EAMC2022

XV Encontro Acadêmico de Modelagem Computacional

Laboratório Nacional de Computação Científica - LNCC/MCTI

22 a 25 de Fevereiro de 2022.

Proceedings do evento

XV Encontro Acadêmico de Modelagem Computacional

Laboratório Nacional de Computação Científica - LNCC/MCTI

22 a 25 de Fevereiro de 2022. Petrópolis - RJ.

Comitê Organizador

Ana Luiza Martins Karl - LNCC

Andressa Alves Machado - LNCC

Guilherme Guilhermino Neto - UFJF

Gustavo Alves Bezerra - LNCC

João Vitor de Oliveira - UFRJ

Luis Alonso Mansilla Alvarez - LNCC

Luis Fernando Mendes Cury - UnB

Matheus Muller Pereira da Silva - LNCC

Rafael Terra - LNCC

Rennan Mendes dos Santos Dias - UFF

Comitê Científico

Caio César Graciani Rodrigues - LNCC

Camila de Oliveira Vieira - UnB

Gregório Kappaun Rocha - IFF

Guilherme Guilhermino Neto - UFJF

Karina Baptista dos Santos - LNCC

Lucas de Asis - IFES

Lucas dos Santos Fernandez - LNCC

Mariza Ferro - LNCC

Weslley da Silva Pereira - UFJF

Apoio

Laboratório Nacional de Computação Científica - LNCC

Ministério da Ciência, Tecnología e Inovações - MCTI

Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ

O Encontro Acadêmico de Modelagem Computacional do LNCC (EAMC/LNCC) é um evento dedicado à Modelagem Computacional, parte do Programa de Verão da mesma instituição, que busca promover maior integração entre os alunos, docentes, pesquisadores e demais profissionais da área de Ciência e Tecnologia. O EAMC é inteiramente organizado por discentes do LNCC e de outras instituições colaboradoras, sendo composto por minicursos, mesas redondas, palestras, além de apresentações de trabalhos em formato pôster e apresentações orais.

Criado em 2007, o EAMC tem como principal objetivo proporcionar um ambiente que promova a divulgação científica, estimule a cooperação entre os profissionais da área, além de fomentar a multidisciplinaridade na formação de novos profissionais. O EAMC abrange diferentes áreas do conhecimento, tais como Computação Científica, Controle e Filtragem de Sistemas Dinâmicos, Modelagem de Biossistemas e Bioinformática, Modelagem de Circulação e Transporte, Modelagem de Equilíbrio e Otimização e, mais recentemente, Ciência de Dados e áreas correlatas, como Inteligência Artificial.

A XV edição do Encontro Acadêmico de Modelagem Computacional foi realizada virtualmente desde as instalações do Laboratório Nacional de Computação Científica, em Petrópolis/RJ, durante os dias 22 a 25 de fevereiro de 2022. Em razão da pandemia do SARS-CoV-2, os minicursos e apresentações orais foram transmitidas de forma virtual visando a segurança de todos os participantes.

O comitê organizador gostaria de agradecer à comunidade do LNCC e instituições parceiras (representadas por pesquisadores, professores, discentes) pela participação e dedicação buscando melhorar o impacto e importâcia do EAMC a cada edição. Particularmente, o nosso agradecimento à organização do Programa de Verão do LNCC, professoras Sandra Malta e Kary Ocaña, os pesquisadores que aceitaram compor o comitê científico e os participantes dessa edição.

Comitê Organizador
XV Encontro Acadêmico de Modelagem Computacional
Laboratório Nacional de Computação Científica - LNCC

Trabalhos completos

Trabalhos em formato pôster

Página

Trabalhos completos

Variações no algoritmo de Yuan para problemas de Equilíbrio de Nash

Amanda Vetorazzi[1] and Luís Felipe Bueno[2]

[1] *ITA-UNIFESP, São José dos Campos/SP, Brasil*
[2] *UNIFESP, São José dos Campos/SP, Brasil*

Resumo

Problemas de Equilíbrio de Nash modelam importantes situações em diversos ramos do conhecimento. Entretanto, os algoritmos para resolver este tipo de problema não são muito bem estabelecidos, principalmente para problemas não convexos. Por este motivo, é importante o estudo de particularidades do comportamento numérico de algoritmos para esta finalidade. Sendo assim, este é um trabalho de caráter exploratório, que almeja compartilhar investigações sobre as consequências práticas de algumas variações no algoritmo de Yuan para problemas de Equilíbrio de Nash.

Palavras-Chaves:Equilíbrio de Nash, algoritmo de Yuan, experimentos numéricos, região de confiança

1 INTRODUÇÃO

Problemas de Equilíbrio de Nash (PENs) visam modelar situações onde pessoas ou entidades interagem entre si, cada um buscando melhorar individualmente seu próprio objetivo, mas com a decisão de um influenciando no retorno dos outros. Por simplicidade, consideraremos dinâmicas que envolvam apenas dois agentes, de modo que um problema de Equilíbrio de Nash seja descrito como a interação entre dois jogadores, representados pela notação $v \in \{1, 2\}$, que possuem como propósito otimizar funções objetivo u_v simultaneamente, mas sem cooperação entre eles. Denotando por $x_v \in \mathbb{R}^{n_v}$ a variável que representa a escolha do jogador v, considera-se $n = \sum_v^2 n_v$. Em [7], quando $n_v > 1$, propõe-se que ao conjunto de estratégias x_v seja atribuída uma distribuição de probabilidade, caracterizando jogos de estratégias mistas. Nesse trabalho as estratégias são analisadas puramente, tratando-as de maneira independente. Nos casos em que as escolhas são qualitativas, atribui-se a estas um vetor de probabilidade com as chances da decisão ser tomada. Com isso, o vetor

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n,$$

Contato: Amanda Vetorazzi, agomesvetorazzi@gmail.com

engloba todas as escolhas dos dois jogadores, de modo que as funções $u_v(x) : \mathbb{R}^n \to \mathbb{R}$ dependem não só das variáveis que o jogador v controla, mas também das variáveis do outro jogador.

Define-se como um Equilíbrio de Nash as estratégias conjuntas formadas por decisões que otimizam a função objetivo de cada jogador, fixadas as estratégias adversárias [7]. Isso significa que, estando em equilíbrio, nenhum dos agentes possui incentivo para trocar de estratégia unilateralmente, ou seja, sem que algum outro jogador também o faça. Para evidenciar as escolhas controladas por v, o vetor x, que contém todas as variáveis do jogo, é decomposto em $x = (x_v, x_{-v})$, de forma que x_{-v} represente as variáveis não controladas pelo jogador v. O ponto $x^* = (x_v^*, x_{-v}^*)$ é declarado como um Equilíbrio de Nash se o jogador v escolhe x_v^* como estratégia ótima para o conjunto fixo de estratégias x_{-v}^*. Assim, para todo jogador, em um Equilíbrio de Nash devemos ter que x_v^* resolve o problema de otimização da forma

$$\min_{x_v \in \mathbb{R}^{n_v}} u_v(x_v, x_{-v}^*). \tag{1}$$

Para verificar se um ponto x^* é um equilíbrio de Nash, resolve-se os problemas de otimização em (1), para cada jogador. Contudo, obter o ponto de equilíbrio não simples, já que é desconhecido x_{-v}^* à priori. Dentre a variedade de algoritmos desenvolvidos para resolver problemas de Equilíbrio de Nash, como por exemplos os revisado em [3], [4] e [5], destacam-se aqueles que usam estratégias do tipo Newton para satisfazer condições necessárias de primeira ordem de (1) e as do tipo Jacobi ou Gauss-Seidel, usando um processo iterativo de melhor resposta. Considerando $g_v(x) = \frac{\partial u_v(x)}{\partial x_v}$, as condições de primeira ordem consistem na resolução do sistema

$$F(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \end{pmatrix} = 0. \tag{2}$$

Usando uma filosofia do tipo Newton, dada uma estimativa de solução x^k, procura-se pelo vetor x que resolva (2), fazendo aproximações lineares de g_v baseadas em x^k. Essas linearizações podem ser interpretadas como derivadas de aproximações quadráticas de u_v. Já em métodos do tipo Jacobi, a estimativa x^{k+1} é obtida de forma a reduzir $u_v(x_v, x_{-v}^k)$.

Os métodos do tipo Newton costumam apresentar bom desempenho computacional, convergindo rapidamente para uma solução do sistema $F(x)$, pelo menos quando um bom ponto inicial é fornecido. Entretanto, para problemas não convexos, não há predileção por minimizadores de (1) em relação a outros pontos estacionários. Já os métodos do tipo Jacobi levam em conta informações de descenso, mas as condições de convergência são limitadas.

Considera-se de especial interesse para nosso trabalho o algoritmo apresentado por Y. X. Yuan em [10], que é elaborado com elementos do tipo Jacobi e aproximações quadráticas de u_v. Com o intuito de garantir, sob certas hipóteses, convergência a partir de um ponto inicial arbitrário, é empregado o uso de regiões de confiança. Neste trabalho, o estudo do algoritmo de Yuan foi delimitado ao caso com dois jogadores em que o conjunto de estratégias não possui restrições e com uma certa matriz de escalamento sendo a identidade.

2 O ALGORITMO DE YUAN PARA PENs

O algoritmo de Yuan para problemas de equilíbrio de Nash faz uso de técnicas clássicas de otimização. As variações entre os pontos de iteração, por exemplo, são controladas

adicionado ao subproblema uma região de confiança, que varia conforme a qualidade dos modelos quadráticos. Em problemas de otimização que utilizam dessa técnica (veja por exemplo [6], [8] e [9]), o ponto é aceito se houver um progresso na função original e a região de confiança é atualizada comparando a redução real obtida e a prevista pelo modelo, por meio de um parâmetro $r_{k,v}$. Entretanto, como os PENs não possuem um único problema de otimização, Yuan formula a função

$$\psi(x) = ||F(x)||^2 = \sum_{v=1}^{2} ||g_v(x)||^2,$$

para verificar se o progresso conjunto foi obtido. Isto é feito por meio de um parâmetro ρ_k que compara o valor de ψ no ponto obtido com o melhor valor de ψ nos pontos anteriores. Além disto, o raio da região de confiança de Yuan é definido, para cada problema, como um múltiplo do gradiente no ponto de iteração corrente. Esta ideia é inovadora inclusive no contexto de otimização tradicional.

No artigo original de Yuan [10], tem-se o desenvolvimento teórico sobre o algoritmo estudado por esse trabalho, entretanto não são realizados testes numéricos sobre o mesmo. Em [2] os autores exibem resultados computacionais obtidos da implementação do algoritmo de Yuan em seis PENs, comparando-o com o Método de Newton e com uma estratégia do tipo Jacobi para resolução de (2). O comportamento numérico obtido indica que o método de Yuan tende a preferir minimizadores em vez de pontos meramente estacionários e que o método pode convergir em situações onde Jacobi não converge. Entretanto, para jogos simples cujas funções objetivo provêm do valor esperado de jogos de estratégias mistas de soma zero, o algoritmo não conseguiu obter a solução. Nas conclusões de [2] os autores argumentam que isto pode ser decorrente da maneira com que o algoritmo de Yuan usa a região de confiança. Desta maneira, nossa contribuição aqui é investigar o comportamento do algoritmo ao realizarmos algumas modificações na definição da região de confiança e nos critérios para atualizá-la. Propomos também uma nova maneira para aceitar o ponto resultante da minimização dos modelos aproximados e averiguamos a influência disto em testes numéricos.

Para apresentação do algoritmo de Yuan são definidos $g_{v,k} = \frac{\partial u_v(x)}{\partial x_v}|_{x=x^k}$, como sendo o gradiente de u_v avaliado no ponto x^k e $B_{v,k} = \frac{\partial^2 u_v(x)}{\partial x_v^2}|_{x=x^k}$, uma aproximação da Hessiana. Com estas informações é possível construir o problema auxiliar, que consiste em

$$\begin{aligned} &\min\ \phi_{v,k}(d_v) = u_v(x^k) + d_v^T g_{v,k} + \frac{1}{2} d_v^T B_{v,k} d_v \\ &\text{s. a } ||d_v||_2 \leq \Delta_{v,k} \\ &\quad x_v^k + d_v \in \mathbb{R}^{n_v}, \end{aligned} \tag{3}$$

referenciado no algoritmo a seguir.

Algoritmo de Yuan para PENs

P0: Defina $k := 0$ e um ponto inicial factível $x^0 \in \mathbb{R}^{n_1 \times n_2}$
Escolha constantes positivas τ_v, δ_v, $t_{v,0}$ $\forall\ v \in \{1, 2\}$,
Escolha $\beta_1 \in (0, 1)$ e $\beta_2 \in (0, 1)$.
Para todo $v \in \{1, 2\}$, calcule $\Delta_{v,0} = \frac{1}{\tau_v + t_{v,0}} ||g_{v,0}||_2$.

P1: Se $\sum_{v=1}^{2}||g_{v,k}||_2^2 = 0$ então pare.
Senão resolva (3) para obter $d_{v,k}$ para cada v.

P2: Compute $Ared_{v,k} = u_v(x_v^k, x_{-v}^k) - u_v(x_v^k + d_{v,k}, x_{-v}^k)$, $Pred_{v,k} = \phi_{v,k}(0) - \phi_{v,k}(d_{v,k})$ e determine

$$r_{v,k} = \frac{Ared_{v,k}}{Pred_{v,k}}.$$

P3: Defina o próximo ponto da iteração x^{k+1} como

$$x_v^{k+1} = \begin{cases} x_v^k + d_{v,k}, \text{ se } r_{v,k} > 0, \\ x_v^k, \text{ caso contrário.} \end{cases} \tag{4}$$

P4: Compute $\eta_k = \min\limits_{0\leq i\leq k} \psi(x_i)$, $Pred_k = \sum_{v=1}^{2} Pred_{v,k}$ e

$$\rho_k = \frac{\eta_k - \psi(x_{k+1})}{Pred_k}.$$

P5: Se $\rho_k \geq \beta_1$: $t_{v,k+1} = \begin{cases} \max[t_{v,k} - \delta_v, 0], & \text{se } r_{v,k} \geq \beta_2 \\ t_{v,k}, & \text{se } r_{v,k} \in [0, \beta_2); \\ t_{v,k} + \delta_v, & \text{se } r_{v,k} < 0, \end{cases}$
Caso Contrário: $t_{v,k+1} = t_{v,k} + \delta_v$, para todo os dois jogadores.
Atualize o raio da região de confiança tomando

$$\Delta_{v,k+1} = \frac{1}{\tau_v + t_{v,k+1}}||g_{v,k+1}||_2. \tag{5}$$

P6: Defina $k := k + 1$ e vá para o Passo 1.

Os principais parâmetros para o desenvolvimento do algoritmo, que atribuem identidade a este, são os parâmetros $r_{v,k}$ e ρ_k. Esses parâmetros compõem os critérios de decisão para aceitar um novo ponto e atualizar o tamanho da região de confiança. O valor de $r_{v,k}$ mensura a proporção entre a redução real da função objetivo e a redução do modelo quadrático de aproximação no novo ponto de iteração, de modo que quando estas reduções são idênticas então $r_{v,k} = 1$. Quanto ao parâmetro ρ_k, sabendo que a função de mérito ψ dimensiona a distância das iterações até o equilíbrio, este parâmetro utiliza dessa informação para comparar se o novo ponto de iteração está mais próximo de um ponto de equilíbrio em relação a todos os outros pontos de iterações já calculados, de modo que a menor distância é medida por η_k

Uma situação averiguada com cuidado são os casos em que $r_{v,k}$ e ρ_k podem não estar bem definidos. Observa-se que quando $Pred_{v,k} = 0$ então $g_{v,k} = 0$, indicando que o ponto corrente é estacionário, tanto para o problema auxiliar quanto para o problema original, e nestes casos definimos que $r_{v,k} = 1$. Além disto, $Pred_k = 0$ implicaria que $\psi(x^k) = 0$ e então, pelo Passo 1, o algoritmo terminaria, não calculando ρ_k. Em [10] as propriedades teóricas do algoritmo são estudadas em detalhes e o resultado principal é o teorema de convergência. Aqui, este resultado é apresentado de forma adaptada para as limitações consideradas, enunciado a seguir.

Teorema 2.1. *Considere que $u_v(x)$, $v \in \{1, 2\}$, sejam duas vezes continuamente diferenciável e denote por $J(x)$ a matriz Jacobiana de $F(x)$. Se $J(x)$ é uniformemente definida positiva e $B_{v,k}$ uniformemente limitada para todo v, k, e $x \in \mathbb{R}^n$, então os pontos de iteração $\{x_k\}$ gerados pelo algoritmo são tais que*

$$\lim_{k\to\infty} \eta_k = 0. \tag{6}$$

3 PROPOSTAS DE MODIFICAÇÕES NO ALGORITMO DE YUAN

Motivados por casos em que o algoritmo de Yuan não apresentou convergência, sendo um destes um problema apresentado em [1], bem atual no contexto a pandemia de COVID-19, sugerimos algumas variações no algoritmo objetivando ampliar as possibilidades de resoluções. No problema em questão, tem-se uma situação simplificada onde um governo possui dois tipos de vacinas para um vírus com duas variantes, e pretende otimizar a eficácia conjunta dessas vacinas. Matematicamente, os objetivos do vírus e governo são representadas por $u_1(x_1, x_2) = x_1(0,45x_2 - 0,3)$ e $u_2(x_1, x_2) = -x_2(0,45x_1 - 0,2)$. Embora muito simples, este exemplo não cumpre as condições de convergência do Teorema 2.1, pois $J(x)$ não é definida positiva. Ao analisar o comportamento numérico do algoritmo de Yuan neste problema, obtemos a sequência divergente ilustrada na Fig. 1. Examinando

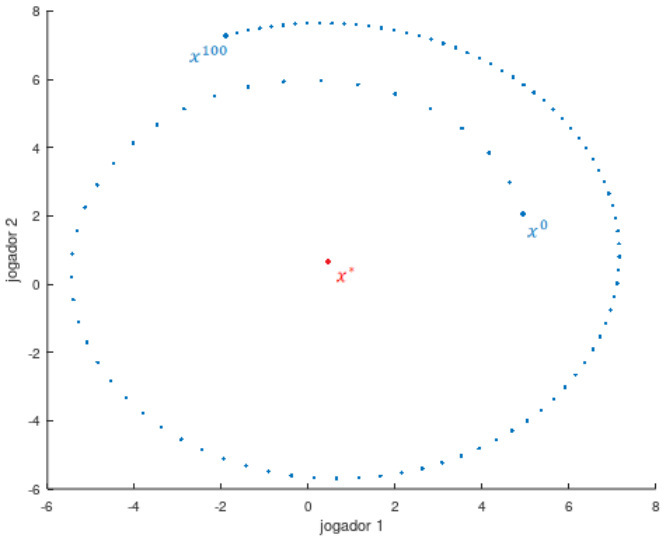


Fig. 1: Iterações do algoritmo de Yuan para exemplo da vacina.

com mais cuidado este exemplo, vemos que o sistema de primeira ordem é descrito pelo conjunto de equações

$$F(x) = \begin{pmatrix} 0,45x_2 - 0,3 \\ -0,45x_1 + 0,2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Observe que o gradiente da função objetivo do jogador v só depende da variável de seu adversário. Isso faz com que quando obtemos, por exemplo, $x_1^k = x_1^*$ temos que $g_{2,k} = 0$, implicando que a região de confiança do segundo jogador seja nula, impedindo mudanças

em x_2. Porém o valor de x_1, que está no valor de equilíbrio, volta a ser atualizado. Esta observação em particular nos motivou a reavaliar a estratégia de região de confiança usada no algoritmo de Yuan, visando componentes como a construção da região de confiança no subproblema e a definição do critério que decide se a solução do problema auxiliar deve ser aceita ou rejeitada.

3.1 Primeira modificação

Como primeira modificação para o algoritmo propomos não usar o raio da região de confiança como sendo proporcional ao gradiente da função no ponto. Neste caso podemos usar algo como a maneira mais tradicional de otimização para o cálculo de regiões de confiança, que tem o tamanho atualizado conforme constantes pré estabelecidas de acordo com critérios dispostos no algoritmo. Assim, adotamos $\Delta_{v,0} = 1$ e o final do Passo 5 é substituído por:

P5: Atualize o raio da região de confiança tomando

$$\Delta_{v,k+1} = \begin{cases} 0,5\Delta_{v,k}, & \text{se } \rho_k \leq \beta_1 \text{ ou } r_{v,k} \leq 0. \\ \Delta_{v,k}, & \text{se } \rho_k > \beta_1 \text{ e } r_{v,k} < \beta_2. \\ 2\Delta_{v,k}, & \text{caso contrário.} \end{cases} \tag{7}$$

Isso significa que quando ρ_k e $r_{v,k}$ são suficientemente grandes classificamos a iteração como bem sucedida e a região de confiança tende a aumentar.

3.2 Segunda modificação

O exemplo da vacina também destaca que é possível que o algoritmo aceite a nova estratégia para o jogador v, devido à melhora de u_v, mesmo que isso acarrete piora da função objetivo do outro jogador. No artigo original [10], Yuan já ressaltava isso, inclusive dizendo que a função objetivo de todos os jogadores pode piorar no novo iterando. Este é um dos argumentos de Yuan para incluir o decréscimo de ψ como critério para atualização da região de confiança, embora não influencie no aceite do ponto. Entretanto, nós acreditamos que rejeitar os pontos que não ocasionem melhora conjunta para os jogadores pode beneficiar o desempenho do algoritmo. Sendo assim, nossa segunda proposta de modificação nesta linha é testar a estratégia de declarar sucesso para o aceite do ponto apenas se os dois objetivos são melhorados. Ou seja,

P3: Defina o próximo ponto da iteração x^{k+1} como

$$x_v^{k+1} = \begin{cases} x_v^k + d_{v,k}, & \text{se } r_{1,k} \text{ e } r_{2,k} > 0, \\ x_v^k, & \text{caso contrário.} \end{cases} \tag{8}$$

Entretanto, este requisito pode ser demasiadamente forte, não permitindo que o algoritmo se mova. Por exemplo, caso x_1^k seja minimizador de $u_1(x_1, x_2^k)$, mas x_2^k não seja solução do problema do segundo jogador, o algoritmo não aceitaria o novo ponto, pois $r_{1,k} \leq 0$, o que impossibilitaria regular x_2^k.

3.3 Terceira modificação

As limitações na modificação anterior nos levou a propor uma nova forma para considerar a influência conjunta na mudança das variáveis, de modo que passamos a atualizar as funções do parâmetro de aceite já com a atualização das variáveis do jogador adversário. Isso significa modificar o cálculo do parâmetro $r_{v,k}$ de modo que a redução real $Ared_{v,k}$ seja calculada como

$$Ared_{v,k} = u_v(x_v^k, x_{-v}^k + d_{-v}^k) - u_v(x_v^k + d_v^k, x_{-v}^k + d_{-v}^k).$$

Dessa forma não estamos considerando a redução no momento atual, mas sim incluindo a expectativa de decisão futura do adversário.

A Tabela 1 explicita a maneira como os experimentos foram realizados, indicando as características mantidas originais e as características que foram modificadas, de modo que, no Teste 1 foi considerado o algoritmo de Yuan original; no Teste 2, variou-se o cálculo da região de confiança; no Teste 3, alterou-se o critério para aceitar o novo ponto de iteração; no Teste 4, foi modificando o cálculo de $Ared_{v,k}$; e por fim, no Teste 5, foram combinadas as modificações propostas em 3.2 e 3.3. Outras possibilidades de variações, que comparam testes numéricos entre o algoritmo de Yuan com métodos do tipo Newton e métodos do tipo Jacobi podem ser conferidos em [2].

Tabela 1: Tipos de testes

	Forma de $\Delta_{v,k}$	aceite x^{k+1}	cálculo $Ared_{v,k}$
Teste 1	original	original	original
Teste 2	modificado	original	original
Teste 3	original	modificado	original
Teste 4	original	original	modificado
Teste 5	original	modificado	modificado

4 RESULTADOS E DISCUSSÕES

A Tabela 2 contém as dinâmicas utilizadas para implementação do algoritmo, que se diferenciam pela composição de convexidade das funções objetivo dos jogadores, nesta tabela, também se encontram os pontos de equilíbrio de Nash associado a dinâmica e o tipo de composição entre as funções. Esses exemplos também foram estudados em [2].

Para os testes numéricos utilizando os Exemplos de 1 a 3, foi atribuído um ponto inicial $x^0 = (5; 1)$, e para o Exemplo 4 um ponto inicial $x^0 = (0,3; 0,7)$. Em todos os casos, manteve-se as constantes $\beta_1 = \frac{1}{2}$, que nivela a tolerância de ρ_k, e $\beta_2 = \frac{1}{2}$ que compõe o critério para atualização de $t_{v,k}$. Nos testes envolvendo a região de confiança original foram usados $\delta_v = 0,1; \tau_v = 1; t_{v,1} = 1$, e os valores atribuídos as constante foram selecionados proporcionalmente a escala do ponto inicial. Também foi estabelecido como critério de parada um limite de 100 iterações, ou até que a soma em módulo dos gradientes assumissem um valor de tolerância próximo de zero, sinalizando tendência a um ponto estacionário. Os resultados obtidos, indicando o ponto final encontrado e o número de iterações realizadas, para os cinco testes nos quatro exemplos são disponibilizados nas Tabelas 3 a 6.

Tabela 2: Funções objetivo e ponto de Equilíbrio de Nash (x^*) associado.

	$u_1(x)$	$u_2(x)$	Convexidade	x^*
Exemplo 1	$\frac{x_1^2}{4} + x_1x_2 - 5x_1$	$\frac{x_2^2}{6} - x_1x_2 - x_2$	ambas convexas	(0,5714 ; 4,7142)
Exemplo 2	$x_1^2 + x_1x_2 - 5x_1$	$-\frac{3x_2^2}{2} - x_1x_2 - x_2$	convexa e não convexa	Não existe
Exemplo 3	$\frac{x_1^3x_2^2}{3} + \frac{x_1^2}{2}$	$\frac{x_1^2x_2^3}{3} + \frac{x_2^2}{2}$	convexidade múltipla	(0 ; 0)
Exemplo 4 (vacina)	$x_1(0,45x_2 - 0,3)$	$-x_2(0,45x_1 - 0,2)$	funções lineares	(0,4444 ; 0,6666)

Tabela 3: Exemplo 1 - $x^* = (0{,}5714\ ;\ 4{,}7142)$

	Iterações	x_{final}
Teste 1	95	(0,5712 ; 4,7114)
Teste 2	Não convergiu	(-2,2500 ; 4,6250)
Teste 3	95	(0,5712 ; 4,7114)
Teste 4	86	(0,5701 ; 4,7122)
Teste 5	Não convergiu	(-1,2119 ; 5,8538)

Tabela 4: Exemplo 2 - Não existe x^*

	Iterações	x_{final}
Teste 1	Não convergiu	(-9,0115e+03 ; 1,9171e+04)
Teste 2	Não convergiu	(3,0000 ; 3,0000)
Teste 3	Não convergiu	(-9,0115e+03 ; 1,9171e+04)
Teste 4	Não convergiu	(-9,0115e+03 ; 1,9171e+04)
Teste 5	Não convergiu	(-9,0115e+03 ; 1,9171e+04)

Tabela 5: Exemplo 3 - $x^* = (0\ ;\ 0)$

	Iterações	x_{final}
Teste 1	8	(8,6462e-04 ; 2,7857e-04)
Teste 2	4	(1,7080e-04 ; 1,7036e-04)
Teste 3	8	(8,6462e-04 ; 2,7857e-04)
Teste 4	8	(1,7797e-03 ; 1,0200e-03)
Teste 5	8	(1,7797e-03 ; 1,0200e-03)

Tabela 6: Exemplo 4 (Vacina) - $x^* = (0{,}4444\ ;\ 0{,}6666)$

	Iterações	x_{final}
Teste 1	Não convergiu	(-1,8865 ; 7,2896)
Teste 2	Não convergiu	(2,9802e-08 ; 5,5000e+00)
Teste 3	Não convergiu	(-1,8865 ; 7,2896)
Teste 4	Não convergiu	(-1,8181 ; 7,0709)
Teste 5	Não convergiu	(-5,3164 ; 0,7031)

Na Tabela 7 consolidamos os resultados, apresentando conjuntamente o número de iterações realizadas em cada um dos testes para os exemplos estudados, e indicamos com a letra N, os casos onde não há convergência.

Tabela 7: Desempenhos dos algoritmos com as modificações propostas

	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5
Exemplo 1	95	N	95	86	N
Exemplo 2	N	N	N	N	N
Exemplo 3	8	4	8	8	8
Exemplo 4	N	N	N	N	N

Das Tabelas 3 a 6 percebemos que os pontos finais encontrados são basicamente os mesmos em todas as versões em que o algoritmo converge, de modo que os desempenhos das versões propostas são comparáveis. Dos dados apresentados na Tabela 7, podemos ver que não houve convergência das sequências geradas pelas variações propostas em nenhum problema não resolvido pelo algoritmo original. Em particular, todas as modificações atingiram o máximo de iterações permitidas quando aplicadas no Exemplo 2. Uma vez que este exemplo não possui Equilíbrio de Nash, este é um fato positivo, indicando que as alterações não tornaram o ponto estacionário $(3, 2; -1, 4)$ um atrator.

Por outro lado, algumas das modificações acabaram por acarretar falha onde o algoritmo original era bem sucedido. Isso aconteceu no Exemplo 1, tanto para o Teste 2, que modifica a região de confiança, quanto para o Teste 5, que combina a modificação no cálculo de $r_{v,k}$ com o critério de aceite de x^{k+1}. Ainda no Teste 2, este apresentou resultados variados, apresentando piora no desempenho para resolver o Exemplo 1, mas melhora para resolver o Exemplo 3.

Passamos então a comparar as versões que resolveram dois dos quatro exemplos. O desempenho da implementação usada no Teste 3 foi exatamente igual à do algoritmo original, sendo que em todos os exemplos os pontos finais obtidos são idênticos. Já a alteração no cálculo de $Ared_{v,k}$ isoladamente, usada no Teste 4, pareceu ser a mais favorável para uma performance melhor do método. Houve uma melhora de cerca de 10% no número de iterações em relação ao Exemplo 1, e não houve alterações para o Exemplo 3.

Embora exemplos específicos não sejam suficientes para que sejam atribuídas conclusões generalizadas, buscou-se através dos resultados desenvolver a percepção de como algumas sugestões de variação poderiam influenciar no comportamento do algoritmo. Ao final, não conseguimos perceber impactos relevantes, originados pelas modificações, de modo a consolidar o desempenho prático do algoritmo de Yuan. Entretanto, a nova proposta de atualização do cálculo de $Ared_{v,k}$ parece ser a que merece maior atenção em futuras investigações neste sentido.

5 CONCLUSÕES

Neste trabalho discutimos algumas características do Algoritmo de Yuan para PENs e constatamos algumas características que impedem a convergência do mesmo inclusive para problemas simples. Em vista disso, propomos algumas modificações na estratégia de região de confiança do algoritmo para tentar contornar estes problemas. Infelizmente, nenhuma das alternativas apresentadas foi eficaz a ponto de ocasionar a convergência para

problemas em que o algoritmo original não convergia. Entretanto, a modificação usada no Teste 4 parece indicar que pode haver algum ganho de desempenho em relação ao algoritmo original.

Esta estratégia consiste em analisar o progresso obtido não olhando para a função $u_v(x_v, x^k_{-v})$, mas sim para $u_v(x_v, x^k_{-v} + d^k_{-v})$. De certa forma isso pode ser interpretado como um expectativa do jogador v quanto a escolha futura do oponente, não se baseando no instante atual. Esta observação é muito interessante e pode motivar novas interpretações acerca de algoritmos para PENs. Além disto, as discussões aqui feitas levantam a necessidade de pesquisa futura para investigação de novas propostas para o uso de regiões de confiança em PENs.

6 *Agradecimentos*

Os autores agradecem à CAPES, ao CNPq e à FAPESP (processos 2013/07375-0 e 2018/24293-0) pelo apoio financeiro para realização deste trabalho.

Referências

[1] H. Bortolossi, G. Garbagio, and B. Sartini. *Uma introdução à Teoria Econômica dos Jogos*. 26º Colóquio Brasileiro de Matemática, IMPA, 2007.

[2] L. Bueno and A. Vetorazzi. Experimentos numéricos sobre o método de Yuan para problemas de Equilíbrio de Nash. *INTERMATHS*, 2(2):59–74, 2021.

[3] F. Caruso, M. Ceparano, and J. Morgan. An inverse-adjusted best response algorithm for Nash equilibria. *SIAM Journal on Optimization*, 30:1638–1663, 01 2020.

[4] F. Facchinei and C. Kanzow. Generalized Nash equilibrium problems. *Annals of Operations Research*, 175(1):177–211, 2010.

[5] A. Fischer, M. Herrich, and K. Schonefeld. Generalized Nash Equilibrium Problems - Recent Advances and Challenges. *Pesquisa Operacional*, 34(3):521–558, 2014.

[6] J. M. Martínez and S. A. Santos. *Métodos Computacionais de Otimização*. Departamento de Matemática Aplicada, IMECC-UNICAMP, 1998.

[7] J. Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

[8] Y. xiang Yuan. A review of trust region algorithms for optimization. *ICM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics*, 1999.

[9] Y. xiang Yuan. Recent advances in trust region algorithms. *Mathematical Programming*, 151:249–281, 2015.

[10] Y.-x. Yuan. A trust region algorithm for Nash equilibrium problems. *Pacific Journal of Optimization*, 7(1):125–138, 2011.

Estudo de Caso da Regressão por Processos Gaussianos para Previsão de COVID-19

André Igor Nóbrega da Silva[1], Elloá B. Guedes[2] e Edmar Candeia Gurjão[1]

[1] *Universidade Federal de Campina Grande, Campina Grande, PB*
[2] *Universidade do Estado do Amazonas, Manaus, AM*

Resumo

Com vistas a colaborar na vigilância epidemiológica da pandemia do COVID-19, este trabalho baseia-se na utilização de regressão por Processos Gaussianos, uma abordagem bayesiana não-linear e não-paramétrica de Aprendizagem de Máquina para séries temporais, com vistas a prever o acumulado de casos de COVID-19 na cidade de Campina Grande, Paraíba. Para tanto, considerou-se uma abordagem auto-regressiva e também a decomposição aditiva da série temporal para efetuar hipóteses acerca das melhores estruturas de *kernels* para modelar o problema. A partir da análise das métricas de desempenho aferidas, foi possível validar a solução, com especial destaque para as hipóteses iniciais elencadas e boa qualidade de previsão perante poucos dados de treinamento, o que é essencial para cenários práticos de surtos de doenças.

Palavras-chave: COVID-19, Regressão por Processos Gaussianos, Aprendizagem de Máquina, Séries Temporais, Modelagem Preditiva

1 INTRODUÇÃO

Desde o início do ano de 2020, a humanidade vêm passando por uma pandemia do COVID-19 que, no início de Novembro de 2021 já vitimou 5.054.257 pessoas ao redor do mundo [11]. Essa doença é causada pelo agente SARS-CoV-2, um vírus altamente infeccioso e de transmissão de pessoa para pessoa, cujo contágio pode ser reduzido com medidas de distanciamento social, uso de máscara e protocolos de higienização [16]. Devido à suas proporções continentais e cenário de desigualdade social, o Brasil foi especialmente afetado pela doença, chegando a representar o epicentro global da pandemia em certos momentos [3]. Até o início de novembro de 2021, 21.880.439 pessoas haviam sido infectadas no País e 609.447 destas vieram a óbito [11].

Contato: André Igor Nóbrega da Silva, andre.nobrega@ee.ufcg.edu.br

Um monitoramento epidemiológico efetivo que inicia no nível de comunidades locais e então abrange regiões maiores é a chave para encontrar soluções que combatam o contágio de doenças. No entanto, o sistema de vigilância do COVID-19 precisa ser melhorado usando Inteligência Artificial e Tecnologia da Informação [16]. Nesse cenário, modelos inteligentes podem ser avaliados para apoiar decisões estratégicas em Saúde Coletiva, auxiliando autoridades e gestores a (*i*) criar, adotar, revisar e sustentar políticas de distanciamento social (reabertura de escolas, fechamento temporário de lojas e redução de jornada, por exemplo); (*ii*) racionalizar os testes rápidos e RT-PCR, especialmente quando houver potencial de surto; (*iii*) melhorar a logística dos hospitais, antecipando a demanda por leitos, medicamentos, equipamentos, etc.; (*iv*) planejar, implementar e reforçar campanhas de vacinação, especialmente com foco na população vulnerável, dentre outras. Em geral, promover estratégias de prevenção é economicamente mais barato do que implementar intervenções terapêuticas [8].

Diante do exposto, o presente trabalho se propôs a investigar o uso da Regressão por Processos Gaussianos (GPR, do inglês *Gaussian Process Regression*) na previsão do número de casos acumulados de COVID-19 na cidade de Campina Grande, Paraíba. O GPR é uma abordagem Bayesiana não-paramétrica para problemas de regressão que pode ser utilizada em cenários de exploração e extrapolação, com capacidade de capturar uma ampla variedade de relações entre entradas e saídas, utilizando um número teoricamente infinito de parâmetros e permitindo que os dados determinem o nível de complexidade [15, 14]. Campina Grande, por sua vez, é um município brasileiro localizado no Estado da Paraíba com 411.807 habitantes, a qual foi recentemente eleita Cidade Criativa pela UNESCO [13]. O cenário deste estudo de caso foi escolhido levando-se em consideração as demandas práticas para a previsão de casos de COVID-19 nessa localidade para ajudar a conscientizar a população sobre a pandemia, incentivando-a a respeitar as políticas de distanciamento social e saneamento.

Para apresentar o que se propõe, este trabalho está organizado como segue. A fundamentação teórica, apresentada na Seção 2, considera os conceitos elementares a respeito de GPR e a decomposição de séries temporais. Em seguida, na Seção 3, tem-se a metodologia do trabalho, na qual são descritos os materiais e métodos utilizados. Os resultados obtidos encontram-se na Seção 4. Por fim, as considerações finais são apresentadas na Seção 5, juntamente com perspectivas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão apresentados os principais conceitos envolvidos no treinamento e na realização de previsões utilizando GPR.

2.1 Regressão por Processos Gaussianos

Uma tarefa de regressão se caracteriza como o problema de encontrar uma função que mapeia entradas e saídas a partir de um conjunto de dados de treinamento, ou seja, dado um conjunto $D = \{(x_i, y_i)|i = 1, 2, ...n\}$, em que x é a variável de entrada e y é a saída (alvo da predição), deseja-se encontrar uma função f tal que $f(x) \approx y$. A partir desse processo, utiliza-se a função f para realizar predições em dados de entrada não antes vistos x_*, obtendo y_* [14]. Existem diversas técnicas e algoritmos para realização de tal tarefa,

dentre as quais podemos destacar a regressão por Processos Gaussianos.

De acordo com Rasmussen & Williams, um processo gaussiano é uma coleção de variáveis aleatórias, em que qualquer número finito das quais possui uma distribuição conjunta de Gauss [14]. Pode-se afirmar que um processo gaussiano é completamente definido pela sua função de média $m(x)$ e sua função de covariância $K(x, x')$, esta última também chamada de *kernel*. Tanto x quanto x' são pontos quaisquer no domínio da função, com seus respectivos valores observados y e y'. Podemos especificar tais funções de acordo com as Eqs. (1) e (2), nas quais o o símbolo $\mathbb{E}$ indica o valor esperado das variáveis aleatórias.

$$m(x) = \mathbb{E}[f(x)], \tag{1}$$
$$K(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \tag{2}$$

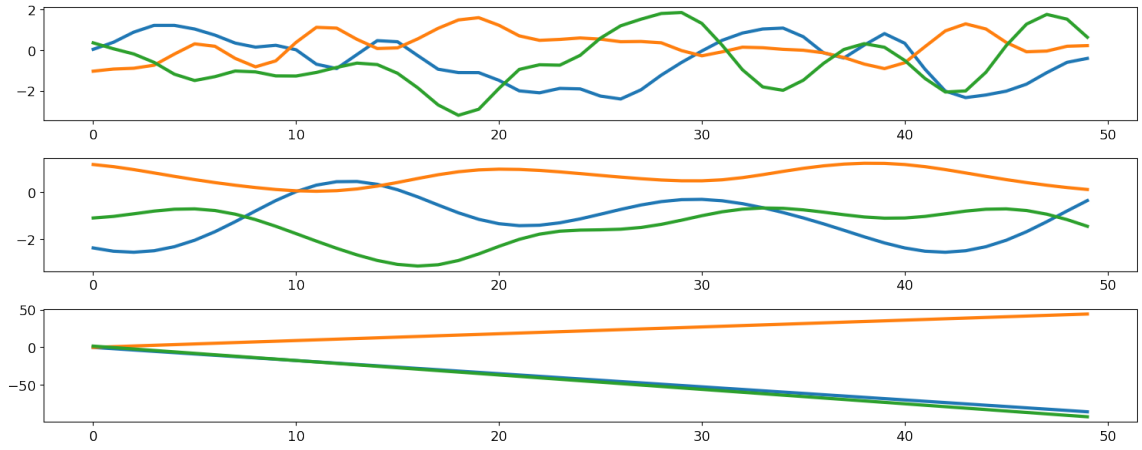
Uma forma comum de representação do processo gaussiano é dada conforme segue:

$$f(x) \sim \mathcal{GP}(m(x), K(x, x')). \tag{3}$$

Por razões práticas e também para evitar a introdução de um viés sobre a tendência do modelo, geralmente se utiliza uma função de média igual a zero [1].

Depois de levar a média em consideração, o tipo de estrutura que pode ser capturado pelo método GPR é determinado inteiramente pelo *kernel* adotado. A função *kernel* determina como o modelo generaliza, i.e., extrapola para novos dados. Alguns exemplos de funções *kernel* podem ser observados na Fig. 1. No primeiro gráfico, observa-se um *kernel* com características de suavidade. O segundo apresenta um *kernel* com características periódicas e o último, por sua vez, mostra um *kernel* com características lineares. Um maior detalhamento acerca dessas funções de covariância e suas características é dado na seção seguinte.

Fig. 1: Representação de exemplos de funções Kernel. Adaptação de [17]



A maior dificuldade de se realizar uma tarefa de aprendizagem por meio de um GPR é a escolha de uma função de covariância (*kernel*) adequada para o problema. Dentre as principais funções *kernel* utilizadas no GPR, pode-se destacar o *kernel* função exponencial

quadrática (SE, do inglês *squared exponential*), o *kernel* periódico e o *kernel* linear [6].

- ***Kernel* SE**: O *kernel* SE é um *kernel* do tipo estacionário. Isso significa que o valor da sua função K_{se} depende apenas da diferença $x - x'$. Na Eq. (4), o parâmetro ℓ, chamado *lengthscale*, determina o tamanho das ondulações observadas nas amostras. A variância σ^2, por sua vez, determina a distância de um valor da função da sua média.

$$K_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \tag{4}$$

- ***Kernel* Periódico**: O *kernel* periódico permite modelar funções que se repetem ao longo do tempo. A sua utilização é recomendada em problemas de séries temporais que apresentam sazonalidade. Tal função *kernel* é dada pela Eq. (5).

$$K_{\mathrm{Per}}(x, x') = \sigma^2 exp\left(-\frac{2sin^2(\pi|x - x'|/p)}{\ell^2}\right). \tag{5}$$

  Nessa estrutura, o parâmetro p determina a distância entre os ciclos da função e o parâmetro ℓ possui o mesmo significado anteriormente mencionado.

- ***Kernel* Linear**: O *kernel* linear, ao contrário dos anteriores é um *kernel* não estacionário. Isso significa que o modelo GPR produzirá diferentes predições caso haja um deslocamento linear nos dados [6].

  Ao utilizar apenas o *kernel* linear para um determinado problema, implementa-se uma regressão Bayesiana simples. O valor da função desse *kernel* é dado pela Eq. (6)

$$K_{\mathrm{Lin}}(x, x') = \sigma(x - c)(x' - c), \tag{6}$$

  em que o parâmetro c é chamado de *offset* e determina um ponto em que todas as funções amostras da distribuição *a posteriori* passam.

Além da utilização de *kernels* individuais para tarefas de regressão, também é possível combiná-los por meio de operações de adição e multiplicação, obtendo estruturas cada vez mais complexas [6]. Também é possível modelar possíveis ruídos utilizando uma função desconhecida e de rápida variação. Essa estrutura pode ser incorporada adicionando um *kernel* SE local com um *lengthscale* muito pequeno. Quando esse parâmetro tende a zero, esse sinal tende a um ruído branco [6].

2.2 Decomposição Aditiva de Séries Temporais

Uma série temporal é uma sequência de observações tomadas sequencialmente no tempo [4]. Diversos fenômenos constituem séries temporais, tais como quantidade de vendas em uma loja por mês, a evolução do preço de ações no mercado financeiro, a temperatura diária de uma determinada cidade, a quantidade de pacientes infectados por uma doença, dentre outros.

Uma das principais técnicas para análise da série temporal é a decomposição aditiva da mesma. Uma maneira comum de decomposição é em uma parcela sistemática e em uma parcela não-sistemática [7]. A parcela sistemática é subdividida em nível, tendência e sazonalidade. A parcela não-sistemática, por sua vez, é chamada de ruído. O nível descreve o valor médio da série, a tendência é a variação da série entre períodos e a sazonalidade descreve um comportamento cíclico de curta duração que pode ser observado diversas vezes ao longo da série. Por fim, o ruído é a variação aleatória advinda de erros de medição ou outras razões diversas. A Eq. (7) mostra como uma série temporal y com $t \in 0, \ldots, n$ observações discretas pode ser decomposta a partir desse método.

$$y[t] = \text{nível}[t] + \text{tendência}[t] + \text{sazonalidade}[t] + \text{ruído}[t]. \tag{7}$$

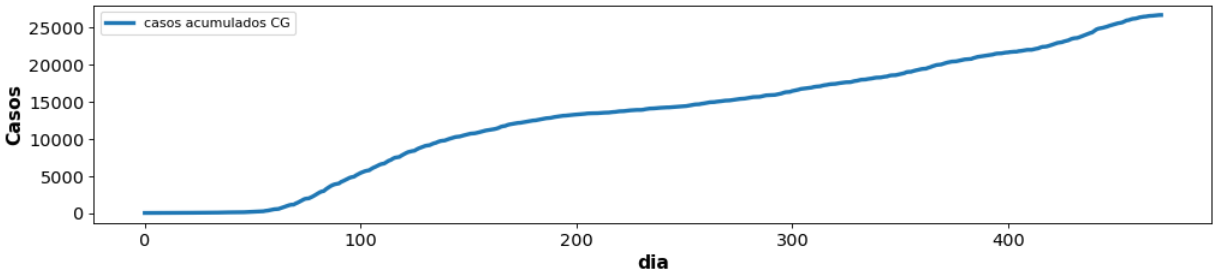
3 METODOLOGIA

Nesta seção apresenta-se o conjunto de dados obtidos para a modelagem do problema, bem como a sequência de etapas seguidas para o desenvolvimento do modelo de regressão por Processos Gaussianos.

3.1 Dados Experimentais

Os dados utilizados para o problema foram fornecidos pela Secretaria de Saúde da cidade de Campina Grande, correspondendo ao acumulado total de casos de COVID-19 no período de 02 de fevereiro de 2020 à 01 julho de 2021[1]. Ao todo, 472 dias de evolução da doença foram monitorados e a evolução pode ser observada na Fig. 2.

Fig. 2: Evolução do acumulado de casos de COVID-19 em Campina Grande, PB.

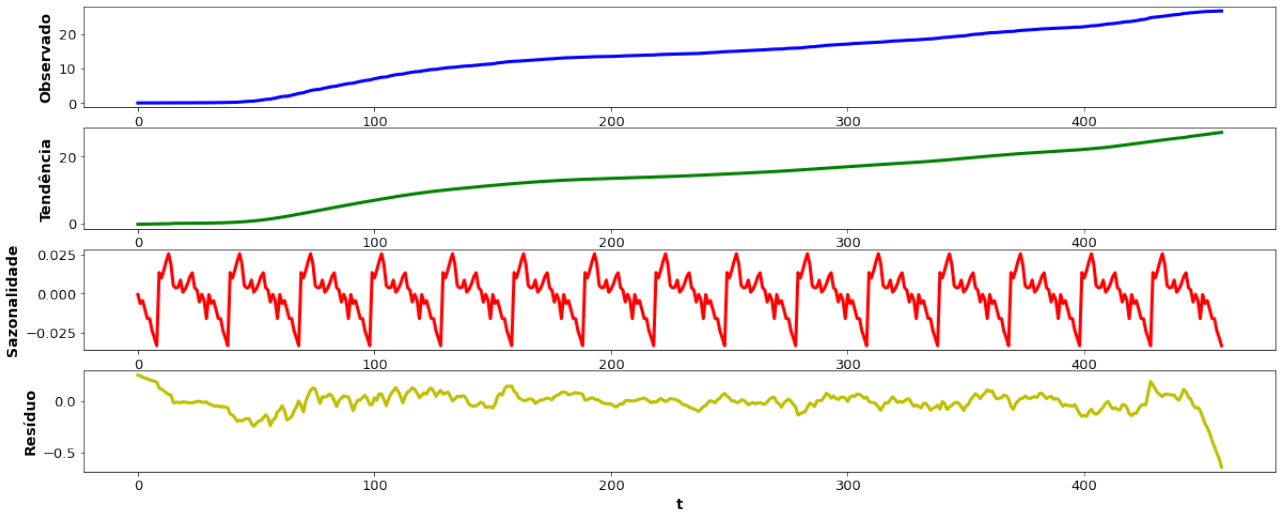


Qualitativamente, podemos observar que trata-se de uma série temporal não-estacionária com uma forte componente de tendência. De fato, pela própria característica do cumulativo de casos, tem-se uma função monotonicamente crescente. Na Figura 3, é possível observar a decomposição dessa série. Percebe-se que a componente de sazonalidade parece ser fraca em relação à componente de tendência, visto que a primeira possui uma ordem de grandeza muito inferior quando comparada à segunda.

Uma vez obtido o conjunto de dados que representava o acumulado de casos ao longo do tempo, definiu-se então o conjunto de variáveis preditoras e a variável alvo. A variável alvo foi definida como o acumulado de casos no dia $d + 7$, i.e., almeja-se a previsão da evolução da curva com horizonte de 7 dias.

[1]Dados podem ser obtidos em https://github.com/andreigor/TCC-GPR/blob/master/Data

Fig. 3: Decomposição aditiva da série temporal do acumulado de casos



Para o conjunto de variáveis preditoras, utilizou-se a prática comum de obter características a partir de valores passados da própria série temporal [7]. Dessa forma, as variáveis usadas foram a média móvel do acumulado de casos no conjunto de dias $\{d-7, d-6, ..., d\}$ e a própria quantidade de casos no dia d.

3.2 Proposição de Estruturas de Kernels

A partir da análise da decomposição da série temporal, duas estruturas de *kernels* foram propostas para o desenvolvimento do modelo GPR: a estrutura para regressão polinomial e a estrutura de regressão semi-paramétrica, conforme sugerido por Duvenaud [6]. Tal escolha justifica-se em virtude de ambas possuírem componentes lineares capazes de capturar adequadamente séries temporais com tendências. A implementação do modelo foi feita utilizando a biblioteca *scikit-learn* do Python. Nesta biblioteca, o *kernel* SE, o *kernel* periódico e o *kernel* linear são nomeados, respectivamente, *RBF*, *DotProduct* e *ExpSineSquared*. As duas estruturas consideradas são detalhadas como segue:

- **Estrutura de Regressão Polinomial**: essa estrutura é dada como segue:

$$K = c + DotProduct^{e} + r. \tag{8}$$

  O valor da constante c, que indica o nível da série temporal, foi variado no conjunto $\{1, 2, 3, 4, 5\}$. O parâmetro e, com $e \in \{1, 2, 3, 4, 5, 6, 7\}$, indica a ordem do polinômio considerado no *kernel* *DotProduct*. O valor r denota o ruído.

- **Estrutura de Regressão Semi-Paramétrica**: essa estrutura é dada pela Eq. (9):

$$K = DotProduct(\sigma) + RBF(\ell) + r, \tag{9}$$

  em que o espaço de busca dos parâmetros σ e ℓ foi definido da seguinte maneira: $\sigma \in \{1, 2, \ldots, 10\}$, e ℓ variou no espaço logarítmico de $[-1, 1]$ coletando 5 amostras

igualmente espaçadas. Essa estrutura possui a característica de modelar tendências com variações suaves ao longo do tempo.

3.3 Avaliação de Desempenho

As métricas de desempenho consideradas para a tarefa em questão foram:

1. $\mathbf{R^2}$. Denominado coeficiente de determinação, denota a proporção da variância da variável dependente que é prevista a partir das variáveis independentes. Possui valor máximo igual a 1 e pode assumir valores negativos para modelos arbitrariamente piores que prever uma constante ou valores aleatórios;

2. ***Mean Absolute Percentage Error* (MAPE)**. Essa métrica reflete a acurácia de modelos de regressão. Considerando $\mathbf{y}$ como o vetor dos valores observados e $\hat{\mathbf{y}}$ como o vetor dos valores previstos por um modelo, ela é calculada conforme segue: $\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$.

3. ***Mean Squared Error* (MSE)**. É um estimador derivado da distância Euclidiana que sempre assume valores positivos com mínimo igual a zero para um regressor perfeito, sendo calculado conforme segue: $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

A avaliação das estruturas de *kernels* foi efetuada conforme a validação cruzada para séries temporais denominada *Time Series Split* com $n = 10$. Oriunda de uma adaptação da validação cruzada *k-folds*, nela os dados são divididos em n intervalos fixos, denominados *splits*, que respeitam a ordem temporal de observação dos eventos. A cada iteração os índices de teste são maiores que na iteração anterior, refletindo assim a capacidade do modelo em prever o fenômeno desde as observações iniiciais e também conforme mais dados tornam-se disponíveis. As métricas de desempenho previamente mencionadas são aferidas no conjunto de teste de cada *split* [10].

4 RESULTADOS E DISCUSSÃO

Para a busca em grade proposta, os seguintes *kernels* e parâmetros obtiveram os melhores resultados para a estrutura de regressão polinomial e regressão semi-paramétrica, respectivamente:

1. Valor de constante $c = 2$ somado ao *kernel* linear com grau $e = 1$ somado ao *kernel* de ruído com variância igual a 1;

2. *Kernel* RBF com $\ell = 1$ somado ao *kernel* linear com grau 1 somado ao *kernel* de ruído com variância igual a 1.

As Fig. 4 a seguir mostra as métricas de avaliação R^2, MAPE e MSE para 10 *splits* de validação cruzada utilizados no processo de treinamento com a série temporal. A Fig. 5, por sua vez, compara os dados reais acumulados com as previsões obtidas pelo modelo GPR gerado pela estrutura semi-paramétrica que apresentou o melhor resultado.

Ambas as estruturas apresentaram resultados muito bons nas métricas de avaliação desde os *splits* iniciais, o que fica evidenciado ao verificar que o maior valor possível

Fig. 4: Comparação das métricas de avaliação ao longo dos splits

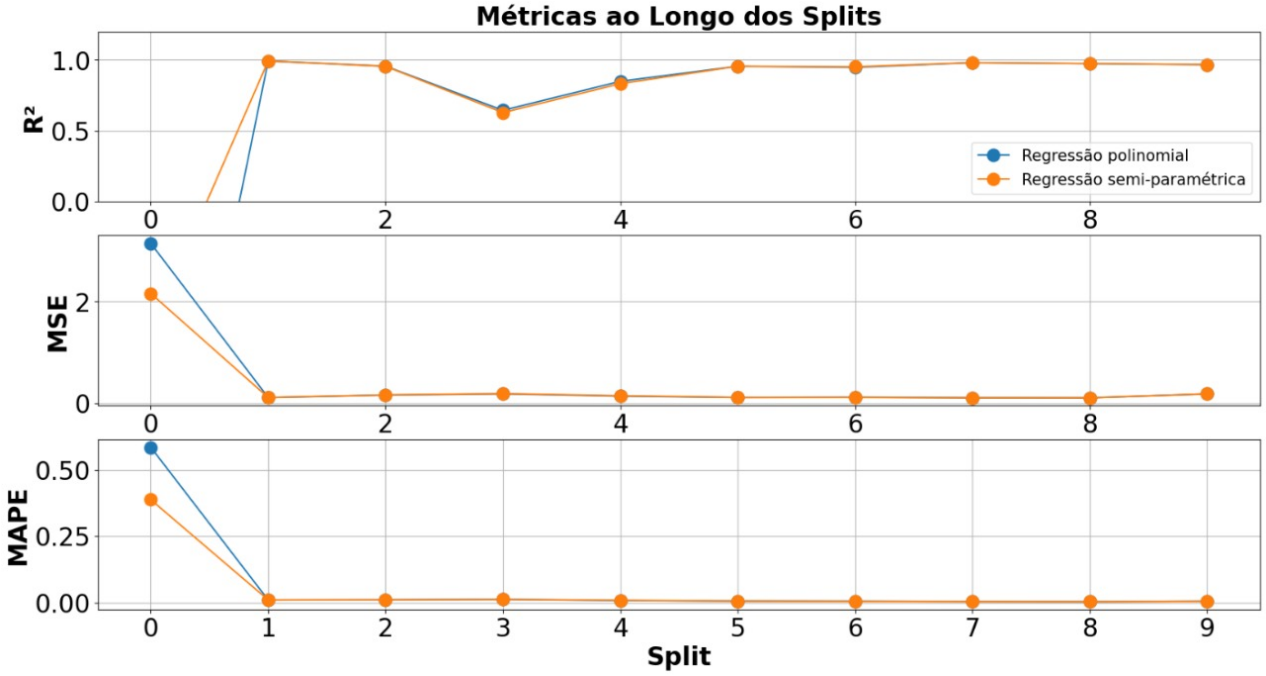
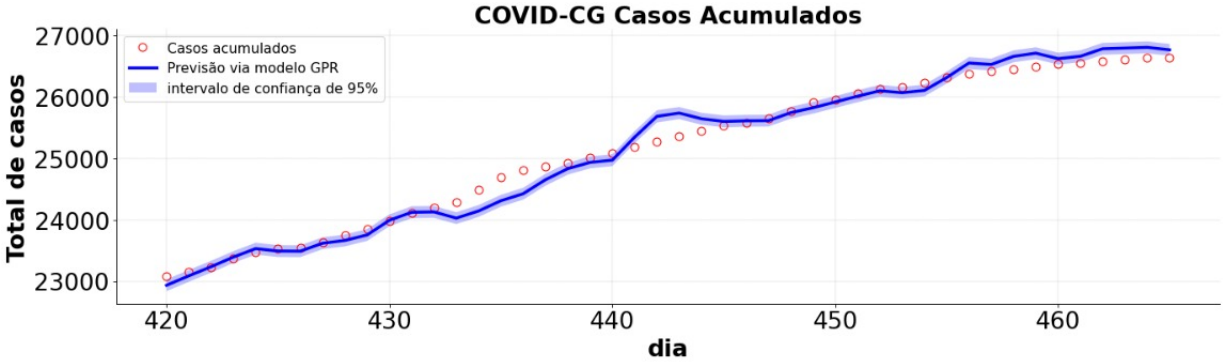


Fig. 5: Predições e dados reais para o último *split* de teste



para o R^2 é igual a 1. A partir dos gráficos, é possível afirmar que, dado o problema do acumulado de casos, as estruturas de *kernel* para regressão polinomial e para a regressão semi-paramétrica demonstram uma necessidade de poucos dados para a extrapolação, uma vez que no *Split* 1 tem-se apenas 20% dos dados disponíveis para o treinamento. Os valores das métricas MAPE e MSE, que indicam a distância entre as saídas previstas e os valores observados também refletem a qualidade das previsões ao longo do tempo, pois mostram-se próximas de 0 em todo o cenário experimental.

Embora muito trabalhos na literatura abordem o problema da previsão de COVID-19 com Aprendizado de Máquina [5, 9, 12], são escassos os trabalhos que consideram a previsão exclusivamente de natureza auto-regressiva, sem variáveis exógenas. No tocante ao uso do GPR para previsão de COVID-19, Ahmad *et al.* consideraram a previsão de casos no Oriente Médio e na Ásia, tendo verificado experimentalmente a adequação do

kernel Matérn para esta tarefa [2]. Velásquez & Lara utilizaram GPR em dados de 82 dias de casos de COVID-19 nos Estados Unidos e concluíram a efetividade deste método na obtenção de uma estimativa acurada do espalhamento de casos em tal localidade [18]. Até o momento, no melhor dos esforços de busca por trabalhos relacionados, não foram encontradas publicações de terceiros que utilizem o GPR de maneira auto-regressiva para previsão de COVID-19 em nível municipal no Brasil.

5 CONSIDERAÇÕES FINAIS

Neste trabalho descreveu-se os princípios de funcionamento da regressão por Processos Gaussianos aplicado à previsão da evolução do acumulado de casos de COVID-19 na cidade de Campina Grande-PB. Para tanto, foram propostas 2 estruturas de *kernels* (estrutura de regressão polinomial e estrutura de regressão semi-paramétrica) e definiu-se um espaço de busca em grade a partir da variação dos parâmetros ajustáveis. O treinamento e avaliação do modelo se deram por meio de uma validação cruzada adaptada para séries temporais. Como resultado, foi possível observar que desde os *splits* iniciais, em que uma menor quantidade de dados de treino é utilizada, o método proposto mostra-se robusto perante as métriicas de avaliação observadas

É importante destacar que a curva de evolução do acumulado de casos depende de diversos fatores sociais, políticos e econômicos. A utilização apenas de variáveis autorregressivas para a previsão de novos casos é uma grande simplificação do problema em questão. Apesar disso, o R^2 de ambas as estruturas propostas ficou muito próximo de 1, indicando uma alta explicabilidade dos modelos construídos para o cenário prático considerando, com um horizonte temporal de 7 dias, indicando uma grande capacidade de previsão da curva de casos acumulados de COVID-19 em Campina Grande, PB. A partir de tais técnicas de Aprendizagem de Máquina aplicadas à séries temporais, é possível propôr políticas públicas mais assertivas para controle do espalhamento da doença, já que existem expectativas acerca dos casos futuros baseadas em dados.

Como trabalhos futuros, sugere-se investigar a proposição de estruturas de *kernels* baseadas na análise da decomposição da série temporal para prever a quantidade diária de novos casos. Trata-se de um problema muito mais desafiador devido à grande variância e à natureza não-monotônica dessa série.

Referências

[1] A. B. Abdessalem, N. Dervilis, D. J. Wagg, and K. Worden. Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo. *Frontiers in Built Environment*, 3:52, 2017.

[2] F. Ahmad, S. N. Almuayqil, M. Humayun, S. Naseem, W. A. Khan, and K. Junaid. Prediction of COVID-19 cases using machine learning for effective public health management. *Computers, Materials & Continua*, 66(3):2265–2282, 2021.

[3] C. Almeida, L. Lüchmann, and C. Martelli. A pandemia e seus impactos no Brasil. *Middle Atlantic Review of Latin American Studies*, 4(1):20–25, 2020.

[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis, Forecasting and Control*. John Wiley & Sons, Inc, 2016.

[5] R. G. da Silva, M. H. D. M. Ribeiro, V. C. Mariani, and L. dos Santos Coelho. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos, Solitons & Fractals*, 139:110027, Oct. 2020.

[6] D. K. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

[7] S. Galit and L. Kenneth. *Practical Time Series Forecasting with R*. Axelrod, 2016.

[8] E. B. Guedes, C. Martins, P. R. L. Júnior, and E. C. Gurjão. A case study on forecasting new daily cases of COVID-19 at different scales in Brazil. pages 1–5, Fortaleza, Ceará. SBrT.

