-43 fosforiladas a longas distâncias pelo

sistema nervoso central, através de transporte axonal, transmissão pelas sinapses e terminais de axônios (BRETTSCHNEIDER et al., 2014; FEILER et al., 2015; NONAKA et al., 2013).

1.1.6.3 Conhecendo a FUS/TLS: funções e relação com a ALS

A proteína FUS/TLS (*fused in sarcoma/translocated in liposarcoma*), assim como a TDP-43, é uma ribonucleoproteína heterogênea nuclear (RATTI; BURATTI, 2016). Sua estrutura é composta por 526 resíduos de aminoácidos, uma região N-terminal com um domínio rico em Glu-Gly-Ser-Tyr, um motivo de reconhecimento de RNA altamente conservado (RRM), um motivo *zinc-finger* e um domínio C-terminal de múltiplas repetições Arg-Gly-Gly e a sequência núcleo, domínio que, juntamente com o RRM, reúnem a maior parte das mutações (DENG; GAO; JANKOVIC, 2014; RATTI; BURATTI, 2016; RENTON; CHIÒ; TRAYNOR, 2014). Os domínios são necessários para mediar interações proteína-RNA e proteína-proteína em várias atividades em níveis transcricional e pós-transcricional, como processamento de RNAs (RATTI; BURATTI, 2016; VANCE et al., 2009).

A FUS/TLS, assim como a SOD1 e a TDP-43, é universalmente expressa pelas células e se localiza predominantemente no núcleo celular. A FUS/TLS possui similaridade funcional com a TDP-43, se liga à proteínas motoras como cinesina e miosina, está envolvida no reparo de DNA, regulação da transcrição, controle de mRNAs em neurônios, participa do metabolismo de RNAs não codificantes, *splicing* de RNA, exportação de RNAs para o citoplasma, bem como transporte de mRNA, no qual defeitos no transporte axonal são características presentes no mecanismo patológico da ALS (DE VOS et al., 2008; RENTON; CHIÒ; TRAYNOR, 2014; VANCE et al., 2009a, 2009b).

Embora exista grande semelhança funcional entre a FUS/TLS e a TDP-43, essas duas proteínas possuem papéis distintos em diferentes tipos de RNAs e participam de maneiras diferentes da patogenicidade em ALS (DENG et al., 2014; RATTI; BURATTI, 2016).

Em 2009 a FUS/TLS foi descoberta como um fator comum e abundante em vários subtipos de ALS e FTD (RATTI; BURATTI, 2016; VANCE et al., 2009). Uma ampla taxa de pacientes de ALS e FTD apresentam formas agregadas da proteína, em alguns casos juntamente com a presença de TDP-43, no cérebro e na medula espinhal (DENG et al., 2014; RATTI; BURATTI, 2016). Mutações na FUS/TLS são responsáveis por cerca de 5% dos casos de ALS, considerando FALS e SALS (RATTI; BURATTI, 2016; ROTUNNO; BOSCO, 2013).

As mutações na FUS/TLS ocorrem principalmente na região C-terminal da proteína, na qual se localiza a NLS (*nuclear localization signal*). Essas mutações causam prejuízo funcional dessa região e podem levar à alteração da localização normal da proteína no citoplasma e ao estresse celular, o qual é sugerido em muitas pesquisas como o primeiro passo na patogênese da ALS (BENTMANN; HAASS; DORMANN, 2013; DE SANTIS et al., 2017; EMDE et al., 2015; RATTI; BURATTI, 2016; VANCE et al., 2009).

Contudo, pesquisadores têm sugerido que essa disfunção observada na FUS/TLS sozinha não é suficiente para causar a degeneração dos neurônios motores. Portanto, uma hipótese é que o mecanismo patológico envolvendo a ausência dessa proteína no núcleo e sua toxicidade no citoplasma, é apenas parte do seu papel na ALS (KINO et al., 2015; LING; POLYMENIDOU; CLEVELAND, 2013; SCEKIC-ZAHIROVIC et al., 2016).

Devido à participação da FUS/TLS na biogênese e metabolismo de RNA, mutações nesta proteína podem causar alterações no transcriptoma (DE SANTIS et al., 2017). É possível que, associado a esse fato, miRNAs possuem papel no desenvolvimento da ALS, já que essas moléculas também desempenham função na saúde dos neurônios motores (CAPAUTO et al., 2018; DINI MODIGLIANI et al., 2014; EMDE et al., 2015; KYE; GONÇALVES, 2014). Neste contexto é importante ressaltar que, de modo geral, proteínas associadas à ALS estão envolvidas na biogênese de miRNAs (DE SANTIS et al., 2017). Adicionalmente, mutações na FUS/TLS também podem afetar sua própria expressão, já que se trata de uma proteína autorregulada (VANCE et al., 2009).

1.1.7 Abordagens Computacionais no Estudo e Análise dos Efeitos de Mutações

Atualmente existem três medicamentos aprovados pelo *Food and Drug Administration* (FDA) para o tratamento da ALS. São eles o riluzole (LEMIESZEK et al., 2018), edaravone (CRUZ, 2018; MIYAJI et al., 2015) e dextrometorfano/quinidina (GREDAL et al., 1997). Apenas o riluzole tem se mostrado mais eficaz no aumento da sobrevida de pacientes e ainda assim existem muitas exceções. Dessa forma além dos tratamentos paliativos (WIJESEKERA; LEIGH, 2009), muitas pesquisas têm se voltado ao desenho de novas moléculas e reposicionamento de drogas, com o objetivo de alcançar um maior espectro da doença, atendendo a um maior número de pessoas.

Retrato desse esforço científico é o Drugbank (WISHART et al., 2018), que aponta 121 fármacos em teste para utilização em ALS, além do composto CuATSM, sob estudo experimental em ratos (WILLIAMS et al., 2016).

O número de moléculas em estudo demonstra a importância clínica de ALS, considerando seu amplo espectro fenotípico, muitos dos tratamentos atingem apenas uma limitada parcela da comunidade portadora da doença. Novos agentes no combate dessa doença podem trazer aumento na qualidade de vida de pacientes, principalmente aqueles refratários aos medicamentos atualmente em uso.

Mas a pesquisa científica da ALS não se resume às pesquisas medicamentosas. Muitos grupos têm trabalhado no sentido de tentar compreender fatores envolvidos nos mecanismos moleculares que causam a doença, de modo a buscar melhores alternativas de tratamento e critérios para segregação de pacientes (AL-CHALABI; VAN DEN BERG; VELDINK, 2017; BRAHIMI et al., 2016; MACHTOUB; KASUGAI, 2015).

A ALS possui muitas variações fenotípicas, que podem estar relacionadas à características e ocorrência de mutações *missense* e os mecanismos pelos quais essas podem afetar as proteínas nas quais ocorrem. Por esse motivo, a comunidade científica tem devotado um esforço significativo para o desenvolvimento de plataformas *in silico* para análise e predição dos efeitos moleculares de mutações na estrutura e função de proteínas, habilidade que tem se mostrado altamente relevante para o processo de elucidação de mecanismos de doenças e relação com fenótipos e severidade (ALBANAZ et al., 2017). Como exemplo é possível citar fatores importantes como a alteração da estabilidade estrutural, a capacidade de interação de proteínas com outras moléculas, a flexibilidade, entre muitos outros aspectos (NEMETHOVA et al., 2016; PEREIRA et al., 2019; PIRES; ASCHER, 2016; PIRES; BLUNDELL; ASCHER, 2015, 2016; RODRIGUES; ASCHER; PIRES, 2018; RODRIGUES; PIRES; ASCHER, 2018; TREZZA et al., 2017; USHER et al., 2015; VEDITHI et al., 2018).

Esse esforço se deve à necessidade de entender os mecanismos moleculares de doenças, de forma a direcionar terapias e medicamentos. Como exemplo é possível citar estudos de mecanismos moleculares de doenças como a síndrome de von Hippel-Lindau, onde foi possível a otimização da triagem de pacientes com base nas mutações *missense* com dados clínicos e experimentais, analisando *in silico* os efeitos funcionais dessas mutações ao fenótipo (GOSSAGE et al., 2014), prevendo o risco de desenvolvimento de câncer renal associado a determinadas mutações. Outro estudo que exemplifica a relevância do entendimento do

mecanismo molecular de patogenias envolve a Alcaptonúria. Esta é uma rara desordem autossomal recessiva, na qual mutações no gene da enzima homogentisate 1,2-dioxigenase, são responsáveis por desestabilizar a estrutura e interações da proteína, afetando sua função e a relação com fenótipos clínicos. Utilizando abordagens computacionais foi possível destacar dados distintos de pacientes que reagem ao tratamento, daqueles que não reagem (USHER et al., 2014). Além do estudo de mutações em humanos, métodos computacionais tem ainda importante aplicação em estudos de mutações que causam resistência a fármacos em microrganismos, como no *Mycobacterium tuberculosis*. A habilidade de predição computacional do efeito de mutações permitiu identificar *hotspots* para mutações potencialmente causadoras de resistência à fármacos, otimizando inclusive o trabalho de validação experimental (PHELAN et al., 2016).

O trabalho de Kumar e colaboradores (KUMAR et al., 2017) é um exemplo de estudos que buscam correlacionar mutações *missense* em proteínas com fenótipos observados em pacientes portadores da ALS. Kumar e colaboradores (KUMAR et al., 2017) buscaram mostrar a importância de um fator estrutural sob o efeito de mutações, a estabilidade. Foram analisadas correlações entre a alteração na estabilidade estrutural da SOD1, experimentalmente aferidas e descritas como a variação na energia livre de Gibbs (ΔΔG, em kcal/mol) sob o efeito de 30 mutações *missense* isoladamente, com os respectivos dados de pacientes, no quesito idade de início da doença, idade de óbito e tempo de sobrevivência.

A complexidade de doenças como a ALS requer análises de fatores que vão além da estabilidade. Para melhor entender o contexto dessa desordem, buscou-se adicionar outros atributos além da estabilidade, bem como aumentar o número de casos clínicos coletados e utilizados para as análises de correlação. Dessa forma buscou-se ampliar o plano de análise e aproximá-la da complexidade multifatorial real da doença.

O acesso à um banco de dados completo e atualizado também é de grande relevância para estudos de doenças. Neste contexto, apesar do banco de dados em ALS existente, o ALSoD (ABEL et al., 2013) ser um repositório para mutações de vários tipos em ALS, este banco encontra-se desatualizado desde 2015. Além disso muitos dos casos clínicos depositados não possuem referência rastreável na literatura. Adicionalmente, essa base de dados não conta com dados estruturais e efeito de mutações na proteína (ABEL et al., 2013). Sendo assim, há uma carência por um banco de dados atualizado e que compreenda informações necessárias ao estudo de causas e efeitos moleculares da doença.

2 Justificativa

Apesar das funções das proteínas SOD1, TDP-43 e FUS/TLS estarem bem estabelecidas na literatura, ainda se sabe pouco sobre o mecanismo molecular que estas proteínas desempenham na ALS e sua relação com o amplo espectro de fenótipos clínicos.

Considerando que a ALS pode ser desencadeada por um ou mais fatores, como uma ou mais mutações por exemplo, é importante considerar sua complexidade tanto molecular quanto fenotípica, levando, também em conta os diversos efeitos moleculares que uma mutação pode desencadear.

A complexidade observada na ALS se torna um grande desafio a ser vencido. Ainda existem muitas limitações a serem enfrentadas, como por exemplo a carência de estruturas elucidadas para a maior parte das proteínas que participam do processo patológico, bem como a carência de um banco de dados amplo ligando mutações e seus efeitos moleculares à diferentes fenótipos.

Estudar a estrutura de proteínas envolvidas na ALS é de alta relevância na trajetória para decifrar o mecanismo molecular do qual disfunções proteicas fazem parte.

Sendo assim, este presente trabalho busca preencher parte dessa lacuna, buscando analisar as consequências moleculares e fenotípicas de mutações *missense* em proteínas associadas à ALS, em especial na SOD1, TDP-43 e FUS/TLS, discriminando onde mutações *missense* associadas a fenótipos críticos estão localizadas estruturalmente e buscando predizer acuradamente os fenótipos clínicos da ALS. Isso, por sua vez, poderá guiar alterações na conduta clínica e gerenciamento de pacientes, possibilitando a adoção de tratamentos mais eficazes e personalizados. Considerando que a ALS é uma desordem idade-dependente, a relação genótipo-fenótipo e a predição da idade de início, idade de óbito e tempo de sobrevivência, podem ser peças-chave para tratamentos personalizados e mais eficazes, em tempo hábil.

3 Objetivos

3.1 Objetivos Gerais

Elucidar os mecanismos moleculares da ALS e utilizar características estruturais de proteínas envolvidas nessa desordem para desenvolver uma plataforma para predizer acuradamente seus fenótipos clínicos.

3.2 Objetivos Específicos

- Ampliação da base de conhecimento da literatura em relação às mutações em ALS e a coleta de dados de pacientes, pela construção de um banco de dados relacional.
- Caracterização de efeitos moleculares de mutações em ALS, utilizando abordagens computacionais.
- Identificação de relações entre mutações e seus potenciais efeitos e participação nos mecanismos moleculares que causam a doença.
- Desenvolvimento de métodos preditivos capazes de relacionar mutações a diferentes fenótipos da doença, por meio de aprendizado de máquina supervisionado.

4 Metodologia

A metodologia foi dividida em nove passos principais em alinhamento aos objetivos propostos, como ilustrado na Figura 3. Iniciou-se pela revisão da literatura, de modo a coletar dados de casos clínicos de portadores da ALS com mutações nos genes das proteínas SOD1, TDP-43 e FUS/TLS. Subsequentemente foram obtidas as sequências de aminoácidos das proteínas SOD1, TDP-43 e FUS/TLS, bem como suas estruturas tridimensionais e modelos obtidos por modelagem comparativa. Utilizando plataformas preditivas *in silico* foram feitas predições baseadas em sequência e em estrutura 3D.

Com base nas sequências de aminoácidos foram feitas predições de conservação, taxa de tolerância a mutações *missense*, efeito funcional e tolerância às mutações *missense* específicas, desordem estrutural e busca pela frequência dos alelos das mutações na população.

As predições baseadas em estrutura 3D, até o momento foram realizadas apenas para a SOD1, devido à baixa qualidade dos modelos estruturais obtidos para as proteínas TDP-43 e FUS/TLS, sendo necessário dar prioridade a otimizações das estruturas. Dessa forma, os resultados apresentados referem-se em sua maioria a SOD1. As predições incluem as interações moleculares do homodímero da SOD1, efeitos das mutações na estabilidade, flexibilidade, conformação, interação entre os monômeros da estrutura, distância dos resíduos à interface de contato, tanto entre os monômeros quanto à superfície exposta, distância dos resíduos aos metais e acessibilidade relativa ao solvente (RSA).

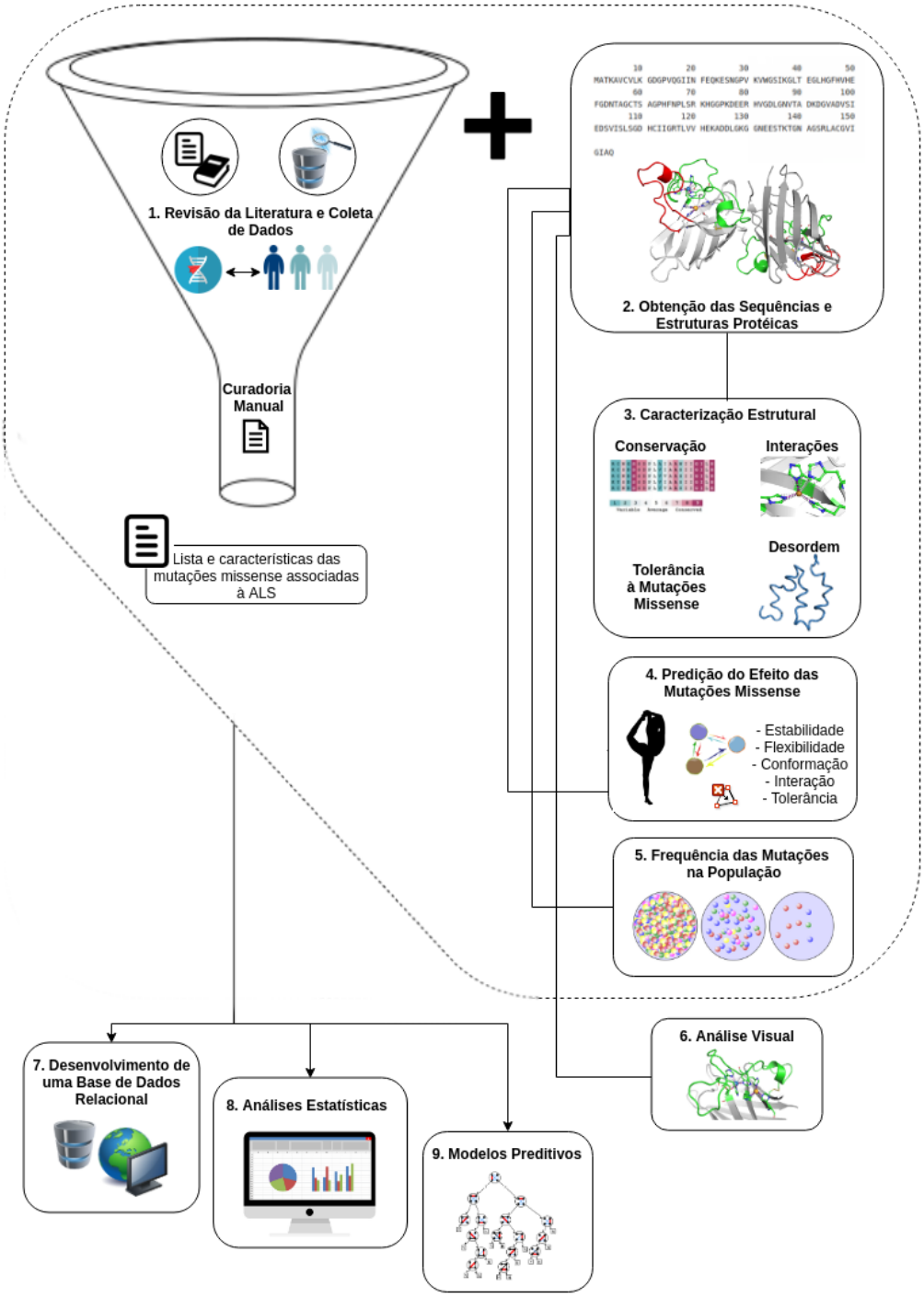
Os possíveis efeitos das mutações *missense* na estrutura da SOD1 foram avaliados visualmente, bem como sua relação com as características estruturais, como por exemplo a alteração de um resíduo hidrofílico acessível ao solvente por um hidrofóbico.

Todos as informações coletadas e geradas passaram por uma curadoria manual e foram armazenadas na base de dados desenvolvida, DynAMISM.

Os atributos gerados pelas predições baseadas em sequência e estrutura foram analisados estatisticamente em busca de relações entre estes e os fenótipos de idade de início, idade de óbito e tempo de sobrevivência dos pacientes. Esta etapa teve como finalidade identificar dados relevantes para a compreensão do mecanismo molecular da doença, caracterizados pelo comportamento diferenciado conforme o atributo testado.

Os atributos de relevância foram utilizados para o desenvolvimento de modelos capazes de predizer a idade de início da ALS no paciente, idade de óbito e o tempo de sobrevivência.

Figura 3 - Metodologia



A metodologia seguiu o curso exemplificado nesta figura, onde primeiramente foi realizada a revisão da literatura com a simultânea coleta de dados de pacientes de ALS com mutações nas proteínas SOD1, TDP-43 e FUS/TLS. Esses dados passaram por curadoria manual e constituíram uma lista com características das mutações *missense* e seus respectivos casos clínicos associados (1). Foram então obtidas as sequências e estruturas das proteínas (2), as quais foram caracterizadas quanto à conservação, tolerância à mutações, ordem e desordem estrutural e interações dentro da estrutura, além das predições dos efeitos das mutações na estrutura (4). A frequência dos alelos das mutações foi reportada (5) e as principais mutações foram analisadas visualmente (6). Os dados gerados entre os passos 1 e 6 foram depositados na base de dados desenvolvida (DynAMISM) (7), sendo posteriormente analisados estatisticamente (8) e utilizados para a construção de modelos preditivos (9) dos fenótipos clínicos de idade de início, óbito e tempo de sobrevivência de pacientes.

4.1 Revisão da Literatura e Coleta de Dados - Mutações *Missense* em Portadores de ALS

Termos-chave relacionados a ALS (Tabela 7, aApêndice II - Palavras-chave Utilizadas na Revisão Bibliográfica) foram utilizados para guiar a busca pela literatura científica. As buscas foram realizadas no *Google Scholar* e *PubMed*, visando principalmente publicações compreendidas entre 2015 e 2018, visto que a base de dados *Amyotrophic Lateral Sclerosis Online Genetics Database* (ALSoD) (ABEL et al., 2013) teve sua última atualização em 2015. Essa base de dados será detalhada no decorrer do texto.

Os artigos encontrados durante a busca foram primeiramente selecionados com base na avaliação do resumo. A pesquisa se concentrou na descrição da ALS, com características específicas da SALS e FALS, incluindo dados clínicos de pacientes portadores de mutações *missense* nas proteínas SOD1, TDP-43 e FUS/TLS. Também foram coletadas informações sobre proteínas e mutações *missense* que possam ter papel chave na patogênese da doença e no contexto geral do distúrbio da ALS.

Os dados clínicos coletados incluem: a classificação em SALS ou FALS; a idade de início da doença (do inglês, *age of onset*); idade de óbito (do inglês, *age of death*) e tempo de sobrevivência do paciente (do inglês, *survival time*); gênero; país de origem; sítio inicial da doença; presença ou ausência de sintomas nos membros superiores, inferiores, sinais bulbares e respiratórios; além de informações adicionais como presença de outras síndromes, doenças e causa da morte.

A ALSoD (ABEL et al., 2013), mencionada anteriormente na introdução (Seção 1.1.7) é uma base de dados criada em 1999 para armazenar informações clínicas de mutações e pacientes de ALS. A base fornece informações clínicas (classificação SALS/FALS, gênero, país, etnia e idade de início da doença), genéticas e preditivas do efeito funcional de mutações (PROVEAN (CHOI; CHAN, 2015), SIFT (SIM et al., 2012) e pph2 (ADZHUBEI et al., 2010)). Apesar da variedade de genes abordados na ALSoD (ABEL et al., 2013), essa base teve sua última atualização em 2015 e apresenta limitação de informações clínicas, estruturais e preditivas.

Dessa forma foram coletadas todas as mutações *missense* reportadas nas proteínas SOD1, TDP-43 e FUS/TLS, disponíveis no acervo desse banco de dados e na literatura, incluindo todas as informações disponíveis acerca dos pacientes portadores destas mutações e seus fenótipos associados, como mencionado anteriormente.

4.2 Ampliação da Base de Conhecimento da Literatura em Relação às Mutações em ALS

As informações geradas em cada etapa desta metodologia para as mutações coletadas da SOD1, foram dispostas na base de dados relacional desenvolvida, a DynAMISM[8]. Um subconjunto desses dados, predições baseadas em sequência e estrutura, foram utilizados nas análises de correlações com fenótipos da ALS e como evidência para treinar e testar modelos preditivos capazes de predizer fenótipos de ALS e sua relação com as mutações *missense*.

A base de dados desenvolvida, DynAMISM, é dividida basicamente em dois grandes grupos de características, o grupo de características clínicas (descrito anteriormente no item 4.1) e de características estruturais e preditivas de mutações *missense* baseadas em sequência e em estrutura, como descrito no início da (item 4.1). Os dados de frequência populacional das mutações e contagem de interações de cada resíduo, são utilizados apenas na formação da base de dados.

Na DynaMISM cada dado clínico é representado por um código de mutação *missense*, portanto, cada linha da base de dados é constituída pelo código da mutação, os componentes do grupo de características clínicas e de características estruturais e preditivas de mutações *missense*.

Em relação às idades de início, óbito e tempo de sobrevivência (Tabela 1), pode-se observar que há casos onde é encontrado um "~" antes da idade, o que significa, por exemplo, que aquela foi a idade com a qual o paciente se apresentou no hospital e foi diagnosticado, mas não se sabe ao certo se coincide com o início dos sintomas. Também é possível observar o termo "*not described*" quando dada informação não foi disponibilizada e ">" quando, como no caso 13 na (Tabela 1), o paciente ainda estava vivo até os 47 anos, com sobrevivência mínima de 36 meses, quando o estudo foi publicado e, portanto, seu falecimento é desconhecido.

[8] https://github.com/AmandaAlbanaz/dynamism

Tabela 1 - Exemplificação da Representação de Dados Clínicos na Base de Dados

Caso	Idade de início (anos)	Idade de óbito (anos)	Tempo de sobrevivência (meses)
5	*not described*	57	*not described*
13	44	>47	>36
47	~40	40,10	10

Exemplos da disposição dos dados de idade de início, idade de óbito e tempo de sobrevivência e situações onde casos são reportados com a ausência de um ou mais desses dados. O caso número 5 não possui dado de idade de início e tempo de sobrevivência descrito na literatura. No caso 13 o paciente ainda estava vivo quando o estudo foi publicado, nesse caso o símbolo de maior "**>**" é utilizado. O paciente do caso 47 foi reportado após o início da doença, nesse caso o símbolo til "**~**" é utilizado, para reportar a idade de início aproximada. Os dados reportados com os símbolos "**>**" ou "*not described*" não são utilizados nas análises estatísticas.

4.3 Coleta das Sequências de Aminoácidos e Obtenção das Estruturas Tridimensionais

As sequências de aminoácidos das proteínas SOD1, TDP-43 e FUS/TLS foram obtidas do UniProt (UNIPROT CONSORTIUM, 2018) no formato fasta (Apêndice 10.6). A estrutura 3D da SOD1 foi extraída do Protein Data Bank (PDB) (BERMAN et al., 2000), PDB ID: 2C9V.

As estruturas da TDP-43 e FUS/TLS disponíveis no PDB não possuem cobertura e qualidade satisfatórias para sua utilização nas predições baseadas em estrutura, de modo que se utilizou modelagem por homologia para a obtenção de modelos estruturais mais completos. Isto fez-se necessário, visto a importância de estruturas com qualidade e resolução satisfatórias para os objetivos de predição do efeito de mutações em propriedades estruturais da proteína, como interações, estabilidade, flexibilidade, RSA, distância da interface de contato e, portanto, compreensão do mecanismo molecular.

A modelagem por homologia, ou modelagem comparativa, é a construção da estrutura tridimensional de proteínas a partir de estruturas já resolvidas, que são utilizadas como moldes. Os moldes são proteínas relacionadas ao modelo alvo (proteína a ser modelada), evolutivamente ou funcionalmente por exemplo, devido à conservação da sequência de aminoácidos estar relacionada com a conservação estrutural. Esse método é um dos mais difundidos para predição de estruturas protéicas tridimensionais (TRAMONTANO; MOREA, 2003; VYAS et al., 2012).

Devido à importância de gerar estruturas de alta qualidade, existem várias formas de aumentar acurácia da modelagem de modo a aproximá-la o máximo possível da estrutura nativa, como etapas de refinamento, baseadas em campos de força *physics-based* (ZHANG, 2008).

De forma prática, a metodologia da modelagem por homologia empregada se resumiu em quatro passos, utilizando o software Chimera (PETTERSEN et al., 2004) como ferramenta e interface para o programa de modelagem MODELLER (WEBB; SALI, 2016). Os passos executados foram:

1. **Busca por proteínas de estrutura conhecida**: Etapa crucial para uma modelagem de qualidade, já que moldes adequados fornecem informações como restrições espaciais[9], essencial ao correto enovelamento, dinâmica e funcionalidade da estrutura (WEBB; SALI, 2016).
    a. **Critérios para seleção dos moldes**: os moldes foram selecionados de acordo com critérios de identidade e similaridade mínimos de 35% com a sequência alvo. Foram utilizados critérios de busca padrão utilizados pelo MODELLER, considerando e-value máximo de 3, optando-se pela estrutura de menor e-value e maior escore.
    b. **Modelagem da TDP-43**: foram utilizados como moldes as estruturas resolvidas por NMR que cobrem trechos da sequência alvo com 100% de identidade: 5MRG (resíduos 1 a 10), 4BS2 (resíduos 102 a 269, domínios RRM1 e RRM2) e 2N3X (resíduos 311 a 360).
    c. **Modelagem da FUS/TLS**: foram utilizadas as estruturas estruturas resolvidas por NMR que cobrem trechos da sequência alvo: 2LA6 (resíduos 1 a 99, 100% de identidade), 2LCW (resíduos 278 a 385, 100% de identidade), 5YVG (resíduos 509 a 526, 100% de identidade) e 2CPE (resíduos 346 a 458, 56,3% de identidade), também resolvidas por NMR.
2. **Alinhamento entre as sequências molde e alvo** (HILBERT; BÖHM; JAENICKE, 1993; VYAS et al., 2012): Etapa que permitiu a verificação da escolha dos moldes, sua cobertura e identidade com a sequência alvo, como descritos no tópico anterior.
3. **Construção da estrutura alvo e refinamento:** Munidos com alinhamento satisfatório, realizou-se a construção e refinamento do modelo utilizando os parâmetros padrão do MODELLER:
    - Foram gerados 20 modelos para cada modelagem ou refinamento realizado. O refinamento foi realizado no programa Chimera. Esse refinamento consiste na utilização de modelo matemático para reduzir a energia do modelo construído e chegar a um mínimo global de energia. Para isso foram utilizados os critérios padrão. O refinamento

[9] https://proteinstructures.com/Modeling/homology-modeling.html

feito através do Chimera se baseia no uso do campo de força Amber no caso de resíduos padrões e no módulo Antechamber, no caso de resíduos não padrões (SALOMON-FERRER; CASE; WALKER, 2013).

4. **Validação e Critérios para Seleção dos Modelos:** a validação dos modelos gerados foi feita utilizando o PROCHECK (LASKOWSKI et al., 1993). A escolha por um único modelo foi feita seguindo os critérios abaixo:
    a. **Avaliação de energia do modelo:** avaliou-se o DOPE de cada modelo, sendo que quanto menor o valor, mais confiável e próximo o modelo está da estrutura nativa (SHEN; SALI, 2006);
    b. **Avaliação do gráfico de Ramachandran para cada modelo:** foi possível verificar a qualidade estereoquímica de dos modelos estruturais gerados. Nos gráficos de Ramachandran foi possível visualizar as combinações dos ângulos diédricos Ψ (psi) e Φ (phi). Ângulos que determinam a conformação das cadeias principais das proteínas. O ângulo Φ define a rotação em torno da ligação Carbono α - Nitrogênio do resíduo, e o Ψ define a rotação em torno da ligação Carbono α - Carbono do mesmo resíduo. Através da avaliação deste gráfico é possível escolher modelos nos quais os resíduos estão dispostos de modo a não causar impedimento estérico das cadeias laterais dos aminoácidos. É esperado que modelos de boa qualidade possuam acima de 90% dos resíduos em regiões favorecidas (LASKOWSKI et al., 1993).

