Dinâmica molecular aplicada no estudo de doença autoimune e aprendizado de máquina aplicado a uma base pública de dados da saúde

Bolsista Mariangela Dametto (CTI) mdametto@cti.gov.br, Rodrigo Bonacin (CTI) rodrigo.bonacin@cti.gov.br

Resumo

A interação do domínio SAND da proteína AIRE com a SIRT1 está envolvida na seleção de clones de timócitos auto agressivos, consequentemente, atuando em mecanismos de doenças autoimunes. A mutação G228W no domínio SAND tem sido descrita como responsável pelo mecanismo que causa uma síndrome autoimune conhecida como APS-1. Neste trabalho, realizamos simulações de dinâmica molecular do complexo SAND-AIRE mutado e sem mutação para compreensão de como as alterações moleculares interferem com a ligação e o desligamento da interação entre as duas proteínas. Essas informações foram reunidas junto com experimentos de bancada de um grupo do Departamento de Genética e Biologia Molecular da Medicina da USP/Ribeirão Preto.

Concomitantemente aos trabalhos supracitados, nosso grupo realizou simulações de aprendizado de máquina com uma base de dados nacional depositada na Fundação ONCOCENTRO de São Paulo (FOSP). Em relação ao ano passado a base foi atualizada e analisada por outros algoritmos supervisionados e não supervisionados. Este tem sido um trabalho contínuo com a finalidade de construir modelos de classificação para predição de recorrência dos diversos tipos de câncer com base em características chaves dos pacientes. Além disso, a mesma base foi utilizada com algoritmos de clusterização para o câncer de próstata, com o intuito de obter informações sobre padrões que possam ser relacionados envolvendo as características dos pacientes com câncer de próstata e recidiva.

Palavras-chave: Aprendizado de Máquina, Algoritmos Supervisionados, Classificação, Recidiva de tipos de Câncer, Clusterização, Simulações de Dinâmica Molecular, SAND-AIRE e Síndrome autoimune APS-1, Interação entre proteína ACE Humana e Proteína SPIKE do Coronavírus.

1. Introdução

A proteína regulatória autoimune (do inglês, AIRE) atua juntamente com outras proteínas para formar o complexo AIRE, que é um controlador de transcrição genética, promovendo a expressão de um conjunto de genes nos tecidos periféricos envolvidos com a seleção negativa de timócitos autorreativos (AALTONEN et al, 1994; BLECHSCHMIDT et al, 1999; NAGAMINE et al, 1997). Esta seleção contribui para evitar a ocorrência de doenças autoimunes. Neste trabalho, estudou-se especificamente a interação entre o domínio SAND da proteína AIRE e a proteína Sirtuina 1 (SIRT-1), envolvida também na regulação da ativação de transcrição gênica. Trabalhos anteriores (CETANI et al., 2001; ILMARINEN et al, 2005;

SU et al, 2008) indicaram que a interação entre a SIRT1 e o domínio da AIRE com a mutação G228W causava uma doença autoimune conhecida como síndrome poliglandular autoimune tipo 1 (do inglês, APS-1). Portanto, o grupo do Dr. Geraldo Aleixo Passos (Departamento de Genética e Biologia Molecular da Medicina da USP/Ribeirão Preto), que estuda doenças genéticas autoimunes, formou uma colaboração com nosso grupo para que fizéssemos a parte de simulações de dinâmica molecular do complexo domínio SAND da AIRE e a SIRT1. Como consequência, os resultados computacionais corroboraram os experimentos de bancada (SANTOS et al, 2022). Com as simulações atomísticas, foi observada molecularmente a importância da mutação na AIRE-SAND no sentido de promover a desacetilação de uma lisina, importante para a atividade enzimática funcional da AIRE, além de se observar a aproximação do sítio catalítico de uma histidina (H363) que não era descrita ter importância no processo de interação das moléculas, o que acaba por originar a doença autoimune acima mencionada. Os resultados das simulações estão descritos na seção de resultados deste artigo.