[9] E. Z. Martinez, D. C. Aragon, and A. A. Nunes. Short-term forecasting of daily COVID-19 cases in Brazil by using the holt's model. *Revista da Sociedade Brasileira de Medicina Tropical*, 53, 2020.

[10] A. Nielsen. *Practical Time Series Analysis – Prediction with Statistics & Machine Learning*. O'Reilly, Canada, 1 edition, 2020.

[11] OMS. Coronavirus disease (COVID-19) pandemic. Disponível em `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`, 2020. Acessado em 18 de fevereiro de 2022.

[12] I. G. Pereira, J. M. Guerin, A. G. S. Júnior, G. S. Garcia, P. Piscitelli, A. Miani, C. Distante, and L. M. G. Gonçalves. Forecasting COVID-19 dynamics in Brazil: A data driven approach. *International Journal of Environmental Research and Public Health*, 17(14):5115, July 2020.

[13] Prefeitura Municipal de Campina Grande. Portal da Prefeitura Municipal de Campina Grande, 2021. Disponível em `https://campinagrande.pb.gov.br/`. Acessado em 18 de fevereiro de 2022.

[14] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology Press, Estados Unidos, 1 edition, 2006.

[15] E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1 – 16, 2018.

[16] N. Srivastava, P. Baxi, R. K. Ratho, and S. K. Saxena. Global trends in epidemiology of coronavirus disease 2019 (COVID-19). In *Medical Virology: From Pathogenesis to Disease Control*, pages 9–21. Springer Singapore, 2020.

[17] B. A. Swastanto. Gaussian process regression for long-term time series forecasting. Dissertação de mestrado, Delft University of Technology, Delft, 2016.

[18] R. M. A. Velásquez and J. V. M. Lara. Forecast and evaluation of COVID-19 spreading in USA with reduced-space gaussian process regression. *Chaos, Solitons & Fractals*, 136:109924, July 2020.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Avaliação de Técnicas de Redução de Dimensionalidade de Dados para Problemas de Classificação

Carla Nascimento Neves[1], Mariza Ferro[1], Francisco Bruno Souza Oliveira[2] e Paulo Eduardo Ambrósio[2]

[1] *Laboratório Nacional de Computação Científica, Petrópolis/RJ, Brazil*
[2] *Universidade Estadual de Santa Cruz, Ilhéus/BA, Brazil*

Abstract

O presente trabalho tem como objetivo investigar quais técnicas de redução de dimensionalidade apresentam melhor desempenho para certos tipos de bases de dados comuns na literatura. Foram selecionados cinco conjuntos de dados com diferentes características, e treze das técnicas mais utilizadas de redução de dimensionalidade comumente aplicadas sobre os mesmos. A classificação dos conjuntos foi feita com o algoritmo *Random Forest*. Os resultados alcançados auxiliam, principalmente, na escolha de um método de redução de dimensionalidade para diferentes tipos de dados.

Keywords:Redução de dimensionalidade, Classificação, Aprendizado de máquina.

1 INTRODUÇÃO

Conjuntos de dados consistem em várias observações medidas de um grupo de amostras. Essas observações são chamadas de atributos ou características, que constituem dimensões no espaço de dados [15]. Em conjuntos multidimensionais é comum haver uma quantidade significativa de dados ruidosos ou redundantes, que podem afetar a eficácia de algoritmos de Aprendizado de Máquina (AM). Portanto, em muitas situações, é necessário produzir uma representação dos dados com dimensionalidade reduzida. A redução do número de características empregadas para representar um determinado conjunto de dados é chamada de redução de dimensionalidade. Essa estratégia transfere os dados originais para um espaço de menor dimensão por remoção de atributos ou transformação matemática e tenta preservar os atributos essenciais nesse espaço de menor dimensão [9].

Contato: Carla Neves, cneves@lncc.br

Em muitas tarefas de AM, sendo a classificação o foco deste artigo, a redução de dimensionalidade é aplicada para melhorar a qualidade dos atributos dos dados, facilitar sua interpretação, reduzir a complexidade computacional e melhorar a precisão do reconhecimento de padrões. Para se beneficiar das técnicas de redução de dimensionalidade com o objetivo de maximizar a precisão do algoritmo de classificação, é necessário saber qual técnica se adapta melhor para cada tipo de dados.

Tendo isso em vista, o objetivo deste trabalho é avaliar o desempenho de diferentes técnicas de redução de dimensionalidade e determinar qual técnica melhor se adapta a cada tipo de conjunto de dados. Para avaliação foram selecionados diferentes conjuntos de dados, os quais são comumente utilizados em trabalhos de redução de dimensionalidade para a tarefa de classificação.

Na Seção 2 estão as técnicas utilizadas nesta pesquisa, bem como alguns trabalhos relacionados. Na Seção 3 é apresentada a metodologia seguida pelo trabalho. Os resultados obtidos estão na Seção 4. Por fim, na Seção 5 são feitas as considerações finais.

2 Redução de Dimensionalidade em Problemas de Classificação

As técnicas de redução são utilizadas em tarefas de classificação com o objetivo de melhorar a precisão do reconhecimento de padrões ou minimizar as perdas nessa precisão utilizando um número reduzido de atributos. Os conjuntos de dados utilizados neste artigo, são bastante utilizados na literatura. Na Subseção 3.1 há uma descrição desses conjuntos, onde são também mencionados diversas referências bibliográficas que os utilizaram. Sendo assim, foi possível selecionar quais as técnicas que são mais empregadas nestes conjuntos, a fim de fazer uma comparação do desempenho das mesmas, como proposto neste artigo. A seguir, são feitas breves descrições destas técnicas, acompanhadas de referências que as detalham. Além disso são citados alguns trabalhos relacionados que fazem a comparação entre diferentes técnicas de redução de dimensionalidade.

2.1 Técnicas de redução de dimensionalidade

A Decomposição de Valores Singulares (SVD) está entre as técnicas de fatorações de matrizes mais importantes da era computacional. Para realizar a redução de dimensionalidade, expõe a subestrutura dos dados originais mais claramente e os ordena da maior variação para a menor [22].

A técnica do Filtro Qui-Quadrado (CHI) é utilizada no campo da redução de dimensionalidade para calcular se um atributo é importante ou não para a classificação. É baseada na comparação dos valores obtidos da frequência de uma classe por causa da divisão com a frequência esperada da classe [12]. A Trasformada Discreta de Fourier (DFT) é sugerida para redução da dimensionalidade de séries temporais. A abordagem representa este tipo de dados e realiza truncamentos no domínio da frequência [24].

A Aproximação Simbólica de Fourier (SFA) também é utilizada na redução de dimensionalidade de séries temporais, se constitui em duas fases: aproximação, em que ocorre a aplicação de um filtro passa baixa com uma frequência de corte, para eliminar os ruídos, e discretização, que consiste na aplicação de uma função para mapear um intervalo de valores numéricos para valores discretos (símbolos) [18].

No Mapeamento Isométrico de Atributos (ISOMAP), a medida de distância inter-

pontos é realizada de forma a capturar as não-linearidades do *manifold* (hiper-superfície formada pelos atributos de entrada) imerso no espaço de alta dimensão original. A ideia desta técnica é utilizar, como medida, uma aproximação razoável da distância geodésica entre os pontos de entrada [23].

A Análise de Componentes Principais (PCA) escolhe um novo sistema de coordenadas para o conjunto de dados que é uma representação das direções em que a variância dos dados é mais alta, ou seja, mantendo os componentes principais de ordem superior e ignorando os de ordem inferior [22]. A PCA baseada em *Kernel* (KPCA) é uma técnica construída utilizando uma função *kernel*. Ela calcula os autovetores principais da matriz do *kernel*, em vez dos da matriz de covariância, como ocorre na PCA [9].

Na Análise de Fatores (FA), o objetivo é explicar as intercorrelações entre variáveis, identificando conjunto de dimensões subjacentes, ou fatores. A técnica representa um modelo para os dados, que relaciona as variáveis originais a um conjunto menor de fatores subjacentes não observáveis [5]. A Análise de Componentes Independentes (ICA) descreve um modelo para dados multivariados. As variáveis deste modelo são consideradas não gaussianas e mutuamente independentes. Esta técnica tenta encontrar os componentes ou fontes originais por meio de algumas suposições simples de suas propriedades estatísticas [3].

A Análise Discriminante Linear (LDA) determina um subespaço de dimensão inferior à da amostra de dados original, com boa separabilidade de classes. Essa separabilidade é definida em termos de medidas estatísticas de valor médio e variância [19].

Foram utilizados três algoritmos de seleção de atributos baseados em busca sequencial. Em cada iteração, um classificador é treinado para avaliar a importância dos atributos do conjunto de dados no problema, para assim adicionar e/ou excluir variáveis.

Conforme descrito em [14], a Seleção Sequencial à Frente (SFS) é iniciada com um conjunto vazio e adiciona um atributo do conjunto de dados a cada iteração, até que o melhor subconjunto de atributos seja obtido . Já a Seleção Sequencial Inversa (SBS) começará com um conjunto completo de atributos posteriormente removerá gradualmente os atributos irrelevantes, um por um. Por fim, a Eliminação Recursiva de Atributos (RFE) é um procedimento iterativo semelhante à SBS, porém há a possibilidade de eliminar um grupo de atributos a cada iteração.

Alguns exemplos de trabalhos recentes que fazem a comparação de desempenho de técnicas de redução de dimensionalidade em problemas de classificação são: A pesquisa de [20], em que há a comparação entre SFS, PCA e KPCA, o artigo [1], que compara as técnicas PCA e *Auto-Encoder* (AE) e o artigo [19], que compara LDA e PCA.

Neste trabalho, é realizada a comparação da redução de dimensionalidade com uma variedade maior de técnicas e diferentes tipos de bases de dados, de forma a auxiliar as pesquisas que utilizarão estes métodos para tarefas de classificação.

3 METODOLOGIA

O processo geral seguido nesta pesquisa pode ser descrito da seguinte maneira: i) primeiramente, foi feita a classificação dos conjuntos de dados utilizando todos os atributos. Após essa tarefa, ii) foi feita a redução de dimensionalidade dos dados originais com diferentes técnicas. Em seguida, iii) os subconjuntos obtidos com a redução de dimensionalidade foram submetidos ao mesmo algoritmo de classificação usados em (i). Finalmente, os

resultados obtidos serão avaliados usando o classificador *Random Forest*.

Para a realização dessas etapas foi utilizada uma máquina com processador Intel Core i5-8250U CPU @ 1.60GHz x 8 e memória RAM de 4GB.

3.1 Descrição dos Conjuntos de Dados

Neste trabalho, serão apresentados os resultados da redução de dimensionalidade em cinco conjuntos de dados. Quatro deles foram obtidos no repositório de bases de dados *UCI Machine Learning Repository* [1]. Essas bases de dados são amplamente utilizadas na literatura de redução de dimensionalidade e possuem diferentes características. Portanto, as técnicas de redução de dimensionalidade aplicadas não serão iguais para todos os conjuntos de dados. Com base na literatura foram analisados quais técnicas são aplicadas nas mesmas, e as técnicas encontradas (já apresentadas na Subseção 2.1) serão aplicadas no presente trabalho, para comparação do desempenho na classificação. A descrição das bases utilizadas está a seguir. Também são identificadas as técnicas que são usualmente utilizados com exemplos de referências que fazem a aplicação nos referidos conjuntos.

O conjunto de dados *Wisconsin Diagnostic Breast Cancer* é usado para classificar os exemplos em malignos (cancerígenos) ou benignos. São 569 exemplos, 357 benignos e 212 malignos. Cada exemplo possui 30 atributos do tipo real, os quais descrevem características dos núcleos celulares presentes em imagens de massa mamária. As técnicas usualmente aplicadas nessa base, com exemplos de uso na literatura são: CHI, utilizada, por exemplo, em [12], ICA [17], PCA [21],SVD em [22], SBS e SFS [6].

Heart Dataset tem 13 atributos, que foram extraídos de exames de 270 pacientes. Desses atributos, 6 são do tipo real, 3 do tipo binário e 4 do tipo nominal. Os pacientes foram divididos em 2 classes, que dizem a respeito da presença ou ausência de uma doença cardiológica. Das 270 amostras, 150 pertencem à primeira classe e 120 à segunda classe. As técnicas usualmente aplicadas na base *Heart* são: CHI [10], SBS [6], SFS [6], KPCA [8], RFE [16] e LDA [21].

O conjunto *Epileptic Seizure Recognition Data Set* é utilizado para o reconhecimento de crises epilépticas e está caracterizado como série temporal. O conjunto de dados corresponde à gravações de atividade cerebral, tem 11500 amostras, cada informação contém 178 (atributos) pontos de dados por segundo. As 11500 informações estão divididas em 5 classes, cada uma com 2300 amostras do tipo real. Todos os sujeitos das classes 2, 3, 4 e 5 são sujeitos que não tiveram convulsão epiléptica. Somente os sujeitos da classe 1 apresentam crise epiléptica. As técnicas usualmente aplicadas nesta base são: ICA [3], PCA e SVD em [13], DFT [24] e SFA [18].

O conjunto de dados *Lymphography Domain* diz respeito à exames de linfografia. O problema é distinguir as amostras analisadas em quatro classes diferentes. Os dados têm 148 amostras, sendo 2 da primeira classe, 81 da segunda, 61 da terceira e 4 da quarta classe. Cada exemplo possui 18 atributos do tipo categórico. As técnicas usualmente aplicadas na base *Lymphography* são: CHI [12], SBS e SFS em [6], LDA [11] e FA [5].

A base de dados *RIM-One* está disponível no repositório do Grupo de Análise de Imagens Médicas [2], da Universidade de Laguna (ULL). Esta base é constituída de 455

[1]Disponível em https://archive.ics.uci.edu/
[2]Disponível em http://medimrg.webs.ull.es/

imagens de retinografia, sendo que 255 são de pessoas que não tem glaucoma, e 200 de pessoas diagnosticadas com a doença. Para utilizar essa base de dados neste trabalho, foi criado um vetor de atributos com características de textura postuladas em [7] para representar as imagens coletadas. Nove recursos de textura (segundo momento angular, contraste, média da soma, variância, correlação, variação da soma, momento da diferença inversa, entropia e medida de correlação) foram submetidos a quatro ângulos de uma matriz de co-ocorrência, resultando em 36 atributos para cada imagem. Na base de dados RIM-ONE, as técnicas escolhidas foram: SVD, PCA, ICA, KPCA, FA e ISOMAP.

3.2 Pré-Processamento

Na fase de pré-processamento foi feita a redução da dimensionalidade dos conjuntos de dados descritos na seção anterior. As técnicas de redução de dimensionalidade utilizados nessa pesquisa foram implementados na linguagem de programação *Python*.

Para auxiliar a implementação dos algoritmos SVD, PCA, ICA, FA, KPCA, ISOMAP, LDA, CHI e RFE, foi utilizada a biblioteca *Scikit-learn*[3], que contém uma coleção de algoritmos para tarefas de análise de dados, dentre elas, redução de dimensionalidade. Outra biblioteca utilizada com os mesmos tipos de algoritmos foi a *MLxtend*[4], que auxiliou a implementação da SFS e SBS. Já a implementação dos algoritmos SFA e DFT, foi feita a partir do *Pyts*[5], um pacote Python dedicado à classificação de séries temporais.

3.3 Classificação

O objetivo do trabalho é reduzir a quantidade de atributos dos conjuntos de dados utilizados para melhorar o desempenho da classificação, e o algoritmo escolhido para esta tarefa foi o *Random Forest* [2], que induz um conjunto de árvores de classificação a partir de um conjunto de dados de entrada, que posteriormente são utilizadas para classificação de um novo exemplo. Cada uma das árvores de decisão utiliza um subconjunto de atributos aleatórios a partir do conjunto original e atribui uma classificação. A classificação que será escolhida é a que tiver o maior número de votos.

O algoritmo *Random Forest* foi executado com o *software* WEKA [6], na versão 3.8.4, com a configuração padrão e utilizando o método de validação cruzada (*cross-validation*) com 10 subconjuntos para avaliar a capacidade de previsão dos modelos obtidos.

As medidas de desempenho utilizadas para comparar o desempenho dos conjuntos de dados foram Acurácia e *F-measure*, indicando respectivamente o desempenho geral do modelo, e a média harmônica entre as métricas *Recall* (razão entre o número de exemplos positivos corretamente classificados e o número total de exemplos positivos originais) e *Precision* (razão entre o número de exemplos positivos corretamente classificados e o número total de exemplos identificados como positivos pelo classificador) [4].

A exceção é a base *Lymphography*, pois a separabilidade entre as classes dessa base apresenta significativa complexidade, o que gerou, nos experimentos, inconsistências ao calcular as métricas *Precison*, *Recall*, e, consequentemente, a *F-measure*. Nessas bases, serão avaliadas as métricas de Acurácia, bem como as taxas de Verdadeiro-Positivo (TPr)

[3]Disponível em scikit-learn.org
[4]Disponível em rasbt.github.io/mlxtend/
[5]Disponível em pyts.readthedocs.io
[6]Disponível em https://www.cs.waikato.ac.nz/ml/weka/

e Falso-Positivo (FPr). As duas últimas métricas, assim como a *F-measure*, avaliam o desempenho da classe positiva [4].

4 RESULTADOS

Na fase de pré-processamento, os dados originais foram submetidos às técnicas de redução de dimensionalidade até que restassem somente dois atributos, criando assim subconjuntos com dimensionalidade reduzida. Com base nos resultados da classificação, o melhor subconjunto produzido por cada técnica e os dados originais são comparados a seguir.

Na Tabela 1 é apresentado o resultado da classificação dos dados originais e com as técnicas CHI, ICA, PCA, SBS, SFS e SVD para o conjunto de dados *Breast Cancer*.

Tabela 1: Resultados da Classificação para a Base Breast Cancer.

Técnica	Nº de Atributos	Acurácia (%)	*F-measure* (%)
Dados Originais	30	96,1	96,1
CHI	23	**97,0** ↑	**97,0**
ICA	22	94,7	94,7
PCA	23	94,9	94,9
SBS	24	96,8	96,8
SFS	13	96,3	96,3
SVD	23	**93,1** ↓	93,1

A base original alcançou 96,1% de acurácia e 96,1% para a métrica *F-measure*. Conforme exposto na Tabela 1, as técnicas CHI, SBS e SFS afetaram positivamente a classificação ao reduzir o número de atributos, ainda que a classificação da base de dados original já fosse alta. Já os subconjuntos produzidos pelas técnicas ICA, PCA e SVD obtiveram um desempenho inferior na classificação em relação aos dados originais. Com o CHI, que obteve o melhor desempenho, foi alcançado 97% de acurácia e *F-measure* com o subconjunto que tinha 23 atributos, ou seja com 76,7% dos atributos originais.

Na Tabela 2 estão os resultados obtidos na classificação do conjunto de dados *Heart* e também os resultados com as técnicas CHI, KPCA, LDA, RFE, SBS e SFS.

Tabela 2: Resultados da Classificação para a Base Heart.

Técnica	Nº de Atributos	Acurácia (%)	*F-measure* (%)
Dados Originais	13	81,1	81,1
CHI	9	82,6	82,5
KPCA	8	82,9	82,9
LDA	1	81,1	81,1
RFE	11	81,9	81,8
SBS	3	**83,7** ↑	83,6
SFS	9	82,6	82,6

O classificador obteve 81,1% para as métricas Acurácia e *F-measure* com os dados originais. Todos as técnicas, com exceção da LDA, tiveram ganhos nas métricas avaliadas. Contudo, a LDA manteve a mesma classificação dos dados originais com um subconjunto que continha apenas 1 atributo, que equivale a 7,7% no número de atributos original. Das técnicas que afetaram positivamente os classificador, a SBS teve o melhor desempenho (resultado destacado com ↑). Alcançou 83,7% de acurácia e 83,6% para *F-measure*, no subconjunto com 3 atributos, que correspondem a 23,1% dos dados originais.

Na Tabela 3 são exibidos os resultados da classificação para o conjunto de dados *Epileptic Seizure Recognition*, bem como o resultado com a redução de dimensionalidade realizada pelas técnicas DFT, ICA, PCA, SFA, SVD e KPCA.

Tabela 3: Resultados para a Base Epileptic Seizure Recognition.

Técnica	Nº de Atributos	Acurácia (%)	*F-measure* (%)
Dados Originais	178	70,1	70,0
DFT	62	**74,8** ↑	74,4
ICA	45	73,7	73,3
PCA	89	73,8	73,0
SFA	45	**64,1** ↓	63,1
SVD	53	74,0	73,3

O classificador obteve 70,1% para as métricas Acurácia e 70,0% para *F-measure* com os dados originais. Com exceção da SFA, todos as técnicas afetaram positivamente o classificador. A SFA obteve um desempenho pior que os dados originais na classificação (destacado com ↓) ao reduzir o número de atributos, alcançando 64,1% de acurácia e 63,1% para a *F-measure*. O melhor desempenho para as métricas avaliadas foi obtido com a DFT (resultado em destaque com ↑), que alcançou 74,8% de acurácia e 74,4% para a *F-measure*. O subconjunto que obteve essa classificação possui 62 atributos, que corresponde a 34,8% do número de atributos originais.

Na Tabela 4 é apresentada a classificação do conjunto de dados *Lymphography* e os resultados com as técnicas CHI, FA, RFE, SBS, SFS e LDA.

Tabela 4: Resultados da Classificação para a Base Lymphography.

Técnica	Nº de Atributos	Acurácia (%)	TPr (%)	FPr (%)
Dados Originais	18	83,1	83,1	16,6
CHI	17	85,8	85,8	15,4
FA	4	83,8	83,8	15,8
RFE	15	85,1	85,1	14,5
SBS	17	85,1	85,1	15,6
SFS	16	87,8	87,8	13,0
LDA	3	87,2	87,2	13,1

Este conjunto de dados obteve na classificação 83,1% de acurácia, 83,1% para a taxa TPr, e 16,6% para a taxa FPr. Todas as técnicas aplicadas conseguiram melhorar as taxas de acurácia TPr e FPr, sendo que a técnica que menos afetou o classificador, em termos de métricas avaliadas, foi a FA. Porém, a técnica alcançou esses resultados com apenas 4 atributos, que correspondem a 22,2% dos atributos originais. A técnica SFS alcançou as melhores taxas nas métricas avaliadas, obtendo 87,8% de acurácia, 87,8% de TPr e 13,0% de FPr. A técnica LDA alcançou resultados semelhantes (87,2% de acurácia, 87,2% de TPr e 13,1% de FPr) com apenas 3 atributos, que correspondem a 16,7% do número de atributos originais, enquanto o melhor subconjunto produzido pela SFS continha 16 atributos, que são 88,9% dos atributos originais.

Finalmente, na Tabela 5, são apresentados os resultados da classificação do conjunto de dados *RIM-One*, bem como pelas técnicas SVD, PCA, ICA, FA, KPCA e ISOMAP.

De acordo com a Tabela 5, o classificador foi afetado principalmente pela técnica KPCA. A acurácia aumentou de 79,6% para 90,9%, e a *F-measure* de 79,5% para 91,0%.

Tabela 5: Resultados da Classificação para a Base RIM-One.

Técnica	Nº de Atributos	Acurácia (%)	*F-measure* (%)
Dados Originais	36	79,6	79,5
SVD	31	84,2	84,1
PCA	16	82,9	82,8
ICA	13	81,5	81,5
FA	14	82,6	82,6
KPCA	35	**90,9** ↑	91,0
ISOMAP	8	**61,3** ↓	61,2

Com exceção do ISOMAP, todas as técnicas de redução de dimensionalidade afetaram positivamente a classificação dos dados em relação ao conjunto original.

Na Tabela 6 é apresentado um resumo dos resultados obtidos neste trabalho, considerando, o tipo de atributos, a técnica que apresentou o melhor resultado nas métricas avaliadas e o ganho obtido na acurácia após a aplicação desta técnica.

Tabela 6: Resumos dos resultados obtidos.

Conjunto de Dados	Tipo de atributos	Melhor técnica	Ganho na acurácia
Breast Cancer	Real	CHI	+0.9%
Heart	Real, binário e nominal	SBS	+2.6%
Lymphography	Categórico	SFS	+4.7%
Epileptic Seizure Recognition	Série temporal	DFT	+4.4%
RIM-One	Real (atributos de textura)	KPCA	+11.3%

5 CONSIDERAÇÕES FINAIS

Este trabalho procurou investigar quais técnicas de redução de dimensionalidade tinham melhor desempenho para certos tipos de bases de dados comuns na literatura. Foram apresentados os resultados para cinco conjuntos, bem como quais eram os métodos de redução de dimensionalidade mais utilizados para os mesmos, ou bases semelhantes.

Na base *Breast Cancer*, composta de atributos do tipo real, a classificação original já era alta. Contudo, com a aplicação da técnica CHI foi possível incrementar os resultados utilizando 76,7% dos atributos originais. Já na base mista, *Heart*, a classificação foi melhorada com a técnica SBS, utilizando apenas 23,1% dos dados originais. Outra técnica de seleção sequencial, SFS, teve o melhor desempenho no conjunto de dados categóricos *Lymphography*, enquanto na base composta por uma série temporal, a *Epilepitic Seizure Recognition*, o melhor resultado foi obtido com a DFT, que com 34,8% do número de atributos originais melhorou consideravelmente a classificação. Por fim, no conjunto de dados *RIM-One*, que é composto 36 atributos extraídos de imagens de retinografia, a técnica KPCA alcançou resultados significativamente superiores nas métricas avaliadas.

Além de investigar quais técnicas tem melhor desempenho para aprimorar a tarefa de classificação em diferentes tipos de conjuntos de dados, este trabalho mostrou que algumas técnicas frequentemente utilizadas na literatura podem não afetar ou até mesmo afetar negativamente a classificação dos dados. Os resultados obtidos nessa pesquisa também

auxiliam a decidir qual técnica utilizar em pesquisas que necessitem da utilização de redução de dimensionalidade, a depender do tipo de base de dados com que se trabalha.

A principal sugestão de trabalho futuro é aprimorar a metodologia aqui utilizada, com o fim de estabelecer uma heurística para a seleção de técnicas de redução de dimensionalidade. Outra alternativa é comparar os resultados desta pesquisa com a classificação por outra abordagem, como a de Redes Neurais Artificiais. Mais uma possibilidade é aplicar este trabalho a problemas complexos de alta dimensionalidade, considerando tempos de execução dos algoritmos de redução de dimensionalidade e de classificação.

6 *Agradecimentos*

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e colaboração do Grupo de Pesquisa em Biologia Computacional e Reconhecimento de Padrões da Universidade Estadual de Santa Cruz (UESC).

Referências

[1] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour, and A. Abuzneid. Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electronics*, 8(3):322, 2019.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] L. Cao, K. S. Chua, W. Chong, H. Lee, and Q. Gu. A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1-2):321–336, 2003.

[4] C. L. d. Castro and A. P. Braga. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, 22(5):441–466, 2011.

[5] O. Claveria and A. Poluzzi. Positioning and clustering of the world's top tourist destinations by means of dimensionality reduction techniques for categorical data. *Journal of Destination Marketing & Management*, 6(1):22–32, 2017.

[6] P. Domingos. Context-sensitive feature selection for lazy learners. In *Lazy learning*, pages 227–253. Springer, 1997.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, (6):610–621, 1973.

[8] C.-C. Hsu and J.-W. Wu. Visualized mixed-type data analysis via dimensionality reduction. *Intelligent Data Analysis*, 22(5):981–1007, 2018.

[9] X. Huang, L. Wu, and Y. Ye. A review on dimensionality reduction techniques. *Inter. Journal of Pattern Recognition and Artificial Intelligence*, 33(10):1950017, 2019.

[10] M. Jabbar, B. Deekshatulu, and P. Chandra. Computational intelligence technique for early diagnosis of heart disease. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 1–6. IEEE, 2015.

[11] S. G. Jacob, R. G. Ramani, and P. Nancy. Discovery of knowledge patterns in lymphographic clinical data through data mining methods and techniques. In *Advances in computing and information technology*, pages 129–140. Springer, 2013.

[12] E. M. Karabulut, S. A. Özel, and T. Ibrikci. A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, 1:323–327, 2012.

[13] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

[14] S. Khalid, T. Khalil, and S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378. IEEE, 2014.

[15] R. Liu and D. F. Gillies. Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognition*, 53:73–86, 2016.

[16] M. Mafarja and S. Mirjalili. Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, 62:441–453, 2018.

[17] A. Mert, N. Kılıç, and A. Akan. An improved hybrid feature reduction for increased breast cancer diagnostic performance. *Biomedical Eng. Letters*, 4(3):285–291, 2014.

[18] K. G. M. Quispe, W. S. Lima, and E. J. P. Souto. Human activity recognition on smartphones using symbolic data representation. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pages 93–100, 2018.

[19] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker. Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8:54776–54788, 2020.

[20] R. T. Sousa. Avaliação de classificadores na classificação de radiografias de tórax para o diagnóstico de pneumonia infantil. Master's thesis, Universidade Federal de Goiás, Goiânia, 2013.

[21] R. Sujatha, E. Ephzibah, S. Dharinya, G. U. Maheswari, V. Mareeswari, and V. Pamidimarri. Comparative study on dimensionality reduction for disease diagnosis using fuzzy classifier. *International Journal of Engineering & Technology*, 7(1):79–84, 2018.

[22] S. Tanwar, T. Ramani, and S. Tyagi. Dimensionality reduction using pca and svd in big data: A comparative case study. In *International Conference on Future Internet Technologies and Trends*, pages 116–125. Springer, 2017.

[23] J. B. Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in neural information processing systems*, pages 682–688, 1998.

[24] S. Vijay Kartik, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux. A fourier dimensionality reduction model for big data interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 468(2):2382–2400, 2017.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

TÉCNICAS DE BIOFÍSICA COMPUTACIONAL PARA AVERIGUAÇÃO DE POSSÍVEIS INIBIDORES DA GLUTAREDOXINA A1

Charlene Marcondes Avelar[1], *Marcos Serrou do Amaral*[1], *Danilo da Silva Olivier*[2]
1 - Universidade Federal de Mato Grosso do Sul - UFMS/INFI, Campo Grande/MS, Brazil.
2 - Universidade Federal de Tocantins - UFT, Araguaína/TO, Brazil.

RESUMO

Este trabalho aborda elementos a respeito da bactéria *Corynebacterium pseudotuberculosis* (Cp) que causa uma doença chamada Linfadenite Caseosa, a qual afeta animais de diversas raças e sexo. A Cp, assim como qualquer ser vivo, sofre mutações em seu DNA através da interação com Espécies Reativas de Oxigênio (ERO). Por esse motivo, os organismos desenvolveram vários sistemas de defesa para proteger seus genomas de danos oxidativos. Assim, neste trabalho o sistema Glutaredoxina A1 (GrxA1) codificado pela bactéria Cp foi estudado.

A caracterizacação bioquímica e estrutural desta proteína tem potencial para o desenvolvimento de inibidores, seguidos pelo desenvolvimento de agentes farmacológicos contra *Corynebacterium pseudotuberculosis*. A partir de métodos *in silico* como Modelagem por Homologia, Dinâmica Molecular, *Docking Molecular* e *Virtual Screening*, buscou-se encontrar moléculas que sejam promissoras no bloqueio do acesso ao sítio ativo da GrxA1 e, consequentemente, que sejam capazes de destruir a bactéria. Encontrou-se um conjunto com 26 candidatos a inibidores da *C. pseudotuberculosis*. Os resultados servem para estudos futuros, permitindo desenvolver medicamentos eficientes, trazendo benefícios em aplicações biotecnológicas.

Palavras-chave: *Corynebacterium pseudotuberculosis*, Dinâmica Molecular, Homologia, Linfadenite Caseosa, *Virtual Screening*.

[1]Contato: Charlene Marcondes Avelar, marcondesavelar@yahoo.com.br

1 - INTRODUÇÃO

A criação de caprinos e ovinos tem sido uma alternativa de alimentação para boa parte dos brasileiros. Além da carne e do leite, o couro e/ou a lã tem permitido também a obtenção de uma renda extra para a agricultura de subsistência [1]. O Brasil tem um grande potencial de crescimento em relação à essa criação, porém esses animais exigem cuidados específicos com a saúde, além de zelo pela higiene dos produtos [1].

O Brasil conta com um rebanho de cerca de 11,3 milhões de caprinos e 19,7 milhões de ovinos de modo que essa criação tem papel relevante na economia familiar, no agronegócio e na agroindústria, segundo dados do Instituto Brasileiro de Geografia e Estatísitica (IBGE) [2]. Os rebanhos estão distribuídos em todas as regiões, porém de forma desigual. A região Nordeste concentra 57,5% do total de ovinos, seguida pelas regiões Sul (29,3%), Centro-Oeste (5,5%), Sudeste (4,0%) e Norte (3,6%). Quanto aos caprinos, a distribuição refere-se a 91,6% na Região Nordeste, 3,5% na região Sul, 2,2% Sudeste, 1,6% no Norte e 1,0% na região Centro-Oeste [3].

Uma das características principais na produção destes tipos de animais é que são formados em pequenas propriedades caracterizadas pela agricultura familiar, gerando ambém emprego e renda nas regiões de criação. Os maiores produtores de ovinos e caprinos estão situados nas regiões nordeste e sul do país, entretanto, podemos observar a criação desses animais por todo o território brasileiro, fazendo do agronegócio um importante segmento econômico, responsável por grande parte do Produto Interno Bruto [4].

Apesar dos pontos positivos, devemos ressaltar que existem fraquezas que devem ser sanadas, tais como as infecções que afetam os rebanhos e causam imensas perdas. Dentre as doenças que acometem os rebanhos, podemos destacar a Linfadenite Caseosa. Doença crônica e infectuosa, ela tem sido um incômodo significativo na maioria das regiões por mais de um século e tem se mostrado difícil de controlar e a prevalência é alta em muitas partes do mundo [5]. A partir de dados epidemiológicos, verifica-se que dentre as regiões afetadas pela doença, encontram-se países como África do Sul, Austrália, Argentina, Brasil, Canadá, Chile, Estados Unidos, França, Itália e Inglaterra [6].

Glutaredoxinas (Grx) são pequenas proteínas enzimáticas oxidorredutases que auxiliam na conservação de ambientes intracelular, importantes para desintoxicar a oxidação de agentes. As Grx são de baixo peso molecular (9-12 kDa) consistindo de uma folha β antiparalela de quatro fitas central rodeado por três hélices α (Figura 1) [7].

As Grx possuem o sítio catalítico na sequência de cisteínas -CXXC- e tem a função de atuar

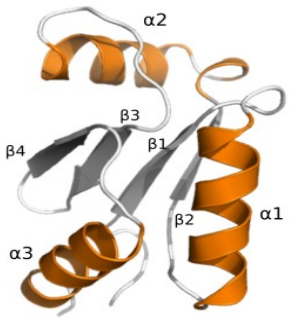


Figura 1: Enzima Glutaredoxina com nomeação das estruturas secundárias.
Adaptado de D. S. Olivier, 2021.

na redução de outras proteínas no meio intracelular e, dessa forma, manter o equilíbrio celular dentro das condições normais. Para que a molécula esteja ativa, suas cisteínas precisam estar reduzidas, sendo que a interação enzimática fará com que a proteína alvo seja reduzida e a Glutaredoxina oxidada, gerando uma ligação dissulfeto entre as cisteínas C14 e C17, a oxidação das cisteínas inativam a proteína.

Para que as Grx se tornem ativas novamente é necessário que sejam reduzidas pela ação específica da Glutationa (GSH) ou *Mycothiol* (MSH). Grande parte das bactérias gram positivas tais como *C. pseudotuberculosis* produz apenas a MSH, neste caso, a glutaredoxina A1 atua como uma *mycoredoxina* (proteína exclusiva dos actinomicetos) em interação com MSH. Alguns fatores que alteram o valor de pKa das cisteínas catalíticas são a geometria do sítio catalítico, pH do ambiente, aminoácidos circundantes, bem como a interação com ligantes e interação com outras proteínas [8, 9, 10].

Assim, identificações de ligantes que bloqueiam a atividade da proteína podem ser úteis para o desenvolvimento de fármacos [11]. A inibição da proteína pode resultar em várias disfunções, como o aumento em toxicidade de EROs. Deste modo, o controle do metabolismo *redox* é uma das abordagens mais promissoras para medicamentos, uma vez que é de importância primária em patógenos de crescimento rápido, como *C. pseudotuberculosis*.

2 - METODOLOGIA

A estrutura da proteína estudada, por não estar disponível no banco de dados PDB, foi obtida por meio de modelagem por homologia, através da sequência da base de dados UniProt - D9Q987. Um modelo para a estrutura 3D da Cp-GrxA1 foi proposto baseado na estrutura de *Mycoredoxin*-1 (Mrx1 - reduzida) de *Mycobacterium tuberculosis* (PDB ID: 2LQO), *template* indicado pelo servidor *Swiss-Model*, devido a sua melhor identidade.

Para uma representação estabilizada da estrutura, foi realizada dinâmica molecular utilizando o pacote computacional Amber18, em pH 7,4 com o modelo reduzido.

A proteína foi centrada e solvatada em caixa d'água octaédrica, cuja superfície tem distância de 15 Å de qualquer átomo de proteína e preenchida com moléculas de água com o modelo TIP3P [12]. As cargas do sistema foram neutralizadas com íons de sódio (Na_+), enquanto o campo de força $ff19SB$ foi usado para representação dos potenciais interatômicos, originando os arquivos de topologias e coordenadas.

Minimizações iniciais de energia do sistema com proteína completa foram realizadas em dois estágios, sendo que na etapa de aquecimento e equilibração os resíduos da proteína foram restringidos e simulados e, num segundo momento, uma outra equilibração com número de partículas, pressão e temperaturas constantes foi realizada sem forças ou quaisquer restrições. A temperatura foi ajustada pelo termostato Langevin em 298 K, obtendo um relaxamento termodinâmico.

Por fim, a etapa de produção foi realizada em duas etapas sequenciais de 100 ns cada, com temperatura de 298 K, pressão de 1 atm, intervalo de tempo de 2 fs e sem qualquer restrição da conformação da proteína, sendo os dados relativos à trajetória desses tempos coletados a cada 10 ps. Com o mesmo estado inicial a fim de fazer uma avaliação estatística da estabilidade da proteína, foram realizadas três réplicas da dinâmica molecular, com temperatura a 298 K.

O *docking* molecular, feito por meio da técnica de *Virtual Screening* (VS) através do software AutoDock Vina, foi realizado para encontrar moléculas que sejam capazes de interagir com o sítio catalítico -CXXC- da proteína GrxA1. Utilizou-se uma biblioteca com 8.823 compostos, proveniente do *DrugBank*, buscando moléculas Drug-like. A triagem foi baseada na estrutura do receptor (proteína GrxA1), o qual permaneceu rígido, tendo os compostos flexíveis e restritos nas dimensões da caixa em volta do sítio, onde cada conformação gerou uma energia de afinidade diferente. Para projeção da caixa, foi considerado apenas as extensões do sítio ativo da proteína, visto que possíveis cofatores possam se ligar em áreas diferentes podendo influenciar a energia total de ligação e espontaneidade ou não do encaixe. As coordenadas tridimensionais do centro foram $centro\_x = 18,526$ Å, $centro\_y = -7,449$ Å, $centro\_z = -3,713$ Å e o tamanho da caixa em: x = 16 Å, y = 20 Å, z = 16 Å.

3 - RESULTADOS E DISCUSSÕES

A proteína GrxA1 tem um sítio ativo com duas cisteínas catalíticas que podem estar em

dois estados: reduzidas ou oxidadas. No estado reduzido a proteína está ativa e assim, o ideal é bloquear a proteína de modo a impedir que ela atue reduzindo outra proteína, sendo que a determinação da estrutura em sua forma reduzida destaca a importância fisiológica na manutenção do metabolismo antioxidante. Assim, o template escolhido para a estrutura tridimensional Cp-GrxA1, tem sua forma no estado reduzido com uma sequência de 83 resíduos, com o sítio catalítico preservado nas cisteínas da sequência CPFC - C14 e C17, conforme a Figura 2.

Figura 2: Alinhamento das sequências 2lqo.1 e o modelo D9Q987, em amarelo a posição das císteínas.

Fonte: Servidor web Swiss-Model.

Para validação do modelo tridimensional da proteína GrxA1, utilizamos as pontuações do servidor Swiss-Model, a estrutura apresentou em torno de 46% de identificação entre sequências e uma cobertura aproximadamente em 99% com o template *Mycoredoxin-1* (Mrx1 - reduzida). As pontuações acima permitiram avaliar quantitativamente a confiabilidade do template e a verificação da qualidade do modelo, então foi gerada a estrutura tridimensional da Figura 3.

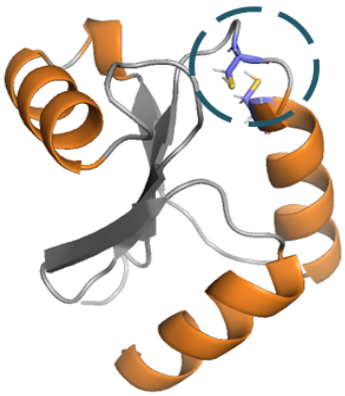


Figura 3: Estrutura gerada por homologia para Glutaredoxina A1 de *C. pseudotuberculosis*, em destaque o sítio ativo, cisteínas em sua forma reduzida.

Pela observação, deve-se escolher uma das réplicas e a conformação representativa, dentre todas as diversas conformações geradas ao longo do tempo pela dinâmica molecular dessa réplica. Esta conformação representará o modelo refinado que será utilizado para *Virtual screening*.

Observa-se na Figura 4 a evolução e a estabilidade das réplicas ao longo do tempo.

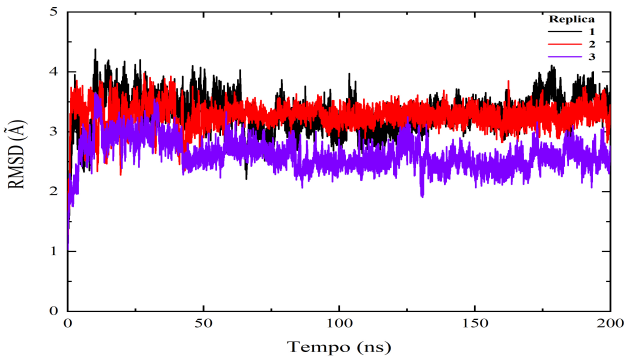


Figura 4: Gráficos comparativo entre os valores de RMSD obtidos para as três primeiras réplicas da proteína GrxA1 em função do tempo, na temperatura 298 k.

O RMSD da trajetória, mostra um equilíbrio na estabilidade, não ocorrendo mudanças bruscas em relação as posições iniciais. Ainda assim, a réplica 2 em vermelho, apresenta os menores picos de oscilação entre as estrutura, correspondendo a melhores posições de equilíbrio do sistema ao longo de 200 ns.

Já os raios de giração resultaram em valores referentes à compactação estrutural das réplicas e assim a dimensão geral da proteína (Figura 5). Tais valores foram obtidos a partir da medição da distância entre os átomos de carbono alfa (considerando sua massa e posição) e o centro de massa da molécula.

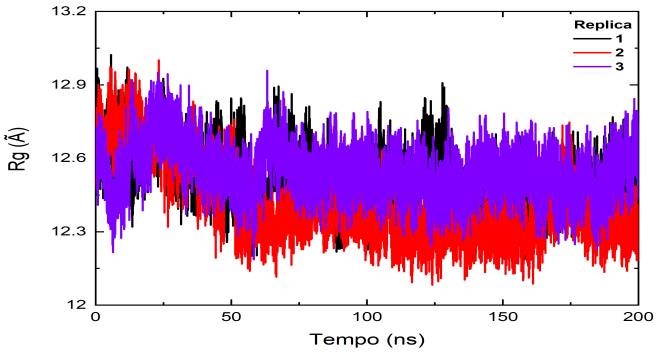


Figura 5: Gráfico comparativo dos valores de Raio de giro (Rg) para estudo da compactação das estruturas proteicas durante a simulação da dinâmica.

Na análise do raio de giro não obtivemos expressivas alterações de valores, o que indica que o enovelamento se manteve estável durante a simulação nas três réplicas.

Para estudo da flexibilidade dos diferentes resíduos o RMSF é calculado (Figura 6), comparando-se as estruturas obtidas a cada passo da simulação com as estruturas iniciais antes da Dinâmica Molecular, utilizando os átomos de carbono alfa da proteína.

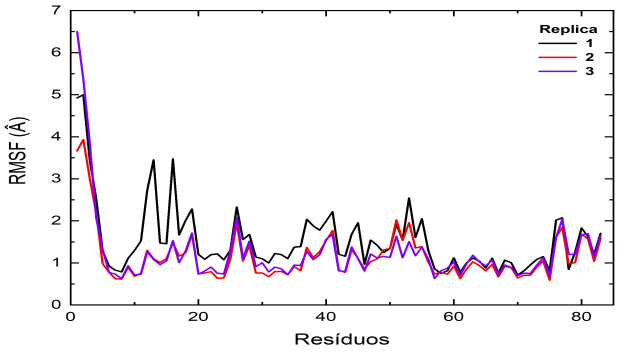


Figura 6: Gráfico comparativo da flutuação dos 83 resíduos das réplicas.

Quando comparamos os RMSF das réplicas, vemos que as estruturas 2 e 3, na região ativa da proteína entre os resíduos 10-20, apresentam pequenas flutuações, ou seja, suas posições são conservadas ao longo do tempo de simulação. Isso sinaliza que o sítio catalítico apresenta estabilidade estrutural, mesmo após ser submetido a transformações físicas que definem a dinâmica.

A partir das observações evidenciadas, vemos que não obtivemos significativas desconformidades entres às réplicas, seria necessário análises mais particularizadas. Ainda assim, foi considerada a réplica 2 como sendo o sistema mais constante dentre as três e ela será utilizada para as etapas posteriores do trabalho. Para escolher uma conformação representativa da proteína GrxA1, utilizamos a análise de *cluster* com o algoritmo *K-means*. A avaliação foi realizada variando o número de *clusters* de 2 a 10 e a partir de métricas de qualidade *Davies Bouldin Index* (DBI), pseudo Eststítica F (pSF) e SSR/SST da conformação.

Após as etapas da triagem virtual, o conjunto individual de simulações identificou com maior precisão a capacidade de cada ligante em atingir a menor energia de interação e, consequentemente, a conformação de mínima energia. Com esse refinamento, foi possível classificar os melhores compostos conforme a Tabela 1. Essas moléculas formam um conjunto de 26 candidatos a inibidores da proteína GrxA1 de *C. pseudotuberculosis*. A tabela enumera dados como a posição no *ranking* de cada composto e a identificação de cada molécula. Já a classificação acompanha o critério de menor energia na interação proteína-ligante, em caso de repetição dessa energia, seguindo a ordem da melhor taxa de sucesso na obtenção do valor de menor energia

em 100 simulações, junto da menor quantidade de átomos pesados, trazendo ainda o número de torções entre átomos.

Tabela 1: Resíduos-chave obtidos nos complexos entre a GrxA1 e os 26 compostos triados.

Classificação	Composto	Torções	Átomos*	Energia(kcal/mol)	Taxa de Sucesso %
1	DB13014	8	38	-8,6	100%
2	DB14878	3	28	-7,7	87%
3	DB08006	3	25	-7,1	97%
4	DB07435	3	23	-6,8	100%
5	DB12886	5	30	-6,7	86%
6	DB13991	4	24	-6,7	81%
7	DB04064	6	28	-6,6	100%
8	DB07453	1	21	-6,6	99%
9	DB15039	6	38	-6,6	96%
10	DB06732	1	21	-6,5	100%
11	DB07949	4	25	-6,5	88%
12	DB15305	4	30	-6,4	92%
13	DB07430	4	27	-6,4	82%
14	DB07993	5	24	-6,4	81%
15	DB07201	6	21	-6,3	96%
16	DB07247	5	27	-6,3	84%
17	DB07296	3	25	-6,3	81%
18	DB12211	5	26	-6,3	80%
19	DB12379	0	20	-6,2	100%
20	DB04716	4	23	-6,2	100%
21	DB08707	2	23	-6,2	100%
22	DB11636	5	24	-6,2	97%
23	DB15308	3	30	-6,2	91%
24	DB14885	5	35	-6,2	82%
25	DB12868	8	26	-6,1	92%
26	DB04059	2	20	-6,1	87%

*Átomos pesados sem hidrogênio.

Foram estudados os tipos de interação proteína-ligante para cada molécula encontrada. Assim, apresentamos os resultados para uma das estruturas encontradas, Figura 7.

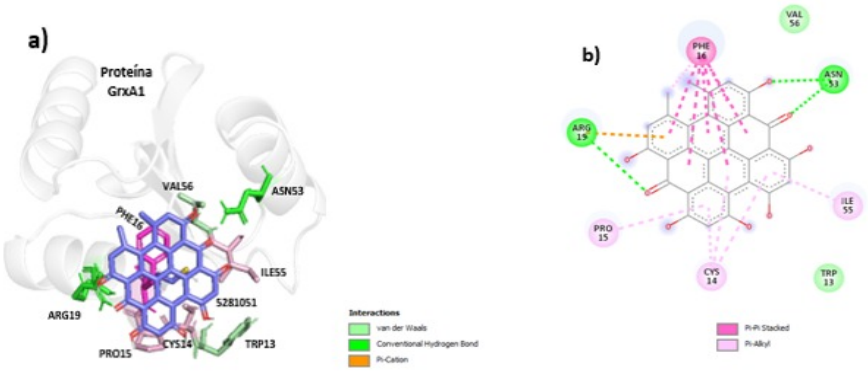


Figura 7: Representação da interação entre o composto DB13014 e a proteína GrxA1.
A Figura a) em 3D e b) 2D, apresentam os aminoácidos mais relevantes para a interação, bem como os tipos de forças intermoleculares atuantes.

Análises indicam que a maioria das interações são favoráveis e o conjunto de 26 compostos selecionados pode vir a ser um forte candidato a bloquear o acesso ao sítio catalítico da enzima GrxA1. Destaca-se nele a afinidade na interação receptor-ligante, as interações com os resíduos

onde ocorre a atividade da proteína. Além disso, temos a possibilidade dos compostos formarem um grupo farmacofórico.

4 - CONCLUSÕES

O presente trabalho foi divido em duas etapas, na primeira os estudos foram focados em propor uma estrutura tridimensional, por meio de modelagem por homologia, otimização e simulação do modelo por Dinâmica Molecular (DM) por 200 ns, sendo possível obter a conformação mais estável, mantendo o sítio catalítico preservado nas cisteínas reduzidas da sequência CPFC - C14 e C17 da proteína GrxA1 de *C. pseudotuberculosis*.

Com a estrutura representativa, na segunda etapa realizamos ancoragem molecular entre a proteína e 8.823 compostos do banco de dados *DrugBank*, que já são drogas comerciais. Assim, utilizando a técnica de triagem virtual das pequenas moléculas, conseguimos estabelecer um conjunto de 26 candidatos com características promissoras para bloquear o acesso ao sítio catalítico da enzima GrxA1. É importante ressaltar que não há paralelo direto entre a energia de interação obtida via *docking molecular* e dados experimentais. Além disso, o posicionamento efetivo da molécula em relação a proteína é apenas uma sugestão baseada em cálculos de energia.

Demos um passo significativo, a partir desse refinamento e combinações de técnicas para que outros estudos possam indicar e ratificar se algumas das moléculas selecionadas serão capazes de inibir a enzima *in vitro*.

Pretendemos aprimorar a metodologia descrita neste estudo através de técnicas computacionais, como *Machine Learning* e Dinâmica Molecular, para desenvolver simulações da enzima GrxA1 complexadas com os candidatos a ligante e analisar cada interação. Também almejamos constituir parcerias com demais grupos de pesquisas, para comparação dos resultados, no sentido de obtermos moléculas com mais alto potencial de se tornarem fármacos, trazendo benefícios em aplicações biotecnológicas e contribuindo para auxiliar no tratamento das infecções ou a cura da Linfadenite Caseosa, evitando grandes perdas na pecuária e gerando assim grande interesse econômico.

5 - AGRADECIMENTOS

Apoio: FUNDECT, CNPq, CAPES e PROPP/UFMS.

Referências Bibliográficas

[1] EMBRAPA, ***Boletim do Centro de Inteligência e Mercado de Caprinos e Ovinos***, vol. 9. https://www.bdpa.cnptia.embrapa.br, 2019.

[2] EMBRAPA, "**Novo Censo Agropecuário:** ," *Crescimento de Efetivo de Caprinos e Ovinos no Nordeste*, 2019.

[3] M. Balan, J. Daltio, and C. de LUCENA, "**Uso de painéis interativos para publicação de dados espaço-temporais: caso do Centro de Inteligência e Mercado de Caprinos e Ovinos da Embrapa.**," in *Embrapa Territorial-Artigo em anais de congresso (ALICE)*, In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA, 20., 2019. São José dos Campos . . . , 2019.

[4] FNP, "**Anuário da Pecuária Brasileira**," *ANUALPEC*, vol. , no. 20, 2018.

[5] G. Baird and M. Fontaine, "**Corynebacterium pseudotuberculosis and its role in ovine caseous lymphadenitis**," *Journal of comparative pathology*, vol. 137, no. 4, pp. 179–210, 2007.

[6] D. F. Alves, P. L. G. Carvalho, O. S. Costa, M. Anderson, and A. Vasco, "**Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence**," *Veterinary research*, vol. 37, no. 2, pp. 201–218, 2006.

[7] J. L. Martin, "**Thioredoxin—a fold for all reasons**," *Structure*, vol. 3, no. 3, pp. 245–250, 1995.

[8] D. A. Mavridou, J. M. Stevens, S. J. Ferguson, and C. Redfield, "**Active-site properties of the oxidized and reduced C-terminal domain of DsbD obtained by NMR spectroscopy**," *Journal of molecular biology*, vol. 370, no. 4, pp. 643–658, 2007.

[9] D. A. Mavridou, J. M. Stevens, A. D. Goddard, A. C. Willis, S. J. Ferguson, and C. Redfield, "**Control of periplasmic interdomain thiol: disulfide exchange in the transmembrane oxidoreductase DsbD**," *Journal of Biological Chemistry*, vol. 284, no. 5, pp. 3219–3226, 2009.

[10] A. T. Setterdahl, P. T. Chivers, M. Hirasawa, S. D. Lemaire, E. Keryer, M. Miginiac-Maslow, S.-K. Kim, J. Mason, J.-P. Jacquot, C. C. Longbine, *et al.*, "**Effect of pH on the Oxidation- Reduction Properties of Thioredoxins**," *Biochemistry*, vol. 42, no. 50, pp. 14877–14884, 2003.

[11] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, *et al.*, "**swiss-model: homology modelling of protein structures and complexes**," *Nucleic acids research*, vol. 46, no. W1, pp. W296–W303, 2018.

[12] R. W. Hockney and J. W. Eastwood, ***Computer simulation using particles***. crc Press, 2021.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Modelagem dos Elementos Parasitas em um Conversor Buck

Gustavo Eckhardt[1], Leonardo Luan Moreira Serpa Sá[1], Paulo Sérgio Sausen[1], Maurício de Campos[1], Airam Teresa Zago Romcy Sausen[1] e João Manoel Lenz[1]

[1] *Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí/RS, Brasil*

Resumo

Este trabalho apresenta a modelagem computacional e análise dos elementos parasitas de um conversor CC-CC. Tradicionalmente, os circuitos de conversores são estudados considerando os componentes como ideais. Contudo, na prática, existem diversas não-idealidades e elementos parasitas dos componentes passivos, ativos e da placa de circuito impresso que interferem no desempenho eletromagnético do conversor. Para aumentar a eficiência de conversores operando em alta frequência é preciso conhecer o impacto destes parasitas, especialmente na comutação nas chaves semicondutoras. Para isso, este trabalho apresenta a discussão e modelagem das não-idealidades dos capacitores e indutores considerando modelos e parâmetros fornecidos pelos fabricantes. Ainda, são analisados os elementos parasitas das trilhas de uma placa de circuito impresso para um conversor Buck, realizada por simulação em método de elementos finitos no software Ansys Q3D. Resultados de comparação do conversor Buck sem e com os elementos parasitas são apresentados.

Palavras-chave: Modelagem computacional, Elementos parasitas, Conversores estáticos, Comutação de MOSFETs.

1 Introdução

Nos últimos anos, grandes esforços estão sendo empregados na área da Eletrônica de Potência para que os conversores estáticos apresentem maior eficiência e confiabilidade, e ainda se tornem mais baratos e compactos [10]. Existem alguns métodos, para realizar esta otimização, que são normalmente utilizados, tais como, aumentar a frequência de chaveamento, minimização de elementos parasitas na estrutura da placa de circuito impresso (PCI), técnicas de gerenciamento térmico, e ainda a utilização de novas tecnologias de dispositivos semicondutores, como por exemplo, os transistores de *Galium Nitride* (GaN) e *Silicon Carbide* (SiC) [10, 14].

Contato: Gustavo Eckhardt, eckhardt.gustavo@gmail.com

Entretanto, um aumento significativo na velocidade de chaveamento, implica em um aumento da influência dos elementos parasitas característicos da PCI e dos componentes utilizados no funcionamento do circuito, adicionando oscilações indesejadas [7, 14]. Ainda vale ressaltar, que estas oscilações em alta frequência, são as principais fontes de interferência eletromagnética (EMI) [7].

Neste contexto, a presença destes elementos parasitas não pode mais ser desprezada, visto que eles afetam o desempenho, a estabilidade e a eficiência do conversor [13, 14]. Portanto, é necessário obter com precisão os elementos parasitas, para que possam ser analisados e minimizados [7, 14].

Deste modo, este trabalho tem como objetivo demonstrar uma metodologia para extração dos elementos parasitas de uma PCI, de um conversor CC-CC Buck. E ainda, realizar uma análise comparativa do funcionamento do circuito, sem e com os elementos parasitas, com o intuito de demonstrar a influência destes elementos no circuito.

2 Modelagem de Elementos Parasitas

Os componentes utilizados em circuitos elétricos não são ideais e estas não idealidades tornam-se mais significativas com o aumento da frequência [12]. Componentes como indutores, capacitores, resistores, fios e trilhas, muitas vezes considerados ideais, também possuem impedâncias que variam com a frequência, devido as suas não idealidades [12].

Essas ocorrem devido ao processo de fabricação dos componentes, conforme o material utilizado. Por exemplo, em um resistor, indutâncias e capacitâncias parasitas podem estar acopladas [2]. Esta mudança devido ao material utilizado, pode ser percebida em resistores de fio, que apresentam maiores indutâncias parasitas, se comparado a resistores de carbono [2]. E ainda, dependendo da frequência utilizada, a característica do elemento pode mudar completamente, passando de um comportamento capacitivo para um indutivo e vice-versa [12].

Além das características internas do componente, seus terminais também podem exercer influência nas não linearidades acopladas ao circuito [2]. Isto ocorre de maneira mais acentuada no caso de alguns componentes do tipo *pin through hole* (PTH), que possuem terminais mais longos. Entretanto, para contornar este problema, normalmente são utilizados elementos do tipo *surface-mount device* (SMD), que possuem pequenas placas para serem soldadas diretamente sobre a PCI [2].

De maneira similar, em altas frequências as características das trilhas da PCI também começam a impactar o funcionamento do circuito [2]. Sua resistência e indutância associadas variam de acordo com as características construtivas de uma PCI, como a distribuição das trilhas e dos materiais utilizados nesta. Entretanto, para testes de conformidade de EMI, a indutância apresenta uma influência maior [2].

2.1 Conversor CC-CC Buck

Como neste trabalho serão discutidos os efeitos dos elementos parasitas em um conversor CC-CC Buck, convém apresentar sua topologia. Este realiza basicamente a conversão de um nível CC maior, para uma menor amplitude. Devido a esta característica ele também é comumente chamado de conversor abaixador [4]. Ainda será realizado um chaveamento síncrono no circuito, onde uma segunda chave é inserida no lugar do diodo.

Com esta técnica, obtém-se uma maior eficiência do circuito [4]. A topologia do conversor é demostrada na Fig. 1, onde

- C_{in} e C_o são os capacitores de entrada e saída, respectivamente;
- L é o indutor de potência;
- Ro é a carga;
- V_{in} é a fonte de tensão de entrada; e
- MOS_HS e MOS_LS são as chaves semicondutoras *high-side* e *low-side*, respectivamente.

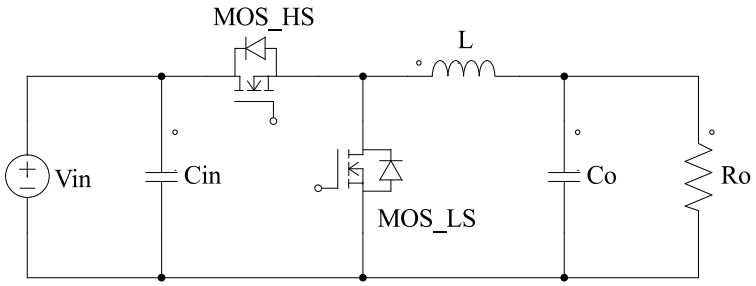


Fig. 1: Conversor CC-CC Buck considerando componentes ideais.

Vale salientar que a Fig. 1 demonstra a topologia do conversor considerando os componentes como ideais, sem elementos parasitas. Entretanto, como comentado anteriormente, a consideração destes elementos no circuito é de suma importância, visto os efeitos que estes podem causar no funcionamento do circuito.

2.2 Conversor CC-CC Buck com Elementos Parasitas

Como descrito na seção 2, os elementos parasitas podem ser oriundos principalmente de três locais, dos componentes, dos terminais de conexão dos componentes e da PCI, sempre devido as características construtivas.

Neste trabalho, somente dois destes locais serão abordados, os elementos parasitas intrínsecos aos componentes e a PCI. Visto que em projetos de conversores onde é preferível um menor volume, sempre são utilizados componentes do tipo SMD, e como dito anteriormente, seus efeitos na instauração de parâmetros parasitas no circuito, podem ser desprezados.

De modo a considerar elementos parasitas intrínsecos aos componentes no conversor, primeiramente serão abordados os capacitores e indutores. Para realizar um refinamento no modelo destes componentes, é necessário adicionar um resistor em série [11].

A resistência interna, nos capacitores, chamada de *equivalent series resistance* (ESR), se deve principalmente pelo dielétrico utilizado em sua construção [11]. Por exemplo, os capacitores cerâmicos possuem uma ESR baixa e podem ser desconsiderados em algumas aplicações [6]. Isso não ocorre nos demais tipos de capacitores, que possuem grandes valores de capacitância, mas como contraponto, uma ESR maior [11].

Nos caso dos indutores, a resistência interna é chamada de *direct current resistance* (DCR) e seu valor é inerente ao metal utilizado no condutor, que é bobinado [11]. Este é um parâmetro importante, visto que uma grande resistência também pode vir a alterar o funcionamento do circuito [11].

Para realizar o controle da taxa de trabalho do circuito, foram escolhidos transistores do tipo MOSFET. Estes também possuem seus elementos parasitas, devido as características de construção deste semicondutor [12]. Neste caso, capacitâncias são acopladas entre os terminais da chave, ou seja, entre *gate*, *drain* e *source*, e por fim, ainda existem resistências internas associadas em série em cada terminal [12].

Por fim, as trilhas de uma PCI, que tem como função interligar os componentes inseridos na mesma, também adicionam elementos parasitas ao circuito. Como essas trilhas entre componentes são muito curtas, o comportamento dos elementos parasitas pode ser modelado por uma resistência e uma indutância em série [2]. Os valores destes elementos variam de acordo com a frequência utilizada, devido ao efeito pelicular [2].

Finalmente, a Fig. 2 mostra o conversor CC-CC apresentado na Fig. 1, mas agora com a adição dos elementos parasitas intrínsecos aos componentes e a PCI, conforme foram descritos acima.