4.4 Caracterização de Efeitos Moleculares de Mutações *Missense* em ALS

Visando conhecer melhor as características das proteínas em estudo, SOD1, TDP-43 e FUS/TLS, foram utilizados métodos computacionais (descritos nos tópicos subsequentes) para predição de características como conservação da sequência, estrutura secundária, desordem estrutural e interações moleculares. Esses dados fazem parte da composição do banco de dados e são de grande importância para compreender o impacto de mutações.

4.4.1 Predições Baseadas em Sequência

As tarefas de predição de características das proteínas e dos efeitos de mutações foram divididas em predições baseadas em sequência e predições baseadas em estruturas.

As predições baseadas em sequência utilizou informações da sequência linear de resíduos da proteína.

4.4.1.1 Dados de Conservação de Sequência

A fim de se obter os dados de conservação, foi utilizado o ConSurf Web Server[10] (CELNIKER et al., 2013). A análise da conservação de proteínas pode mostrar a relevância dos aminoácidos nos aspectos funcionais e estruturais (ASHKENAZY et al., 2016), fornecendo maior suporte e compreensão a processos que podem ocorrer ao longo da evolução. A taxa de conservação evolutiva pode indicar um equilíbrio entre uma tendência natural de mutação e a necessidade global de preservar a integridade estrutural e a função da proteína. Nesse sentido, uma ferramenta para estimar e visualizar a conservação evolutiva de macromoléculas é de alta relevância (ASHKENAZY et al., 2016).

Na execução do ConSurf foram utilizadas as sequências da SOD1, TDP-43 e FUS/TLS, bem como a estrutura da SOD1 (PDB ID: 2C9V) e o modelo estrutural da TDP-43. As estruturas foram utilizadas para visualização da conservação.

A partir da sequência de aminoácidos o ConSurf identifica os homólogos pela construção do alinhamento múltiplo de sequências. Utilizando esses dados e modelos evolutivos probabilísticos, o ConSurf estima os graus de conservação para cada posição da sequência fornecida (ASHKENAZY et al., 2016).

Os escores de conservação foram mapeados na estrutura fornecida e também nas sequências do alinhamento múltiplo, juntamente com as predições sítio-específicas de acessibilidade ao solvente (ASHKENAZY et al., 2016).

Apesar do procedimento automático e simples do ConSurf, em alguns passos se fez necessária a escolha por algoritmos de acordo com as necessidades e objetivos do estudo. Sendo assim, para a etapa de busca por homólogos, o algoritmo HMMER[11], sugerido como padrão, foi utilizado e a base de dados escolhida foi o Clean UniProt. O método *Bayesian* e MAFFT-L-INS-i (para alinhamentos de até 200 sequências) foram escolhidos para o cálculo dos escores de conservação e alinhamento múltiplo de sequências, respectivamente. Os demais procedimentos ocorreram de forma padrão como sugerido na interface web.

Os escores de conservação são divididos e traduzidos em uma escala discreta de nove graus, para visualização de cores que variam de azul (grau 1 - resíduos mais variáveis), branco (grau 5 - média) e vermelho (grau 9 - mais conservado).

[10] http://consurf.tau.ac.il/2016/
[11] http://hmmer.org/

4.4.1.2 Dados de Frequência Populacional de Mutações e Tolerância à Mutações *Missense*

Os dados de frequência populacional de alelos de mutações podem indicar quando uma mutação é benigna ou maligna, conforme sua prevalência.

A frequência populacional dos alelos mutantes foi obtida pela interface web do banco de dados gnomAD (*Genome Aggregation Database*) (EXOME AGGREGATION CONSORTIUM et al., 2016), fornecendo o nome dos genes das proteínas SOD1, TDP-43 e FUS/TLS. Essa base de dados tem por objetivo agregar e harmonizar uma ampla variedade de dados de projetos de sequenciamento de exoma e genoma. São dados provenientes de estudos de doenças e genética de populações (EXOME AGGREGATION CONSORTIUM et al., 2016).

Para obtenção de dados de variações genéticas em proteínas relacionadas à ALS em populações sadias foi utilizada a interface web do método MTR (*Missense Tolerance Ratio*) (TRAYNELIS et al., 2017), fornecendo-se o código do transcrito: SOD1 - ENST00000270142; TDP-43 - ENST00000240185 e FUS/TLS - ENST00000380244.

Essa ferramenta resume os dados de variações genéticas disponíveis em humanos dentro dos genes, reunindo a amplitude de variação genética no nível populacional. Este fornece dados acerca do panorama de tolerância de mutações *missense* ao longo das posições das sequências protéicas, importante para predição de mutações patogênicas com base na depleção preferencial de mutações *missense*, dada a variação total observada (TRAYNELIS et al., 2017).

4.4.1.3 Dados de Desordem Estrutural

Como mencionado na introdução, é importante destacar regiões desordenadas dentro de proteínas, regiões que não apresentam formação de estrutura secundária, visto que essas regiões podem causar dificuldades no processo de cristalização – como por exemplo regiões N e C-terminais e domínios específicos, sendo informações úteis para a validação de estruturas preditas (LINDING et al., 2003).

O estudo das regiões de ordem e desordem estrutural são importantes também para a entender a funcionalidade da proteína dado sua flexibilidade (muitos sítios funcionais são motivos lineares), além de vias e trajetórias de enovelamento (VERKHIVKER et al., 2003).

Para a predição dessas regiões foram utilizadas a interface web para a ferramenta IUPRED2A[129], fornecendo à interface as sequências de aminoácidos de cada proteína de interesse.

O IUPRED2A se baseia em um método de estimativa de energia, um potencial estatístico de baixa resolução para caracterizar tendências de aminoácidos em formar contatos. (MÉSZÁROS; ERDOS; DOSZTÁNYI, 2018).

Entre os tipos de predição do IUPRED2A estão o *long disorder* e o *short disorder*. Ambos foram utilizados. A predição pelo método *long disorder* visa a predição de regiões desordenadas de pelo menos 30 resíduos consecutivos, independentemente de contexto, enquanto o *short disorder* prediz regiões curtas, contexto-dependente e, possui tendência à predição de desordens nas regiões terminais das proteínas (MÉSZÁROS; ERDOS; DOSZTÁNYI, 2018).

4.4.1.4 Predição do Efeito de Mutações na Funcionalidade Proteica

Considerando a importância das mutações *missense* em vários aspectos que podem afetar a funcionalidade proteica, é de alta relevância utilizar plataformas preditivas que possam discriminar mutações neutras daquelas prejudiciais.

Para identificar quando a mutação é funcionalmente importante foi utilizada a interface web da ferramenta PROVEAN (*Protein Variation Effect Analyzer*)[13] (CHOI et al., 2012), que disponibiliza predições do impacto de mutações na função biológica da proteína se baseando na sequência de aminoácidos.

O PROVEAN utiliza uma abordagem de escores baseada em alinhamentos e possui três funções de predição, o *Protein for any organism*, *Protein Batch* e *Genome Variants*. A função utilizada foi o *Protein Batch*, que oferece suporte a predições de um grande número de variações, mutações *missense* no caso, fornecendo escores que representam prejuízo funcional ou não, sendo classificadas como *deleterious* ou *neutral*, respectivamente.

A interface *web* do PROVEAN também fornece predições de tolerância de mutações do SIFT (*Sorting Intolerant From Tolerant*) (SIM et al., 2012) e informação acerca da diversidade das sequências utilizadas para a predição, denominada *median sequence information* (CHOI; CHAN, 2015). Para obtenção de todas essas informações foi fornecida a lista de variantes

[12] https://iupred2a.elte.hu/
[13] http://provean.jcvi.org/genome_submit_2.php

contendo o identificador Ensembl da proteína, a posição de mutação, o aminoácido selvagem e o mutante.

Além do PROVEAN e SIFT, o PolyPhen2 (pph2) é um método para predição dos efeitos de mutações *missense* na estrutura e função de proteínas (ADZHUBEI; JORDAN; SUNYAEV, 2013). Essa ferramenta foi utilizada via instalação em máquina local.

O pph2 extrai algumas características preditivas da região de mutação, se baseando na sequência e na estrutura. Essas características são selecionadas por um algoritmo guloso iterativo e a partir disso alimenta um classificador probabilístico para que sejam feitas as predições (ADZHUBEI et al., 2010; ADZHUBEI; JORDAN; SUNYAEV, 2013).

O método também constrói um alinhamento múltiplo de sequências homólogas utilizando para isso escores baseados em perfil e identidade, os quais fornecem informações sobre a frequência do resíduo mutante no alinhamento e alterações no volume da cadeia lateral (ADZHUBEI et al., 2010; ADZHUBEI; JORDAN; SUNYAEV, 2013).

Ao utilizar o algoritmo *Naive Bayes* para as predições, o pph2 mensura a probabilidade de uma dada mutação ser prejudicial para a proteína - sendo classificada como *damaging* ou *probably damaging*, ou não apresentar efeito prejudicial - classificada como *benign* (ADZHUBEI et al., 2010).

O *workflow* utilizado para utilização local do pph2 se baseou em passar uma lista de mutações com o código de acesso para o UniProt para o *script* "*run_pph.pl*" (fornecido pelo pph2), o qual gera um arquivo de características relacionadas às mutações *missense*. Este arquivo gerado foi então passado como parâmetro para o *script* "*run_weka.pl*" (também fornecido pelo pph2), que fez as predições, *possibly damaging* ou *benign*, se baseando por padrão, no modelo de classificação HumDiv. Além desse modelo, foi utilizado também a alternativa para o modelo HumVar (ADZHUBEI; JORDAN; SUNYAEV, 2013).

A grande diferença entre esses modelos é que o HumDiv é utilizado para estudo de alelos raros, mapeamento de regiões identificados por estudos de *genome-wide association* e análises que envolvem seleção natural. Enquanto o HumVar é sugerido para casos de análises de alelos deletérios e doenças mendelianas, onde se faz necessária a distinção entre mutações que causam drásticos efeitos e àquelas consideradas neutras, por exemplo (ADZHUBEI; JORDAN; SUNYAEV, 2013).

4.4.2 Predições Baseadas em Estrutura

Nosso grupo vem buscando encontrar correlações entre características estruturais capazes de relacionar mutações *missense* à fenótipos de doenças, em exemplos de sucesso como o estudo de mutações em VHL e alcaptonúria, como anteriormente mencionado na introdução. Através das correlações entre genótipo e fenótipo é possível estudar o desenvolvimento de modelos preditivos baseados nesses atributos e determinar fatores-chave que requerem atenção mais específica.

4.4.2.1 Interações Não-covalentes

Entender quais interações são quebradas, enfraquecidas ou originadas quando ocorre uma mutação *missense* é de suma importância no processo de compreensão de mecanismos moleculares de doenças, visto que interações podem afetar a flexibilidade, a estrutura secundária e funcionalidade da proteína (JUBB et al., 2017).

O cálculo das interações moleculares da SOD1 foi realizado utilizando o servidor web Arpeggio[14] (JUBB et al., 2017). O Arpeggio recebeu como parâmetro o arquivo PDB, contendo a estrutura proteica e também a lista de resíduos de interesse para calcular interações como van der Waals, iônicas, metálicas, hidrofóbicas, contatos de halogênio, pontes de hidrogênio, interações envolvendo anéis aromáticos (cátion-pi, doador-pi, halogênio-pi, carbono-pi e pi-pi). Esses cálculos são baseados no tipo de átomo, distância e ângulos torcionais.

Além do cálculo de interações, a etapa de visualização das estruturas 3D das proteínas, de suas características e interações é crucial para a compreensão de parte do mecanismo molecular que pode levar à uma doença. Esta etapa foi realizada com o auxílio do *software* PyMol[15].

4.4.2.2 Predição do Efeito de Mutações *Missense* na Estabilidade Estrutural e Afinidade Proteína-Proteína

A fim de mensurar o efeito das mutações na estabilidade estrutural da SOD1, foram utilizados os preditores: mCSM[16] e DUET[17]. O mCSM prediz, além da estabilidade estrutural, a afinidade entre proteínas (Equação (2)), que no caso da SOD1 é calculada entre os monômeros.

O mCSM (*mutation Cutoff Scanning Matrix*) é um método baseado em aprendizado de máquina e assinaturas de grafos, que utiliza padrões de distância no ambiente do resíduo do tipo

[14] http://biosig.unimelb.edu.au/arpeggioweb/
[15] https://pymol.org/2/
[16] http://biosig.unimelb.edu.au/mcsm/
[17] http://biosig.unimelb.edu.au/duet/

selvagem, como evidência para treinar modelos preditivos. Para a predição do impacto de mutações *missense* na afinidade entre as cadeias da proteína SOD1 foi utilizado o mCSM-PPI, enquanto para mensurar a alteração de estabilidade, fez-se uso do mCSM-Stability (PIRES; ASCHER; BLUNDELL, 2014).

O DUET é um método integrado que agrega o mCSM-*Stability* e SDM (*Site Directed Mutator*) em uma predição consenso pela combinação dos resultados destes métodos em um preditor otimizado utilizando *Support Vector Machines* (PIRES; ASCHER; BLUNDELL, 2014).

Já o SDM é um método *knowledge-based* e *structure-based*. Este utiliza tabelas de substituição de aminoácidos ambiente-específicas para obter uma função potencial de energia estatística (WORTH; PREISSNER; BLUNDELL, 2011).

4.4.2.3 Predição do Efeito de Mutações *Missense* em Flexibilidade e Conformação

O DynaMut[18] é um método disponível na *web* (RODRIGUES; PIRES; ASCHER, 2018), que possibilita analisar e visualizar o impacto de mutações *missense* na dinâmica e estabilidade protéica. Para que isso seja possível, o método integra assinaturas baseadas em grafos com a abordagem de *normal mode analysis*, de modo a gerar uma predição consenso acerca do impacto de mutações no repertório conformacional proteico (RODRIGUES; PIRES; ASCHER, 2018). Portanto, o DynaMut permite acessar as alterações de mutações (em kcal/mol) na conformação, flexibilidade e dinâmica proteicas (RODRIGUES; PIRES; ASCHER, 2018).

Para utilização dos métodos preditivos baseados em estrutura, mCSM-Stability, mCSM-PPI, DUET e DynaMut, é necessário fornecer o arquivo contendo a estrutura tridimensional da proteína de interesse e outro arquivo com a lista de mutações *missense*. Da mesma forma, apresentam o resultado da predição como uma variação na energia livre de Gibbs (ΔΔG, em kcal/mol). A energia livre de Gibbs (ΔG) é a energia que pode ser utilizada para desempenhar trabalho em uma reação química. É uma quantificação termodinâmica equivalente a entalpia (H), de um sistema ou processo, menos o produto da entropia pela temperatura absoluta (TS). Cálculos de energia livre de Gibbs permitem determinar se uma reação será termodinamicamente favorável ou não. Dessa forma, ΔG equivale à energia livre utilizada para enovelamento e formação de complexos protéicos, calculada segundo a Equação 1, onde ΔH é

[18] http://biosig.unimelb.edu.au/dynamut/

a variação na entalpia e ΔS é a variação na entropia de um sistema ou processo (BERG; STRYER; TYMOCZKO, 2014).

Equação (1) $\Delta G = \Delta H - T\Delta S$

Dessa forma, o cálculo dos métodos *in silico* utilizados seguem a Equação 2, a variação desta energia (ΔΔG) dada a diferença entre energia livre de Gibbs da proteína selvagem e mutante. Assim o ΔΔG se traduz na diferença da estabilidade, afinidade ou flexibilidade, dado uma alteração de aminoácido (BERG; STRYER; TYMOCZKO, 2014; PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L., 2014)

Equação (2) $\Delta\Delta G = \Delta G_{mutante} - \Delta G_{selvagem}$

O valor ΔΔG será negativo quando a estrutura selvagem for mais estável e será positivo se a estrutura mutante o for. Indicando quando uma mutação contribuiu para o aumento (ΔΔG positivo) ou perda (ΔΔG negativo) da estabilidade.

4.5 Identificação de Relações entre Mutações *Missense* e Seus Potenciais Efeitos e Fenótipos da Doença

O teste estatístico de Mann-Whitney, foi utilizado para buscar relações entre os atributos gerados pelos preditores baseados em sequência e estrutura e os dados fenotípicos de idade de início, idade de óbito e tempo de sobrevivência de pacientes. A visualização desses dados foi feita por meio de *boxplots* (gerados com a biblioteca ggplot2[19] e apresentados na seção de resultados) para visualizar e comparar a distribuição dos dados fenotípicos apresentados na Tabela 2. Os *boxplots* fornecem um resumo visual do padrão de variação em um conjunto de dados, o que é especialmente útil quando a comparação é feita entre diferentes distribuições (BAKKER; BIEHLER; KONOLD).

Tabela 2 - Atributos Preditos e seus Pontos de Corte de Classificação para Análise de Mann-Whitney

Atributo	Ponto de Corte	Classificação
RSA	≥ 17	Exposto
	< 17	Enterrado

[19] https://cran.r-project.org/web/packages/ggplot2/index.html

Stability (SDM, DUET, mCSM-Stability, mCSM-PPI, Dynamut)	< - 0,5 or > 0,5	Alto efeito
	≥ -0,5 or ≤ 0,5	Baixo efeito
pph2 (HumDiv e HumVar)	< 0,5	Tolerada
	> 0,5	Não tolerada
Conservation (ConSurf)	≤ 0,000	Conservado
	> 0,000	Não conservado
Distance to Contact Interface	> 6Å	Distante
	≤ 6Å	Próximo
Distance to Metals	> 3Å	Distante
	≤ 3Å	Próximo
IUPRED (Short, Long)	> 0,5	Desordenado
	≤ 0,5	Ordenado
MTR	≥ 0,8	Tolerante
	< 0,8	Não tolerante
SIFT	≤ 0,06	Não tolerada
	> 0,06	Tolerada
PROVEAN	≤ -2,5	Não tolerada
	> -2,5	Tolerada

Cada atributo disposto nesta tabela foi classificado conforme pontos de corte descritos na literatura, de modo a separar os dados em duas classes para análise estatística contra os dados fenotípicos de idade de início da doença, idade de morte e tempo de sobrevivência dos pacientes.

4.6 Desenvolvimento de Modelos para Predição de Fenótipos da ALS

Os modelos preditivos foram construídos utilizando a plataforma weka[20] para aprendizado de máquina e mineração de dados. O weka disponibiliza diversos algoritmos de aprendizado supervisionado e, dentre eles, o algoritmo de árvores de regressão M5P, o qual foi utilizado para construção dos modelos preditivos (QUINLAN, 1992), por ser um algoritmo simples, de rápida performance e intuitivo, devido ao método de árvores de regressão. A geração de árvores de regressão torna intuitiva a interpretação do modelo preditivo gerado e a importância dos atributos nele utilizados.

5 Resultados e Discussão

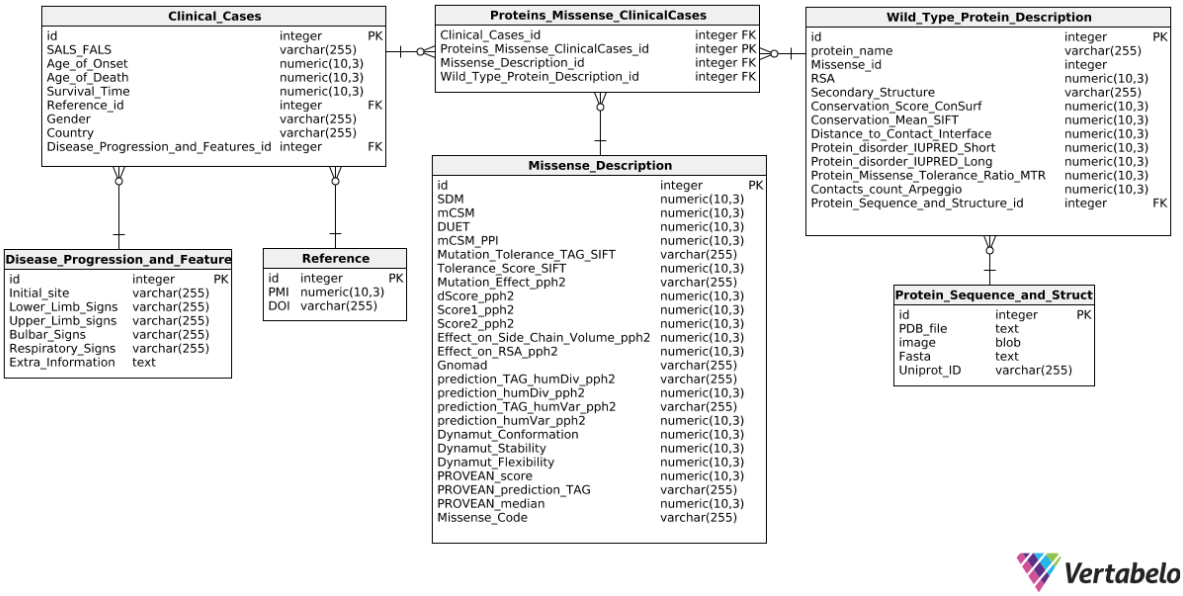
Os resultados obtidos até o momento correspondem majoritariamente à dados da SOD1, visto que as estruturas da TDP-43 e FUS/TLS requerem mais estudos para otimização da qualidade. Além disso, o maior número de casos de ALS está associado à mutações na SOD1, havendo portando maiores informações disponíveis.

5.1 DynAMISM: Dados Clínicos e Mutações *Missense* Associadas à ALS

A base de dados desenvolvida segue o modelo da Figura 4, de modo a reunir harmonicamente todas as informações coletadas na literatura e geradas por preditores. Dividido em sete seções, o modelo busca proporcionar a geração futura de uma interface web organizada e de fácil acesso, proporcionando praticidade ao usuário que a acessa em busca de dados relevantes e diferentes informações acerca da ALS, seja no âmbito puramente clínico ou em nível protéico.

[20] https://www.cs.waikato.ac.nz/ml/weka/

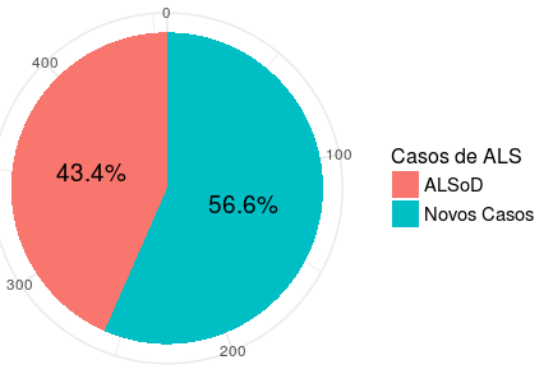
Figura 4 - Modelo da Base de Dados



Modelo da base de dados relacional DynAMISM. Modelagem representando a organização e disposição dos dados na base de dados DynAMISM. Disponível em https://github.com/AmandaAlbanaz/dynamism. Modelo gerado utilizando o software Vertabelo[21].

A coleta de dados clínicos de pacientes portadores da ALS proporcionou o desenvolvimento de uma nova base de dados, a DynAMISM. Esta, amplia o repertório de mutações documentadas em relação às bases relacionadas como a ALSoD e agrega atributos estruturais e de conservação que, por sua vez, podem auxiliar na elucidação de mecanismos moleculares da patogênese e da relação genótipo-fenótipo.

Figura 5 - Percentual de Novos Casos de ALS Associados à SOD1 na Base de Dados



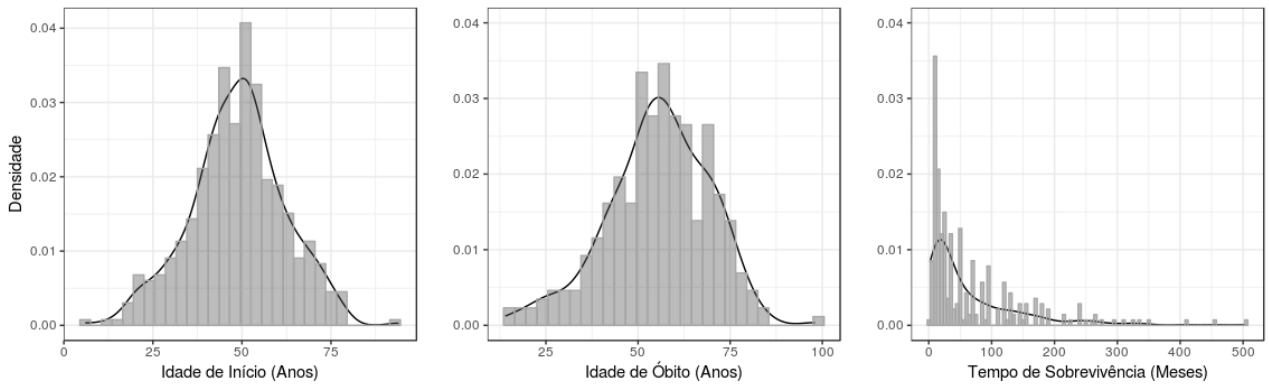
A nova base de dados DynAMISM compreende 456 casos de ALS associados à SOD1. Em comparação à ALSoD isso representa um aumento de 56,6% no número de casos (257 novos casos) e 199 já disponíveis na ALSoD.

[21] https://my.vertabelo.com

Até o momento da publicação deste trabalho, as informações contidas na DynAMISM se referem majoritariamente à SOD1. Optou-se por concentrar os resultados na SOD1 considerando a maior disponibilidade de dados da literatura e de estrutura 3D de qualidade para serem trabalhados – perante a carência de estrutura 3D e dados referentes às proteínas TDP-43 e FUS/TLS. Sendo assim, a DynAMISM possui mais que o dobro dos casos de ALS associados às mutações *missense* na SOD1, contando com 456 casos de mutações, das quais 257 são novos casos (Figura 5), um aumento de 56% quando comparada à ALSoD.

Os histogramas da Figura 6 mostram a frequência relativa de casos de acordo com a idade de início da doença, idade de óbito e sobrevivência dos pacientes. Como é possível observar, os dados coletados condizem com a estimativa descrita pela literatura. A maior parte dos pacientes descritos tiveram a doença iniciada entre 40 e 60 anos de idade, grande parte faleceu com idade entre 50 e 70, mostrando como o tempo de sobrevivência com a doença é curto, na maior parte dos casos o paciente sobrevive de poucos meses a 3 anos.

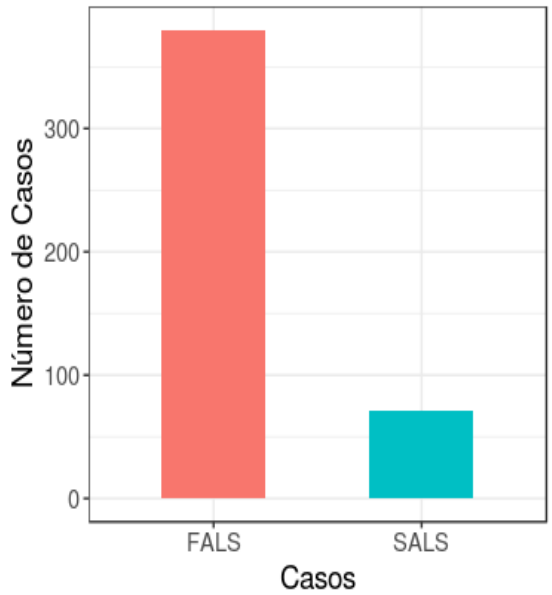
Figura 6- Frequência relativa das idades de início, óbito e tempo de sobrevivência de pacientes reportados na base de dados.



Histogramas de densidade mostrando a frequência relativa da idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes na base de dados DynAMISM.

De acordo com a literatura a maior parte dos casos de ALS são esporádicos, sem histórico familiar (Figura 2). Apesar disso, dos 456 casos de mutações *missense* descritos na SOD1, a maior parte possui histórico familiar (380 casos, 84,26%), sendo classificados como FALS (Figura 7). É possível que essa reversão de proporção tenha ocorrido devido à maior facilidade em constatar casos onde já existe um histórico familiar. Casos diagnosticados podem levar à investigação da família e ao longo do tempo facilitar a descoberta de mais casos, bem como a maior atenção quando já se conhece alguma ocorrência da doença na família.

Figura 7 - Percentual de casos de SALS e FALS na base de dados



Relação entre o número de casos de FALS e SALS na base de dados DynAMISM. FALS contabilizam 380 casos e SALS 71. Os 5 casos faltantes não foram classificados na literatura.

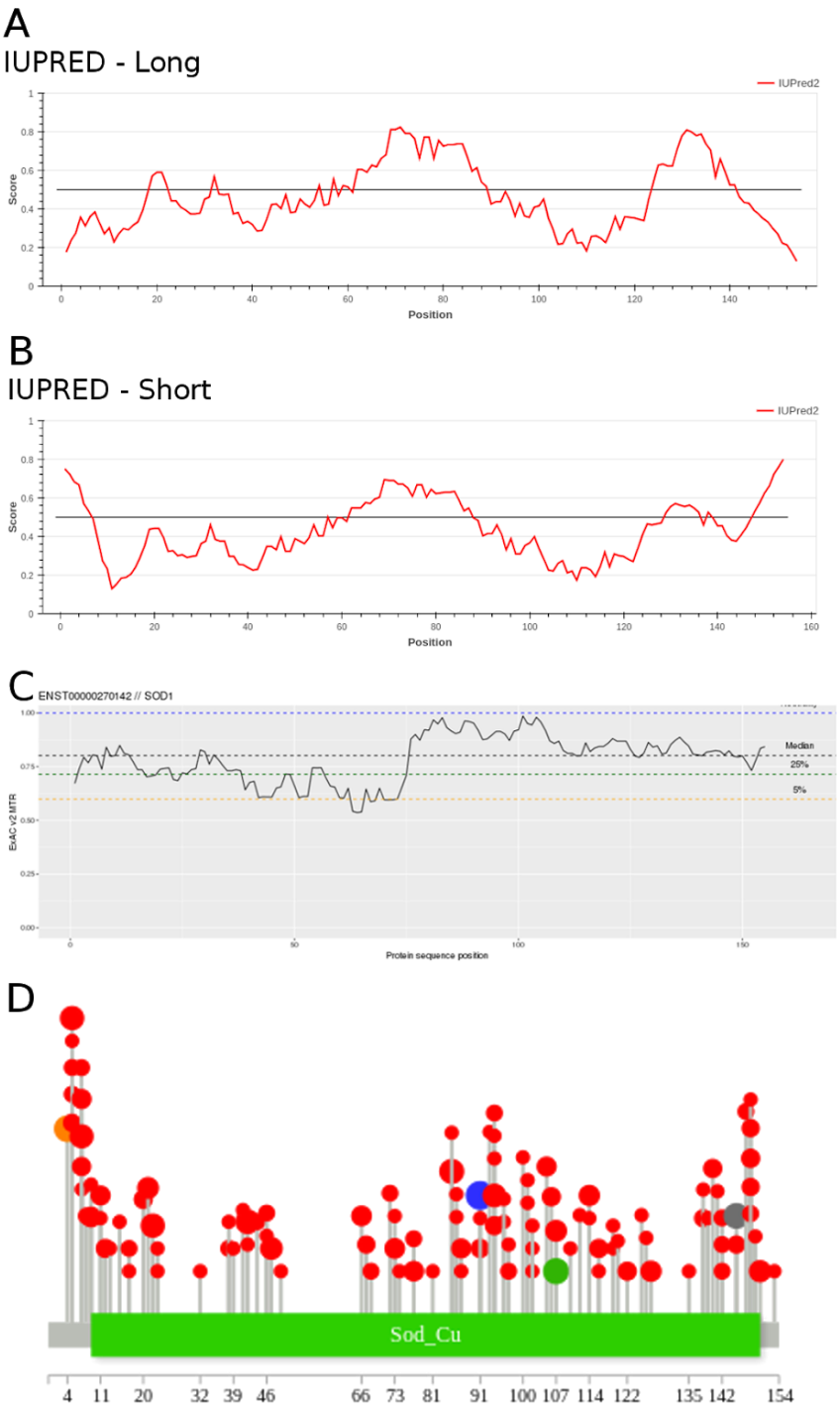
As Figura 8 a 11 mostram como as mutações reportadas na DynAMISM estão distribuídas ao longo da sequência da SOD1, TDP-43 e FUS/TLS respectivamente, bem como qual mutação predomina em cada posição, além da comparação com as regiões preditas como desordenadas e tolerantes às mutações *missense.*

Na Figura 9 é possível observar que na SOD1, as mutações predominantes são para os aminoácidos Alanina (A), Glutamato (E), Fenilalanina (F), Serina (S) e Valina (V) e majoritariamente nas posições 3 a 6, 90, 106 e 144, não conservadas na proteína conforme pode ser observado na Figura 14 e no alinhamento múltiplo de sequência[22].