Na parte do projeto relacionada à ciência de dados, algoritmos supervisionados de classificação (Naive Bayes, SVM, dentre outros) e de clusterização têm sido aplicados na base da Fundação ONCOCENTRO de São Paulo (FOSP) (http://www.fosp.saude.sp.gov.br/), após pré-processamento dos dados brutos. Com relação aos algoritmos de classificação, o Naive Bayes e o SVM apresentaram resultados mais promissores (MAEDA et al, 2022), sendo que com o Naive Bayes foram obtidos acurácia de 63% e F1 score 0.76 e com o SVM a acurácia foi de 66% e F1 de 0.78. Estes valores foram calculados após ser realizado o balanceamento dos dados com o método SMOTE, visto que os dados se encontram bastante desbalanceados em relação à classe (recidiva e não-recidiva). Espera-se que, após a criação de modelo de classificação, seja possibilitado a construção de modelo de predição de recidiva de acordo com as características dos pacientes, incluindo o tipo de câncer específico. Já para o estudo de clusterização, foi analisado o câncer de próstata e foi aplicado o algoritmo K-Prototypes com separação inicial em 5 clusters, e interessantemente, foram encontrados 2 clusters que apresentam 57% dos casos de recidiva de todo o conjunto de dados analisados nesta parte do trabalho (CROCCO et al, 2022). E ainda mais, cada cluster descrevendo predominantemente duas morfologias diferentes, o Adenocarcinoma SOE (faixa etária entre 64 e 75 anos) e o Carcinoma de Células Acinosas (faixa etária entre 58 e 63 anos). Estes resultados são muito recentes e ainda sob interpretação.

2. Metodologia

As estruturas tridimensionais dos complexos SIRT1 e o domínio SAND da AIRE (tanto a selvagem quanto a mutante) foram cedidos pelo grupo do Dr. Passos para que fossem realizadas as simulações biomoleculares para verificar a estabilidade e a afinidade dos dois diferentes complexos. A lisina (posição 222 nas estruturas tridimensionais utilizadas) foi acetilada usando os parâmetros descritos no pacote de dinâmica molecular AMBER18 (CASE et al, 2022), e a conectividade entre este resíduo específico (nomeada ACK no software) e o restante da molécula foi gerado por meio do módulo tleap. Os dois sistemas foram solvatados por meio do modelo TIP3P e o campo de forças ff19SB foram utilizados para preparar os sistemas para posterior realização das simulações de dinâmica molecular. Ainda antes das simulações de produção em si, foram feitos a minimização de energia dos sistemas e a fase de termalização do sistema para a temperatura de 300K, usando termostato Langevin e o algoritmo SHAKE para restringir a vibração das interações onde há hidrogênio envolvido. A dinâmica de produção foi realizada com pressão constante e time step de 2 fs, sendo que o tempo total de simulação foi de 2100 ns com frames gravados a cada 20 ps. A extração dos

resultados ocorreu nos últimos 2050 ns, ou seja, após a fase de equilibração dos sistemas, e foi feita por meio dos scripts presentes no pacote de ferramentas AMBERTOOLS21.

A metodologia básica no estudo da COVID19 foi a mesma descrita acima, com a diferença que o tempo total de simulação de cada um dos 5 sistemas foi de 500 ns e não foi feita alteração nas estruturas tridimensionais cedidas pelo grupo do Dr. Passos.

No treinamento do modelo de classificação utilizando-se um dataset da base da FOSP, foi utilizada a proporção 80/20 para treino e teste, a técnica de K-fold para validação cruzada e o método SMOTE para balanceamento das instâncias da classe (recidiva e não-recidiva). Já para o estudo de clusterização do câncer de próstata (também realizado com a base de dados da FOSP), o algoritmo aplicado foi o K-prototypes numa faixa de 1 a 5 clusters. Nestas duas partes do trabalho foi utilizada a plataforma scikit-learn (PEDREGOSA et al, 2011).

3. Resultados

Os resultados da dinâmica molecular até 1μs de tempo de simulação não demonstraram a existência de estrutura estável do complexo com o domínio SAND-AIRE sem mutação, mesmo com a lisina K222 acetilada. Por outro lado, o domínio mutado G228W da SAND-AIRE mostrou uma interação com maior afinidade com a histidina (H363) do sítio catalítico da SIRT1, sendo que esta histidina apresenta importante papel na reação enzimática necessária para desencadeamento do mecanismo natural do organismo contra prejudiciais reações autoimunes.

A Figura 1 mostra o cálculo de RMSD (do inglês, Root Mean Square Deviation) que se refere ao posicionamento e à movimentação dos átomos das moléculas (e das interações entre as moléculas) durante o tempo de simulação atomística de dinâmica molecular. Este gráfico mostra relativa maior estabilidade do complexo com a forma mutada G228W da SAND-AIRE e a SIRT1, sendo que um valor maior de RMSD (eixo y) significa menor estabilidade.