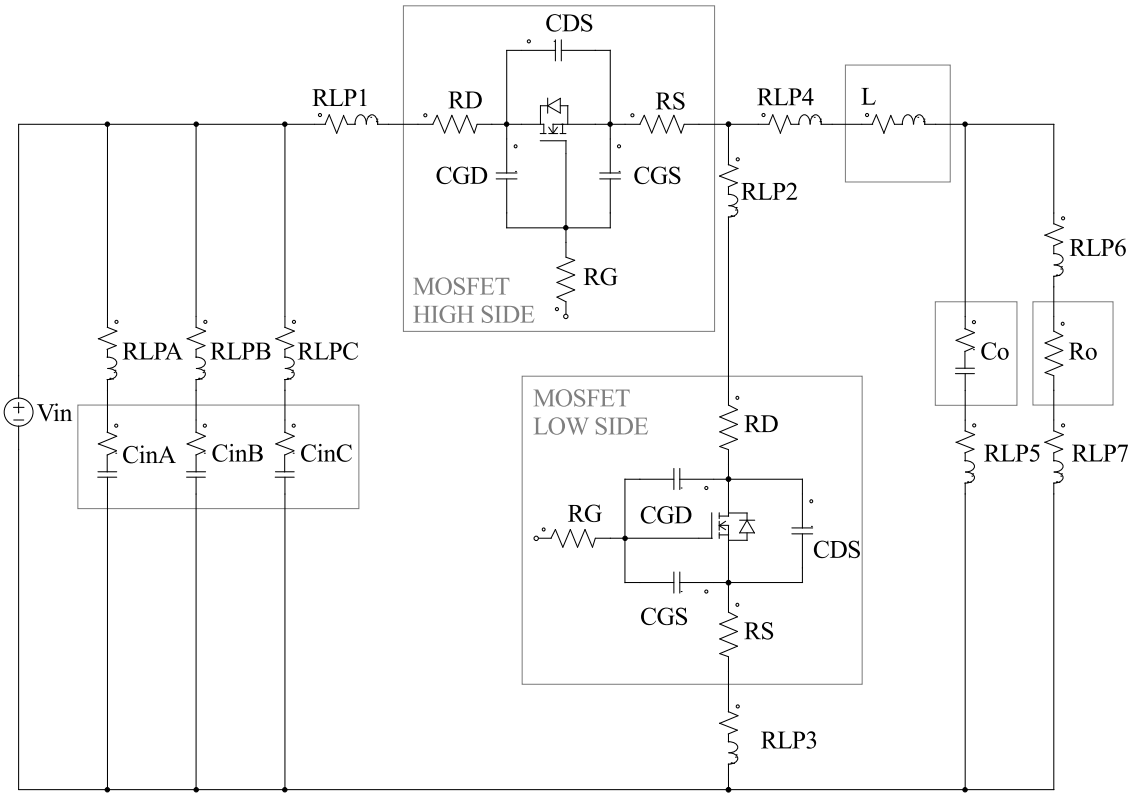


Fig. 2: Conversor CC-CC Buck considerando os elementos parasitas.

2.3 Extração dos Elementos Parasitas

Os parâmetros parasitas podem ser estimados experimentalmente ou por simulação eletromagnética, utilizando métodos numéricos que levam em conta a geometria e os materiais

utilizados [7]. Para isso, existem dois métodos principais de simulação, o método de circuito equivalente de elemento parcial (PEEC, *partial element equivalent circuit*) e o método de análise de elemento finito (FEA, *finite element analysis*) [7].

Neste trabalho, será utilizado o segundo método, através da ferramenta Ansys Q3D Extractor disponível no software Ansys Electronics. Esta ferramenta calcula os parâmetros parasitas presentes nas trilhas condutoras da PCI como resistência, indutância, capacitância dependentes da frequência [1]. O *layout* da PCI utilizada neste trabalho é mostrada na Fig. 3. Os resultados da modelagem de parasitas obtidas pelo Ansys Q3D são apresentados na próxima seção.

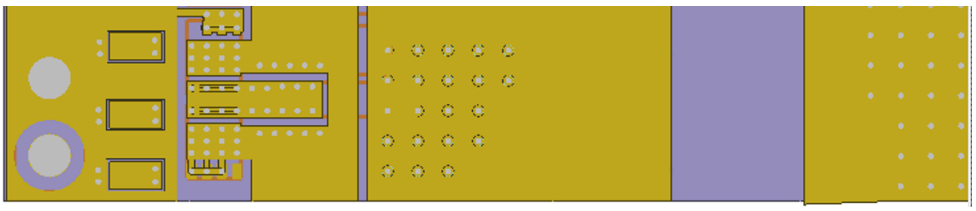


Fig. 3: PCI do conversor Buck utilizada na ferramenta computacional Ansys Q3D para extração dos elementos parasitas.

Mas é importante salientar que esta análise foi realizada somente para os elementos parasitas intrínsecos a PCI. Visto que os elementos intrínsecos aos componentes são fornecidos nas folhas de dados dos mesmos.

3 Resultados e discussão

Para avaliar a modelagem de elementos parasitas e o impacto destas no desempenho do conversor, o conversor CC-CC Buck foi simulado no software PSIM assumindo os componentes passivos e a PCI como ideais. Os parâmetros considerados para o conversor são descritos na Tabela 1.

Tabela 1: Especificações do conversor considerado

Parâmetro	**Valor**
Potência	200 W
Tensão de entrada	13.5 V
Tensão de saída	10.5 V
Razão cíclica ideal	0.778
Frequência de Chaveamento	500 kHz
L	1 uH
C_{in}	2x 82 uF + 1x 220 uF
C_o	1x 33 uF
MOSFETs	BUK7S2R5-40H

A Fig. 4 mostra as curvas de tensão e potência de saída, indicando que o conversor projetado atende a tensão e a potência média de saída desejada. Já a Fig. 5 mostra

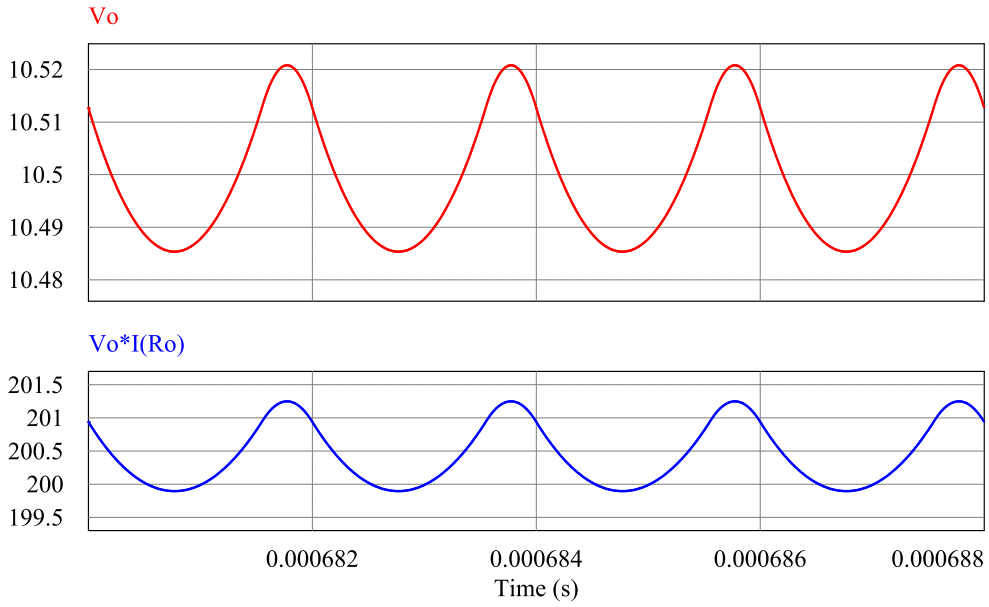


Fig. 4: Formas de onda em regime permanente da tensão de saída (Vo) e a potência de saída (Vo*I(Ro)).

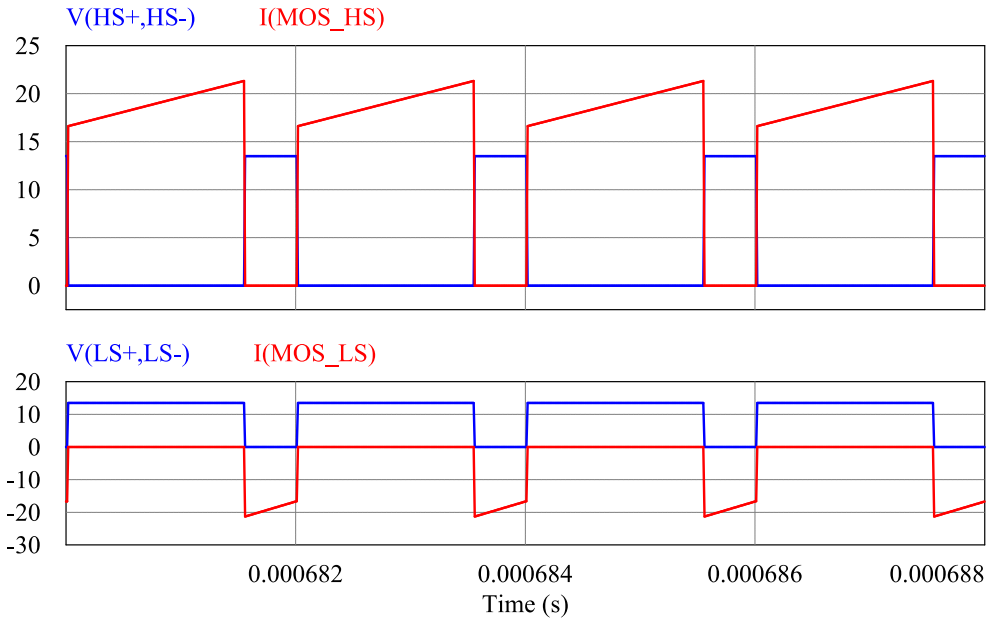


Fig. 5: Formas de onda em regime permanente da tensão (V(HS+,HS-)) e corrente I(MOS HS) da chave *high-side*, acima, e tensão (V(LS+,LS-)) e corrente I(MOS LS) da chave *low-side*, abaixo.

as curvas de tensão e corrente nas chaves *high-side* e *low-side*. Destaca-se que, devido a ausência de elementos parasitas, as comutações ocorrem sem resposta oscilatória.

Após validar o conversor ideal, o impacto dos elementos parasitas no conversor Buck foi analisado. Para isso, realizou-se a simulação computacional no software LTspice com as não-idealidades dos capacitores de entrada e saída, do indutor de potência, das trilhas

da PCI e do MOSFET, conforme a Fig. 2. Para este último, um modelo SPICE foi obtido utilizando as informações obtidas pelo fabricante [9] e o método de modelagem de [3, 8]. A Tabela 2 mostra os valores de todos os elementos parasitas modelados conforme método descrito na Seção 2.

Tabela 2: Valores dos elementos parasitas

Parâmetro	Valor	Parâmetro	Valor
RL_{PA}	$149\,\mu\Omega,\ 0,5\,nH$	ESR_{CAin}	$70\,m\Omega$
RL_{PB}	$149\,\mu\Omega,\ 0,5\,nH$	ESR_{CBin}	$70\,m\Omega$
RL_{PC}	$149\,\mu\Omega,\ 0,5\,nH$	ESR_{CCin}	$14\,m\Omega$
RL_{P1}	$390\,\mu\Omega,\ 0,9\,nH$	ESR_{Co}	$25\,m\Omega$
RL_{P2}	$31\,\mu\Omega,\ 35\,pH$	DCR_L	$4,9\,m\Omega$
RL_{P3}	$290\,\mu\Omega,\ 0,75\,nH$	R_D	$1\,n\Omega$
RL_{P4}	$626\,\mu\Omega,\ 326\,pH$	R_G	$76\,m\Omega$
RL_{P5}	$705\,\mu\Omega,\ 3,34\,nH$	R_S	$3\,u\Omega$
RL_{P6}	$303\,\mu\Omega,\ 6,9\,nH$	C_{DS}	$602\,nF$
RL_{P7}	$615\,\mu\Omega,\ 13,2\,nH$	C_{GS}	$2582\,nF$
		C_{GD}	$127\,pF$

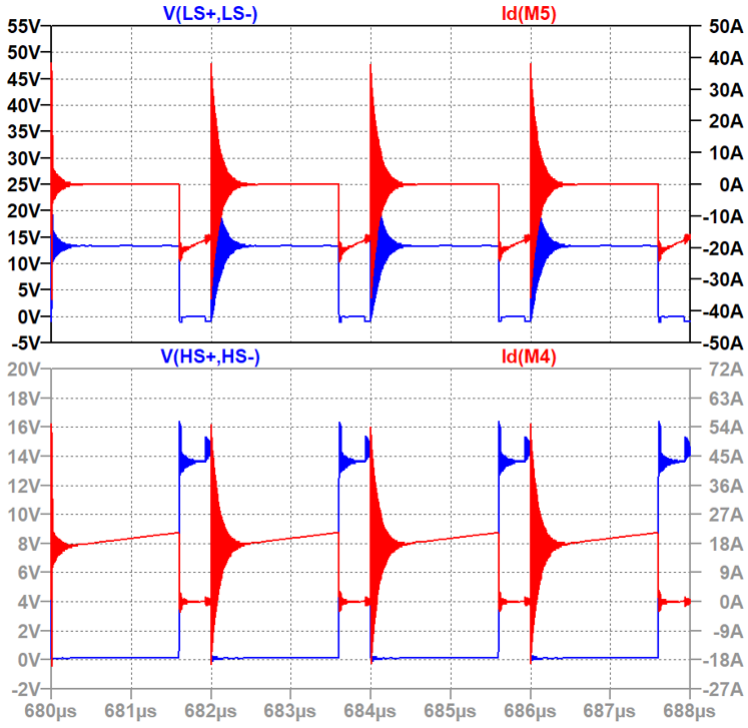


Fig. 6: Formas de onda em regime permanente da tensão (V(LS+,LS-)) e corrente Id(M5) da chave *low-side*, acima, e da tensão (V(HS+,HS-)) e corrente Id(M4) da chave *high-side*, abaixo.

O principal impacto negativo dos elementos parasitas no desempenho do conversor, ocorre durante a comutação das chaves semicondutoras. Por operarem com elevada taxa

de variação nos valores de tensão e de corrente elétrica, e devido aos elementos parasitas possuírem tipicamente valores pequenos, as comutações causam oscilações entre os elementos reativos [5] em frequências tipicamente na ordem de dezenas a centenas de MHz, como pode ser visto na Fig. 6 e, em detalhe ampliado, na Fig. 7. Tais oscilações distorcem os sinais de tensão e corrente, diminuem a eficiência do conversor, e causam ruídos indesejados de modo comum e modo diferencial.

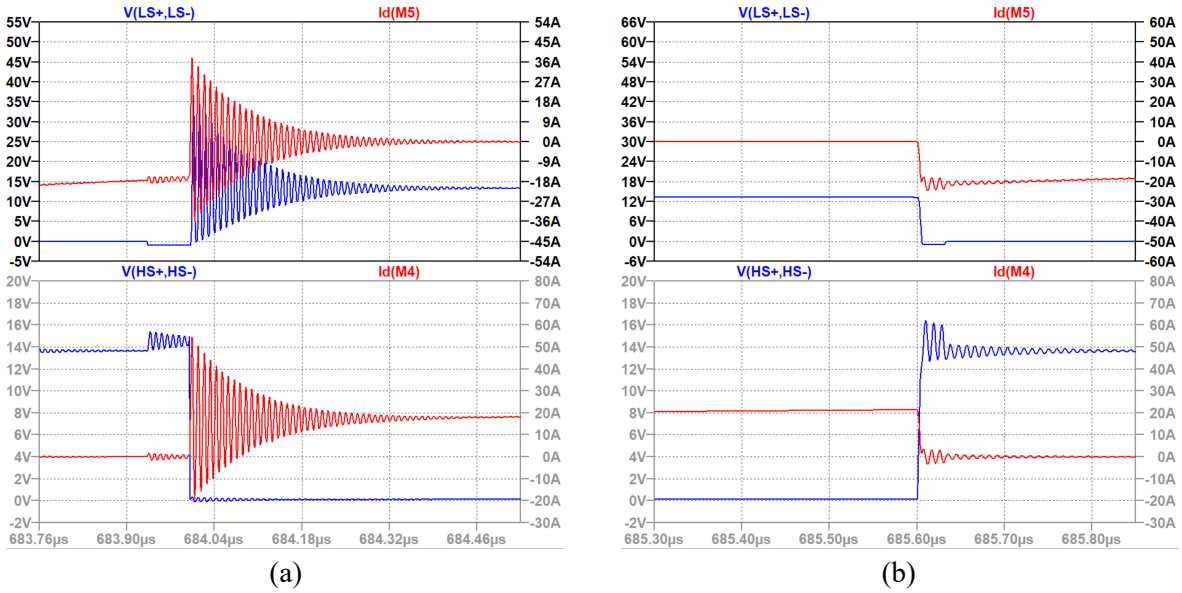


Fig. 7: Detalhe ampliado das formas de onda da tensão (V(HS+,HS-)) e corrente Id(M4) da chave *high-side* e da tensão (V(LS+,LS-)) e corrente Id(M5) da chave *low-side* durante (a) *turn-on* e (b) *turn-off* da chave principal.

4 Conclusões

Neste trabalho foi apresentada uma metodologia para a obtenção dos valores dos elementos parasitas da PCI de um conversor CC-CC Buck. A obtenção destes utilizando a ferramenta Ansys Q3D Extractor do software Ansys Electronics foi considerada satisfatória, visto que o software é amplamente utilizado para tal função.

A simulação do conversor CC-CC Buck sem e com os elementos parasitas deixou evidente as diferenças que estes elementos causam em um conversor CC-CC Buck. Foi possível observar, que são adicionados grandes níveis de oscilações na resposta do circuito, especialmente nos momentos em que ocorre a comutação da chave. É importante salientar que estas oscilações com grandes taxas de variação, podem afetar negativamente os níveis de EMI do conversor, o que pode atrasar o tempo de inserção de um produto no mercado, visto que existem normas que estabelecem limites para este tipo de interferência.

5 *Agradecimentos*

Os autores agradecem às empresas ESSS e Ansys pelo fornecimento de licenças temporárias do software Ansys Electronics 2021 R2.

Referências

[1] Ansys. Ansys q3d extractor: multiphysics parasitic extraction & analysis, 2021.

[2] L. R. Ávila. Simulação de emissões radiadas e conduzidas de placas de circuito impresso. 2019.

[3] S. Bramble. Ltspice tutorial: Part 6, 2021.

[4] D. W. Hart. *Eletrônica de potência: análise e projetos de circuitos*. McGraw Hill Brasil, 2016.

[5] T. Instruments. Minimizing switching ringing at tps53355 and tps53353 family devices. Technical report, 2018.

[6] A. Kingatua. Determining the equivalent series resistance (esr) of capacitors, 2017.

[7] A. Llamazares, M. García-Gracia, and S. Martín-Arroyo. Characterization of parasitic impedance in pcb using a flexible test probe based on a curve-fitting method. *IEEE Access*, 9:40695–40705, 2021.

[8] S. T. Microelectronics. Spice model tutorial for power mosfets - user manual. Technical report, 2013.

[9] Nexperia. Buk7s2r5-40h - product data sheet, 2021.

[10] T. Piovesan, H. C. Sartori, V. C. Bender, and J. R. Pinheiro. Método para quantificação de perdas em semicondutores aplicados a conversores estáticos devido aos elementos parasitas da placa de circuito impresso. *SOBRAEP*, 26(1):42–52, 2021.

[11] F. Sbrogio. Caracterização de parâmetros de indutores e capacitores aplicados ao modelamento de resposta em frequência de fontes chaveadas. Trabalho de conclusão de curso de engenharia elétrica, Universidade Estadual de Londrina, Londrina, SC, 208.

[12] L. C. M. Schlichting. Contribuição ao estudo da compatibilidade eletromagnética aplicada aos conversores estáticos. 2003.

[13] N. K. Trung, T. Ogata, S. Tanaka, and K. Akatsu. Attenuate influence of parasitic elements in 13.56-mhz inverter for wireless power transfer systems. *IEEE Transactions on Power Electronics*, 33(4):3218–3231, 2018.

[14] Y. Zeng, Y. Yi, and P. Liu. An improved investigation into the effects of the temperature-dependent parasitic elements on the losses of sic mosfets. *Applied Sciences*, 10(20), 2020.

Análise de Métodos de Aprendizado de Máquina para Classificação de Imagens com Poucos Dados

Ítalo M. Félix Santos[1], Mariza Ferro[1], Gilson A. Giraldi[1] e Paulo Sérgio Rodrigues[2]

[1] *Laboratório Nacional de Computação Científica, Petrópolis/RJ, Brasil*
[2] *Centro universitário FEI, São Bernardo do Campo/SP, Brasil*

Resumo

Neste artigo é feita uma análise comparativa entre os métodos clássicos de Aprendizado de Máquina (AM), como a Floresta Aleatória e a Máquina de Vetores de Suporte baseada em kernel e o recente método das Redes Neurais Convolucionais em um problema de classificação de imagens. Além disso, o método de análise das componentes principais baseado em kernel é utilizado como pré-processamento das imagens de modo a extrair um vetor de características para o uso nos métodos de AM. As Redes Neurais Convolucionais são o estado arte na classificação de imagens, entretanto é requerida uma quantidade grande de dados para se obter boa generalização do algoritmo. Neste sentido, foi escolhido um cenário onde existem poucos dados para efetuar o treinamento dos métodos e foi verificado qual o mais adequado nesta situação.

palavras-chave: Análise de componentes principais, Máquina de vetores de suporte, Florestas aleatórias, Redes Neurais Convolucionais, Aprendizado de máquina

1 INTRODUÇÃO

Na última década, as redes neurais convolucionais (CNN)[1] têm sido constantemente utilizadas na área de processamento de imagens, seja em classificação/segmentação, extração de características, super-resolução, redução de dimensionalidade, além de servirem como a base para a construção de redes mais complexas [18]. Um dos motivos se deve ao avanço tecnológico e o sucesso de arquiteturas de redes como ResNet [11], SeNet [12] e VggNet [24] no desafio ImageNet [21]. Apesar das conquistas das redes neurais, estes métodos necessitam de grande volume de dados para produzir bons resultados e este problema se

Contato: Italo, italo.messias.felix@gmail.com
[1]do inglês Convolutional Neural Networks

torna mais crítico quando estes dados são de alta dimensão, como no caso das aplicações em imagens devido ao risco de sobreajuste do modelo aos dados de treinando. Nestes casos, é necessário reduzir a complexidade dos modelos e utilizar técnicas adicionais para aumentar o seu poder preditivo [28, 16].

Em particular, para problemas de classificação, existem algoritmos muito utilizados, como as florestas Aleatórias (RF)[2] [4] e as Máquinas de Vetores de Suporte (SVM)[2] [26]. Entretanto, estes algoritmos necessitam de dados estruturados, isto é, um espaço/vetor de características disposto em linhas ou colunas. Em vista disso, em processamento de imagens é usual efetuar uma extração de características utilizando algoritmos como a Análise de Componentes Principais (PCA)[2] [8], a Análise de Discriminantes Lineares (LDA)[2] [10], dentre outros. Estes algoritmos permitem uma redução de dimensionalidade eliminando características redundantes nas imagens de uma base de dados, consequentemente, facilitando a classificação de imagens com metodologias semelhantes às RFs e ao SVM.

Baseado na desvantagem das CNNs em um cenário com poucos dados, este trabalho visa uma análise comparativa entre uma CNN simples, uma RF e uma variação do SVM baseado em Kernel [26, 23] em uma base de dados com poucas imagens.

O presente texto está organizado como se segue: na Seção 2 é apresentada uma breve descrição da base de dados e as métricas de avaliação utilizadas, e uma descrição geral do Kernel PCA (KPCA); na Seção 3 são descritos brevemente os modelos que serão comparados no presente artigo, isto é, a RF, o Kernel SVM (KSVM) e a CNN; posteriormente na Seção 4 são apresentados os resultados de cada modelo separadamente; por fim, são expostas as considerações e conclusões na Seção 5.

1.1 Trabalhos relacionados

No trabalho [27], foi efetuada uma comparação entre RF, KSVM e uma rede convolucional profunda em um problema de classificação multi-classe. Em geral, foi descoberto que o RF se sobressai em dados tabulares e não-estruturados (imagem e áudio) com tamanhos de amostra pequenos, enquanto a CNN tem um desempenho melhor em dados não-estruturados com tamanhos de amostra maiores. Entretanto, os testes foram realizados em uma base de dados grande (10000 imagens), mas de baixa resolução (32,32). Além disso, não foram utilizadas técnicas de extração de características ou redução de dimensionalidade para aplicação da RF e do KSVM.

No artigo [2], é efetuado uma comparação entre uma RF e um KSVM para classificação de câncer de mama como benignos e malignos. Diferente do presente trabalho, os autores efetuaram os experimentos em uma base de dados com 114 observações cada uma delas compostas por nove características fisiológicas.

No trabalho [22], foi feita uma classificação multi-classe de tumor cerebral em imagens neurológicas usando uma RF como modelo classificador. Posteriormente, o PCA é usado para redução de dimensionalidade e os experimentos mostraram que a combinação PCA-RF apresentou um desempenho superior em predições.

Por fim, no artigo [13], foi realizado um acoplamento PCA-SVM para reconhecimento digital de imagens médicas de modo a aprimorar a eficiência do algoritmo.

[2]Siglas do inglês Random Forest, Support Vector Machines, Principal Component Analysis e Linear Discriminant Analysis

No presente trabalho será realizada uma comparação entre os modelos RF, KSVM e CNN como é feito em [27]. Entretanto, diferente dos trabalhos [2] e [27] a comparação será realizada em uma base de imagens e com uma pequena quantidade de dados, respectivamente. Além disso, é feito um acoplamento KPCA-RF e KPCA-KSVM objetivando aprimorar os resultados dos métodos clássicos, assim como nos trabalhos [22] e [13].

2 DESCRIÇÃO DOS DADOS

Um problema comum relacionado a bases de dados de imagens médicas é a pouca quantidade de imagens e/ou desbalanceamento de classes. Sendo assim é adequado utilizar uma base de imagens com esse cenário e padrões para avaliar os métodos RF, KSVM e CNN. A base de dados escolhida foi fornecida pelo Dr. Paulo Sérgio do Centro Universitário FEI - Campus SBC (disponível em [20]) contendo 250 imagens de tomografia de câncer de mama (quatro imagens por paciente) em tons de cinza e com resolução (128,128). Além disso, esta base de dados possui 100 imagens de tumores benignos e 150 de tumores malignos. Nas Figuras 1 (a) - (d) ilustramos quatro imagens da classe benigno e nas Figuras 1 (e) - (h) quatro imagens da classe maligno, sendo todas de pacientes distintos.

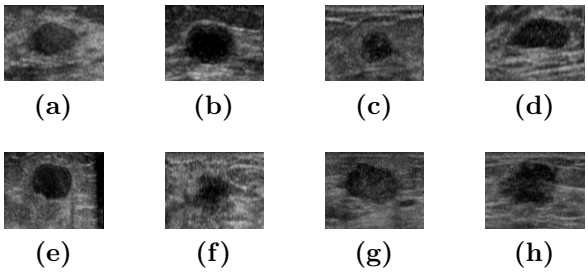


Fig. 1: (a)-(d) Amostras de imagens da classe benigno e (e)-(h) de imagens da classe maligno.

O número ideal de dados ainda é uma medida voltada para a combinação do domínio, tipos de dados sendo utilizados e modelo de aprendizado, não tendo sido possível encontrar na literatura um valor que defina exatamente o que são poucos ou muitos dados. Normalmente só é possível definir o número ideal de exemplos após etapas de treinamento, ajustes do modelo e a verificação de sobreajuste ou subajuste do modelo aos dados. Para este trabalho, dada a complexidade do domínio e o treinamento realizado para o modelo o número de exemplos é considerado pouco.

Conforme a distribuição de classes e o total de imagens do banco de dados, o erro majoritário para esta base de dados é $Erro\ majoritário = 1 - 0.6 = 0.4$. Portanto, o erro do classificador não deve ficar abaixo de 40%, pois seria a mesma acurácia obtida com um classificador que simplesmente seleciona todo exemplo como maligno, isto é, uma acurácia de 60%. Além disso, é importante considerar que os exemplos classificados como falsos positivos (resultados malignos classificados erroneamente como benignos) são mais prejudiciais que os falsos negativos (resultados benignos classificados erroneamente como malignos) do ponto de vista médico [5]. Em vista disso, será utilizada a área da curva ROC (AUC)[3] [7] além da acurácia para avaliar os modelos descritos na Seção 3.

[3]do inglês Area Under the ROC Curve

2.1 Pré-processamento de dados com KPCA

O KPCA é uma Análise de Componentes Principais não-linear, realizada em um espaço de dimensionalidade limitada pelo número de dados, o qual é mapeado implicitamente de forma não linear através da função Kernel. [8, 23]. O KPCA é muito utilizado para redução de dimensionalidade quando a quantidade M de dados é muito menor que a dimensão N do espaço. Em resumo, o KPCA transforma dados $\mathbf{x}$ de dimensão N em um vetor de características $\hat{\mathbf{x}}$ de dimensão M sem perda de informações na base de dados em que o algoritmo foi treinado. O espaço de características $\hat{\mathbf{x}}$ é ordenado em relação à relevância das características [8, 19], o que permite selecionar uma quantidade específica das primeiras M características.

Em vista disso, o KPCA será utilizado para extrair um espaço de características mais adequado e simples para aplicação da RF e o KSVM. Entretanto, para aplicação do KPCA em uma base de imagens é necessário converter a imagem g com resolução (n,m) em um vetor com dimensão (n*m), usualmente realizada através do mapeamento $g(i,j) \equiv g(i + j * y)$, com $i = 1, \ldots, n$ e $j = 1, \ldots, m$.

O tipo de kernel utilizado no KPCA influencia consideravelmente no espaço de características que será obtido, logo afeta as soluções da RF e o KSVM, uma vez, que os treinaremos com o espaço obtido pelo KPCA. Neste sentido, iremos utilizar os três kernels mais populares na literatura na avaliação dos modelos de classificação presentes na seção 3: o polinomial, a função de base radial (RBF) e a função sigmoide [23, 19].

3 MODELOS

Esta Seção visa contextualizar de maneira sucinta os modelos que serão explorados neste trabalho, deste modo: na Subseção 3.1 são abordadas as Rfs e suas vantagens mediante aos modelos de árvores tradicionais; na Subseção 3.2, descrevemos o KSVM e sua principal diferença entre o modelo SVM tradicional; na Subseção 3.3, introduzimos as CNN e as camadas que serão utilizadas para o modelo específico adotado por este trabalho.

3.1 Floresta Aleatória

A Floresta Aleatória é um algoritmo de *ensemble* do tipo *Bagging* [4], o qual reduz a variância das previsões combinando os resultados de várias Árvores de Decisão (AD) modeladas em diferentes subamostras da mesmas bases de dados. Ou seja, a RF pode ser descrita como um modelo formado por um conjunto de AD $h(X, y)$, onde y são vetores aleatórios amostrados de forma independentes, distribuídos igualmente em todas as árvores da floresta. O resultado é uma classe X com o maior número de votos entre todas as árvores consideradas [14]. As AD frequentemente se sobreajustam aos dados (overfitting) se nenhum tipo de controle for utilizado. Por outro lado, as RF resolvem parte deste problema se há árvores suficientes na floresta, isto é, o número de estimadores [4]. Entretanto, é possível impor um controle sobre o número mínimo de amostras que cada árvore da floresta aleatória deve caracterizar durante o treinamento. Portanto, o número mínimo de amostras por nós será variado durante os experimentos objetivando obter a RF com maior poder preditivo.

3.2 Máquina de Vetores de Suporte baseada em Kernel

Diferente do SVM que separa as observações positivas e negativas com a margem máxima através de um hiperplano (linear), o KSVM cumpri o mesmo objetivo procurando por uma hiperfície de separação com margem máxima representada pela expressão (1):

$$\sum_{i=1}^{M} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + \beta = 0, \tag{1}$$

sendo k uma função Kernel [23], M o número de dados, $\mathbf{x}_i$ os dados de entrada, y_i os resultados desejados para cada dado $\mathbf{x}_i$, e α_i e β valores reais a se determinar [26]. Segundo a Expressão 1 a hiperfície de separação está intrinsecamente relacionada a função Kernel k escolhida. Neste sentido, iremos testar três tipos de Kernel para o KSVM (Seção 2.1).

3.3 Redes Neurais Convolucionais

Uma Rede Neural Convolucional é um algoritmo de aprendizado que pode aprender filtros que conseguem extrair mapas de características apropriados para uma determinada tarefa de processamento de imagens [17]. Sua principal inspiração são os filtros clássicos que eram desenvolvidos manualmente pelos cientistas de processamento de imagens [9].

Redes neurais, em geral, necessitam de muitos dados para produzir bons resultados principalmente quando esses dados são de alta dimensão como as imagens. Na ausência de grandes bases de dados [20] não é aconselhado utilizar uma CNN muito complexa, pois pode gerar um *overfitting* da rede convolucional. Desta forma, serão utilizados uma CNN com camadas de convoluções simples [17], normalização de *batch* [15], camadas totalmente conectadas ao final da rede [3] e uma função sigmoide como ativação, dado que desejamos classificar apenas duas classes [23]. Além disso, será adicionada uma camada Dropout [25] com probabilidade 0.5 na penúltima camada da rede para evitar o *overfitting* do modelo. Neste artigo uma camada de convolução simples terá a estrutura dada pelo diagrama 2, o qual foi escrito com a gramática de uma biblioteca de código aberto utilizada para criar e desenvolver modelos de aprendizagem de máquina, o tensorflow [1].

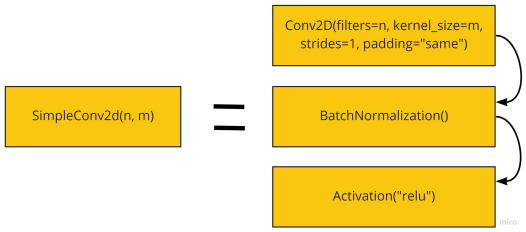


Fig. 2: Estrutura de uma convolução simples (Fonte: Elaborada pelo autor). [4]

4 RESULTADOS E EXPERIMENTOS

Para verificar qual o melhor modelo para classificação das imagens presente na base de dados descrita na Seção 2 foi utilizada a validação de amostragem de *bootstrap* [6]. Este método de validação consiste em treinar cada modelo um número K de vezes com conjuntos de treino e testes selecionados aleatoriamente. Ao fim do processo é considerado

[4]Descrito em: `https://www.tensorflow.org/api_docs/python/tf/keras/layers/`

como resultado a média das métricas resultantes de cada um dos treinos realizados (em nosso caso, a acurácia e AUC sobre os conjuntos de treino e teste de cada interação).

A base de dados (Seção 2) é composta de 250 imagens a qual foram separadas 60% para treino e 40% para testes em cada iteração da validação por *bootstrap*. Deste modo, em cada iteração, são utilizadas 150 imagens para a construção dos KPCAs, isto é, cada dado com dimensão 128*128 é convertido em um vetor com 150 características que serão utilizados para treinar as RFs e os KSVMs. A utilização das 50 primeiras características somente (das 149 fornecidas pelo KCPA), não afetaram significativamente a acurácia e AUC das classificações realizadas nos conjuntos de treino e teste. Portanto, os treinamentos das RFs e os KSVMs utilizaram um vetor com 50 características.

Nas seções seguintes são apresentados os resultados de cada método tendo como base as acurácias e AUCs médias fornecidas pela validação por *bootstrap* com $K = 20$.

4.1 Experimentos com florestas aleatórias

O algoritmo RF utilizado neste artigo foi implementado na biblioteca scikit-learn*sklearn* [19], onde os parâmetros utilizados para os diferentes modelos de RF e kernels do KPCA estão presentes na Tabela 1 (os demais parâmetros foram o valor padrão). Os resultados referentes as RFs testadas via *bootstrap* estão presentes nas Tabelas 2 e 3, onde variamos o número mínimo de amostras por folha como: sem controle (isto é, podem existir nós cobrindo uma única amostra), 5% da quantidade de dados da classe minoritária (número mínimo 5) e 10% dos dados da classe minoritária (número mínimo 10).

Parâmetros	**Espaço de escolhas**		
Kernels do KPCA	RBF	Polinomial	Sigmoide
Critério de ganho	"gini"		
Número mínimo de amostras para folha	1	5	10
Número de estimadores	100		

Tabela 1: Espaço de parâmetros escolhidos para a RF e para o KCPA.

RFs	**KPCA (Kernel)**	**Mínimo de amostras por folha**	**Acurácia média e desvio padrão (Conjunto de treino)**	**Acurácia média e desvio padrão (Conjunto de teste)**
Modelo 1	RBF	1	100.00% ±0.0000%	99.40% ±1.3191%
Modelo 2	RBF	5	100.00% ±0.0000%	97.80% ±2.6944%
Modelo 3	RBF	10	99.73% ±0.3887%	96.95% ±2.8013%
Modelo 4	Polinomial	1	100.00% ±0.0000%	98.95% ±1.9098%
Modelo 5	Polinomial	5	99.93% ±0.2000%	97.40% ±2.6153%
Modelo 6	Polinomial	10	99.70% ±0.5764%	97.05% ±3.2323%
Modelo 7	Sigmoide	1	100.00% ±0.0000%	99.35% ±1.2757%
Modelo 8	Sigmoide	5	100.00% ±0.0000%	98.55% ±1.7168%
Modelo 9	Sigmoide	10	99.43% ±0.6064%	95.60% ±2.9394%

Tabela 2: Acurácias dos modelos de RF avaliados com validação por *bootstrap*.

Os resultados presentes nas Tabelas 2 e 3 indicam que a melhor RF na classificação das imagens da base de dados [20] foi o modelo 1, visto que produziu maior acurácia e AUC médias em ambos os conjuntos de treino e teste com o menor desvio padrão dentre todos os modelos testados. Note que o modelo 1 utilizou kernel RBF para o KPCA e não

RFs	KPCA (Kernel	Mínimo de amostras por folha	AUC média e desvio padrão (Conjunto de treino)	AUC média e desvio padrão (Conjunto de teste)
Modelo 1	RBF	1	100.00 ±0.0000	0.9933 ±0.0154
Modelo 2	RBF	5	1.0000 ±0.0000	0.9758 ±0.0300
Modelo 3	RBF	10	0.9968 ±0.0047	0.9653 ±0.0316
Modelo 4	Polinomial	1	1.0000 ±0.0000	0.9897 ±0.0180
Modelo 5	Polinomial	5	0.9993 ±0.0022	0.9734 ±0.0302
Modelo 6	Polinomial	10	0.9965 ±0.0064	0.9664 ±0.0348
Modelo 7	Sigmoide	1	1.0000 ±0.0000	0.9925 ±0.0156
Modelo 8	Sigmoide	5	1.0000 ±0.0000	0.9833 ±0.0199
Modelo 9	Sigmoide	10	0.9932 ±0.0076	0.9485 ±0.0353

Tabela 3: Resultados referentes as AUCs dos modelos de RF avaliados com técnica de validação por *bootstrap*.

utilizou restrições para o número de mínimo de amostras por folha para a RF, entretanto as predições efetuadas no conjunto teste indicam que não houve *overfitting*.

4.2 Experimentos com a Máquina de vetores de suporte baseada em Kernel

O algoritmo KSVM utilizado neste artigo foi implementado na biblioteca scikit-learn [19], onde os kernels utilizados na construção dos modelos e do KPCA estão presentes na Tabela 4. Os resultados com KSVM via *Bootstrap* estão expostos nas Tabelas 5 e 6.

Parâmetros	**Espaço de escolhas**		
Kernels do KPCA	RBF	Polinomial	Sigmoide
Kernels do KSVM	RBF	Polinomial	Sigmoide

Tabela 4: Espaço de parâmetros escolhidos para o KSVM e para o KCPA.

KSVMs	KPCA (Kernel)	KSVM (Kernel)	Acurácia média e desvio padrão (Conjunto de treino)	Acurácia média e desvio padrão (Conjunto de teste)
Modelo 1	RBF	RBF	100.00% ±0.0000%	99.55% ±1.0712%
Modelo 2	RBF	Polinomial	99.90% ±0.2380%	98.20% ±2.0881%
Modelo 3	RBF	Sigmoide	94.40% ±1.5261%	88.15% ±3.9784%
Modelo 4	Polinomial	RBF	100.00% ±0.0000%	99.85% ±0.4770%
Modelo 5	Polinomial	Polinomial	99.90% ±0.2380%	99.10% ±2.0952%
Modelo 6	Polinomial	Sigmoide	94.17% ±1.9422%	88.25% ±4.1458%
Modelo 7	Sigmoide	RBF	100.00% ±0.0000%	99.55% ±1.2031%
Modelo 8	Sigmoide	Polinomial	99.90% ±0.2380%	98.20% ±2.5020%
Modelo 9	Sigmoide	Sigmoide	93.67% ±2.1029%	88.15% ±3.0541%

Tabela 5: Acurácias dos modelos de KSVM avaliados com a técnica de validação por *bootstrap*.

Conforme as Tabelas 5 e 6, o modelo 4 obteve a maior acurácia e AUC médias em ambos os conjuntos de treino e teste com o menor desvio padrão dentre todos os modelos experimentados. Sendo assim, o modelo 4 foi o melhor modelo encontrado para classificação das imagens da base [20].

KSVMs	KPCA (Kernel)	KSVM (Kernel)	AUC média e desvio padrão (Conjunto de treino)	AUC média e desvio padrão (Conjunto de teste)
Modelo 1	RBF	RBF	1.0000 ±0.0000	0.9946 ±0.0129
Modelo 2	RBF	Polinomial	0.9987 ±0.0032	0.9795 ±0.0233
Modelo 3	RBF	Sigmoide	0.9367 ±0.0176	0.8741 ±0.0387
Modelo 4	Polinomial	RBF	1.0000 ±0.0000	0.9981 ±0.0061
Modelo 5	Polinomial	Polinomial	0.9988 ±0.0030	0.9887 ±0.0259
Modelo 6	Polinomial	Sigmoide	0.9340 ±0.0205	0.8719 ±0.0393
Modelo 7	Sigmoide	RBF	1.0000 ±0.0000	0.9941 ±0.0163
Modelo 8	Sigmoide	Polinomial	0.9987 ±0.0032	0.9795 ±0.0278
Modelo 9	Sigmoide	Sigmoide	0.9288 ±0.0258	0.8743 ±0.0341

Tabela 6: Resultados referentes as AUCs dos modelos de KSVM avaliados com a técnica de validação por *bootstrap*.

4.3 Experimentos com a rede neural convolucional

A CNN utilizada foi implementada com auxílio da biblioteca Tensorflow [1], seus hiperparâmetros, parâmetros de treinamento, e estrutura foram selecionados de forma empírica objetivando aumentar o poder preditivo do modelo. Para o treinamento da rede foram utilizados taxa de aprendizado 0.0001, otimizador Adam com momentum padrão, tamanho de batch 32, número máximo de épocas 300 e a entropia cruzada binária como função de custo. A CNN escolhida contém um total de 24929 parâmetros, sendo 24689 treináveis e 240 não treináveis. Ademais, sua estrutura pode ser visualizada no diagrama 3 escrito com a gramática do Tensorflow [1].

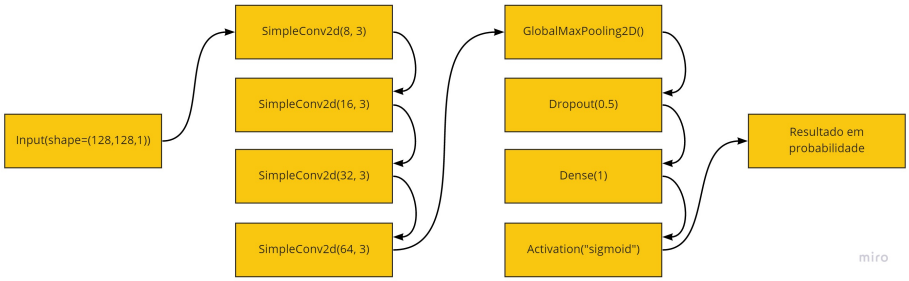


Fig. 3: Estrutura da CNN utilizada nos experimentos (Fonte: Elaborada pelo autor). [5]

Os resultados referentes a CNN via *bootstrap* estão presentes na Tabela 7. Além disso, foi selecionado um dos 20 treinamentos realizados durante a validação por *bootstrap* de forma aleatória para avaliar a evolução do treinamento ao longo das épocas. O gráfico referente a esse treinamento pode ser visualizado na Figura 4, onde podemos perceber que não esta ocorrendo *overfitting* durante o treinamento do modelo. Entretanto, nota-se pela Figura 4 instabilidade na convergência da função de custo, acurácia e AUC do modelo no conjunto de treino e teste principalmente nas primeiras épocas. Este comportamento ocorre devido a rede totalmente conectada ser rasa e possuir uma camada do tipo *dropout* com probabilidade 0.5, além da baixa quantidade de dados utilizadas no treinamento.

[5]Descrito em: `https://www.tensorflow.org/api_docs/python/tf/keras/`

Acurácia média e desvio padrão (Conjunto de treino)	**Acurácia média e desvio padrão (Conjunto de teste)**
99.97% ±0.1453%	98.45% ±1.7457%
AUC média e desvio padrão (Conjunto de treino)	**AUC média e desvio padrão (Conjunto de teste)**
0.9600 ±0.0081	0.9601 ±0.0081

Tabela 7: Resultados da acurácia e AUC da CNN avaliadas com *bootstrap*.

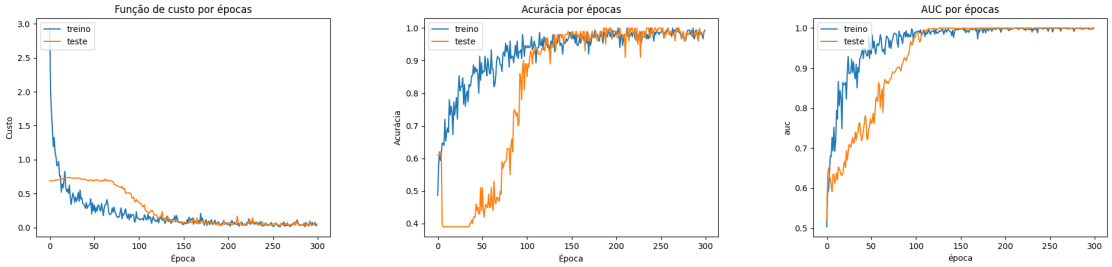


Fig. 4: Evolução das funções de perda, acurácia e AUC ao longo das 300 épocas.

5 CONCLUSÕES

No presente trabalho foram efetuados o treinamento e validação de diversos modelos de RF e KSVM objetivando comparar com os resultados de uma CNN na classificação de uma base de dados pequena de imagens. Os experimentos realizados na Seção 4 mostraram que os melhores modelos de RF e KSVM obtiveram acurácias e AUCs médias acima de 99% e 0.99, respectivamente, com desvios padrões baixos. Por outro lado, a melhor CNN dentre aquelas obtidas pela técnica *bootstrap* obteve acurácia e AUC médias dadas por 98.45% e 0.9601, respectivamente, sendo portante inferiores aos modelos tradicionais testados. Entretanto, não é possível afirmar que o mesmo irá ocorrer para quaisquer outras bases de dados de imagens, mas essas análises e experimentos podem ajudar e dar um caminho ao leitor na escolha de métodos para este tipo específico de classificação.

Apesar das conclusões aqui apresentadas, existem metologias para aumento de dados que podem ser aplicadas, permitindo aumentar a complexidade da CNN e consequentemente aumentar o seu poder preditivo. Além disso, metodologias de transferência de aprendizado podem ser utilizadas para aprimorar a eficiência da CNN. Abordagens para a CNN baseada nas metodologias aqui citadas e testes em várias bases de dados diferentes serão exploradas em trabalhos futuros.

6 *Agradecimentos*

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Software disponível em: `tensorflow.org`.

[2] C. Aroef, Y. Rivan, and Z. Rustam. Comparing random forest and support vector machines for

breast cancer classification. *TELKOMNIKA Telecommunication Computing Electronics and Control*, 18:815–821, 2020.

[3] D. Barber and T. Heskes. An introduction to neural networks. 2004.

[4] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2004.

[5] J. Brodersen and V. Siersma. Long-term psychosocial consequences of false-positive screening mammography. *The Annals of Family Medicine*, 11:106 – 115, 2013.

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[7] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[8] K. Fukunaga. Introduction to statistical pattern recognition-second edition. 1990.

[9] R. C. Gonzales and P. Wintz. Digital image processing (2nd ed.). Upper Saddle River, N.J., 1987.

[10] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. 2001.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2020.

[13] L. Hu and J. Cui. Digital image recognition based on fractional-order-pca-svm coupling algorithm. *Measurement*, 145:150–159, 2019.

[14] M. M. Ibañez, F. M. Ramos, and A. R. C. . Uso de redes neurais nebulosas e florestas aleatórias na classificação de imagens em um projeto de ciência cidadã. Master's thesis, 2016.

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, 2015.

[16] R. Keshari, M. Vatsa, R. Singh, and A. Noore. Learning structure and strength of CNN filters for small sample size training. *CoRR*, abs/1803.11405, 2018.

[17] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *ArXiv*, 2015.

[18] A. Patil and M. Rane. Convolutional neural networks: An overview and its applications in pattern recognition. In *Information and Communication Technology for Intelligent Systems*, pages 21–30, Singapore, 2021. Springer Singapore.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and et al. Scikit-learn: Machine learning in Python.

[20] P. S. Rodrigues. Breast cancer images (fei), 2021. Disponível em: `https://goo.gl/ld5dF0`.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, and et al.

[22] V. Saraswathi and D. Gupta. Classification of brain tumor using pca-rf in mr neurological images. *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pages 440–443, 2019.

[23] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[26] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[27] H. Xu, M. Ainsworth, Y.-C. Peng, M. Kusmanov, S. Panda, and J. Vogelstein. When are deep networks really better than random forests at small sample sizes? *ArXiv*, 2021.

[28] W. Zhao. Research on the deep learning of the small sample data based on transfer learning. volume 1864, page 020018, 08 2017.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Regressor de floresta aleatória para o cálculo de propriedades equivalentes em reservatórios de petróleo fraturados

Iury Coimbra[1], Eduardo Garcia[1], Tuane Lopes[1] e Eduardo Krempser[2]

[1] *Laboratório Nacional de Computação Científica (LNCC-MCT), Petrópolis/RJ, Brazil*
[2] *Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro/RJ, Brasil*

Resumo

Este estudo aplicou um algoritmo de aprendizado de máquina baseado no regressor de floresta aleatória para computar propriedades físicas equivalentes em reservatórios de petróleo naturalmente fraturados. As metodologias atuais disponíveis podem ser divididas em dois grupos distintos que apresentam um dilema entre a precisão da solução e o custo computacional. O método numérico NDFM (Non-conforming Discrete Fracture Model) possui alta precisão e elevado custo computacional, enquanto o método analítico proposto por Oda em 1985, apresenta baixo custo computacional, porém, superestima os valores calculados dependendo da disposição física das fraturas, chegando em até 55% acima do valor calculado pela metodologia numérica em casos particulares. Este estudo tem o objetivo de criar uma alternativa que possua baixo custo computacional e alta precisão, explorando as técnicas de aprendizado de máquina. A metodologia proposta utilizou para o treinamento do regressor, 1000 células de 10 reservatórios fraturados que foram gerados pelo programa de modelagem geológica Skua-Gocad. O treinamento supervisionado utilizou como dados de entrada informações geométricas das fraturas e como valores alvo (dados de saída) resultados de simulações numéricas via método NDFM. O estudo contou com busca de hiperparâmetros e validação cruzada. O modelo regressor de floresta aleatória foi desenvolvido com o uso das bibliotecas do Scikit-Learn.

Palavras-chave:Aprendizado de máquina, regressor de floresta aleatória, propriedades físicas equivalentes, reservatórios fraturados.

1 INTRODUÇÃO

Para a indústria do petróleo, o conhecimento sobre as redes de fraturas impacta diretamente no volume de óleo produzido, pois podem alterar as propriedades petrofísicas originais da rocha, como porosidade e permeabilidade [7, 3].

Contato: Iury Coimbra, coimbra@lncc.br

No último meio século, com o foco em desenvolver técnicas para lidar com a simulação de fluxo em rochas fraturadas, alguns pesquisadores ganharam grande relevância. Uma revisão mais detalhada pode destacar as contribuições de Barenblatt et al. [2] em 1960, a de Warren e Root [16] em 1963 e a de Karimi et al. [12] em 2004. Esses autores desenvolveram as principais soluções numéricas para realizar o *Flow based upscaling*, metodologia usada para obter a propriedade equivalente que será a base para a simulação de fluxo em uma escala acima (meso ou macro). Entretanto, essas técnicas fazem uso de malhas especiais, que discretizam as fraturas no domínio, alinhando elementos e nós das fraturas com elementos e nós da matriz (rocha). A geração dessas malhas demanda esforço e tempo, por se trata de um processo que carece de automatização.

Em 2020, Ziyao Xu e Yang Yang[17] propuseram uma nova metodologia chamada *Non-conforming Discrete Fracture Model* (NDFM), que utiliza malhas simples e de fácil geração sem comprometer a precisão do valor calculado. No entanto, o método NDFM ainda não pode ser implementado em simuladores de fluxo comerciais, pois na escala de simulação, o número de células pode chegar a centenas de milhares, sendo inviável a geração de malhas, mesmo não se tratando de malhas especiais. Para contornar esse impasse, as soluções analíticas são cada vez mais exploradas e aplicadas nos simuladores comerciais.

Em 1985, Oda [14] propôs uma solução analítica para o cômputo do tensor de permeabilidade física equivalente de um maciço rochoso (matriz + fraturas), que calculava a contribuição das redes de fraturas e somava essa contribuição à permeabilidade da rocha. A solução analítica de Oda tem boa capacidade de calcular as propriedades físicas equivalentes para um grande número de células em tempos computacionais aceitáveis. O preço exigido por se utilizar essa metodologia é aceitar que a solução de Oda superestima o valor calculado, dependendo de como as fraturas estão dispostas nas células.

Podemos então criar dois grupos e classificar as soluções numéricas como sendo altamente precisas, mas de alto custo, enquanto as soluções analíticas tem baixa precisão e baixo custo. Fazendo uso de técnicas de aprendizado de máquina, este estudo explora as vantagens das duas metodologias.

A regressão é uma tarefa básica do aprendizado de máquina, onde se destaca o algoritmo de floresta aleatória (RF - *Random Forest*) que faz parte da família dos algoritmos de comitê de classificação (*Ensemble Learning*) e é uma técnica avançada de árvore de decisão (DT - *Decision Tree*), podendo ser usada para classificação ou regressão [5] [11]. A RF surgiu para superar as deficiências da DT, em especial a sua fácil susceptibilidade de produzir modelos com sobreajuste [5]. Essencialmente, a solução consiste em criar uma diversidade de árvores de decisão, onde as propriedades das árvores (profundidade, número de árvores, número de divisões em cada nó) são inicializadas com valores aleatórios e convergem para um número ideal, diferente em cada árvore.

Hidayat e Astsauri (2021) [11] realizaram um estudo com o algoritmo de floresta aleatória (RF), para avaliar a correlação entre os atributos do processo de injeção de água de baixa salinidade, LSWI (*Low Salinity Water Injection*), com a recuperação avançada de petróleo. No estudo, foram utilizados dados de 1000 projetos experimentais contendo os parâmetros para LSWI como sódio, cálcio, magnésio, sulfato, cloreto, carbonato e bicarbonato. Com a RF, Hidayat e Astsauri chegaram à conclusão que apenas 3 parâmetros eram os principais responsáveis pelo volume recuperado, enquanto outros 10 parâmetros

abordados não apresentavam efeito significativo na recuperação.

Guo et al. (2021) [9] testaram 6 algoritmos de aprendizado supervisionado, entre eles, árvore de decisão (DT), floresta aleatória (RF) e k-vizinho mais próximo (KNN - *K-Nearest Neighbor*). O estudo envolveu a predição da saturação em reservatórios de petróleo com o objetivo de superar as limitações e necessidades especiais do método eletromagnético tradicional. Os ajustes durante o treinamento foram feitos seguindo a curva de aprendizado alinhada à busca em grade por hiperparâmetros e validação cruzada (GridSearchCV). O modelo desenvolvido consegue prever com precisão superior a 90% a saturação vertical de cada camada geológica do reservatório do bloco 2 do campo petrolífero de Gangdong (Huanghua, China).

Larestani et al. (2022) [13] realizaram um estudo detalhado para detecção de danos à formação durante a inundação de água em reservatórios de petróleo. A inundação com água é uma das principais opções empregadas pela indústria do petróleo na recuperação avançada de hidrocarbonetos, mas é altamente propensa a causar danos à formação, ativação de falhas e mudanças físico-químicas que podem comprometer completamente as regiões inundadas. O estudo proposto envolveu uma exaustiva combinação de técnicas de aprendizado de máquina como árvore de decisão otimizadas com aumento de gradiente (GBDT - *Gradient-Boosted Decision Trees*), rede de retropropagação em cascata (CFBPN - *Cascade-Forward Back-Propagation Network*) e redes neurais de regressão generalizada (GRNN - *Generalized Regression Neural Networks*). O desempenho foi superior para o modelo GBDT em comparação com os outros modelos desenvolvidos, onde o GBDT pode estimar mais de 90% dos pontos com erro médio absoluto inferior a 0,5%. A análise de tendência mostrou alta capacidade dos modelos em detectar danos à formação.

Apesar de focos de atuação diferentes, os citados trabalhos indicam a ampliação da aplicação de métodos de aprendizado de máquina na indústria de petróleo e gás.

O presente trabalho aplicou um algoritmo de aprendizado de máquina baseado no regressor de floresta aleatória com o objetivo de treinar um modelo para prever os valores de propriedades físicas equivalentes (K_{eq}). O foco é treinar um modelo com os resultados numéricos de simulações realizadas via metodologia NDFM, com os mesmos valores de entrada da solução analítica de Oda. Para isso, foi criado um banco de dados contendo informações de 1000 células fraturadas (amostras), onde cada amostra contém informações geométricas da rede de fraturas e o valor de K_{eq} previamente simulado via NDFM, isto é, o banco de dados tem as informações geométricas que a solução de Oda utiliza e também os valores de referência que foram numericamente calculados.

O conjunto de dados foi dividido em subconjuntos de treinamento e teste. Em seguida, uma pesquisa exaustiva sobre valores de hiperparâmetros junto com a técnica de validação cruzada foi realizada via pesquisa em grade (GridSearchCV) [15].

2 METODOLOGIA

O método NDFM (*Non-conforming Discrete Fracture Model*), proposto por Ziyao Xu e Yang Yang em 2020 [17], consiste em uma extensão ao método DFM (*Discrete Fracture Model*), que discretiza cada uma das fraturas do reservatório. A principal diferença entre o DFM e o NDFM é que este último utiliza malhas regulares e cartesianas, como apresentado na Figura 1. Para o DFM, os nós e elementos que representam as fraturas precisam estar alinhados aos nós e elementos da matriz, o que torna esse tipo de malha difícil de ser gerada

dependendo do número de fraturas presentes no domínio. O método NDFM descarta essa necessidade e agiliza a etapa de geração de malhas, resumindo o processo na criação de subdomínios.

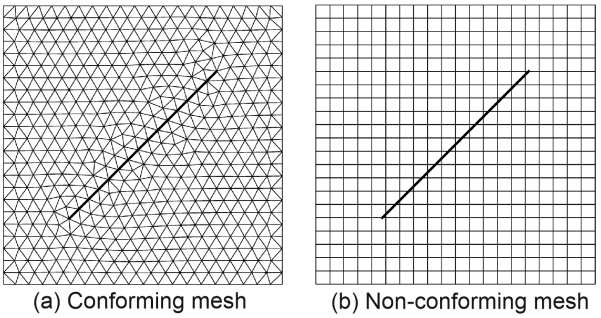


Fig. 1: Malha conforme (DFM) versus malha não-conforme (NDFM) - Fonte: [17].

Métodos numéricos apresentam uma maior acurácia dos valores calculados quando comparamos seus resultados com os de soluções analíticas [6]. A justificativa para se optar por uma solução menos precisa, como é o caso da metodologia de Oda, é o custo computacional envolvido nos cálculos. Métodos numéricos são computacionalmente caros e a etapa de geração carece de automatização nos cenários com grande número de fraturas, tornando seu uso em simuladores de fluxo inviável, levando os usuários a aceitarem menor precisão em troca de agilidade do cálculo.

A solução analítica proposta por Oda demanda informações relativas a abertura, comprimento, permeabilidade e orientação de cada fratura, assumindo impermeabilidade na direção ortogonal. Antes da permeabilidade equivalente ser calculada, computamos o tensor de fratura F_{ij} da seguinte forma

$$F_{ij} = \frac{1}{V} \sum_{k=1}^{N} A_k d_k K_k n_{ik} n_{jk}, \tag{1}$$

onde V é o volume da célula, N o número de fraturas no interior da célula, A_k a área, d_k a abertura, K_k a permeabilidade e n_{ik}, n_{jk} as componentes normais de cada fratura.

Após o cômputo do tensor de permeabilidade das fraturas F_{ij}, a permeabilidade equivalente do meio pode ser calculada da forma [8]

$$K_{M_{ij}}^{eq} = (F_{kk}\delta_{ij} - F_{ij}) + K \tag{2}$$

com K a permeabilidade da matriz (rocha), F_{kk} o traço de F_{ij} e δ_{ij} o delta de Kroenecker.

Na prática, a metodologia de Oda é eficiente para cômputo de propriedades equivalentes, porém, existem situações onde o cálculo de Oda superestima os resultados. A Figura 2 é um exemplo dessa situação. Fraturas desconectadas da rede (isoladas), são consideradas como conectadas por Oda, o que resulta em um valor acima do real. Onde pela Tabela 1 e possível observar que, ao assumir a solução numérica como valor de referência, a solução analítica de Oda superestimou a propriedade equivalente em aproximadamente 55%.

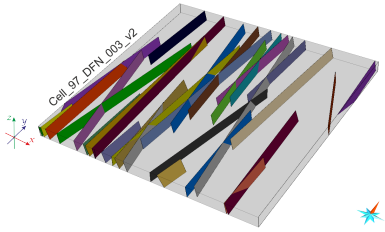


Fig. 2: Exemplo de célula com previsão Oda superestimada.

Tabela 1: Propriedade equivalente computada

K equivalente (mD)	
Oda	1048.11
NDFM	669.46

2.1 Construção do Banco de Dados

Para este estudo, 10 reservatórios fraturados foram criados através do programa de modelagem geológica SKUA-GOCAD [1], com a ferramenta de geração de rede de fraturas discretas, FracMV™. Os reservatórios possuem estrutura de grade no formato 10x10x1, isto é, 10 células no eixo x, 10 células no eixo y e 1 célula no eixo z (100 células por reservatório). Em todos os reservatórios, as células são ortogonais de dimensão 200x200x10m (comprimento, largura e altura), conforme a Figura 3.

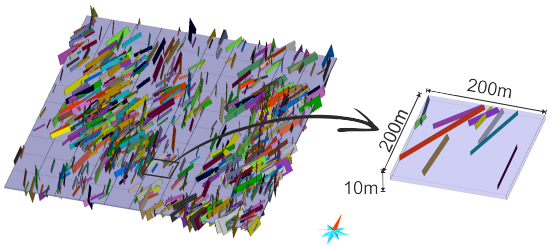


Fig. 3: Rede de fraturas discretas do reservatório 1.