A SOD1 é uma proteína predita, de modo geral, como tolerante a mutações *missense*, dado os valores de escore fornecidos pelo MTR e de estrutura moderadamente desordenada Figura 8. A tolerância à ocorrência de mutações *missense* pode estar relacionada às regiões de desordem de modo que estas tendem a ser menos conservadas que as estruturas secundárias de alfa hélice e folha-β, tolerando melhor a variabilidade de aminoácidos. Concomitantemente, como é mostrado na Figura 8C, a maior variabilidade de aminoácidos se concentra nessas regiões (Figura 8).

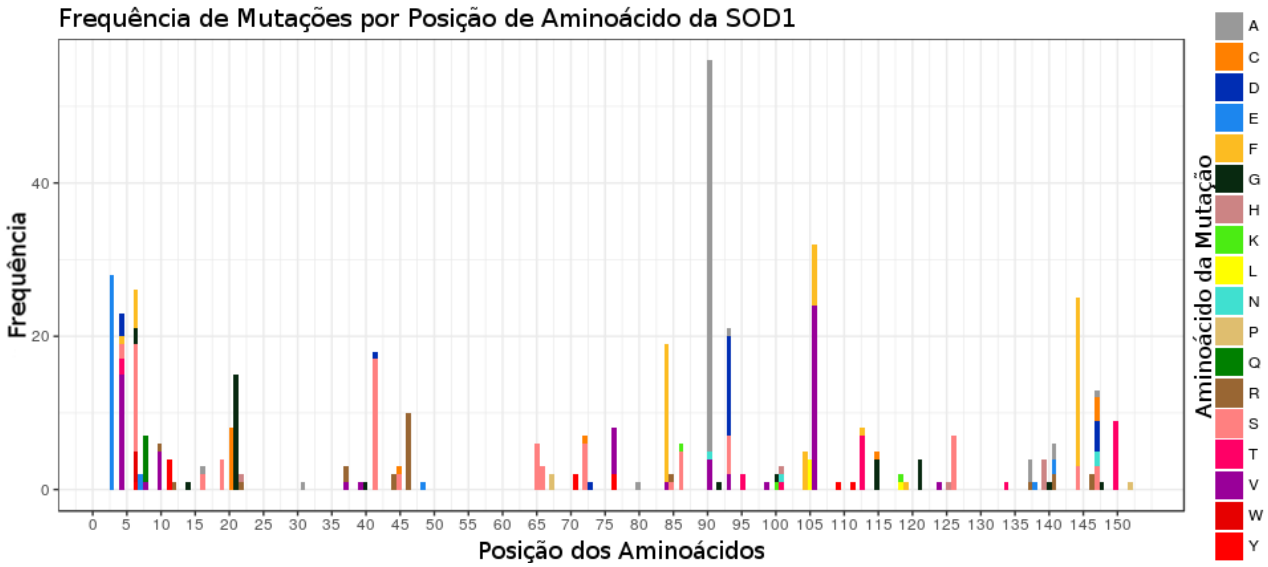
[22] https://drive.google.com/drive/u/0/folders/1-ttgdF-ct7A-ZhatTVDz-boS15OnNAZV

Figura 8 - Distribuição das mutações missense na SOD1



A linha central dos gráficos A e B representa o ponto de corte entre dados de regiões preditas como desordenadas e ordenadas. O eixo x representa a sequência de aminoácidos da SOD1 e o eixo y representa o escore de desordem. Portanto, a linha vermelha representa a variação de desordem estrutural. Quando esta linha se encontra acima da linha reta preta as regiões da proteína são preditas como desordenadas e, quando abaixo da linha preta, são ordenadas. O gráfico C representa a predição da taxa média de tolerância à mutações *missense* e a figura D representa a distribuição das mutações ao longo da SOD1. Em destaque estão as mutações predominantes na SOD1, K3E em laranja, D90A em azul, L106F em verde e L144F em cinza.

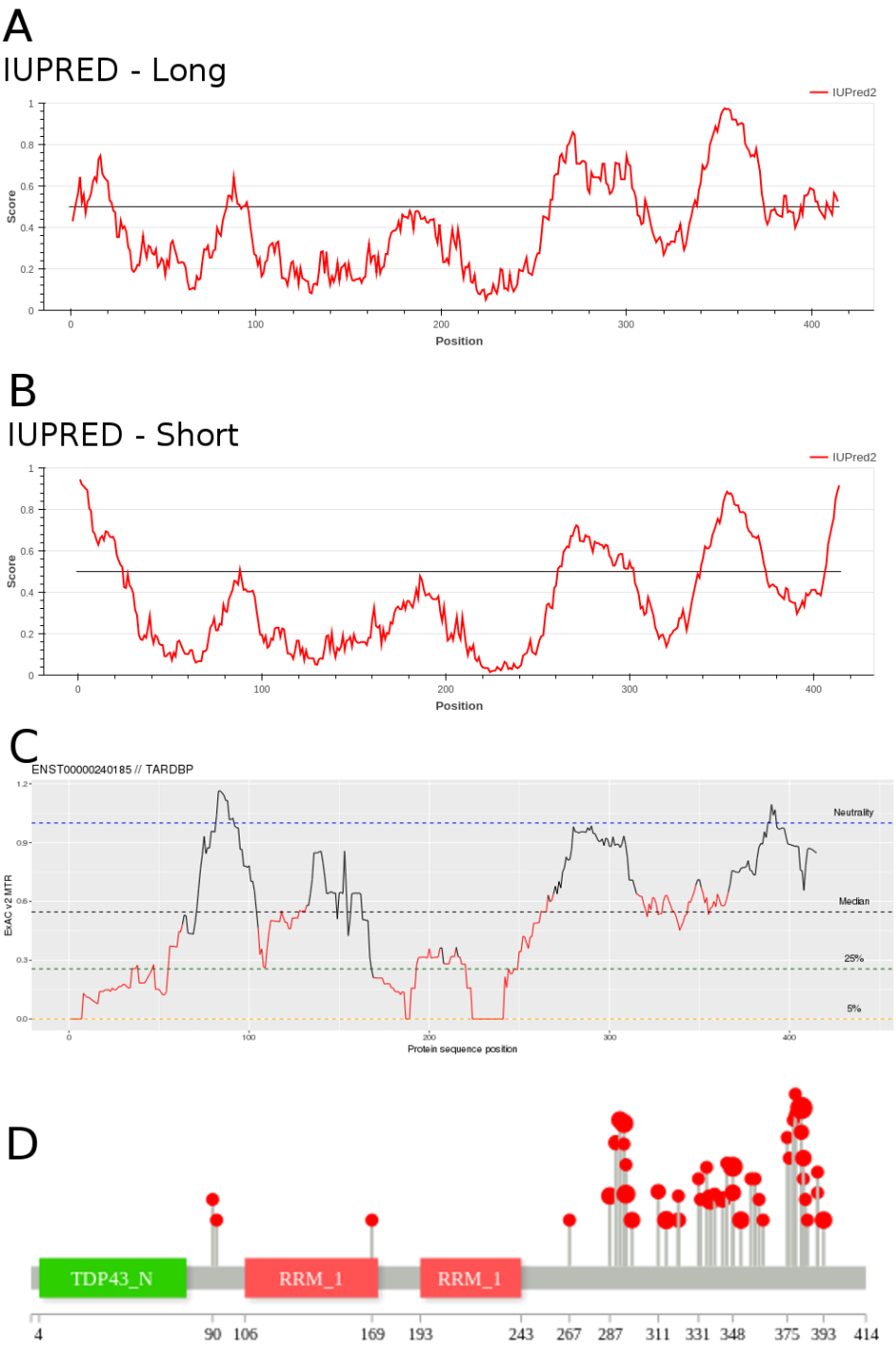
Figura 9 – Frequência de Mutações por Posição de Aminoácidos na Sequência da SOD1



Frequência de mutações *missense* por posição da sequência de aminoácidos da SOD1. Cada aminoácido resultante de mutação está representado pelo código de uma letra e por diferentes cores: A (alanina, cinza), C (cisteína, laranja), D (aspartato, azul escuro), E (glutamato, azul claro), F (fenilalanina, laranja claro), G (glicina, preto), H (histidine, marrom claro), K (lisina, verde claro), L (leucina, amarelo), N (asparagine, azul céu), P (prolina, bege), Q (glutamine, verde escuro), R (arginina, marrom), S (serina, rosa claro), T (treonina, rosa), V (valina, roxo), W (triptofano, vermelho escuro), Y (tirosina, Vermelho). O eixo X do gráfico se refere a cada posição da sequência e o eixo Y, se refere a quantidade de mutações em cada posição.

A discordância na predição de desordem entre os métodos *Long* e *Short* do IUPRED (Figura 8 A e B) se deve à especificidade e precisão de cada um, sendo que o método *Short* o mais preciso, por ser contexto-dependente.

Figura 10 - Distribuição das mutações missense na TDP-43

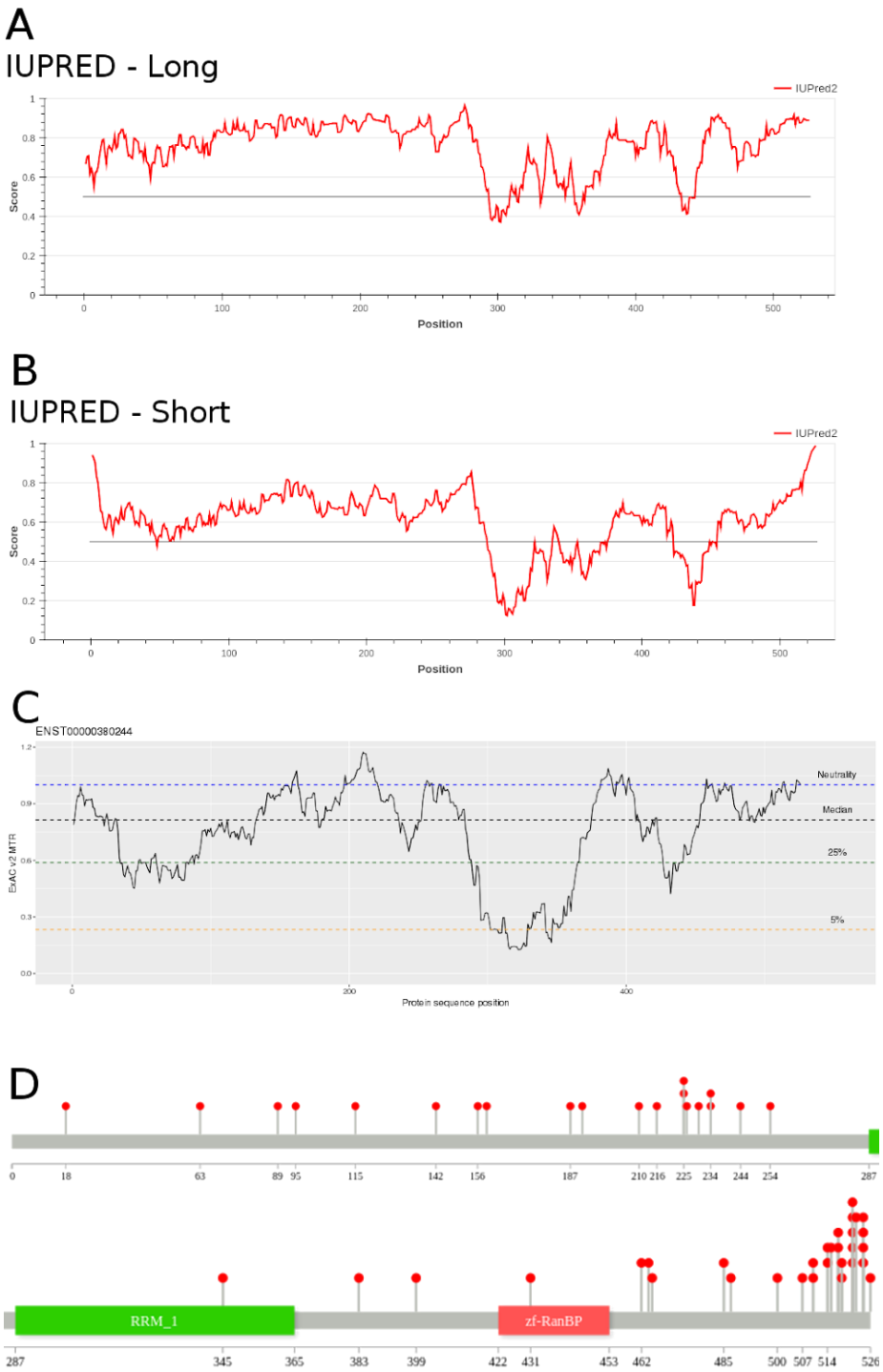


A linha central dos gráficos A e B representa o ponto de corte entre dados de regiões preditas como desordenadas e ordenadas. O eixo x representa a sequência de aminoácidos da TDP-43 e o eixo y representa o escore de desordem. Portanto, a linha vermelha representa a variação de desordem estrutural. Quando esta linha se encontra acima da linha reta preta as regiões da proteína são preditas como desordenadas e, quando abaixo da linha preta, são ordenadas. O gráfico C representa a predição da taxa média de tolerância a mutações *missense* e a figura D representa a distribuição das mutações ao longo da TDP-43.

A proteína TDP-43 possui regiões de alta intolerância a mutações *missense*, conforme predição do MTR (Figura 10), correspondendo muitas vezes às regiões de maior ordenação estrutural. A FUS/TLS é bastante tolerante a ocorrência de mutações missense e possui uma região intolerante entre os resíduos 350 e 400 que compreende uma das regiões ordenadas da estrutura e a maior parte do motivo de reconhecimento de RNA (do inglês, *RNA recognition motif* (RRM)). Nesta região intolerante apenas 3 mutações são reportadas. O mesmo ocorre com TDP-43, onde as regiões intolerantes a mutações coincidem com os RRMs, bem como coincidem com regiões ordenadas.

Ao comparar as características proteicas referentes à ordem/desordem estrutural, tolerância/intolerância às mutações e conservação da sequência de aminoácidos, é possível delinear importantes relações. Frequentemente a intolerância de mutações *missense* predomina em regiões ordenação estrutural devido à maior especificidade de aminoácidos que determinam dada estrtura e por sua vez, devido à essa maior especificação de aminoácidos, essas regiões tendem a se manter conservadas durante o processo evolutivo. A intolerância a mutações em regiões de desordem estrutural, como alças, pode seguir o mesmo preceito em algumas proteínas onde essas regiões são cruciais à função exercida, como por exemplo, as alças próximas aos íons de cobre e zinco na estrutura da SOD1 (Figura 14).

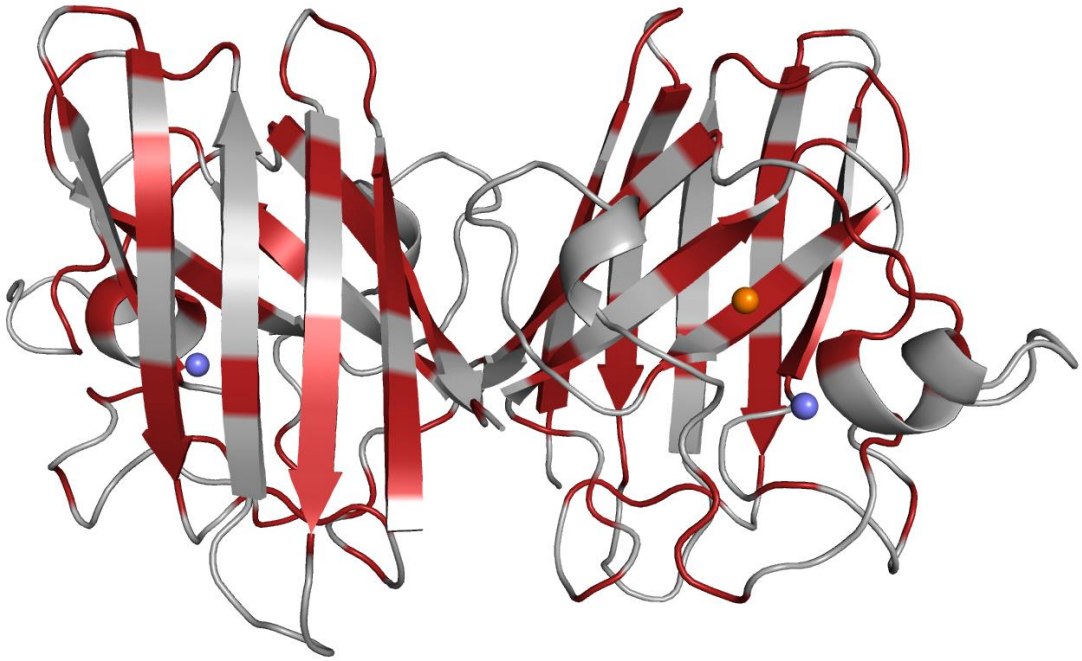
Figura 11 - Distribuição das mutações missense na FUS/TLS



A linha central dos gráficos A e B representa o ponto de corte entre dados de regiões preditas como desordenadas e ordenadas. O eixo x representa a sequência de aminoácidos da FUS/TLS e o eixo y representa o escore de desordem. Portanto, a linha vermelha representa a variação de desordem estrutural. Quando esta linha se encontra acima da linha reta preta as regiões da proteína são preditas como desordenadas e, quando abaixo da linha preta, são ordenadas. O gráfico C representa a predição da taxa média de tolerância a mutações *missense* e a figura D representa a distribuição das mutações ao longo da FUS/TLS.

A Figura 12 mostra como essas mutações coletadas estão espalhadas por toda a estrutura da SOD1, especialmente nas folhas beta (estrutura predominante na SOD1) e em alças, estrutura de alta flexibilidade e importância funcional para SOD1. As mutações descritas na estrutura atingem, portanto, regiões importantes para a estabilidade da SOD1, para sua função catalítica e áreas de interação com outras moléculas, causando uma grande disfunção.

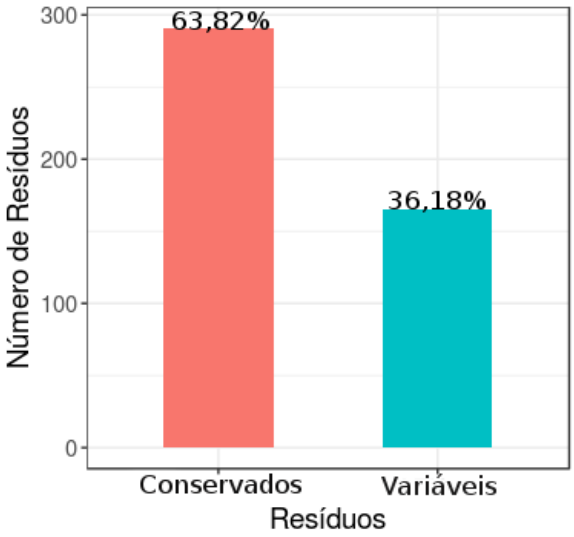
Figura 12 - Mapeamento estrutural das mutações missense na SOD1



Mapeamento das mutações *missense* na SOD1 reportadas na base de dados DynAMISM. As mutações estão destacadas em vermelho.

A grande maioria das mutações *missense* nos casos de ALS associados à SOD1 ocorrem em resíduos conservados, como é observado na Figura 8.

Figura 13 - Percentual de resíduos da SOD1 conservados e variáveis na base de dados



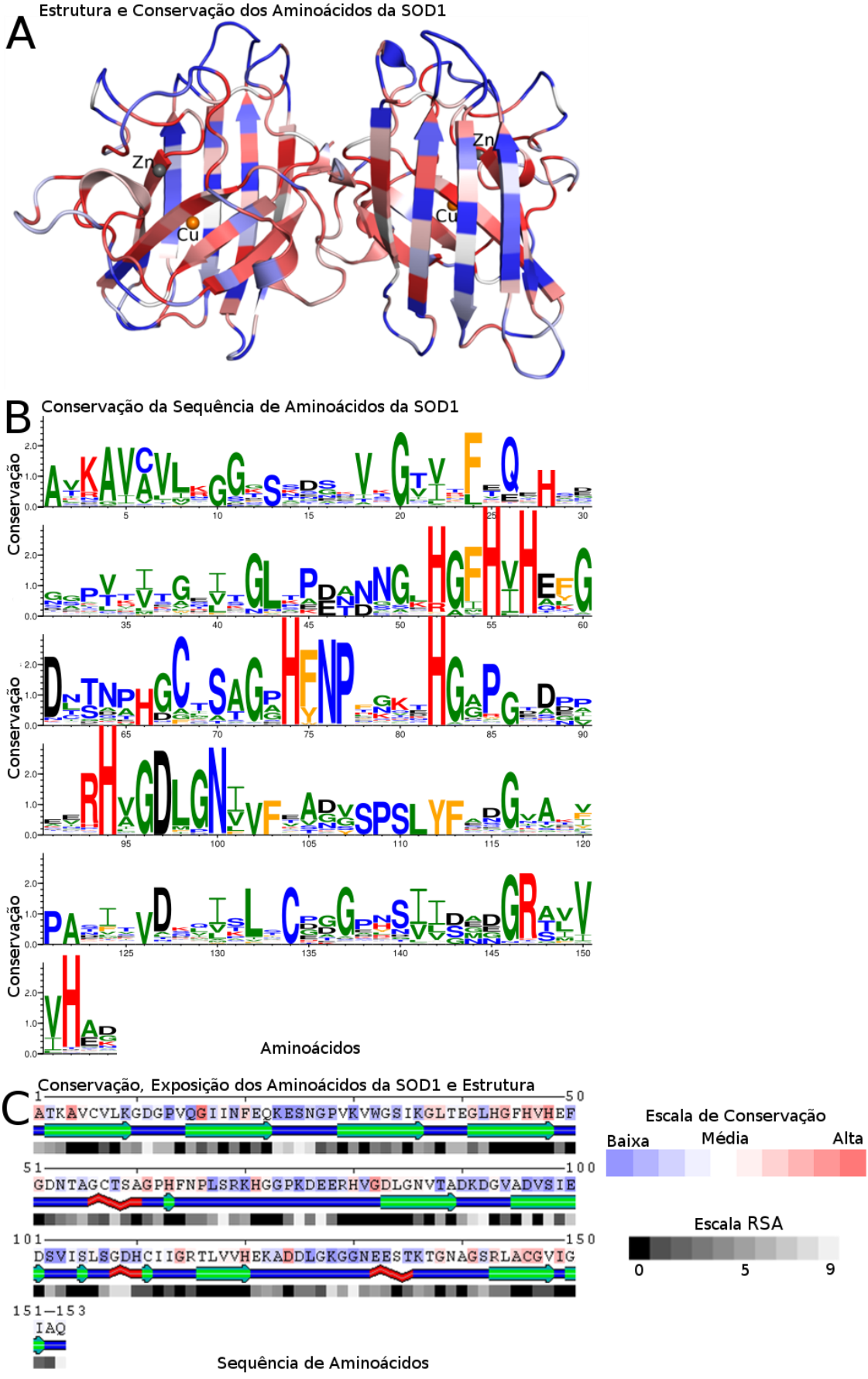
Relação entre a quantidade de mutações *missense* na SOD1 que ocorrem em resíduos conservados, preservados durante o processo evolutivo e em resíduos e não conservados.

A estrutura tridimensional da SOD1 (Figura 14) é predominantemente conservada na região de contato entre as cadeias da proteína e de interação com os metais cobre e zinco. Visualmente foi possível notar que as regiões de alça e possivelmente desordenadas, mais externas na estrutura são menos conservadas.

Pela figura também é possível notar que a maior parte dos resíduos conservados são histidinas que coordenam os metais (Figura 30). Adicionalmente, é possível reafirmar a tendência da conservação do núcleo protéico, especialmente a região de contato entre os monômeros (Figura 23) e a predominância de resíduos pouco conservados nas proximidades da superfície.

O núcleo estrutural, crucial à estabilidade e conformação, evolui lentamente de modo geral. Apesar da tendência à variabilidade, na superfície existem *hotspots* de resíduos conservados (ASHKENAZY et al., 2016).

Figura 14 - Conservação da SOD1



As figuras B e C foram feitas utilizando a ferramenta Polyview 3D[23]. A estrutura (A) foi colorida utilizando os escores calculados pelo ConSurf, sendo que o nível de conservação segue de azul escuro para resíduos variáveis, azul claro para resíduos pouco variáveis, vermelho claro para resíduos pouco conservados, até vermelho escuro para resíduos altamente conservados. Os íons de cobre e zinco são mostrados em esferas, na coloração laranja e cinza respectivamente.

5.1.1 Correlacionando Idade e Sobrevivência dos Pacientes com Características Estruturais e Impacto de Mutações *Missense* na SOD1

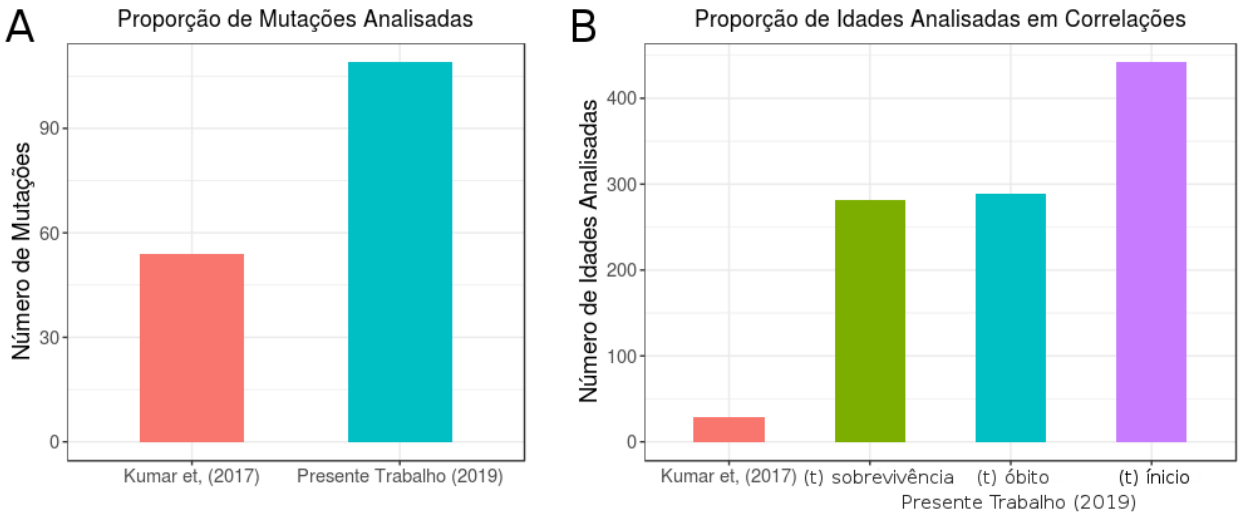
O trabalho de Kumar e colaboradores (KUMAR et al., 2017) procurou correlacionar a alteração da estabilidade estrutural sob o efeito de 30 mutações *missense* com dados fenotípicos de idade de início, óbito e tempo de sobrevivência de pacientes portadores de ALS. Desse modo obtiveram correlações moderadas (correlação de Pearson de até 0,4) entre o tempo de sobrevivência do portador da doença e alterações de estabilidade estrutural.

Embora seja um resultado animador, os dados representam uma taxa limitada da população portadora da ALS e esta, por ser uma complexa desordem, requer uma análise mais completa que envolva outros atributos além de estabilidade estrutural.

Neste trabalho procurou-se representar uma porção maior da população portadora da ALS, coletando todas as mutações descritas em pacientes, com os dados de idade e tempo de sobrevivência. A Figura 15A representa em número as mutações *missense* coletadas e as mutações únicas independente de casos. Foram coletadas mais que o dobro de mutações *missense* (107 mutações) que o trabalho anteriormente mencionado (50 mutações). Quando se compara o número de casos portadores dessas mutações (Figura 15B), o número total coletado por Kumar e colaboradores (2017) foi de 30, enquanto que no presente trabalho foram coletados 456 dados no total, sendo que alguns casos não possuem dados de tempo de sobrevivência, ou idade de início e óbito, portanto os números absolutos variam entre esses dados.