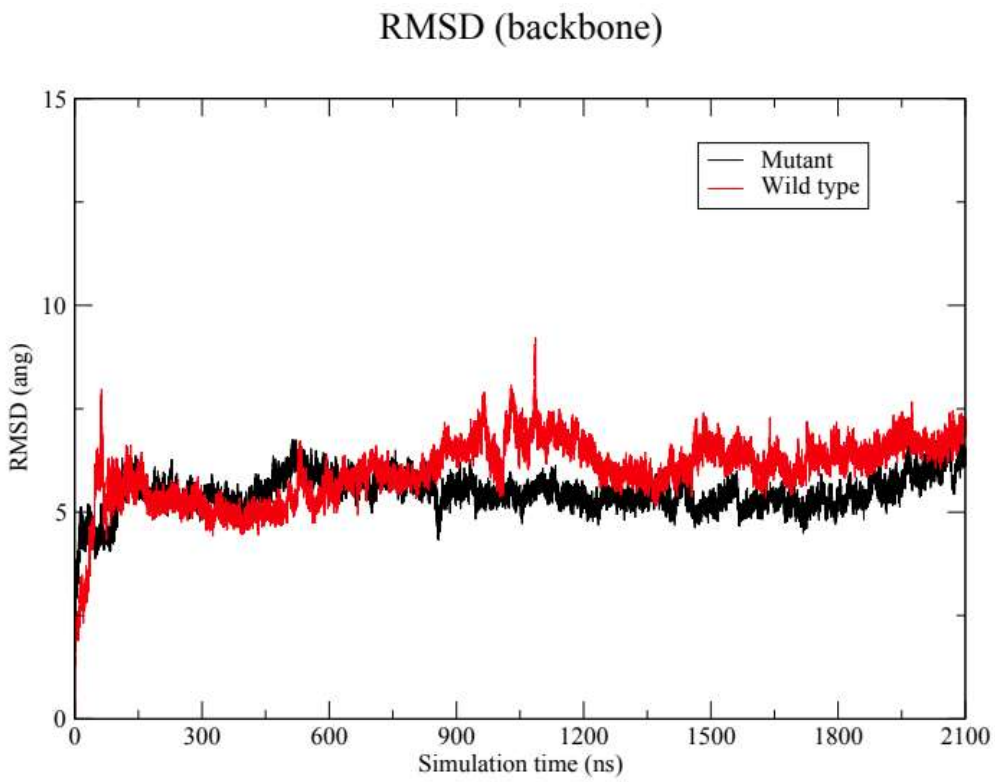

Figura 1 - RMSD após resultado da obtenção análise da trajetória da dinâmica molecular da interação entre o domínio da AIRE-SAND selvagem (em vermelho) e a mutada G229W (em preto). O gráfico mostra que ambos os complexos apresentam comportamento semelhante até os 600 ns iniciais da simulação. Entre os 600 e 900 ns, os complexos apresentam certa diferença na estabilidade, mas ambos tendem para a estabilidade após 1200 ns, apesar de que o complexo com a mutação apresenta maior instabilidade durante maior parte da simulação em relação ao complexo sem mutação (Figura extraída de SANTOS, 2020).

Em relação aos resultados referentes à base de dados da FOSP com aplicação de aprendizado de máquina, o algoritmo de classificação que gerou melhor modelo com a maior e mais recente base de pacientes foi o Naive Bayes (Tabela 1). Nesta parte do trabalho, foram considerados todos os tipos de câncer existentes na base de dados.

Naive Bayes	***80/20***	***K-fold (Média)***	***Smote***
Acurácia	0.9100436805756901	0.9109144402564749	0.6344758093264445
Precisão	0.9104839630179203	0.9118629773790585	0.9484022372589667
F1 Score	0.9528981359348945	0.9533671866601054	0.7592123513922243

Fonte: Extraída de MAEDA et al, 2022

Tabela 1 - Resultados obtidos para treinamento do modelo com Naive Bayes

Contudo, o SVM também mostrou bons indicadores agora com esta base maior (Tabela 2).

SVM	*80/20*	*K-fold (Média)*	*SMOTE*
Acurácia	0.9104286780328056	0.9920838183934806	0.6654894933718648
Precisão	0.9104286780328056	0.9920838183934807	0.9469494651420141
F1 Score	0.9531145428263634	0.9960261598875692	0.7848399246704332

Fonte: Extraída de MAEDA et al, 2022.

Tabela 2 - Resultados obtidos para cada estratégia de treinamento do modelo com SVM

A aplicação de algoritmo de clusterização para 1 a 5 clusters, considerando-se somente o câncer de próstata, indicou a formação de 2 clusters mais predominantes (Tabela 3).

CLUSTER	IDADE	TRATCON	DIAGTRA	ESCOLAI	MORFO	BASEDIAG	EC	CATEATEN	PERDASEG	T	N	M	G	PSA	GLEASON	TRATHOSP	TRATFAPOS	RECLOCAL
1	63	126	167	9	85503	3	IIB	2	0	2C	0	0	8	1	2	A	J	0
2	68	86	143	9	85503	3	IIA	2	0	1C	0	0	8	1	1	I	J	0
3	68	120	176	2	85503	3	III	2	0	3	0	0	8	8	8	I	J	0
4	64	90	104	2	85503	3	II	9	0	2C	0	0	8	8	8	A	J	0
5	70	67	109	9	81403	3	II	2	0	2	0	0	8	8	8	I	J	0

Fonte: Extraída de CROCCO et al, 2022.