A soma de todas as células dos 10 reservatórios resulta em um conjunto de dados com 1000 células (amostras), que formam o banco de dados que será utilizado nas etapas de treinamento e teste do regressor de floresta aleatória.

A Tabela 2 apresenta informações sobre o conjunto de dados. Os atributos são apresentados em relação à cada célula, contendo o número de fraturas (numFrat), a soma de áreas de fraturas (tAreaFrat), a soma da componente normal das direções X, Y e Z (normalX, normalY e normalZ respectivamente), a soma das aberturas de fraturas (tAbertura) e a soma das permeabilidades das fraturas (tPermFrat). O estudo considerou fluxo isotrópico, porosidade da rocha (0.0758) e permeabilidade da rocha (343.29mD) a mesma em todas as células, levando a descartar esses atributos.

As linhas **mean**, **min** e **max** são auto explicativas, o desvio padrão (**std**) mede a dispersão dos valores, as linhas **25%**, **50%** e **75%** mostram o número de pontos com o mesmo valor em determinada faixa de observação, por exemplo, 25% das células possuem 7 fraturas (numFrat), enquanto 75% dos valores da soma das aberturas é de 3.42E-03 (tAbertura).

2.1.1 Traço do tensor de fraturas (F_{kk})

A fórmula analítica de Oda descrita na Equação (2), consiste na soma entre o valor de permeabilidade da matriz (rocha), o traço do tensor das fraturas (F_{kk}) e o próprio tensor das fraturas(F_{ij}).

Tabela 2: Informações sobre os atributos do conjunto de dados.

	numFrat	**tArea**	**normalX**	**normalY**	**normalZ**	**tAbertura**	**tPermFrat**
mean	18.2	1.46E+04	1.66E+01	5.31E+00	3.27E+00	2.41E-03	8.94E-08
std	14.7	1.35E+04	1.36E+01	4.55E+00	2.69E+00	2.23E-03	9.54E-08
min	1	3.60E+01	6.87E-01	5.79E-03	1.81E-02	1.93E-07	3.07E-15
25%	7	4.68E+03	6.50E+00	2.13E+00	1.29E+00	7.42E-04	1.64E-08
50%	13.5	1.03E+04	1.23E+01	3.91E+00	2.44E+00	1.79E-03	6.17E-08
75%	25	2.04E+04	2.32E+01	7.38E+00	4.52E+00	3.42E-03	1.31E-07
max	88	8.01E+04	8.22E+01	2.55E+01	1.62E+01	1.43E-02	6.33E-07

Observando as Equações (1) e (2) juntamente com as informações da Tabela 2 e levando em consideração as simplificações consideradas neste estudo (fluxo isotrópico, porosidade e **permeabilidade da rocha** iguais em todas as células) é possível perceber que da equação de Oda o único atributo que não está presente no conjunto de dados é o traço do tensor de fraturas (F_{kk}), que pode ser facilmente calculado e inserido no conjunto de dados.

2.2 Busca por Hiperparâmetros e Validação Cruzada

Quatro baterias de tamanhos diferentes são utilizadas para verificar o desempenho com relação ao tamanho do conjunto, isto é, para verificar o quão dependente da quantidade de dados disponíveis é o sucesso do algoritmo. As baterias contêm 400, 600, 800 e 1000 amostras.

Em todas as baterias, o conjunto de amostras são divididos em subconjuntos de treinamento e teste, nas proporções 80% e 20% (respectivamente). A divisão é realizada de modo aleatório para auxiliar na prevenção de sobreajuste [10] e os subconjuntos gerados são disjuntos, ou seja, em nenhum momento do processo de treinamento o algoritmo tem acesso ao subconjunto de teste.

O algoritmo de floresta aleatória possui uma série de parâmetros que podem ser otimizados pela busca por hiperparâmetros [11], embora se admita sua utilização com valores *default* nas situações em que não há tempo ou competências suficientes disponíveis para essa tarefa, como é aconselhado por [4].

Uma pesquisa exaustiva sobre valores de parâmetros específicos foi realizada via pesquisa em grade (GridSearchCV), com o objetivo de otimizar os hiperparâmetros do algoritmo regressor de floresta aleatória.

A Tabela 3 descreve os valores mínimos e máximos dos hiperparâmetros que foram explorados para otimização do modelo de regressão. Ao todo são 300 candidatos a hiperparâmetros.

Tabela 3: Hiperparâmetros explorados no regressor de floresta aleatória.

Hyperparâmetro	código	min	max
Profundidade Máxima	max_depth	10	1000
Número de recursos (Max)	max_features	2	10
Número de Árvores (Max)	n_estimators	10	1000

Além da busca por hiperparâmetros em grade, a técnica de validação cruzada é realizada com o ajuste de 5 partes (*folds*) para as 300 combinações candidatas à hiper-

parâmetros, o que gerou um total de 1500 modelos de ajustes que foram aleatoriamente construídos e testados.

3 RESULTADOS E DISCUSSÕES

O aprendizado de máquina ganhou rápida popularidade na ciência de dados a medida que a tecnologia se desenvolvia, principalmente, devido à sua capacidade de facilmente lidar com grandes e complexos conjuntos de dados. A análise da correlação entre os atributos é uma etapa essencial para limpeza dos dados e identificação da existência de associação entre atributos. As métricas utilizadas para medir a eficiência do modelo proposto auxiliam na verificação da capacidade de generalização do regressor, uma vez que o modelo treinado não teve acesso ao conjunto separado para teste e não conhece os valores esperados para a solução. A seguir apresentamos a análise da dispersão e do erro realizadas neste estudo.

3.1 Análise da dispersão

Para verificar a correlação entre os vários atributos na busca de *insights* ou identificar a necessidade de manipulação dos dados (como remoção de atributos insignificantes), se conduziu uma busca por correlação entre os atributos via *Análise de Coeficiente de Correlação Padrão* ou **r** *de Pearson* [10]. A Figura 4 apresenta a correlação entre os atributos. Valores perto de 0% significam baixa correlação, valores próximos a -100% ou $+100\%$ siginificam alta correlação.

	numFrat	tArea	normalX	normalY	normalZ	tAbertura	tPermFrat	traco	K_eq
numFrat	100%	97%	100%	87%	98%	91%	80%	93%	57%
tArea	97%	100%	97%	82%	96%	89%	78%	92%	57%
normalX	100%	97%	100%	85%	98%	91%	80%	93%	55%
normalY	87%	82%	85%	100%	85%	79%	69%	81%	63%
normalZ	98%	96%	98%	85%	100%	89%	78%	92%	56%
tAbertura	91%	89%	91%	79%	89%	100%	96%	87%	69%
tPermFrat	80%	78%	80%	69%	78%	96%	100%	77%	75%
traco	93%	92%	93%	81%	92%	87%	77%	100%	54%
K_eq	57%	57%	55%	63%	56%	69%	75%	54%	100%

Fig. 4: Correlação entre atributos.

Embora o atributo K_eq apareça na Figura 4, ele não é um atributo de entrada para o regressor de floresta aleatória (RFR), ele é utilizado somente no treinamento como valor alvo. O objetivo de inserir o K_eq na busca por correlação, foi para confirmar que isoladamente nenhum atributo de entrada é capaz de sozinho descrever o comportamento da propriedade equivalente. Por sinal, todos os atributos de entrada da RFR tem baixa correlação com os valores de propriedade equivalente.

3.2 Análise do erro

Para avaliar o desempenho do algoritmo regressor de floresta aleatória, este estudo concentrou esforços na comparação dos resultados entre o modelo regressor de floresta aleatória (RFR) com os resultados dos cálculos analíticos (Oda). O objetivo é criar um modelo de aprendizado que atinja valores próximos aos resultados numéricos, para isso, os valores reais (ou valores observados) utilizados vieram das simulações numéricas via NDFM.

As métricas de desempenho utilizadas neste estudo e apresentadas na Tabela 4, são: *Mean Absolute Error* (MAE) - cálculo do erro absoluto médio, *Mean Squared Error* (MSE) - Média da diferença quadrática e *Root Mean Squared Error* ($RMSE$) - Raiz quadrada do erro médio quadrático.

Tabela 4: Métricas utilizadas para avaliar a regressão.

Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=0}^{N}\lvert y_i - \hat{y}\rvert$
Mean Squared Error	$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y})^2$
Root Mean Squared Error	$RMSE = \sqrt{(MSE)} = \sqrt{(\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y})^2)}$

Para essas 3 métricas, y é o valor predito (Oda ou RFR) e $\hat{y}$ o valor real (NDFM). Valores próximos de 0 (zero) indicam melhor desempenho. Além disso, podemos interpretar o MAE como uma medida da média residual no conjunto de dados, o MSE como uma medida de variância residual e o $RMSE$ como uma medida do desvio padrão dos resíduos.

Tabela 5: Medidas de erro para oda.

Baterias	**400**	**600**	**800**	**1000**
MAE	11.49	10.60	9.949	11.17
MSE	19.52	19.96	19.52	21.20
RMSE	381.1	398.6	381.2	449.5

Tabela 6: Medidas de erro para rfr.

Baterias	**400**	**600**	**800**	**1000**
MAE	6.242	5,333	5.033	4.426
MSE	9.6720	9,599	9.605	7.984
RMSE	93.55	92.14	92.25	63.74

As Tabelas 5 e 6 apresentam as medidas de erro para a solução encontra pelo método de Oda e pelo regressor de floresta aleatória (respectivamente). Com base nesses valores, é possível perceber que o modelo RFR erra menos que a solução analítica de Oda em todas as baterias propostas. Para uma melhor percepção, a Figura 5 apresenta de modo gráfico $RMSE$ de cada metodologia.

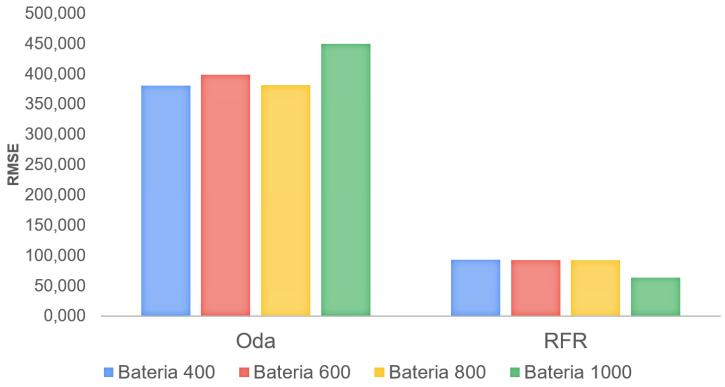


Fig. 5: RMSE das 4 baterias, Oda vs RFR.

4 CONCLUSÃO

O estudo focou no treinamento de um modelo de aprendizado de máquina baseado no algoritmo regressor de floresta aleatória (RFR), para prever valores de propriedades equivalentes em reservatórios fraturados. A estratégia para o treinamento do modelo utilizou

dados de entrada de soluções analíticas (Oda) e resultados das simulações numéricas via NDFM como valores de referência (valores de conceito alvo ou dados de saída). Foram construídos 10 reservatórios fraturados (conceituais), para criação de um conjunto de dados com 1000 amostras. O conjunto de dados foi dividido em 4 baterias de tamanho 400, 600, 800 e 1000 amostras. Uma busca por hiperparâmetros com validação cruzada foi conduzida sobre as 4 baterias e os resultados baseados nas métricas MAE, MSE e $RMSE$ apontam que o modelo regressor de floresta aleatória é mais preciso que a solução analítica de Oda. As 4 baterias também auxiliam para mostrar que o modelo de aprendizado de máquina tende a errar menos à medida que o conjunto de treinamento aumenta. Para a bateria 1, contendo 400 amostras, o RFR apresentou uma $RMSE$ 93.5 enquanto a solução analítica de Oda tem $RMSE$ de 381. Para a bateria 4, contendo 1000 amostras, o modelo de RFR apresentou uma $RMSE$ de 63.7, enquanto a solução analítica de Oda tem $RMSE$ de 449.5.

Com base nos resultados da busca por correlação, Figura 4, notamos que as variáveis são altamente correlacionadas, o que em alguns casos pode comprometer a generalização de modelos de regressão. Este estudo focou em utilizar as mesmas variáveis da solução analítica de Oda, porém em trabalhos futuros, é factível testes para a redução de dimensionalidade dos atributos de entrada. Além da análise de correlação não-linear entre os atributos de entrada e o de saída.

O estudo mostrou que modelos de aprendizado de máquina são ferramentas com capacidade de contribuir e melhorar significativamente metodologias já estabelecidas e consolidadas na indústria do petróleo e que a disponibilidade de dados é uma etapa importante para desenvolvimento e aplicação de novas estratégias baseadas em aprendizado de máquina.

5 Agradecimentos

Os autores apreciam o apoio financeiro da bolsa de mestrado fornecida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o número 133796/2020-5. Agradecemos as contribuições da professora Dsc. Mariza Ferro do LNCC e ao geocientista Mathieu Moriss da empresa Emerson Paradigm.

Referências

[1] Emerson Paradigm Holding LLC. Software de modelagem geológica skua-gocad™.

[2] G. Barenblatt, I. Zheltov, and I. Kochina. Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]. *Journal of Applied Mathematics and Mechanics*, 24(5):1286–1303, 1960.

[3] A. U. Chaudhry. Chapter 7 - well testing methods for naturally fractured reservoirs. In A. U. Chaudhry, editor, *Oil Well Testing Handbook*, pages 254–286. Gulf Professional Publishing, Burlington, 2004.

[4] R. Couronné, P. Probst, and A.-L. Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19:1471–2105, 2018. The role of fluids in faulting and fracturing in carbonates and other upper crustal rocks.

[5] A. Cutler, D. R. Cutler, and J. R. Stevens. *Ensemble Machine Learning: Methods and Applications*. Springer, Boston, MA, 2012.

[6] P. De Bièvre. The 2007 international vocabulary of metrology (vim), jcgm 200: 2008 [iso/iec guide 99]: Meeting the need for intercontinentally understood concepts and their associated intercontinentally agreed terms. *Clinical biochemistry*, 42(4-5):246–248, 2009.

[7] F. Felici, A. Alemanni, D. Bouacida, and P. de Montleau. Fractured reservoir modeling: From well data to dynamic flow. methodology and application to a real case study in illizi basin (algeria). *Tectonophysics*, 690:117–130, 2016. The role of fluids in faulting and fracturing in carbonates and other upper crustal rocks.

[8] P. K. Ghahfarokhi. The structured gridding implications for upscaling model discrete fracture networks (dfn) using corrected oda's method. *Journal of Petroleum Science and Engineering*, 153(Complete):70–80, 2017.

[9] Q. Guo, T. Zhuang, Z. Li, and S. He. Prediction of reservoir saturation field in high water cut stage by bore-ground electromagnetic method based on machine learning. *Journal of Petroleum Science and Engineering*, 204:108678, 2021.

[10] A. Géron. *Hands-On Machine Learning with Scikit-Learn TensorFlow*. O'Reilly Media; 2nd ed. edição, 2017.

[11] F. Hidayat and T. M. S. Astsauri. Applied random forest for parameter sensitivity of low salinity water injection (lswi) implementation on carbonate reservoir. *Alexandria Engineering Journal*, 2021.

[12] M. Karimi-Fard, L. Durlofsky, and K. Aziz. An Efficient Discrete-Fracture Model Applicable for General-Purpose Reservoir Simulators. *SPE Journal*, 9(02):227–236, 06 2004.

[13] A. Larestani, S. P. Mousavi, F. Hadavimoghaddam, and A. Hemmati-Sarapardeh. Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. *Journal of Petroleum Science and Engineering*, 208:109315, 2022.

[14] M. Oda. Permeability tensor for discontinuous rock masses. *Géotechnique*, 35(4):483–495, 1985.

[15] Scikit Learn - Machine Learning in Python. Overview, 2021. Access date: 1 set. 2021.

[16] J. E. Warren and P. J. Root. The behavior of naturally fractured reservoirs. *Society of Petroleum Engineers Journal*, 3:245–255, 1963.

[17] Z. Xu and Y. Yang. The hybrid dimensional representation of permeability tensor: A reinterpretation of the discrete fracture model and its extension on nonconforming meshes. *Journal of Computational Physics*, 415:109523, 2020.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Ressurgimento superdifusivo em caminhadas aleatórias não-Marcovianas

Antonio Adrielson dos S. Carvalho[1], Silvério Sirotheau Corrêa Neto[1], Jair Rodrigues Neyra[1, 2] e Thiago Rafael da Silva Moura[3, 4]

[1] *Faculdade de Engenharia/ Campus Universitário de Salinópolis, Universidade Federal do Pará, Salinópolis/PA, Brazil*

[2] *Programa de Pós-Graduação em Engenharia Química/ Laboratório de Ciência e Engenharia de Petróleo/ Instituto de Tecnologia, Universidade Federal do Pará, Belém/PA, Brazil*

[3] *Faculdade de Física/ Laboratório de Inovação Interdisciplinar/ Campus Universitário de Salinópolis, Universidade Federal do Pará, Salinópolis/PA, Brazil*

[4] *Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para Inovação, Universidade Federal do Sul e Sudeste do Pará, Marabá/PA, Brazil*

Abstract

Propomos um modelo de caminhadas aleatórias guiados pelo princípio de interação cinética (KIP). Segundo o KIP, a evolução temporal da função de distribuição de partículas idênticas sujeitas a colisões binárias nos remete a um funcional sempre crescente com o tempo e as caminhadas aleatórias do tipo com memória, que registram suas decisões ao longo do tempo. Neste contexto, buscamos analisar caminhadas aleatórias do tipo com memória dentro do quadro geral da teoria do princípio de interação cinética (KIP). Primeiramente, encontramos uma distribuição κ-exponencial discreta para construir nosso modelo de caminhadas aleatórias. Prosseguimos, realizando experimentações numéricas, quantificando os regimes difusivos típicos das caminhadas aleatórias. Em particular, evidenciamos duas regiões do parâmetro deformador κ. Para destacar os resultados da primeira região, partimos do limite máximo de anti-equilíbrio ($\kappa \to 3/2$), afastando-se da região de anti-equilíbrio. Afastando-se da região de máximo anti-equilíbrio, nossa segunda região de interesse, surpreendentemente, encontramos que os regimes difusivos são sensíveis as escalas típicas de κ. Afastando-se da região de máximo anti-equilíbrio para a região de equilíbrio, encontramos um região de transição entre os regimes difusivo ordinário, subdifusivo e superdifusivo, respectivamente.

Keywords:κ-exponencial, caminhadas aleatórias, difusão, regimes difusivos, transição.

Contato: Antonio, asc274@yahoo.com.br

1 INTRODUÇÃO

Sistemas desordenados podem apresentar uma característica chamada de auto-similaridade. A auto-similaridade está presente em diversos objetos matemáticos e naturais. Na geometria fractal, a auto-similaridade é a propriedade de um objeto ser semelhante a si mesmo em diferentes escalas. Para a matemática, estes objetos são chamados de fractais. Na natureza, diversos objetos satisfazem esta propriedade, por exemplo, as costas dos continentes, as macroestruturas de galáxias, fraturas geológicas, terremotos, agregações moleculares, superfícies, interfaces, polímeros, coloides, internet, trafego de cidades, estruturas topológicas, dentre outros [10] [11] [12] [13] [15] [16] [28] [32] [33] [38] [39] [40].

Um modelo que apresenta estruturas auto-similares são as caminhadas aleatórias. As caminhadas aleatórias são onde emergem objetos auto-similares. Elas exibem padrões de estruturas fractais presentes em diversos fenômenos: fenômenos de transporte, transporte anômalo, movimento Browniano, difusão em aglomerados, difusão em sistemas percolantes, difusão enviesada, agregação limitada por difusão, fenômenos de crescimento biológico, transporte através de interfaces, eletrodos fractais, membranas, fluxo multifásico em meios porosos, multifractais [3].

Na ótica da matemática, o estudo do fenômeno de difusão é abordado utilizando diversas técnicas teóricas como Martingales, equações diferenciais estocásticas, integrais estocásticas de Itô , sigma-álgebras, cálculo fracionário, análise combinatória, séries temporais, caminhadas aleatórias, equações de Fokker–Planck, etc [4] [17] [19] [20] [24] [25] [29] [30] [34] [37] [41].

Um outro grupo específico de caminhadas aleatórias acrescentou a característica de possuir memória. Isto é, a capacidade inerente do caminhante aleatório gravar todos os passos dados ao longo de toda a sua história. Em razão desta propriedade, as caminhadas com esta característica receberam o rótulo de caminhadas aleatórias do tipo (ou classe) com memória. O primeiro modelo deste tipo foi proposto por Schütz e Trimper [36].

Inspirados no modelo de Schütz e Trimper, diversos outros foram construídos utilizando novas características, tais como elementos que simulam perdas de memória ou danos, incluindo distribuições de probabilidade para construir os perfis de memória [1] [2] [5] [6] [7] [8] [9] [14] [27] [31].

Inspirados no desenvolvimento de modelos de caminhadas aleatórias do tipo com memória, propomos nosso modelo de caminhadas aleatórias com perfil de memória κ-exponencial. A distribuição κ-exponencial é resultante do princípio da interação cinética de Kaniadakis [21] [22]. Então, para construir o nosso modelo, primeiramente, encontramos uma distribuição κ-exponencial discreta. Prosseguimos, modelando as caminhadas aleatória com o perfil de memória κ-exponencial. Para ultimar, realizamos experimentações numéricas para caminhadas aleatórias discretas de tamanho finito. Nossas experimentações numéricas abrangeram duas regiões uma próxima do limite máximo de anti-equilíbrio, i.e., ($\kappa \to 3/2$) e outra, afastando-se da região de anti-equilíbrio, para valores discretos de ($\kappa > 3/2$). Para cada uma dessas situações, realizamos medidas do expoente de Hurst (H). Este expoente utilizado, em caminhadas aleatórias, para classificar os regimes difusivos em Os regimes difusivos são classificados em subdifusivo ($H < 1/2$), difusivo ordinário ($H = 1/2$) e superdifusivo ($1/2 < H < 1$) [18]. Neste trabalho, apresentamos nossos resultados.

2 METODOLOGIA

Em 2001, Kaniadakis propôs uma nova estatística que generaliza a estatística Clássica de Maxwell-Boltzmann-Gibbs [21] [22]. O arcabouço teórico usado por Kaniadakis foi o princípio de interação cinética (KIP). Segundo este princípio a evolução temporal da função de distribuição de partículas idênticas sujeitas a colisões binárias nos leva a construção de um funcional sempre crescente no tempo. Este funcional satisfaz a Segunda Lei da Termodinâmica. Este funcional está associado a um tipo de entropia definida por

$$S_\kappa = -\langle ln_\kappa \left[\psi\left(x\right)\right]\rangle = -\int dx\psi\left(x\right) ln_\kappa \left[\psi\left(x\right)\right] \tag{1}$$

sendo $\psi\left(x\right)$ a distribuição de velocidades das partículas, ln_k é o logaritmo deformado pelo parâmetro κ. O ln_k é uma função decrescente, real e válida $\forall x \in R$, dada por:

$$ln_k\left(x\right) = \frac{x^\kappa - x^{-\kappa}}{2\kappa} \tag{2}$$

onde sua inversa é chamada de κ-exponencial [23]. O parâmetro deformador da função, κ, é entendido com base nas possíveis correlações das partículas. O valor minoritário do índice é $\kappa \to 3/2$, associado ao valor do estado de anti-equilíbrio, caracterizado pelas correlações mais altas. O valor majoritário do índice é $\kappa \to +\infty$, está associado ao limite térmico clássico [26].

Para a construção de nosso modelo, vamos, primeiramente, determinar a função de distribuição de probabilidade em sua forma discreta. Para isso, vamos utilizar a técnica empregada em [35].

Propomos a função escrita na seguinte maneira

$$f_k\left(n\right) = A(\kappa)(exp_\kappa(-\lambda))^n \tag{3}$$

onde $n = 0, 1, 2, 3..., \infty$ e λ é um parâmetro: seus limites serão determinados posteriormente. Já que $f_k\left(n\right)$ é uma função densidade de probabilidade. Então, a soma total de $f_k\left(n\right)$ é unitária. A propriedade a seguir deve ser satisfeita

$$\sum_{n=0}^{\infty} f_k\left(n\right) = \sum_{n=0}^{\infty} A(\kappa)(exp_\kappa(-\lambda))^n = 1 \tag{4}$$

para realizar esta soma, vamos realizar a seguinte transformação

$$(exp_\kappa(-\lambda))^n = (exp(-\lambda_\kappa))^n \tag{5}$$

onde $\lambda_\kappa = \frac{1}{\kappa} ln(\sqrt{1+\lambda^2\kappa^2} + \lambda\kappa)$ [23]. Substituindo na função densidade de probabilidade, prosseguindo com a obtenção adequada da função densidade de probabilidade através da soma unitária

$$\sum_{n=0}^{\infty} f_k\left(n\right) = \sum_{n=0}^{\infty} A(\kappa)(exp(-\lambda_\kappa))^n = 1 \tag{6}$$

seguimos com

$$\sum_{n=0}^{\infty} A(\kappa)(exp(-\lambda_\kappa))^n = \frac{A(\kappa)}{1-exp(-\lambda_\kappa)} = 1 \tag{7}$$

Após calcular a constante $(A(\kappa))$, a função κ-exponencial discreta é obtida como uma função densidade de probabilidade para $n = 0, 1, 2, 3..., \infty$ e $\lambda \geq 0$, a saber,

$$f_\kappa(n) = (1 - exp(-\lambda_\kappa))(exp(-\lambda_\kappa))^n \tag{8}$$

Realizando a seguinte mudança de variáveis $n = (t - t')$ para $t > t'$, teremos que $f_\kappa(n) = f_\kappa(t', t)$ e usando as relações da Eq.(5), obtemos

$$f_\kappa(t', t) = (1 - exp_\kappa(-\lambda))(exp_\kappa(-\lambda))^{(t-t')} \tag{9}$$

Para alcançar o próximo passo de nossa construção, a saber, o modelo de caminhadas aleatórias com perfil de memória κ-exponencial, vamos aplicar a Eq.(9) no contexto de caminhadas aleatórias no contexto descrito por Schütz e Trimper [36]. De maneira general, uma caminhada aleatória é um processo constituído por valores aleatórios de posição. A sequência de valores aleatórios de posição formam caminho aleatório. A caminhada aleatória pode ser do tipo com memória, i.e., cada passo é gravado a cada instante de tempo; ou ainda, pode ser sem memória, quando cada passo não é gravado a cada instante de tempo. No modelo proposto por Schütz e Trimper, a memória é formada por um conjunto de variáveis aleatórias $\sigma_{t'}$, onde t' é o tempo escolhido de maneira equiprovável. A cada instante de tempo uma decisão do caminhante é recuperada a partir de uma distribuição uniforme, sendo que t é o tempo atual. Em nosso modelo substituímos a distribuição uniforme $1/t$ pela distribuição κ-exponencial discreta Eq.(9).

Existem dois aspectos fundamentais de nossas caminhadas aleatórias que precisam ser analisados: a dinâmica estocástica e acesso à memória. A dinâmica estocástica remete ao fato de que o caminhante está se movendo seguindo regras pré-estabelecidas. A memória está relacionada ao fato de que recobrar, no tempo atual t, segundo uma determinada distribuição de probabilidade, um passo dado em um instante de tempo anterior $t' < t$. Em nosso modelo esta lembrança é recobrada com probabilidade segundo κ-exponencial Eq.(9).

Quantitativamente, a conexão entre dinâmica estocástica e acesso à memória ocorre da seguinte maneira: o caminhante anda para a direita $+1$ ou para a esquerda -1, tal como em uma caminhada aleatória unidimensional Marcoviana. A equação de evolução estocástica da posição é

$$X_{t+1} = X_t + \sigma_{t+1} \tag{10}$$

No instante de tempo $t + 1$, a variável σ_{t+1} assume o valor $+1$ quando o caminhante executa um passo para a direita e -1 quando o caminhante executa um passo para a esquerda. A memória consiste de um conjunto de variáveis aleatórias $\sigma_{t'}$ para o tempo $t' < t$. Este processo ocorre da seguinte maneira:

1. no tempo $t = 1$, o caminhante, inicialmente na posição $X_0 = 0$, anda para a direita, $\sigma_1 = +1$, com probabilidade q ou para a esquerda, $\sigma_1 = -1$, com probabilidade $(1 - q)$. Segundo esta descrição, a probabilidade do primeiro passo é

$$P[\sigma_1 = \pm 1] = \frac{1}{2}[1 + (2q - 1)\sigma_1] \tag{11}$$

2. no tempo $t+1$, um tempo t' é escolhido do conjunto $\{1, 2, 3, ..., t\}$ com probabilidade $w(t)$;

3. no tempo $t+1$, σ_{t+1} é escolhido estocasticamente pela regra $\sigma_{t+1} = +\sigma_{t'}$ $\sigma_{t+1} = -\sigma_{t'}$ com probabilidade $p(1-p)$;

$$P[\sigma_{t+1} = \pm\sigma_{t'}|\sigma_{t'}] = \frac{1}{2}[1 + (2p-1)\sigma_{t+1}\sigma_{t'}] \tag{12}$$

4. usando as regras (2) e (3), obtemos a probabilidade condicionada para o tempo $t+1$

$$P[\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}] = \frac{1}{2}\sum_{j=1}^{t}[1 + (2p-1)\sigma\sigma_j]w(j) \tag{13}$$

onde $\sigma = \pm 1$ é o valor observado do conjunto $\sigma_{1,2,...,t} = \{\sigma_1, \sigma_2, ..., \sigma_t\}$ para $t \geq 1$.

Resolvendo a Equação (13), calculamos o deslocamento condicional $\langle\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}\rangle$

$$\langle\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}\rangle = \sum_{\sigma=\pm 1}\sigma P[\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}] \tag{14}$$

desenvolvendo a Eq.(14), encontramos

$$\langle\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}\rangle = \frac{1}{2}\sum_{j=1}^{t}[1 + (2p-1)\sigma\sigma_j]w(j) \tag{15}$$

prosseguindo resolvendo a Eq.(15), obtemos

$$\langle\sigma_{t+1} = \sigma|\sigma_{1,2,...,t}\rangle = \frac{1}{2}\sum_{j=1}^{t}\alpha\sigma_j w(j) \tag{16}$$

e usando a Eq.(16), encontramos a equação geral do primeiro momento da posição

$$\langle x_{t+1}\rangle = \langle x_t\rangle + \langle\sum_{j=1}^{t}\alpha\sigma_j w(j)\rangle \tag{17}$$

onde $\alpha = 2p-1$ e $x_t = X_t - X_0$, sendo o deslocamento do caminhante. Se $w(t) = 1/t$ o modelo é o proposto por Schütz e Trimper [36]. Em nosso problema, as caminhadas aleatórias possuem um perfil de memória κ-exponencial. Portanto, basta realizar a seguinte substituição $w(t) = f_\kappa(t', t)$ na Eq.(17)

$$\langle x_{t+1}\rangle = \langle x_t\rangle + \langle\sum_{j=1}^{t}\alpha\sigma_j f_\kappa(j, t)\rangle \tag{18}$$

A Eq.(18) não possui solução analítica fechada. Então, vamos resolvê-la numericamente. Para a solucionar este problemas numericamente, utilizamos os métodos de Monte Carlo para estimar o primeiro momento da posição

$$\langle x_t\rangle = \frac{1}{N}\sum_{j=1}^{N}x_t^j \tag{19}$$

o segundo momento da posição

$$\langle (x_t)^2 \rangle = \frac{1}{N} \sum_{j=1}^{N} (x_t^j)^2 \tag{20}$$

onde j é o índice que soma sobre a quantidade de N caminhantes no tempo t . Usando as Equações (19) e (20) estimamos a variância

$$Var(x_t) = \langle (x_t)^2 \rangle - (\langle x_t \rangle)^2 \tag{21}$$

Em particular, este tipo de passeio aleatório apresenta a característica de que o primeiro passo pode ser macroscopicamente relevante. Isto significa que possui impacto nos regimes de difusão medidos pelo expoente de Hurst [18]. Estimamos o expoente de Hurst (H) usando a lei de escala assintótica do desvio quadrático médio da posição em relação ao tempo

$$Var(x_t) = \langle (x_t)^2 \rangle - (\langle x_t \rangle)^2 = At^{2H} \tag{22}$$

onde A uma constante e H o expoente de Hurst. Para caminhadas aleatórias primeiro momento da posição cresce mais lentamente do que o segundo momento, a seguinte aproximação é pertinente $Var(x_t) \approx \langle (x_t)^2 \rangle = At^{2H}$. Os regimes difusivos são classificados em subdifusivo ($H < 1/2$), difusivo ordinário ($H = 1/2$) e superdifusivo ($1/2 < H < 1$) [18].

3 RESULTADOS E DISCUSSÃO

Realizamos experimentações numéricas de caminhadas aleatórias com perfil de memória κ-exponencial. As experimentações numéricas foram realizadas para 10^4 caminhadas e 10^5 passos. Exploramos as caminhadas aleatórias segundo o significado físico de κ, compreendido à partir de correlações das partículas. O menor valor do índice está no limite de $\kappa \to 3/2$, corresponde ao valor do estado de anti-equilíbrio. Este estado é mais distante do equilíbrio térmico clássico e, é caracterizado por correlações mais altas. No limite de $\kappa \to +\infty$, o maior valor do índice, corresponde ao comportamento do sistema no limite térmico clássico. Utilizamos, objetivando analisar o impacto de κ sobre os regimes difusivos, o valor da constante de decaimento $\lambda = 1$ e κ nas seguintes escalas: 10^1, 10^2, 10^3 e 10^4.

Na Fig. 1 são exibidas medidas do expoente de Hurst em função de κ e p. No diagrama, tons mais escuros mostram valores de H mais baixos, ($H < 1/2$) e ($H = 1/2$), evidenciando a emergência do regime subdifusivo e do regime difusivo ordinário, respectivamente. Tons mais claros mostram valores de H mais altos, $H \approx 1$, representando o regime superdifusivo. Como realizamos experimentações numéricas para medir H são para o intervalo do parâmetro deformador no intervalo $3/2 < \kappa < 8x10^5$. Notamos que não é visualmente perceptível, em razão das mudanças de escala de κ, observar os três regimes difusivos em um único gráfico. Para alcançar o objetivo de observar mais de perto a emergência dos três regimes difusivos e as transições entre os regimes difusivos, dividimos a apresentação dos nossos resultados em escalas do parâmetro κ, a saber, escalas de 10^1, 10^2, 10^3 e 10^4, respectivamente.

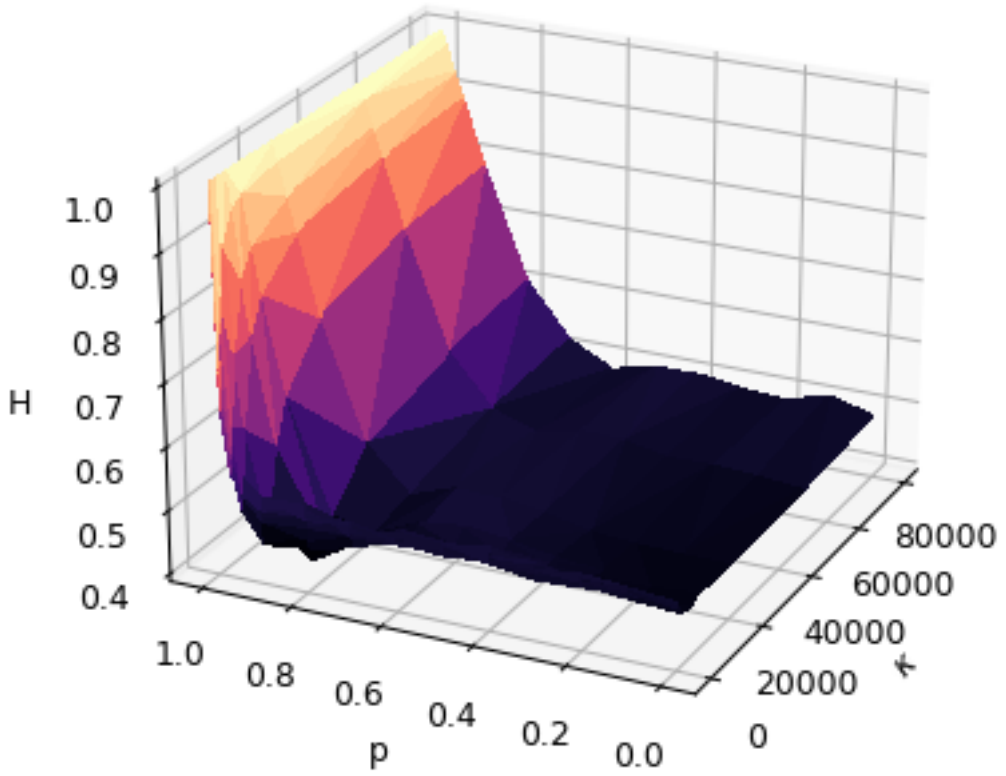


Fig. 1: Diagrama de difusão 3D: medidas típicas do expoente de Hurst em função de p e $3/2 < \kappa < 8x10^4$. Tons mais escuros, mostram valores de Hurst mais baixos, tipicamente $H \approx 0,5$, representando o regime difusivo. Tons mais claros mostram valores de Hurst mais altos, tipicamente $H \approx 1$, representando o regime superdifusivo.

A transição do regime difusivo apresenta um comportamento diferente para diferentes ordens de grandeza de κ. A Fig. 2 mostra os valores do expoente de Hurst em função da probabilidade para valores de κ da ordem de 10^1. Na região de anti-persistência $p < 1/2$, observamos o regime subdifusivo $H < 1/2$. Em $p = 1/2$, tipicamente emerge o difusão ordinária, havendo, portanto, neste intervalo uma transição do regime subdifusivo para o regime difusivo ordinário na medida que p cresce. Este comportamento é observado para valores de κ na ordem de 10^2, 10^3 e 10^4 apresentados nas Figuras 3, 4 e 5, respectivamente. Diante desta conclusão geral dos regimes difusivos para a zona de anti-persistência, prosseguiremos com nossa análise na zona de persistência ($p > 1/2$). Não obstante, para a Fig. 2, notamos que aumentando os valores de p, conseguimos observar outra transição, agora do regime difusivo ordinário para o superdifusivo na medida que ($p \to 1$). Notamos essa transição para valores próprios de $p = 0.9$.

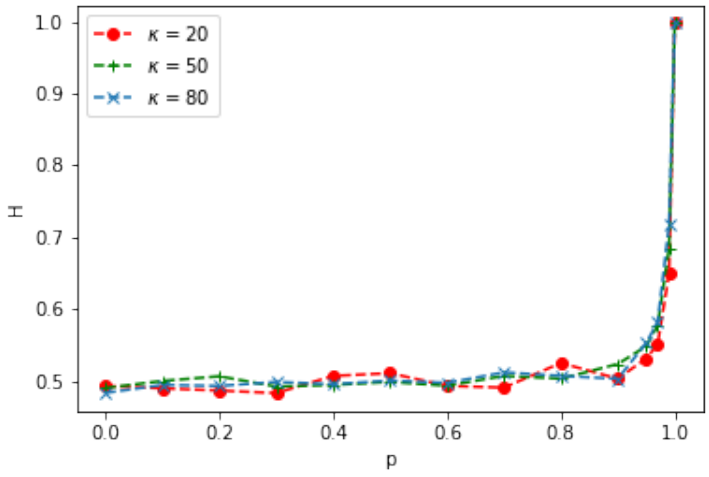


Fig. 2: edidas típicas do expoente de Hurst em função de p, mantendo κ fixo.

Para ordem de grandeza de $\kappa = 10^2$ verificamos um comportamento diferente da transição do regime difusivo com o aumento da probabilidade na região de persistência ($p > 1/2$). Relembramos que destacaremos os resultados da região de persistência uma vez que realização a conclusão geral dos regimes difusivos para todas as ordens de κ. A Fig. 3, em $0,6 < p < 0,95$, exibe valores característicos de transição do superdifusivo para o regimes subdifusivo para valores de κ mais elevados. Por exemplo, para $p = 0.9$, $\kappa = 80$ e $\kappa = 800$, temos $H > 1/2$ e $H < 1/2$, os regimes superdifusivo e subdifusivo, respetivamente.

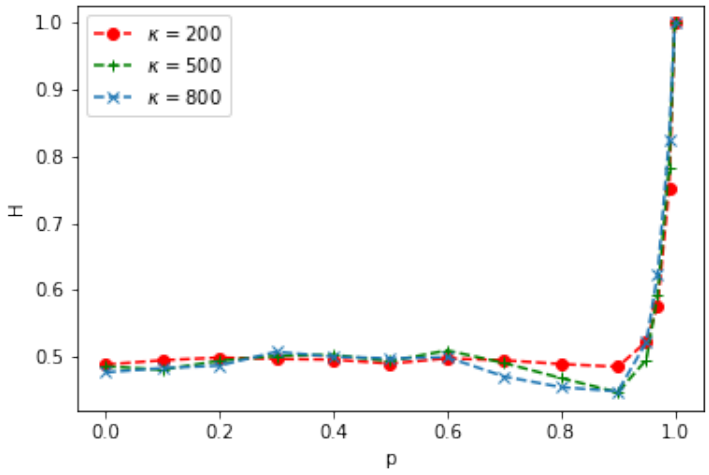


Fig. 3: Medidas do expoente de Hurst em função da probabilidade com $\kappa = 10^2$.

Observamos na Fig. 4, apresentamos os valores próprios de H para κ na ordem de 10^3. Aumentando a ordem de grandeza de $\kappa = 10^2$ (Fig. 3) para 10^3 (Fig. 4), observamos que o regime subdifusivo fica mais pronunciado na região de persistência, caracterizado por valores do expoente de Hurst menores. Para destacar a maior intensidade do regime subdifusivo, destacamos o comportamento de H para $\kappa = 2000$, onde em $p = 0,8$ exibe o regime subdifusivo mais intenso.

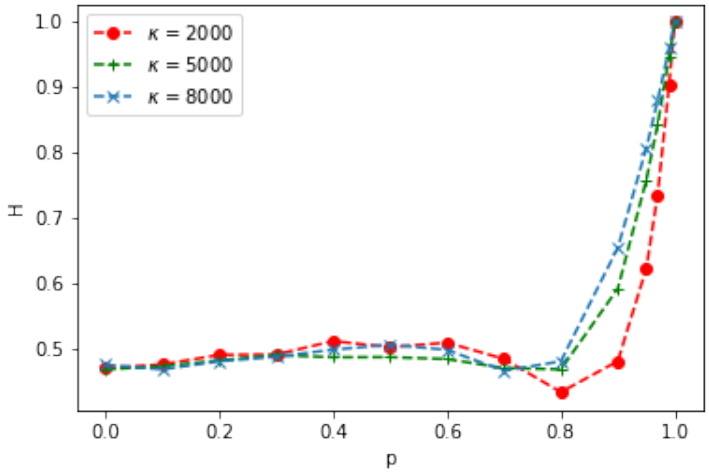


Fig. 4: Medidas do expoente de Hurst em função da probabilidade com $\kappa = 10^3$.

Na Fig. 5, apresentamos medidas típicas de H para κ na ordem de 10^4. Notamos que os valores do expoente de Hurst aumentam na medida que κ aumenta. Nesta tendência, os valores característicos de H são próprios de regime superdifusivo. Comparando as Figuras 4 e 5, notamos que exite uma transição do regime subdifusivo (Fig. 4) para o

superdifusivo (Fig. 5), passando pelo regime difusivo ordinário. Portanto, analisando κ a transição da ordem de grandeza de 10^3 para 10^4, as caminhadas aleatórias devem passar por duas transições de regime difusivo. A primeira, originando do regime subdifusivo para o difusivo ordinário e a segunda transição, originando do regime difusivo ordinário para o regime superdifusivo. No entanto, o ponto intermediário do regime difusivo ordinário não foi determinando. Ainda, outra hipótese é que ocorre apenas uma transição abrupta do regime subdifusivo para o superdifusivo, mas que requer análise numérica mais cuidadosa para ser determinado.

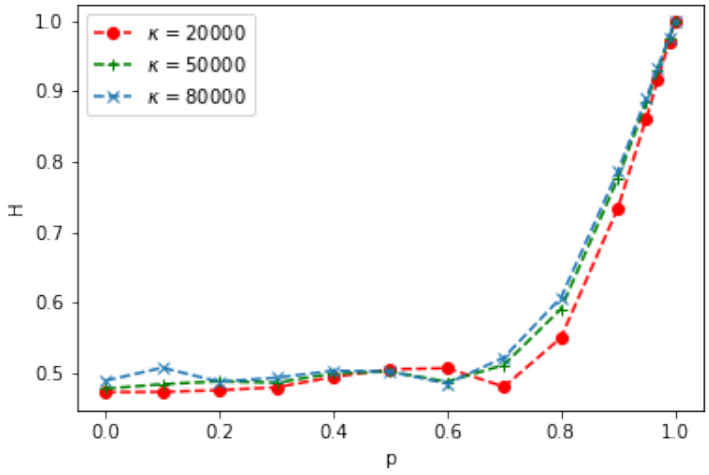


Fig. 5: Medidas do expoente de Hurst em função da probabilidade com $\kappa = 10^4$.

4 CONCLUSÕES

Propomos um modelo de caminhadas aleatórias com perfil de memória κ-exponencial discreto. Construímos a distribuição κ-exponential discreta para recobrar as decisões em caminhadas aleatórias do tipo com memória. Realizamos experimentações numéricas para quantificar os regimes difusivos típicos de nosso modelo de caminhadas aleatórias. Apresentamos nossos resultados na ordem crescente do parâmetro deformador da distribuição κ-exponencial. Verificamos a região da ordem máxima do anti-equilíbrio ($\kappa \to 3/2$), região em que as correlações são mais intensas, prosseguindo nos afastamos de região de anti-equilíbrio, onde as correlações são menos intensas. Destacamos que os regimes difusivos dependem da ordem de grandeza do parâmetro deformador κ. Surpreendentemente, encontramos diversas transições entre regimes difusivos: (i) Na região de anti-persistência $p < 1/2$, observamos uma transição do regime subdifusivo $H < 1/2$ para o regime de difusão ordinária, na medida que p cresce, em $p = 1/2$. Observamos este comportamento para todos os valores de κ utilizados em nossas simulações; (ii) na região de persistência ($p > 1/2$) transições entre regimes difusivos emergem ao longo das mudanças de ordem de κ. Encontramos as transições entre os valores de κ iguais a 10^1, 10^2, 10^3 e 10^4. Destacamos três momentos de transição: a) Para κ da ordem de 10^1, a curva de Hxp cresce suavemente, onde encontramos os regimes difusivo ordinário e superdifusivo; b) em 10^2 o regime muda para subdifusivo e aumenta de intensidade em 10^3 e c) para valores de típicos

de κ na ordem de 10^4, sofre outra transição para o regime superdifusivo; (iii) não obstante, observando a transição da ordem de grandeza de 10^3 para 10^4, notamos que as caminhadas aleatórias podem transitar entre os regimes difusivos subdifusivos de duas maneiras: a) a primeira maneira é composta por duas transições, originalmente do regime subdifusivo para o difusivo ordinário e a segunda transição do regime difusivo ordinário para o regime superdifusivo. Destacamos que o ponto intermediário do regime difusivo ordinário não foi determinado; e a segunda maneira b) ocorre apenas uma transição abrupta do regime subdifusivo para o superdifusivo. A determinação das características dessa transição requer mais investigação. Deixaremos estes pontos para serem determinados em um trabalho futuro, quando este trabalho ganhará continuidade.

5 *Agradecimentos*

Nós agradecemos a FAPESPA pelo suporte financeiro.

Referências

[1] G. A. Alves, J. M. de Araújo, J. C. Cressoni, L. R. da Silva, M. A. A. da Silva, and G. Viswanathan. Superdiffusion driven by exponentially decaying memory. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(4):P04026, 2014.

[2] G. Borges, A. Ferreira, M. da Silva, J. Cressoni, G. Viswanathan, and A. Mariz. Superdiffusion in a non-markovian random walk model with a gaussian memory profile. *The European Physical Journal B*, 85(9):1–5, 2012.

[3] A. Bunde and S. Havlin. *Fractal and Disordered Media*. Springer-Verlag, Berlin Heidelberg, second edition, 1995.

[4] C. C. Chen and K.-M. Koh. *Principles and techniques in combinatorics*. World Scientific, 1992.

[5] J. Cressoni, G. Viswanathan, and M. da Silva. Log-periodicity in piecewise ballistic superdiffusion: Exact results. *Physical Review E*, 98(5):052102, 2018.

[6] J. Cressoni, G. Viswanathan, A. Ferreira, and M. da Silva. Alzheimer random walk model: Two previously overlooked diffusion regimes. *Physical Review E*, 86(4):042101, 2012.

[7] M. A. A. da Silva, G. Viswanathan, A. Ferreira, and J. Cressoni. Spontaneous symmetry breaking in amnestically induced persistence. *Physical Review E*, 77(4):040101, 2008.

[8] K. de Lacerda, J. Cressoni, G. Viswanathan, and M. A. da Silva. Log-periodicity can appear in a non-markovian random walk even if there is perfect memory of its history. *EPL (Europhysics Letters)*, 130(2):20004, 2020.

[9] R. Diniz, J. Cressoni, M. da Silva, A. Mariz, and J. de Araújo. Narrow log-periodic modulations in non-markovian random walks. *Physical Review E*, 96(6):062143, 2017.

[10] G. Z. et al. *Switch between critical percolation modes in city traffic dynamics.* Proc Natl Acad Sci U S A, 2019.

[11] P. E. et al. *Uncovering space-independent communities in spatial networks.* Proc Natl Acad Sci U S A, 2011.

[12] S. C. et al. *A model of Internet topology using k-shell decomposition.* Proc Natl Acad Sci U S A, 2007.

[13] S.-H. Y. et al. *Modeling the Internet's large-scale topology.* Proc Natl Acad Sci U S A, 2002.

[14] M. Felisberto, F. Passos, A. Ferreira, M. da Silva, J. Cressoni, and G. Viswanathan. Sudden onset of log-periodicity and superdiffusion in non-markovian random walks with amnestically induced persistence: exact results. *The European Physical Journal B*, 72(3):427–433, 2009.

[15] J. Gaite. *Scaling Laws in the Stellar Mass Distribution and the Transition to Homogeneity.* Advances in Astronomy, 2021.

[16] S. Havlin, A. Barabási, S. Buldyrev, C. Peng, M. Scwartz, H. Stanley, T. Sander, and P. Meakin. *NATO Advanced Research Workshop.* Plenum, New York, 1993.

[17] R. Hilfer. *Applications of fractional calculus in physics.* World scientific, 2000.

[18] H. E. HURST, R. P. BLACK, and Y. M. SIMAIKA. Long term storage. *An experimental study*, 1965.

[19] E. T. Jaynes. *in E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics.* Dordrecht, Holland, 1983.

[20] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge, 2003.

[21] G. Kaniadakis. Non-linear kinetics underlying generalized statistics. *Physica A: Statistical mechanics and its applications*, 296(3-4):405–425, 2001.

[22] G. Kaniadakis. Statistical mechanics in the context of special relativity. *Physical review E*, 66(5):056125, 2002.

[23] G. Kaniadakis. Theoretical foundations and mathematical formalism of the power-law tailed statistical distributions. *Entropy*, 15(10):3983–4010, 2013.

[24] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus.* Springer, 2008.

[25] D. C. Lay. *Linear Algebra and Its Applications.* Addison Wesley, Tappan, NJ,, 2005.

[26] G. Livadiotis. *Kappa distributions: Theory and applications in plasmas.* Elsevier, 2017.

[27] T. R. Moura, G. Viswanathan, M. da Silva, J. Cressoni, and L. da Silva. Transient superdiffusion in random walks with a q-exponentially decaying memory profile. *Physica A: Statistical Mechanics and its Applications*, 453:259–263, 2016.

[28] U. Nakaya. Snow crystals. *Harvard University Press*, 1954.

[29] M. Newman and R. C. Thompson. Math. Comput., Tappan, NJ,, 1987.

[30] J. A. Oteo. Math. Phys, 1991.

[31] F. N. Paraan and J. Esguerra. Exact moments in a continuous time random walk with complete memory of its history. *Physical Review E*, 74(3):032101, 2006.

[32] R. Pynn and T. Riste. Time-dependent effects in disordered materials. *Plenum*, 1987.

[33] R. Pynn and A. Skjeltorp. *Scaling phenomena in disordered systems*. Plenum, New York, 1986.

[34] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 1 and 2. Cambridge university press, 2000.

[35] H. Sato, M. Ikota, A. Sugimoto, and H. Masuda. A new defect distribution metrology with a consistent discrete exponential formula and its applications. *IEEE Transactions on Semiconductor Manufacturing*, 12(4):409–418, 1999.

[36] G. M. Schütz and S. Trimper. Elephants can always remember: Exact long-range memory effects in a non-markovian random walk. *Physical Review E*, 70(4):045101, 2004.

[37] A. Stanislavsky and K. Weron. Numerical scheme for calculating of the fractional two-power relaxation laws in time-domain of measurements. *Computer Physics Communications*, 183(2):320–323, 2012.

[38] D. Stauffer and H. Stanley. From newton to mandelbrot: A primer in theoretical physics. 1995.

[39] D. L. Turcotte. *Scaling in geology: landforms and earthquakes*. Proc Natl Acad Sci U S A, 1995.

[40] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. *Scaling phenomena in the Internet: Critically examining criticality*. Colloquium, 2002.

[41] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, New York, 2007.

SIMULAÇÃO DO PROCESSO DE AQUECIMENTO DA ÁGUA POR MICRO-ONDAS

Janaína Monteiro Pedrosa[1] e Arley Silva Rossi[1]

[1] *Universidade Federal do Amazonas, Manaus/AM, Brazil*

RESUMO

A tecnologia utilizando as micro-ondas tem sido muito empregada em ambientes domésticos, laboratoriais e industriais. O aquecimento por micro-ondas tem como base o acoplamento de dois fenômenos físicos: a transferência de calor e o eletromagnetismo. Neste caso, a cinética de aquecimento e a interação do campo eletromagnético com o material estão associadas pelas propriedades dielétricas. A utilização do *software* COMSOL *multiphysics* permitiu o estudo do processo de aquecimento da água pura, para a validação dos dados experimentais e simulados descritos neste trabalho. Os testes foram realizados em 6 tempos (1-6min) com uma massa de água de 300g em todos os ensaios, aquecidos em um forno micro-ondas da marca eletrolux modelo MEC41, em que foram realizados em duplicata medindo a temperatura do início e no final de cada tempo. Comparando os dados obtidos através da simulação, o erro ficou abaixo de 7%, isso permitiu a validação do processo de simulação através dos dados experimentais. A simulação computacional mostrou uma forte ligação das propriedades dielétricas da água com relação a absorção de energia, através da geração do campo elétrico e a dissipação dessa energia na forma de calor.

Palavras-chaves: Eletromagnetismo, Transferência de Calor, Micro-ondas, Propriedades Dielétricas.

1 INTRODUÇÃO

A utilização de simuladores de processos tornou-se evidente nos últimos anos, por ser uma ferramenta que permite o estudo de diversos parâmetros em diferentes processos industriais. Uma das alternativas de simuladores eficientes para o estudo em engenharia é o COMSOL *Multiphysics* que, por meio de informações sobre a física do processo e as características do material, simula o processo que está sendo estudado. O processo a qual se transmite energia de um meio, ou material, a outro de menor temperatura é conhecido como transferência de calor. A

transferência de calor ocorre por três diferentes mecanismos: condução, convecção e radiação. A condução pode ser explicada pelo contato direto das partículas de um material com outro, o que pode ser visto em um fogão elétrico, já a convecção é a transferência de calor que ocorre através do movimento do fluido. O único que pode ocorrer na ausência de contato é a radiação, onde a energia radiante é transferida de uma fonte a um receptor [1].

No mecanismo da radiação, o receptor absorve parte da energia causando um aumento da temperatura, como quando esquentamos um alimento no forno micro-ondas. Esse aquecimento ocorre devido a interação do campo eletromagnético com o material a ser aquecido, por meio de mecanismos dielétricos. Quando um dielétrico é exposto a um campo elétrico externo, este sofre um efeito denominado polarização, este efeito consiste em um rearranjo ou reorientação das moléculas que compõem o dielétrico [2].

A polarização do dipolo é o principal mecanismo de aquecimento por micro-ondas em moléculas polares, onde o dipolo é uma característica natural do material dielétrico por apresentarem em sua estrutura atômica uma separação natural de cargas. Em moléculas apolares, o dipolo é induzido por um campo elétrico externo, através da distorção da nuvem eletrônica, onde os dipolos se realinham conforme o campo varia [3].

As propriedades dielétricas de um material apresentam parâmetros importantes para a compreensão do comportamento destes materiais, especificamente dos materiais dielétricos, visando melhorar ou desenvolver processos. Essas propriedades podem ser influenciadas por alguns fatores, como por exemplo a natureza do material, temperatura e frequência do campo elétrico aplicado [4].

2 METODOLOGIA

2.1 – Cinética de aquecimento da água

O forno para aquecimento utilizado no presente trabalho pertence a marca ELECTROLUX modelo MEC41 (Figura 1) com potência de 1000W e frequência nominal de 2,45GHz. A cavidade interna possui as seguintes medidas 370mm de comprimento e profundidade, e 258mm de altura. Já o guia de ondas fica posicionado na lateral direita superior no meio da cavidade interna do forno e possui as seguintes medidas: 69mm de comprimento, 107mm de profundidade e 25,8mm de altura. É interessante perceber que a altura do guia de ondas é proporcional à altura da cavidade. Isso ocorre para garantir que as paredes do forno não absorvam energia gerada pelas micro-ondas. Garantindo, assim, que o material a ser aquecido receba grande parte da energia disponível.

Figura 1– Micro-ondas utilizado no trabalho
Fonte: Manual de instruções MEC41, 2013.

O forno ainda possui um disco de vidro giratório localizado na parte inferior da cavidade, onde é depositado o material a ser aquecido. Esse disco giratório garante que o material sofra um aquecimento mais homogêneo possível. O equipamento ainda possui um sistema de exaustão, responsável pela retirada dos vapores gerados na cavidade do forno. Foi utilizado um copo medidor de 400ml de volume para o aquecimento da água. Para medidas de temperatura foi empregado um termômetro culinário digital da marca Clink com range de medida de -50 a 300ºC para medidas de temperatura da operação de aquecimento.

2.2 – Procedimento de aquecimento

Foram realizados 12 conjuntos de ensaio experimental neste trabalho com o objetivo de aquecer o material desejado. A massa de água utilizada foi de 300g em todos os testes. Com intuito de coletar vários pontos de medidas os experimentos foram realizados em diferentes tempos de aquecimento (0,5 a 6min) em que cada ponto foi coletado de 30 em 30 segundos, utilizando a potência máxima do equipamento. A temperatura foi medida antes e depois de cada processo de aquecimento e todos os testes foram realizados em duplicada, totalizando ao final 24 experimentos. Os valores obtidos na operação de aquecimento da água via micro-ondas foram compilados em um gráfico e comparados com os valores obtidos através da simulação computacional.

2.3 - Simulação eletromagnética

A título de comparação foi realizado um estudo de fluidodinâmica computacional utilizando, para isso, o *software* COMSOL Multiphysics versão 5.5. Para esse estudo foi gerada uma malha da mesma geometria do forno utilizado nos experimentos. Trata-se de um forno da marca

Electrolux, com potência de 1000W distribuído pelo guia de ondas posicionado em sua lateral direita. Para a simulação eletromagnética utilizou-se a mesma massa de 300g de água dos testes experimentais. A Figura 3 exibe o forno utilizado nos testes experimentais, bem como o modelo desenvolvido no simulador COMSOL para realizações das simulações eletromagnéticas.

Figura 3 – (a) Cavidade do forno micro-ondas; (b) modelo computacional para simulação
Fonte: copilado pelo autor[1]

A Figura 3(a) exibe a cavidade interna do forno empregado para o aquecimento. A Figura 3(b) mostra o modelo desenvolvido no programa COMSOL para simulação com seu respectivo guia de ondas posicionado na lateral direita da cavidade. Para que não aconteça interferência destrutiva das micro-ondas o prato giratório fica posicionado no centro da cavidade abaixo do elemento a ser aquecido. Deste modo, ocorre melhor distribuição das ondas, garantindo que o centro do forno receberá a maior parte da energia liberada. A Figura 4 traz as principais dimensões calculas em milímetros e considerações observadas no modelo desenvolvido no simulador.

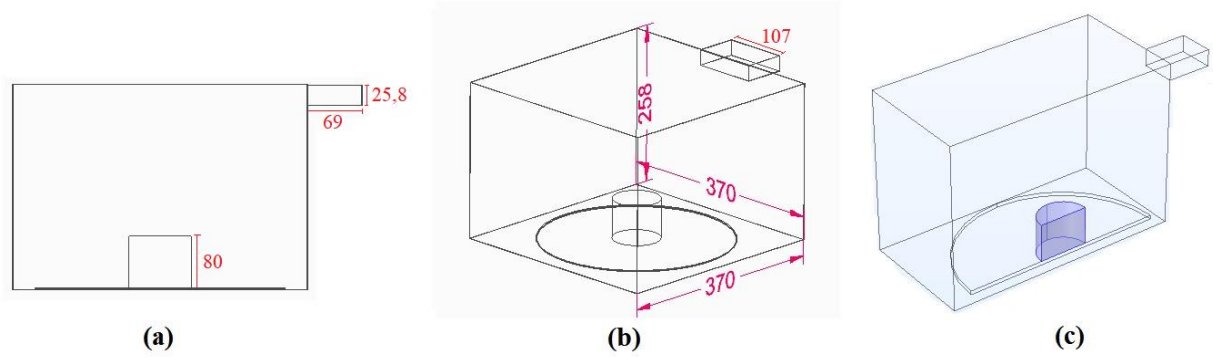


Figura 4 – (a) Vista frontal; (b) Vista isométrica (c) Modelo computacional COMSOL

[1] Montagem a partir de imagens coletadas no manual de instruções MEC41 (a) e simulação realizada pelo autor no COMSOL *multiphysics* (b)

Fonte: Autoria própria, 2021

As vistas frontal e isométrica (Figura 4 a e b) mostram como a geometria da cavidade do forno é simétrica. Deste modo, devido a essa simetria, os testes de aquecimento através da simulação eletromagnética foram realizados apenas com a metade da cavidade (Figura 4c), diminuindo assim, o esforço computacional.

2.4 – Metodologia e malha computacional

A simulação computacional de aquecimento via micro-ondas pelo *Software* COMSOL engloba a conexão de dois fenômenos físicos: a transferência de calor e o eletromagnetismo. Partindo-se do módulo geral *Radio Frequency Module* (*RF Module*) adiciona-se o módulo *Microwave Heating*. Neste caso, trata-se de um módulo específico para aquecimento via micro-ondas onde são resolvidas as equações de eletromagnetismo e de transferência de calor. A parte do eletromagnetismo é definida pela equação de Maxwell determinando, assim, a distribuição de campo elétrico na cavidade do micro-ondas (Eq. 1).

$$\nabla\left(\frac{1}{\mu}\nabla\times\vec{E}\right)-\frac{\omega^2}{c}(\varepsilon'-\varepsilon''j)\vec{E}=0 \quad (1)$$

$\vec{E}$ é a intensidade do campo elétrico [V/m], ω é a frequência de onda angular [2πf ,rad/s], μ é a permeabilidade relativa do material, c é a velocidade da luz no espaço ($3x10^8$m/s), constante dielétrica (ε') e o fator de perda dielétrica (ε''). A intensidade do campo elétrico obtido, a partir das Eq. de Maxwell e as propriedades dielétricas do material são indispensáveis para se calcular a potência volumétrica gerada pelo material devido a exposição as micro-ondas (Eq. 2). Desta forma, a fase final da simulação é a resolução da Eq. de Balanço de Energia para se calcular o gradiente de temperatura do material devido a condução. Como simplificação do modelo não foi adotado os termos referentes a transferência por convecção. As variáveis dependentes do módulo e resolvidas no presente trabalho são o perfil de temperatura T[^{0}C] da água a ser aquecida e a distribuição de campo elétrico E[V/m] na cavidade do forno [5].

$$Q=\sigma|\overrightarrow{E}|^2=2\pi\varepsilon_0\varepsilon''f|\vec{E}|^2 \quad (2)$$

Em que Q é a densidade de potência [W/m³], que é a energia absorvida pela amostra, σ é a condutividade elétrica de material [S/m], ε_0 é a permissividade no espaço livre ($8{,}854x10^{12}$F/m). Para calcular a distribuição de temperatura do material devido a condução e convecção é necessário resolver a equação de balanço de energia (Eq. 3).

$$\rho C_p \frac{\partial T}{\partial t} + \rho C_p \vec{u} \nabla T = k \nabla^2 T + Q \quad (3)$$

Em que, $\vec{u}$ [m/s] é o vetor velocidade que faz parte do termo de transferência de calor por convecção, não utilizado nos modelos da simulação desse trabalho, k [W/mK] que é a condutividade térmica, Cp é a capacidade calorífica [J/kgK] e ρ é a densidade [kg/m³].