[23] http://polyview.cchmc.org/

Figura 15 - Proporção de mutações analisadas



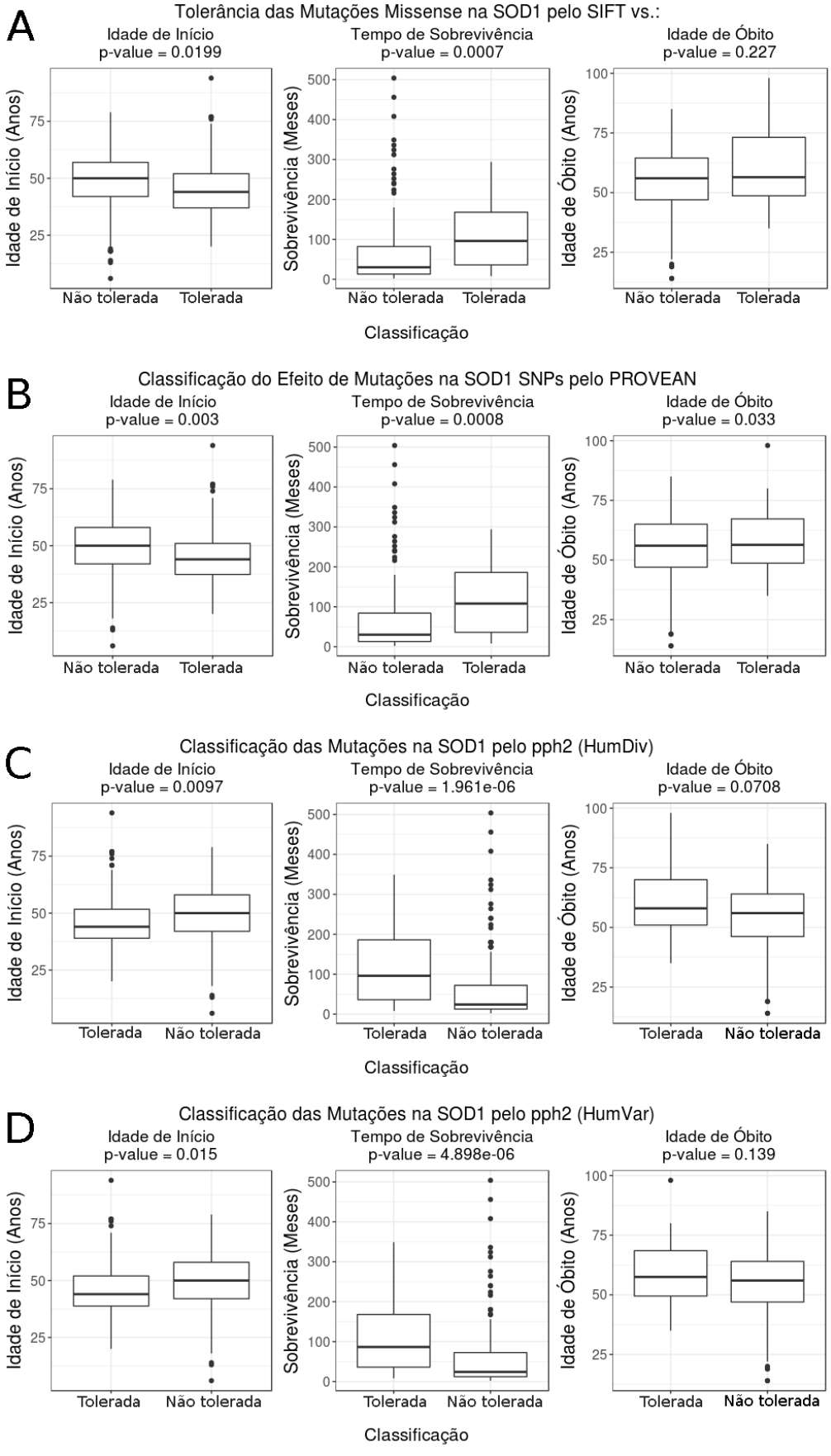
Analisou-se uma quantidade de mutações *missense* na SOD1 consideravelmente superior (107) mutações) ao que foi utilizado pelo trabalho de Kumar e colaboradores (2017) em sua correlação entre tempo de sobrevivência e alterações em estabilidade (50 mutações *missense*) (A). Ao comparar o número de casos de mutações *missense* com dados de idade de início, idade de óbito e tempo de sobrevivência, foram utilizados um total de 456 casos, sendo que deste total há 442 com dados de idade de início da doença, 289 com idade de óbito e 281 com dados de tempo de sobrevivência, enquanto o trabalho de Kumar e colaboradores (2017) utilizou 30 casos com todas essas informações. É importante ressaltar que o número de mutações e o número de idades analisadas não correspondem ao número de casos de pacientes portadores de mutações *missense*, o qual totaliza 456 casos.

Esses dados foram utilizados também para correlacionar com atributos, dados estruturais como acessibilidade relativa ao solvente (RSA), distância da interface de contato e dados acerca dos efeitos de mutações, analisados nos *boxplots* a seguir.

Analisando os *boxplots* e o p-valor gerado pelo teste estatístico U de Mann-Whitney, buscou-se encontrar atributos que podem estar correlacionados ao início da doença em idade mais jovem ou mais tardias, e correlacionados também ao maior ou menor tempo de sobrevivência e à morte precoce.

O escore calculado pelo SIFT (SIM et al., 2012) denota quando uma mutação é tolerada ou não pela proteína, considerando consequências funcionais. Os *boxplots* e os p-valores da Figura 16 permitem visualizar a relação existente entre a tolerância das mutações e a idade de início e tempo de sobrevivência dos pacientes, de modo que mutações preditas como não toleradas, pelo SIFT, estão correlacionadas ao início precoce da doença e ao menor tempo de sobrevivência.

Figura 16 - Relação entre a Classificação das mutações missense em análise quanto a tolerância e fenótipos de idade e sobrevivência

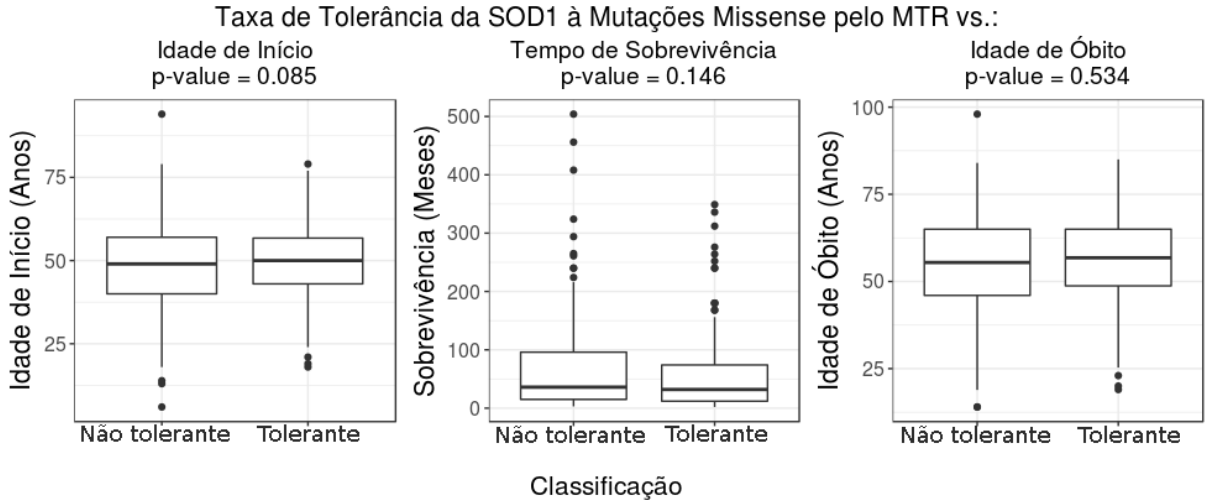


Boxplots mostrando relações entre predições de tolerância de mutações *missense* e dados fenotípicos de idade de início, óbito e tempo de sobrevivência para métodos baseados em sequência.

Adicionalmente ao escore do SIFT, foram calculados os escores do PROVEAN e para os modelos de classificação HumDiv e HumVar (Figura 16) do pph2. De acordo com essas predições, foi observada forte correlação entre a presença de mutações não toleradas e o início da ALS em idade mais jovem, bem como com o pouco tempo de sobrevivência. Contudo a relação dessas mutações com uma idade de morte mais jovem é observada apenas nas predições provenientes do PROVEAN.

Apesar da forte relação existente entre a tolerância de mutações e o tempo de sobrevivência de pacientes, não há associação com a taxa de tolerância a mutações no contexto geral da proteína (Figura 17). Isso possivelmente ocorre devido à relação direta da doença com mutações *missense* específicas e não com o quanto a estrutura suporta ou não mutações como um todo. Portanto, existem certas mutações que não se associam à fenótipos da ALS, mesmo quando ocorrem nessas regiões de baixa tolerância média. A taxa de tolerância a mutações *missense* calculada pelo MTR se refere ao contexto geral na proteína no sentido em que a taxa é calculada em relação a qualquer possibilidade de mutação *missense* segundo o código genético e não apenas àquelas descritas e analisadas neste presente trabalho.

Figura 17 - Relação entre a Tolerância da SOD1 às mutações *missense* e os fenótipos de idade e sobrevivência

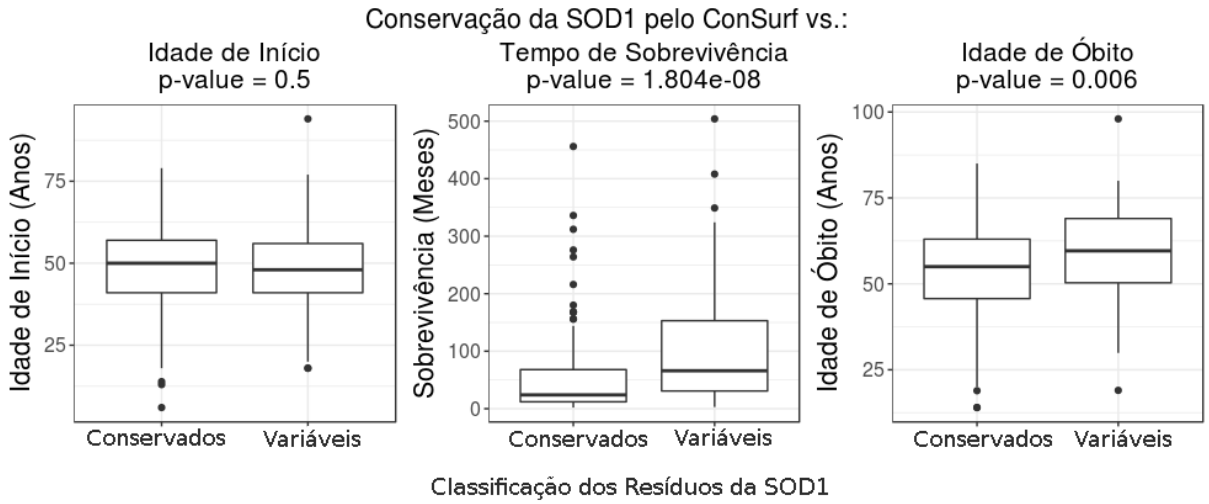


Boxplots relacionando os fenótipos de idade de início, óbito e tempo de sobrevivência com a taxa média de tolerância da SOD1 a mutações *missense*.

Mutações não toleradas e, portanto, prejudiciais à estrutura e função protéica tendem a ocorrer em regiões conservadas dentro da proteína (Figura 18). Considerando a alta importância

funcional dessas regiões conservadas é esperado que mutações que nela ocorram acarretem problemas funcionais e, portanto, estejam relacionadas à disfunções e mecanismos patogênicos. A relação entre mutações não toleradas e sua ocorrência em regiões conservadas pode ser observada na ALS, onde as mutações que ocorrem em regiões conservadas se relacionam ao baixo tempo de sobrevivência do paciente e sua morte precoce.

Figura 18 - Relação entre a conservação da SOD1 e fenótipos de idade e sobrevivência



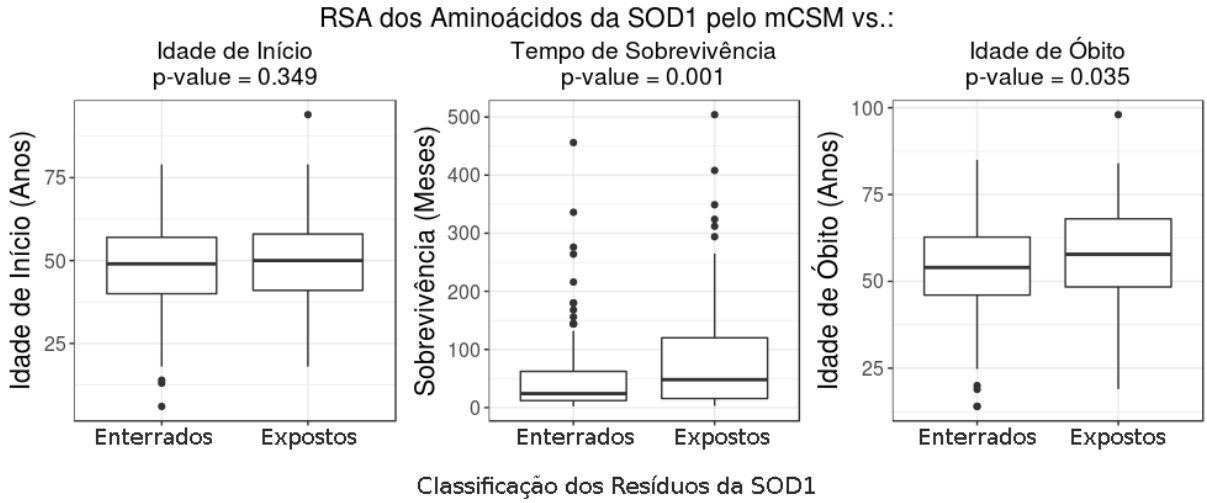
Boxplots relacionando dados fenotípicos de idade de início, óbito e tempo de sobrevivência com mutações reportadas em regiões conservadas (*conserved*) e não-conservadas (*variable*).

Seguindo este raciocínio, resíduos conservados possuem tendência em se localizar em regiões mais ao centro da estrutura e, portanto, estão menos acessíveis ao solvente. Sendo assim as mutações que ocorrem nessas regiões, também estão associadas ao menor tempo de sobrevivência e morte precoce (Figura 19). No caso da enzima SOD1, mutações ao centro da estrutura estão menos acessíveis ao solvente e na região de interface entre os monômeros, de modo que mutações nessa região afetam diretamente a conformação de dímero da enzima, importante também para sua funcionalidade. Sendo assim, mutações nessas regiões conservadas se relacionam ao menor tempo de sobrevivência (Figura 20). O que pode estar associado ao prejuízo funcional causado pela redução de capacidade de interação, que pode gerar uma cascata de disfunções. Dessa forma, é esperado que exista forte correlação entre mutações que afetam a interação proteína-proteína (Figura 21), outro potencial mecanismo molecular de patogenicidade em ALS. Contudo tal relação não é observada pelas predições obtidas pelo método mCSM-PPI, neste caso provavelmente devido à pouca quantidade de dados de interação proteína-proteína disponíveis para análise ou limitações do modelo preditivo. Apenas podem ser considerados como resíduos participantes da interação, àqueles que estão próximos da interface de contato (nesse caso foi utilizado o ponto de corte de 6Å), como representado na

estrutura da SOD1 (Figura 22), muito embora fatores alostéricos também possam interferir na afinifade proteína-proteína.

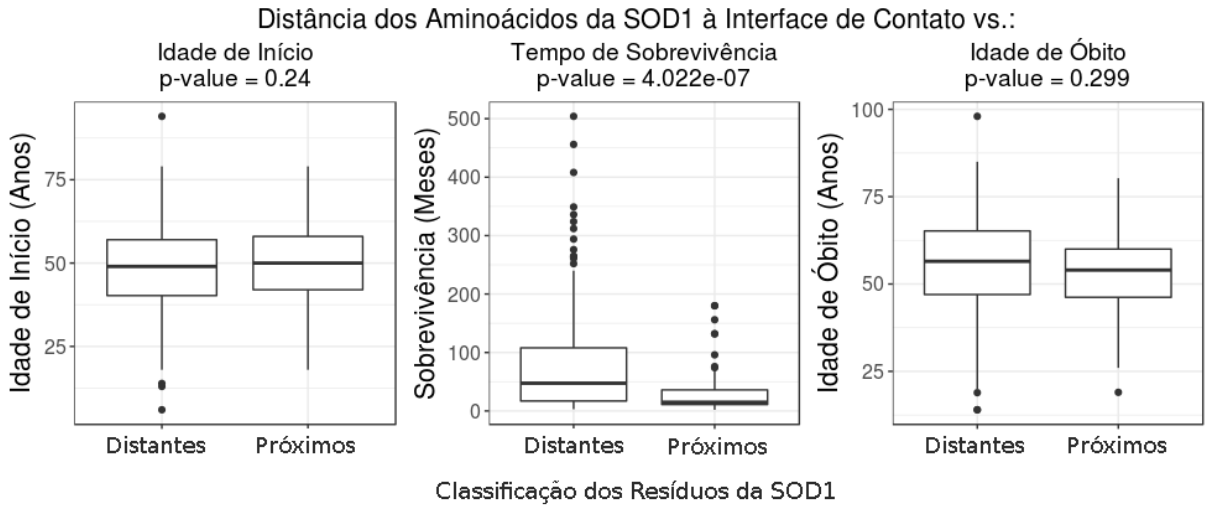
Embora não haja correlação evidente, pela análise dos *boxplots* é possível notar que as medianas sugerem uma possível tendência ao menor tempo de sobrevivência quando as mutações têm alto impacto na interação entre os monômeros da proteína e uma idade de morte mais avançada, podendo representar um início tardio da doença.

Figura 19 - Relação entre a acessibilidade relativa ao solvente e os fenótipos de idade e sobrevivência



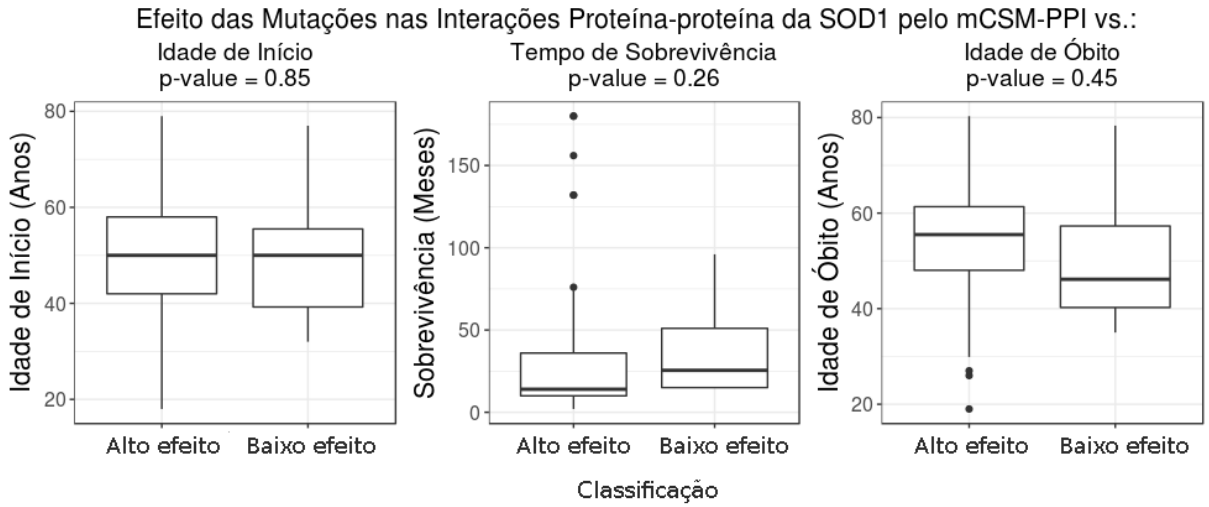
Boxplots relacionando dados fenotípicos de idade de início, óbito e tempo de sobrevivência com mutações reportadas em regiões expostas ao solvente (*Exposed*) e enterradas na estrutura (*Buried*).

Figura 20 - Relação entre a distância à interface de contato e fenótipos de idade e sobrevivência



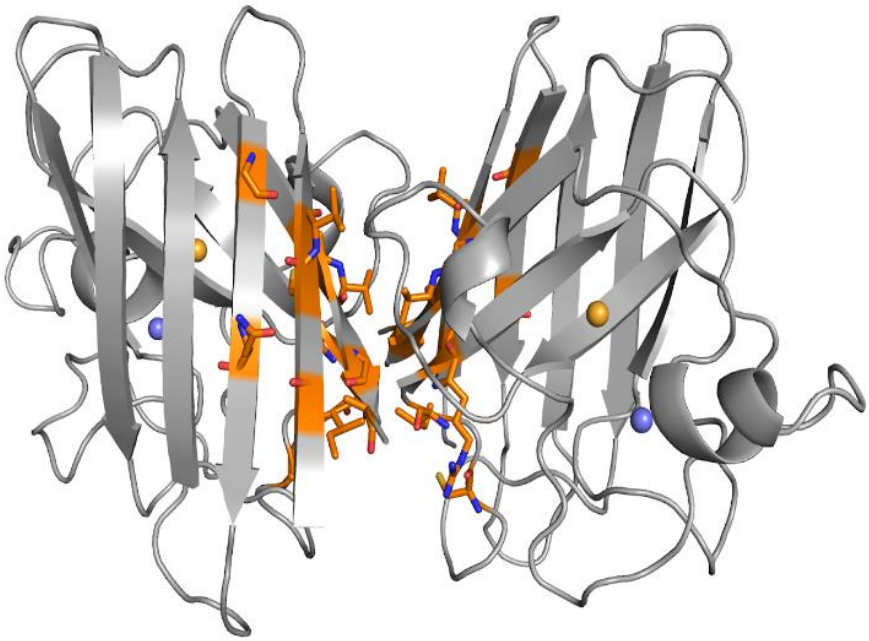
Boxplots relacionando dados fenotípicos de idade de início, óbito e tempo de sobrevivência com mutações reportadas em regiões distantes (*distant*) e próximas (*proximal*) à interface de contato.

Figura 21 - Relação entre: o efeito das mutações *missense* na interação entre os monômeros da SOD1 e os fenótipos da ALS relativos à idade e sobrevivência.



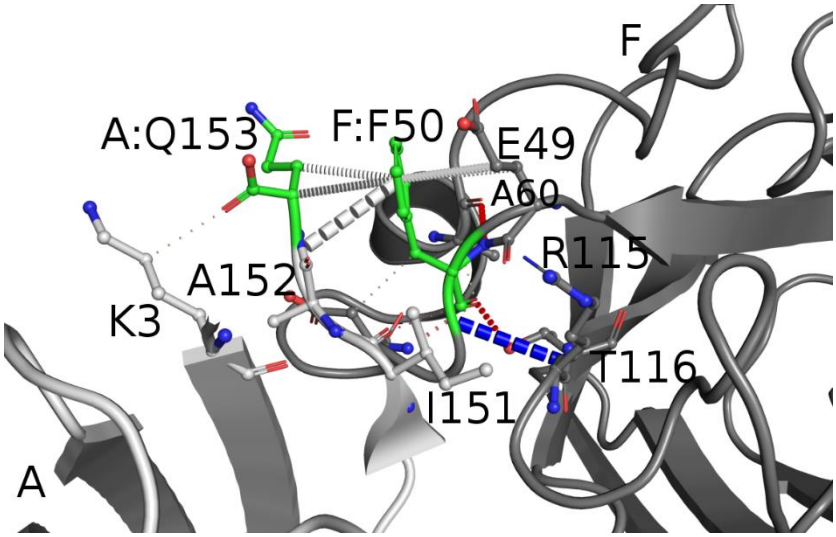
Boxplots relacionando dados fenotípicos de idade de início, óbito e tempo de sobrevivência com mutações que possuem alto efeito e baixo efeito nas interações entre os monômeros da SOD1.

Figura 22 - Resíduos na interface de contato entre os monômeros da SOD1



Destacados em laranja estão os resíduos dentro da distância máxima de 6Å da interface de contato entre os monômeros da SOD1, utilizados na análise estatística representada na (Figura 21). Este ponto de corte de distância contempla os resíduos que participam da interface de contato.

Figura 23 - Interações dos resíduos glutamina 153 e fenilalanina na interface de contato entre os monômeros da SOD1: posições onde mutações afetam a interação entre as cadeias

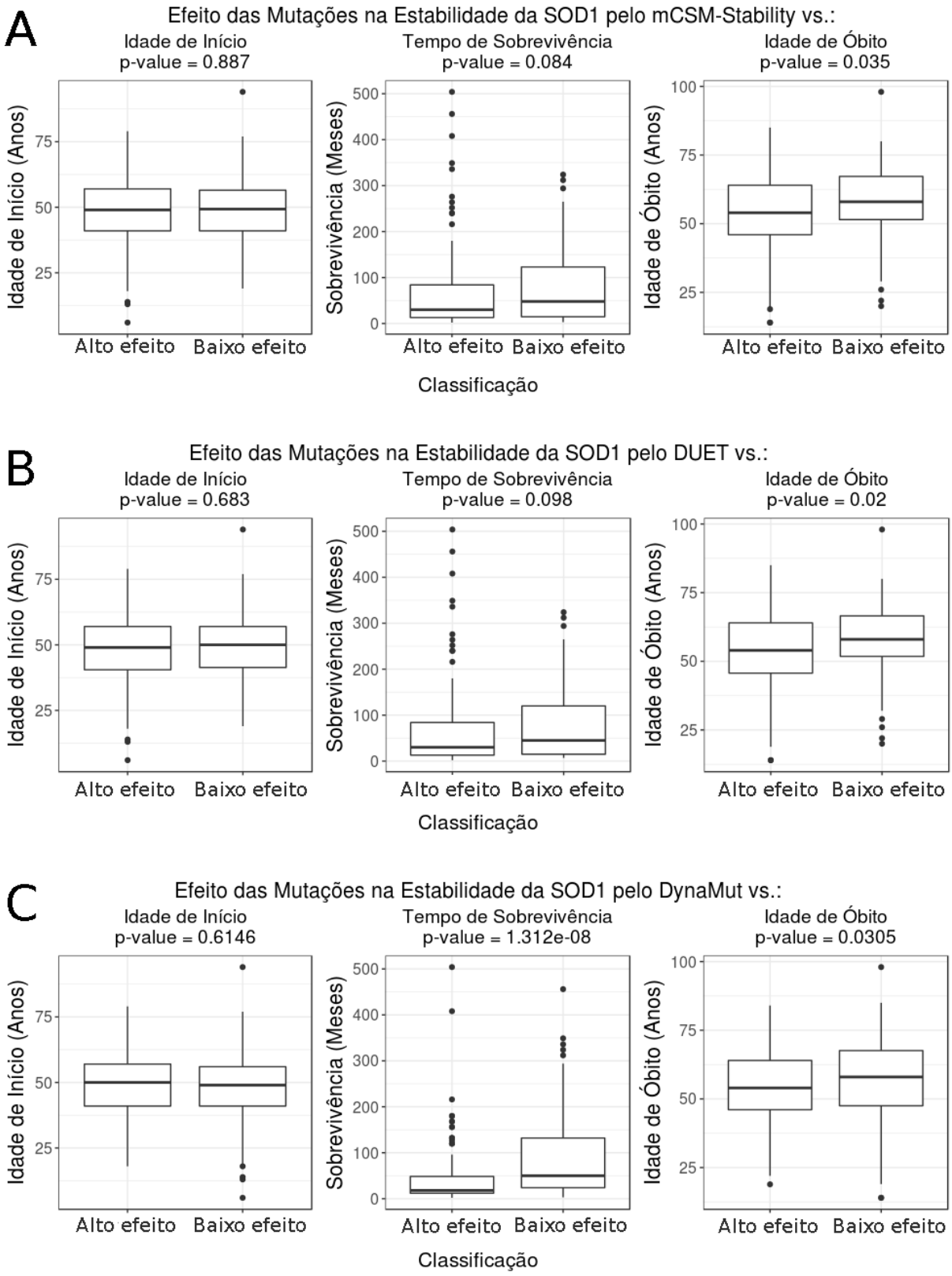


Representação da interação entre os monômeros da SOD1, como exemplo de uma região onde mutações afetam a interação proteína-proteína. Nesta representação está ilustrada a interação entre o resíduo.

A glutamina na posição 153 e fenilalanina 50 são exemplos de resíduos importantes para PPI, bem como os demais resíduos destacados na Figura 23. Mutações nessas posições podem afetar a formação do dímero

Dentro do contexto de estabilidade, é consenso entre os preditores do efeito de mutações em estabilidade (mCSM e DUET - (Figura 24)) que, mutações *missense* que tem alto efeito se relacionam fortemente com o fenótipo de morte precoce.

Figura 24 - Relação entre efeitos das mutações na estabilidade estrutural e os fenótipos de idade e sobrevivência



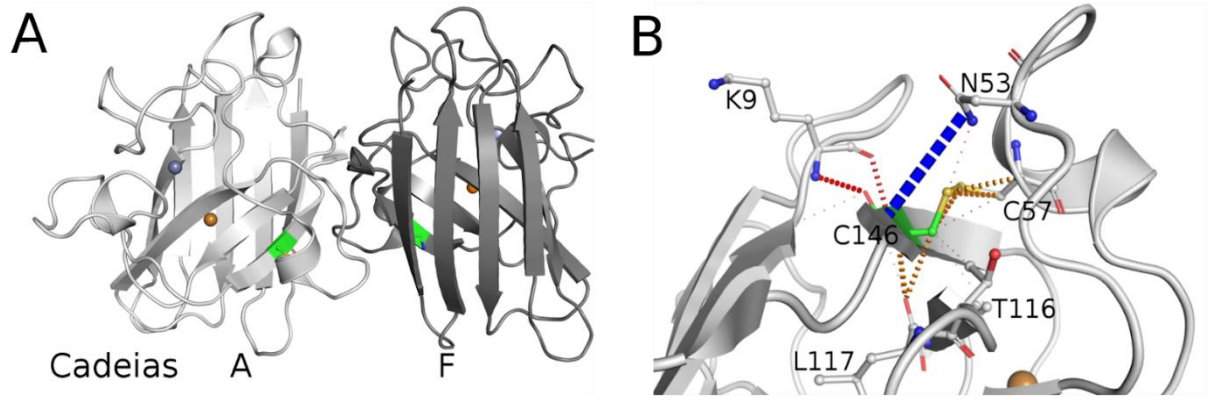
Boxplots relacionando as predições dos efeitos das mutações *missense* na estrutura da SOD1 com os fenótipos de idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes. É

importante enfatizar que a predição dos efeitos em estabilidade pelo DynAMISM (C) é uma combinação das predições dos efeitos das mutações na flexibilidade e conformação (Figura 26).

Outra importante interação para a estabilidade e funcionalidade da SOD1 são as pontes dissulfeto formadas pelas cisteínas na estrutura. A mutação C146R rompe essa ponte e influencia negativamente na estabilidade estrutural e habilidade de interação entre os monômeros apesar de não estar compreendida dentro do ponto de corte de proximidade da interface de contato.

Essa mutação reduz as interações de hidrogênio locais, alterando a conformação, estabilidade e flexibilidade protéica, causando desenovelamento (SRINIVASAN; RAJASEKARAN, 2017).

Figura 25 - Interação dos resíduos de cisteína 146 e 57 formando ponte dissulfeto



A ponte dissulfeto entre as cisteínas 57 e 146 em cada monômero é de extrem importância para a estabilidade da estrutura. É uma ponte fazendo ligação entre a folha β a pequena alfa hélice, possivelmente exercendo um importante papel estrutural que estabiliza a região de alças entre os monômeros (A). Estes resíduos estão acessíveis ao solvente e sustentam outras interações com resíduos em alça como visualizado na figura (B).

A mutação C146R (Figura 25) é predita como deletéria, corroborando com sua descrição na literatura (SRINIVASAN; RAJASEKARAN, 2017).