Tabela 3 - Centroides gerados para os 5 clusters

4. Conclusões e Discussão

Com relação ao trabalho realizado com o complexo AIRE-SIRT1, não foi observada variação na estabilidade da interação durante a simulação de dinâmica, seja do complexo com o aminoácido G228W mutado, seja do complexo sem esta referida mutação. Assim, o domínio SAND-AIRE mutado pode interagir da mesma forma com a SIRT1, contudo, ele provoca uma outra interação com a histidina H363, que está descrita ser importante no desencadeamento da castata contra autoimunoreatividade (SANTOS, 2020). Esta interação com a H363 somente pôde ser observada por meio da simulação atomística, que estuda os complexos biológicos, no caso, molecularmente.

Os resultados referentes ao modelo de classificação, para todos os tipos de câncer e com as características dos pacientes mencionadas anteriormente, são promissores no sentido de se construir um modelo que classifique corretamente os casos de recidiva e de não-recidiva de acordo com estas características específicas. A criação deste modelo auxiliará na tomada de decisão pelos profissionais de saúde visando um melhor tratamento no sentido que não haja recidiva futura pós tratamento.

O trabalho de clusterização dos pacientes com câncer de próstata também apresentou resultados interessantes, pois indicou a formação de 2 clusters, cada um deles com predominância de diferentes morfologias tumorais e idades dos pacientes. Sendo este um resultado muito recente, está sendo analisado e interpretado do ponto de vista biomédico.

Referências

AALTONEN J, BJÖRSES P, SANDKUIJL L, PERHEENTUPA J, PELTONEN L. *An autosomal locus causing autoimmune disease: autoimmune polyglandular disease type 1 assigned to chromosome 21*. Nat Genet 8:83–7, 1994.

BLECHSCHMIDT K, SCHWEIGER M, WERTZ K, POULSON R, CHRISTENSEN HM, ROSENTHAL A, et al. The mouse aire gene: comparative genomic sequencing, gene organization, and expression. Genome Res 9:158–66, 1999.

CASE DA, AKTULGA HM, BELFON K, BEN-SHALOM IY, BERRYMAN JT, BROZELL SR, et al. AMBER. San Francisco: University of California (2022). Available at: https://ambermd.org/AmberTools.php

CETANI F, BARBESINO G, BORSARI S, PARDI E, CIANFEROTTI E, PINCHERA A, et al. A novel mutation of the autoimmune regulator gene in an Italian kindred with autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy, acting in a dominant fashion and strongly cosegregating with hypothyroid autoimmune thyroiditis. J Clin Endocrinol Metab 86:4747–52, 2001.

CROCCO PF, MAEDA AE, RUPPERT GCS, DAMETTO M, BONACIN *Clusterização de Dados Abertos em Oncologia Usando Técnicas de Aprendizagem de Máquina: um estudo preliminar sobre recidiva de câncer de próstata*, XXIV Jornada de Iniciação Científica do Centro de Tecnologia da Informação Renato Archer - JICC´2022. PIBIC/CNPq/CTI

ILMARINEN T, ESKELIN P, HALONEN M, RUPPELL T, KILPIKARI R, TORRES GD, et al. Functional analysis of SAND mutations in AIRE supports dominant inheritance of the G228W mutation. Hum Mutat 26(4):322–31, 2005.

LIU Z, XIAO X, WEI X, et al. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. J Med Virol. 92:595–601, 2020.

MAEDA AE, CROCCO PF, RUPPERT GCS, DAMETTO M, BONACIN R *Um Estudo sobre a Predição da Recidiva de Câncer Usando Técnicas de Aprendizagem de Máquina*, XXIV Jornada de Iniciação Científica do Centro de Tecnologia da Informação Renato Archer - JICC´2022. PIBIC/CNPq/CTI

NAGAMINE K, PETERSON P, SCOTT HS, KUDOH J, MINOSHIMA S, HEINO M, et al. Positional cloning of the APECED gene. Nat Genet 17(4):393–8, 1997.

PEDREGOSA et al, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011

SANTOS JC, DAMETTO M, MASSON AP, FAÇA VM, BONACIN R, DONADI EA and PASSOS GA . Front. Immunol. 13:948419, 2022.

SU MA, GIANG K, ZUMER K, JIANG H, OVEN I, RINN JL, et al. Mechanisms of an autoimmunity syndrome in mice caused by a dominant mutation in aire. J Clin Invest (2008) 118(5):1712–26.