É importante ressaltar que as propriedades dielétricas do elemento a ser aquecido tem grande importância em todo processo de simulação eletromagnética. Assim, a permissividade complexa (ε_r) que engloba a constante dielétrica (ε') e o fator de perda dielétrica (ε''), foi utilizada de maneira constante, isto é, sem variação com a temperatura. Isto gerou um menor esforço computacional na busca de respostas sobre a formação do campo elétrico e o gradiente de temperatura da água. A título de comparação foi realizada uma única simulação eletromagnética com a ε_r variando com a temperatura. Deste modo, todos dos testes se iniciaram com a temperatura da água de 25ºC e potência nominal de 1000W.

Todas as propriedades físico-químicas dos materiais da cavidade do forno, bem como da água a ser aquecida foram retiradas do banco de dados do próprio simulador COMSOL. Para a malha de controle foi utilizada uma configuração no formato tetraédrico com tamanhos de elementos (S_{max}) aplicando o critério de Nyquist (Eq. 4) descrito nos trabalhos de Rossi et.al., 2016 e Vaz et al., 2014. Tal malha é largamente empregada em problemas eletromagnéticos em geral.

$$S_{max} = \frac{\lambda}{2} = \frac{c}{2f\sqrt{\varepsilon'\mu}} \quad (4)$$

Onde λ representa o comprimento de onda (m), c é a velocidade da luz no vácuo ($3x10^8$m/s), f é a frequência de oscilação do campo (Hz) ε' é a constante dielétrica do material e μ é a permeabilidade do material (H/m).

3 RESULTADOS E DISCUSSÕES

3.1 Malha computacional

A malha computacional foi aplicada em todo o volume da cavidade obedecendo o tamanho máximo de elementos seguindo o critério apresentado pelo modelo de Nyquist (Eq.4).

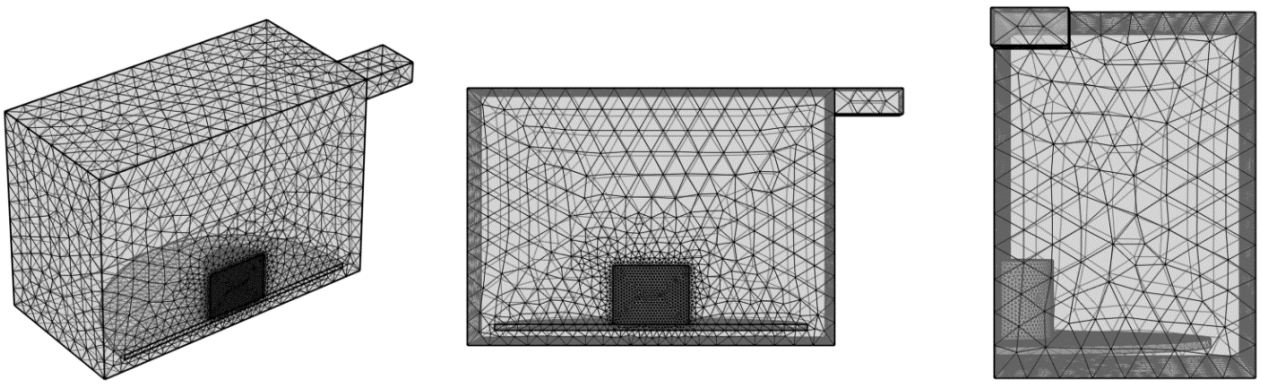


Figura 4 – Malha computacional empregada no forno

Fonte: Autoria própria, 2021

A parte menor e mais escura, em formato cilíndrico (Fig. 4) representa o elemento a ser previamente aquecido (água pura). Desta maneira, esse elemento apresenta um maior refinamento de sua malha de controle devido a sua complexibilidade. O objetivo, neste caso, é apresentar resultados mais precisos e com uma maior confiabilidade. As demais propriedades físico-químicas como densidade, capacidade calorífica e condutividade térmica foram utilizadas a partir do banco de dados do próprio simulador. Devido a limitações computacionais a maioria dos testes de simulação eletromagnética foram feitas com as propriedades dielétricas constantes e obtidas a partir do banco de dados do COMSOL.

3.2 – Cinética de aquecimento e simulação computacional

As simulações computacionais foram realizadas nos mesmos intervalos de tempo e frequência dos testes experimentais executados no forno micro-ondas. Assim, após os testes de aquecimento experimentais foram realizadas as simulações eletromagnéticas e os dados coletados tanto na fase experimental quanto na simulação podem ser observados na Figura 5.

Tabela 1 - Propriedades dielétricas da água pura (SALVI et al., 2011; ROSSI, 2017)

	20ºC	40 ºC	60 ºC	80 ºC	100 ºC
ε'	80,1	69,7	57,9	44,8	37,7
ε''	12,1	11,5	9,2	5,2	2,6

Com os dados obtidos da Tabela 1 foi realizada uma regressão linear com cada propriedade dielétrica separadamente. O intuito dessa ferramenta foi obter equações onde se pudesse observar os valores de propriedades dielétricas variando-se com a temperatura. Deste modo, garantindo resultados mais fidedignos e próximos a um experimento convencional de aquecimento por micro-ondas. As melhores regressões forneceram um quadrado do coeficiente

de correlação ou R^2 = 99,3% para constante dielétrica e R^2 = 95% para o fator de perda dielétrica. Deste modo, a permissividade complexa obtida é:

$$\varepsilon_{r=}(90{,}95 - 0{,}5476T) - (15{,}701 - 0{,}1261T)i \quad (5)$$

Desta maneira, a Equação complexa (Eq. 5) foi inserida no programa com permissividade complexa variando com a temperatura.

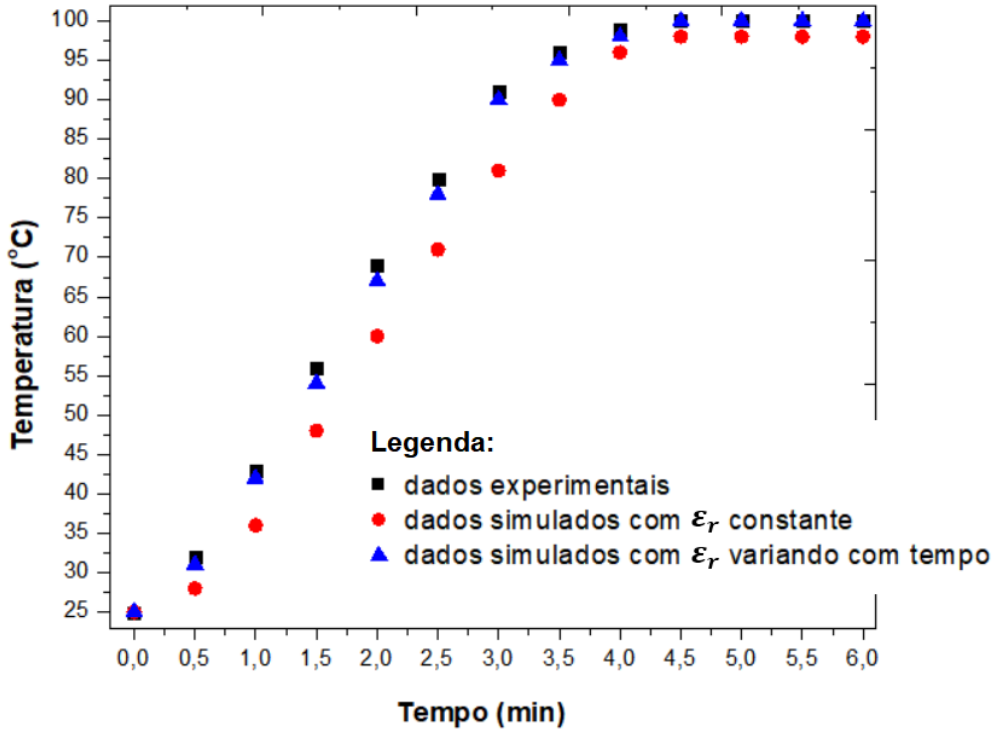


Figura 5: Dados experimentais e simulados para o aquecimento de água

Fonte: Autoria própria, 2021

Quando comparamos os dados experimentais aos dados simulados com ε_r variando com a temperatura, percebemos uma grande proximidade. Isso se deve ao fato que as principais propriedades envolvidas em processos de aquecimento de micro-ondas (propriedades dielétricas) são levadas em consideração. E há um substancial diferença entre dos dados medidos em 20ºC e 100ºC. No caso da constante dielétrica (ε') ocorreu uma diminuição de aproximadamente 53% do seu valor. Por outro lado, para o fator de perda dielétrica (ε'') a queda foi ainda mais expressiva, em torno de 78,5%. Sem dúvidas esses valores impactam no processo de simulação, e quando levados em consideração no processo computacional, geram valores muito próximos

aos encontrados experimentalmente. Levando em consideração que se trata de um estudo preliminar de aquecimento e os dados simulados com ε_r constante apresenta boa proximidade com os dados experimentais. Calculou-se o erro relativo para cada par de pontos do trabalho e no total o erro ficou abaixo de 7%. Isso demostra a eficiência do processo de simulação eletromagnética em retratar experimentos de aquecimento em micro-ondas. Devido a esses resultados os próximos estudos no presente trabalho levam em consideração apenas os dados obtidos com as propriedades dielétricas obtidas em temperatura constante.

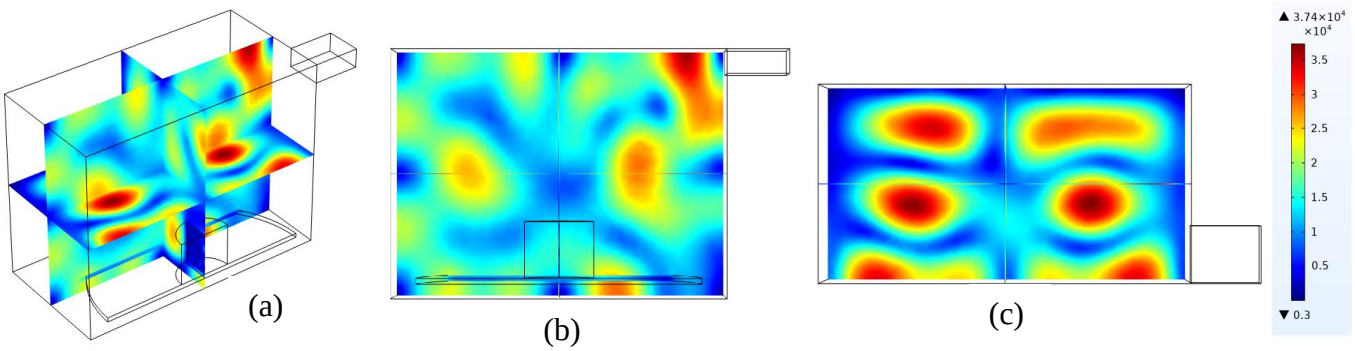


Figura 6 – Distribuição de campo elétrico (V/m) no forno micro-ondas

Fonte: Autoria própria, 2021

A vista isométrica (a) mostra o comportamento do campo elétrico em toda a cavidade do forno. A vista frontal (b) e a vista superior (c) apresentam essa distribuição do campo elétrico na cavidade do forno, onde é possível observar uma assimetria do campo elétrico no interior do forno. Essa assimetria pode ser explicada pela posição do guia de ondas do micro-ondas, em que essa se encontra posicionada apenas na parte superior direita, como é possível observar em (b), a presença do campo é muito maior próximo a saída do guia de ondas do que no centro da cavidade. Essa defasagem do campo elétrico é comum em processos de aquecimento em micro-ondas, visto que a variação do seu valor é responsável pela agitação das moléculas internas da amostra causando o efeito de rotação dipolar. Deste modo, o alinhamento e desalinhamento das moléculas libera calor, como resultado da fricção dessas moléculas.

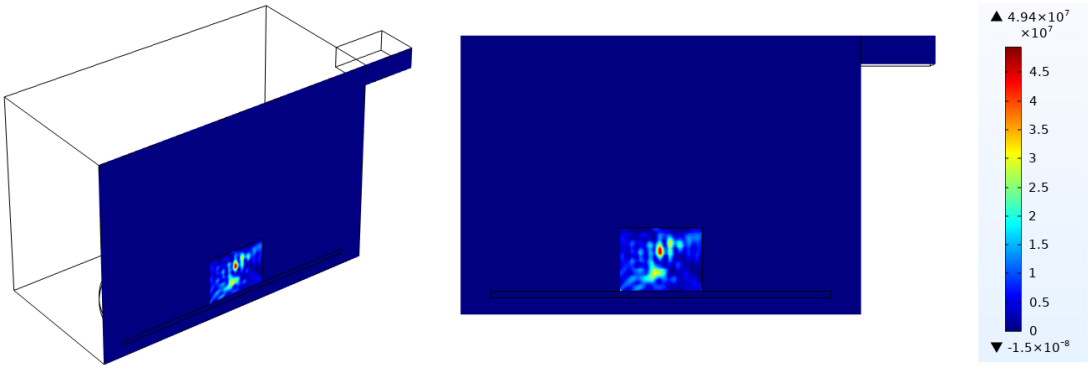


Figura 7 – Densidade de potência (W/m^3) para aquecimento de água em micro-ondas
Fonte: Autoria própria, 2021

Na operação de aquecimento ocorre uma variação no padrão de oscilação na formação do campo elétrico isto resulta em um padrão comum de densidade de potência. Em outras palavras, as regiões com maior densidade de potência encontram-se no centro da amostra que está sendo aquecida, sugerindo que o aquecimento se dará do centro para a periferia. Esse resultado já foi percebido em outros trabalhos como de Rossi, 2017 e Santos, 2014, 2018. Os Pesquisadores realizaram ensaios experimentais de secagem de cascalhos contaminados com fluidos de perfuração em micro-ondas. Deste modo, a depender do tamanho da amostra a ser seca com relação ao tamanho da cavidade o aquecimento pode se dar de maneiras distintas. Para uma amostra pequena em relação a cavidade do forno o aquecimento se dá do centro para periferia, para amostras grandes em relação a cavidade, o aquecimento se dá das bordas para o centro.

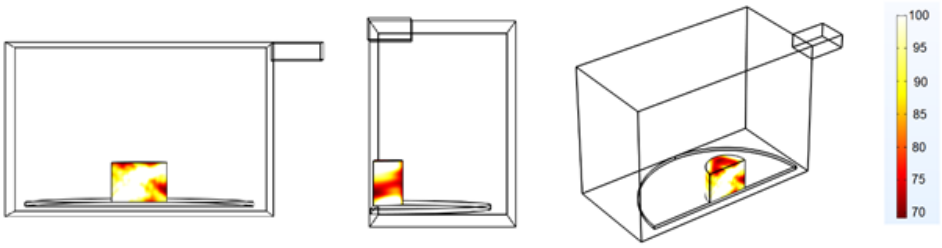


Figura 8 –Distribuição de temperatura (ºC) para aquecimento de água em micro-ondas
Fonte: Autoria própria, 2021.

O padrão observado na formação da densidade de potência se repete aqui. Observou-se, maiores temperaturas nas regiões centrais, indicando que a dissipação de energia na forma de calor se dá do centro para periferia da amostra analisada. Foi observado uma variação razoável de temperatura da Figura 8, esse fenômeno pode ser explicado pela formação das correntes de

convecção que em um experimento convencional de aquecimento por micro-ondas é observado. Contudo, devido a dificuldades do modelo não foram consideradas no presente trabalho.

4 CONCLUSÃO

Como visto no decorrer do presente trabalho, o aquecimento por micro-ondas é um mecanismo que envolve as propriedades dielétricas do material a ser aquecido. A água por sua vez, apresenta propriedades favoráveis ao aquecimento, visto que quanto maior for a quantidade de água no material, mais favorável ao aquecimento o material vai ser.

A simulação do aquecimento da água pura no COMSOL apresentou resultados satisfatórios, em que os dados obtidos experimentalmente se mostraram bem próximos aos obtidos através da simulação. A densidade de potência apresentada pela simulação descreveu com eficiência o que era esperado teoricamente, onde o aquecimento descrito aconteceu da parte central para as extremidades, assim como o gradiente de temperatura e o campo elétrico, apesar da pequena assimetria apresentada.

Assim, através deste trabalho foi possível validar a simulação eletromagnética de aquecimento da água a partir dos dados experimentais, obtidos a partir dos experimentos de cinética de aquecimento da água pura em micro-ondas. Deste modo os dados experimentais se mostraram bem próximos aos simulados. A única ressalva está ligada ao uso de propriedades dielétricas da substância a ser aquecida. No caso da água ficou comprovado que a diferença entre dados experimentais e simulados ficou abaixo de 7%. Isso justifica o uso das propriedades dielétricas constantes nesse tipo de simulação, garantindo resultados confiáveis e com baixo esforço computacional.

4 AGRADECIMENTOS

Primeiramente agradecer a Deus por ser a razão de tudo. Agradecer ao Dr. Arley Rossi por todo suporte durante a realização deste trabalho.

5 REFERÊNCIAS

[1] RAFFINO, M. E., transferência de calor. *Concept.de*. Disponível em: https: //conósito.de/transferencia-de-calor/. Acessado em: 16 de agosto de 2020.

[2] FILHO, O. C., A inclusão dos efeitos de temperatura na função resposta dielétrica do modelo de dissado – Hill e seu potencial para implicações em materiais de interesse para eletrônica. Dissertação de mestrado. Pato Branco/PR, UTFPR, 2014.

[3] RAYMUNDO, L. M., Estudo de aquecimento via micro-ondas. Trabalho de conclusão de curso. Porto Alegre/RS, UFRS, 2013.

[4] CREMASCO, P. F. M., Instrumentação para a caracterização dielétrica de filmes biodegradáveis. Dissertação de mestrado. Pirassununga/SP, FZEA/USP, 2016

[5] ROSSI, A. S., Cinética de aquecimento e secagem, propriedades dielétricas e simulação computacional aplicado ao tratamento de cascalho de perfuração por micro-ondas. Tese de doutorado. Uberlândia/MG, UFU, 2017.

[6] SALVI, D.; BOLDOR, D.; AITA, G.M.; SABLIOV, C.M. COMSOL Multiphysics model for continuous flow microwave heating of liquids. Journal of Food Engineering 104, p.422–429, 2011.

[7] SANTOS, J. M., Remediação de sólidos de perfuração via aquecimento por micro-ondas. Dissertação de Mestrado. Uberlândia/MG, UFU, 2014.

[8] SANTOS, J. M; PETRI, I. J; MOTA, A. C. S; MORAIS, A. S; ATAÍDE, C. H., Optimization of the batch decontamination process of drill cuttings by microwave heating. Journal of Petroleum Science and Engineering 163, p. 349-358 , 2018.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

NAZCA: a machine-learning based methodology for performance prediction and configuration recommendation of multiscale numerical simulations

Juan H. L. Fabian[1], Antônio T. A. Gomes[1] e Eduardo Ogasawara[2]

[1] *Laboratório Nacional de Computação Científica (LNCC), Petrópolis/RJ, Brazil*
[2] *Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro/RJ, Brazil*

Abstract

Simulating multiscale phenomena requires computationally robust methods. The proper use of these methods demands an understanding of their inherent complexity. We propose a machine-learning based methodology called NAZCA, aimed at helping users of this type of simulations. We consider the Multiscale Hybrid Mixed (MHM) finite element method as our case study. For now, we have used NAZCA to estimate the execution time and recommend numerical parameters of MHM simulations. We developed specific learning techniques that explore knowledge about the numerical method. We show that these techniques obtain smaller errors than other state-of-the-art techniques and a high level of interpretability. We expect to also use NAZCA with other numerical methods with similar computational characteristics.

Keywords:Multiscale methods, Machine learning, Performance evaluation

1 INTRODUCTION

Phenomena with multiscale characteristics require sophisticated numerical methods to deal with these characteristics in terms of the quality of approximation and computational performance. The so-called *multiscale numerical methods* tackle both issues; they achieve low approximation error rates and incorporate the granularity of the new generations of massively parallel architectures. For this paper, we consider multiscale numerical methods for finite element (FE) analysis. From a mathematical viewpoint, these methods are composed of: (i) a global formulation defined in a coarse partition of the domain; and (ii) a collection of local problems, partition by partition, guided by the problem data –

Contato: Juan H. L. Fabian, juanhlf@lncc.br

which is inherently multiscale [5]. In computational terms, this formulation induces a process with two stages: an asynchronous and a coupled stage. Local problems are solved in the asynchronous stage, and a global problem is solved in the coupled stage. Crucially, the global problem is typically smaller than those found in classical numerical methods and, therefore, computationally advantageous.

One drawback of the aforementioned two-stage process is that it increases the number of numerical parameters and configuration possibilities. Properly configuring numerical simulations for the intended approximation error rates and the number of computing resources involved is particularly important in shared computing infrastructures such as clusters in supercomputing centers. In these infrastructures, workload management systems are responsible for regulating users' access to computing nodes. These systems implement scheduling strategies that arbitrate resource contention, managing queues of jobs sent by users. Typically, users and the supercomputing centers benefit from job specifications that provide accurate estimates of total execution time. Nevertheless, providing accurate estimates for simulations based on multiscale numerical methods is not easy; each configuration possibility impacts the quality of approximation and computational performance achieved.

To help the users of simulators based on multiscale FE methods, we present a methodology called NAZCA, which employs machine learning for performance prediction and configuration recommendation of these simulations. We use the MHM method proposed in [1] as a use case for training and testing the prediction models and the recommenders. Nonetheless, NAZCA may also be applied to other methods, notably, those with the same parallel execution pattern as MHM (e.g., [9, 3]).

We have not found any work in the literature that aims at performance prediction and configuration recommendation of simulations based on (either classical or multiscale) FE methods. As far as we know, research on predicting the execution time of high-performance computing (HPC) applications in shared computing infrastructures [16, 13, 12, 14] has only targeted general-purpose code kernels and parallel execution patterns. For the recommendation of parameters, none of the approaches we are aware of [18, 19] use domain-specific information about the applications.

We organized the remainder of this paper as follows. In Section 2, we briefly describe the MHM method. In Section 3, we present our NAZCA methodology. In Section 4, we compare the machine learning techniques developed within NAZCA with some other state-of-the-art techniques. In Section 5, we present some concluding remarks.

2 THE FAMILY OF MHM METHODS

The MHM methodology encompasses a family of FE methods aimed at solving large problems with multiple scales. It departs from a hybrid FE formulation that enforces the continuity of the solution space of a partial differential equation (PDE) using Lagrange multipliers. The hybrid formulation is then rewritten to obtain two types of problems: global and local. They are then discretized to obtain proper numerical approximations to the solution of the original PDE. The global problem is solved on the skeleton of a fixed FE mesh that partitions the domain of the PDE. The local problems are defined in an independent way for each mesh element. Each local problem considers its corresponding element of the mesh a domain of its own, possibly partitioning it with a "sub-mesh".

There are various validated problem examples of the MHM methodology in the literature [1, 11, 2, 4]. For the sake of illustration, we consider a boundary value problem for a diffusive process in a two-dimensional domain Ω,[1] with a multiscale $\mathcal{K}$ coefficient:
Diffusion problem: *Find the pressure* $u : \Omega \to \mathbb{R}$ *in the domain* Ω *such that:*

$$\begin{cases} -\mathcal{K}\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

A rigorous definition of the MHM method applied to diffusion problems is in [10, 1]. In the following, we briefly present the main ingredients of the method that help build predictive models for both execution time and numerical error. First, the following (infinite dimensional) function spaces are defined:

- $\mathbf{V}$: the space of u living over Ω; and
- $\mathbf{\Lambda}$: the space of Lagrange multipliers living over the skeleton formed by a decomposition of Ω. It is associated with the normal fluxes over the subdomains' boundaries.

In the MHM method, a solution u can be characterized as:

$$u = u_0 + \boldsymbol{T}(\lambda, f), \text{ with } u_0 \in V_0, \boldsymbol{T} : \mathbf{\Lambda} \times L^2(\Omega) \to \tilde{V}, \lambda \in \mathbf{\Lambda}, \text{ and } \mathbf{V} = V_0 \bigoplus \tilde{V},$$

where V_0 is the space in which the kernel of the differential operator lives – in our illustrative example, the Laplacian (Δ) operator. This hybrid formulation is rewritten to obtain two types of problems: global and local, which are then discretized.

In the discretization procedure, the global problem is solved on the skeleton of a mesh of elements $\mathcal{T}_H = \{K\}$. The skeleton is defined by $\mathcal{E}_H = \{\partial K\}_{K \in \mathcal{T}_H}$, where ∂K is the boundary of K, *i.e.*, the set of element faces in $\mathcal{T}_H$. $H > 0$ is the characteristic measure (*i.e.*, the measure of the largest face) of $\mathcal{T}_H$ and reflects its level of refinement. In our implementation of MHM, a local problem is associated with each $K \in \mathcal{T}_H$, which may be solved in parallel, independent from the other local problems.

The approximate (finite dimensional) function spaces are then:

$$\Lambda_H = \Lambda_l^m \subset \Lambda \text{ and } \tilde{V}_h = \bigoplus\nolimits_{K \in \mathcal{T}_H} \tilde{V_K} \subset \tilde{V}.$$

The parameter m in the space of Lagrange multipliers defines the number of partitions of each face of $\mathcal{E}_H$. The parameter l defines the degree of Lagrange polynomials in each such partition. Notably, the finite set of basis functions that span Λ_l^m are not known *a priori*; the local problems compute and "upscale" them to the global problem. This is how MHM (and multiscale methods in general) captures multiscale features. At the local level, each K has its space $\tilde{V}_K$ formed by Lagrange polynomials of degree k that live on a "sub-mesh" within K. $h > 0$ is the characteristic measure of this sub-mesh.

In computational terms, the MHM method (and again multiscale methods in general) can be seen as a two-stage process: (i) the *asynchronous stage* solves the local problems independently of each other, without communication among the involved processors; (ii) the

[1] Much of the description in this section also applies to a three-dimensional domain setting.

coupled stage collects the solutions to the local problems (the upscaling procedure) to build a single, coupled problem that uses all available processors synchronously.

Since the local problems are much cheaper computationally than the global problem, they may be performed offline when the upscaling procedure can be done only once for a single simulation setup. This is often the case for stationary (e.g., the diffusive process illustrated above) and transient linear problems, but not for nonlinear problems (e.g., phenomena governed by the Navier-Stokes equations). In the latter, the multiscale basis functions that span Λ_l^m must be computed at each step of an iterative linearization process. It renders an algorithm in which the local and global problems must be solved online, *i.e.*, within the same simulator instantiation.

2.1 On the estimation of the execution time of MHM simulations

It is well known in the literature (see, for instance, [8]) that the time spent on FE simulations is mainly due to the solution of their underlying linear system of equations. Therefore, these linear systems are the main target of our learning approach to estimating the execution time of MHM simulations. Nevertheless, the matter becomes somewhat more complex for MHM, as different linear systems appear globally and locally. For the sake of argument, we only consider direct approaches to solving these linear systems.

Because of the hybridization procedure, the linear system associated with the global problem has the general saddle-point form below, with the dimension of A being determined by l, m, and $\#\mathcal{E}_H$, and the dimensions of B and B^T being proportional to $\#\mathcal{T}_H$:

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ u_0 \end{pmatrix} = \begin{pmatrix} g_f \\ g_0 \end{pmatrix}.$$

Importantly, the larger are the parameters l and m defining the level of refinement of the mesh, the more challenging is the linear system for the solvers. Moreover, these parameters affect the linear system differently: refining the mesh – *i.e.*, increasing $\#\mathcal{E}_H$ and $\#\mathcal{T}_H$ only – increases the dimensions of the matrix; increasing l or m makes the matrix not only bigger but also denser. Considering these aspects has an important impact on the quality of the predictions for the time to run simulations based on the MHM method.

For each local problem, there is also a set of $N_I + 1$ linear systems of the form:

$$\begin{aligned} LC &= F, \\ LD_i &= N_i, \ \forall i \in \{0, ..., N_I - 1\}. \end{aligned}$$

The dimension of L is determined by k and h. N_I is defined by the number of faces in the element K and by l and m. These systems share the same matrix L, which can be factorized only once. Hence, increasing l or m and the number of faces in the element K only increases the matrix-vector multiplications, which are less computationally expensive than the factorizations (up to a limit). Mind that we have as many local problems as the number of elements in $\mathcal{T}_H$, and these local problems are independent of each other. So, we can distribute the computation of their corresponding linear systems across all the cores made available by the shared computing infrastructure. Estimating the execution time for solving the local problems must therefore take into account this distribution.

2.2 Convergence analysis of the MHM method

In [10], the MHM method is presented for the diffusion problem. In [1], the analysis of this method is detailed. Also, in [17], an analysis of this method is presented for a more complex configuration of the diffusion problem. Convergence analysis of the method is described in all cases, showing the expected behavior of the approximate solution to the diffusion problem to have an error in the L^2-norm of $\mathcal{O}(H^{l+2})$. There are some other error estimates in the literature, in which approximations of local problems are considered.[2]

3 THE NAZCA METHODOLOGY

NAZCA aims to assist users in the configuration of multiscale simulations and the computing resources used for these simulations. The hypothesis is that NAZCA can provide such assistance by learning from performance measures obtained from previous simulations.

NAZCA departs from a set of three *parameter spaces*: (i) the simulation characterization, (ii) the computational architecture, and (iii) the performance metrics. Figure 1 illustrates a possible dataset in NAZCA considering these three parameter spaces. Some of the parameters may be interrelated: *e.g.*, data about RAM usage and execution time in the performance metrics space is only present if the simulation ends successfully.

For each problem related to using multiscale simulations, we envision an specific task within NAZCA. So far, the task types considered in NAZCA are prediction and recommendation. They can be viewed as workflows of conceptual components in the methodology, as depicted in Figure 2. Any user requirement must be translated to a specific NAZCA task of one of these two task types. In this paper, we consider two specific scenarios within which we define NAZCA tasks. These scenarios are described below.

[2]In this paper, the error norms are the usual ones for Sobolev spaces.

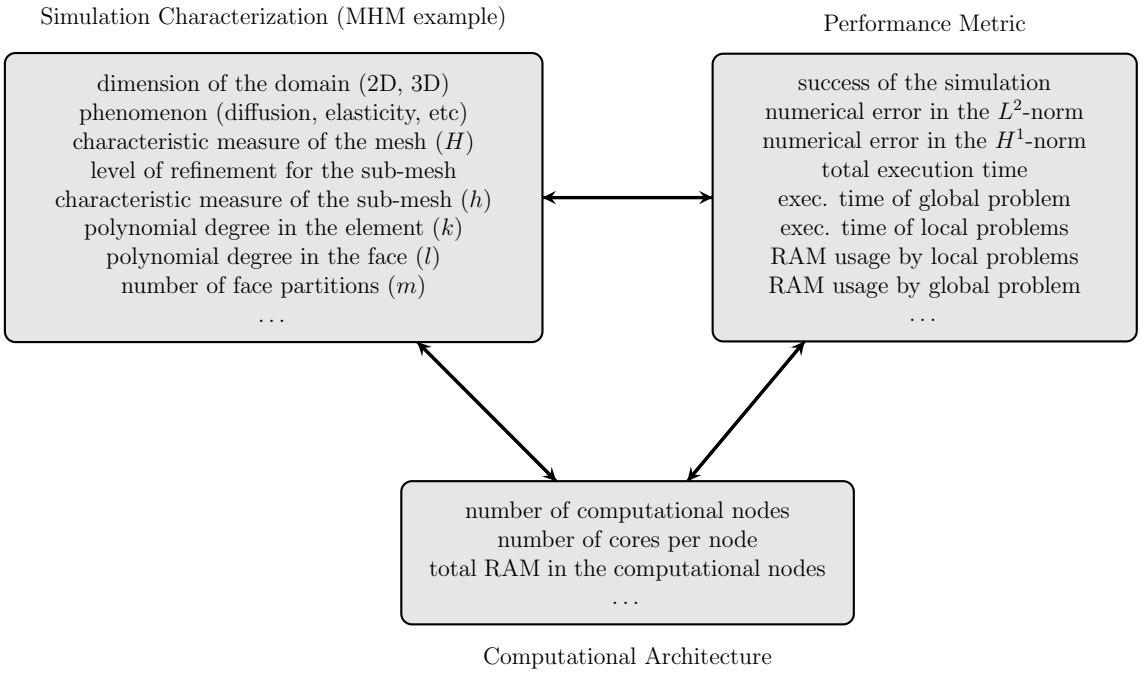


Fig. 1: Examples of NAZCA parameter spaces.

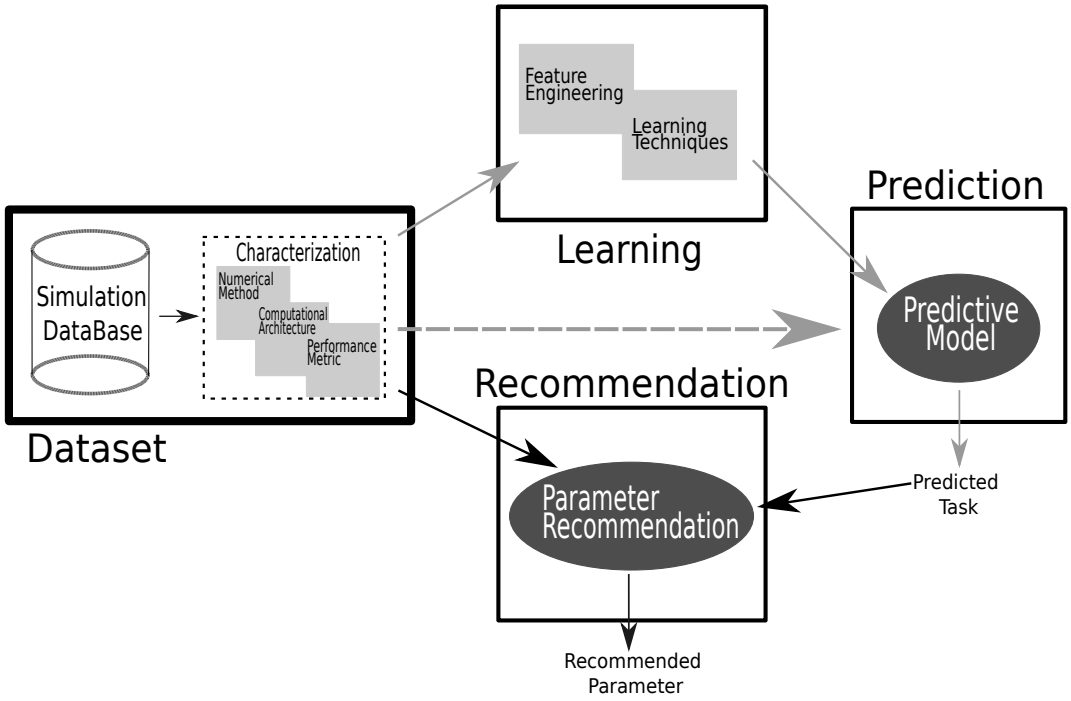


Fig. 2: Conceptual components of the NAZCA methodology.

3.1 Scenario 1: Prediction of MHM simulator performance

As explained in Section 2, the execution time of MHM simulations encompasses the time used in the asynchronous and coupled stages. The execution time of the global problem in the coupled stage (TPG) is influenced by the parameters l, m, $\#\mathcal{E}_H$, and $\#\mathcal{T}_H$. In particular, l and m are determinants for the sparsity pattern of the system of linear equations associated with the global problem. Thus, we devised a tree-based prediction model that handles each possible combination of l and m. Moreover, we derived a latent attribute (GLG) that represents the total number of degrees of freedom in the linear system solved by the global problem. We employed GLG and TPG respectively as the predictor and the target variable of several univariate regression models, each one living on a different leaf of the tree.

The time used in the asynchronous stage (TMPL) is influenced by the number of local problems and computational architecture. For a specific local problem, k is determinant for the sparsity pattern of its linear system of equations. Thus, we devised a tree-based prediction model that handles each value of k. Moreover, we derived a latent attribute (GLLT) that represents the total number of degrees of freedom of all local linear systems allocated to a single core. GLLT is computed using the Equation 1, where GLL is the number of degrees of freedom in a specific local linear system:

$$GLLT = \lceil \frac{\#LocalProblems}{\#CoresInArchitecture} \rceil \times GLL \tag{1}$$

We employed GLLT and TMPL respectively as the predictor and the target variable of several univariate regression models, each one living on a different leaf of the tree.

At each stage, selecting the best univariate regression models on tree leaves is performed using the empirical analysis procedure presented in [6, 7]. Finally, the actual target variable for the execution time of MHM simulations is defined as TEL = TPG + TMPL.

Building predictive models for the error in the L^2-norm is performed in a similar way as described for the execution time. Considering what was described in Section 2 about the convergence analysis of the MHM method, we devised a tree-based architecture that handles each possible combination of k, l, and m. The attribute H, which is a characteristic measure of the mesh, is employed as the predictor of several univariate regression models, each one living on a different leaf of the tree. L2, the numerical error in the L^2-norm, is used as the target variable. Unlike what we did for the execution time, the selection of models in the leaves of the tree was made using analysis of variance.

3.2 Scenario 2: Recommendation of MHM numerical parameters

In Algorithm 1, we show the behavior of the proposed method for recommending numerical parameters for an MHM simulation. The recommendation is based on the execution time and the numerical accuracy targeted by the user. The algorithm starts by loading the training dataset of simulations (`loadDataset()`) and checking whether it contains all possible combinations of numerical parameters; if not, the predictive models from Scenario 1 are employed (`completeDataset()`) to enrich the dataset. After that, the recommender generates an euclidean space with normalized TEL as the X-axis and normalized L2 as the Y-axis, so as to select an entry in the enriched dataset with the pair (TEL, L2) having the smallest distance to the pair targeted by the user (`euclideanDistance()` and `minDistance()`). The recommender then retrieves the numerical parameters associated with the selected entry in the enriched dataset (`selectParameters()`).

Algorithm 1 Recommendation of numerical parameters for an MHM simulation

Require: execution time (TEL) and numerical accuracy (L2) targeted by the user
  $Sims_{train} \leftarrow$ `loadDataset`()
  $Sims_{total} \leftarrow$ `completeDataset`($Sims_{train}$)
  $Sims_{dist} \leftarrow \emptyset$
  for all $sim \in Sims_{total}$ **do**
    $sim_{dist} \leftarrow$ `euclideanDistance`($(sim\$TEL, sim\$L2); (\text{TEL}, \text{L2})$)
    $Sims_{dist} \leftarrow Sims_{dist} \cup sim_{dist}$
  end for
  $Sim_{selected} \leftarrow$ `minDistance`($Sims_{dist}$)
  $Par_{selected} \leftarrow$ `selectParameters`(Sim_{select})

4 Experimental Evaluation

For our experiments, we set a single configuration for the computational architecture, consisting of a workstation with two 12-core sockets and 320 GB of RAM. All the simulations used 2 MPI processes to collect performance metric data. We varied the numerical parameters of the MHM method (see [6, 7] for details) to amount to a total of 1800 simulations in our experimental dataset.

We randomly split the dataset into training and test datasets with a holdout of 80-20. For Scenario 1, we trained the models with the training dataset and assessed them with the test dataset. We did not optimize the hyperparameters of the models; therefore, a validation set was not defined. For Scenario 2, we used the complete dataset for the recommendation process. To evaluate the recommendation, we only used the test dataset.

4.1 Scenario 1

We compared our prediction models presented in Section 3 with state-of-the-art models automatically selected by the Auto-WEKA tool.[3] For the execution time, Auto-WEKA selected a Gaussian Process Regression. For the error in the L^2-norm, it selected an Additive Regression with Random Forest as the base learner. Table 1 shows the quality of fit obtained for each case by our prediction models and those selected by Auto-WEKA.

Table 1: Prediction of execution time and L^2-norm error

	Technique	RMSE
TEL	NAZCA	0.440 seg
	Gaussian Processes	1.488 seg
L2	NAZCA	0.036
	Additive Regression (Random Forest)	0.006

Importantly, Auto-WEKA implements a time-sensitive approach to finding the best model. We let Auto-WEKA run for approximately one week in the same ordinary laptop computer where we ran our model selection procedure.

For the execution time, we conclude that using domain-specific information related to the numerical method allowed the NAZCA models to reach a better generalization. The same did not apply to the L^2-norm error; nevertheless, all NAZCA models employed much less computational effort during the learning process than Auto-WEKA.

4.2 Scenario 2

In these experiments, we are interested in recommending the numerical parameters k, l, and m. Since the evaluation of the parameter recommendation is performed using the test dataset, each of the simulations belonging to this dataset will be considered a target simulation in the recommendation process. We consider, for the TEL and L2 variables in the test set, three possible approaches for predictive models using NAZCA and Auto-WEKA. For each variable, in the first approach, only NAZCA models are used. In contrast, in the second, only Auto-WEKA models are used. In the third, NAZCA models are used for TEL and Auto-WEKA models are used for L2 (since in the previous subsection, we showed that Auto-WEKA selected a model with an smaller error).

We calculated the number of times a given approach recommended simulation parameters with the smallest distance from the target simulation. The results of this calculation

[3]Auto-WEKA [15] searches the better models between the learning algorithms and their hyperparameters implemented in the WEKA workbench [20]

are shown in Table 2. We observed that the NAZCA models allowed us to get the simulation parameters closest to the target simulations.

Table 2: Recommendation of parameters

	Technique	Accuracy
shortest distance	NAZCA	85.39% (304/356)
	WEKA	13.20% (47/356)
	Hybrid (NAZCA+WEKA)	1.41% (5/356)

5 Conclusion

In this work, we presented NAZCA, a methodology for performance prediction and parameter recommendation of multiscale simulators. We gathered performance data from simulations based on the MHM method applied to a diffusive process to evaluate our methodology. By comparing the predictive models obtained with NAZCA with those automatically selected by the Auto-WEKA tool, we concluded that our proposed techniques are competitive both for performance evaluation and parameter recommendation.

The results presented herein are valid for the specific dataset gathered in our experiments. Using a different dataset of MHM simulations may render different learned models, albeit preserving their tree-based structure and relevant parameters.

References

[1] R. Araya, C. Harder, D. Paredes, and F. Valentin. Multiscale Hybrid-Mixed Method. *SIAM Journal on Numerical Analysis*, 51(6):3505–3531, 2013.

[2] R. Araya, C. Harder, A. H. Poza, and F. Valentin. Multiscale hybrid-mixed method for the Stokes and Brinkman equations – The method. *Computer Methods in Applied Mechanics and Engineering*, 324:29–53, 2017.

[3] T. Arbogast, G. Pencheva, M. F. Wheeler, and I. Yotov. A multiscale mortar mixed finite element method. *Multiscale Modeling & Simulation*, 6(1):319–346, 2007.

[4] T. Chaumont-Frelet and F. Valentin. A multiscale hybrid-mixed method for the Helmholtz equation in heterogeneous domains. *SIAM Journal on Numerical Analysis*, 58(2):1029–1067, 2020.

[5] Y. Efendiev and T. Y. Hou. *Multiscale Finite Element Methods*. Springer, 2009.

[6] J. H. L. Fabian, A. T. A. Gomes, and E. Ogasawara. Estimating the execution time of fully-online multiscale numerical simulations. In *Anais do XXI Simpósio em Sistemas Computacionais de Alto Desempenho (WSCAD)*, pages 191–202, 2020.

[7] J. H. L. Fabian, A. T. A. Gomes, and E. Ogasawara. Estimating the execution time of the coupled stage in multiscale numerical simulations. In *Latin-American High Performance Computing Conference (CARLA)*, pages 86–100, 2021.

[8] I. Farmaga, P. Shmigelskyi, P. Spiewak, and L. Ciupinski. Evaluation of computational complexity of finite element analysis. In *11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, pages 213–214, 2011.

[9] R. T. Guiraldello, R. F. Ausas, F. S. Sousa, F. Pereira, and G. C. Buscaglia. The multiscale Robin coupled method for flows in porous media. *Journal of Computational Physics*, 355:1–21, 2018.

[10] C. Harder, D. Paredes, and F. Valentin. A family of Multiscale Hybrid-Mixed finite element methods for the Darcy equation with rough coefficients. *Journal of Computational Physics*, 245:107–130, 2013.

[11] C. Harder, D. Paredes, and F. Valentin. On a multiscale hybrid-mixed method for advective-reactive dominated problems with heterogeneous coefficients. *Multiscale Modeling & Simulation*, 13(2):491–518, 2015.

[12] D. N. Hieu, T. Tieu Minh, T. Van Quang, B. X. Giang, and T. Van Hoai. A machine learning-based approach for predicting the execution time of CFD applications on cloud computing environment. In *Future Data and Sec. Eng.*, pages 40–52, 2016.

[13] L. Huang, J. Jia, B. Yu, B.-g. Chun, P. Maniatis, and M. Naik. Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression. In *Advances in Neural Information Processing Systems 23*, pages 883–891. 2010.

[14] S. Kim, Y. Suh, and J. Kim. EXTES: An Execution-Time Estimation Scheme for Efficient Computational Science and Engineering Simulation via Machine Learning. *IEEE Access*, 7:98993–99002, 2019.

[15] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown. Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. *Journal of Machine Learning Research*, 18(25):1–5, 2017.

[16] A. Matsunaga and J. A. B. Fortes. On the Use of Machine Learning to Predict the Time and Resources Consumed by Applications. In *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pages 495–504, 2010.

[17] D. Paredes, F. Valentin, and H. M. Versieux. On the robustness of multiscale hybrid-mixed methods. *Mathematics of Computation*, 86(304):525–548, 2017.

[18] S. Pellegrini, J. Wang, T. Fahringer, and H. Moritsch. Optimizing MPI runtime parameter settings by using machine learning. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, pages 196–206, Berlin, Heidelberg, 2009.

[19] D. D. C. Silva, A. Paes, E. Pacitti, and D. C. Oliveira. FReeP: towards parameter recommendation in scientific workflows using preference learning. In *2018 SBC 33rd Brazilian Symposium on Databases (SBBD)*, pages 211–216, 2018.

[20] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3 edition, 2011.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Predição do tempo de vida de baterias através do modelo híbrido de Kim

Julia Dammann[1], Airam Teresa Zago Romcy Sausen[2] e Marcia de Fátima Brondani Binelo[3]

[1] *Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí/RS, Brasil*
[2] *Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí/RS, Brasil*
[3] *Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí/RS, Brasil*

Resumo

As baterias recarregáveis são capazes de fornecer energia para operar dispositivos móveis sem que eles estejam conectados a uma fonte fixa de energia, mesmo por um tempo limitado. Com os avanços tecnológicos e as inúmeras possibilidades de utilização desses dispositivos, cresce a busca por alternativas que possam atender as necessidades dos usuários. Para isso, destaca-se que o uso desses dispositivos depende do estado de carga da bateria e da necessidade de prever sua vida útil. Assim, a modelagem matemática é uma ferramenta fundamental para modelar o tempo de vida, pois simula um processo real de descarga. Nesse sentido, o objetivo deste trabalho é realizar a modelagem matemática da vida útil da bateria utilizando o modelo híbrido Kim. Em seguida, é realizado um estudo teórico nas baterias e no modelo híbrido de Kim, e a simulação de diferentes correntes de descarga é comparada, mostrando que o híbrido Kim tem boa precisão para correntes de descarga baixas, médias e altas.

Palavras-chave: Modelagem Matemática, Baterias, Modelos Híbridos de Kim.

1 INTRODUÇÃO

A fonte de energia de diversos dispositivos são as baterias recarregáveis, permitindo que o dispositivo seja utilizado sem depender de uma fonte externa de energia, mesmo que por um período finito de tempo. Pesquisas sobre esse tema tem sido desenvolvidas, principalmente sobre os fatores que influenciam no desempenho da bateria durante o seu uso, sendo eles, por exemplo, a capacidade e o número de ciclos [2].

Existem diferentes tecnologias de baterias sendo utilizadas no mercado, mas o objeto de estudo desse trabalho são as de Lítio Íon Polímero (Li-Po), usadas em telefones celulares do tipo *smartphone* pois acabam se destacando das demais e vêm sendo muito empregadas,

Contato: Julia Dammann, julia_dammann@hotmail.com

devido a sua densidade de energia ser alta, possuir massa reduzida e longa vida útil [1]. Muitos dispositivos apresentam uma demanda maior de energia armazenada [3], pois passam por várias atualizações e aperfeiçoamentos, assim como cresce o uso desses dispositivos em diferentes atividades.

Um fator muito importante para o desempenho e uso dessas baterias, é o número de ciclos, pois eles caracterizam a quantidade de descargas e recargas que a bateria suporta antes que sua capacidade comece a reduzir significativamente [5]. Destaca-se, ainda, que essa redução de capacidade afeta a conduta de armazenamento e utilização da bateria [8]. Outro aspecto relevante é o tempo de vida das baterias, uma vez que representa o tempo que a mesma leva para alcançar o nível mínimo de carga para o funcionamento do dispositivo, o chamado nível de *cutoff*, impossibilitando-a de fornecer energia elétrica ao sistema. Além disso, o tempo de vida tem o papel de apresentar aos usuários o tempo que o dispositivo ficará operacional, sem a necessidade de efetuar uma recarga.

Para calcular o tempo de vida das baterias, assim como avaliar outros aspectos importantes, recorre-se à modelagem matemática, pois através de diferentes modelos é possível simular o processo de descarga real e capturar as características não lineares do seu funcionamento. Os modelos matemáticos de descarga de baterias mais utilizados na literatura são: elétricos, eletroquímicos, estocásticos, analíticos, Modelos via Teoria de Identificação de Sistemas e os modelos híbridos, estes modelos capturam as características reais de operação da bateria e podem ser utilizados para prever o comportamento da mesma, sob várias condições de carga e descarga.

Dentre os modelos presentes na literatura, nesse artigo serão apresentados os resultados das simulações do modelo híbrido de Kim, considerando ensaios experimentais de baterias do tipo Li-Po a partir de uma plataforma de teste. O modelo híbrido de Kim é formados pela união de um modelo elétrico para Predizer Runtime e Característiva V-I [4] com o modelo analítico de KiBaM [10], a união desses dois modelos agregam as vantagens de ambos, fornecendo maiores informações sobre o processo de descarga.

2 MATERIAL E MÉTODOS

Os procedimentos metodológicos adotados seguem três etapas principais. Em um primeiro momento realiza-se um estudo teórico acerca da descarga de baterias e do modelo híbrido de Kim, por meio de uma revisão da literatura técnica. Num segundo momento apresenta-se os dados experimentais através de gráficos, que são resultados da descarga de oito baterias de Li-Po do modelo PL 383562 2C com capacidade de nominal de 800 mA e tesão nominal de 4,2 V. Essas baterias foram carregadas até alcançarem 4,2 V, e descarregadas com correntes constantes, em três momentos distintos, utilizando primeiro uma corrente de 175 mA, depois 400 mA e por fim 700 mA, durante as descargas a tensão foi registrada por um sistema de aquisição de dados. No terceiro e último momento, os dados experimentais são comparados às curvas de descarga obtidas pela simulação do modelo híbrido de Kim, por meio da ferramenta Simulink/Matlab, considerando os parâmetros de Gomes [7] para a parte analítica, analisando também os resultados com as diferentes correntes.

3 MODELAGEM MATEMÁTICA

Os modelos híbridos reúnem as vantagens de dois modelos, uma vez que representam a união de um modelo elétrico e um modelo analítico. Pode-se destacar, dentro da categoria dos modelos híbridos: o modelo de Kim [9], que por compreenderem as vantagens de dois modelos, possibilitam simulações acuradas que informam o processo de descarga. Logo, pode-se obter modelos matemáticos que consideram os principais efeitos não lineares presentes em um processo de descarga real, como também registram as características elétricas do sistema [11].

O foco desse estudo é o modelo híbrido de Kim, utilizado para predizer o tempo de vida de baterias, caracteriza-se por uma união entre o modelo elétrico para Predizer Runtime e Características V-I com o modelo analítico KiBaM.

Neste caso, o modelo elétrico foi escolhido por fornecer todas as características dinâmicas do circuito da bateria como a tensão de circuito aberto, a tensão terminal e a resposta transiente. Já o modelo analítico se destaca por capturar os efeitos não lineares do processo de descarga, (i.e., efeito de recuperação e efeito taxa de capacidade) que não são capturados pelo modelo elétrico [6]. O desdobramento do modelo híbrido ocorre através da substituição dos componentes responsáveis pelo estado de carga e o tempo de vida da bateria, no modelo elétrico e pelas equações que constituem o modelo analítico KiBaM.

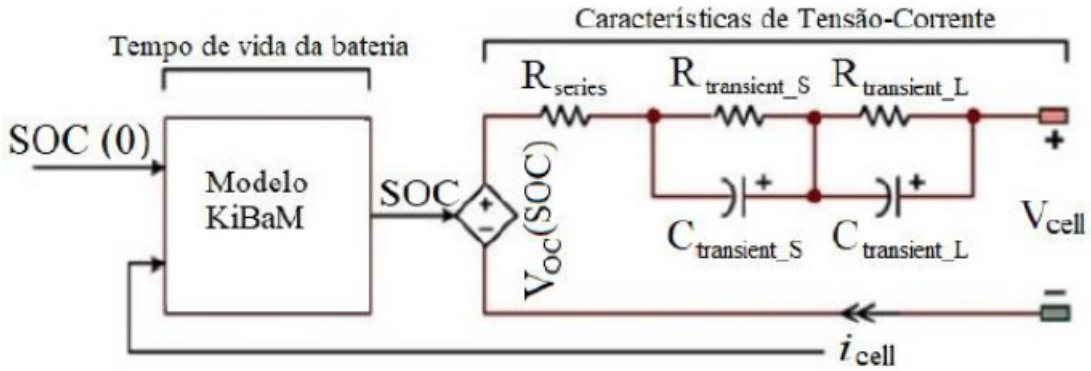


Fig. 1: Esquema do Modelo Híbrido de Kim [9]

O estado de carga (SOC) pode ser descrito pela seguinte equação:

$$SOC(t) = \frac{C_{available}(t)}{C_{max}}, \tag{1}$$

em que: $C_{available}(t)$ representa a capacidade disponível e C_{max} a nominal da bateria. A capacidade disponível da bateria, $C_{available}(t)$, é determinada por:

$$C_{available}(t) = C_{initial} - l(t) - C_{unavailable}(t), \tag{2}$$

em que: $C_{initial}$ é a capacidade inicial da bateria e $l(t)$ é a carga total consumida pelo sistema, dada por:

$$l(t) = \int i_{c_{ell}}(t)dt \tag{3}$$

A capacidade indisponível no tempo t é representada por $C_{unavailable}(t)$ e pode ser descrita pela carga indisponível $u(t)$, oriunda da substituição das equações do modelo de KiBaM, obtendo-se:

$$u(t) = \begin{cases} (1-c)[\delta(t_0)e^{-k(t-t_0)} + \frac{I}{c}.\frac{1-e^{(t-t_0)}}{k'}], t_0 < t < t_d \\ (1-c)\delta(t_d)e^{(-k'(t-t_d))}, t_d < t < t_r \end{cases}, \quad (4)$$

em que: c é a fração da capacidade total da bateria C, δt_0 é a diferença entre as alturas das fontes do modelo KiBaM no início da descarga, $k^{'}$ é a constante que representa a taxa de difusão de energia entre as fontes de cargas, $\delta(t_d)$ é a diferença entre as alturas das fontes do modelo KiBaM no tempo final de descarga, I é a corrente de descarga, t_0 é o tempo inicial, $\delta(t_d)$ é o tempo final de descarga, e t_r é o tempo que resta para terminar o período.

Durante o processo de descarga, ou seja, no intervalo $t_0 < t < t_d$, a carga indisponível $u(t)$ aumenta, o que representa o efeito taxa de capacidade. No intervalo $t_d < t < t_r$, a carga indisponível $u(t)$ diminui, porque a carga da fonte limitada flui para a fonte de carga disponível representando assim o efeito de recuperação. Com isso, $C_{unavailable}(t)$ também pode ser representado por:

$$C_{unavailable}(t) = \begin{cases} C_{unavailable}(t_0)e^{-k(t-t_0)} + (1-c)\frac{I}{c}.\frac{1-e^{(t-t_0)}}{k'}, t_0 < t < t_d \\ C_{unavailable}(t_d)e^{(-k'(t-t_d))}, t_d < t < t_r \end{cases}, \quad (5)$$

onde: $C_{unavailable(t_0)}$ é a capacidade indisponível da bateria no início da descarga, e $C_{unavailable}(t_d)$ é a capacidade indisponível da bateria no final do tempo de descarga.

Portanto, o estado de carga (SOC) é dado por:

$$SOC(t) = SOC_{initial} - \frac{1}{C_{max}}[\int [i_{cell}(t)dt + C_{unavailable}(t)]], \quad (6)$$

onde: $SOC_{initial}$ é o SOC estimado antes de t_0, $i_{cell}(t)$ é a corrente de descarga, e $C_{unavailable}(t)$ é a capacidade indisponível da bateria que é oriunda do modelo KiBaM. A tensão desse modelo pode ser expressa por:

$$V_{cell}(t) = VOC[SOC(t)] - i_{cell}(t).R_{series}.V_{transient}(t), \quad (7)$$

onde: $V_{cell}(t)$ é a tensão, $VOC[SOC(t)]$ é a tensão de circuito aberto, R_{series} é a resistência em série, e $V_{transient}(t)$ é a tensão transiente e são obtidos através das seguintes equações:

$$V_{oc}[SOC(t)] = a_0e^{-a_1[SOC(t)]} + a_2 + a_3[SOC(t)] - a_4[SOC(t)]^2 + a_5[SOC(t)]^3, \quad (8)$$

$$R_{series}[SOC(t)] = b_0e^{-b_1[SOC(t)]} + b_2 + b_3[SOC(t)] - b_4[SOC(t)]^2 + b_5[SOC(t)]^3, \quad (9)$$

$$V_{transient}(t) = V_{transientS}(t) + V_{transientL}(t), \quad (10)$$

onde: $V_{transientS}(t)$ é a tensão transiente de curta duração, e $V_{transientL}(t)$ é a tensão de longa duração, dadas pelas seguintes equações respectivamente:

$$V_{transientS}(t) = \begin{cases} R_{transientS}.i_{cell}(t)1 - e^{-\frac{(t-t_0)}{\tau S}}, t_0 < t < t_d \\ V_{transientS}(t_d)e^{-\frac{(t-t_d)}{\tau S}}, t_d < t < t_r \end{cases}, \tag{11}$$

onde: $R_{transientS}$ é a resistência transiente de curta duração, $V_{transientS}(t_d)$ é a tensão transiente de curta duração no final da descarga, e S é o produto entre $R_{transientS}$ e $C_{transientS}$, que representa a capacitância transiente de curta duração

$$V_{transientL}(t) = \begin{cases} R_{transientL}.i_{cell}(t)1 - e^{-\frac{(t-t_0)}{\tau S}}, t_0 < t < t_d \\ V_{transientL}(t_d)e^{-\frac{(t-t_d)}{\tau S}}, t_d < t < t_r \end{cases}, \tag{12}$$

em que: $R_{transientL}$ é a resistência transiente de longa duração, $V_{transientL}(t_d)$ é a tensão transiente de longa duração no tempo final da descarga e L o produto entre $R_{transientL}$ e $C_{transientL}$, que é a capacitância transiente de longa duração. Os parâmetros que modelam a tensão transiente são funções do SOC e dadas pelas quatro equações abaixo:

$$R_{transientS}[SOC(t)] = c_0 e^{-c1[SOC(t)]} + c_2, \tag{13}$$

$$C_{transientS}[SOC(t)] = d_0 e^{-d1[SOC(t)]} + d_2, \tag{14}$$

$$R_{transientL}[SOC(t)] = e_0 e^{-e1[SOC(t)]} + e_2, \tag{15}$$

$$C_{transientL}[SOC(t)] = f_0 e^{-f1[SOC(t)]} + f_2. \tag{16}$$

onde: $a_{0..5}$, $b_{0..5}$, $c_{0..2}$, $d_{0..2}$, $e_{0..2}$ e $f_{0..2}$ são os coeficientes das funções,ou seja, os parâmetros do modelo que precisam ser estimados e são apresentados na tabela 1.

Tabela 1: Parâmetros do modelo

Coeficiente	Valor
a_0	-0.852
a_1	63.867
a_2	3.6297
a_3	0.559
a_4	0.51
a_5	0.508
b_0	0.1463
b_1	30.27
b_2	0.1037
b_3	0.0584
b_4	0.1747
b_5	0.1288
c_0	0.1063
c_1	62.49
c_2	0.0437
d_0	-200
d_1	138
d_2	300
e_0	0.0712
e_1	61.4
e_2	0.0288
f_0	-3083
f_1	180
f_2	5088

Tais parâmetros foram obtidos por meio do ajuste de curvas realizado na ferramenta *curve fitting* do Matlab. Os dados experimentais são importados para o software em formato de matriz e depois para a ferramenta, aí então encontrou-se o modelo que melhor se ajustou aos dados, com os coeficientes acima listados.

4 RESULTADOS E DISCUSSÕES

Para os resultados apresentados nesta seção foram considerados oito experimentos de descarga de bateria de Li-Po, com a aplicação de três correntes: 175 mA, 400 mA e 700 mA. Também foi realizada a simulação da descarga pelo modelo híbrido de Kim.

Nas Figuras 2, 3 e 4, as curvas em vermelho representam as curvas experimentais, a curva em azul refere-se à simulação, a curva verde é a mais próxima do tempo de vida experimental médio e o círculo mostra o tempo de vida experimental médio.