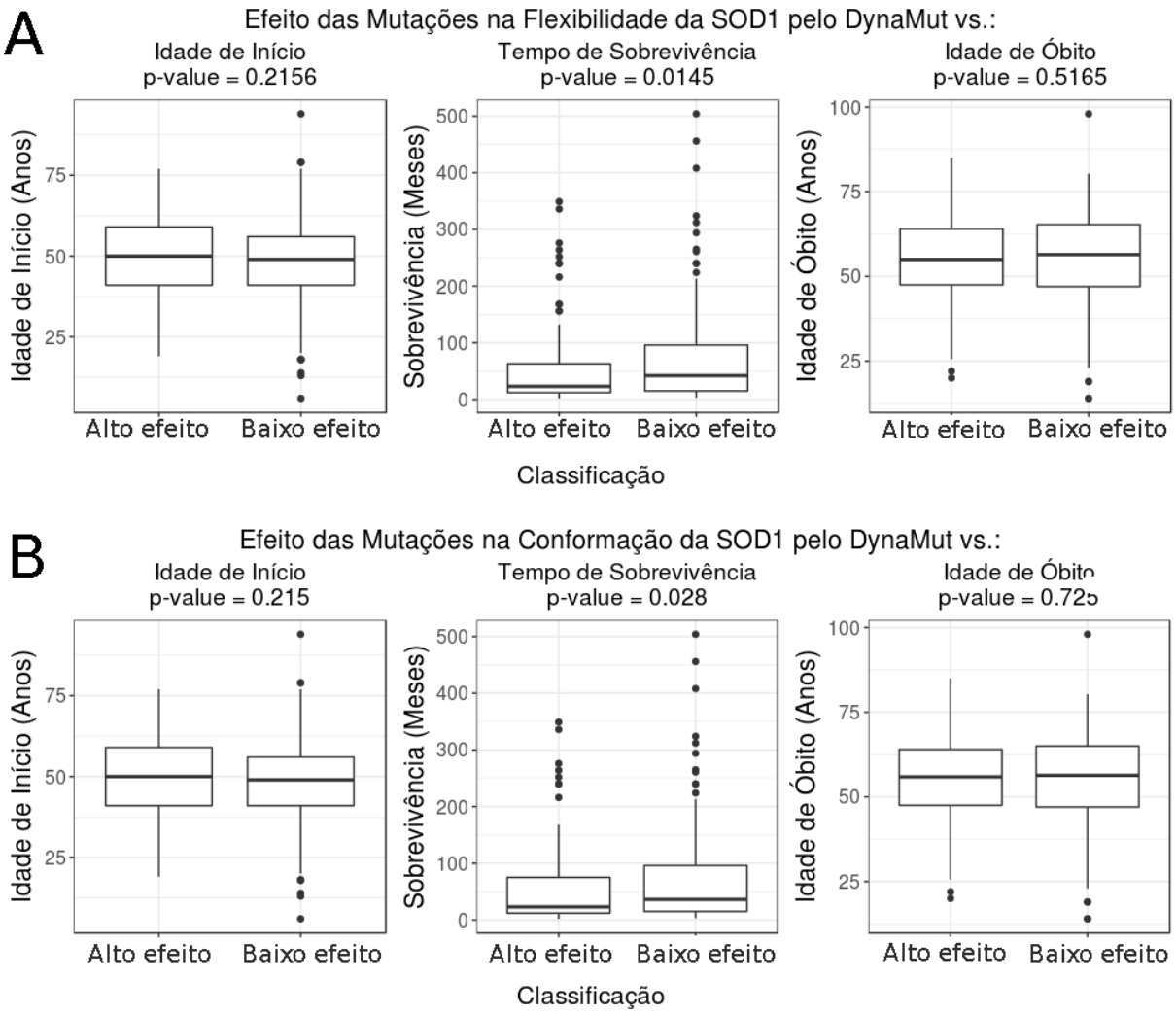
As mutações que têm grande efeito na flexibilidade e conformação da proteína se relacionam ao curto tempo de sobrevivência de pacientes com ALS.

O trabalho de Pereira e colaboradores (PEREIRA et al., 2019) utilizou análises de dinâmica molecular com a SOD3, indicando que a flexibilidade protéica pode ser aumentada por algumas mutações, enquanto a estabilidade em interações pode ser reduzida.

Isso implica no fato da flexibilidade estrutural interferir na habilidade de formação de interações mais estáveis. Avaliou-se que o alto efeito de mutações na flexibilidade, predito pelo método

DynaMut, se relaciona ao tempo de sobrevivência reduzido, contudo é necessário avaliar se esse efeito está em sua maioria, aumentando ou reduzindo a flexibilidade.

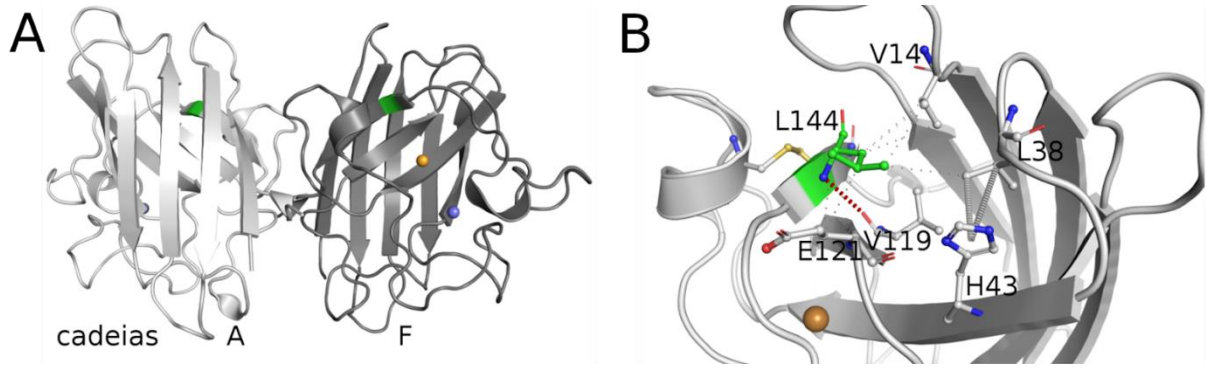
Figura 26 - Relação entre o efeito das mutações na flexibilidade e conformação da SOD1 e fenótipos de idade e sobrevivência



Boxplots relacionando as predições dos efeitos das mutações *missense* na flexibilidade (A) e conformação (B) da SOD1 com os fenótipos de idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes. Essas predições combinadas constituem a predição de alteração em estabilidade.

A estabilidade estrutural, flexibilidade e habilidade de conformação (Figura 26) podem estar relacionadas entre si e com os demais atributos. Se um resíduo, ou região, é de grande importância para essas três características, é provável que haja uma forte tendência que esses resíduos se mantenham conservados e estáveis durante o processo evolutivo.

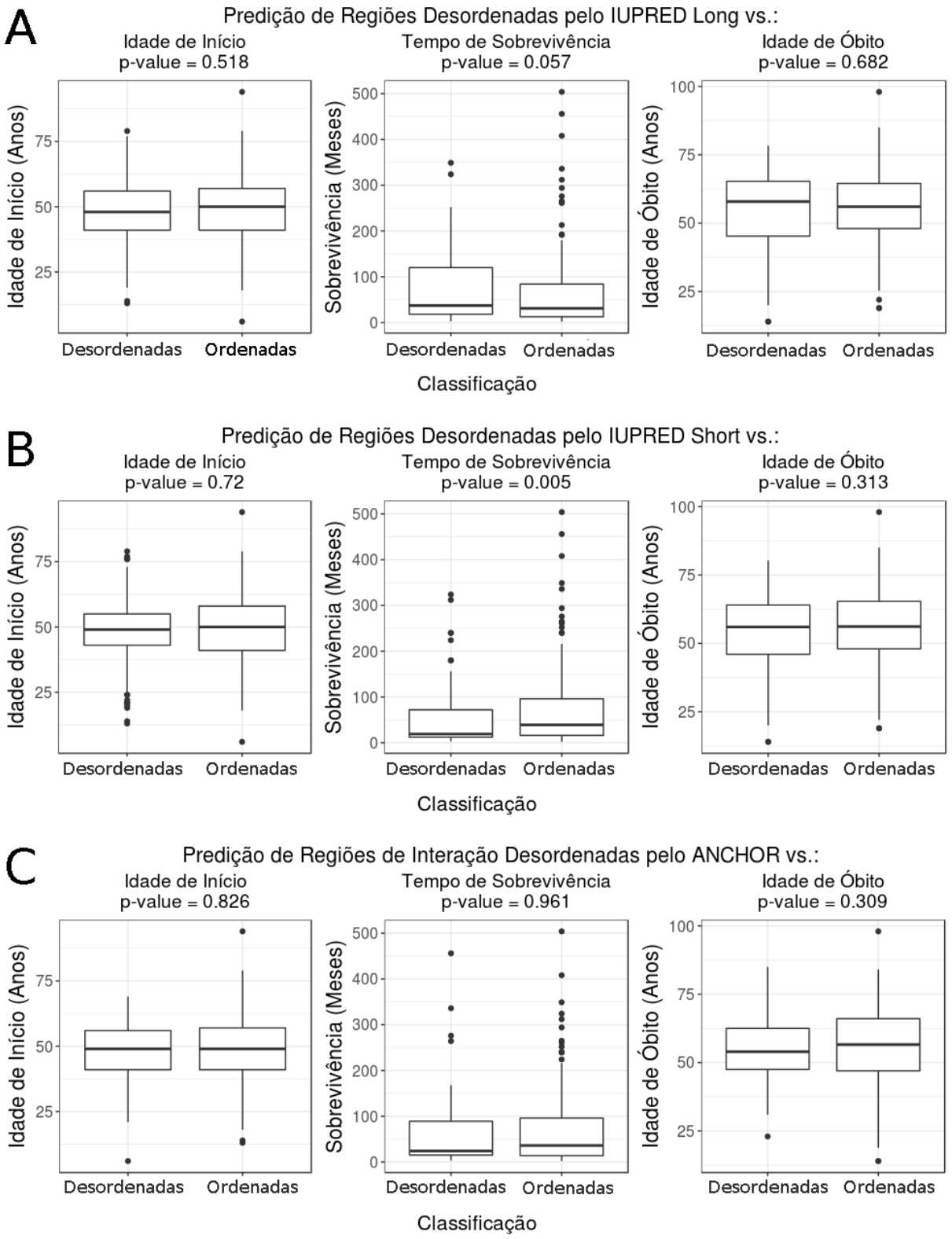
Figura 27 - Interações do Resíduo da Leucina 144



Localização estrutural da leucina 144 (A). Um resíduo pouco acessível ao solvente que interage com os resíduos de valina 14, leucina 38 e glutamato 121 por interações hidrofóbicas (em cinza claro) e uma forte interação de hidrogênio com a valina 119 (B, em vermelho).

A Figura 27 (A e B) representa o resíduo leucina 144, onde há uma mutação frequente para fenilalanina. Essa alteração que pode causar um choque estérico dentro da região de interação da proteína, devido ao aumento da cadeia lateral proporcionado pela inserção do anel aromático, o qual também interfere na flexibilidade, aumentando a rigidez e alterando a estabilidade da estrutura e, portanto, causando também alterações conformacionais. Essa mutação é de herança dominante, predita como deletéria por todos os métodos utilizados, embora os pacientes que a possuem, de modo geral, não tenham apresentado o início da ALS precocemente ou morte precoce. Senso assim, essa mutação pode estar associada à lenta progressão da doença.

Figura 28 - Relação entre regiões desordenadas e fenótipos de idade e sobrevivência



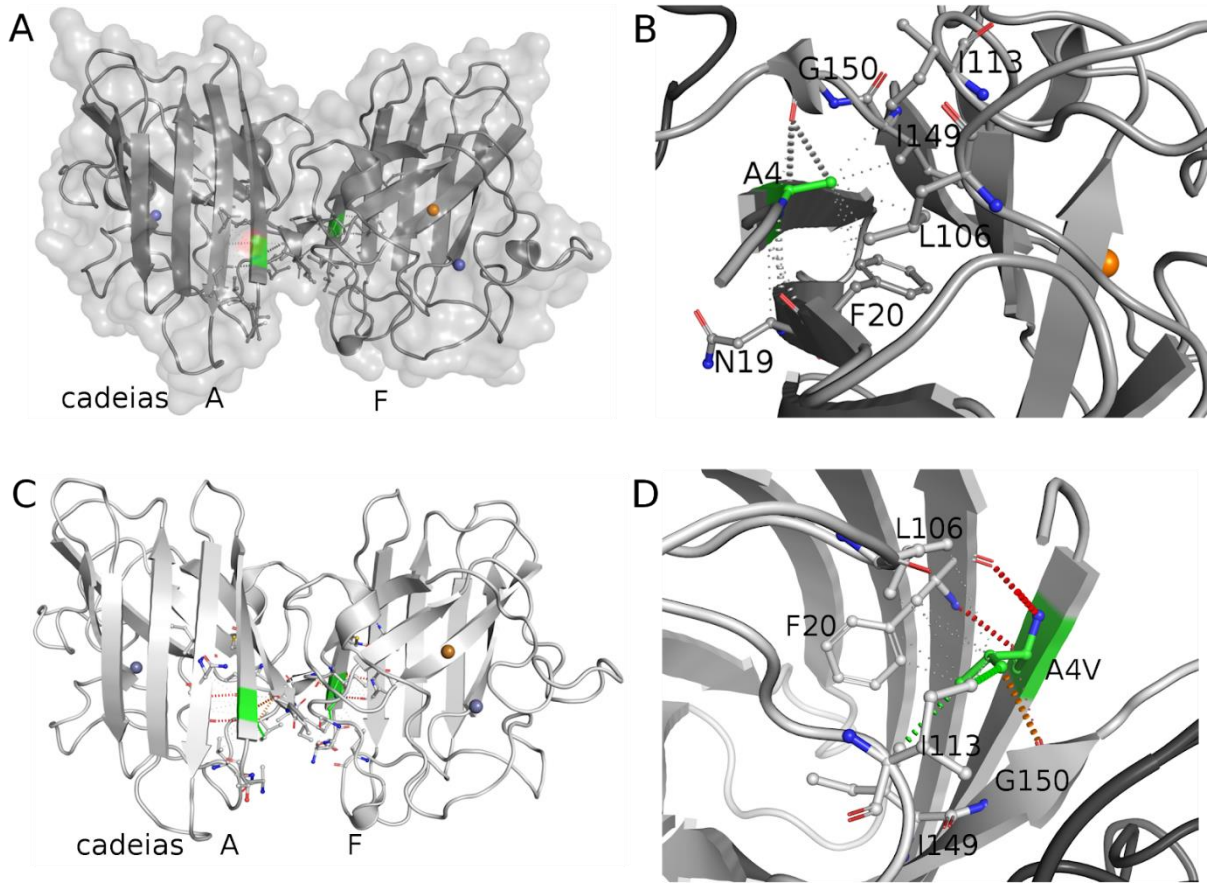
Boxplots relacionando as predições de regiões desordenadas (*disordered*) e ordenadas (*ordered*) da SOD1 com os fenótipos de idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes.

A via patogênica de enovelamento errôneo também pode ter contribuição da própria SOD1 nativa, a qual é classificada como moderadamente desordenada, considerando que cerca de 30% de seus resíduos são preditos como desordenados (SANTAMARIA et al., 2017).

Outra mutação importante e frequente mutação é a A4V (Figura 29) Causadora de um fenótipo agressivo de ALS, que geralmente leva o paciente a morte dentro de um ano após o início dos sintomas (TANG et al., 2018). Está situada em uma região ordenada (Figura 8) e suas predições indicam uma redução da flexibilidade, um aumento na estabilidade estrutural e conformação. Possui fenótipo deletério e características relacionadas à rápida progressão da doença, sinais de Hoffman, além de se relacionar a herança dominante autossomal[24].

A Figura 29 (A e B) representam a estrutura selvagem com a alanina 4, enquanto a Figura 29 (C e D) mostra uma valina na mesma posição. Nas figuras é possível notar a adição de fortes interações hidrofóbicas com a leucina 106, justificando o aumento da estabilidade anteriormente discutida.

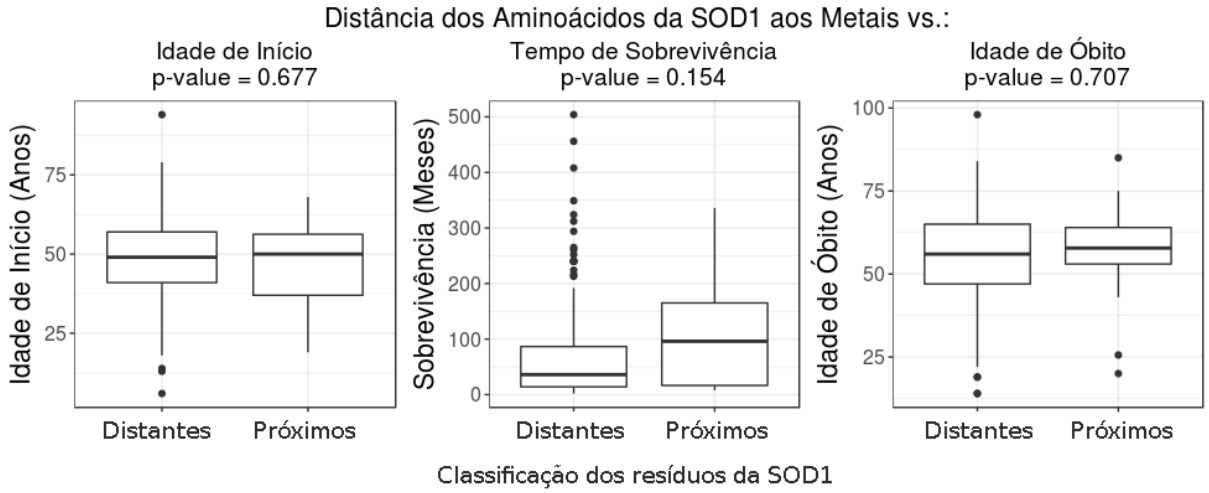
Figura 29 - Regão de Interação da AlaMutação A4V



[24] https://github.com/AmandaAlbanaz/dynamism

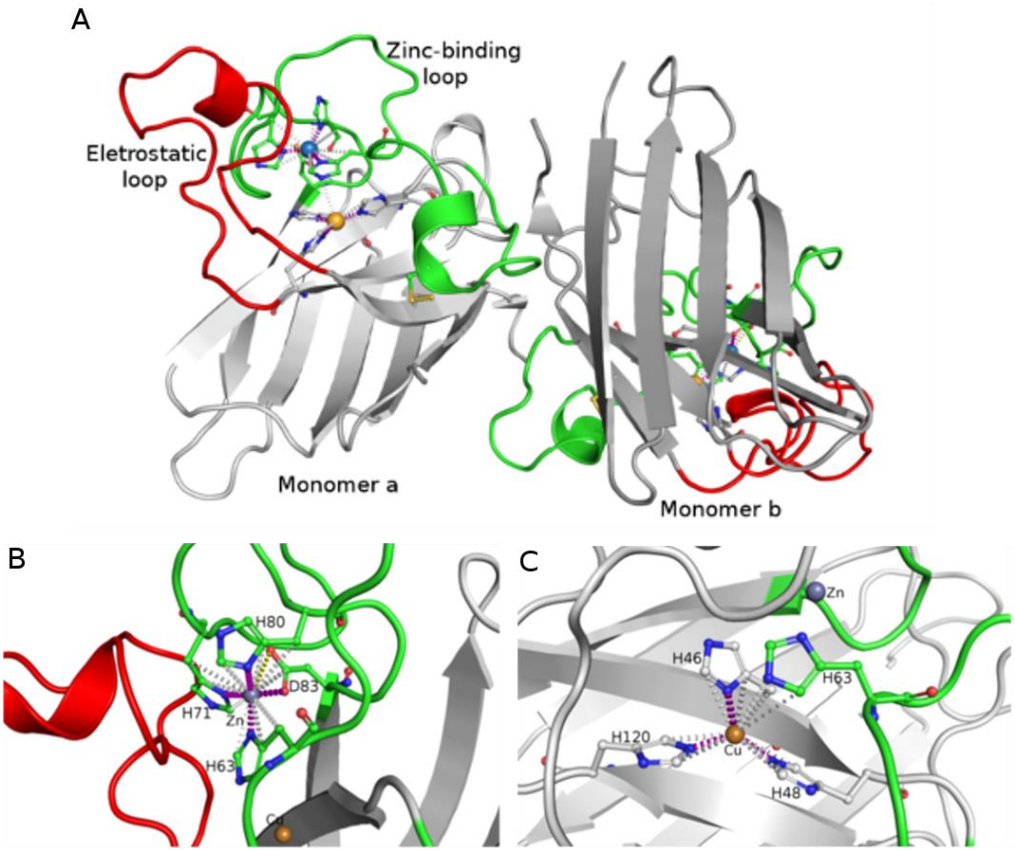
Representação da localização estrutural do resíduo alanina 4 (A e C). A alanina 4 interage com vários resíduos (B) e quando sofre mutação para uma valina (D) são formadas fortes interações de hidrogênio (D, em vermelho) com a leucina 106 e duas fortes interações hidrofóbicas com as isoleucinas 113 e 149 (em verde).

Figura 30 - Relação entre a distância dos aminoácidos da estrutura da SOD1 para metais de cobre e zinco e os fenótipos de idade e sobrevivência



Boxplots relacionando dados fenotípicos de idade de início, óbito e tempo de sobrevivência com mutações reportadas em regiões distantes (*distant*) e próximas (*proximal*) à interface de contato.

Figura 31 - Estrutura da SOD1: domínios e complexos metálicos



PDB 2C9V, (STRANGE et al., 2006). (A) Estrutura da SOD1 selvagem, com 153 aminoácidos em cada monômero. Estrutura bem organizada constituída por oito folhas-beta antiparalelas. Destacado em vermelho está o loop eletrostático, que inclui os resíduos das posições 121 a 142. Em verde está o loop de ligação com o zinco (esfera em azul (A)), constituído pelos resíduos 50 ao 83. A ponte dissulfeto, crucial para a função da SOD1, está colorida em laranja em cada monômero. (B) Representação das quatro histidinas (H80, H71 e H63) e do aspartato (D83), responsáveis pelo estabelecimento do complexo metálico com o íon de zinco (cinza (B), conexões metálicas representadas em roxo). As demais interações, apresentadas em cinza e amarelo, mostram interações de força de Van der Waals e uma conexão iônica, respectivamente. (C) O segundo complexo metálico apresentado é formado por interações entre um cobre e quatro histidinas (H120, H63, H48 e H46). Adicionalmente, são vistas em cinza as interações de força de Van der Waals.

O complexo com o cobre é o centro de catalítico da SOD1, enquanto o complexo com o zinco desempenha um importante papel estrutural no sítio ativo (KAWAMATA; MANFREDI, 2010). Neste contexto foi possível averiguar na Figura 31 importantes interações que podem ser alteradas por mutações e desequilibrar esses importantes domínios. Por exemplo, a mutação H80A, importante resíduo para coordenação dos metais, pode enfraquecer as interações de força de Van der Waals devido a perda da cadeia lateral. Esta é uma mutação relacionada a um fenótipo de início e morte precoce[25] e perda da coordenação do zinco.

A mutação H80 está em uma região altamente conservada e desestruturadas, como já mencionado anteriormente, regiões desestruturadas (alças) podem possuir funções específicas e cruciais e serem conservadas evolutivamente. Neste caso as alças contêm em sua formação as histidinas cruciais à coordenação dos íons de cobre e zinco, importantes para a função da SOD1. Sendo assim, a histidina 80 pertence à um domínio de alças funcionalmente importantes (Figura 8 e Figura ***31***).

Apesar da importância dos metais para a estrutura e funcionalidade da SOD1, não foram obtidas correlações envolvendo os resíduos dentro da distância de 3 Å destes íons e os fenótipos clínicos dos pacientes (Figura 30). Isso pode ter ocorrido devido ao pouco número de resíduos próximos a esses metais (14 resíduos com mutações reportadas), em comparação a todo o restante dos resíduos da proteína, considerando o ponto de corte para distância de 3Å.

5.1.2 TDP-43

A quantidade de mutações na TDP-43 reportada na literatura é menor em comparação à SOD1. Sendo obtidas 114 mutações *missense* dispostas na base de dados ALSoD.

No contexto de estudo estrutural das mutações são encontradas algumas limitações quanto a TDP-43 e a FUS/TLS. A principal limitação é a estrutura tridimensional destas proteínas,

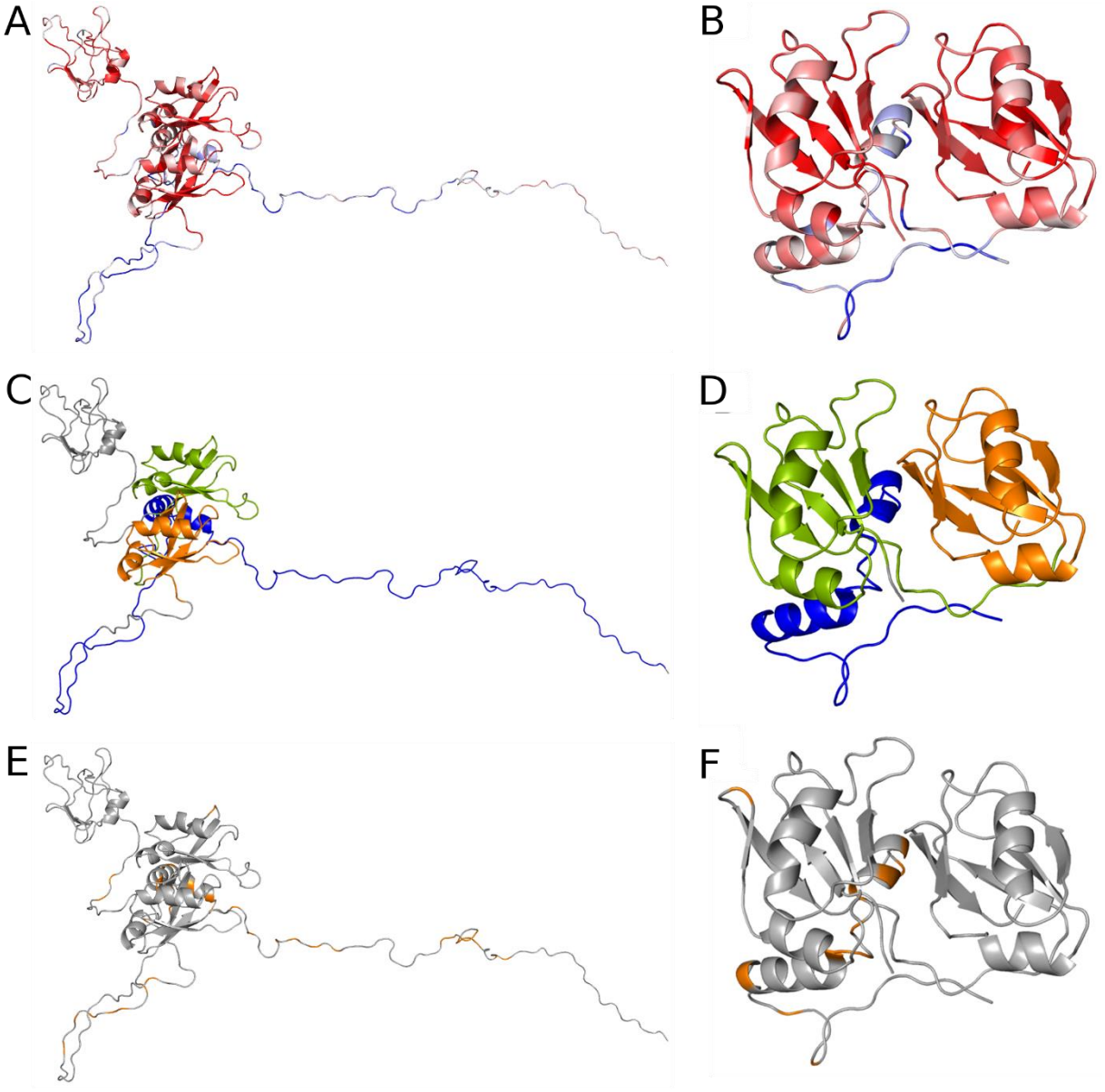
[25] https://github.com/AmandaAlbanaz/dynamism

considerando que estão disponíveis no PDB apenas estruturas resolvidas por NMR e de baixa qualidade que dificultam a obtenção de estruturas via modelagem por homologia, considerando que grandes porções dos trechos disponíveis não possuem estrutura secundária definida. Os detalhes da qualidade dos modelos estão disponíveis nos apêndices Apêndice VII – Modelo Tridimensional da TDP-43e Apêndice VIII – Modelo Tridimensional da FUS/TLS

Embora o modelo tridimensional da TDP-43 seja de baixa qualidade, a estrutura obtida permite a noção inicial da disposição das mutações e sua relação com as regiões conservadas (Figura 32A), domínios cruciais à função (Figura 32B) e regiões mais, ou menos, tolerantes à essas mutações.

Apesar da aparente desorganização estrutural da TDP-43 modelada (Figura 32), essa estrutura serve como partida para que sejam feitas otimizações que permitam estudos mais concretos e aprofundados. Comparando a estrutura obtida com as regiões preditas como desordenadas, é perceptível que a região terminal predita de maior desordem corresponde àquela representada na estrutura, a qual também é composta por resíduos menos conservados.

Figura 32 - Modelo estrutural da TDP-43: domínios, mutações e conservação



Modelo estrutural da TDP-43 obtido via modelagem por homologia. Representação estrutural dos resíduos conservados (A e B); domínios funcionais da TDP-43 (C e D): RRM1 (verde), RRM2 (laranja) e região rica em glicina (azul); mapeamento das mutações *missense* em análise na DynAMISM em laranja (E e F). Modelo obtido com 86,8% dos resíduos em regiões favorecidas, 12,6% em regiões permitidas e 0,6% em regiões proibidas do gráfico de Ramachandran e DOPE: 1.14698. As regiões sem estrutura secundária definida na modelagem foram retiradas nas figuras B, D e F.

Concomitantemente, os domínios de grande importância, RRM1 e RRM2 são altamente conservados, mais estruturados e menos tolerantes às mutações. É provável que mutações que se relacionem com graves fenótipos de ALS ocorram nessas regiões e estejam envolvidas em disfunções de RNAs. Contudo as mutações encontradas ocorrem majoritariamente fora desses domínios (Figura 10 e Figura ***32***), em regiões de maior desordem e menor conservação. As

quais também são regiões de interação com outras proteínas presentes no mecanismo da ALS, a UBQLN2. Isso indica que é necessário coletar mais informações sobre as mutações em ambas as proteínas e otimizações estruturais, já que essa região pode estar relacionada a fenótipos de ALS que envolvam essas duas proteínas, em um mecanismo mais amplo e complexo. Apesar dessa região de interação ser, de modo geral mais tolerante a mutações, existe uma ilha de grande intolerância entre os resíduos de posição 300 e 370, como é visto na Figura 10C, onde há um grande número de mutações (Figura 10D).