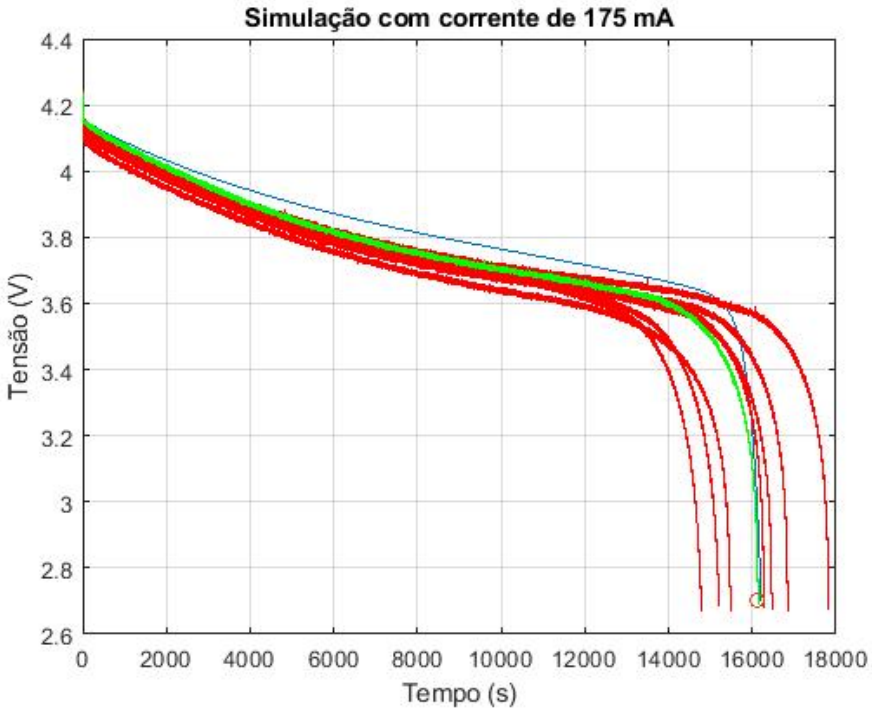


Fig. 2: Simulação com corrente de descarga de 175 mA

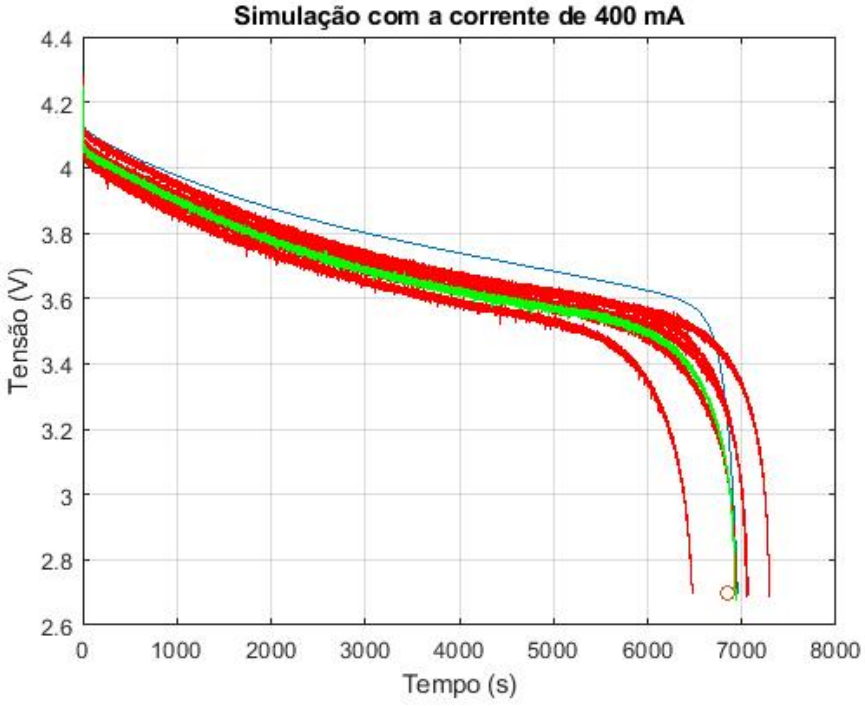


Fig. 3: Simulação com corrente de descarga de 400 mA

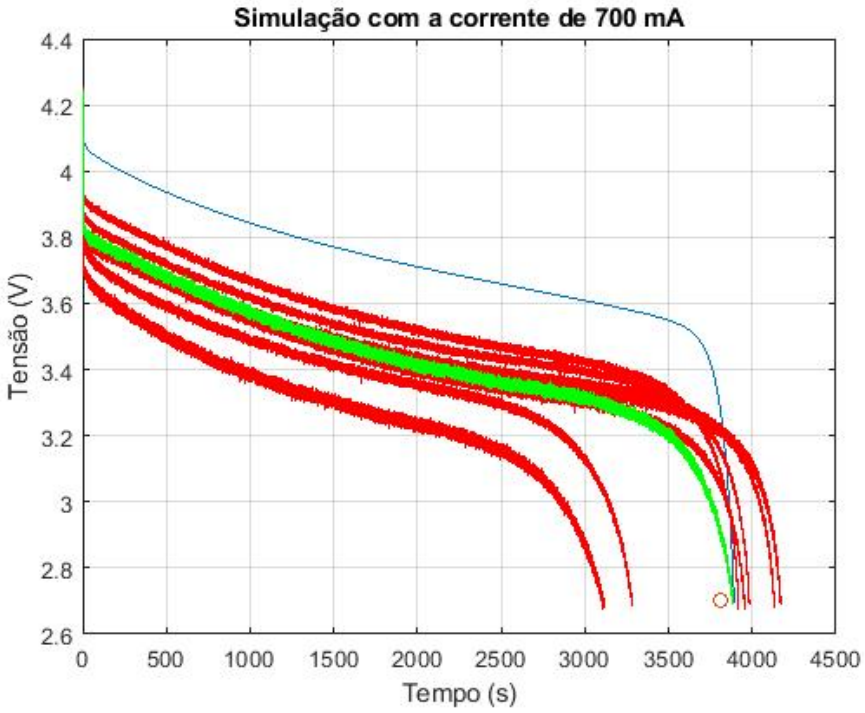


Fig. 4: Simulação com corrente de descarga de 700 mA

Comparando as simulações de diferentes correntes de descarga é possível perceber que quanto menor a corrente aplicada, maior o tempo de vida. Desse modo, quando se aplica a corrente de 175 mA, o tempo de vida simulado é de 16213 s e o tempo de vida experimental médio é de 16127 s, resultando em um erro relativo percentual de 0,53%.

Já na corrente intermediária de 400 mA, o tempo de vida simulado é de 6954 s e o tempo de vida experimental médio é de 6875 s, apresentando um erro relativo percentual de 1,15%.

Por fim, com a corrente mais alta, de 700 mA, tem-se o tempo de vida simulado de 3892 s e o tempo de vida experimental médio de 3811 s, obtendo um erro relativo percentual de 2,13%.

5 CONSIDERAÇÕES FINAIS

Os usuários de dispositivos móveis vêm exigindo cada dia mais recursos de seus aparelhos, gerando a necessidade de baterias recarregáveis com um tempo de vida maior. Desse modo, as pesquisam acerca do tema são relevantes e cada vez mais frequentes, objetivando explorar tecnologias para atender essas necessidades.

Buscou-se investigar a predição do tempo de vida das baterias do tipo Li-Po, por meio do modelo híbrido de Kim, que se destaca por reunir as vantagens do modelo elétrico de Chen e Rincón-Mora e as do modelo analítico de KibaM. Nesse caso, objetivou-se apresentar uma comparação das simulações com diferentes correntes de descarga constantes: 175 mA, 400 mA e 700 mA, aplicadas separadamente.

É possível visualizar que modelo híbrido de Kim tem boa acurácia tanto para correntes baixas, como médias e altas. Com a corrente de descarga constante de 175 mA obteve-se

um erro relativo percentual de 0,53%, com a corrente de 400 mA o erro foi de 1,15%, e a corrente de 700 mA, apresentou erro de 2,13%.

Destaca-se, ainda, que é possível realizar mais estudos considerando outros efeitos importantes, como a diminuição da capacidade da bateria devido ao número de ciclos de carga e descarga.

6 *Agradecimentos*

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pelo financiamento da pesquisa.

Ao Grupo de Automação Industrial e Controle - GAIC.

Referências

[1] T. Ashwin, Y. M. Chung, and J. Wang. Capacity fade modelling of lithium-ion battery under cyclic loading conditions. *Journal of Power Sources*, 328:586–598, 2016.

[2] M. F. B. Binelo, A. T. Z. R. Sausen, and P. Sausen. Método multi-fase de estimação e adaptação de parâmetros de modelos elétricos para a predição do tempo de vida de baterias. 2019.

[3] M. Broussely, S. Herreyre, P. Biensan, P. Kasztejna, K. Nechev, and R. Staniewicz. Aging mechanism in li ion cells and calendar life predictions. *Journal of Power Sources*, 97:13–21, 2001.

[4] M. Chen and G. A. Rincon-Mora. Accurate electrical battery model capable of predicting runtime and iv performance. *IEEE transactions on energy conversion*, 21(2):504–511, 2006.

[5] C. Clemm, C. Sinai, C. Ferkinghoff, N. Dethlefs, N. F. Nissen, and K.-D. Lang. Durability and cycle frequency of smartphone and tablet lithium-ion batteries in the field. In *2016 Electronics Goes Green 2016+(EGG)*, pages 1–7. IEEE, 2016.

[6] K. P. Duarte. Aplicação de um modelo híbrido para predição do tempo de vida de baterias utilizadas em dispositivos móveis. 2014.

[7] L. B. Gomes. Proposição de um modelo híbrido considerando a lei de peukert estendida para a predição do tempo de vida de baterias. 2017.

[8] M. Jafari, K. Khan, and L. Gauchia. Deterministic models of li-ion battery aging: It is a matter of scale. *Journal of Energy Storage*, 20:67–77, 2018.

[9] T. Kim and W. Qiao. A hybrid battery model capable of capturing dynamic circuit characteristics and nonlinear capacity effects. *IEEE Transactions on Energy Conversion*, 26(4):1172–1180, 2011.

[10] J. F. Manwell and J. G. McGowan. Lead acid battery storage model for hybrid energy systems. *Solar energy*, 50(5):399–405, 1993.

[11] J. Zhang, S. Ci, H. Sharif, and M. Alahmad. An enhanced circuit-based model for single-cell battery. In *2010 Twenty-Fifth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, pages 672–675. IEEE, 2010.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Algoritmos para Classificação Estrutural de Proteínas

Lúcio Paccori Lima[1] and Dr. Marcos Augusto dos Santos[2]

[1] *UFMG, Belo Horizonte/MG, Brazil*
[2] *UFMG, DCC, Belo Horizonte /MG, Brasil*

Abstract

Neste trabalho, estruturas de proteínas são mapeadas em um *vector space model* para sua manipulação em ambientes que remetem às modernas maquinas de busca (Yahoo, Google etc). Instrumentos da álgebra linear junto com a regressão logística com modificações esteiam um novo algoritmo, aqui denominado **Adhoc**, para a busca de assinaturas de estruturas tridimensionais. Este método transcende a aplicação aqui discutida; pode ser utilizado de forma geral em outros problemas de recuperação de informação. Para a validação, testamos as ideias para classificar proteínas em seus respectivos grupos e famílias, a partir da estrutura já determinada. Os resultados apresentados são animadores e outros experimentos estão sendo providenciados.

Keywords: Bioinformática, Regressão Logística, SVM

1 INTRODUÇÃO

Este trabalho contempla o uso de algoritmos de classificação para mineração de dados de estruturas proteicas e suas assinaturas. Uma assinatura, na acepção aqui considerada, é uma marca que comprova um conjunto único de atributos para identificar uma entidade (proteína ou parte da proteína) como membro de um conjunto previamente escolhido.

Existem ótimos métodos para classificar e recuperar proteínas em suas respectivas super famílias, famílias e subgrupos diversos. A partir da descrição das proteínas por um *vector space model* (VSM), onde cada entidade é representada por um conjunto de tamanho fixo de atributos numéricos, são aplicados algoritmos tradicionais, a saber, redes neurais, pesquisa em árvores, *vector support machine* e outros [Vapkin et al., 1999], [Frank et al,. 2004], [Wang et al., 2005] e [Ma et al., 2014]. São capazes de predizer com eficiência a que grupo uma dada proteína pertence. Ao que nos consta, em [Pires DE, de Melo-Minardi

Contato: Lúcio, paccori@uol.com.br

RC, dos Santos MA, Santoro MM, et al., 2012] tem-se os melhores resultados, onde são apresentados resultados que permitem construir oráculos perfeitos (considerados por nós quando a média harmônica obtida de validações cruzadas são superiores à 0,95).

Entretanto, os artefatos de mineração citados no paragrafo acima não são, sabidamente, adequados para escolher quais os atributos (e seus valores) que contribuem para um determinado indivíduo estar em um certo grupo. Este processo de escolha de atributos (*feature selection*) configura uma área em aberto na ciência da computação.

Um método interessante que tem como resultado colateral uma avaliação de tais atributos, é a regressão logística [Kleinbaum et al., 2002], [Mesquita, 2014]; uma técnica bem conhecida e muito utilizada na área médica [Fort, et al., 2005], [Hosmer et al., 2000]. Mas a sua utilização genérica como instrumento de classificação traz uma série de inconvenientes; talvez seja esta razão por ele não estar precisamente contemplado na taxonomia do processo de mineração de dados. Nem sempre nos problemas reais tem-se um domínio com equilíbrio adequado entre a quantidade (e variedade) dos membros do grupo que se deseja classificar e a quantidade (e variedade) da população total. Nos próximos trabalhos, estaremos usando superfamílias com 27000 entidades ao lado de outras com apenas algumas dezenas. Estas e outras inconveniências são resolvidas nos algoritmos que estamos propondo e/ou adaptando.

A justificativa e muito da motivação inicial para este trabalho, era encarar um fato que ocorre na natureza. É sabido que às proteínas com sequências similares, correspondem estruturas tridimensionais também similares [Leach, 2001]. Entretanto, existem casos em que estruturas similares não guardam similaridade quanto às sequências primárias. Esperamos encontrar assinaturas que sejam invariantes quanto a este aspecto.

No desenvolvimento do trabalho, acabamos por ter a necessidade de buscar por oráculos perfeitos. Este é um dos principais aspectos tratados neste trabalho.

2 METODOLOGIA

2.1 Conjunto de dados

Os testes foram realizadas no conjunto de superfamílias de enzimas, tidas como um padrão ouro, que utilizam mecanismos distintos para executar suas funções [Brown et al., 2006]. Neste conjunto são consideradas seis superfamílias (amdohydrolase, crotonase, haloacid dehalogenase, isoprenoind synthase type I e vicinal oxygen chelate), compreendendo 47 famílias distribuídas em 896 diferentes cadeias.

2.2 Decomposição por valor singular

A decomposição por valores singulares ou *singular value decomposition* (**svd**) [Eldén, 2006, 2007; Berry, 1995] é uma técnica da álgebra linear utilizada para reduzir a dimensionabilidade das entidades sem comprometer a sua essência. Com o **svd** é possível, ao invés de trabalhar com uma matriz com o posto aproximado, simplesmente usar a combinação linear dos padrões presentes na matriz que aproximam o posto. Em geral, para o propósito de recuperar informação em problemas reais, não são necessários muitos padrões, implicando em uma representação mais econômica.

Mais formalmente, a decomposição por valores singulares é uma fatoração de uma matriz qualquer em três outras matrizes com propriedades importantes. Possui várias

aplicações, tanto diretas, nas quais se aplicam os resultados extraídos de suas matrizes fatores, quanto como um passo em muitos algoritmos.

Definição 1. *Dado $A \in \mathbb{R}^{m\times n}$, não necessariamente de posto completo, a decomposição por valores singulares de A é uma fatoração tal que:*

- $A = USV^T$;
- $U \in \mathbb{R}^{m\times m}$, *são os autovetores de* AA^T *e é ortogonal;*
- $V \in \mathbb{R}^{n\times n}$, *são os autovetores de* A^TA *e é ortogonal;*
- $S \in \mathbb{R}^{m\times n}$, *é diagonal se* $m = n$, *caso contrário adiciona-se* $m - n$ *linhas de zeros em* S *e é formado por a raiz quadrada dos autovalores de* AA^T.

$$S = diag(\sigma_1, \sigma_2, .., \sigma_n)$$

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

Os valores $(\sigma_1, \sigma_2, ...\sigma_n)$ são chamados de valores singulares de A. As colunas de U, $(u_1, u_2, ...u_m)$ são chamadas de vetores singulares à esquerda de A e as colunas da matriz V, $(v_1, v_2, ..., v_n)$ são os vetores singulares à direita de A. As colunas de U podem ser interpretadas como padrões das entidade em A com um peso correspondente σ_i; já V são os padrões das linhas, que têm o mesmo peso associado.

Estamos utilizando o *svd* neste trabalho com três propósitos. Primeiro é uma ferramenta que possibilita a visualização de conjuntos de dados no espaço tridimensional [Marcolino LC, Couto BRGM, dos Santos MA, 2010], o que nos permite ter uma avaliação inicial das eventuais dificuldades no processo de classificação. Outra aplicação vem do fato de que nas melhorias que estamos propondo nos algoritmos, temos de fazer escolhas e recuperar um subconjunto de entidades que são mais próximas a uma dada consulta; usando as técnicas de máquina de busca, conseguimos melhores resultados. Finalmente, usamos diretamente os valores singulares como **VSM**, como será usado futuramente.

2.2.1 Representação de entidades no contexto de estruturas de proteínas

Neste trabalho estamos utilizando a representação de entidades usando *o cutoff scanning* que foi proposto por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., (2012)]; deu origem ao melhor, ou um dos melhores classificadores para esta classe de problemas.

Representação de entidades usando o *cutoff scanning* .

Neste **VSM**, a proteína é descrita a partir do número de átomos que existe em subintervalos de 0,2 ångströms até uma distância máxima de 30Å. Foi usado por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011]; onde são mostrados resultados excepcionais para esta classe de problemas de classificação.

2.3 Regressão Logística

A regressão logística consiste em encontrar valores de $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ para ajustar a equação (1) para cada uma das entidades j. O valor da função, conhecida como *logit*, $P_j(x)$, com $P_j(x) \in [0, 1]$ informa a probabilidade da uma dada entidade ser classificada como pertencente a um subconjunto específico.

$$P_j(x) = \frac{e^{\sum_{i=1}^{m} \alpha_i x_i}}{1 + e^{\sum_{i=1}^{m} \alpha_i x_i}} \tag{1}$$

Observa-se que quando $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para zero, $P_j(x)$ também tende para zero. Por outro lado, se $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para infinito, $P_j(x)$ aproxima-se da unidade.

Para proceder a determinação dos valores de α, faz-se uma transformação linear que remeterá o problema à solução de um problema de álgebra linear. Seja a chance $C_j(x)$ é definido por:

$$C_j(x) = \frac{P_j(x)}{1 - P_j(x)}. \tag{2}$$

A expressão da equação (2) usando (1), temos

$$C_j(x) = e^{\sum_{i=1}^{m} \alpha_i x_i} \tag{3}$$

Tomando-se o logaritmo em ambos os lados de (3), obtemos um sistema de equações lineares para determinar α:

$$b_j = \sum_{i=1}^{n} \alpha_i x_i \tag{4}$$

onde $b_j = \log(C_j)$, para j=1,2,...m.

Associamos ao conjunto de dados uma matriz $A = \{a_{i,j}\} \in \mathbb{R}^{m \times n}$. As linhas representam as entidades e às colunas estão associados os atributos. Assim, o valor de cada $a_{i,j}$ é o valor do atributo j na proteína i. Observar que isto é diferente da forma que tratamos anteriormente, quando discutimos a decomposição por valores singulares (esta matriz é a transposta dessa anterior).

Seja $b = (b_1, b_2, ..., b_m)^T$; o sistema de equações lineares (4) pode ser representado por:

$$A\alpha = b \tag{5}$$

Em (5), cumpre observar que quando o número de equações é inferior ao número de incógnitas, a solução do sistema, fornecida pelo método da minimização da somatória dos quadrados dos resíduos, é indeterminado (infinitas de soluções). Em outras palavras, a abordagem clássica em álgebra linear é minimizar $||A\alpha - b||^2$, o que requer o posto completo de $A^T A$ - uma propriedade não é esperada das matrizes A nos cenários em que estamos trabalhando. Geralmente, para contornar essa dificuldade descarta-se um grupo das variáveis, mantendo apenas um subconjunto das variáveis originais. Este procedimento é chamado de *feature selection* em mineração de dados - uma área de pesquisa em aberto. Nós usamos um termo estabilizador no modelo de regressão logística, encontrado nos trabalhos de [Morais, Rodrigues, F., et al., 2020], que permite a atribuição de valores para os parâmetros α que minimizam a soma dos quadrados dos resíduos $(A\alpha - b)$,

adicionado aos quadrados de α. Assim, para encontrar uma solução a (5), resolvemos um problema otimização quadrática irrestrito dado por

$$Minimize f(\alpha) = ||\alpha||^2 + ||A\alpha - b||^2 \tag{6}$$

Como $f(\alpha)$ é função convexa, o argumento α^* que minimiza (6) é dado pela derivação de $f(\alpha)$ em α e igualando resultado a zero, que resulta em um sistema de equações lineares

$$(I + A^T A)\alpha = A^T b, \tag{7}$$

onde I é a matriz identidade da ordem n.

A solução ótima para α de (6) é obtida pela solução de (7) e é única. Então, dada uma proteína (*query*) $q = (q_1, q_2, ..., q_n)$ com n atributos, a probabilidade de q pertencer a uma classe associada ao sistema relacionado é dado por:

$$P(q) = \frac{e^{q\alpha}}{1 + e^{q\alpha}}. \tag{8}$$

Saliente-se que (7) é para o caso em que o número de incógnitas é maior que o número de equações, que geralmente ocorre com o método **Adhoc**, visto a seguir.

2.3.1 Método Adhoc

Adhoc é uma expressão latina cuja tradução literal é "para isto" ou "para esta finalidade". É usada para designar algo ou alguma coisa que foi formada ou usada para um propósito ou necessidade particular e imediata, sem planejamento prévio.

A metodologia descrita a seguir, aqui intitulada **Adhoc**, usa modelos construídos exclusivamente em resposta a uma única demanda por classificação. Dada uma consulta q, um modelo **Adhoc**, específico para esta consulta, é construído a partir da escolha de k_0 entidades mais próximas à q tais que $P_j(x) = 0$, ao lado de k_1 entidades escolhidas dentre aquelas mais próximas à q tal que $P_j(x) = 1$.

Os valores de k_0 e k_1 são determinados experimentalmente. Nos nossos ensaios constatamos que estes valores são baixos; na validação cruzada verificamos que o desempenho da classificação é superior quando atribuímos valores na ordem de algumas unidades a estes dois parâmetros. Isto tem como consequência a construção de matrizes de atributos $M \in \mathbb{R}^{m \times n}$ nas quais $m < n$ o que impede o uso da regressão logística tradicional. Resolvemos esta limitação usando a regressão logística modificada segundo explicado na seção anterior.

Outra dificuldade está na escolha das entidades $k_{0/1}$ mais próximas, o que poderia impactar o tempo de resposta nos problemas de grande porte. Isto é resolvido organizando as entidades em árvores de pesquisa segundo os recursos usados em máquinas de busca (**svd**, clusterização etc). A matriz de atributos original A é particionada em duas outras, A_0 e A_1, segundo $P_j(x) = 1/0$; com cada uma delas organizada enquanto elementos de uma máquina de busca. Assim, a recuperação das entidades mais próximas a uma consulta q fica extremamente eficiente e não impacta o tempo de processamento de forma perceptível.

Concluindo, o método **Adhoc** é um recurso para a regressão logística que permite a sua aplicação em diferentes cenários. A tabela 1 apresenta o algoritmo utilizado para o método **Adhoc.**

Algoritmo 1: Método Adhoc

Entrada: q : consulta q, A1, A0
Saída: P: Probabilidade de q pertencer ao grupo 1
A1: Conjunto de entidades que pertencem a categoria 1;
A0: Conjunto de entidades que não pertencem a categoria 0;
M1 ← Conjunto dos k_1 elementos mais próximos de q em A1;
M0 ← Conjunto dos k_0 elementos mais próximos de q em A0;
M ← [M1; M0];
$i_1 \leftarrow 1$, $\forall$ elemento de M1;
$i_0 \leftarrow 0$, $\forall$ elemento de M0;
i ← $[i_1; i_0]$;
P ← Aplique a regressão logística modificada para M e i

Tabela 1: Regressão logística para ambientes desbalanceados

3 RESULTADOS E DISCUSSÃO

Nesta seção apresentamos a experiência computacional usando o conjuntos de dados que congrega entidades sob aspectos funcionais (*golden standard*). Para todos os ensaios usamos cristais depositados no PDB para extrair as coordenadas dos átomos para os classificadores.

Os átomos considerados neste trabalho para a construção do VSM foram os da cadeia principal (ou parte dela), constituída, seja pelos C_α, seja pelos átomos (C_α, C, N). Mas também, em alguns casos, utilizamos todos os átomos da proteína. Entretanto, não houve grandes alterações nos resultados (não exibidos no texto).

Planejamos os experimentos para inicialmente responder uma primeira pergunta; aquela que tange a adequação dos nossos algoritmos à classificação de estruturas de proteínas. Os métodos que usamos são interessantes para avaliar atributos e, em outras aplicações, como por exemplo, aquelas que utilizam *microarrays*, o efeito colateral que no caso é identificar marcadores biológicos de uma doença, tem resultado em contribuições interessantes [Abreu A.P. (2019)], [Santos A. dos, (2017)], [Morais, Rodrigues, F., et al., 2020]. Assim, passamos a testar os métodos descritos na seção anterior para seguir à busca de assinaturas estruturais para atender ao nosso objetivo.

No que se segue, apresentamos os resultados da regressão logística em um conjunto, um que se apresenta muito bem balanceado (*golden standard*). Com os resultados que alcançamos, buscamos identificar as assinaturas usando como VSM o *cutoff scanning* , e obtivemos resultados muito próximos ao que consideramos como oráculo perfeito (média harmônica na validação cruzada acima de 0,95).

3.1 Validação da regressão logística como instrumento para classificação estrutural

Neste item estão os resultados com modelos construídos com *cutoff scanning* [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011], no conjunto *golden standard* cuja característica que mais nos chama a atenção é o bom balanceamento entre as entidades em todos os conjuntos. Na referência citada, os melhores resultados foram com a matriz de distâncias entre os carbonos alfa C_α da proteína em que o posto foi reduzido usando **svd**. Vários métodos tradicionais de mineração de dados foram utilizados.

Nossos melhores resultados foram alcançados usando a regressão tradicional com as

coordenadas dos átomos da cadeia principal - ver tabela 2. Usando a regressão logística com a estrategia **Adhoc**, o desempenho foi levemente inferior - ver tabela 3. Resultados de outros experimentos usando C_α encontram-se nos anexos (tabela 5). Possivelmente, embora excepcionais, estes resultados possam ser melhorados usando como em Pires (ref. citada), a decomposição por valores singulares. Neste primeiro momento, como nos interessa avaliar atributos para entender uma possível assinatura, evitamos este recurso pois estaríamos perdendo um certo mapeamento dos atributos originais, dado que o **svd** trabalha em espaços projetados. Entretanto, tão logo seja possível, faremos isto.

Cumpre observar que tivemos dificuldade para comparar diretamente os nossos resultados com aqueles obtidos em Pires (referência citada). O desempenho dos métodos é aferido a partir de 20 *folders* aleatoriamente construídos. Não há na referência citada uma sugestão de *data set* para testes. Além disto, adotamos, por considerar mais adequado ao nosso problema, critérios de avaliação baseados na média harmônica e curva **ROC** . Entretanto, acreditamos que os nossos resultados seguem semelhantes àqueles que nos servem de referência.

Resultados da regressão logística no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9756	0,9581	0,9664	0,9597
crotonase	1,0000	0,9890	0,9944	0,9984
enolase	0,9779	0,9931	0,9850	1,0000
haloacid dehalogenase	0,9571	0,9573	0,9551	0,9587
isoprenoid synthase type I	1,0000	0,9994	0,9997	1,0000
vicinal oxygen chelate	1,0000	0,9968	0,9984	1,0000
Média	0,9851	0,9823	0,9831	0,9861

Tabela 2: Com a regressão tradicional alcançamos a média harmônica (média) 98% , sensibilidade média de 98% e uma especificidade média de 98%. As coordenadas dos átomos para a construção da matriz de distâncias foram os da cadeia principal

Resultados do método Adhoc no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9699	0,9419	0,9553	0,9794
crotonase	0,9818	0,9643	0,9727	0,9836
enolase	0,9351	0,9194	0,9266	0,9584
haloacid dehalogenase	0,9737	0,9559	0,9643	0,9831
isoprenoid synthase type I	1,0000	0,9742	0,9866	0,9926
vicinal oxygen chelate	0,9533	0,9657	0,9587	0,9734
Média	0,9690	0,9535	0,9607	0.9784

Tabela 3: Regressão logística em *data sets* bem balanceados: o método **Adhoc** alcançou 96% de média harmônica (média), sensibilidade média de 96% e uma especificidade média de 95%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos carbono alfa (C_α) da proteína.

3.2 ***Busca de assinaturas estruturais usando o VSM*** *cutoff scanning* ***com o método Adhoc***

Asim, dado que os resultados alcançados foram animadores, passamos à próxima etapa que é a definição da assinatura, escolhendo e procurando interpretar os atributos mais importantes na classificação. A regressão logística indica isto a partir dos valores de α; usam-se os valores de α_i mais positivos e os mais negativos. Por exemplo, no gráfico mostrado na figura (1), escolheríamos alguns atributos que estivessem topologicamente mais distantes do eixo das abcissas. Aqueles atributos cujos valores de α_i são próximos de zero, não têm poder discriminatório.

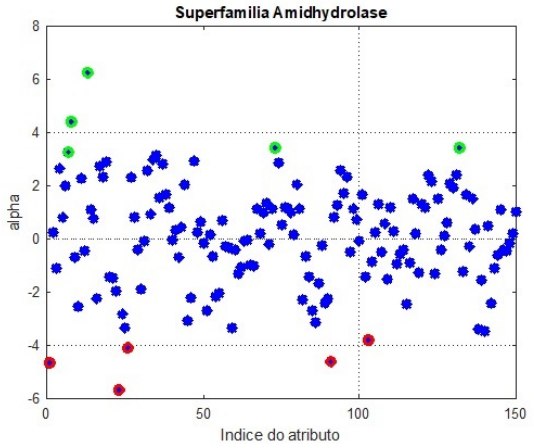


Fig. 1: Valores dos pesos (α_i) obtidos em um modelo de regressão logística, para a superfamília Amidhydrolase do conjunto de dados *gold standard* . Foram usados os átomos do carbono alfa (C_α).

Por exemplo, para o oráculo da superfamília Amidhydrolase, escolheríamos os atributos que são mostrados na tabela (4). O próximo passo consiste em validar esta escolha de atributos. Caso os oráculos construídos somente com estes atributos sejam eficientes, tem-se uma prova de conceito, como utilizamos em outros problemas da área, quando buscávamos marcadores biológicos. [Leite, et al. 2020]e [Morais R. F. et al. 2020].

Alfa	Atributo	Intervalos (ångströms)	Alfa	Atributo	Intervalos (ångströms)
Positivos	7	$(2,2\ ,\ 2,4)$	Negativos	6	$(2,0\ ,\ 2,2)$
	8	$(2,4\ ,\ 2,6)$		3	$(1,4\ ,\ 1,6)$
	4	$(1,6\ ,\ 1,8)$		2	$(1,2\ ,\ 1,4)$
	1	$(1,0\ ,\ 1,2)$		18	$(4,4\ ,\ 4,6)$
	5	$(1,8\ ,\ 2,0)$		17	$(4,2\ ,\ 4,6)$

Tabela 4: Escolha de atributos baseados nos parâmetros α_i da superfamília Amidhydrolase. Foram utilizados os átomos da cadeia inteira

4 CONCLUSSÃO

O objetivo principal deste trabalho é a busca por assinaturas de estruturas tridimensionais de proteínas que passa, necessariamente, por processos que atestem a sua pertinência.

Dentre os caminhos para alcançar este objetivo, escolhemos aqueles que remetem à mineração de dados. Imaginávamos um caminho mais fácil.

Um trabalho [Morais, R. F. et al. 2020] do grupo de pesquisa ao qual pertenço a publicado há alguns anos atrás e obteve um resultado importante. A partir de um modelo para descrever as estruturas (*cutoff scanning*) e usando as técnica usuais de mineração de dados, Pires [Pires, et al., 2012]) conseguiu classificadores que se situam entre os de melhor desempenho para esta classe de problemas. Entretanto, em função dos métodos utilizados, nada foi especulado quanto ao papel dos atributos e sua importância em cada classe em que ele foi aplicado. Imaginávamos que poderíamos utilizar nossos métodos e assim descortinar o papel dos atributos.

Os nossos métodos mostraram-se eficientes para a esta classe de problemas. Como efeito colateral, consideramos que com o método **Adhoc** conseguimos situar a regressão logística como uma alternativa eficaz aos métodos usuais de mineração de dados. Estendemos a possibilidade de usar a regressão logística em cenários desbalanceados em geral. Os resultados transcendem esta aplicação deste nosso trabalho em bioinformática e tem aplicação geral em na ciência da computação. Resolve a grande falha da regressão logística na escalabilidade da sua aplicação nos cenários reais.

Além da eficiência, cumpre chamar a atenção dos recursos computacionais que o método demanda: nada além de um microcomputador. Não são necessários servidoras e nem instalações sofisticadas para usar a metodologia.

Um efeito colateral deste nosso trabalho tem sido a melhoria dos algoritmos de classificação; não é exatamente nosso objetivo primário mas, precisamos desses métodos para mostrar a efetividade das nossas escolhas. Nossos métodos se traduzem também como uma contribuição para área de mineração de dados em ciência da computação. Gostaria de salientar que em nenhum momento este foi o nosso foco. O Laboratório de Bioinformática e Sistemas (**LBS**) ao qual estamos ligados, tem como um dos projetos o reposicionamento de fármacos. Nesta trilha, sempre aparece a necessidade de criar um domínio com as cavidades nas quais os fármacos modulam as proteínas. A dificuldade em mapear estas entidades sempre esteve presente e este nosso trabalho pode vir a ser uma resposta a este problema.

Referências

[1] Abreu, A. P. (2019) *Prospecção de Biomarcadores para Câncer de Mama Subtipo Luminal A em estágio inicial usando Álgebra Linear*. Mestrado em Bioinformáica. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

[2] Berry, M. W.; Dumais, S.T. & O'Brien, G.W.(1995). *Using linear algebra for intelligent information retrieval*. SIAM reviem, 37(4): 573-595.

[3] Brown SD1, Gerlt JA, Seffernick JL, Babbitt PC.(2006) *A gold standard set of mechanistically diverse enzyme superfamilies*. Genome Biology, 7(1):R8

[4] Fort, G. and Lambert-Lacroix, S. (2005) *Classification using partial least squares with penalized logistic regression,* Bioinformatics, 21(7): 1104-1111.

[5] Hosmer D. W.; Lemeshow , S. (2000) *Applied Logistic Regression* . H 2ad ed. New York: John Wiley & Sons.

[6] Jain P1, Hirst JD. *Automatic structure classification of small proteins using random forest.* BMC Bioinformatics. 2010 Jul 1;11:364. doi: 10.1186, 1471-2105-11-364.

[7] Kleinbaum, D. G.; Klein, M. (2002) *Logistic Regression, a Self-Learning Text,* 3ª edição, Londres-ENG, Springer, 701p.

[8] Kloczkowski A., Jernigan R. L., Wu Z., Song G., Yang L., Kolinski A., Pokarowski P.(2009)*Distance matrix-based approach to protein structure prediction.* J Struct Funct Genomics. 2009 Mar;10(1):67-81.

[9] Landwehr, N. Hall, M. & Frank, E (2005). *Logistic model trees.* Machine Learning, 59(1-2): 161-2015.

[10] Leach A. R. (2001) *Molecular Modelling: Principles and Applications.* Prentice Hall; 2nd ed. edição (2001)

[11] Leite, C. F. V. et al.*Milk-Way algorithm for ligand-based virtual screening: CDK2 case study. Trends in Developmental Biology*, v. 13, 2020.

[12] Lorena A. C., De Carvalho, André C. P. L. F. (2007) *Uma introdução às support vector machines,* Revista de Informática Teórica Aplicada, vol. 14, no. 2, pp. 43?67.

[13] Ma, C.; Zhang, H. H., Wang, X. (2014). *Machine learning for big data analytics in plants.* Trends in plant science, 19(12):798–808.

[14] Leandro S. Marcolino, Bráulio R. G. M. Couto, Marcos A. dos Santos. (2010) *Genome visualization in space* . In Procceedings of IWPACCBB, Springer Berlin Heidelberg pp. 225-232.

[15] Mesquita, P. S. B. (2014) *Um Modelo de regressão logística para avaliação dos programas de Pós-graduação no Brasil* . Dissertação (Mestrado) Universidade Estadual do Norte Fluminense, Campos dos Goytacazes.

[16] Morais R.F.; Ortega, J. M. ; Azevedo, V. A.C. ; Dos Santos, M. A., et al., (2020) *Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression* GENE, v. 726, p. 144168

[17] Pires, D. E. V. ; Minardi, R. C. M. ; Santos, M. ; Da Silveira, C. H. ; Santoro, M. M. ; Meira Junior, W. (2011) *Cutoff Scanning Matrix (CSM): funciton prediction and fold recognition by protein inter-residue distance patterns* . BMC Genomics, v. 12 Suppl 4, p. S12.

[18] Santos, Anderson R; Santos, Marcos A ; Baumbach, Jan ; McCulloch, John A ; Oliveira, Guilherme C ; Silva, Artur ; Miyoshi, Anderson ; Azevedo, Vasco. (2011).*A singular value decomposition approach for improved taxonomic classification of biological sequences.* BMC Genomics, v. 12, p. S11.

[19] Santos, Alysson dos (2017) *Early Breast Cancer Detection Using Logistic Regression Models.* Mestrado em Bioinformáica. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

[20] Wang, J. T.; Zaki, M. J.; Toivonen, H. T., Shasha, D. (2005). *Introduction to data mining in bioinformatics.* Em Data Mining in Bioinformatics, pp. 3–8. Springer.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

CFD Simulation of the Saliva Droplet Trajectory for Human Sneeze in Standing, Walking, and Running Positions

Nicolas Lima Oliveira[1] and Patricia Habib Hallak[1]

[1] *Graduate Program in Computational Modeling, Federal University of Juiz de Fora, Brazil*

Abstract

With the spread of the Sars-CoV-2 virus, which causes the pandemic disease COVID19, quarantine and lockdown orders were put in place all over the world. Physical activities, which were practiced mostly in closed environments, such as gyms, began to be practiced outdoors. Taking this trend into account, the present work aimed to understand how the virus spreads in environments in five different cases: standing, walking (**1.2 m/s**), running (**4 m/s**) and another two running cases in formation, aligned and side by side. For the analysis, computational fluid dynamics (CFD) were used through the ANSYS AIM program in its student free version. The physical phenomena of dragging, buoyancy, and dispersion were considered to understand how particles of fluids from the human respiratory system through sneezing behave. The results show that for the standing case, particles can reach up to 4m from the point of origin, while for running and walking cases they can reach up to 5m and 10m, respectively. Both cases studied with human bodies in formation showed the possibility of contagion. The results of the study reinforce the need for social isolation and the use of masks.

Keywords: Aerodynamics, CFD, COVID19, Particle Dispersion.

1 INTRODUCTION

With the spread of the Sars-CoV-2 virus, which causes the COVID19 pandemic disease, quarantine and lockdown orders have been implemented almost all over the world. Physical activities, mainly practiced in enclosed environments such as gyms, began to be practiced outdoors.

The practice of activities, such as walking and jogging, generates a wake of particles that has the potential to transmit the virus to individuals who are in the vicinity of an infected person. The

Contato: Nícolas, nicolas_limaoliveira@yahoo.com.br

range of these particles, which can be understood as human saliva, is an important parameter to understand for determining social safety distances.

This research aims, therefore, to assess under which conditions the greatest spread of human saliva droplets occurs. To achieve this goal, the concepts of computational fluid dynamics (CFD) are used to simulate some situations that are not possible in most official directions determined by the responsible bodies. Among the situations of possible contagion by COVID19, the work considers the situation of sneezing without barriers as the most critical among breathing, speech, and coughing. In this condition, it is investigated how much body movement can enhance the spread of saliva droplets.

In all cases considered, the trajectory of the saliva droplets and their maximum range are observed. Simulations are performed with the ANSYS software, in its academic version of AIM environment. Size distribution of the sneeze droplets exhaled immediately in the mouth follows the proposal of Han *et al* [5]. Nishimura *e*t al [9] propose to mathematically understand the dynamics of particles from the observation of images of a sneeze induced by nasal cavity stimulation. Bourouiba *et al* [2] use CFD combined with experimental observations to assess the path and range of particles. In 2016, Rahiminejad *e*t al [11] used CFD to describe the distributions of velocity, pressure, and turbulence intensity of the airflow during a female patient's sneeze. Blocken *e*t al [1] also use CFD for an aerodynamic analysis of droplet dispersion with the movement of the body of the individual who exhales them, in order to assess whether there is droplet transfer between nearby people.

These analyzes become important, as most droplets are expected to fall to the ground or evaporate before traveling a distance of 1.5 m [1]. However, the distance of 1.5 m is efficient for the condition of stationary people and does not take into account the possible aerodynamic effects induced by the movement of people [1].

This research corroborates the activities of the Federal University of Juiz de Fora, which, in partnership with the city hall, forms the Municipal Committee for Combating and Preventing COVID19 in the region. The relevance of the work is, therefore, to motivate the debate about the minimum distances in collective spaces, about the mandatory use of masks and hygiene practices. The purpose is to contain the contagion by droplets of saliva arising in different situations.

2 METHODOLOGY

2.1 Theoretical Reference

2.1.1 Navier Stokes Equations

To carry out this work, the ANSYS AIM 2020 program was used in its student version. This version of the program has the same functions as its paid version, but with a simulation limitation for computational fluid dynamics problems of meshes with 1 million cells. In all cases studied here, this limitation was respected. For incompressible flows with constant viscosity, the problem is governed by the Navier-Stokes equations 1 and 2. These equations are presented in an Eulerian description, where the characteristic properties of the environment are considered a function of space and time. The problem is defined in a domain Ω with a contour Γ containing n dimensions of Euclidean space. In non-conservative form, that is, using primitive pressure and

velocity variables in Cartesian coordinates and using the summation convention: $a = 1, ..., n$ and $b = 1, ..., n$ these equations take the form:

- Momentum Conservation Equation:

$$\rho \frac{\partial v_{\mathrm{a}}}{\partial t} + \rho v_{\mathrm{b}} \frac{\partial v_{\mathrm{a}}}{\partial x_{\mathrm{b}}} - \frac{\partial \tau_{\mathrm{ab}}}{\partial x_{\mathrm{b}}} + \frac{\partial p}{\partial x_{\mathrm{a}}} = f_{\mathrm{a}}; \quad \tau_{\mathrm{ab}} = \mu \left(\frac{\partial v_{\mathrm{a}}}{\partial x_{\mathrm{b}}} + \frac{\partial v_{\mathrm{b}}}{\partial x_{\mathrm{a}}} \right) \tag{1}$$

- Mass conservation equation (continuity equation)

$$\frac{\partial v_{\mathrm{a}}}{\partial x_{\mathrm{a}}} = 0 \tag{2}$$

  where ρ is the fluid specific mass, τ_{ab} is the viscous stress, μ is the fluid viscosity, v is the velocity components, p is the pressure, f_{a} are the components of forces per unit of volume.

Due to flow random characteristics, it is necessary to consider turbulence models. The model chosen for the treatment of turbulence in the cases discussed here was k-ω SST. The SST (Shear Stress Transport) model was proposed by Menter [7] and is a combination of the traditional k-ϵ models, with good performance in flows far from the boundary layers, and k-ω, which is concerned with the transport of shear stresses in boundary layers. This model considers the turbulent viscosity (ν_t), turbulent kinetic energy k and, the specific rate of dissipation of ω. To understand the adopted turbulence model, some references on the subject are cited [12, 10, 8, 7].
.

2.1.2 Particle Transport Model

Particle transport modeling is a type of multiphase model, in which particles are tracked through the flow following the Lagrangian framework. The field described by the particles is accompanied by a set of transient ordinary differential equations that represent the position and velocity of each particle. These equations are then integrated using a simple integration method to calculate the behavior of particles as they traverse the flow field [3]. In other words, it is a model that "tracks" a small number of particles passing through a fluid in a continuous environment.

Movement variables of particles are obtained as follows:

Displacement: calculated using direct Euler integration for the time increment δt and described by:

$$\mathbf{x}_p^n = \mathbf{x}_p^0 + \mathbf{U}_p^0 \delta t \tag{3}$$

where p refers to the particle, n and 0 the new and old increments, $\mathbf{x}$ is the position and $\mathbf{U}$ is the velocity.

Velocity: corresponds to the particle velocity calculated at the beginning of the increment, that is:

$$m_p \frac{d\mathbf{U}_p}{dt} = \sum \mathbf{F} \tag{4}$$

where $\sum \mathbf{F}$ is the sum of the aerodynamic forces acting on the particle of mass m_p.

transfer interface: it is understood, by Eq. 4, that the fluid exerts influence on the particle's movement. The particle, in turn, affects the fluid's movement. There is, therefore, an information exchange interface between fluid and particle that is seen as a "phase coupling". A robust coupling requires that the terms representing the particles be included in the fluid momentum equations as source terms, associated with dispersion and turbulence.

The aerodynamic force on the particle F_{Dp} is proportional to the relative velocity between the particle and the fluid, U_S. In this way, the drag force is written as:

$$F_{Dp} = \frac{1}{2} C_{Df} \rho_a A_f \|U_S\| U_S \tag{5}$$

where C_{Df} is the drag coefficient of the particle, A_f is its sectional area, and ρ_a the density of the continuous fluid. The buoyancy force is the force that arises from a particle immersed in a fluid. This force is equal to the displaced weight of the fluid, therefore:

$$F_B = \frac{\pi}{6} d_P^3 (\rho_P - \rho_a) \boldsymbol{g} \tag{6}$$

where d_P is the diameter of a given particle, $\boldsymbol{g}$ is the acceleration vector of gravity. Particle density is symbolized by ρ_P. The source term associated with S_p particles are obtained by solving transport equations for the sources:

$$\frac{dS_p}{dt} = C_S \phi_p + R_S \tag{7}$$

where $C_S \phi_p$ is the linear contribution to the transport of the generic property ϕ and R_s would be other contributions.

2.1.3 Numerical Schemes

Discretization was performed using the finite volume method based on elements. The fluid equations are solved in steady state and, for the temporal evolution of the particle equations, the explicit Euler method was used. A second order upwind technique was used for the advective parcel and a coupled pressure-based algorithm to solve the pressure and velocity terms.

2.2 Domains and Boundary Conditions

Three possibilities of movement were considered: standing human body, at walking speed (1.2 m/s), at running speed (4 m/s). In the three conditions, first, only one body was considered in the simulations. In the running condition, two runners were also simulated, one on the other's streamline and side by side. Henceforth, these conditions will be called Standing, Walking, Running and In Line and Side by Side, respectively. Fluid density ρ was 1.1843 kg/m^3, kinematic viscosity ν was 10×10^{-6} Pa, and moderate turbulence intensity.

An example of one of the computational domains, by way of illustration, is shown in Fig. 1 and refers to an individual in an isolated race condition. The length of Standing, Walking, and Running single body domains are 7m, 7m, and 13.5m, respectively, with the bodies positioned at 6m, 2m, and 1.5m in relation to the starting face. The length of In Line and Side by Side two bodies domains are 5.5m. In the In Line domain, the bodies are positioned in 1.5m and 3.5m. In Side by Side domain, both bodies are positioned in 1.5m.

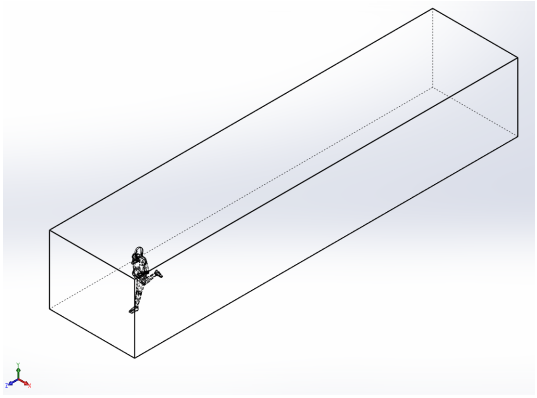


Fig. 1: Computational domain for running condition

In all cases, atmospheric pressure values $P_{atm} = 1$ atm and gravity acceleration $g = 9.81$ m/s^2 were used. The physical model used to solve the problem was the RANS (Reynolds Avareged Navier Stokes). All simulations had as turbulence model $k-\omega\, SST$ and used crossed buoyancy models with turbulence. Escape conditions for the particles were used when there was a collision with the boundary conditions. For the walking and running models, speeds of 1.2 m/s and 4 m/s were used, respectively, through boundary conditions of *inlet* type. Conditions of *open* type were configured as a null pressure difference in relation to the environment outside the domain. The soil and bodies were configured as *wall, no slip* condition. The particle injection condition has an Average Diameter of $360.1\mu m$, a Diameter Standard Deviation of $123\mu m$, and a Minimum Diameter of 105μ, for mass flow the value of 10^{-20} kg/s [5] and the value of 30 m/s was applied was used to describe the average speed of the sneeze action ([5] and [11]). For the forming running model, the mouth region of the second body was modeled with a boundary condition of *outlet* type with a mass flow rate of $5.5859 * 10^{-4} kg/s$ [4] . The values used to describe the droplet diameter range were obtained from [5] and treated as a normal distribution.

2.3 Meshes

All meshes were created by the curvature and proximity generator algorithm. The quality of elements and skewness resulting from the meshes are in Tables 1 and 2, the number of elements and nodes for the five situations are in Table 3. The meshes are displayed in Fig. 2 to 6. The faces close to the mouth of the geometric model underwent a refinement process using the element size of $0.002m$ (Fig. 7).

TABLE 1: SKEWNESS AND QUALITY OF ELEMENTS - SINGLE BODY

		Standing	Walking	Running
Skewness	Average	0.236	0.236	0.232
	Standard Deviation	0.126	0.126	0.123
Quality	Average	0.832	0.833	0.835
	Standard Deviation	0.100	0.100	0.098

TABLE 2: SKEWNESS AND QUALITY OF ELEMENTS - TWO BODIES

		In line	Side by side
Skewness	Average	0.280	0.271
	Standard Deviation	0.145	0.145
Quality	Average	0.805	0.806
	Standard Deviation	0.111	0.110

TABLE 3: NUMBER OF NODES AND ELEMENTS

	Nodes	Elements
Standing	1.6937e+5	8.9291e+5
Walking	1.6014e+5	8.4872e+5
Running	1.3381E+5	9.4927e+5
In Line	1.7766e+5	6.6798E+5
Side by Side	1.3491e+5	6.7365E+5

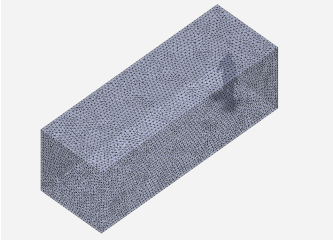


Fig. 2: Standing mesh

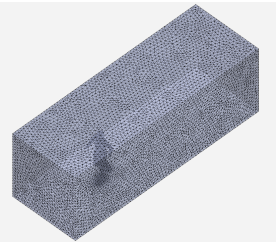


Fig. 3: Walking mesh

Fig. 4: Running Mesh

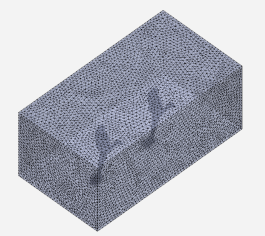


Fig. 5: In Line mesh

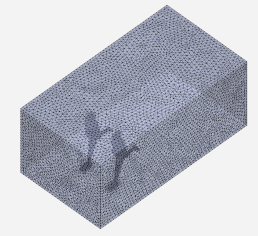


Fig. 6: Side by Side mesh

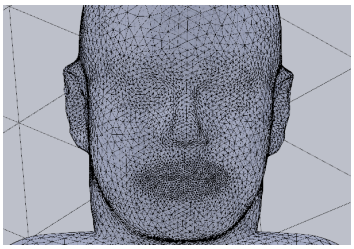


Fig. 7: Nose and mouth region refinement

2.4 Model Validation

To verify the model used in this study, the value of the drag coefficient (C_D) was chosen, given by: $C_D = \frac{2F_D}{\rho V^2 S}$, where F_D is the drag force generated by the fluid in the body of interest, ρ is the air density, V is the velocity (1.2 m/s or 4 m/s) and S is the reference area (0.5 m^2 for walking and 0.42 m^2 for running).

For walking and running, the values of $C_D = 0.72$ and 0.64 were obtained, respectively. According to Lari [6], the value of $C_D = 0.69$ is a valid and reasonable approximation for different speeds. The errors of the values obtained for the drag coefficient, when comparing them with this reference, were 4.3% and 7.2% for the cases of walking and running, respectively. This

result was expected since the walking human body has a larger reference area than the running body. The velocity contours for both cases are graphically displayed in Fig. 8 and 9.

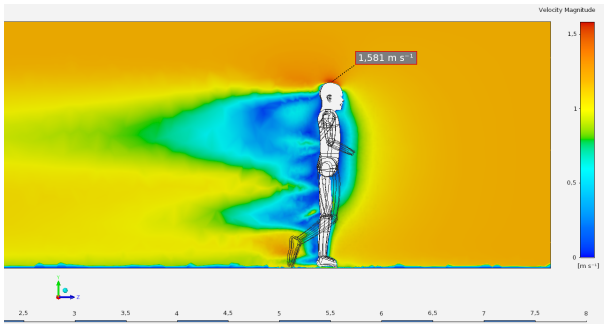


Fig. 8: Velocity contour - Walking ($1.2m/s$)

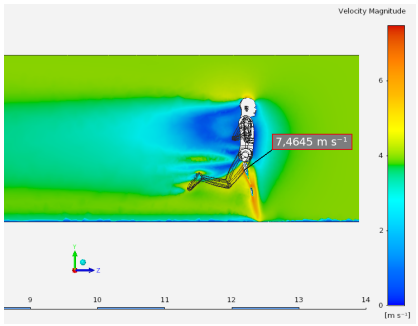


Fig. 9: Velocity contour - Running ($4m/s$)

A comparison was also made of the trajectory of the sneeze particles in Standing condition with the literature [2]. In this case, the velocity of the particles was $6m/s$. Figs. 10 and 11 show images that compare the simulation results with the literature.

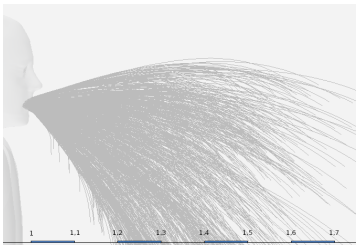


Fig. 10: Particle dispersion - average sneeze speed of $6m/s$

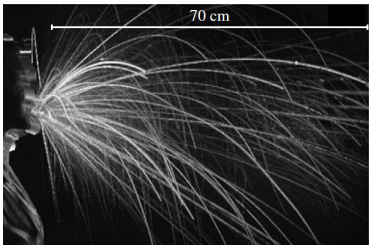


Fig. 11: Experimental particle dispersion test - Adapted from [2]

3 RESULTS AND DISCUSSION

3.1 Single bodies

Fig. 12 displays the graphical results for the scattering of particles in an event of a sneeze for a standing human body. Fig. 13 displays the graphical results for particle dispersion in the event of a sneeze for a walking human body. Fig. 14 displays the graphical results for particle dispersion in the event of a sneeze for a human body in an isolated run situation.

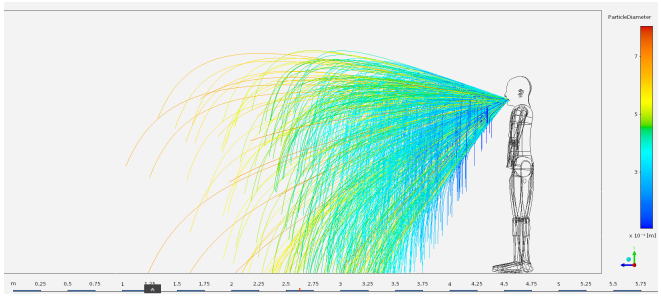


Fig. 12: Particle dispersion - Standing

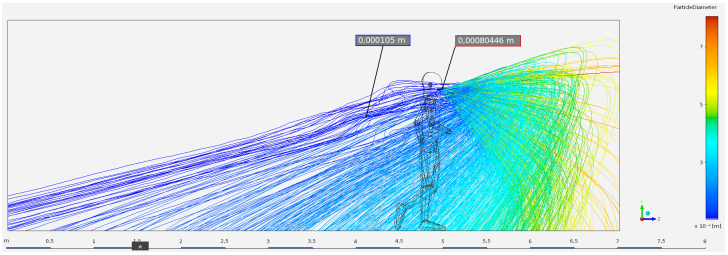


Fig. 13: Particle dispersion - Walking ($1.2m/s$)

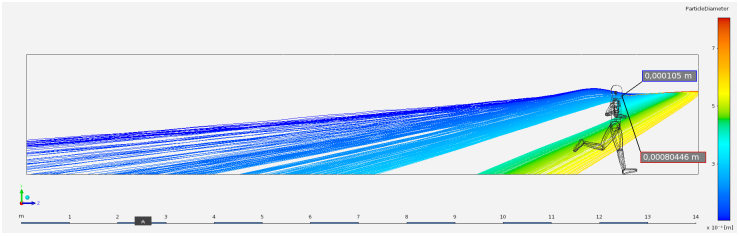


Fig. 14: Particle dispersion - Running ($4m/s$)

Through the results obtained in this subsection, it is possible to make some important observations about how the particles behave in the studied cases. For all the results obtained in this work, conditions related to human sneezing were used, as this produces greater spread than other phenomena of the human airway such as coughing, breathing, and speech ([5] and [9]).

For the first case studied (Fig. 12), where the human body is motionless, there was a scattering distance of up to 4m. In this case, it is clear that the particles that moved farther from their point of origin were the ones with the largest diameter, while these are also the ones that reach the ground more quickly. In this case, the portion of buoyancy over the particles dominates, since the drag over them is nil. This observation is confirmed by the results obtained by Zhu *et al* [13].

For the cases of walking (1.2m/s or 4.32km/h) and running (4m/s or 14.4km/h) it is possible to observe the trajectories of particles along the simulation domains (Fig .13 and 14). Contrary to what was concluded in the first case, it is now observed that the particles that spread the most are those with the smallest diameter due to their greater buoyancy. In addition to buoyancy, drag forces also act on the particles, which increase with movement. There is a direct relation between the speed of the body and the distance of reach of the particles, which reach greater distances because they spend more time in the air due to buoyancy and because they are suffering the action of drag during this time. Based on what was observed in the simulations, for the case of walking, a minimum distance of $5m$ is recommended and for the case of running, $10m$. These safety distances are compatible with the work done by Blocken [1].

Due to the behaviors observed, it is important to understand that individuals who are at a height below the source of particles (mouth and nose region) will be more prone to the streamline region of the analyzed object. This is especially important in the case of children and domestic animals, since these, despite not being mostly affected by COVID19, behave as important vectors of the disease.

3.2 Two bodies

Fig. 15, 16, and 17 show the graphical results for particle dispersion in the case of a sneeze for a human body in a forming-run situation. Note, again, that the smaller particles are those that remain in suspension longer. It is also concluded that in line training it is observed that the corridor behind the streamline is subject to the reception of large amounts of particles, both in the face region, as well as throughout the body. In side by side training, there is also an exchange of material between practitioners. From the analysis of these specific cases, the importance of using masks for the practice of these activities is considered.

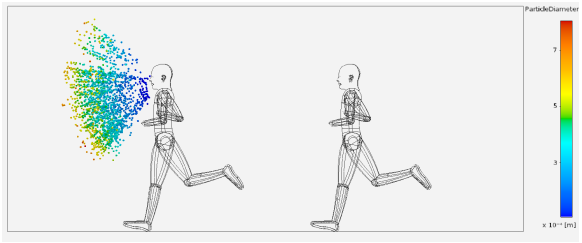


Fig. 15: Particle dispersion for In Line at the time of sneeze.

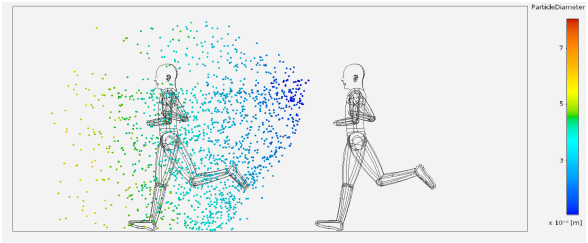


Fig. 16: Particle dispersion for In Line at the beginning of scattering.

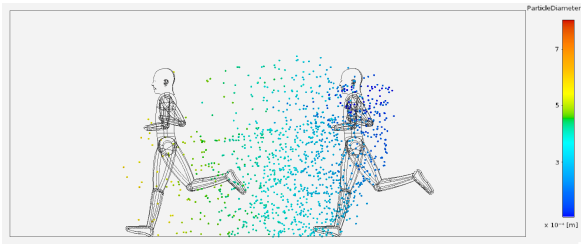


Fig. 17: Particle dispersion for In Line at the time the droplets reach the second individual.

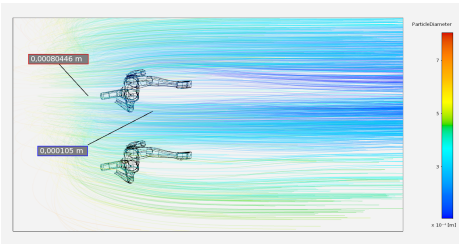


Fig. 18: Particle dispersion streamline - Side by Side.

4 Final considerations

This work aimed to use CFD to follow the path of human saliva droplets in the sneeze condition. Three situations were analyzed, standing static, walking, and running. It was observed that the movement intensifies the distribution of droplets, which reach greater distances in the running condition.

The analyzes show that the minimum safety distance of 1.5 m from one person to another is only valid for the stationary person condition. In motion, the saliva droplets reach greater distances. The streamline produced under these circumstances, although decaying along the way, pose risks to smaller bodies, such as children and pets. Thus, the authors, based on the conclusions of this work, recommend greater distances for the practice of physical activities and the use of masks.

For the continuation of the research, it is intended to invest in the improvement of the computational model. One of the perspectives is about the limitations related to the sneeze's particle diameter spectrum, which despite having higher speed, also end up reaching the ground faster.

Another proposal is to simulate other phenomena, such as breathing or coughing, which generate smaller particles and some of them may remain in suspension. Another issue to be studied is the convection of particles caused by continuous winds in different directions as well as gusts of wind, in addition to different formations of people exercising physical activity. The effectiveness of the use of masks during activities is also an object of study in future research.

5 *Acknowledgements*

The authors wish to thank CAPES (Finance Code 001) Graduate Program in Computational Modeling - UFJF for their support.

REFERENCES

[1] T. v. D. T. M. B. Blocken, F. Malizia. Towards aerodynamically equivalent covid19 1.5 m social distancing for walking and running, 2020.

[2] L. Bourouiba, E. Dehandschoewercker, and J. W. Bush. Violent expiratory events: on coughing and sneezing. *Journal of Fluid Mechanics*, 745:537–563, 2014.

[3] A. CFX. 15.0. 0. *ANSYS CFX-Solver Theory Guide. ANSYS Inc*, 2013.

[4] T. Gemci, V. Ponyavin, Y. Chen, H. Chen, and R. Collins. Computational model of airflow in upper 17 generations of human respiratory tract. *Journal of Biomechanics*, 41(9):2047–2054, 2008.

[5] Z. Han, W. Weng, and Q. Huang. Characterizations of particle size distribution of the droplets exhaled by sneeze. *Journal of The Royal Society Interface*, 10(88):20130560, 2013.

[6] T. A. Lari G. Estimating drag and heat transfer coefficient of a runner by using numerical methods. 2013.

[7] F. Menter. Zonal two equation kw turbulence models for aerodynamic flows. In *23rd fluid dynamics, plasmadynamics, and lasers conference*, page 2906, 1993.

[8] F. R. Menter. Review of the shear-stress transport turbulence model experience from an industrial perspective. *International journal of computational fluid dynamics*, 23(4):305–316, 2009.

[9] H. Nishimura, S. Sakata, and A. Kaga. A new methodology for studying dynamics of aerosol particles in sneeze and cough using a digital high-vision, high-speed video system and vector analyses. *PloS one*, 8(11), 2013.

[10] S. B. Pope and S. B. Pope. *Turbulent flows*. Cambridge university press, 2000.

[11] M. Rahiminejad, A. Haghighi, A. Dastan, O. Abouali, M. Farid, and G. Ahmadi. Computer simulations of pressure and velocity fields in a human upper airway during sneezing. *Computers in biology and medicine*, 71:115–127, 2016.

[12] D. C. Wilcox et al. *Turbulence modeling for CFD*, volume 2. DCW industries La Canada, CA, 1998.

[13] S. Zhu, S. Kato, and J.-H. Yang. Study on transport characteristics of saliva droplets produced by coughing in a calm indoor environment. *Building and environment*, 41(12):1691–1702, 2006.

Anais do XV Encontro Acadêmico de Modelagem Computacional.
Laboratório Nacional de Computação Científica, 22 a 25 de Fevereiro de 2022.

Método de elementos finitos para uma equação de viga com o operador p-biharmônico

Rui M. P. Almeida[1], José C. M. Duque[1], Jorge Ferreira[2] e Willian S. Panni[1]

[1] *Universidade da Beira Interior, Centro de Matemática e Aplicações, Covilhã, Portugal*
[2] *Universidade Federal Fluminense, Departamento de Ciências Exatas, Volta Redonda/RJ, Brasil*

Resumo

Neste artigo, consideramos uma equação de viga não linear com o operador p-biharmônico, onde $1 < p < \infty$. Usando uma mudança de variável, transformamos o problema em um sistema de equações diferenciais e demonstramos a existência, unicidade e regularidade da solução fraca aplicando o teorema de Lax-Milgram e resultados clássicos de análise funcional. Investigamos a formulação discreta desse sistema e, com o auxílio do teorema de Brouwer, mostramos que o problema tem solução discreta. A unicidade e estabilidade da solução discreta são obtidas através de métodos clássicos. Depois de estabelecer a ordem de convergência, aplicamos o método dos elementos finitos para obtermos um sistema de equações algébricas. Por fim, implementamos os códigos computacionais no software Matlab e realizamos a comparação entre teoria e simulações.

Palavras-chaves: Operador p-biharmônico, Solução fraca, Ordem de convergência, Método de elementos finitos, Simulações numéricas.

1 INTRODUÇÃO

Seja Ω um domínio limitado em $\mathbb{R}^N$ $(N \geq 1)$ com fronteira suave $\partial\Omega$. Consideramos o problema de encontrar uma função u tal que

$$\begin{cases} \Delta_p^2 u = f(x), & \text{em } \Omega, \\ u = 0, \Delta u = 0, & \text{em } \partial\Omega, \end{cases} \tag{1}$$

onde Δ_p^2 é o operador de quarta ordem, chamado de p-biharmônico, definido por

$$\Delta_p^2 u = \Delta\left(|\Delta u|^{p-2}\Delta u\right),$$

$p \in \mathbb{R}$ satisfazendo $1 < p < \infty$ e $f \in L^2(\Omega)$.

Contato: Willian S. Panni, willian.panni@ubi.pt

2 PRELIMINARES

Neste trabalho, usamos a notação padrão para os espaços de Lebesgue e Sobolev, $L^p(\Omega)$ e $W^{m,p}(\Omega)$, respectivamente (ver [1]). Denotamos o produto interno em $L^2(\Omega)$ por $(\cdot\,,\cdot)$ e as constantes que são independentes dos parâmetros e das funções envolvidas são denotadas por C. Para mais detalhes sobre o referido artigo, indicamos [2].

Teorema 2.1 *Sejam $0 < \mu \leq 1$ e $\Omega \subset \mathbb{R}^N$. Se $u \in L^2(\Omega)$, então $|u|^\mu \in L^1(\Omega)$ e*

$$\int_\Omega |u(x)|^\mu \, dx \leq C \, \|u\|^\mu_{L^2(\Omega)} \,,$$

onde $C = |\Omega|^{\frac{2-\mu}{2}}$.

Denotamos por $\mathcal{T}_h$ uma partição não degenerada do domínio poligonal $\Omega \subset \mathbb{R}^N$ em simplexes com parâmetro h, ou seja, o conjunto Ω é subdividido em um número finito de subconjuntos T_k, $k = 1, \ldots, n$, chamados de elementos finitos, tais que as seguintes condições são satisfeitas:

i) $\Omega = \cup_{k=1}^n T_k$;

ii) $int(T_k) \neq \emptyset, \forall\ T_k \in \mathcal{T}_h$;

iii) $int(T_i) \cap int(T_j) = \emptyset, \forall\ T_i, T_j \in \mathcal{T}_h$ com $i \neq j$;

iv) Cada lado de T_k ou pertence à fronteira de Ω ou é lado de outro $T_i \in \mathcal{T}_h$;

v) Cada T_k tem fronteira Lipschitz contínua.

Denotamos por V_h o espaço de funções contínuas em $\overline{\Omega}$, que são polinômios de grau $r-1$, com $r \geq 2$, em cada elemento de $\mathcal{T}_h$ e que se anulam em $\partial\Omega$, ou seja,

$$V_h = \left\{ u \in C_0^0(\overline{\Omega}); u_{\big|_{T_k}} \text{ é um polinômio de grau } r-1 \text{ para todo } T_k \in \mathcal{T}_h \right\}.$$

Teorema 2.2 *Para todo $q > 1$ e $\delta \geq 0$, existe uma constante positiva C tal que, para todo $\xi, \kappa \in \mathbb{R}$ com $\xi \neq \kappa$,*

$$\left| |\xi|^{q-2}\,\xi - |\kappa|^{q-2}\,\kappa \right| \leq C\,|\xi - \kappa|^{1-\delta}\,(|\xi| + |\kappa|)^{q-2+\delta}\,.$$

3 EXISTÊNCIA E UNICIDADE DA SOLUÇÃO FRACA

Nesta seção, estabelecemos a existência e a unicidade da solução fraca para o Problema (1). Seguindo Katzourakis e Pryer [3], definimos a variável auxiliar

$$v = |\Delta u|^{p-2}\,\Delta u. \tag{2}$$

Considerando q como o expoente conjugado de p, ou seja, $q = \frac{p}{p-1}$, então

$$|v|^{q-2}\,v = \Delta u. \tag{3}$$

Usando as Eq. (2) e (3), reescrevemos o Problema (1) como o seguinte sistema de equações diferenciais

$$\begin{cases} \Delta v = f, & \text{em } \Omega, \\ \Delta u = |v|^{q-2}\,v, & \text{em } \Omega, \\ u = 0,\ v = 0, & \text{em } \partial\Omega. \end{cases} \tag{4}$$

Definição 3.1 *O par* $(u, v) \in H_0^1(\Omega) \times H_0^1(\Omega)$ *é uma solução fraca para o Problema* (4) *se, para todo* $(\psi, \eta) \in H_0^1(\Omega) \times H_0^1(\Omega)$*, o seguinte sistema for satisfeito*

$$\begin{cases} (\nabla v, \nabla \psi) = -(f, \psi), \\ (\nabla u, \nabla \eta) = -\left(|v|^{q-2} v, \eta\right). \end{cases} \tag{5}$$

A formulação fraca (5) pode ser reescrita como

$$\begin{cases} a(v, \psi) = -(f, \psi), & \forall\ \psi \in H_0^1(\Omega), \\ a(u, \eta) = -\left(|v|^{q-2} v, \eta\right), & \forall\ \eta \in H_0^1(\Omega), \end{cases} \tag{6}$$

onde a forma bilinear $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ é definida por

$$a(u, v) = \int_\Omega \nabla u \nabla v dx,\ \forall\ u, v \in H_0^1(\Omega). \tag{7}$$

A seguir, usamos o teorema de Lax-Milgram e o teorema de imersão de Sobolev com $\Omega \subset \mathbb{R}^N$ para provar a existência e a unicidade da solução fraca. Devido as imersões, consideramos primeiro $N \leq 4$ e, em seguida, $N > 4$.

Teorema 3.1 *Sejam* $\Omega \subset \mathbb{R}^N$*,* $N \leq 4$*, um conjunto aberto e limitado com* $\partial\Omega$ *de classe* C^2*,* $f \in L^2(\Omega)$ *e* $1 < p < \infty$*. Então, existe um único par* (u, v) *que é a solução fraca do Problema* (4)*, no sentido da Definição 3.1, e além disso* $(u, v) \in H^2(\Omega) \times H^2(\Omega)$*.*

Demonstração (esboço): Como a forma bilinear a, definida em (7), é contínua e coerciva, então pelo teorema de Lax-Milgram existe um único $v \in H_0^1(\Omega)$ tal que a primeira equação do Sistema (6) é satisfeita. Uma vez que $f \in L^2(\Omega)$ e $\partial\Omega$ possui classe C^2, tem-se $v \in H^2(\Omega)$.

Nesta etapa, investigamos $|v|^{q-2}v$ através do teorema de imersões de Sobolev [4, p. 44]. Se $N < 4$ e $1 < p < \infty$, então $H^2(\Omega) \subset C(\overline{\Omega})$, assim $|v|^{q-2}v \in L^\infty(\Omega) \subset L^2(\Omega)$. Se $N = 4$ e $1 < p < 3$, então $1 < 2q - 2 < \infty$, assim $v \in H^2(\Omega) \subset L^{2q-2}(\Omega)$ que implica em $|v|^{q-2}v \in L^2(\Omega)$. Se $N = 4$ e $3 \leq p < \infty$, então $0 < 2q - 2 \leq 1$ e pelo Teorema 2.1 vem que $|v|^{q-2}v \in L^2(\Omega)$. Portanto, $|v|^{q-2}v \in L^2(\Omega)$ para $\Omega \subset \mathbb{R}^N$, com $N \leq 4$ e $1 < p < \infty$, finalmente repetimos os passos para a segunda equação. ∎

Teorema 3.2 *Sejam* $\Omega \subset \mathbb{R}^N$*,* $N > 4$*, um conjunto aberto e limitado com* $\partial\Omega$ *de classe* C^2*,* $f \in L^2(\Omega)$ *e* $\frac{2N-4}{N} < p < \infty$*. Então, existe um único par* (u, v) *que é a solução fraca do Problema* (4)*, no sentido da Definição 3.1, e além disso* $(u, v) \in H^2(\Omega) \times H^2(\Omega)$*.*

Demonstração (esboço): Prosseguindo de maneira análoga à demonstração do Teorema 3.1, no entanto agora como $N > 4$, para $|v|^{q-2}v \in L^2(\Omega)$ é necessário utilizar as imersões de Sobolev [4, p. 44] com $\frac{2N-4}{N} < p < \infty$. De fato, se $\frac{2N-4}{N} < p < 3$, então $1 < 2q-2 < \frac{2N}{N-4}$, assim $v \in H^2(\Omega) \subset L^{2q-2}(\Omega)$ implica em $|v|^{q-2}v \in L^2(\Omega)$. Por outro lado, se $3 \leq p < \infty$, então $0 < 2q - 2 \leq 1$ e pelo Teorema 2.1 obtemos $|v|^{q-2}v \in L^2(\Omega)$. ∎

4 PROBLEMA DISCRETO

Nesta seção, estudamos o problema discreto associado ao Problema (4). Por uma questão de simplicidade, consideramos $\Omega \subset \mathbb{R}$. Para $\Omega \subset \mathbb{R}^N$, $N > 1$, as demonstrações são similares, mas dependem de outras imersões de Sobolev e exigem algumas restrições sobre p.

4.1 Existência, unicidade e regularidade da solução discreta

Começamos definindo o conceito de solução discreta para o Problema (4).

Definição 4.1 *O par $(u_h, v_h) \in V_h \times V_h$ é considerado uma solução discreta para o Problema* (4) *se, para cada par $(\psi_h, \eta_h) \in V_h \times V_h$, o seguinte sistema for satisfeito*

$$\begin{cases} (\nabla v_h, \nabla \psi_h) = -(f, \psi_h), & \forall\ \psi_h \in V_h, \\ (\nabla u_h, \nabla \eta_h) = -\left(|v_h|^{q-2} v_h, \eta_h\right), & \forall\ \eta_h \in V_h. \end{cases} \tag{8}$$

A seguir, mostramos que o Problema (4) tem uma única solução discreta.

Teorema 4.1 *Se $f \in L^2(\Omega)$, então existe uma única solução discreta $(u_h, v_h) \in V_h \times V_h$ para o Problema* (4)*, no sentido da Definição 4.1.*

Demonstração (esboço): Usando as desigualdades de Young, Poincaré e Hölder, e o teorema de Brouwer, demonstra-se que existe uma solução discreta $(u_h, v_h) \in V_h \times V_h$ para o Problema (4). A unicidade da referida solução é obtida por contradição supondo que existam duas soluções diferentes. ∎

No próximo teorema, provamos a estabilidade da solução discreta (u_h, v_h) para o Problema (4).

Teorema 4.2 *Seja $(u_h, v_h) \in V_h \times V_h$ a solução discreta do Problema* (4)*, no sentido da Definição 4.1. Então, para toda $f \in L^2(\Omega)$,*

$$\|v_h\|_{H_0^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}, \tag{9}$$

$$\|u_h\|_{H_0^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}^{q-1}. \tag{10}$$

Demonstração (esboço): Considerando $\psi_h = v_h$ e $\eta_h = u_h$, através das desigualdades de Young e Poincaré, e de imersões de Sobolev, obtemos (9) e (10). ∎

4.2 Ordem de convergência

Investigamos agora a ordem de convergência do Problema (8). Para tanto, consideramos primeiro $1 < p \leq 2$, que implica $2 \leq q < \infty$, e depois consideramos $2 < p < \infty$, que corresponde a $1 < q < 2$.

Teorema 4.3 *Sejam $\Omega \subset \mathbb{R}$ um conjunto aberto e limitado com $\partial\Omega$ de classe C^2, (u, v) solução de* (5)*, (u_h, v_h) solução de* (8) *e $1 < p \leq 2$. Se $u, v \in H^s(\Omega)$, com $1 \leq s \leq r$, então*

$$\|\nabla (v - v_h)\|_{L^2(\Omega)} \leq Ch^{s-1} \|v\|_{H^s(\Omega)}, \tag{11}$$

$$\|v - v_h\|_{L^2(\Omega)} \leq Ch^s \|v\|_{H^s(\Omega)}, \tag{12}$$

$$\|\nabla (u - u_h)\|_{L^2(\Omega)} \leq Ch^{s-1} \|u\|_{H^s(\Omega)} + Ch^s \|v\|_{H^s(\Omega)}, \tag{13}$$

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^s \|u\|_{H^s(\Omega)} + C\left(h^s + h^{s+1}\right) \|v\|_{H^s(\Omega)}. \tag{14}$$

Demonstração (esboço): Para a primeira equação do Problema (8), a demonstração é idêntica à de Thomée [5, Teorema 1.1, p. 5], assim obtemos (11) e (12).

Usamos o operador de interpolação no espaço de elementos finitos V_h e através das desigualdades de Minkowski, Hölder, Young e Poincaré, do Teorema 2.2 com $\delta = 0$ e das imersões de Sobolev, concluímos (13) e (14). ∎

Agora, consideramos $2 < p < \infty$, então o seu expoente conjugado satisfaz $1 < q < 2$.

Teorema 4.4 *Sejam $\Omega \subset \mathbb{R}$ um conjunto aberto e limitado com $\partial\Omega$ de classe C^2, (u, v) solução de (5), (u_h, v_h) solução de (8) e $2 < p < \infty$. Se $u, v \in H^s(\Omega)$, com $1 \leq s \leq r$, então*

$$\begin{aligned}
\|\nabla(v - v_h)\|_{L^2(\Omega)} &\leq Ch^{s-1} \|v\|_{H^s(\Omega)}, && (15)\\
\|v - v_h\|_{L^2(\Omega)} &\leq Ch^{s} \|v\|_{H^s(\Omega)}, && (16)\\
\|\nabla(u - u_h)\|_{L^2(\Omega)} &\leq Ch^{s-1} \|u\|_{H^s(\Omega)} + Ch^{\frac{s}{p-1}} \|v\|^{\frac{1}{p-1}}_{H^s(\Omega)}, && (17)\\
\|u - u_h\|_{L^2(\Omega)} &\leq Ch^{s} \|u\|_{H^s(\Omega)} + C\left(h^{\frac{s+p-1}{p-1}} + h^{\frac{s}{p-1}}\right) \|v\|^{\frac{1}{p-1}}_{H^s(\Omega)}. && (18)
\end{aligned}$$

Demonstração (esboço): A primeira equação do Problema (8) não depende de p ou q, sendo assim, a demonstração é idêntica à de Thomée [5, Teorema 1.1, p. 5], assim obtemos (15) e (16).

Prosseguindo de maneira análoga à demonstração do Teorema 4.3 e considerando $\delta = \frac{p-2}{p-1}$ no Teorema 2.2, concluímos (17) e (18). ∎

4.3 Discretização pelo método de elementos finitos

Nesta seção, transformamos o Problema (8) em um sistema de equações algébricas. Para este efeito, consideramos a seguinte base para o espaço de elementos finitos $V_h = \langle \varphi_1, \varphi_2, \cdots, \varphi_n \rangle$. Então, podemos representar

$$u_h = \sum_{i=1}^{n} u_i \varphi_i, \quad v_h = \sum_{i=1}^{n} v_i \varphi_i, \quad \eta_h = \sum_{i=1}^{n} \eta_i \varphi_i, \quad \psi_h = \sum_{i=1}^{n} \psi_i \varphi_i, \tag{19}$$

onde u_i e v_i, com $1 \leq i \leq n$, são os coeficientes a serem determinados. Substituindo (19) no Problema (8) e notando que ψ_i e η_i são arbitrárias, nós obtemos o sistema

$$\begin{cases}
\left(\nabla \sum_{i=1}^{n} v_i \varphi_i, \nabla \varphi_j\right) = -(f, \varphi_j), & j = 1, \cdots, n,\\
\left(\nabla \sum_{i=1}^{n} u_i \varphi_i, \nabla \varphi_j\right) = -\left(\left|\sum_{i=1}^{n} v_i \varphi_i\right|^{q-2} \sum_{i=1}^{n} v_i \varphi_i, \varphi_j\right), & j = 1, \cdots, n.
\end{cases}$$

Como u_i e v_i, $1 \leq i \leq n$, são os coeficientes constantes,

$$\begin{cases}
\sum_{i=1}^{n} v_i (\nabla \varphi_i, \nabla \varphi_j) = -(f, \varphi_j), & j = 1, \cdots, n,\\
\sum_{i=1}^{n} u_i (\nabla \varphi_i, \nabla \varphi_j) = -\left(\left|\sum_{i=1}^{n} v_i \varphi_i\right|^{q-2} \sum_{i=1}^{n} v_i \varphi_i, \varphi_j\right), & j = 1, \cdots, n.
\end{cases} \tag{20}$$

Notemos que,

$$\left(\left| \sum_{i=1}^{n} v_i \varphi_i \right|^{q-2} \sum_{i=1}^{n} v_i \varphi_i, \varphi_j \right) = \sum_{i=1}^{n} v_i \int_{\Omega} \left| \sum_{k=1}^{n} v_k \varphi_k \right|^{q-2} \varphi_i \varphi_j dx.$$

Denotamos os elementos dos vetores $\vec{u} \in \mathbb{R}^{n\times 1}$ e $\vec{v} \in \mathbb{R}^{n\times 1}$ por u_i e v_i, respectivamente, enquanto os elementos das matrizes $M(\vec{v}) \in \mathbb{R}^{n\times n}$ e $K \in \mathbb{R}^{n\times n}$ são denotados por

$$\begin{aligned} M_{ij} &= \int_{\Omega} \left| \sum_{k=1}^{n} v_k \varphi_k \right|^{q-2} \varphi_i \varphi_j dx, \\ K_{ij} &= (\nabla \varphi_i, \nabla \varphi_j), \end{aligned}$$

respectivamente. Os elementos do vetor $\vec{f} \in \mathbb{R}^{n\times 1}$ são denotados por $f_i = -(f, \varphi_i)$. Assim, obtemos o seguinte sistema associado com o Problema (20)

$$\begin{cases} K\vec{v} = \vec{f}, \\ K\vec{u} = -M(\vec{v})\,\vec{v}. \end{cases} \tag{21}$$

Os Teoremas 4.1 e 4.2 garantem a existência, unicidade e regularidade das soluções $\vec{u}$ e $\vec{v}$ do Sistema (21). Além disso, como a matriz K é formada pelo produto interno entre $\nabla\varphi_i$ e $\nabla\varphi_j$, e as funções φ_i e φ_j formam o espaço de elementos finitos V_h, então segue que K é tridiagonal, simétrica e definida positiva. Portanto, (21) é um sistema linear possível e determinado.

5 RESULTADOS NUMÉRICOS

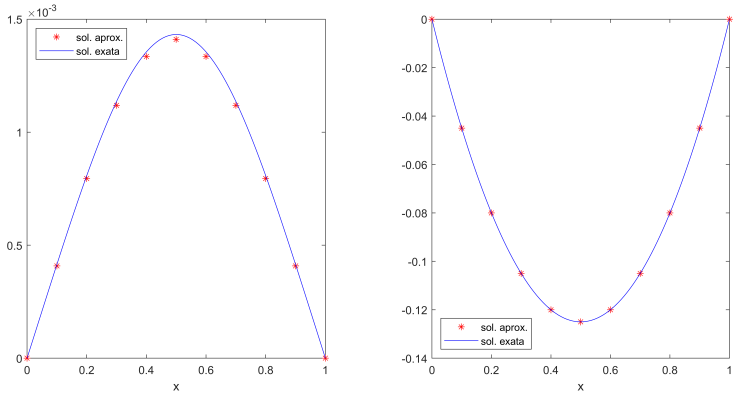
Nesta seção, apresentamos os resultados da implementação da teoria em Matlab. Primeiro validamos o código e, posteriormente, analisamos a ordem de convergência.

5.1 Exemplo 1

Para o primeiro exemplo vamos considerar $\Omega = [0, 1]$, a malha do domínio uniforme, o espaço de elementos finitos formado por funções de bases lineares, $f(x) = 1$ e $p = 1.5$. Então, as soluções exatas são

$$u(x) = \frac{x - x^4(2x^2 - 6x + 5)}{240} \quad \text{e} \quad v(x) = |\Delta u(x)|^{p-2}\,\Delta u(x) = \frac{x\,(x-1)}{2}.$$

As Fig. 1(a) e 1(b) exibem as soluções aproximadas $u_h(x)$ e $v_h(x)$ comparadas com as soluções exatas $u(x)$ e $v(x)$, respectivamente. Nesse caso, usamos $n = 10$ elementos finitos.

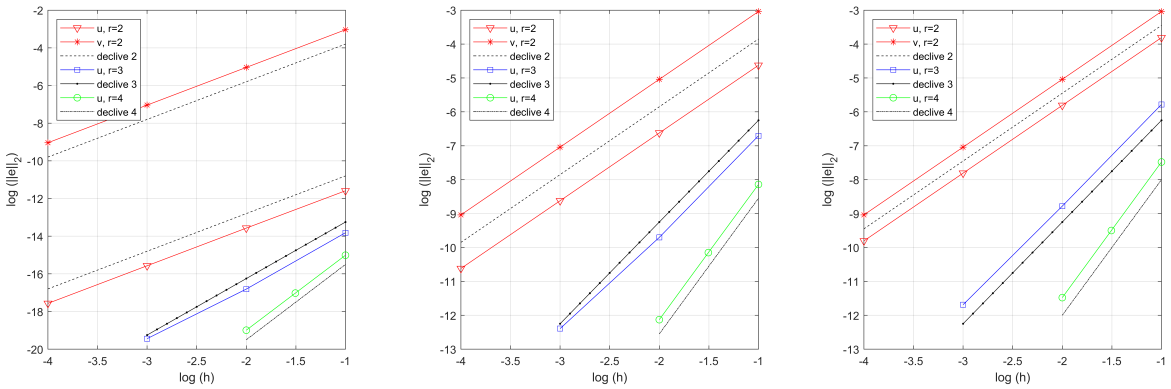


(a) $u(x)$ e $u_h(x)$. (b) $v(x)$ e $v_h(x)$.

Fig. 1: Soluções exatas $u(x)$ e $v(x)$ e soluções aproximadas $u_h(x)$ e $v_h(x)$.

A seguir, consideramos ainda $f = 1$, mas diferentes valores de p. Para cada valor de p, calculamos a solução exata. Nesse caso, as soluções aproximadas $u_h(x)$ e $v_h(x)$ estão suficientemente próximas das soluções exatas $u(x)$ e $v(x)$, portanto, optamos por não exibir os gráficos que realizam a comparação entre as soluções aproximadas e exatas.

Nas Fig. 2(a), 2(b) e 2(c) mostramos os gráficos da ordem de convergência para $p = 1.1, 1.5, 2$. Consideramos $n = 10, 100, 1000, 10000$ e, para cada caso, calculamos o erro na norma $L^2(\Omega)$.



(a) $p = 1.1$. (b) $p = 1.5$. (c) $p = 2$.

Fig. 2: Ordem de convergência para $p = 1.1, 1.5, 2$ com $r = 2, 3, 4$.

Notemos que a função $v(x)$ é um polinômio de grau 2 e, portanto, para $r \geq 3$, a solução aproximada é idêntica à solução exata. Assim, os erros dependem apenas do método utilizado para resolver o sistema, o que significa que não se pode observar a ordem de convergência. Portanto, nas Fig. 2(a), 2(b) e 2(c), o gráfico da ordem de convergência para $v_h(x)$ não é exibido ao utilizarmos as funções de base com grau 2 e 3, respectivamente, $r = 3$ e $r = 4$.

Destacamos que, nas Fig. 2(a), 2(b) e 2(c), os gráficos da ordem de convergência estão de acordo com os resultados do Teorema 4.3, ou seja, as simulações computacionais são

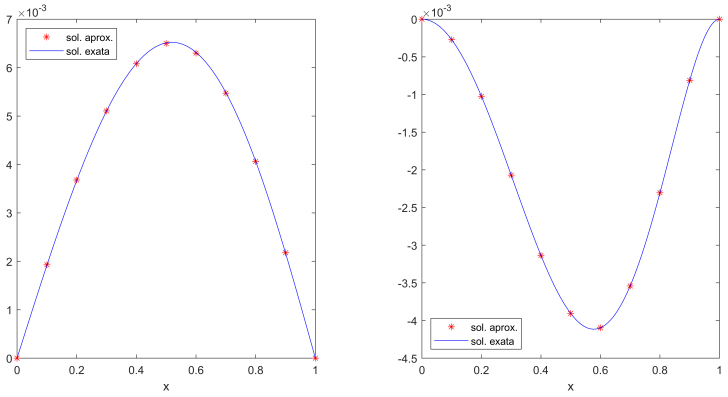
consistentes com o estudo analítico. Além disso, as estimativas (12) e (14) não dependem de p e, para $\|u - u_h\|_{L^2(\Omega)}$ e $\|v - v_h\|_{L^2(\Omega)}$, a ordem de convergência é $O(h^r)$ para funções de base formadas por polinômios de grau $r-1$, com $r \geq 2$.

5.2 Exemplo 2

Neste exemplo, consideramos $\Omega = [0, 1]$, a malha do domínio uniforme, o espaço de elementos finitos formado por funções de base lineares, $f(x) = -\frac{5x^4}{6} + \frac{2x^2}{3} - \frac{1}{18}$ e $p = 3$. Neste caso, as soluções exatas são

$$u(x) = \frac{x^5}{120} - \frac{x^3}{36} + \frac{7x}{360} \quad \text{e} \quad v(x) = \frac{x^2\left(x^4 - 2x^2 + 1\right)}{36}. \tag{22}$$

As Fig. 3(a) e 3(b) comparam as soluções aproximadas $u_h(x)$ e $v_h(x)$ com as soluções exatas $u(x)$ e $v(x)$, respectivamente. Nesse caso, $n = 10$ elementos finitos foram usados novamente.



(a) $u(x)$ e $u_h(x)$. (b) $v(x)$ e $v_h(x)$.

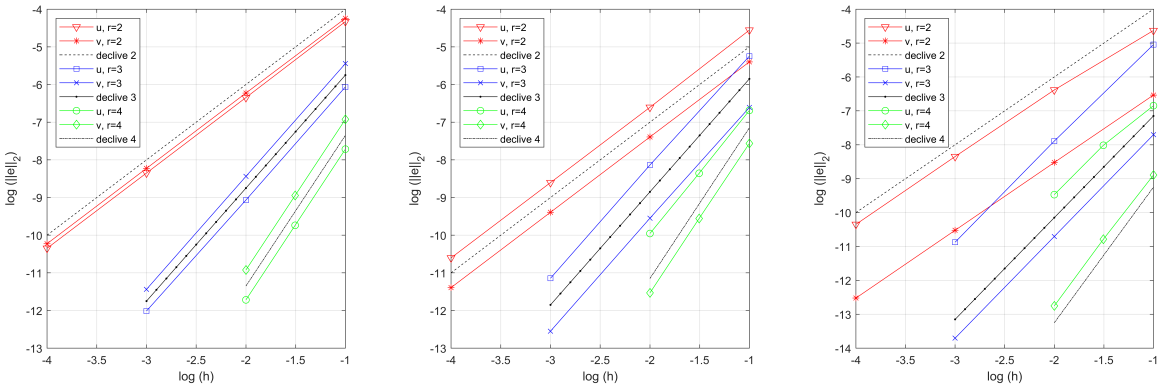
Fig. 3: Soluções exatas $u(x)$ e $v(x)$ e soluções aproximadas $u_h(x)$ e $v_h(x)$.

A seguir, alteramos os valores de p e obtemos a função $f(x)$ dada por

$$f(x) = x\,(p-1)\left(\frac{x}{6} - \frac{x^3}{6}\right)^{p-2} - (p-1)\,(p-2)\left(\frac{x}{6} - \frac{x^3}{6}\right)^{p-3}\left(\frac{x^2}{2} - \frac{1}{6}\right)^2.$$

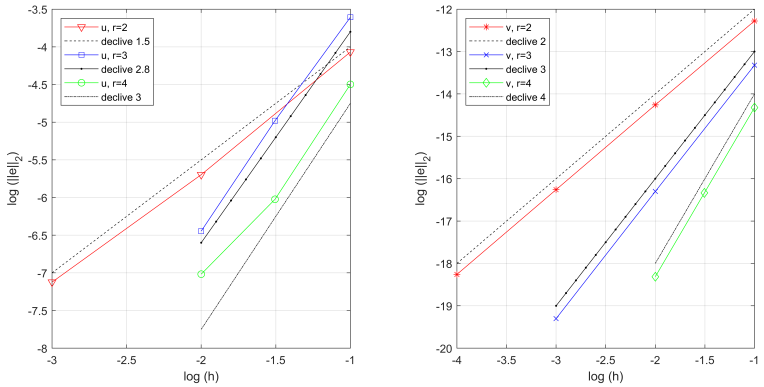
Novamente, as soluções aproximadas estão suficientemente próximas das soluções exatas, então optamos por omitir os gráficos que comparam as soluções aproximadas e exatas.

Nas Fig. 4, 5 e 6 exibimos os gráficos da ordem de convergência para $p = 3, 4, 5, 10, 25$. Novamente, consideramos $n = 10, 100, 1000, 10000$ e, para cada um desses valores, calculamos o erro na norma $L^2(\Omega)$.



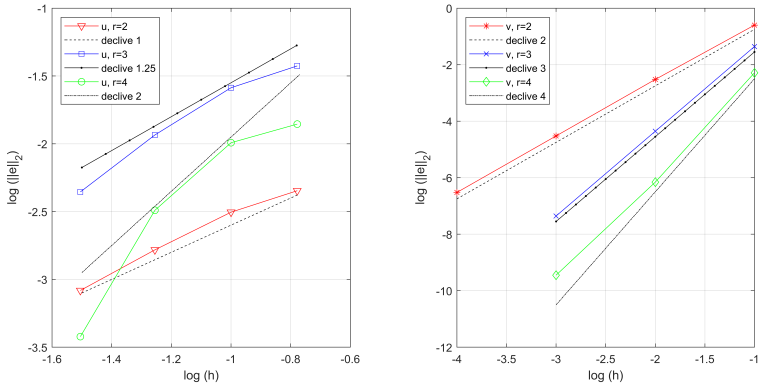
(a) $p = 3$. (b) $p = 4$. (c) $p = 5$.

Fig. 4: Ordem de convergência para $p = 3, 4, 5$ com $r = 2, 3, 4$.



(a) Função $u(x)$. (b) Função $v(x)$.

Fig. 5: Ordem de convergência para $p = 10$ com $r = 2, 3, 4$.



(a) Função $u(x)$. (b) Função $v(x)$.

Fig. 6: Ordem de convergência para $p = 25$ com $r = 2, 3, 4$.

Relembremos que, a estimativa (16) no Teorema 4.4 não depende de p, então nas Fig. 4(a), 4(b), 4(c), 5(b) e 6(b), temos que, para $\|v - v_h\|_{L^2(\Omega)}$, a ordem de convergência é $O(h^r)$ para as funções de base formadas por polinômios de grau $r-1$, com $r \geq 2$.

Na Fig. 4(a), para $\|u-u_h\|_{L^2(\Omega)}$, obtemos que a ordem de convergência é $O(h^r)$ quando usamos polinômios de grau $r-1$, com $r \geq 2$. Essa análise é idêntica para as Fig. 4(b) e 4(c), mas, para $r = 4$, notamos que a ordem de convergência é menor do que $O(h^4)$. Nas Fig. 5(a) e 6(a) a convergência existe mas a ordem de convergência é relativamente baixa, o que está de acordo com a teoria, pois no Teorema 4.4 a estimativa (18) depende de p e quando $p \to +\infty$ a ordem de convergência diminui.

6 CONCLUSÕES

Estudamos uma equação de viga não linear com o operador p-biharmônico. Reescrevendo o Problema (1) como um sistema de equações diferenciais adequadas, provamos a existência, unicidade e regularidade da solução fraca. Além disso, provamos a existência, unicidade e estabilidade da solução discreta usando elementos finitos formados por funções de base de graus arbitrários. Estabelecemos condições suficientes nos dados para obtermos as ordens de convergência ótimas para $1 < p < 2$ e provamos ordens de convergência subótimas, dependendo de p, para $p > 2$ na norma $L^2(\Omega)$. Por fim, implementamos o método no software Matlab, para $N = 1$, e realizamos simulações que confirmam a teoria.

7 *Agradecimentos*

Este trabalho foi parcialmente apoiado pelos projetos de investigação: FEDER através do Programa Operacional Factores de Competitividade, FCT - Fundação para a Ciência e a Tecnologia [Projeto UIDB/00212/2020].

O quarto autor foi apoiado pela FCT - Fundação para a Ciência e a Tecnologia, através do Centro de Matemática e Aplicações - Universidade da Beira Interior, sob o número de bolsa UI/BD/150794/2020, e também foi apoiado por MCTES, FSE e UE.

Referências

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure Appl. Math. (Amst.)*. Elsevier/Academic Press, Amsterdam, 2^{nd} edition, 2003.

[2] R. M. P. Almeida, J. C. M. Duque, J. Ferreira, and W. S. Panni. Mixed finite element method for a beam equation with the p-biharmonic operator. Preprint submitted to arXiv.org, 2022.

[3] N. Katzourakis and T. Pryer. On the numerical approximation of p-biharmonic and ∞-biharmonic functions. *Numer. Methods Partial Differential Equations*, 35(1):155–180, 2019.

[4] A. Novotný and I. Straškraba. *Introduction to the mathematical theory of compressible flow*, volume 27 of *Oxford Lecture Ser. Math. Appl.* Oxford University Press, Oxford, 2004.

[5] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Ser. Comput. Math.* Springer-Verlag, Berlin, 2^{nd} edition, 2006.

Trabalhos apresentados em formato de pôster

Análises *In Silico* e Dinâmica Molecular de Mutações da Proteína a-Syn Associadas ao Desenvolvimento de Doença de Parkinson

XIII Encontro Acadêmico de Modelagem Computacional

Aloma Nogueira Rebello da Silva[1], Gabriel Rodrigues Coutinho Pereira[1], Tiago Fleming Outeiro[2,3], Joelma Freire de Mesquita[1]

[1] Department of Genetics and Molecular Biology, Bioinformatics and Computational Biology Laboratory, Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Rio de Janeiro, Brazil

[2] Department of Experimental Neurodegeneration, Center for Biostructural Imaging of Neurodegeneration, University Medical Center Göttingen, Göttingen, Germany

[3] Max Planck Institute for Experimental Medicine, Göttingen, Germany

aloma.nogueira@gmail.com, joelma.mesquita@unirio.br

Introdução

A doença de Parkinson (DP) é a segunda doença neurodegenerativa mais comum, atingindo cerca de 6 milhões de pessoas no mundo. A DP não possui cura e é caracterizada por sintomas motores típicos que incluem tremor, instabilidade postural e bradicinesia. 5-10 % dos casos de DP têm origem familial, com mutações no gene SNCA, que codifica a proteína Alfa-sinucleína (a-Syn).

Objetivos

O objetivo deste trabalho foi avaliar os efeitos das mutações na proteína a-Syn associadas ao desenvolvimento da Doença de Parkinson.

Metodologia

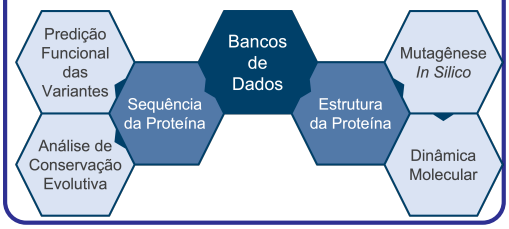
Os efeitos das mutações na função da proteína a-Syn foram preditos utilizando os algoritmos SIFT, PolyPhen-2, PhD-SNP, PANTHER, PMUT, PROVEAN e MutPred.

A análise evolutiva de conservação estrutural do a-Syn foi realizada com o algoritmo ConSurf.

A mutagênese in silico foi realizada utilizando o Mutator Plugin do VMD 1.9.3.

As simulações de dinâmica molecular (DM) da a-Syn WT e de suas variantes A30P, A53T e G51D foram realizadas em triplicatas utilizando o pacote GROMACS 2018.6 com um campo de força AMBER99SB-ILDN, uma caixa triclínica e água TIP3P.

Os sistemas moleculares foram neutralizados pela adição de íons Na+ e Cl e minimizados em 5000 etapas. Após a minimização do sistema, os conjuntos de NVT e NPT foram realizados a 1atm e 300K por 100ps.



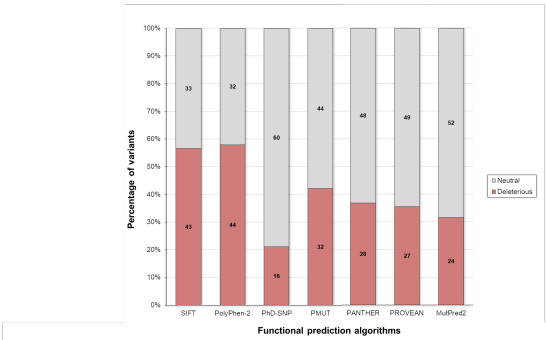
Referências

[1] Popova, B. 2015. *Posttranslational Modifications and Clearing of -Synuclein Aggregates in Yeast*

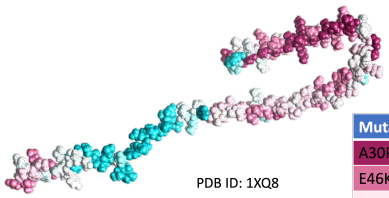
[2] Powers, R. 2015. *Metabolic Investigations of the Molecular Mechanisms Associated with Parkinson ' s Disease*

Resultados e Discussão

A mutação A30P foi considerada deletéria por todos os algoritmos.

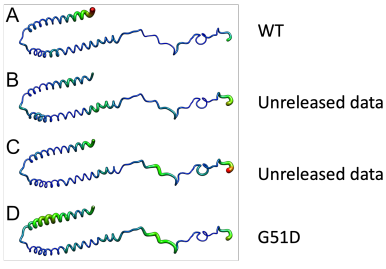


Os resultados do Consurf revelaram que as mutações A30P, E46K, H50Q, G51D, A53E, A53T ocorrem em locais conservados.



Mutation	ConSurf Score
A30P	9
E46K	8
H50Q	6
G51D	6
A53E	6
A53T	6

A análise DM apontou para uma diminuição da flexibilidade na região N-terminal das variantes analisadas e aumento da flexibilidade em suas regiões C-terminal em comparação com a WT. A análise da estrutura secundária sugeriu alterações na proteína com a variante G51D, principalmente pelo aumento do número de folhas formadas. A mutação G51D é conhecida por aumentar a propensão da a-Syn para formar folhas pregueadas, caracterizando no aumento na sua tendência de formar agregados tóxicos de proteínas.



B-factor results

Conclusão

Este trabalho sugere que as mutações da a-Syn afetam a estrutura e função da proteína, o que pode estar relacionado ao desenvolvimento de DP.

Parametrização para Triagem Virtual com SHMT de *Trypanosoma cruzi*

Ana Carolina Silva Bulla[1], Manuela Leal da Silva[1,2]

[1] Programa de Pós-Graduação em Biologia Computacional e Sistemas, Instituto Oswaldo Cruz

[2] Laboratório Integrado de Biologia Computacional e Pesquisa em Ciências Farmacêuticas, Universidade Federal do Rio de Janeiro

anabulla@aluno.fiocruz.br, manuela@macae.ufrj.br

Introdução

A doença de Chagas, também conhecida como Tripanossomíase Americana, é causada por protozoários da espécie *Trypanosoma cruzi* sendo uma das principais causas de doenças cardíacas no mundo com altos índices de morbimortalidade [1]. A transmissão do parasita ocorre principalmente por vetores artrópodes e por via oral, sendo o tratamento baseado nos antiparasitários, benznidazol e nifurtimox, ainda que seus perfis de segurança não sejam ideais e a eficiência pareça diminuir com o tempo da infecção primária. Dessa forma, esforços têm sido empregados tanto na busca de alvos terapêuticos como de potênciais antiparasitários. A triagem virtual (TV) têm se mostrado uma técnica eficiente para este fim uma vez que proporciona a predição de interação de milhares de compostos contra uma estrutura alvo otimizando o processo de seleção de ligantes promissores.

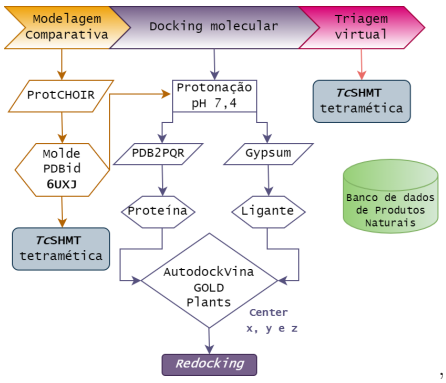
Objetivo

O trabalho teve como objetivo a obtenção de parâmetros para realização da TV com a Serina Hidroximetiltransferase de *T. cruzi* (TcSHMT), enzima envolvida na biossíntese de ácidos nucleicos [2], contra o banco de dados de moléculas da biodiversidade brasileira

Metodologia

Inicialmente um modelo tridimensional tetramérico para TcSHMT foi gerado por modelagem comparativa com ProtCHOIR sendo o molde a PDBid 6UXJ (*Glicine max*) onde o *RMSD* (*Root-mean-square deviation of atomic positions*) foi de 0.35 Å. Para a TV foram selecionados os programas AutoDock Vina (ADT), GOLD e PLANTS onde foram realizados estudos de *redocking* com a PDBid 6UXJ e seu ligante *5-formyltetrahydrofolate*. Para a caixa de simulação foram englobados os resíduos em até 5 Å de distância do ligante encontrados pela ferramenta *Zone* do *CHIMERA*.

Fluxograma do método empregado:



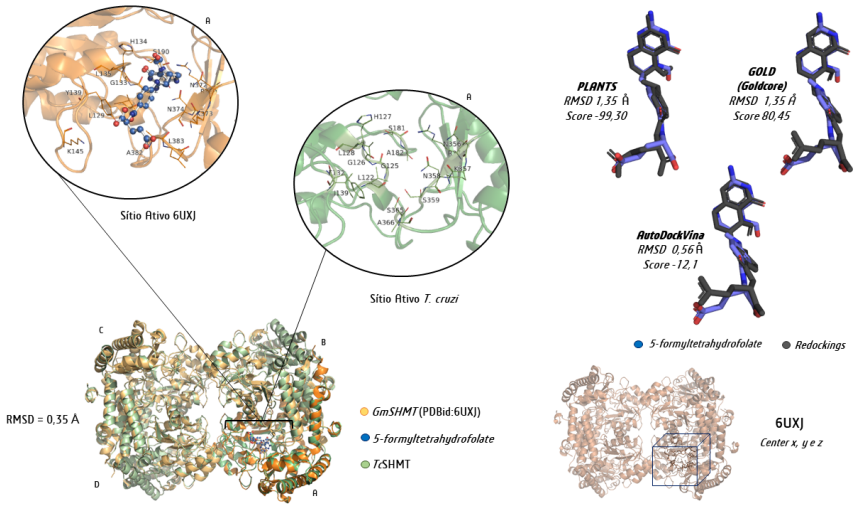
References

[1] World Health Organization. 2022. Chagas disease. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)>. Acesso em: 18/02/2022.

[2] CAPELLUTO, D. G. 2000. Purification and properties of serine hydroxymethyltransferase from *Trypanosoma cruzi*. In: *European Journal Biochemistry* v.267, pp. 712-719.

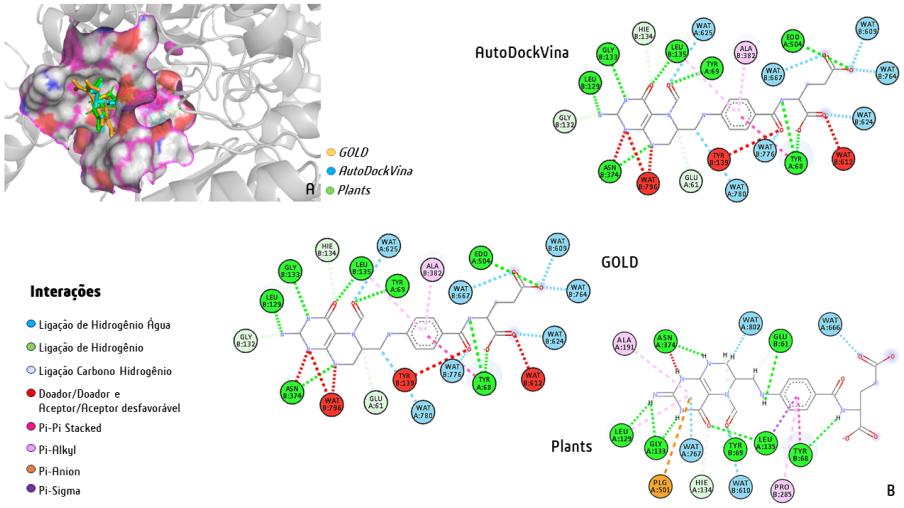
Resultados

Para as três simulações, a pose melhor classificada pela função de pontuação correspondeu a um RMSD < 2 Å. De maneira geral, os programas foram eficientes em reproduzirem a pose cristalográfica.



Resultados

Gráficos das interações gerados pelo *Discovery Studio* (B) após *redocking* com a GmSHMT e o *5-formyltetrahydrofolate*. Em (A), representada pela superfície com carbonos em magenta está a região do sítio ativo da GmSHMT com os ligantes encaixados pelos três programas.



Conclusão

Os resultados de RMSD indicaram a capacidade dos programas em reproduzir as interaçõos encontradas no cristal. Para continuação do trabalho, testaremos a capacidade dos programas em diferenciar moléculas bioativas e inativas. Posteriormente, seguiremos com o estudo de *docking* com a TcSHMT e o banco de dados de produtos naturais.

Mineração de dados aplicado ao uso de software de montanhismo em Petrópolis (RJ) - XIII EAMC.

Bernardo Garcez[1], Luana Pitzer[2]

[1] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Campus Petrópolis
[2] Universidade do Estado do Rio de Janeiro

bernardogarcez@gmail.com; pitzer.luana@hotmail.com

XIII Encontro Acadêmico de Modelagem Computacional

Introdução

O campo da ciência de dados é uma área amplamente estudada e seu crescimento tende a ser de forma exponencial. Empresas e instituições, cada vez mais, tomam o conhecimento da sua importância para traçar decisões mais assertivas e direcionadas. Podemos retirar informações extremamente relevantes para diversas aplicações, somente analisando e trabalhando dados. A aplicabilidade em estudos científicos por exemplo é fortemente utilizada, mesmo em diferentes áreas do conhecimento. Processos que reúnem um grande fluxo de dados (Big Data) podem facilitar a obtenção de informações, que segundo [1] é o "fenômeno da massificação de elementos de produção e armazenamento de dados, bem com os processos e tecnologias para extraí-los e analisá-los". Diante disso, esse estudo utilizou a análise de dados para pro- duzir um panorama acerca de um aplicativo de nave- gação outdoor atrelado à prática do montanhismo e o uso público em unidades de conservação (UC) em Petrópolis (RJ). O município dispõe de um relevo montanhoso em que a prática do montanhismo é cul- turalmente associada, sendo 60% do seu território protegido por UCs [2]. O objeto de estudo foi o site/aplicativo Wikiloc, conhecido entre montanhistas, com a funcionalidade de gravar rotas e trilhas feitas pelos usuários e compartilhá-las publicamente.

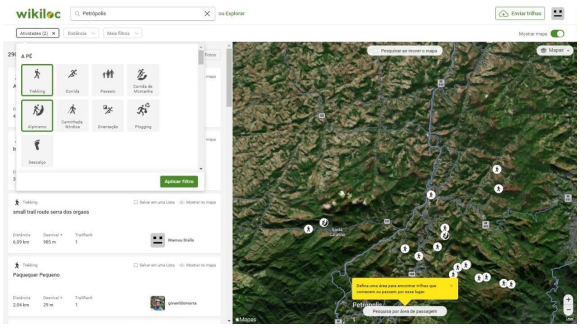
Metodologia

A metodologia foi de caráter descritiva, optou-se por uma abordagem quali-quantitativa, visto que as informações extraídas do site são numéricas e textuais.

A busca ocorreu no dia 22 de outubro de 2021, foi filtrada pelo território Petrópolis, e atividades denominadas "*Trekking*" e "Alpinismo", resultando em 2.904 trilhas.

A mineração dos dados efetuou-se utilizando o método de Web Scraping, coletou-se os dados de todas as trilhas feitas pelos usuários, o número de vezes que o trajeto foi visualizado e que foi efetiva- mente baixado. Tendo essa coleta feita, resultando em um data frame, sucedeu-se para a etapa da ma- nipulação e filtragem dos dados, realizada com o Jupyter como plataforma de desenvolvimento, com a linguagem Python e sua biblioteca Pandas.

Figura 1 – Site Wikiloc



Resultados

Tabela 1 – Amostragem dos dados coletados no Wikiloc

Trilha	UC	Nº de rotas	Baixadas	Visualizadas
Travessia Petrópolis - Teresópolis	PARNASO	487	14673	129857
Castelos do Açu	PARNASO	550	5390	39787
Alcobaça	PARNASO	113	5134	6545
Maria Comprida	APA PETRÓPOLIS	77	4650	5527
Pedra de Itaipava	APA PETRÓPOLIS	114	4466	8204
Monte de Milho	APA PETRÓPOLIS	15	2587	1582
Uricanal	PARNASO	26	2414	3508
Seio de Vênus	APA PETRÓPOLIS	46	2390	5263
Pico do Glória	PARNASO	13	2067	1148
Morro da Mensagem	APA PETRÓPOLIS	18	1405	3476
Travessia Araras - Secretário	APA PETRÓPOLIS	16	1310	1010
Palmares	APA PETRÓPOLIS	42	1117	5785
Pedra do Taquaril	FORA DE UC	28	975	1819
Pedra do Retiro	APA PETRÓPOLIS	59	958	9474
Mãe d'água	PARNASO	42	951	4862
Pedra da Índia	REBIO ARARAS	38	537	3931
Pedra do Elefante	MONA PEDRA DO ELEFANTE	39	437	3299
Morro do Alicate	PARNASO	25	340	3815
Cantagalo	APA PETRÓPOLIS	10	324	632
Pedra do Cone	PARNASO	9	279	1007
Cachoeira Véu da Noiva	PARNASO	71	225	4388
Pedra de Nogueira	APA PETRÓPOLIS	8	174	1455
Pedra da Cuca	FORA DE UC	19	93	1851
Travessia Cuiabá - Brejal	FORA DE UC	7	52	1333
Tapera do Morin	APA PETRÓPOLIS	4	51	855
Pedra do Juriti	FORA DE UC	4	50	505
Morro do Bonet	REBIO TINGUÁ	5	15	206
Total	-	**1885**	**53064**	**251124**
Indefinidos	Não se aplica	1019	8523	59417

As 5 trilhas mais baixadas são: a Travessia Petrópolis – Teresópolis (27,6%), seguidas pelos Castelos do Açu (10,1%), Alcobaça (9,6%), Maria Comprida (8,7%) e Pedra de Itaipava (8,4%). As trilhas filtradas (1.885) possuem 251.124 visualizações

Conclusão

No geral identificamos uma concentração de trilhas específicas no aplicativo/site, e outras com números não tão representativos quando se comparado, são indicados estudos futuros com a temática. A metodologia apresentou-se adequada, e a associação entre gestão ambiental e ciência de dados foi satisfatória.

Referências

[1] F. Amaral . Introdução à Ciência de Dados. Min-eração de Dados e Big Data. Editora Alta Books,2016.

[2] T. Freitas, N. Moura, B. Fateicha, B. C. Santos, L. Pessoa, M. Salomão, M. Porretti, F. Pessoa. Unidades De Conservação em Petrópolis (RJ): um ensaio sobre suas características e potenciais. In 9º Simpósio de Gestão Ambiental e Biodiver- sidade, Três Rios, 2020.

Uma comparação entre o Método das Diferenças Finitas e o Método das Soluções Fundamentais na Equação de Laplace

Bryan Aoliabe Siqueira[1], Wilian Jeronimo dos Santos[2]

[1,2] Programa de Pós−Graduação em Modelagem Matemática e Computacional – Instituto de Ciências Exatas, Seropédica/RJ, Brasil

bryansiqueira.mtm@gmail.com, wilianj@gmail.com

Introdução

A equação de Laplace é uma equação diferencial parcial do tipo elíptica. Neste trabalho, apresenta-se a solução numérica da referida equação em duas dimensões, com condições de contorno do tipo Dirichlet. Para tal, faz-se o uso do Método das Diferenças Finitas e do Método das Soluções Fundamentais. O primeiro, busca aproximar a solução da equação utilizando a série de Taylor em pontos específicos que se originam através da construção de uma malha. Já o segundo método, é um dos diversos tipos de métodos sem malha, onde a solução é calculada em pontos arbitrários sobre o domínio. Os programas foram desenvolvidos no software *MATLAB* ®, comparando a solução numérica com a solução analítica do problema proposto, assim como, o tempo computacional para a execução dos programas criados.

A Equação de Laplace

A equação diferencial parcial considerada é dada por:

$$\nabla^2 u(x,y) \equiv \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = 0, \quad (1)$$

para (x,y) no conjunto $R = \{(x,y)|\ 0 < x < 0.5, 0 < y < 0.5\}$, sujeita às condições de contorno $u(0,y) = 0$, $u(x,0) = 0$, $u(x,0.5) = 200x$ e $u(0.5,y) = 200y$.

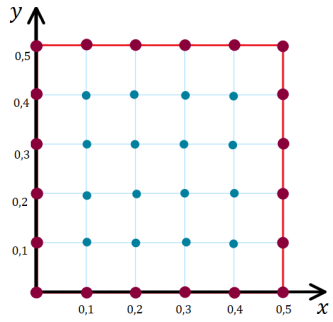
De acordo com [1], a solução exata para a equação acima, com suas respectivas condições de contorno, é dada por:

$$u(x,y) = 400xy. \quad (2)$$

Os resultados obtidos pelos dois métodos numéricos utilizados serão comparados com (2).

Método das Diferenças Finitas

Considera-se a seguinte malha para o domínio de (1), de modo que a solução será calculada nos pontos em azul.



Utilizando diferenças centradas nas variáveis x e y, a fórmula de diferenças finitas para o problema (1), sob as condições da malha já construída é dada por:

$$4w_{ij} - w_{i+1,j} - w_{i-1,j} - w_{i,j+1} - w_{i,j-1} = 0, \quad (3)$$

onde $w(i,j)$ é a solução aproximada.

Referências

[1] Douglas J. Faires Richard L. Burden and Annette M. Burden. *Análise Numérica*. Cengage Learning, São Paulo, 2017.

[2] José Alberto Cuminato and Messias Meneguette Junior. *Discretização de Equações Diferenciais Parciais - Técnicas de Diferenças Finitas*. SBM, Rio de Janeiro, 2013.

[3] Gregory E. Fasshauer. *Meshfree Approximation Methods with Matlab*. World Scientific, USA, 2007.

Método das Soluções Fundamentais

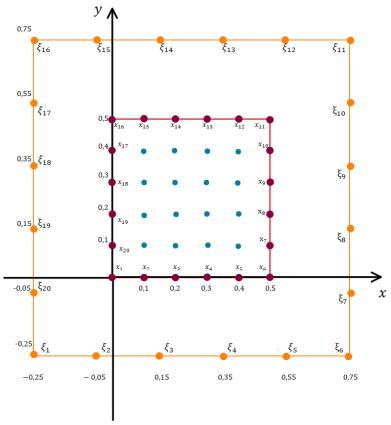
No Método das Soluções Fundamentais, admite-se soluções da forma:

$$U_i = \sum_{j=1}^{N} u^*_{ij} c_j, \quad (4)$$

onde U_i é a solução, i representa o conjunto de pontos de colocação no contorno, j representa o conjunto de pontos fonte ξ externos ao contorno, u^* é a solução fundamental no ponto x_i, c são as constantes à determinar e N é o número de pontos de colocação, igual ao número de pontos fonte. Para a equação de interesse, a solução fundamental é da forma:

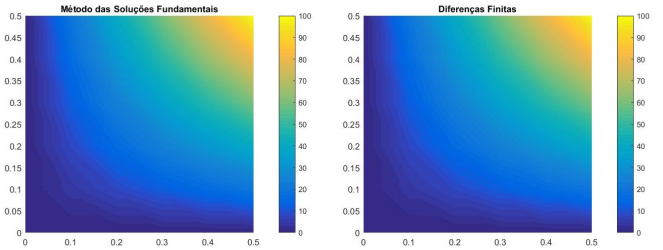
$$u^* = \frac{1}{2\pi} \ln\left[\frac{1}{r(x,\xi)}\right], \quad (5)$$

onde r é a distância euclidiana entre o ponto de colocação x_i e o ponto fonte ξ_j.



Resultados e Tempo de Execução do Programa

Os resultados gráficos com ambos os métodos numéricos foram:



Quando ao tempo de simulação, o programa desenvolvido para o Método das Diferenças Finitas levou 1.653422 segundos, enquanto o programa para o Método das Soluções Fundamentais foi de 3.116092 segundos.
Para verificar a melhor aproximação, utiliza-se a distância induzida pela norma l_2, de onde obtém-se que para o MSF temos $3.4674e^{-04}$ e para o MDF 0.

Conclusões

Ambos os métodos apresentaram boas soluções para o problema proposto, visto sua simplicidade, porém, o Método das Diferenças Finitas se destaca, visto que o tempo para execução do programa foi notavelmente menor, e também, as soluções obtidas com este método não produziram erros quando comparados a solução exata. Este fato se justifica pelo erro de truncamento ser zero em cada passo (devido a expansão em série de Taylor) e também, não houveram erros de arrendamentos visto a simplicidade do sistema gerado.

In Silico Characterization of the A4V and D90A Variants of Human SOD1 protein using Molecular Dynamics and Machine Learning

Gabriel Rodrigues Coutinho Pereira[1], Joelma Freire de Mesquita[1]
[1]Department of Genetics and Molecular Biology, Federal University of the State of Rio de Janeiro.
gabrielkytz@hotmail.com; joelma.mesquita@unirio.br

Introduction

Amyotrophic lateral sclerosis (ALS) is the most frequent motor neurodegenerative disorder in adults [1]. Missense mutations in superoxide dismutase 1 (SOD1), a major cytoplasmic antioxidant enzyme, are associated with the development of ALS. The A4V and D90A variants account for approximately half of all ALS-SOD1 cases in the United States and Europe [2]

Objectives

The objective of this work is to characterize *in silico* the structural and functional effects of A4V and D90A variants on human SOD1 protein. Understanding the effects of SOD1 mutations on protein structure facilitates the design of further experiments and provides relevant information on the molecular mechanism of pathology, which may contribute to improvements in existing treatments for ALS [3].

Materials and Methods

Three dimensional structures of A4V and D90A protein variants were computationally modeled in the VMD-1.9.1 package using the experimentally determined structure of wild-type SOD1 (PDB ID: 2C9V) as the template. Molecular dynamics (MD) simulations of the wild-type SOD1 protein and its variants A4V and D90A were performed in triplicates using the GROMACS-2018.8 package and AMBER99SB-ILDN force-field. TIP3P water molecules were added to a dodecahedral box system, which was neutralized by the addition of Na+ Cl - ions and then minimized. The system also had its temperature and pressure equilibrated at 1atm and 300K before the start of the simulations, which lasted 300ns. The MD trajectories were concatenated, and the following parameters were analyzed using GROMACS distribution programs: RMSD, RMSF, and essential dynamics [4].

References

1- CALLISTER, J. B.; PICKERING-BROWN, S. M. Pathogenesis/genetics of frontotemporal dementia and how it relates to ALS. Experimental Neurology, v. 262, n. PB, p. 84–90, 2014.

2- 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, 2014.

3- DE BAETS, G. et al. SNPeffect 4.0: On-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Research, v. 40, n. Database, p. D935–D939, 2012.

4- PEREIRA, G. R. C.; DE AZEVEDO ABRAHIM VIEIRA, B.; DE MESQUITA, J. F. Comprehensive in silico analysis and molecular dynamics of the superoxide dismutase 1 (SOD1) variants related to amyotrophic lateral sclerosis. PLoS ONE, v. 16, n. 2 February, p. 1–27, 2021

Results

The triplicates for wild-type SOD1 and its variants presented a similar behavior throughout the simulation. The establishment of a plateau in RMSD values after approximately 150ns indicates that the protein structures fluctuate around average stable conformations and, consequently, system equilibration. The MD analyses of variants A4V and D90A pointed to flexibility and essential dynamics alterations at the electrostatic and metal-binding loops, which are functional domains indispensable for SOD1 enzymatic activity, substrate guidance, and structural stability. Considering that structural flexibility and dynamics are key factors that drive protein interactions, our findings indicate that A4V and D90A may affect SOD1 interactions, particularly at the functional loops. A well-accepted hypothesis suggests that dynamics and structural alterations at the electrostatic and metal-binding loops of SOD1 could lead to ALS possibly through a toxic mechanism involving aberrant protein interactions triggering aggregation.

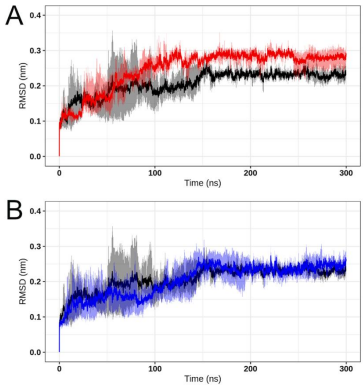


Figure 1. RMSD analysis of wild-type SOD1 and variants. The RMSD values for the backbone atoms at 300K are shown as a function of time. The means (solid lines) and confidence intervals (smooth lines) are displayed for the SOD1 wild-type (black), variant A4V (red), and variant D90A (blue).