As regiões intolerantes às mutações (Figura 10C) são também as mais ordenadas (Figura 10A e B), possivelmente áreas de formação de alfa-hélices, estruturas altamente conservadas e cruciais em proteínas com domínios transmembrana, por sua característica hidrofóbica.

A região C-terminal (entre os resíduos 275 e 310) é uma região rica em glicina (PESIRIDIS; LEE; TROJANOWSKI, 2009), um aminoácido altamente flexível sendo característico de regiões desordenadas e de alça (*loops*), contudo pelas predições essa região é ordenada (Figura 10A e B). Tende a ser um domínio conservado, sendo condizente com o fato de ser um domínio crítico para a TDP-43. Algumas mutações nesta região, principalmente para os resíduos de serina ou treonina, podem ser capazes de hiperfosforilar o estado da TDP-43 (PESIRIDIS; LEE; TROJANOWSKI, 2009). Mutações nesta região também podem estar relacionadas ao ganho de função *splicing* (SIVAKUMAR et al., 2018).

Adicionalmente, mutações na região C-terminal, mais especificamente no domínio *prion-like* (entre os resíduos 263-414), estão relacionadas à perda de enovelamento (SCOTTER; CHEN; SHAW, 2015).

Algumas mutações específicas podem alterar o estado de fosforilação ou aumentar a propensão da proteína em formar agregados, fatores que podem participar do mecanismo de início ou progressão da ALS. Por exemplo, as mutações S379C, S379P e S393L abolem os sítios de fosforilação da caseína quinase I na TDP-43, sendo consideradas mutações patogênicas. (PESIRIDIS; LEE; TROJANOWSKI, 2009). Essas mutações ocorrem em posição inserida em uma região predita como ordenada (Figura 10A e B). A S379C é predita como neutra pelos métodos polyphen2-HumVar e PROVEAN, enquanto seu efeito é deletério pelos métodos SIFT e polyphen2-HumDiv.

As mutações Q331K e N345K no domínio *prion-lik*e da região C-terminal podem criar novos alvos para ubiquitinação da proteína e fazer parte do mecanismo patogênico da doença

(PESIRIDIS; LEE; TROJANOWSKI, 2009). Em experimento em ratos essa mutação causou fenótipos de problemas cognitivos (SIVAKUMAR et al., 2018). A mutação Q331K ocorre em região ordenada (Figura 10A e B), crítica quanto a intolerância de mutações, contudo predita como deletéria apenas pelo método polyphen2-HumDiv, enquanto que a N345K ocorre em região desordenada (Figura 10A e B). Apesar de pertencer a uma região com baixa tolerância de mutações é predita como neutra por todos os métodos anteriormente mencionados.

A mutação Q331K também está relacionada ao fenótipo progressivo idade-dependente, mesmo sem alteração de localização da proteína do núcleo para o citoplasma ou formação de agregados (ARNOLD et al., 2013).

Ao comparar essas informações com o mapeamento de taxa de tolerância de mutações (Figura 10), nota-se que as regiões menos desestruturadas, que também são as mais conservadas, são ao mesmo tempo, as regiões de menor tolerância a mutações.

A mutação G348C é a segunda *missense* mais frequente e ocorre em região desordenada. Esta mutação ocorre em uma região da proteína que embora tenha sido predita como tolerante às mutações em geral (qualquer possibilidade de mutação *missense* segundo o código genético), foi predita como deletéria pelos métodos SIFT e polyphen2-HumDiv, mas como neutra pelos métodos PROVEAN e polyphen2-HumVar. Essas predições de neutralidade são condizentes com a observação da melhoria em déficit motor através da expressão dessa mutação em camundongos transgênicos (DUTTA; SWARUP; JULIEN, 2017).

Dada a complexidade molecular da ALS, etapas do mecanismo patogênico compreendem várias mutações concomitantes, mesmo que de proteínas distintas. Nesse contexto a estabilidade anormal de TDP-43, em sua maioria aumentada devido à presença de mutações, promovem complexos com a FUS/TLS, mesmo que nativa, promovendo a formação e agregados. Essa interação causa prejuízo funcional à FUS/TLS, o que pode ser um evento que precede a alteração de localização protéica e formação de agregados. Esse complexo TDP-43-FUS/TLS tem como característica a mutação Q331K anteriormente abordada (LING et al., 2010).

A maior parte das mutações observadas na TDP-43 aumentam sua estabilidade e estão associadas a formação de agregados (LING et al., 2010).

5.1.3 FUS/TLS

Apesar dos esforços em obter uma estrutura completa da FUS/TLS, não foram obtidas modelos de boa qualidade que tenham sentido biológico (detalhes no Apêndice VIII – Modelo

Tridimensional da FUS/TLS). Apesar da alta desordem estrutural (Figura 11A e B), os curtos trechos ordenados não conferem na estrutura predita.

Os pequenos trechos de regiões ordenadas na estrutura da FUS/TLS são os trechos de menor tolerância à mutações, conforme os gráficos das fFigura 11B e C.

Pode-se dizer que, quando as regiões de interação da FUS/TLS com outras moléculas são mais ordenadas, tem-se interações mais estáveis e portanto, são regiões de grande importância funcional e menos susceptíveis a mutações. Sendo que na Figura 11D são apontadas poucas mutações nessas regiões, especialmente entre os resíduos 290 e 375. Contudo as duas mutações reportadas nesta região, P345S e R383C, ocorrem em posições preditas como conservadas (visível no alinhamento e conservação[26]) e são preditas como mutações deletérias pelo PROVEAN e polyphen2 (dado disponível na base de dados[20]).

O domínio de interação com RNA está localizado entre os aminoácidos 280 e 370 (KWIATKOWSKI et al., 2009) (faixa em verde na Figura 11D), região moderadamente mais ordenada e tolerante a mutações (Figura 11C), onde a mutação P345S, discutida anteriormente, é reportada.

A região com o maior número de mutações (Figura 11D) compreende os resíduos 500 a 526, sendo também a região de maior desordem estrutural na SOD1 e predominantemente composta por resíduos conservados. As predições acerca do efeito de mutações nesta região foram majoritariamente prejudiciais à proteína, contudo a predição do polyphen2 pelo método humVar, prediz essas mutações em sua maioria como neutras, exceto mutações aberrantes como a R524W, onde pode ser causado um choque estérico devido à grande mudança de volume na cadeia lateral. Considerando este conjunto de predições é possível que haja uma limitação metodológica no MTR, que tenha causado a predição de tolerância às mutações, quando na verdade a região está associada à ocorrência de mutações patogênicas, portanto, naturalmente não toleradas.

Apesar dessas mutações estarem acumuladas em uma região tolerante e as próprias alterações serem preditas em sua maioria como neutras, essa região C-terminal possui atividade de tráfico nuclear, a qual é prejudicada, promovendo o deslocamento da proteína ao citoplasma, formando grânulos de estresse citoplasmático e desencadeiando problemas no metabolismo,

[26] https://github.com/AmandaAlbanaz/dynamism

processamento e transporte de mRNAs, resultando em degeneração de neurônios motores (ITO et al., 2011).

Essa região terminal é rica em resíduos de arginina. Esses, devido à sua característica básica são essenciais à participação da FUS/TLS na sinalização de tráfico nuclear (ITO et al., 2011). Portanto as mutações que ocorrem nesta região, apesar de não prejudicarem diretamente a própria estrutura, podem alterar sua carga, crucial à função.

As mutações R514S, R521C e P525L em conjunto, estão diretamente relacionadas à formação dos grânulos subcelulares (ITO et al., 2011). A arginina é um aminoácido carregado positivamente, enquanto a serina é um aminoácido polar. A cisteína é hidrofóbica e a prolina além de hidrofóbica é extremamente rígida. Essa alteração de flexibilidade devido à mutação de prolina para leucina pode comprometer a estrutura protéica.

A R514S é predita como mutação deletéria, exceto pelo método HumVar do polyphen e é um resíduo moderadamente conservado. A R521C é menos conservada e predita como deletéria por todos os métodos.

Algumas das mutações mais encontradas são a R521C, discutida anteriormente, e R521G, a qual foi relacionada à causa de tráfego aberrante da proteína, com retenção no citoplasma (KWIATKOWSKI et al., 2009).

As mutações R521G (predita como deletéria) e H517Q foram relacionadas à alteração da solubilidade protéica da FUS/TLS, aumentando a insolubilidade total da proteína nuclear (KWIATKOWSKI et al., 2009).

5.2 Modelos Preditivos: relacionando mutações *missense* a diferentes fenótipos da doença

A Tabela 3 reúne as melhores correlações de Pearson obtidas até o momento para as correlações entre os dados de idade de início da doença, idade de óbito e tempo de sobrevivência de 30 pacientes com diferentes mutações *missense* na SOD1 e os atributos de predição baseados em sequência (PROVEAN) e estrutura (mCSM-Stability, mCSM-PPI, DynaMut). Ao comparar com as correlações de Pearson na Tabela 4 obtidas por Kumar e colaboradores (2017), onde utilizaram as mesmas 30 mutações, porém relacionando apenas com o atributo de alteração em estabilidade, é possível observar que são encontradas correlações melhores inclusive com outros atributos, apontando para a relevância de outros mecanismos moleculares que podem estar relacionados à mutações em ALS.

Ao combinar atributos baseados em sequência e estrutura selecionados pelo algoritmo M5P na tarefa de conjunto de treino (Tabela 5), foi possível atingir correlações de Pearson satisfatórias, de até 0,7 para predições de idade de óbito e tempo de sobrevivência dos pacientes de ALS portadores de mutações *missense* na SOD1. Contudo para a idade de início da doença não foi encontrada correlação satisfatória. As menores correlações encontradas por Kumar e colaboradores (2017) também foram para os dados de idade de início.

Tabela 3 - Correlações entre fenótipos de 30 casos de mutações missense com atributos preditivos

Fenótipo Clínico	Atributo Preditivo	Correlação de Pearson
Idade de início	mCSM-Stability	0,31
	Dynamut	0,38
Tempo de sobrevivência	mCSM-PPI	0,29
Idade de óbito	mCSM-PPI	0,31
	PROVEAN	0,33
	Dynamut	0,37

Correlações entre os fenótipos clínicos de 30 mutações (conforme trabalho de Kumar e colaboradores (2017) e atributos preditivos adicionais conforme metodologia aqui utilizada.

Tabela 4 - Correlações entre fenótipos de 30 casos de mutações missense com alteração em estabilidade

Fenótipo Clínico	Atributo Preditivo	Correlação de Pearson
Idade de início	Estabilidade experimental	0,1
Tempo de sobrevivência		0,4
Idade de óbito		0,39

Correlações entre os fenótipos clínicos de 30 mutações e a estabilidade experimental das estruturas contendo as mutações. Dados extraídos de Kumar et. al (2017).

Tabela 5 - Modelos Preditivos

Fenótipo Clínico	Atributos Preditivos	Correlação de Pearson
Idade de início		0,383

Tempo de sobrevivência	RSA, distância à interface de contato, escore de conservação, PROVEAN, estabilidade (Dynamut), distância dos metais	0,721
Idade de Óbito		0,702

Correlações entre os fenótipos clínicos de 30 mutações (conforme trabalho de Kumar e colaboradores (2017)) e atributos preditivos adicionais, formando um modelo, conforme metodologia aqui utilizada.

6 Conclusões

A ALS é uma doença complexa e multifatorial na qual mutações *missense* são fatores importantes para análise e compreensão do mecanismo causador da doença. As mutações *missesnse* alteram diversos aspectos dentro da estrutura, causando efeitos que vão além da sua área de ocorrência e até mesmo da proteína. Nesse contexto, a análise das alterações dessas mutações na SOD1 possibilitou a identificação de importantes correlações em relação ao início, óbito e progresso da doença. Esses apontam para outros fatores e mecanismos moleculares que podem ser importantes para um melhor entendimento da relação genótipo-fenótipo em ALS. O estudo da correlação de propriedades proteicas com os fenótipos clínicos possibilitou a formação de um conjunto capaz de predizer os fenótipos clínicos de idade de óbito e tempo de sobrevivência com uma correlação de Pearson de até 0,7. As análises estatísticas e os modelos preditivos ampliam o conhecimento acerca da relação existente entre o genótipo e fenótipo da doença. Estes acrescentam propriedades que são importantes nas proteínas e que sofrem alteração por mutações *missense*, como por exemplo a flexibilidade estrutural, conservação, RSA dos resíduos, bem como sua distância à interface de contato e proximidade dos metais na estrutura da SOD1.

Adicionalmente aos modelos preditivos, a base de dados desenvolvida, DynAMISM, amplia a base de conhecimento sobre ALS e fornece suporte crucial à estudos da doença. O qual carece também de estruturas bem resolvidas para que sejam realizadas análises em múltiplas proteínas.

7 Perspectivas

Com base nos resultados obtidos, será construída a interface web para o banco de dados relacional DynAMISM, provisoriamente hospedado no *github*, facilitando e otimizando a navegação pelas informações. Além de expandir as informações acerca da TDP-43 e FUS e de outras proteínas relacionadas à ALS (Apêndice 10.1), conforme as informações são geradas. Informações relevantes serão adicionadas sempre que necessário, adicionalmente à otimização das estruturas da TDP-43 e FUS/TLS.

Em relação aos modelos preditivos, serão feitos múltiplos testes utilizando ferramentas de aprendizado de máquina e mineração de dados, para que sejam desenvolvidos modelos com a melhor acurácia possível.

Será estudado o acréscimo de novos atributos de modo a refinar os modelos preditivos conforme conhecimentos baseados na literatura. Sabe-se que algumas mutações na região C-terminal de proteínas como a TDP-43 podem aumentar a fosforilação da região, aumentando a propensão à formação de agregados e podendo causar alterações na localização subcelular (KABASHI et al., 2008). Dessa forma será estudada a inclusão de predições de localização subcelular, fosforilação, e de efeitos de mutações *missense* na interação de proteínas com ácido nucléico, como no caso das proteínas TDP-43 e FUS/TLS.

Outro fator importante a ser avaliado é a relação entre as predições baseadas em sequência e estrutura e os fenótipos clínicos de local de início da doença no paciente e sua progressão. Considerando a dificuldade enfrentada ao se tratar os pacientes devido ao prognóstico e espectro de variações da doença.

Após o estabelecimento do modelo preditivo de melhor performance, estes serão disponibilizados em uma plataforma *online* de fácil utilização.

8 Material de Suporte

- Alinhamentos múltiplos de sequência com conservação disponíveis em:

https://drive.google.com/drive/folders/1-ttgdF-ct7A-ZhatTVDz-boS15OnNAZV?usp=sharing

- Banco de dados DynAMISM disponível em:

https://github.com/AmandaAlbanaz/dynamism

9 Referências

ABEL, O. et al. Development of a Smartphone App for a Genetics Website: The Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD). **JMIR mhealth and uhealth**, v. 1, n. 2, p. e18, 4 set. 2013a.

ABEL, O. et al. Development of a Smartphone App for a Genetics Website: The Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD). **JMIR mhealth and uhealth**, v. 1, n. 2, p. e18, 4 set. 2013b.

ADZHUBEI, I. A. et al. A method and server for predicting damaging missense mutations. **Nature Methods**, v. 7, n. 4, p. 248–249, abr. 2010.

ADZHUBEI, I.; JORDAN, D. M.; SUNYAEV, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. **Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]**, v. 0 7, p. Unit7.20, jan. 2013.

ALBANAZ, A. T. S. et al. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. **Expert Opinion on Drug Discovery**, v. 12, n. 6, p. 553–563, 3 jun. 2017.

AL-CHALABI, A.; VAN DEN BERG, L. H.; VELDINK, J. Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. Nature Reviews Neurology, v. 13, n. 2, p. 96–104, fev. 2017.

ANDERSEN, P. M. Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. **Current Neurology and Neuroscience Reports**, v. 6, n. 1, p. 37–46, 1 fev. 2006.

ARAI, T. et al. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. **Biochemical and Biophysical Research Communications**, v. 351, n. 3, p. 602–611, 22 dez. 2006.

ARNESANO, F. et al. The Unusually Stable Quaternary Structure of Human Cu,Zn-Superoxide Dismutase 1 Is Controlled by Both Metal Occupancy and Disulfide Status. **Journal of Biological Chemistry**, v. 279, n. 46, p. 47998–48003, 12 nov. 2004.

ARNOLD, E. S. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. **Proceedings of the National Academy of Sciences**, v. 110, n. 8, p. E736–E745, 19 fev. 2013.

ARTHUR, K. C. et al. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. **Nature Communications**, v. 7, 11 ago. 2016.

ASHKENAZY, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. **Nucleic Acids Research**, v. 44, n. W1, p. W344–W350, 8 jul. 2016.

BEGHI, E. et al. The epidemiology of ALS and the role of population-based registries. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease**, v. 1762, n. 11–12, p. 1150–1157, nov. 2006.

BEGHI, E. et al. The epidemiology and treatment of ALS: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials. **Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases**, v. 12, n. 1, p. 1–10, jan. 2011.

BENTMANN, E.; HAASS, C.; DORMANN, D. Stress granules in neurodegeneration--lessons learnt from TAR DNA binding protein of 43 kDa and fused in sarcoma. **The FEBS journal**, v. 280, n. 18, p. 4348–4370, set. 2013.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 1 jan. 2000.

BERMAN, H. M. The Protein Data Bank: a historical perspective. **Acta Crystallographica Section A Foundations of Crystallography**, v. 64, n. 1, p. 88–95, 1 jan. 2008.

BEYER, K.; ARIZA, A. Alpha-Synuclein Posttranslational Modification and Alternative Splicing as a Trigger for Neurodegeneration. **Molecular Neurobiology**, v. 47, n. 2, p. 509–524, 1 abr. 2013.

BOLOGNESI, B. et al. ANS Binding Reveals Common Features of Cytotoxic Amyloid Species. **ACS Chemical Biology**, v. 5, n. 8, p. 735–740, 20 ago. 2010.

BRAHIMI, F. et al. The Paradoxical Signals of Two TrkC Receptor Isoforms Supports a Rationale for Novel Therapeutic Strategies in ALS. PLOS ONE, v. 11, n. 10, p. e0162307, 3 out. 2016.

BRETTSCHNEIDER, J. et al. TDP-43 pathology and neuronal loss in amyotrophic lateral sclerosis spinal cord. **Acta Neuropathologica**, v. 128, n. 3, p. 423–437, 1 set. 2014.

BUNTON-STASYSHYN, R. K. A. et al. SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology: New and Renascent Themes. **The Neuroscientist**, v. 21, n. 5, p. 519–529, out. 2015.

BURATTI, E. et al. Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. **The EMBO Journal**, v. 20, n. 7, p. 1774–1784, 2 abr. 2001.

BURATTI, E. Functional Significance of TDP-43 Mutations in Disease. In: **Advances in Genetics**. [s.l.] Elsevier, 2015. v. 91p. 1–53.

BURATTI, E.; BARALLE, F. E. Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease. **Frontiers in Bioscience: A Journal and Virtual Library**, v. 13, p. 867–878, 1 jan. 2008.

CAIRNS, N. J. et al. TDP-43 in familial and sporadic frontotemporal lobar degeneration with ubiquitin inclusions. **The American Journal of Pathology**, v. 171, n. 1, p. 227–240, jul. 2007.

CAPAUTO, D. et al. A Regulatory Circuitry Between Gria2, miR-409, and miR-495 Is Affected by ALS FUS Mutation in ESC-Derived Motor Neurons. **Molecular Neurobiology**, v. 55, n. 10, p. 7635–7651, 1 out. 2018.

CELNIKER, G. et al. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. **Israel Journal of Chemistry**, v. 53, n. 3–4, p. 199–206, abr. 2013.

CHIÒ, A. et al. Global Epidemiology of Amyotrophic Lateral Sclerosis: A Systematic Review of the Published Literature. **Neuroepidemiology**, v. 41, n. 2, p. 118–130, 2013.

CHOI, Y. et al. Predicting the Functional Effect of Amino Acid Substitutions and Indels. **PLOS ONE**, v. 7, n. 10, p. e46688, 8 out. 2012.

CHOI, Y.; CHAN, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. **Bioinformatics**, v. 31, n. 16, p. 2745–2747, 15 ago. 2015.

CIRYAM, P. et al. Spinal motor neuron protein supersaturation patterns are associated with inclusion body formation in ALS. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 20, p. E3935–E3943, 16 2017.

COZZOLINO, M. et al. Cysteine 111 affects aggregation and cytotoxicity of mutant Cu,Zn-superoxide dismutase associated with familial amyotrophic lateral sclerosis. **The Journal of Biological Chemistry**, v. 283, n. 2, p. 866–874, 11 jan. 2008.

CRUZ, M. P. Edaravone (Radicava). **Pharmacy and Therapeutics**, v. 43, n. 1, p. 25–28, jan. 2018.

D'AMICO, E. et al. Clinical Perspective of Oxidative Stress in Sporadic ALS. **Free radical biology & medicine**, v. 65, dez. 2013.

DAS, A.; PLOTKIN, S. S. Mechanical Probes of SOD1 Predict Systematic Trends in Metal and Dimer Affinity of ALS-Associated Mutants. **Journal of Molecular Biology**, v. 425, n. 5, p. 850–874, 11 mar. 2013.

DE SANTIS, R. et al. FUS Mutant Human Motoneurons Display Altered Transcriptome and microRNA Pathways with Implications for ALS Pathogenesis. **Stem Cell Reports**, v. 9, n. 5, p. 1450–1462, 14 nov. 2017.

DE VOS, K. J. et al. Role of Axonal Transport in Neurodegenerative Diseases. **Annual Review of Neuroscience**, v. 31, n. 1, p. 151–173, 2008.

DEL AGUILA, M. A. et al. Prognosis in amyotrophic lateral sclerosis: a population-based study. **Neurology**, v. 60, n. 5, p. 813–819, 11 mar. 2003.

DENG, H.; GAO, K.; JANKOVIC, J. The role of FUS gene variants in neurodegenerative diseases. **Nature Reviews. Neurology**, v. 10, n. 6, p. 337–348, jun. 2014.

DENG, Q. et al. FUS is Phosphorylated by DNA-PK and Accumulates in the Cytoplasm after DNA Damage. **The Journal of Neuroscience**, v. 34, n. 23, p. 7802–7813, 4 jun. 2014.

DINI MODIGLIANI, S. et al. An ALS-associated mutation in the FUS 3′-UTR disrupts a microRNA–FUS regulatory circuitry. **Nature Communications**, v. 5, n. 1, dez. 2014.

DUTTA, K.; SWARUP, V.; JULIEN, J.-P. Potential Therapeutic Use of Withania somnifera for Treatment of Amyotrophic Lateral Sclerosis. In: KAUL, S. C.; WADHWA, R. (Eds.). . **Science of Ashwagandha: Preventive and Therapeutic Potentials**. Cham: Springer International Publishing, 2017. p. 389–415.

DUTTA, S. et al. Data deposition and annotation at the worldwide protein data bank. **Methods in Molecular Biology (Clifton, N.J.)**, v. 426, p. 81–101, 2008.

EMDE, A. et al. Dysregulated miRNA biogenesis downstream of cellular stress and ALS-causing mutations: a new mechanism for ALS. **The EMBO journal**, v. 34, n. 21, p. 2633–2651, 3 nov. 2015.

EXOME AGGREGATION CONSORTIUM et al. Analysis of protein-coding genetic variation in 60,706 humans. **Nature**, v. 536, n. 7616, p. 285–291, ago. 2016.

FEILER, M. S. et al. TDP-43 is intercellularly transmitted across axon terminals. **J Cell Biol**, v. 211, n. 4, p. 897–911, 23 nov. 2015.

FLOUDAS, C. A. et al. Advances in protein structure prediction and de novo protein design: A review. **Chemical Engineering Science**, v. 61, n. 3, p. 966–988, fev. 2006.

FOROSTYAK, S.; SYKOVA, E. Neuroprotective Potential of Cell-Based Therapies in ALS: From Bench to Bedside. **Frontiers in Neuroscience**, v. 11, 24 out. 2017.

FORSBERG, K. et al. Novel Antibodies Reveal Inclusions Containing Non-Native SOD1 in Sporadic ALS Patients. **PLoS ONE**, v. 5, n. 7, p. e11552, 14 jul. 2010.

GOMES, C. et al. Establishment of a cell model of ALS disease: Golgi apparatus disruption occurs independently from apoptosis. **Biotechnology Letters**, v. 30, n. 4, p. 603–610, 1 abr. 2008.

GOSSAGE, L. et al. An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. **Human Molecular Genetics**, v. 23, n. 22, p. 5976–5988, 15 nov. 2014.

GRAD, L. I. et al. Intercellular propagated misfolding of wild-type Cu/Zn superoxide dismutase occurs via exosome-dependent and -independent mechanisms. **Proceedings of the National Academy of Sciences**, v. 111, n. 9, p. 3620–3625, 4 mar. 2014.

GREDAL, O. et al. A clinical trial of dextromethorphan in amyotrophic lateral sclerosis. **Acta Neurologica Scandinavica**, v. 96, n. 1, p. 8–13, 1997.

HART, A. Mann-Whitney test is not just a test of medians: differences in spread can be important. **BMJ : British Medical Journal**, v. 323, n. 7309, p. 391–393, 18 ago. 2001.

HAYDEN, E.; CONE, A.; JU, S. Supersaturated proteins in ALS. **Proceedings of the National Academy of Sciences**, v. 114, n. 20, p. 5065–5066, 16 maio 2017.

HILBERT, M.; BÖHM, G.; JAENICKE, R. Structural relationships of homologous proteins as a fundamental principle in homology modeling. **Proteins: Structure, Function, and Genetics**, v. 17, n. 2, p. 138–151, out. 1993.

ITO, D. et al. Nuclear transport impairment of amyotrophic lateral sclerosis-linked mutations in FUS/TLS. **Annals of Neurology**, v. 69, n. 1, p. 152–162, jan. 2011.

JOYCE, P. I. et al. A novel SOD1-ALS mutation separates central and peripheral effects of mutant SOD1 toxicity. **Human Molecular Genetics**, v. 24, n. 7, p. 1883–1897, 1 abr. 2015.

JUBB, H. C. et al. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. **Journal of Molecular Biology**, v. 429, n. 3, p. 365–371, 03 2017.

KABASHI, E. et al. TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. **Nature Genetics**, v. 40, n. 5, p. 572–574, maio 2008.

KANOUCHI, T.; OHKUBO, T.; YOKOTA, T. Can regional spreading of amyotrophic lateral sclerosis motor symptoms be explained by prion-like propagation? **Journal of Neurology, Neurosurgery, and Psychiatry**, v. 83, n. 7, p. 739–745, jul. 2012.

KAWAMATA, H.; MANFREDI, G. Import, Maturation, and Function of SOD1 and Its Copper Chaperone CCS in the Mitochondrial Intermembrane Space. **Antioxidants & Redox Signaling**, v. 13, n. 9, p. 1375–1384, 1 nov. 2010.

KINO, Y. et al. FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. **Acta Neuropathologica Communications**, v. 3, n. 1, dez. 2015.

KUMAR, V. et al. Computing disease-linked SOD1 mutations: deciphering protein stability and patient-phenotype relations. **Scientific Reports**, v. 7, n. 1, dez. 2017.

KUO, P.-H. et al. The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids. **Nucleic Acids Research**, v. 42, n. 7, p. 4712–4722, abr. 2014.

KWIATKOWSKI, T. J. et al. Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. **Science**, v. 323, n. 5918, p. 1205–1208, 27 fev. 2009.

KYE, M. J.; GONÇALVES, I. DO C. G. The role of miRNA in motor neuron disease. **Frontiers in Cellular Neuroscience**, v. 8, p. 15, 2014.

LAFERRIERE, F.; POLYMENIDOU, M. Advances and challenges in understanding the multifaceted pathogenesis of amyotrophic lateral sclerosis. **Swiss Medical Weekly**, 30 jan. 2015.

LASKOWSKI, R. A. et al. PROCHECK: a program to check the stereochemical quality of protein structures. **Journal of Applied Crystallography**, v. 26, n. 2, p. 283–291, 1 abr. 1993.

LEMIESZEK, M. K. et al. Riluzole Inhibits Proliferation, Migration and Cell Cycle Progression and Induces Apoptosis in Tumor Cells of Various Origins. **Anti-Cancer Agents in Medicinal Chemistry**, v. 18, n. 4, p. 565–572, 17 jul. 2018.

LINDING, R. et al. Protein Disorder Prediction. **Structure**, v. 11, n. 11, p. 1453–1459, nov. 2003.

LING, S.-C. et al. ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. **Proceedings of the National Academy of Sciences**, v. 107, n. 30, p. 13318–13323, 27 jul. 2010.

LING, S.-C.; POLYMENIDOU, M.; CLEVELAND, D. W. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. **Neuron**, v. 79, n. 3, p. 416–438, 7 ago. 2013.

LOGROSCINO, G. et al. Descriptive epidemiology of amyotrophic lateral sclerosis: new evidence and unsolved issues. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 79, n. 1, p. 6–11, 1 jan. 2008.

LOGROSCINO, G. et al. Amyotrophic Lateral Sclerosis: An Aging-Related Disease. **Current Geriatrics Reports**, v. 4, n. 2, p. 142–153, jun. 2015.

LUKAVSKY, P. J. et al. Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. **Nature Structural & Molecular Biology**, v. 20, n. 12, p. 1443–1449, dez. 2013.

MACHTOUB, L.; KASUGAI, Y. Amyotrophic Lateral Sclerosis: Advances and Perspectives of Neuronanomedicine. [s.l.] Pan Stanford, 2015.

MACKENZIE, I. R. A.; RADEMAKERS, R. The role of TDP-43 in amyotrophic lateral sclerosis and frontotemporal dementia. **Current opinion in neurology**, v. 21, n. 6, p. 693–700, dez. 2008.

MAKAR, A. B. et al. Formate assay in body fluids: application in methanol poisoning. **Biochemical Medicine**, v. 13, n. 2, p. 117–126, jun. 1975.

MARIN, B. et al. Age-specific ALS incidence: a dose–response meta-analysis. **European Journal of Epidemiology**, v. 33, n. 7, p. 621–634, jul. 2018.

MARKWICK, P. R. L.; MALLIAVIN, T.; NILGES, M. Structural Biology by NMR: Structure, Dynamics, and Interactions. **PLoS Computational Biology**, v. 4, n. 9, p. e1000168, 26 set. 2008.

MCLAUGHLIN, R. L. et al. Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. **Nature Communications**, v. 8, p. 14774, 21 mar. 2017.

MÉSZÁROS, B.; ERDŐS, G.; DOSZTÁNYI, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. **Nucleic Acids Research**, v. 46, n. W1, p. W329–W337, 2 jul. 2018.

MIYAJI, Y. et al. Effect of Edaravone on Favorable Outcome in Patients with Acute Cerebral Large Vessel Occlusion: Subanalysis of RESCUE-Japan Registry. **Neurologia medico-chirurgica**, v. 55, n. 3, p. 241–247, mar. 2015.

MORAHAN, J. M.; PAMPHLETT, R. Amyotrophic Lateral Sclerosis and Exposure to Environmental Toxins: An Australian Case-Control Study. **Neuroepidemiology**, v. 27, n. 3, p. 130–135, 2006.

MÜNCH, C.; O'BRIEN, J.; BERTOLOTTI, A. Prion-like propagation of mutant superoxide dismutase-1 misfolding in neuronal cells. **Proceedings of the National Academy of Sciences**, v. 108, n. 9, p. 3548–3553, 1 mar. 2011.

NARUSE, H. et al. Familial amyotrophic lateral sclerosis with novel A4D SOD1 mutation with late age at onset and rapid progressive course. **Neurology and Clinical Neuroscience**, v. 1, n. 1, p. 45–47, 2013.

NARUSE, H. et al. Burden of rare variants in causative genes for amyotrophic lateral sclerosis (ALS) accelerates age at onset of ALS. **Journal of Neurology, Neurosurgery & Psychiatry**, p. jnnp-2018-318568, 24 out. 2018.

NEMETHOVA, M. et al. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease'' in Italy'. **European Journal of Human Genetics**, v. 24, n. 1, p. 66–72, jan. 2016.

NEUMANN, M. et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. **Science (New York, N.Y.)**, v. 314, n. 5796, p. 130–133, 6 out. 2006.

NICOLAS, A. et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. **Neuron**, v. 97, n. 6, p. 1268- 1283.e6, mar. 2018.

NONAKA, T. et al. Prion-like properties of pathological TDP-43 aggregates from diseased brains. **Cell Reports**, v. 4, n. 1, p. 124–134, 11 jul. 2013.

OU, S. H. et al. Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. **Journal of Virology**, v. 69, n. 6, p. 3584–3596, jun. 1995.

PAPA, L.; MANFREDI, G.; GERMAIN, D. SOD1, an unexpected novel target for cancer therapy. **Genes & Cancer**, v. 5, n. 1–2, p. 15–21, jan. 2014.

PARASURAMAN, S. Protein data bank. **Journal of Pharmacology & Pharmacotherapeutics**, v. 3, n. 4, p. 351–352, 2012.

PASINELLI, P. et al. Amyotrophic Lateral Sclerosis-Associated SOD1 Mutant Proteins Bind and Aggregate with Bcl-2 in Spinal Cord Mitochondria. **Neuron**, v. 43, n. 1, p. 19–30, 8 jul. 2004.

PELED, S. et al. Single cell imaging and quantification of TDP-43 and α-synuclein intercellular propagation. **Scientific Reports**, v. 7, n. 1, p. 544, 28 mar. 2017.

PEREIRA, G. R. C. et al. In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. **Journal of Cellular Biochemistry**, v. 120, n. 3, p. 3583–3598, mar. 2019.

PESIRIDIS, G. S.; LEE, V. M.-Y.; TROJANOWSKI, J. Q. Mutations in TDP-43 link glycine-rich domain functions to amyotrophic lateral sclerosis. **Human Molecular Genetics**, v. 18, n. R2, p. R156–R162, 15 out. 2009.

PETTERSEN, E. F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. **Journal of Computational Chemistry**, v. 25, n. 13, p. 1605–1612, out. 2004.

PHELAN, J. et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. **BMC Medicine**, v. 14, n. 1, dez. 2016.

PHUKAN, J. et al. The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. **Journal of Neurology, Neurosurgery, and Psychiatry**, v. 83, n. 1, p. 102–108, jan. 2012.

PHUKAN, J.; PENDER, N. P.; HARDIMAN, O. Cognitive impairment in amyotrophic lateral sclerosis. **The Lancet Neurology**, v. 6, n. 11, p. 994–1003, nov. 2007.

PIRES, D. E. V.; ASCHER, D. B. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. **Nucleic Acids Research**, v. 44, n. W1, p. W469–W473, 8 jul. 2016.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics (Oxford, England)**, v. 30, n. 3, p. 335–342, 1 fev. 2014a.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. **Nucleic Acids Research**, v. 42, n. W1, p. W314–W319, 1 jul. 2014b.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. **Journal of Medicinal Chemistry**, v. 58, n. 9, p. 4066–4072, 14 maio 2015.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. **Scientific Reports**, v. 6, p. 29575, 7 jul. 2016.

POKRISHEVSKY, E. et al. Aberrant Localization of FUS and TDP43 Is Associated with Misfolding of SOD1 in Amyotrophic Lateral Sclerosis. **PLoS ONE**, v. 7, n. 4, p. e35050, 6 abr. 2012.

POLAZZI, E. et al. Copper-Zinc Superoxide Dismutase (SOD1) Is Released by Microglial Cells and Confers Neuroprotection against 6-OHDA Neurotoxicity. **Neurosignals**, v. 21, n. 1–2, p. 112–128, 2013.

QUINLAN, J. R. Learning With Continuous Classes. World Scientific, 1992

RATTI, A.; BURATTI, E. Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. **Journal of Neurochemistry**, v. 138, p. 95–111, ago. 2016a.

RENTON, A. E.; CHIÒ, A.; TRAYNOR, B. J. State of play in amyotrophic lateral sclerosis genetics. **Nature Neuroscience**, v. 17, n. 1, p. 17–23, jan. 2014.

RODRIGUES, C. H.; ASCHER, D. B.; PIRES, D. E. Kinact: a computational approach for predicting activating missense mutations in protein kinases. **Nucleic Acids Research**, v. 46, n. W1, p. W127–W132, 2 jul. 2018.

RODRIGUES, C. H.; PIRES, D. E.; ASCHER, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. **Nucleic Acids Research**, v. 46, n. W1, p. W350–W355, 2 jul. 2018.

ROSEN, D. R. et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. **Nature**, v. 362, n. 6415, p. 59, mar. 1993.

ROTHSTEIN, J. D. Current hypotheses for the underlying biology of amyotrophic lateral sclerosis. **Annals of Neurology**, v. 65, n. S1, p. S3–S9, jan. 2009a.

ROTHSTEIN, J. D. Current hypotheses for the underlying biology of amyotrophic lateral sclerosis. **Annals of Neurology**, v. 65, n. S1, p. S3–S9, jan. 2009b.

ROTUNNO, M. S.; BOSCO, D. A. An emerging role for misfolded wild-type SOD1 in sporadic ALS pathogenesis. **Frontiers in Cellular Neuroscience**, v. 7, 16 dez. 2013.

ROWLAND, L. P.; SHNEIDER, N. A. Amyotrophic lateral sclerosis. **The New England Journal of Medicine**, v. 344, n. 22, p. 1688–1700, 31 maio 2001.

ROYSTON, J. P. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. **Applied Statistics**, v. 31, n. 2, p. 115, 1982.

ROYSTON, P. Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. **Applied Statistics**, v. 44, n. 4, p. 547, 1995.

SALOMON-FERRER, R.; CASE, D. A.; WALKER, R. C. An overview of the Amber biomolecular simulation package: Amber biomolecular simulation package. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 3, n. 2, p. 198–210, mar. 2013.

SANTAMARIA, N. et al. Intrinsic disorder in proteins involved in amyotrophic lateral sclerosis. **Cellular and Molecular Life Sciences**, v. 74, n. 7, p. 1297–1318, 1 abr. 2017.

SCEKIC-ZAHIROVIC, J. et al. Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. **The EMBO journal**, v. 35, n. 10, p. 1077–1097, 17 2016.

SCHRÖDINGER, LLC. **The PyMOL Molecular Graphics System, Version 1.8**. nov. 2015a.

SCHRÖDINGER, LLC. **The JyMOL Molecular Graphics Development Component, Version 1.8**. nov. 2015b.

SCHRÖDINGER, LLC. **The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8**. nov. 2015c.

SCOTT, K. M. et al. The association between ALS and population density: A population based study. **Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases**, v. 11, n. 5, p. 435–438, out. 2010.

SCOTTER, E. L.; CHEN, H.-J.; SHAW, C. E. TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. **Neurotherapeutics**, v. 12, n. 2, p. 352–363, abr. 2015a.

SCOTTER, E. L.; CHEN, H.-J.; SHAW, C. E. TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. **Neurotherapeutics**, v. 12, n. 2, p. 352–363, 1 abr. 2015b.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3/4, p. 591, dez. 1965.

SHEN, M.; SALI, A. Statistical potential for assessment and prediction of protein structures. **Protein Science**, v. 15, n. 11, p. 2507–2524, nov. 2006.

SILVERMAN, J. M. et al. Disease Mechanisms in ALS: Misfolded SOD1 Transferred Through Exosome-Dependent and Exosome-Independent Pathways. **Cellular and Molecular Neurobiology**, v. 36, n. 3, p. 377–381, 1 abr. 2016.

SIM, N.-L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. **Nucleic Acids Research**, v. 40, n. W1, p. W452–W457, 1 jul. 2012.

SIVAKUMAR, P. et al. TDP-43 mutations increase HNRNP A1-7B through gain of splicing function. **Brain**, v. 141, n. 12, p. e83–e83, 1 dez. 2018.

SRINIVASAN, E.; RAJASEKARAN, R. Computational investigation of the human SOD1 mutant, Cys146Arg, that directs familial amyotrophic lateral sclerosis. **Molecular BioSystems**, v. 13, n. 8, p. 1495–1503, 2017.

STRANGE, R. W. et al. Variable Metallation of Human Superoxide Dismutase: Atomic Resolution Crystal Structures of Cu–Zn, Zn–Zn and As-isolated Wild-type Enzymes. **Journal of Molecular Biology**, v. 356, n. 5, p. 1152–1162, mar. 2006.

BERG, J.M.; STRYER, L.; TYMOCZKO, J.L. **Bioquímica**: capítulos 8 e 17**,** páginas 185-188, 443-444. 7ª edição. Rio de Janeiro: Guanabara Koogan, 2014.

SUNDARAMOORTHY, V. et al. Extracellular wildtype and mutant SOD1 induces ER–Golgi pathology characteristic of amyotrophic lateral sclerosis in neuronal cells. **Cellular and Molecular Life Sciences**, v. 70, n. 21, p. 4181–4195, 1 nov. 2013.

TANG, L. et al. Identification of an A4V SOD1 mutation in a Chinese patient with amyotrophic lateral sclerosis without the A4V founder effect common in North America. **Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration**, v. 19, n. 5–6, p. 466–468, 3 jul. 2018.

TAYLOR, J. P.; BROWN, R. H.; CLEVELAND, D. W. Decoding ALS: from genes to mechanism. **Nature**, v. 539, n. 7628, p. 197–206, 9 nov. 2016.

TAYLOR, J. P.; BROWN, R. H.; CLEVELAND, D. W. Decoding ALS: from genes to mechanism. **Nature**, v. 539, n. 7628, p. 197–206, 9 nov. 2016.

The Universal Protein Resource (UniProt). **Nucleic Acids Research**, v. 36, n. Database issue, p. D190–D195, jan. 2008.

TRAMONTANO, A.; MOREA, V. Assessment of homology-based predictions in CASP5. **Proteins**, v. 53 Suppl 6, p. 352–368, 2003.

TRAYNELIS, J. et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. **Genome Research**, v. 27, n. 10, p. 1715–1729, out. 2017.

TREZZA, A. et al. A Computational Approach From Gene to Structure Analysis of the Human ABCA4 Transporter Involved in Genetic Retinal Diseases. **Investigative Ophthalmology & Visual Science**, v. 58, n. 12, p. 5320–5328, 1 out. 2017.

UNIPROT CONSORTIUM, T. UniProt: the universal protein knowledgebase. **Nucleic Acids Research**, v. 46, n. 5, p. 2699–2699, 16 mar. 2018.

USHER, J. L. et al. Analysis of HGD Gene Mutations in Patients with Alkaptonuria from the United Kingdom: Identification of Novel Mutations. In: ZSCHOCKE, J. et al. (Eds.). . **JIMD Reports, Volume 24**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. v. 24p. 3–11.

USHER, J. L. et al. Analysis of HGD Gene Mutations in Patients with Alkaptonuria from the United Kingdom: Identification of Novel Mutations. **JIMD Reports**, v. 24, p. 3–11, 15 fev. 2015.

VAN ZUNDERT, B.; BROWN, R. H. Silencing strategies for therapy of SOD1-mediated ALS. **Neuroscience Letters**, v. 636, p. 32–39, jan. 2017a.

VAN ZUNDERT, B.; BROWN, R. H. Silencing strategies for therapy of SOD1-mediated ALS. **Neuroscience Letters**, ALS: New Molecular Mechanisms and Emerging Therapeutic Targets. v. 636, p. 32–39, 1 jan. 2017b.

VANCE, C. et al. Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. **Science (New York, N.Y.)**, v. 323, n. 5918, p. 1208–1211, 27 fev. 2009.

VEDITHI, S. C. et al. Structural Implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. **Scientific Reports**, v. 8, n. 1, p. 5016, 22 mar. 2018.

VENU, S.; LOLLA, G.; HOBEROCK, L. On Selecting The Number Of Bins For A Histogram. 2019.

VERKHIVKER, G. M. et al. Simulating disorder-order transitions in molecular recognition of unstructured proteins: Where folding meets binding. **Proceedings of the National Academy of Sciences**, v. 100, n. 9, p. 5148–5153, 29 abr. 2003.

VYAS, V. et al. Homology modeling a fast tool for drug discovery: Current perspectives. **Indian Journal of Pharmaceutical Sciences**, v. 74, n. 1, p. 1, 2012.

WANG, H.-Y. et al. Structural diversity and functional implications of the eukaryotic TDP gene family. **Genomics**, v. 83, n. 1, p. 130–139, jan. 2004.

WANG, I.-F. et al. TDP-43, the signature protein of FTLD-U, is a neuronal activity-responsive factor. **Journal of Neurochemistry**, v. 105, n. 3, p. 797–806, maio 2008a.

WANG, X.; GULBAHCE, N.; YU, H. Network-based methods for human disease gene prediction. **Briefings in Functional Genomics**, v. 10, n. 5, p. 280–293, 1 set. 2011.

WANG, I.-F.; REDDY, N. M.; SHEN, C.-K. J. Higher order arrangement of the eukaryotic nuclear bodies. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 21, p. 13583–13588, 15 out. 2002.

WANG, Q. et al. MicroRNA-377 is up-regulated and can lead to increased fibronectin production in diabetic nephropathy. **The FASEB Journal**, v. 22, n. 12, p. 4126–4135, 20 ago. 2008b.

WEBB, B.; SALI, A. Comparative Protein Structure Modeling Using MODELLER. **Current Protocols in Bioinformatics**, v. 54, p. 5.6.1-5.6.37, 20 2016.

WIJESEKERA, L. C.; LEIGH, P. N. Amyotrophic lateral sclerosis. **Orphanet Journal of Rare Diseases**, v. 4, p. 3, 3 fev. 2009.

WILLIAMS, J. R. et al. Copper delivery to the CNS by CuATSM effectively treats motor neuron disease in SODG93A mice co-expressing the Copper-Chaperone-for-SOD. **Neurobiology of Disease**, v. 89, p. 1–9, maio 2016.

WISHART, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. **Nucleic Acids Research**, v. 46, n. D1, p. D1074–D1082, 4 jan. 2018.

WORTH, C. L.; PREISSNER, R.; BLUNDELL, T. L. SDM--a server for predicting effects of mutations on protein stability and malfunction. **Nucleic Acids Research**, v. 39, n. Web Server issue, p. W215-222, jul. 2011.

WROE, R. et al. ALSOD: The Amyotrophic Lateral Sclerosis Online Database. **Amyotrophic Lateral Sclerosis**, v. 9, n. 4, p. 249–250, jan. 2008.

WU, P. et al. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. **Statistics in Medicine**, v. 33, n. 8, p. 1261–1271, 15 abr. 2014.

YAP, B. W.; SIM, C. H. Comparisons of various types of normality tests. **Journal of Statistical Computation and Simulation**, v. 81, n. 12, p. 2141–2155, dez. 2011.

ZEINEDDINE, R. et al. Addition of exogenous SOD1 aggregates causes TDP-43 mislocalisation and aggregation. **Cell Stress and Chaperones**, v. 22, n. 6, p. 893–902, 1 nov. 2017.

ZEINEDDINE, R. et al. SOD1 protein aggregates stimulate macropinocytosis in neurons to facilitate their propagation. **Molecular Neurodegeneration**, v. 10, n. 1, dez. 2015.

ZHANG, Y. Progress and challenges in protein structure prediction. **Current opinion in structural biology**, v. 18, n. 3, p. 342–348, jun. 2008.

ZUFIRÍA, M. et al. ALS: A bucket of genes, environment, metabolism and unknown ingredients. **Progress in Neurobiology**, v. 142, p. 104–129, jul. 2016.

10 Apêndice

10.1 Apêndice I – Proteínas Associadas à ALS

Tabela 6 - Proteínas associadas à ALS

Protein Gene	Protein Name	Cromossomo	Uniprot	Gene Card	GCID	Localização Subcelular
SOD1	Superoxide Dismutase 1	21	P00441	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SOD1	GC21P031659	Cytoplasm, Mitochondrion, Nucleus
FUS/TLS/TLS	DNA-/RNA-binding protein fused in sarcoma/translocated in liposarcoma	16	P35637	http://www.genecards.org/cgi-bin/carddisp.pl?gene=FUS/TLS	GC16P031265	Nucleus
TDP-43	Transactive response DNA binding protein 43 kDa	1	Q13148	http://www.genecards.org/cgi-bin/carddisp.pl?gene=TARDBP	GC01P011013	Nucleus
KIF5A	(Kinesin Family Member 5A) is a Protein Coding gene for Kinesin heavy chain isoform 5A	12	Q12840	http://www.genecards.org/cgi-bin/carddisp.pl?gene=KIF5A	GC12P057549	Cytoskeleton, perinuclear region
VAPB	Vesicle-associated membrane protein-associated protein B/C	20	O95292	http://www.genecards.org/cgi-bin/carddisp.pl?gene=VAPB	GC20P058389	Endoplasmic reticulum, Golgi apparatus
TUBA4A	Tubulin alpha-4A chain	2	P68366	http://www.genecards.org/cgi-bin/carddisp.pl?gene=TUBA4A	GC02M219249	Cytoskeleton

TRPM7	Transient Receptor Potential Cation Channel Subfamily M Member 7	15	Q96QT4	http://www.genecards.org/cgi-bin/carddisp.pl?gene=TRPM7	GC15M050552	Membrane
TAF15	TATA-Box Binding Protein Associated Factor 15	17	Q92804	http://www.genecards.org/cgi-bin/carddisp.pl?gene=TAF15	GC17P035713	Cytoplasm, Nucleus
SYNE1	Synaptic nuclear envelope protein 1 or Nesprin-1	6	Q8NF91	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SYNE1	GC06M152121	Cytoskeleton, Nucleus
SS18L1	synovial sarcoma translocation gene on chromosome 18-like 1 encoding Calcium-responsive transactivator protein	20	O75177	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SS18L1	GC20P062143	Nucleus, kinetochore
SQSTM1	Sequestosome 1	5	Q13501	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SQSTM1	GC05P179806	Endoplasmic reticulum, Cytosol, Endosome, Lysosome, Nucleus, autophagosome, sarcomere
SPG11	Spatacsin or Spastic paraplegia 11 protein	15	Q96JI7	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SPG11	GC15M044562	Cytosol, Nucleus, Axon, Dendrite
SOX5	Transcription factor SOX-5	12	P35711	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SOX5	GC12M023452	Nucleus

UBQLN2	Ubiquilin-2	X	Q9UHD9	http://www.genecards.org/cgi-bin/carddisp.pl?gene=UBQLN2	GC0XP056606	Nucleus, Cytoplasm, Membrane, Autophagosome
SETX	Probable helicase senataxin	9	Q7Z333	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SETX	GC09M132261	Nucleus, Nucleoplasm, Nucleolus, Cytoplasm, chromosome, Telomere, Axon, Growth cone
PLEKHG5	Pleckstrin homology domain-containing family G member 5	1	O94827	http://www.genecards.org/cgi-bin/carddisp.pl?gene=PLEKHG5	GC01M006526	Plasma membrane, Cytoplasm, perinuclear region, Cell junction, lamellipodium
PCP4	Calmodulin regulator protein PCP4	21	P48539	http://www.genecards.org/cgi-bin/carddisp.pl?gene=PCP4	GC21P039867	Cytosol, Nucleus, protein-containing complex
PARK7	Protein/nucleic acid deglycase DJ-1	1	Q99497	http://www.genecards.org/cgi-bin/carddisp.pl?gene=PARK7	GC01P007957	Mitochondrion, Nucleus, Cytoplasm
OPTN	Optineurin	10	Q96CV9	http://www.genecards.org/cgi-bin/carddisp.pl?gene=OPTN	GC10P013141	Golgi apparatus
OMA1	Metalloendopeptidase mitochondrial	1	Q96E52	http://www.genecards.org/cgi-bin/carddisp.pl?gene=OMA1	GC01M058415	Mitochondrion inner membrane, pass membrane protein
VCP	Transitional endoplasmic reticulum ATPase	9	P55072	http://www.genecards.org/cgi-bin/carddisp.pl?gene=VCP	GC09M035056	cytosol, Endoplasmic reticulum, Nucleus, Stress granule

NETO1	Neuropilin and tolloid-like protein 1	18	Q8TDF5	http://www.genecards.org/cgi-bin/carddisp.pl?gene=NETO1	GC18M072742	Plasma membrane
NEFH	Neurofilament heavy polypeptide	22	P12036	http://www.genecards.org/cgi-bin/carddisp.pl?gene=NEFH	GC22P029480	Cytoplasm
MATR3	Matrin-3	5	P43243	http://www.genecards.org/cgi-bin/carddisp.pl?gene=MATR3	GC05P139274	Nucleus matrix
GLE1	Nucleoporin GLE1	9	Q53GS7	http://www.genecards.org/cgi-bin/carddisp.pl?gene=GLE1	GC09P128504	Nucleus, Cytoplasm, Nuclear pore complex
BCL6	B-cell lymphoma 6 protein	3	P41182	http://www.genecards.org/cgi-bin/carddisp.pl?gene=BCL6	GC03M187721	Nucleus
EWSR1	RNA-binding protein EWS (Ewing sarcoma)	22	Q01844	http://www.genecards.org/cgi-bin/carddisp.pl?gene=EWSR1	GC22P029269	Cell membrane, Cytoplasm, Nucleus
BCL11B	B-cell lymphoma/leukemia 11B	14	Q9C0K0	http://www.genecards.org/cgi-bin/carddisp.pl?gene=BCL11B	GC14M099169	Nucleus
DIAPH3	Protein diaphanous homolog 3	13	Q9NSV4	http://www.genecards.org/cgi-bin/carddisp.pl?gene=DIAPH3	GC13M059665	Nucleus, Cytoplasm
CRIM1	Cysteine-rich motor neuron 1 protein	2	Q9NZV1	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CRIM1	GC02P036324	Cell membrane, Membrane protein

ALS2	ALSin Rho Guanine Nucleotide Exchange Factor	2	Q96Q42	http://www.genecards.org/cgi-bin/carddisp.pl?gene=ALS2	GC02M201701	Cytoskeleton,centrosome, Cytosol, Endosome, dendrite
NEK1	NIMA ((never in mitosis gene a)-related kinase 1	4	Q96PY6	http://www.genecards.org/cgi-bin/carddisp.pl?gene=NEK1	GC04M169393	Nucleus, Cytoplasm
CHGB	Secretogranin-1 or Secretory granule protein chromogranin B	20	P05060	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CHGB	GC20P005891	Extracellular region or secreted
ANG	Angiogenin	14	P03950	http://www.genecards.org/cgi-bin/carddisp.pl?gene=ANG	GC14P020763	Nucleus, Nucleolus
TBK1	TANK Binding Kinase 1 or Serine/threonine-protein kinase TBK1	12	Q9UHD2	https://www.genecards.org/cgi-bin/carddisp.pl?gene=TBK1	GC12P064451	Cytoplasm
SYT9	Synaptotagmin-9	11	Q86SS6	http://www.genecards.org/cgi-bin/carddisp.pl?gene=SYT9	GC11P007238	synaptic vesicle membrane, Single-pass membrane protein
SIGMAR1	Sigma non-opioid intracellular receptor 1	9	Q99720	https://www.genecards.org/cgi-bin/carddisp.pl?gene=SIGMAR1	GC09M034634	Endoplasmic reticulum membrane, Cell membrane, postsynaptic density, Nucleus inner membrane, Nucleus outer membrane
RNASE2	Non-secretory ribonuclease	14	P10153	http://www.genecards.org/cgi-bin/carddisp.pl?gene=RNASE2	GC14P021091	Lysosome

RAMP3	Receptor activity-modifying protein 3	7	O60896	http://www.genecards.org/cgi-bin/carddisp.pl?gene=RAMP3	GC07P045163	Cell membrane, membrane protein, Membrane, Single-pass type I membrane protein
PFN1	Profilin-1	17	P07737	http://www.genecards.org/cgi-bin/carddisp.pl?gene=PFN1	GC17M004945	Cytoskeleton
NIPA1	Magnesium transporter NIPA1	15	Q7RTP0	http://www.genecards.org/cgi-bin/carddisp.pl?gene=NIPA1	GC15P022773	Endosome, Plasma membrane
LUM	Lumican	12	P51884	http://www.genecards.org/cgi-bin/carddisp.pl?gene=LUM	GC12M091102	Extracellular matrix region or secreted
LMN1	Prelamin-A/C	1	P02545	http://www.genecards.org/cgi-bin/carddisp.pl?gene=LMNA	GC01P156082	Nucleus, Nucleus envelope
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	12	P09651	http://www.genecards.org/cgi-bin/carddisp.pl?gene=HNRNPA1	GC12P054280	Nucleus, Cytoplasm
GRB14	Growth factor receptor-bound protein 14	2	Q14449	http://www.genecards.org/cgi-bin/carddisp.pl?gene=GRB14	GC02M164492	Endosome membrane, membrane protein, Cytoplasm
FEZF2	Fez family zinc finger protein 2	3	Q8TBJ5	http://www.genecards.org/cgi-bin/carddisp.pl?gene=FEZF2	GC03M062355	Nucleus
CDH13	Cadherin-13	16	P55290	http://www.genecards.org/cgi-	GC16P082660	Plasma membrane, GPI-anchor

				bin/carddisp.pl?gene=CDH13		
ERBB4	Receptor tyrosine-protein kinase erbB-4	2	Q15303	http://www.genecards.org/cgi-bin/carddisp.pltreat?gene=ERBB4	GC02M211375	Cell membrane, membrane protein, Mitochondrion, Nucleus
DCTN1	Dynactin subunit 1	2	Q14203	http://www.genecards.org/cgi-bin/carddisp.pl?gene=DCTN1	GC02M074361	Cytoskeleton, centrosome, centriole, spindle, Nucleus envelope, Cytoplasm, cell cortex
DOC2B	Double C2-like domain-containing protein beta	17	Q14184	http://www.genecards.org/cgi-bin/carddisp.pl?gene=DOC2B	GC17M000142	Plasma membrane, Cytoplasm, Cytoplasmic granule
DAO	D-amino-acid oxidase	12	P14920	http://www.genecards.org/cgi-bin/carddisp.pl?gene=DAO	GC12P108859	Peroxisome
CX3CR1	CX3C chemokine receptor 1	3	P49238	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CX3CR1	GC03M039279	Cell membrane
CRYM	Ketimine reductase mu-crystallin	16	Q14894	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CRYM	GC16M021238	Cytoplasm
CNTN6	Contactin-6	3	Q9UQ52	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CNTN6	GC03P001063	Plasma membrane

CHMP2B	Charged multivesicular body protein 2b	3	Q9UQN3	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CHMP2B	GC03P087277	Cytosol, Late endosome membrane, membrane protein
CHCHD10	Coiled-coil-helix-coiled-coil-helix domain-containing protein 10, mitochondrial	22	Q8WYQ3	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CHCHD10	GC22M023765	Mitochondrion intermembrane space
CDH22	Cadherin-22	20	Q9UJ99	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CDH22	GC20M046173	Mitochondrion intermembrane space
FIG4	Polyphosphoinositide phosphatase	6	Q92562	http://www.genecards.org/cgi-bin/carddisp.pl?gene=FIG4	GC06P109691	Endosome membrane

Esta tabela contém as 58 proteínas descritas na literatura como associadas à patogênese da ALS, nome do gene, nome da proteína codificada, identificador do UniProt, GeneCard e localização subcelular.

10.2 Apêndice II - Palavras-chave Utilizadas na Revisão Bibliográfica

Tabela 7 - Exemplos de Palavras-chave Utilizadas na Busca pela Literatura

ALS AND new AND mutations
ALS AND mutations AND patients
SOD1 AND new AND mutations
SOD1 AND mutations
SOD1 AND patients
TDP-43 AND new AND mutations
TDP-43 AND mutations
TDP-43 AND patients
FUS/TLS AND new AND mutations
FUS/TLS AND mutations
FUS/TLS AND patients
ALS AND proteins
ALS AND protein AND mutations
ALS AND pathogenesis
ALS AND mechanism
ALS AND treatment

10.3 Apêndice III - Metodologia - Uniprot e PDB

UniProt

As seqüências das proteínas foram obtidas em formato fasta por meio do UniProt

(Universal Protein Resource) (UNIPROT CONSORTIUM, 2018), o qual também fornece informações sobre funções, links para estruturas 3D depositadas no PDB (BERMAN et al., 2000) (já com resolução e tamanho), regiões relevantes dentro da proteína, domínios, família e localização subcelular.

O UniProt pode ser compreendido como um recurso central, curado, de sequências de aminoácidos e anotação funcional, que também fornece informações adicionais de valor agregado através de referências cruzadas com outros bancos de dados e fontes (UNIPROT CONSORTIUM, 2018).

A base de conhecimento UniProt (UniProtKB) é um dos quatro componentes do site UniProt. É um banco de dados com curadoria, um ponto de acesso central para informações de proteínas, com referências cruzadas para várias fontes. Mais especificamente, a anotação UniprotKB subsiste na descrição de funções proteicas, domínios e sítios biologicamente relevantes, modificações pós-traducionais, localização subcelular, especificidade tecidual, interações, estrutura 3D, doenças relacionadas, anormalidades (“The Universal Protein Resource (UniProt)”, 2008).

Protein Data Bank

A estrutura cristalográfica 3D da SOD1 selvagem foi extraída do *Protein Data Bank* (PDB), o repositório global de dados estruturais de moléculas biológicas (BERMAN et al., 2000).

Em 1970 foi iniciado um grupo de jovens cristalográfos, que viriam a ser os colaboradores do PDB (RSCBS PDB - Research Collaborators for Structural Bioinformatics Protein Data Bank), responsáveis pela manutenção da base de dados (BERMAN, 2008). E em 1971 este grupo iniciou o projeto PDB, a fim de listar as estruturas de todos os aminoácidos e, desde então, o PDB têm crescido e se consolidado como um importante recurso internacional para a biologia estrutural, incluindo em seu acervo, proteínas, ácidos nucléicos, complexos de proteínas e DNAs, de todos os organismos, entre outras moléculas (PARASURAMAN, 2012).

As estruturas 3D disponíveis são determinadas por métodos experimentais tais como cristalografia de raios-X, ressonância magnética nuclear (RMN), microscopia eletrônica e

microscopia de crio-eletrõnica (Cryo-EM) (DUTTA et al., 2008). Cada estrutura do PDB pode ser baixada e visualizada usando programas de visualização molecular como PyMol (SCHRÖDINGER, LLC, 2015a, 2015b) e Chimera (PETTERSEN et al., 2004). O arquivo da estrutura também possui número de identificação e nome da macromolécula com data de determinação, nome do autor, o método utilizado para determinar a estrutura 3D e a publicação referente à sua resolução (PARASURAMAN, 2012).

10.4 Apêndice IV - Procedimento do Teste para Verificação da Normalidade dos Dados

O Teste de Normalidade Shapiro-Wilk

Conhecer a distribuição dos dados é crucial para a escolha correta dos procedimentos estatísticos, como por exemplo, a decisão por métodos paramétricos, no caso de distribuição normal, ou não paramétricos, quando a distribuição não é normal.

Existem muitas formas de investigar se os dados, são ou não, normalmente distribuídos, uma delas se baseia em análises de gráficos, como o Q-Q plot, histograma e boxplot, ou em testes formais de normalidade, como o Shapiro-Wilk, sendo comum e apropriado unir ambas as formas (YAP; SIM, 2011).

O teste de Shapiro-Wilk, desenvolvido por Shapiro e Wilk (SHAPIRO; WILK, 1965), é considerado o mais poderoso teste de normalidade para uma ampla faixa de distribuições assimétricas de dados e para níveis de distribuições simétricas. O teste é tido como poderoso quando possui alta probabilidade de rejeitar a hipótese nula de normalidade quando a amostra provém de uma população de distribuição não normal (SHAPIRO; WILK, 1965; YAP; SIM, 2011).

O teste de Shapiro-wilk se baseia na regressão e correlação entre os dados fornecidos e seu escore normal correspondente[27]. Possuía originalmente um limite para amostras de tamanho 50, sendo posteriormente expandido, por Royston (ROYSTON, 1982, 1995), para amostras de tamanho compreendido entre 3 e 5000.

A hipótese nula para esse teste de normalidade é que a população possui distribuição normal. Sendo assim, um valor significante para a estatística de teste W e o valor p (nível de significância do teste) indica que a hipótese nula deve ser rejeitada e que os dados não possuem distribuição normal[28]. Para que a hipótese nula seja aceita, o valor de W deve ser igual ou próximo a 1, geralmente não deve ser menor que 0,99[29].

Para testar a normalidade dos dados foi utilizado o pacote Shapiro-wilk da biblioteca ggplot2, do *software* r de estatística. Exemplos de códigos utilizados estão dispostos na tabela A3.

[27] http://dergipark.gov.tr/download/article-file/129239
[28] http://dergipark.gov.tr/download/article-file/129239
[29] http://emilkirkegaard.dk/en/?p=4452

Tabela 8 - Códigos utilizados para testar a normalidade dos dados

Tarefa	Código
Histograma de Densidade	qplot(dataDeath$age.of.death, geom = 'blank', xlab = "Age of Death (Years)", ylab = "") + geom_line(aes(y = ..density..), stat = 'density') + geom_histogram(aes(y = ..density..), alpha = 0.4, binwidth = 3, col=I("darkgrey")) + theme_bw() + labs(title="Frequencies of Age of Death") + theme(plot.title = element_text(hjust = 0.5))
Shapiro-Wilk	shapiro.test(data$age.of.onset)
QQ-plot	qqnorm(dataOnset$age.of.onset, pch=1, cex=1, main="Age of Onset", xlab = "") + qqline(dataOnset$age.of.onset)
Teste U de Mann-Whitney	wilcox.test(data[which(data$DISTANT == "DISTANT"),]$age.of.death, data[which(data$DISTANT != "DISTANT"),]$age.of.death)
Boxplot	ggplot(dataOnset, aes(x=dataOnset$DISTANT, y=dataOnset$age.of.onset)) + geom_boxplot() + theme_bw() + labs(title="Age of Onset\np-value = 0.24", x="\n", y="Age of Onset (Years)") + theme(axis.text=element_text(size=12), axis.title=element_text(size=16)) + theme(plot.title = element_text(hjust = 0.5)) + theme(text = element_text(size = 12))

Exemplo de códigos utilizados para execução do teste de normalidade, correlações estatísticas de Mann-Whitney e construção dos *boxplots.*

Com o objetivo de analisar a distribuição, os dados referentes às idades e tempo de sobrevivência foram plotados.

Para essa tarefa o histograma é provavelmente a forma mais simples, antiga e comum de visualização gráfica, dividindo o intervalo de dados em barras e os plota de modo que a altura de cada barra corresponde a quantidade de dados, permitindo resumir a distribuição dos dados[30].

O problema de tirar conclusões apenas com esse tipo de representação é que a distribuição de dados pode ser arbitrária, dependendo da escolha da largura das barras (VENU; LOLLA; HOBEROCK, 2019).

Foi utilizada a função qplot da biblioteca ggplot2 do *software* r para plotar os histogramas de densidade (exemplo de código mostrado na Tabela 8).

[30] http://faculty.nps.edu/rdfricke/Stats_for_Biosurv_Cse/Chapter%204.pdf

Um histograma de dados normalmente distribuídos deve expressar uma curva simétrica em forma de sino, a curva é centralizada na média dos dados e se propaga pelo desvio padrão[31].

O gráfico quantil-quantil (Q-Q plot) é provavelmente a melhor representação gráfica para testar se a distribuição é normal ou não. Através deste tipo de gráfico é possível estimar a distribuição dos dados, a qual depende em como a dispersão dos pontos (dados) se difere da linha da distribuição normal. Para que os dados sejam considerados normais os pontos onde todos os pontos devem sobrepor a linha ou estarem perto desta, de modo a seguir sua forma e ficam sobre ou próximo da linha reta traçada no meio dos pontos de dados. O Q-Q plot fornece uma representação na qual os quantis da distribuição dos dados são postos contra os quantis da distribuição normal[32,33]. Para isso foi utilizada a função qqplot da biblioteca ggplot2 do *software* r de estatística (exemplo de código disposto na Tabela 8).

Resultados do Teste de Normalidade dos Dados

Considerando a curva da distribuição de dados normais, é possível notar que a curva das idades de início da ALS é mais alongada e estreita em comparação à normal (Figura 6). Enquanto a curva da idade de morte foge à forma da normal especialmente na sua extremidade superior e na cauda esquerda, enquanto a curva para tempo de sobrevivência difere totalmente da curva da distribuição da normal.

Apesar da densidade dos dados apresentar divergência à distribuição normal ainda se faz necessária a análise dos valores p (Tabela 9) do teste de normalidade e os gráficos Q-Q (Figura 33), considerando a arbitrariedade da distribuição dos dados apresentada nos histogramas e curvas de densidade dos dados.

Observando os gráficos Q-Q (Figura 33), os quais testam o quantil dos dados de idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes contra o quantil teórico da distribuição normal e desenha uma linha na diagonal onde a maior parte dos dados se concentram. Os dados são então dispersos ao longo da linha e esta por sua vez é comparada com a linha da distribuição normal. Dessa forma é possível notar que a linha dos dados de idade

[31] http://www.stat.yale.edu/Courses/1997-98/101/normal.htm
[32] http://dergipark.gov.tr/download/article-file/129239
[33] http://support.sas.com/documentation/cdl/en/procstat/67528/HTML/default/viewer.htm#procstat_univariate_syntax30.htm

de início se aproxima da linha da normal e o p valor é maior que 0,05 respectivamente (Tabela 9), dessa forma é possível concluir que a hipótese nula da distribuição normal é aceita.

A dispersão dos dados de idade de óbito se aproximam da linha da distribuição normal e o p valor menor que 0,05 (Tabela 9 e Figura 33), sendo assim, a hipótese nula de normalidade é rejeitada. Assim como ocorre para os dados de tempo de sobrevivência, para os quais a dispersão destoa totalmente da linha da normal.

Figura 33 - Gráficos para Teste de Normalidade dos Dados: QQ plot

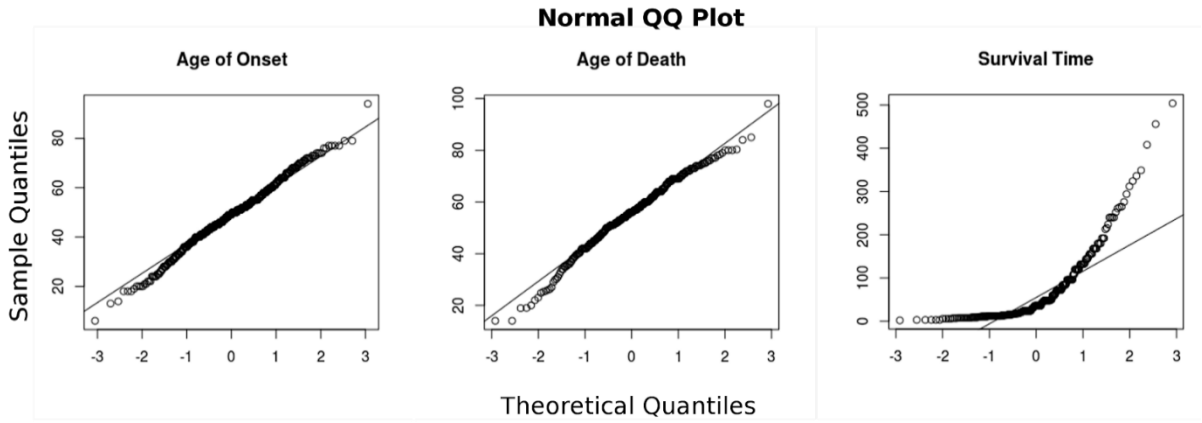


Gráfico Q-Q do teste de normalidade dos dados de idade de início da doença, idade de óbito e tempo de sobrevivência dos pacientes portadores de mutações *missense* na SOD1 descritos na base de dados DynAMISM.

Tabela 9 - Teste de Normalidade Shapiro-Wilk

Dados	p-value
Age of Onset	0,0689
Age of Death	0,01866
Survival Time	2,2e-16

Nível de significância (valor p) do teste de normalidade de Shapiro-Wilk. O nível de significância considerado é de 0,05. Sendo assim, quando o valor calculado for maior que 0,05 a hipótese nula de normalidade dos dados é aceita ao nível de 5% de significância. Se o valor for menor que 0,05 então a hipótese nula é rejeitada ao nível de 5% de significância.

10.5 Apêndice V - Teste U de Mann-Whitney

O teste U de Mann-Whitney, também conhecido como Wilcoxon *rank sum test,* é frequentemente utilizado em dados que apresentam distribuição não normal, como uma alternativa não-paramétrica ao teste t e lidando bem com *outliers* (HART, 2001; WU et al., 2014).

O teste U se baseia no princípio que se duas amostras são extraídas de uma população idêntica e os escores são distribuídos em uma sequência única, então o maior e menor posto devem estar localizados igualmente entre as duas amostras. Mas se a predominância entre esses postos maior e menor não é igual, então provavelmente as populações não são idênticas[34].

Neste teste a comparação acontece entre dois grupos, em termos de localização e forma, de modo a testar se um dado apresenta valores maior que o outro, computando a diferença em médias. Se as médias são similares para a maior parte dos dados, o teste pode também reportar diferenças na dispersão, o que é muito importante, além de definir o valor p (HART, 2001).

Se o valor de p é pequeno (menor que 0,05), então a hipótese nula, que diz que qualquer possível diferença entre os dados se deve à uma amostragem aleatória e considera as populações como distintas uma da outra, é rejeitada. Por outro lado, se o p valor é alto (maior que 0,05), então a hipótese nula é aceita e se considera que os dados não apresentaram diferenças significativas.[35]

O teste U de Mann-Whitney está disponível como uma função para o teste de Wilcox da biblioteca *stats* do *software* r de estatística e foi utilizado com base no código disponível na Tabela 8.

[34] https://psych.unl.edu/psycrs/handcomp/hcmann.PDF
[35] https://www.graphpad.com/guides/prism/7/statistics/how_the_mann-whitney_test_works.htm?toc=0&printWindow

10.6 Apêndice VI - Sequências Fasta de Aminoácidos das Proteínas SOD1, TDP-43 e FUS/TLS

SOD1

>sp|P00441|SODC_HUMAN Superoxide dismutase [Cu-Zn] OS=Homo sapiens OX=9606 GN=SOD1 PE=1 SV=2

MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTS
AGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLVV
HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ

TDP-43

>sp|Q13148|TADBP_HUMAN TAR DNA-binding protein 43 OS=Homo sapiens OX=9606 GN=TARDBP PE=1 SV=1

MSEYIRVTEDENDEPIEIPSEDDGTVLLSTVTAQFPGACGLRYRNPVSQCMRGVRLVEGI
LHAPDAGWGNLVYVVNYPKDNKRKMDETDASSAVKVKRAVQKTSDLIVLGLPWKTTEQDL
KEYFSTFGEVLMVQVKKDLKTGHSKGFGFVRFTEYETQVKVMSQRHMIDGRWCDCKLPNS
KQSQDEPLRSRKVFVGRCTEDMTEDELREFFSQYGDVMDVFIPKPFRAFAFVTFADDQIA
QSLCGEDLIIKGISVHISNAEPKHNSNRQLERSGRFGGNPGGFGNQGGFGNSRGGGAGLG
NNQGSNMGGGMNFGAFSINPAMMAAAQAALQSSWGMMGMLASQQNQSGPSGNNQNQGNMQ
REPNQAFGSGNNSYSGSNSGAAIGWGSASNAGSGSGFNGGFGSSMDSKSSGWGM

FUS/TLS

>sp|P35637|FUS_HUMAN RNA-binding protein FUS OS=Homo sapiens OX=9606 GN=FUS PE=1 SV=1

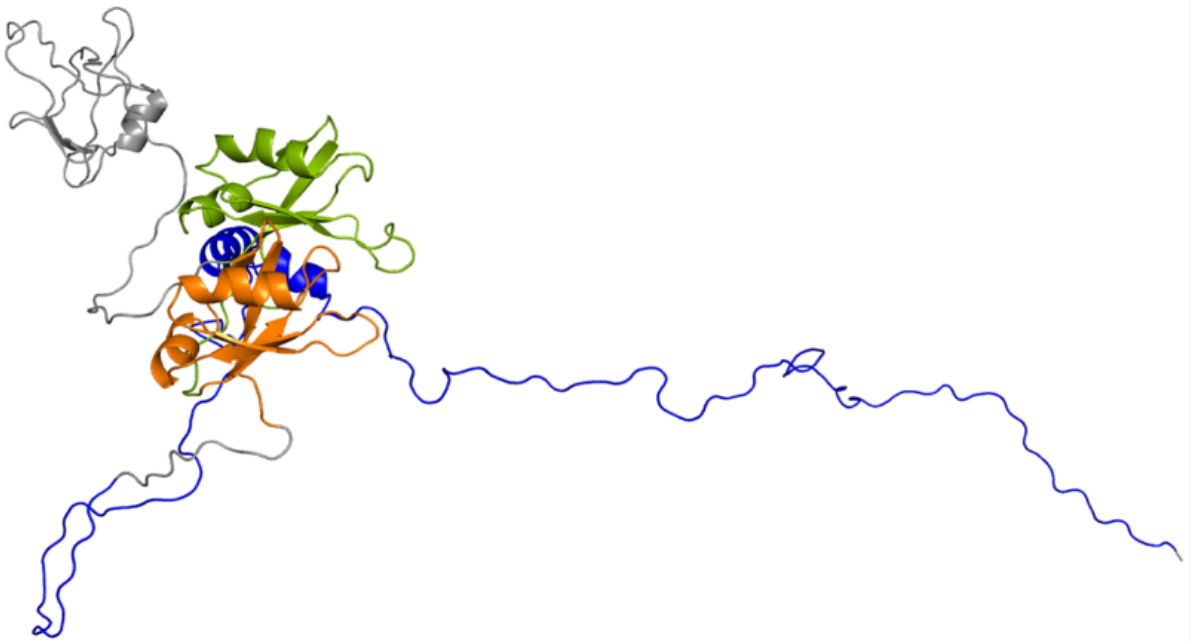
MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQ
SQNTGYGTQSTPQGYGSTGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSS
SYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNSSSGGGGGGGGGGNYGQD
QSSMSSGGGSGGGYGNQDQSGGGGSGGYGQQDRGGRGRGGSGGGGGGGGGGYNRSSGGYE
PRGRGGGRGGRGGMGGSDRGGFNKFGGPRDQGSRHDSEQDNSDNNTIFVQGLGENVTIES
VADYFKQIGIIKTNKKTGQPMINLYTDRETGKLKGEATVSFDDPPSAKAAIDWFDGKEFS
GNPIKVSFATRRADFNRGGGNGRGGRGRGGPMGRGGYGGGGSGGGGRGGFPSGGGGGGGQ
QRAGDWKCPNPTCENMNFSWRNECNQCKAPKPDGPGGGPGGSHMGGNYGDDRRGGRGGYD
RGGYRGRGGDRGGFRGGRGGGDRGGFGPGKMDSRGEHRQDRRERPY

10.7 Apêndice VII – Modelo Tridimensional da TDP-43

A estrura tridimensional da TDP-43 (Figura 34) foi obtida via modelagem por homologia utilizando como molde estruturas disponíveis no PDB resolvidas por NMR, conforme as respectivas coberturas na sequência alvo com 100% de identidade: 2N3X (resíduos 311 a 360), 5MRG (resíduos 1 a 10) e 4BS2(resíduos 102 a 269, domínios RRM1 e RRM2).

A metodologia utilizada para a modelagem por homologia permitiu gerar 20 modelos tridimensionais, dentre os quais o modelo final foi escolhido conforme o menor valor do DOPE (1.14698) e maior percentual de resíduos dentro de regiões favorecidas do gráfico de Ramachandran (Figura 35) e menor percentual dentro de regiões proíbidas.

Figura 34 - Modelo Tridimensional da TDP-43 e Domínios Funcionais



Modelo tridimensional da TDP-43. A figura mostra o modelo tridimensional da estrutura da proteína TDP-43 obtido através de modelagem por homologia. Em diferentes cores estão destacados os domínios funcionais da proteína: RRM1 (verde), RRM2 (laranja) e região rica em glicina (azul). Escore do DOPE: 1.14698.

Figura 35 – Gráfico de Ramachandran para o Modelo Tridimensional da TDP-43

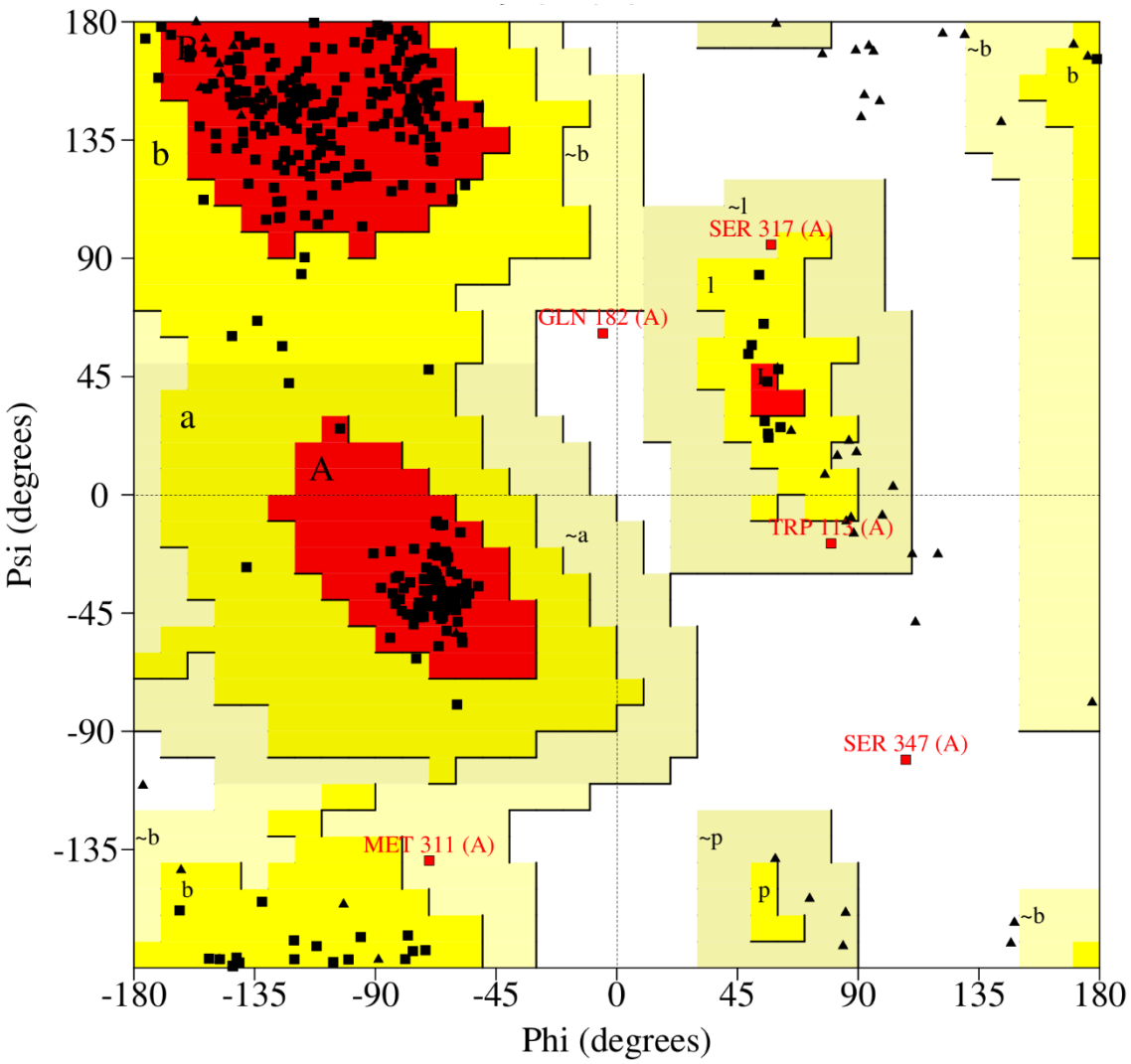


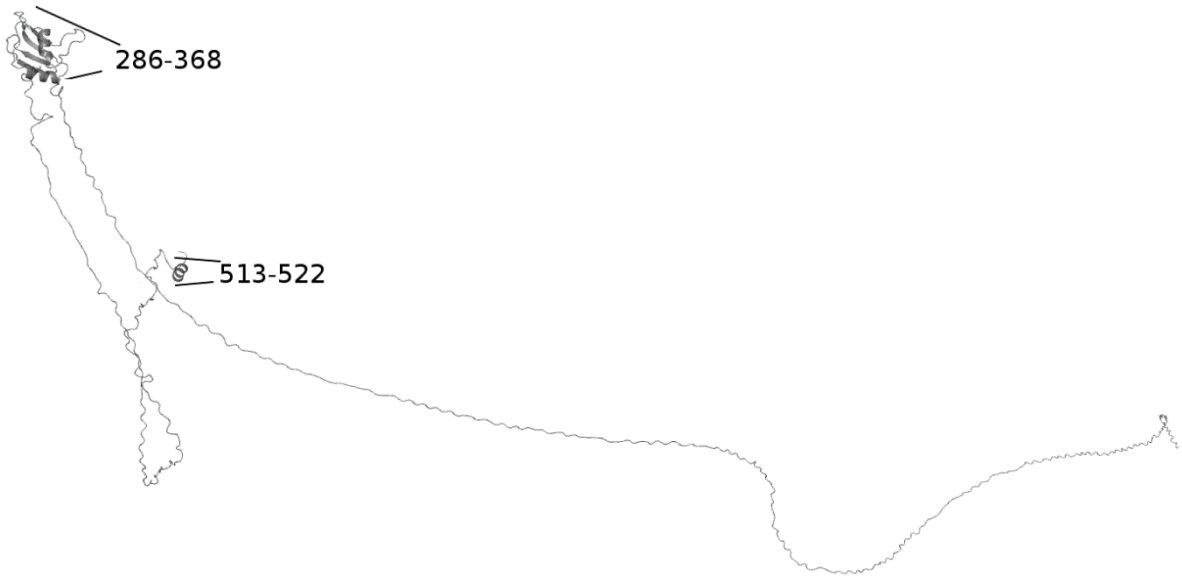
Gráfico de Ramachandran mostrando as restrições espaciais para os resíduos do modelo obitido, conforme os graus de rotação phi-psi. O modelo obtido possui 86,8% dos resíduos em regiões favorecidas, 12,6% em regiões permitidas e 0,6% em regiões proibidas do gráfico de Ramachandran. As regiões em vermelho contendo a letra ‘A’, ‘B’ ou ‘L’ representam as regiões favorecidas; as regiões permitidas estão em amarelo com as letras ‘a’,b’,’l’ ou‘p’; regiões pouco permitidas (ou menos favoráveis) estão em amarelo claro com a representação ‘~a’,’~b’, ‘~l’ ou ‘~p’ e regiões proibidas são representadas em branco. Triângulos representam resíduos de glicina e quadrados representam os demais resíduos. Os resíduos destacados em vermelho estão em regiões pouco favoráveis (conforme as combinações dos ângulos diédricos phi-psi e ocorrência de impedimento estérico) ou proibidas, enquanto os resíduos em símbolos na cor preta estão em suas regiões favoráveis.

10.8 Apêndice VIII – Modelo Tridimensional da FUS/TLS

A estrura tridimensional da FUS/TLS (Figura 36**Figura *36***) foi obtida via modelagem por homologia utilizando como molde estruturas disponíveis no PDB resolvidas por NMR, conforme as respectivas coberturas e identidade com a sequência alvo: 2LA6 (resíduos 1 a 99, 100% de identidade), 2LCW (resíduos 278 a 385, 100% de identidade), 5YVG (resíduos 509 a 526, 100% de identidade) e 2CPE (resíduos 346 a 458, 56,3% de identidade).

A metodologia utilizada para a modelagem por homologia permitiu gerar 20 modelos tridimensionais, dentre os quais o modelo final foi escolhido conforme o menor valor do DOPE (1.66751) e maior percentual de resíduos dentro de regiões favorecidas do gráfico de Ramachandran (Figura 37) e menor percentual dentro de regiões proíbidas.

Figura 36 – Modelo Tridimensional da FUS/TLS



Modelo tridimensional da FUS/TLS. A figura mostra o modelo tridimensional da estrutura da proteína FUS/TLS obtido através de modelagem por homologia. As regiões destacadas na figura mostram as únicas porções do modelo onde foi possível obter estrutura. Escore do DOPE: 1.66751.

Figura 37 – Gráfico de Ramachandran para o Modelo Tridimensional da FUS/TLS

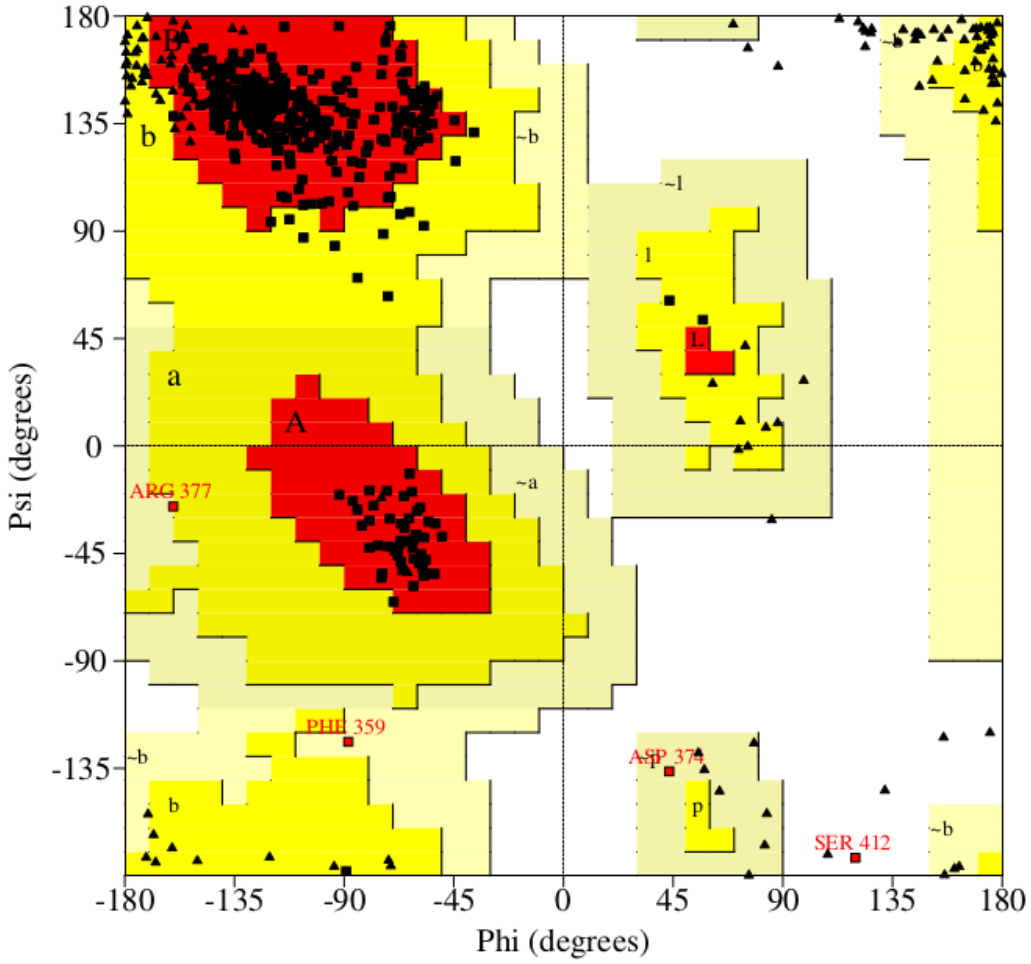


Gráfico de Ramachandran mostrando as restrições espaciais para os resíduos do modelo obtido, conforme os graus de rotação phi-psi. O modelo obtido possui 94,8% dos resíduos em regiões (ou ângulos) favorecidas, 5% em regiões permitidas e 0,3% em regiões proibidas do gráfico de Ramachandran. As regiões em vermelho contendo a letra 'A', 'B' ou 'L' representam as regiões favorecidas; as regiões permitidas estão em amarelo com as letras 'a','b','l' ou'p'; regiões pouco permitidas (ou menos favoráveis) estão em amarelo claro com a representação '~a','~b', '~l' ou '~p' e regiões proibidas são representadas em branco. Triângulos representam resíduos de glicina e quadrados representam os demais resíduos. Os resíduos destacados em vermelho estão em regiões pouco favoráveis (conforme as combinações dos ângulos diédricos phi-psi e ocorrência de impedimento estérico) ou proibidas, enquanto os resíduos em símbolos na cor preta estão em suas regiões favoráveis.