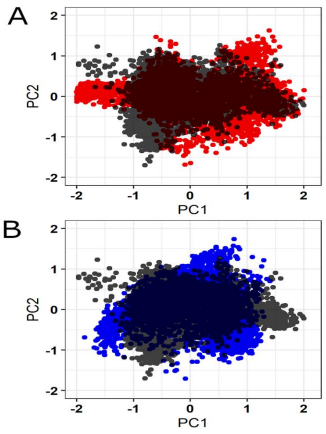


Figure 3. Projections for the first two principal components extracted from the MD trajectories. The projections for the PC1 and PC2 are shown as a scatter plot. Wild-Type SOD1 is represented in black, variant A4V is represented in red, and D90A is represented in blue.

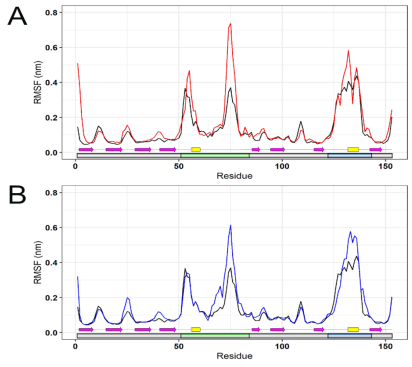


Figure 2: RMSF for the backbone atoms. RMSD for the backbone atoms at 300K are shown per residue. Wild-type SOD1 is represented in red, and variant D90A is represented in blue. A schematic representation of SOD1 secondary structure and functional domains are also shown for further comparison. Beta-sheets are represented by pink arrows, while alpha-helices are represented by the yellow barrels. The metal-binding and electrostatic loops of SOD1 are highlighted in green and blue, respectively.

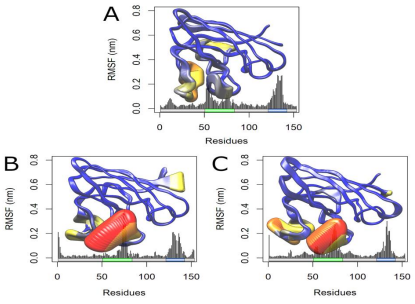


Figure 4. RMSF contribution to PC1. The RMSF contribution of each amino-acid to PC1 is shown simultaneously as a line plot and a projection on the protein structure. The amino-acids were colored and sized according to their RMSF values, following a coloring-thickness scale that varies from blue and thin (low fluctuations) to red and thick (high fluctuations). Schematic representations of the SOD1 functional loops and are shown to further comparison. The metal-binding and electrostatic loops of SOD1 are highlighted in green and blue, respectively. (A) Wild-type SOD1 to PC1. (B) Variant A4V to PC1. (C) Variant D90A to PC1.

Conclusions

Our findings pointed to flexibility and essential dynamics alterations at the A4V and D90A variants at the electrostatic and metal-binding loops. These alterations may have harmful implications for SOD1 and possibly explain their association with the development of ALS, given that these functional regions are known to be involved in protein aggregation.

XV Encontro Acadêmico de Modelagem Computacional

ANÁLISE *IN SILICO* DE VARIANTES GENÉTICAS DA TRIPTOFANO HIDROXILASE 2 HUMANA

Gabriela Fontoura Borges[1], Gabriel Rodrigues Coutinho Pereira[1], Joelma Freire de Mesquita[1]
[1] Departmento de Genética e Biologia Molecular
gfbfontoura@gmail.com, gabrielkytz@hotmail.com, jomesquita@gmail.com

Introdução

A proteína Triptofano Hidroxilase 2 (TPH2) está envolvida diretamente na primeira etapa da síntese da serotonina, neurotransmissor capaz de interferir em diversas funções biológicas, desde apetite, homeostase metabólica, sono e processos neuroendócrinos, até controle do comportamento, emoções e respostas induzidas de estressores ambientais/externos. Mutações na proteína TPH2 estão associadas ao desenvolvimento de diferentes transtornos psiquiátricos que apresentam elevada incidência global (PORTAS; BIORVATN; URSIN, 2000). Entretanto, grande parte dessas mutações ainda não foram caracterizadas quanto à seus efeitos. Caracterizar tais mutações poderia beneficiar o diagnóstico e auxiliar o tratamento dessas patologias altamente subnotificadas (RITCHIE et al., 2018).

Objetivos

O objetivo desse trabalho é caracterizar utilizando simulações computacionais possíveis efeitos funcionais e de estabilidade ocasionados por mutações do tipo variantes de nucleotídeo único na proteína Triptofano Hidroxilase 2 Humana (TPH2).

Material e Métodos

Seguindo a metodologia desenvolvida por nosso grupo (PEREIRA; ABRAHIM-VIEIRA; DE MESQUITA, 2021), métodos in silico foram aplicados ao estudo de variantes genéticas da TPH2 compiladas a partir de banco de dados, como UNIPROT, OMIM e dbSNP, a fim de elucidar seus efeitos funcionais e estruturais. Os efeitos funcionais foram preditos utilizando dez algoritmos que utilizam diferentes estratégias, dentre eles: SNAP2, SNP&GO, PolyPhen-2, PMUT, PhD-SNP, MutPred2, Provean, SIFT, Panther e Predict-SNP. Além disso, o impacto das mutações na estabilidade da proteína também foi predito utilizando três algoritmos: I-MUTANT 3.0, INPS e FoldX. Por fim a conservação evolutiva da TPH2 foi predita no servidor ConSurf utilizando a sequência da proteína nativa.

Referências

[1] PORTAS, C.M.; BIORVATN, B.; URSIN, R. Serotonin and the sleep/wake cycle:special emphasis on microdialysis studies. Prog. Neurobiol., v.60, n.1, p.13-35, 2000.

[2] Ritchie, H.; Dattani, S.; Roser, M. Mental Health. Our World in Data, 2018. Disponível: https://ourworldindata.org/mental-health

[3] Pereira GRC, Vieira BAA, De Mesquita JF. Comprehensive in silico analysis and molecular dynamics of the superoxide dismutase 1 (SOD1) variants related to amyotrophic lateral sclerosis. PLoS One. 2021 Feb 25;16(2):e0247841. doi: 10.1371/journal.pone.0247841. PMID: 33630959; PMCID: PMC7906464.

Resultados

Trezentas e oitenta e quatro mutações da proteína TPH2 humana foram compiladas. A análise de predição funcional apontou para uma elevada taxa de predição deletéria para as variantes da TPH2 humana. Dentre elas, 48 foram preditas como deletérias por todos os algoritmos, sugerindo que tais mutações poderiam ser danosas para a TPH2. Além disso, a análise de predição de estabilidade indicou que a maioria das mutações foi preditas de forma consenso por reduzir a estabilidade da proteína, o que poderia afetar sua atividade enzimática. A análise do SNPEffect, por sua vez, sugeriu que 26 mutações afetaram a propensão amiloide da proteína, 37 mutações afetaram a agregação proteica, e 15 mutações afetaram a tendência de ligação a chaperonas. Por fim, a análise de conservação evolutiva indicou que as mutações da TPH2 afetaram majoritariamente posições conservadas da proteína.

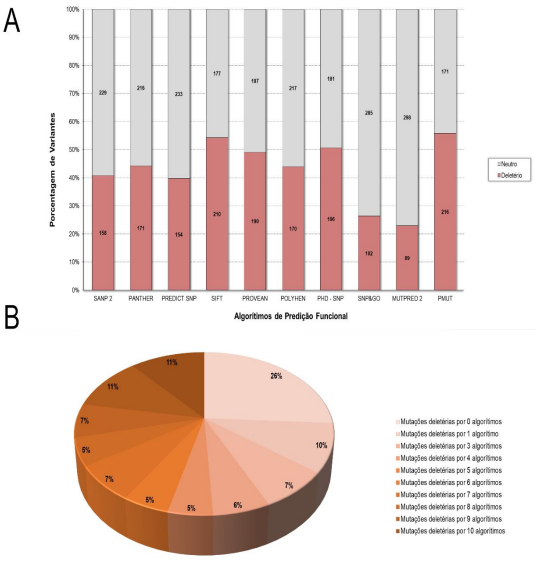


Figura 1. Análise de predição funcional A) O gráfico de barras mostra a porcentagem de variantes preditas como deletérias (vermelho) e neutras (cinza) em cada um dos algoritmos utilizados. (B) O gráfico de setores mostra a porcentagem total de mutações em cada faixa de predição deletéria, variando de zero (laranja claro) até 10 predições deletérias (laranja escuro).

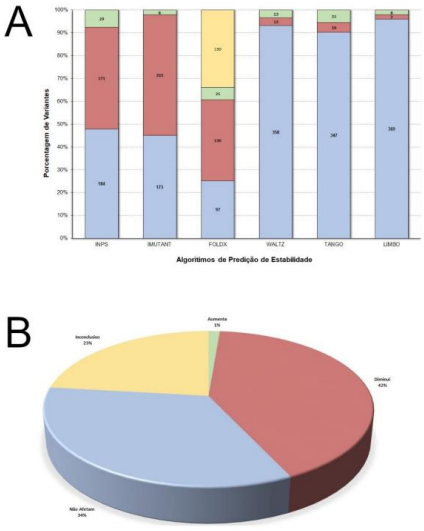


Figura 2. Análise de predição de estabilidade (A) O gráfico de barras mostra a quantidade de predições de estabilidade nos algoritmos INPS, IMutant e FoldX, bem como a quantidade de predições pelos algoritmos WALTZ (propensão amiloide), TANGO (agregação proteica) e LIMBO (ligação a chaperonas) do SNPEffect. As barras azuis representam predições do tipo "Não afeta", as barras vermelhas representam predições do tipo "diminui", as barras verdes representam predições do tipo "aumenta", enquanto que as barras amarelas representam predições do tipo "sem resultado". (B) O gráfico de setores mostra os resultados da análise de predição de estabilidade consenso. O setor azul representa mutações 38 preditas por não afetar a estabilidade, o setor vermelho representa mutações preditas por diminuir a estabilidade, o setor verde representa mutações preditas por aumentar a estabilidade, ao passo que o setor amarelo representa classificações inconclusivas para as mutações.

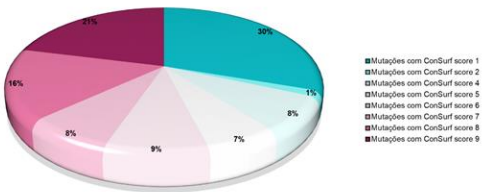


Figura 3. Análise de conservação evolutiva no servidor ConSurf O gráfico de setores mostra a proporção de mutações que afetaram desde posições muito variáveis (grau de conservação 1, representada pela cor ciano) até posições muito conservadas (grau de conservação 9, representada pela cor bordô).

Conclusão

Vinte e sete mutações receberam grau máximo de conservação na análise do ConSurf e também foram preditas como deletérias por todos os algoritmos de predição funcional utilizados, sendo, consequentemente, um importante alvo para futuros experimentos.

Implementação de diferentes modelos para a condutividade hidráulica na solução numérica da Equação de Richards

Caroline da Costa Souza[1], João Gabriel de Souza Debossam[1], Grazione de Souza[1], Helio Pedro Amaral Souto[1]

[1] Instituto Politécnico, Universidade do Estado do Rio de Janeiro

carolinecostasouza08@gmail.com, joaodebossam@gmail.com, gsouza@iprj.uerj.br, helio@iprj.uerj.br

Introdução

Tem-se aqui resultados preliminares de um estudo da dinâmica de infiltração de água no subsolo utilizando a Equação de Richards [2, 3]. O método de diferenças finitas, uma formulação numérica totalmente implícita e uma linearização pelo método de Picard foram aplicados, considerando-se o escoamento unidimensional. Nesta etapa, o modelo de Gardner [1]) foi adotado.

Modelo Físico-matemático

No ciclo hidrológico há a precipitação, em que parte da água da chuva se infiltra no solo. O modelo mais comum para o escoamento de infiltração em meios porosos não-saturados, é a equação Richards, uma equação diferencial parcial (EDP) não-linear e parabólica, que pode ser expressa para a carga hidráulica, no caso 1D na direção x, como

$$C(h)\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(K_x(h)\frac{\partial h}{\partial x}\right) \tag{1}$$

no qual,

$$C(h) = \frac{\partial \theta}{\partial h} \tag{2}$$

onde, $C(h)$ é a capacidade específica, θ é a umidade do solo, $K(h)$ é a condutividade hidráulica do solo não saturado e h é o potencial hidráulico.

Alguns modelos para K e C

Na EDP supracitada, é possível se considerar, e.g.,

- Gardner

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \exp\left(\frac{h}{h_g}\right) \tag{3}$$

$$K(h) = K_s \exp\left(\frac{h}{h_g}\right) \tag{4}$$

$$C(h) = \frac{(\theta_s - \theta_r)}{h_g} = \exp\left(\frac{h}{h_g}\right) \tag{5}$$

- Van Genuchten

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} = [1 + |\alpha.h|^n]^{-m} \tag{6}$$

$$K(h) = K_s {S_e}^{0.5}[1 - (1 - {S_e}^{1/m})^m]^2 \tag{7}$$

$$C(h) = \frac{\alpha.n.m(\theta_s - \theta_r)|\alpha.h|^{n-1}}{(1 + |\alpha.h|^n)^{m+1}} \tag{8}$$

onde S_e é a saturação efetiva, θ_r a umidade residual, θ_s a umidade de saturação, K_s a condutividade hidráulica da situação de saturação, α é relacionado ao tamanho médio dos poros e n e m são relacionados à distribuição do tamanho dos poros, $m = 1 - n^{-1}$.

Referências

[1] D. Gasiorowski and T. Kolerski, 2020. Numerical Solution of the Two-Dimensional Richards Equation Using Alternate Splitting Methods for Dimensional Decomposition *Water*, vol. 12, 1-20.

[2] M. Kuraz and P. Mayer and P. Pech, P. 2014. Solving the nonlinear Richards equation model with adaptive domain decomposition. *Journal of Computational and Applied Mathematics*, vol. 270, 2-11.

[3] M. Farthing and F. Ogden, 2017. Numerical solution of Richards' equation: A review of advances and challenges, *Soil Science Society of America Journal*, vol. 81, 1257-1269.

Metodologia numérica

Discretiza-se a Eq. (1) pelo método CVFD (*Control Volume-Finite Differences*), aplicando-se uma malha numérica espacial como a da Fig. 1.

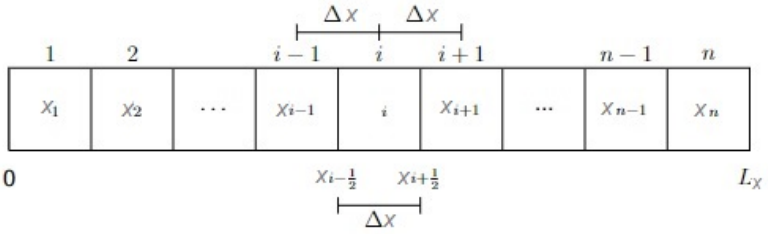


Fig. 1: Malha numérica no espaço.

Foram aplicadas aproximações atrasada no tempo e centrada no espaço, obtendo-se uma formulação totalmente implícita:

$$\frac{C_i^{n+1}\Delta x^2 h_i^{n+1}}{\Delta t} = K_{i+1/2}^{n+1}h_{i+1}^{n+1} + K_{i-1/2}^{n+1}h_{i-1}^{n+1} - h_i^{n+1}(K_{i+1/2}^{n+1} + K_{i-1/2}^{n+1}) + \frac{C_i^{n+1}\Delta x^2 h_i^n}{\Delta t}. \tag{9}$$

O sistema de equações algébricas não-lineares é resolvido usando os métodos de Picard e SOR (Fig. 2).

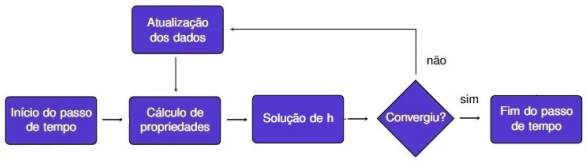


Fig. 2: Fluxograma para um passo de tempo.

Resultados

Nesta etapa do trabalho apresenta-se resultados com o uso do modelo de Gardner, em estudos de refinamento de malha computacional e de variações do tempo máximo de infiltração, de K_s e de h_g, apresentados na Fig. 3. Na análise de sensibilidade à mudança de parâmetros foi utilizado $n_x = 100$, enquanto que no refinamento de malha $h_g = 2$ m, $K_s = 10^{-5}$ m/s e $t = 1000$ s.

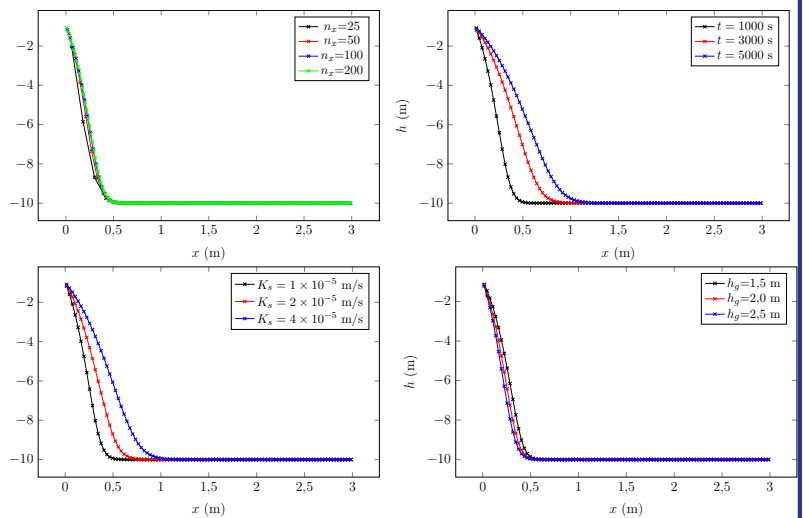


Fig. 3: Resultados para h usando o modelo de Gardner.

Conclusões

Até o momento foi possível se implementar os modelos de Gardner e de van Genuchten, sendo aqui apresentados resultados para Gardner que foram considerados satisfatórios. Outros modelos estão em estudo, e deve-se prosseguir no projeto em etapas de verificação e maior discussão dos resultados. Um caso de interesse futuro é àquele no qual uma heterogeneidade em K_s é considerada.

Avaliação de limitadores de fluxo TVD na simulação do escoamento bifásico em reservatórios de petróleo

Gillyan Macário da Silva[1], Juan Diego dos Santos Heringer[1], Grazione de Souza[1] Helio Pedro Amaral Souto[1]

[1] Instituto Politécnico, Universidade do Estado do Rio de Janeiro

gillyan.silva@gmail.com, jheringer21@gmail.com, gsouza@iprj.uerj.br, helio@iprj.uerj.br

Introdução

Realiza-se aqui um estudo de simulação numérica para a injeção de água para promover a recuperação de óleo [1]. O método IMPES, no qual a pressão do óleo é determinada implicitamente enquanto a saturação da água é obtida explicitamente, é adotado para se determinar a dinâmica do escoamento [2]. O código computacional foi desenvolvido aplicando-se a linguagem Python [4], cuja utilização tem crescido recentemente. Escolheu-se o caso particular do escoamento unidimensional na direção x, para se estudar o efeito de diferentes limitadores de fluxo TVD (*Total Variation Diminishing*) [3] na solução para a frente de avanço de água.

Modelagem

As equações de balanço que modelam o escoamento bifásico, unidimensional na direção x, isotérmico, na ausência de pressão capilar, para meio totalmente saturado e rocha incompressível, são

$$\frac{\partial}{\partial t}\left[k_x A_x \frac{k_{ro}}{\mu_o B_o}\frac{\partial P_o}{\partial x}\right]dx = V_b\phi\frac{\partial}{\partial t}\left[\frac{(1-S_w)}{B_o}\right] \quad (1)$$

para a fase não-molhante, considerada ligeiramente compressível, e

$$\frac{\partial}{\partial t}\left[k_x A_x \frac{k_{rw}}{\mu_w B_w}\frac{\partial P_o}{\partial x}\right]dx = \phi V_b\frac{\partial S_w}{\partial t} \quad (2)$$

para a fase molhante, considerada incompressível. Tem-se que k_x é a permeabilidade absoluta na direção x, A_x a área de seção transversal, k_{rl} a permeabilidade relativa da fase l, μ_l a viscosidade da fase l, B_l o fator-volume-formação da fase l, P_o a pressão do óleo, V_b o volume total (rocha mais fluido), S_w a saturação da fase água e ϕ a porosidade.

Adotou-se as permeabilidades relativas das fases água e óleo, respectivamente, dadas por

$$k_{rw} = 0,6S_{wn}^2 \quad \text{e} \quad k_{ro} = \left(1 - S_{wn}^2\right) \quad (3)$$

sendo

$$S_{wn} = \frac{S_w - S_{iw}}{1 - S_{iw}}, \quad (4)$$

que se refere à saturação normalizada da fase molhante e S_{iw} é a saturação irredutível da fase molhante.

Do ponto de vista de condições auxiliares, tem-se os valores iniciais de P_o e S_w e as condições de contorno dadas na Fig. 1.

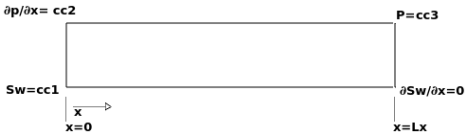


Fig. 1: Condições de contorno.

Metodologia numérica

Utilizou-se a metodologia CVFD (*Control Volume-Finite Difference*) e expansões conservativas na discretização das equações governantes, considerando a malha espacial apresentada na Fig. 2.

Fig. 2: Malha numérica.

Obteve-se as equações discretizadas para as fases não-molhante e molhante, respectivamente,

$$\Delta\left[T_o\left(\Delta P_o\right)\right]_i = C_{op,i}\left(P_{o,i}^{n+1} - P_{o,i}^{n}\right) + C_{os,i}\left(S_w^{n+1} - S_w^{n}\right) \quad \text{e} \quad \Delta\left[T_w\left(\Delta P_o\right)\right]_i = C_{ws,i}\left(S_w^{n+1} - S_w^{n}\right) \quad (5)$$

onde

$$C_{op,i} = \frac{\phi V_b}{\Delta t}\left(\frac{1}{B_o}\right)_i\left(1 - S_w^n\right)_i, \qquad C_{os,i} = -\frac{\phi V_b}{\Delta t}\left(\frac{1}{B_o}\right)_i^{n+1}, \qquad C_{ws,i} = \frac{\phi V_b}{\Delta t}\left(\frac{1}{B_w}\right)_i. \quad (6)$$

sendo Δ um operador de diferenças e T_l a transmissibilidade da fase l.

Utilizando-se o método IMPES (*Implicit Pressure, Explicit Saturation*, tem-se a seguinte equação P_o, onde A^n é um coeficiente utilizado na construção do método e função dos fatores-volume-formação,

$$P_{o_P}^{n+1}\left(T_{o_e}^n + A^n T_{w_e}^n + A^n T_{w_w}^n + C_{op_P}\right) - P_{o_E}^{n+1}\left(T_{o_e}^n + A^n T_{w_e}^n\right) - P_{o_W}^{n+1}\left(T_{o_w}^n + A^n T_{w_w}^n\right) = P_{o_P}^n C_{op,P}, \quad (7)$$

a qual já considera a aplicação de operadores diferenciais. Para a saturação da água, tem-se

$$S_w^{n+1} = \frac{1}{C_{ws}}\left[T_{w_e}^n\left(P_{o_E}^{n+1} - P_{o_P}^{n+1}\right) + T_{w_w}^n\left(P_{o_W}^{n+1} - P_{o_P}^{n+1}\right)\right] + S_w^n. \quad (8)$$

As transmissibilidades são aproximadas utilizando médias harmônica (rocha e geometria), aritmética (dependência na pressão) e no caso da permeabilidade relativa, com o uso do *upwind* de 1^a ordem, $k_{rl_e} = k_{rl_P}$, se $v_e \geq 0$, ou k_{rl_E}, caso contrário. Esta aproximação apresenta expressiva difusão numérica, motivando o estudo feito neste trabalho.

Resultados

Testes foram realizados para diferentes métodos encontrados nos trabalhos [3, 2], e também para a aproximação convencional utilizando o *upwind* de 1^a ordem. Primeiros estudos foram os de refinamento da malha computacional. Neles foi possível notar a difusão numérica para o *upwind* e a sua redução via refinamento de malha e quando do uso dos limitadores de fluxo.

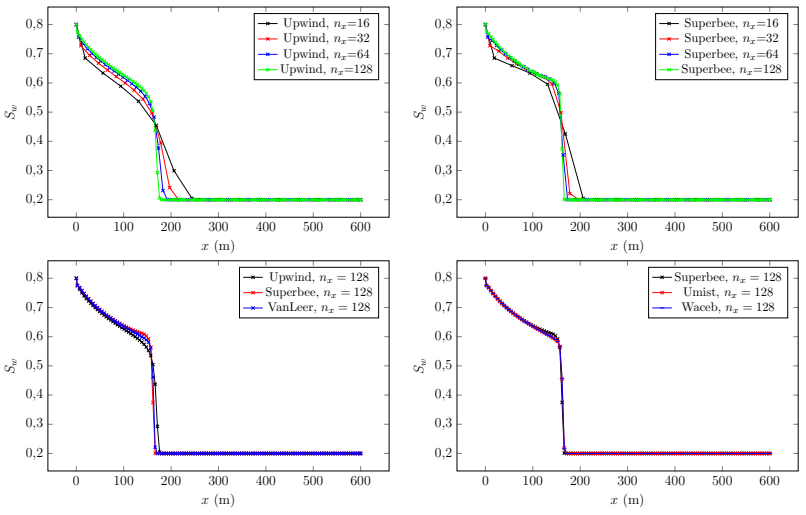


Fig. 3: Resultados para a frente de avanço da água.

Conclusões

Conseguiu-se a redução da difusão numérica e os limitadores WACEB e UMIST (ao nosso conhecimento, não usados no tema do trabalho) apresentaram resultados satisfatórios. Almeja-se estudar casos no plano xy.

Referências

[1] A. Arabzai and S. Honma. 2013. Numerical simulation of the Buckley-Leverett problem. *Proceedings of the School of Engineering*. 38, 9-14

[2] C. Redondo and G. Rubio and E. Valero. 2018. On the efficiency of the IMPES method for two phase flow problems in porous media. *Journal of Petroleum Science and Engineering*. 164, 427-436

[3] D. Zhang and C. Jiang and D. Liang and L. Cheng. 2015. A review on TVD schemes and a refined flux-limiter for steady-state calculations. *Journal of Computational Physics*. 302, 114-154.

[4] G. van Rossum. 2018. Python Tutorial. *Python Software Foundation*.

Revisão bibliográfica e adaptação de simulador numérico no contexto da injeção de CO_2

Gustavo Gomes de Moura[1], Grazione de Souza[1] Helio Pedro Amaral Souto[1]

[1] Instituto Politécnico, Universidade do Estado do Rio de Janeiro

gustavogomesm13@gmail.com, gsouza@iprj.uerj.br, helio@iprj.uerj.br

Introdução

Uma das estratégias para mitigar o problema da liberação de gás carbônico é a sua captura e o sequestro em formações subterrâneas, que podem ter outros objetivos além do armazenamento [1, 2, 3]. Neste sentido, a simulação numérica de escoamentos em meios porosos é de extrema relevância, já que as equações diferenciais parciais que modelam o escoamento de gás em meio poroso são não-lineares. Neste contexto, os objetivos deste trabalho são: 1) a revisão bibliográfica de escoamentos associados a processos de injeção de CO_2 em reservatórios em subsuperfície (e.g., aquíferos e reservatórios de óleo) e 2) o início da construção de uma simulador numérico, baseado no Método de Diferenças Finitas, para ser aplicado em estudos de injeção de CO_2.

Modelagem

Em geral, a baixas velocidades, a equação usada para conservação da quantidade de movimento no escoamento em meios porosos é a lei de Darcy,

$$\mathbf{v} = -\frac{\mathbf{k}}{\mu}\left(\nabla p - \rho g \nabla D\right), \tag{1}$$

onde $\mathbf{v}$ é a velocidade superficial do fluido, g é a magnitude da aceleração da gravidade e D é a profundidade. Já a equação de conservação da massa pode ser expressa por

$$\frac{\partial}{\partial t}\left(\frac{\rho_{sc}\phi}{B}\right) + \nabla\cdot\left(\frac{\rho_{sc}\mathbf{v}}{B}\right) - \frac{q_{sc}\rho_{sc}}{V_b} = 0. \tag{2}$$

onde ρ_{sc} é a massa específica em condições de padrão, ϕ é a porosidade, V_b é o volume total, B é o fator-volume-formação e q_{sc} um termo fonte. Substituindo-se a Eq. (1) na Eq. (2), e como ρ_{sc} é constante,

$$\nabla\cdot\left[\frac{\mathbf{k}}{\mu B}\left(\nabla p - \rho g\nabla D\right)\right] + \frac{q_{sc}}{V_b} = \frac{\partial}{\partial t}\left(\frac{\phi}{B}\right). \tag{3}$$

O termo $\partial(\phi/B)/\partial t$ pode ser reescrito usando-se as propriedades do fluido e da rocha, levando a

$$\Gamma_p = \frac{1}{B}\frac{d\phi}{dp} + \phi\frac{d}{dp}\left(\frac{1}{B}\right), \tag{4}$$

De forma que, para efeitos gravitacionais negligenciáveis

$$\nabla\cdot\left(\frac{\mathbf{k}}{\mu B}\nabla p\right) + \frac{q_{sc}}{V_b} = \Gamma_p\frac{\partial p}{\partial t}. \tag{5}$$

Para o escoamento 2D no plano xy, e um tensor diagonal de permeabilidades, tem-se a partir da Eq. (5),

$$\frac{\partial}{\partial x}\left(\frac{k_{ax}}{\mu B}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{k_{ay}}{\mu B}\frac{\partial p}{\partial y}\right) + \frac{q_{sc}}{V_b} = \Gamma_p\frac{\partial p}{\partial t}. \tag{6}$$

Como condição inicial utiliza-se $p(x,y,t=0) = p_{ini}(x,y) = p_{inic}$, onde a pressão inicial antes do reservatório ser perturbado pela produção/injeção é representada por p_{inic}. As condições de contorno externas são as de fluxo nulo nas fronteiras.

Referências

[1] Y. Ghomian and G. A. Pope and K. Sepehrnoori. 2008. Reservoir simulation of CO_2 sequestration pilot in Frio brine formation, *Energy*. 33, 1055-1067

[2] R. Gupta, R. and S. C. Peter 2020. CO_2 capture and sequestration - A solution for enhanced recoveries of unconventional gasses and liquids, *Energy and Climate Change*

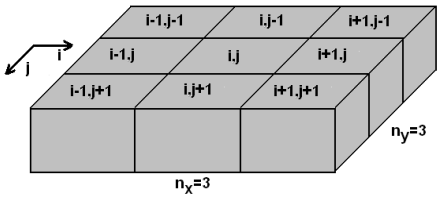
[3] A. Hamza and I. A. Hussein and M. J. Al-Marri and M. Mahmoud and R. Shawabkeh and S. Aparicio. 2021. CO_2 enhanced gas recovery and sequestration in depleted gas reservoirs: A review, *Journal of Petroleum Science and Engineering*, 196

Metodologia numérica

Utilizando a técnica CVFD (*Control Volume-Finite Difference*), discretiza-se a equação da pressão, na célula numerada por i,j (Fig. 1) e no nível temporal $n+1$, no qual as pressões são as incógnitas primitivas, como

$$\left\{\frac{\partial}{\partial x}\left(\mathbb{T}'_x\frac{\partial p}{\partial x}\right)dx + \frac{\partial}{\partial y}\left(\mathbb{T}'_y\frac{\partial p}{\partial y}\right)dy\right\}_{i,j}^{n+1} = \left\{\left(\Gamma_p\right)\frac{\partial p}{\partial t} + q_{sc}\right\}_{i,j}^{n+1}, \tag{7}$$

onde $(V_b)_{i,j} = (\Delta x\Delta y)_{i,j}L_z$ e $\mathbb{T}'_x \equiv A_x k_x/\mu B$ e $\mathbb{T}'_y \equiv A_y k_y/\mu B$, sendo $(A_x)_{i,j} = \Delta y_{i,j}L_z$ e $(A_y)_{i,j} = \Delta x_{i,j}L_z$, onde $\Delta x_{i,j}$ e $\Delta y_{i,j}$, são respectivamente, os espaçamentos da malha nas direções x e y na célula i,j e L_z o comprimento da formação rochosa na direção z.



Introduz-se a transmissibilidade na direção x,

$$\mathbb{T}^{n+1}_{x,i\pm\frac{1}{2},j} = \left(\frac{A_x k_{a,x}}{\mu B\Delta x}\right)^{n+1}_{i\pm\frac{1}{2},j}, \tag{8}$$

onde é utilizada, para os termos de área e de permeabilidade, uma média harmônica para determinar-se os valores na posição $i\pm 1/2, j$, a partir dos valores conhecidos em i,j e em $i\pm 1, j$, enquanto que para as propriedades de fluido uma média aritmética é aplicada. Uma expressão análoga pode ser obtida para a transmissibilidade na direção y. Tem-se para o termo Γ_p,

$$\Gamma_p^{n+1} = \frac{V_i}{\Delta t}\left(\frac{1}{B^n}\frac{\partial\phi}{\partial p} + \phi^{n+1} + \frac{\partial}{\partial p}\left(\frac{1}{B}\right)\right). \tag{9}$$

Utilizando-se uma formulação totalmente implícita no tempo,

$$\begin{aligned}&\mathbb{T}_x\Big|^{n+1}_{i+1/2,j}\left(p^{n+1}_{i+1,j} - p^{n+1}_{i,j}\right) - \mathbb{T}_x\Big|^{n+1}_{i-1/2,j}\left(p^{n+1}_{i,j} - p^{n+1}_{i-1,j}\right) + \mathbb{T}_y\Big|^{n+1}_{i,j+1/2}\left(p^{n+1}_{i,j+1} - p^{n+1}_{i,j}\right)\\ &-\mathbb{T}_y\Big|^{n+1}_{i,j-1/2}\left(p^{n+1}_{i,j} - p^{n+1}_{i,j-1}\right) = \left(\Gamma_p + \Gamma_s\right)^{n+1}_{i,j}\left(p^{n+1}_{i,j} - p^{n}_{i,j}\right) + \left(q_{sc}\right)^{n+1}_{i,j}\end{aligned} \tag{10}$$

As equações para as células formam um sistema de equações algébricas que quando resolvido conduz aos valores de p. Como as equações algébricas são não-lineares, adota-se uma estratégia iterativa para calcular os coeficientes nas equações. Na solução para p foi usado o método dos Gradientes Conjugados.

Resultados

Apresentam-se resultados para a injeção de gás (Fig. 2, com eixos diferentes para p para cada tempo), para a etapa atual de ambientação com um código legado, que trata do escoamento de gás natural.

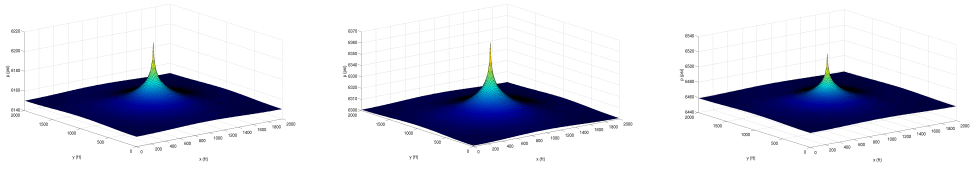


Fig. 2: Da esquerda para a direita, resultados para 120, 240 e 360 dias, respectivamente.

Conclusões

Neste início de trabalho avançou-se na modelagem do escoamento, em métodos numéricos, na ambientação com o código legado e na revisão bibliográfica sobre o escoamento de CO_2 em meios porosos.

Avaliação da aplicação BEAST em Ambientes multiCPU/GPU do SDumont - EAMC 2022.

Guillherme Freire[1,2], Micaella Coelho[1], Carla Osthoff[1], Kary Ocaña[1]
[1] Laboratório Nacional de Computação Científica (LNCC) RJ/ Brasil
[2] Faculdade de Educação Tecnológica do Estado do Rio de Janeiro (FAETERJ) RJ/ Brasil
gfreire, micaella, osthoff, karyann (@lncc.br)

Introdução

As plataformas de Computação de Alto Desempenho (CAD) como o supercomputador Santos Dumont SDumont permitem executar de uma maneira eficiente tarefas com grande demanda de processamento. O SDumont possui uma arquitetura híbrida com CPU multi core e dispositivos com arquitetura many core, GPU e MIC, é usado de forma intensiva por vários grupos de pesquisa brasileiros.

O projeto trabalha em colaboração entre os centros de pesquisa CENAPAD e LABINFO pertencentes ao LNCC, são realizadas diversas pesquisas no apoio às análises computacionais envolvendo bioinformática, biologia computacional e CAD.

Objetivos

• O presente trabalho visa levantar um estudo de desempenho de aplicações de filogenômica e evolução molecular computacional em multi GPU no ambiente de computação de alto desempenho (CAD), mais especificamente nos recursos computacionais do SDumont.

• Estabelecer uma análise comparativa do Beast acoplado a BEAGLE 3, nas suas versões CPU, GPU e multi- CPU/GPU.

• A utilização de um ou mais GPU pode se mostrar mais eficiente do que o uso de CPU multi core para analisar grandes conjuntos de dados.

Metodologia

• Esse projeto está realizando a integração de ferramentas de bioinformática em clusters de supercomputadores nos sistemas de CPU e GPU.

• O BEAST 1.8 e BEAGLE 3 foram acoplados no SDumont e executados em diferentes cenários nas filas CPU, GPU e multi-CPU/GPU levando em consideração as características e natureza dos dados e parametrização das aplicações.

Referências

[1] G Baele, D Ayres, A Rambaut, M Suchard, and P Lemey. High performance computing in Bayesian Phylogenetics and Phylodynamics using BEAGLE, volume 1910 of Methods in Molecular Biology pages 691 722 Springer, 2° edition July 2019

[2] K Ocaña, M Coelho, G Freire, C Osthof. High-Performance Computing of BEAST/BEAGLE in Bayesian Phyloienetics using SDumont Hybrid Resources. In: (BreSci2020)

[3] ERAD RJ - Exploração de Módulos Paralelo Híbrido de Bioinformática para Ambientes GPU de Supercomputação - 1 de dezembro de 2020 Autores: G Dornelas, M Coelho, K Ocaña, C Osthof

[4] Ocaña K, Coelho M, Terra R, Freire G, Santos M, Cruz L, Galheigo M, Carneiro A, Fagundes B, Carvalho D, Cardoso D, Meneses E, Gadelha L, Osthoff C, DEVELOPING EFFICIENT SCIENTIFIC GATEWAYS FOR BIOINFORMATICS INSUPERCOMPUTER ENVIRONMENTS SUPPORTED BY ARTIFICIAL INTELLIGENCE. In: (ISC High Performance 2021)

Aplicações

• O BEAST é um programa de análise filogenômica baseada em inferência Bayesiana, multi-plataforma de sequências moleculares utilizando os métodos de Markov Chain Monte Carlo (MCMC).

• A BEAGLE é uma biblioteca de alto desempenho, que faz uso de processadores altamente paralelos, como aqueles em placas gráficas (GPUs). Está acoplada ao programa (BEAST) para tornar mais eficiente a paralelização em escala fina de cálculos de probabilidade filogenética.

• O VTUNE é uma ferramenta (perfilador) de análise de desempenho, que é utilizado em sistemas operacionais Linux e Windows, para demonstrar o uso da CPU em determinada operação.

• O NSIGHT NVIDIA permite criar e depurar kernels de GPU, o perfilador também têm a função de inspecionar o estado da GPU e da memória.

Resultados

• Dados moleculares do vírus da Dengue no formato XML foram extraídos do diretório benchmark do BEAST 1.8 e usados nas análises.
• Para os cálculos foi usado como variabilidade o parâmetro chainLength fixado em “100000” e “10000000”. • As execuções foram de CPU 24 threads, CPU 40 threads, 1 GPU, CPU 40 threads /1 GPU, CPU 40 threads /8 GPU, e 8 GPU.
• O melhor desempenho foi obtido usando 8 GPU.
• Os experimentos sugerem que as características como tamanho dos dados e configuração de parâmetros no BEAST, como o chainLength, influenciam no tempo computacional.

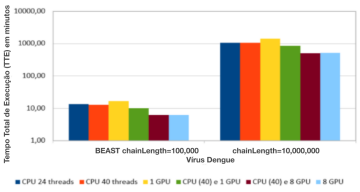


Figura 1. Análise de desempenho do BEAST/BEAGLE no CPU/GPU do SDumont

Conclusão e Trabalhos Futuros

O presente estudo viabiliza a exploração e análise de desempenho do BEAST/BEAGLE em ambientes de CAD com a especificação do ambiente computacional que leve a um desempenho mais eficiente. Permitindo assim que usuários possam usufruir dessas informações e realizar execuções garantindo um uso racional do ambiente do SDumont.
Para analisar o tipo de recursos que estão sendo usados e alocados serão realizadas análises dos recursos CPU, GPU e CPU multi-GPU, por meio de perfiladores como o VTUNE para CPU, NSIGHT NVIDIA para GPU. Dessa maneira, teremos informações sobre e.g., que função do BEAST consome maior tempo de execução e como os recursos estão sendo alocados em CPU/GPU.

• Realizar análises de desempenho e escalabilidade em larga escala nos ambientes CPU e GPU do SDumont.

• Melhorar o desempenho, escalabilidade e usabilidade dos programas com a biblioteca de computação de alto desempenho para filogenética estatística (BEAGLE).

• Como trabalhos futuros visamos uma análise de desempenho em vários nós em paralelo, que levem a uma melhora na execução.

• O BEAST/BEAGLE está integrado ao Portal Bioinfo (https://bioinfo.lncc.br/), como uma aplicação de bioinformática, apresentando o desenvolvimento de portais científicos verdes e eficientes.

Correção topográfica para validação dos dados de reanálise do ERA5-L/ECMWF- XIII EAMC.

Kécia Maria Roberto da Silva[1], Helber Barros Gomes[2], Henrique de Melo Jorge Barbosa[3]

[1,2] Universidade Federal de Alagoas, Maceió/AL, Brasil

[3] Physics Department, University of Maryland Baltimore County, *USA*

kecia.roobert7@gmail.com, helber.gomes@icat.ufal.br, hmjbarbosa@gmail.com

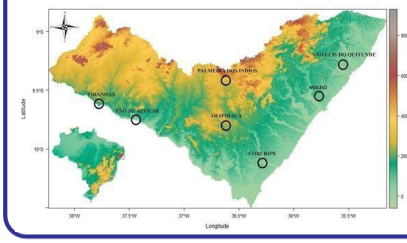
XIII Encontro Acadêmico de Modelagem Computacional

Introdução

O presente trabalho teve como objetivo validar os dados de temperatura e pressão do ERA5-Land no estado de Alagoas (8°S a 11°S de latitude e 39°W a 34°W de longitude), através de comparação com os dados observados pelas estações meteorológicas do Instituto Nacional de Meteorologia (INMET) de 2009 a 2019. Como o modelo topográfico do ERA5-Land difere da altitude real dos dados observados, foi necessário aplicar correções para a temperatura através do lapse-rate, ou taxa de lapso ambiental, e para a pressão através da equação hidrostática. Os resultados mostram como a correção da diferença de altitude entre um modelo e as observações são importantes para a validação dos resultados do modelo.

Área de Estudo

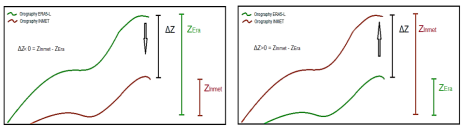
A região de Alagoas, localizada no NEB, ocupa uma faixa costeira de 220 km de extensão, com uma área territorial de 27.779,343 km² e localiza-se entre os paralelos 8,81° S e 10,50° S e os meridianos 35,15 W e 38,23 W (IBGE, 2010).



Dados

Foram utilizados dados meteorológicos das variáveis pressão atmosférica, temperatura do ar de 7 estações meteorológicas automáticas de superfície (EMS) do Instituto Nacional de Meteorologia (INMET), no período de 2009 a 2011.

O ERA5L é um produto de assimilação derivado da componente terrestre do ERA5, que é a), é a última geração de reanálise atmosférica produzida pelo produzida European Centre for Medium-Range Weather Forecasts (ECMWF), porém com o grande diferencial de ter resolução espacial de 9 km (0,1° de grade regular latitude- longitude), incluindo uma correção de altitude para o estado termodinâmico próximo à superfície, o que o torna a primeira reanálise global a alcançar uma resolução tão refinada (Sabater et al., 2021).



Referências

[1] COSGROVE, B. A., "Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) Project". Journal of Geophysical Research: Atmospheres, v. 108, n. D22 2003. DOI: https://doi.org/10.1029/2002JD003118

[2] DEE, D. P., UPPALA, S. M., SIMMONS, A. J., BERRISFORD, P., POLI, P., KOBAYASHI, S., VITART, F., "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system". Quarterly Journal of the royal meteorological society. 2011. DOI: D137(656), 553-597.

Métodos

Sabendo que o geopotencial de superfície Φ é o trabalho que precisa ser realizado contra o campo gravitacional da terrapara elevar uma unidade de massa a uma altitude acima do nível médio do mar (Holton, 2004), dado numa altura Z é definido, portanto, como:

$$\Phi(z) = \int_0^z g\,dz \qquad (1)$$

Sendo assim, a altura geopotencial z pode ser obtida por:

$$Z \equiv \frac{\Phi(z)}{g_0} = \frac{1}{g_0}\int_0^z g\,dz \qquad (2)$$

Dessa forma, a orografia do ERA5-L foi calculada dividindo o geopotencial de superfície (m²/s²) pela aceleração de gravidade (m/s²), para obter a altura geopotencial (m) acima do nível médio do mar.

Lapse Rate (TLR) = - 6,5 K/km :

$$\gamma = -\frac{dT}{dZ}$$

De acordo com Cosgrove *et al.* (2003), a temperatura e pressão pode ser corrigida através da eq.s a seguir:

$$T.Corrigida(K) = T.ERA5L + [(\Delta z) \times (\gamma)] \qquad (3)$$

$$T.Corrigida(K) = T.ERA5L + [(\Delta z) \times (-0{,}0065)] = T.ERA5L + \Delta t \qquad (4)$$

Para o ajuste de pressão, de acordo com a qual a densidade do ar diminui com a altura, temos que:

$$P.ERA5L = \rho RT.Corrigida \qquad (5)$$

$$\rho = P/R/T.corrigida \qquad (6)$$

$$\Delta p = -\rho g \Delta z \qquad (7)$$

$$P.Corrigida\ (hPa) = P.ERA5L + \Delta p \qquad (8)$$

$$P.Corrigida = P.ERA5L + \Delta p \qquad (9)$$

Resultados

Para a Temperatura e Pressão foi deduzido que:

$$\Delta z > 0 \rightarrow \Delta t < 0 \qquad (10)$$

e

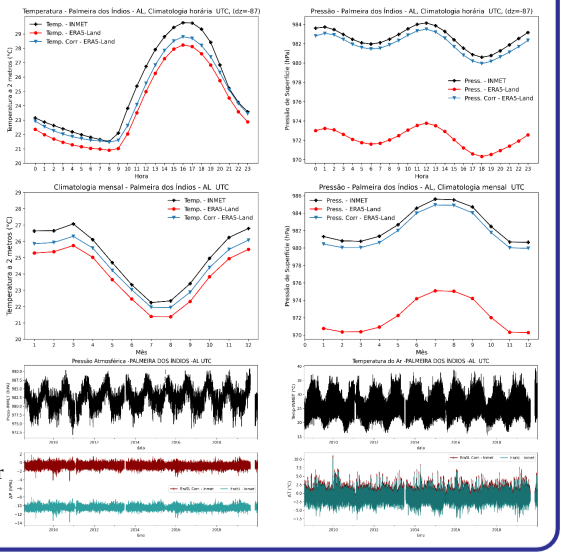
$$\Delta z < 0 \rightarrow \Delta t > 0 \qquad (11)$$

$$\Delta z > 0 \rightarrow \Delta p < 0 \qquad (12)$$

e

$$\Delta z < 0 \rightarrow \Delta p > 0 \qquad (13)$$

Em todos os casos, a temperatura se manteve bem representada pelo modelo, e quanto maior o ΔZ maiores foram as diferenças nas correções e maior a tendência do modelo subestimar os dados reais. Por outro lado, quanto menor o ΔZ , menores as diferençase maior a tendência do modelo superestimar os dados reais.



Conclusões

A reanálise do ERA5-L foi validada na região através de um ajuste topográfico para correção de dados de temperatura do ar com um BIAS de até 0,7°C e RMSE de 0,34°C, mantendo uma correlação muito forte. Enquanto os dados de pressão de superfície tiveram um erro sistemático corrigido em até 11,93 hPa. Uma representação mais realística da pressão atmosférica foi obtida, enquanto a temperatura representou bem os dados observados, mesmo sem correção.

Post-processing techniques for the MHM method: Application to the Darcy Equation

Larissa Martins[1], Weslley Pereira[2], Frédéric Valentin[3]

[1] Computational Modeling Post-Graduate Program, LNCC
[2] Department of Mathematical and Statistical Sciences, University of Colorado Denver
[3] Department of Computational and Mathematical Methods, LNCC

larissam@lncc.br, weslley.spereira@gmail.com, valentin@lncc.br

XIII Encontro Acadêmico de Modelagem Computacional

Introduction

Numerical methods must provide good quality approximations of the velocity in the Darcy model. Notably, the approximate velocity field should preserve flux continuity (e.g., conformity in $H(\mathrm{div},\Omega)$) at the interelement boundary and locally ensure mass conservation. This work aims to introduce a novel post-processing algorithm to reconstruct a velocity field originated from the two-level MHM (Multiscale Hybrid-Mixed) method proposed in [1]. It extends the work in [2] to multiscale problems and provide an alternative to adopt mixed second level solvers within the MHM methodology (c.f. [3]).

Darcy Problem

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ be an open, bounded, polytopal domain with a Lipschitz boundary $\partial\Omega$. Given $f \in L^2(\Omega)$, the standard weak formulation reads: Find $u \in H^1_0(\Omega)$ such that

$$\int_\Omega A\nabla u\cdot\nabla v dx = \int_\Omega f\, v dx \quad \text{for all } v \in H^1_0(\Omega)\,. \tag{1}$$

Here, $A \in L^\infty(\Omega)^{d\times d}$ is a symmetric matrix and may involve *multiscale* features. It is supposed to be uniformly elliptic in Ω,i.e., $\exists\, A_{\min}, A_{\max} > 0$ s.t.

$$A_{\min}|\boldsymbol{\xi}|^2 \le \boldsymbol{\xi}^T A(\boldsymbol{x})\boldsymbol{\xi} \le A_{\max}|\boldsymbol{\xi}|^2\,, \tag{2}$$

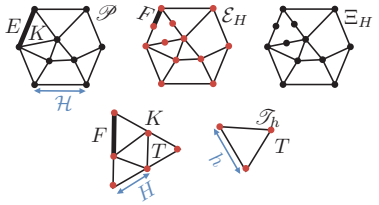
for all $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \Omega$.

Settings for the MHM method

We start by introducing $\mathscr{P}$, a collection of closed, bounded, disjoint polytopes, K, such that $\bar{\Omega} = \cup_{K\in\mathscr{P}} K$. The shapes of the polytopes K are, a priori, arbitrary, but we will suppose that they satisfy a minimal angle condition. We also introduce $\partial\mathscr{P}$ as the set of boundaries ∂K, and $\mathcal{E}$ the set of the faces of $\mathscr{P}$, that is

$$\mathcal{E} := \{E = K\cap K' \text{ or } K\cap\partial\Omega \,:\, K, K' \in \mathscr{P}, \text{ and it is not reduced to a } (d-2) \text{ variety}\}\,,$$

and $\mathcal{E}_0$ the set of internal faces.



The following spaces will be useful in the sequel.

$$V := H^1(\mathscr{P})\,,$$
$$V_0 := \{v \in L^2(\Omega) : v|_K \in \mathbb{P}_0(K),\ \forall K \in \mathscr{P}\}$$
$$\tilde{V} := \{v \in V \,:\, v|_K \in H^1(K)\cap L^2_0(K),\ K \in \mathscr{P}\}\,,$$
$$\Lambda := \{\boldsymbol{\tau}\cdot\boldsymbol{n}^K|_{\partial K} \,:\, \boldsymbol{\tau} \in H(\mathrm{div},\Omega),\ \forall K \in \mathscr{P}\}\,.$$

$\Lambda_H \subset \Lambda$ is the space of discontinuous polynomial functions on $F \subset E$ of degree $\ell \ge 0$.
$\tilde{V}_h \subset \tilde{V}$ is the space of piecewise continuous polynomial functions in each K of degree $k \ge \ell + d$.
$\mathbb{P}_0(K)$ is the space of constant functions on $K \in \mathscr{P}$.

Conclusion

(1) We introduced a novel post-processing way to reconstruct the approximate flow, such that $\boldsymbol{\sigma}_h \in H(\mathrm{div},\Omega)$. (2) The reconstructed flux is superconvergent in the $H(\mathrm{div})-$norm. (3) It remains to validate numerically the theoretical results .

And thanks to the IPES group for all the joint work.

MHM method

We introduce the approximate mappings $T_h : \Lambda \to \tilde{V}_h$ and $\hat{T}_h : L^2(\Omega) \to \tilde{V}_h$, on each $K \in \mathscr{P}$, by the following local problems:

given $\mu \in \Lambda$, $T_h\mu \in \tilde{V}_h$ is the unique solution of

$$(A\nabla T_h\mu, \nabla v_h)_K = -\langle\mu, v_h\rangle_{\partial K}, \quad \forall v_h \in \tilde{V}_h\,. \tag{3}$$

given $q \in L^2(\Omega)$, $\hat{T}_h q \in \tilde{V}_h$ is the unique solution of

$$(A\nabla \hat{T}_h q, \nabla v_h)_K = (q, v_h)_K, \quad \forall v_h \in \tilde{V}_h\,. \tag{4}$$

MHM method: Find $(\lambda_H, u_0^h) \in \Lambda_H \times V_0$ such that

$$\begin{cases} \displaystyle\sum_{K\in\mathscr{P}} \langle\mu_H, T_h\lambda_H\rangle_{\partial K} + \langle\mu_H, u_0^h\rangle_{\partial K} = -\sum_{K\in\mathscr{P}} \langle\mu_H, \hat{T}_h f\rangle_{\partial K} & \text{for all } \mu_H \in \Lambda_H\,, \\ \displaystyle\sum_{K\in\mathscr{P}} \langle\lambda_H, v_0\rangle_{\partial K} = (f, v_0)_\Omega & \text{for all } v_0 \in V_0\,. \end{cases} \tag{5}$$

MHM solution: $u_{Hh} = u_0^h + T_h\lambda_H + \hat{T}_h f$.

Flux Recovery

In the current MHM method it is possible to compute the approximate flow $\boldsymbol{\sigma}_h$ through the solution u_{Hh}, i.e., $\boldsymbol{\sigma}_h = -A\nabla u_{Hh}$, but with this approach we can not guarantee that $\boldsymbol{\sigma}_h \in H(\mathrm{div},\Omega)$. So, since $\boldsymbol{\sigma}_h \notin H(\mathrm{div},\Omega)$, we can not estimate $\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{0,\Omega}$. We lose one order of convergence in the $H(\mathrm{div},\mathscr{P})$-norm.

One element submesh on simplexes

Once the solution (u_{Hh}, λ_H) of the MHM is computed, one can construct a unique $\boldsymbol{\sigma}_H \in \mathcal{RT}_m(K)$ on each $K \in \mathscr{P}$:

$$\boldsymbol{\sigma}_H\cdot\boldsymbol{n}^K = \lambda_H\,, \quad \text{on } \partial K\,,$$
$$\int_K \boldsymbol{\sigma}_H\cdot\boldsymbol{\tau} = \int_K -A\nabla u_{Hh}\cdot\boldsymbol{\tau}\,, \quad \forall\boldsymbol{\tau} \in \mathbb{P}^d_{m-1}(K),\ (m \ge 1)\,.$$

- Error estimate:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_H\|_{\mathrm{div},\Omega} \le C\left(H^k|u|_{k+1,\mathscr{P}} + H^{\ell+1}|\boldsymbol{\sigma}|_{\ell+1,\mathscr{P}} + H^{\ell+1}|f|_{\ell+1,\Omega}\right)\,.$$

Submesh on polytopes

Once the solution (u_{Hh}, λ_H) of the MHM is computed, one can construct a unique $\boldsymbol{\sigma}_h \in \mathcal{RT}_m(T)$ on each $T \in \mathscr{T}_h^K$:

$$\boldsymbol{\sigma}_h\cdot\boldsymbol{n}_F^T = \lambda_H\,, \quad \text{if } F \subset \partial K\,, \tag{6}$$
$$\int_F (\boldsymbol{\sigma}_h\cdot\boldsymbol{n}_F^T)\mu = \int_F -\{A\nabla u_{Hh}\}\cdot\boldsymbol{n}_F^T\,\mu\,, \quad \forall\,\mu \in \mathbb{P}_m(F),\ \text{if } F \in \mathcal{F}_0\,, \tag{7}$$
$$\int_T \boldsymbol{\sigma}_h\cdot\boldsymbol{\tau} = \int_T -A\nabla u_{Hh}\cdot\boldsymbol{\tau}\,, \quad \forall\,\boldsymbol{\tau} \in \mathbb{P}^d_{m-1}(T)\,,\ (m \ge 1)\,. \tag{8}$$

where $\mathcal{F}_0$ is the set of internal faces of T, $\boldsymbol{n}_F^T$ is the unit outward normal to F and $\{\cdot\}$ represents the average value.

- Error estimate:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_H\|_{\mathrm{div},\Omega} \le C\left(h^k|u|_{k+1,\mathscr{P}} + H^{\ell+1}|A\nabla u|_{\ell+1,\mathscr{P}} + h^{\ell+1}|f|_{\ell+1,\Omega}\right)\,.$$

References

[1] Rodolfo Araya, Christopher Harder, Diego Paredes, and Frédéric Valentin. Multiscale hybrid-mixed method. *SIAM Journal on Numerical Analysis*, 51(6):3505–3531, 2013.

[2] So-Hsiang Chou, Do Y Kwak, and Kwang Y Kim. Flux recovery from primal hybrid finite element methods. *SIAM Journal on Numerical Analysis*, 40(2):403–415, 2002.

[3] O. Duran, P. R. B Devloo, S. M. Gomes, and F. Valentin. A multiscale hybrid method for darcy's problems using mixed finite element local solvers. *Comp. Meth. Appl. Mech. Eng.*, 354:213–244, 2019.

Análises por simulação computacional das mutações na proteína FXN humana na Ataxia de Friedreich.

Departamento de genética e biologia molecular da Universidade Federal do Estado do Rio de Janeiro - UNIRIO
[1]Loiane Mendonça Abrantes da Conceição (IC-UNIRIO) – loianemendoncaac@gmail.com
[1]Gabriel Rodrigues Coutinho Pereira (Doutorando PPGNeuro-UNIRIO) – gabrielkytz@hotmail.com
[1]Joelma Freire de Mesquita (orientadora) – jomesquita@gmail.com

Introdução

A Ataxia de Friedreich (FRDA) é uma doença genética autossômica recessiva neurodegenerativa, sendo a mais comum dentre as ataxias hereditárias, acometendo cerca de 1 em 50.000 pessoas no mundo, causando sintomas como a perda progressiva dos movimentos e massa muscular com início nos membros inferiores. Além de deformidades ósseas e disartria, a FRDA pode causar propensão a diabetes melitus tipo II e a cardiomiopatia hipertrófica. A Frataxina é uma proteína codificada pelo gene FXN composto por um total de 210 aminoácidos, sendo responsável pela regulação da homeostase de ferro intracelular. Mutações *missense* no gene FXN levam a perda de função proteica, gerando radicais livres.

Objetivos

Esse estudo teve como objetivo usar métodos computacionais seguindo a metodologia previamente descrita por nosso grupo, para analisar o efeito estrutural e funcional das mutações na proteína frataxina humana associadas à Ataxia de Friedreich. Conhecer as alterações resultantes das variantes genéticas da frataxina humana poderia fornecer informações relevantes sobre as bases estruturais da FRDA e ainda servir de ponto de partida para o desenho de novos estudos de fármacos e intervenções terapêuticas para essa doença.

Materiais e Métodos

A compilação das variantes genéticas do gene FXN foi feita, utilizando os bancos de dados: UNIPROT e o OMIM, e uma revisão da literatura através do PubMed.

As mutações foram submetidas à predição funcional usando os algoritmos: PolyPhen2, Provean, SNP&GO, Panther, SIFT, SNAP, PHD-SNP, MAPP, PREDICT-SNP, MutPred2, SNPEffect4.0 e I-mutant3.0.

A obtenção do fragmento experimental, foi selecionado a partir do alinhamento ProteinBlast visando a busca de sequências de proteínas relacionadas à da frataxina e de estrutura conhecida.

O modelo teórico completo da FXN nativa foi feito utilizando modelagem por threading, comparativa e *ab initio* nos servidores SwissModel, Robetta, I-TASSER, MholLine, e Raptor-X.

A validação do modelo gerado foi realizada utilizando os servidores ProSa-Web, QMEAN, PROCHECK, Verify3D e ERRAT.

O modelo validado foi submetido à análise de conservação evolutiva no servidor ConSurf.

Referências

[1]COSSÉE, M. et al. Friedreich's ataxia: Point mutations and clinical presentation of compound heterozygotes. **Annals of Neurology**, v. 45, n. 2, p. 200–206, 1999.
[2]SCHOENFELD, R. A. et al. Frataxin deficiency alters heme pathway transcripts and decreases mitochondrial heme metabolites in mammalian cells. **Human Molecular Genetics**, v. 14, n. 24, p. 3787–3799, 2005.
[3]CLARK, E. et al. Identification of a novel missense mutation in Friedreich's ataxia –FXNW168R. **Annals of Clinical and Translational Neurology**, v. 6, n. 4, p. 812–816, 2019.
[4]PEREIRA, G. R. C.; TELLINI, G. H. A. S.; DE MESQUITA, J. F. In silico analysis of PFN1 related to amyotrophic lateral sclerosis. **PLoS ONE**, v. 14, n. 6, p. 5–10, 2019.
[5]DE CARVALHO, M. D. C.; DE MESQUITA, J. F. Structural Modeling and In Silico Analysis of Human Superoxide Dismutase 2. **PLoS ONE**, v. 8, n. 6, 2013.
[6]MOREIRA, L. G. A. et al. Structural and functional analysis of human SOD1 in amyotrophic lateral sclerosis. **PLoS ONE**, v. 8, n. 12, p. 1–9, 2013.

Resultados

Após a revisão bibliográfica, as mutações analisadas eram sabidamente deletérias. Por tanto maior parte dos algoritmos de predição funcional obteve uma alta taxa de predição deletéria, e consequentemente acerto em suas predições, como mostrado na Figura 1, somente o algoritmo SNPs&GO apresentou uma baixa taxa de predição deletéria para essas mutações sabidamente associadas ao desenvolvimento de FRDA. Mostrando a importância do uso combinado de algoritmos para a predição funcional. Além disso, todas as mutações analisadas foram preditas como deletérias por pelo menos oito algoritmos de predição funcional utilizados como visto na Figura 2.
A FXN humana teve apenas parte da sua estrutura determinada experimentalmente por cristalografia de raios-X de forma que somente a porção final da proteína, contendo os aminoácidos 89-210, possui uma estrutura tridimensional determinada. Assim, desenvolveu-se um modelo teórico completo da proteína utilizando modelagem *in sílico*. Dos fragmentos experimentais disponíveis no PDB para a FXN humana, o fragmento 3S4M foi selecionado, pois apresentou maior resolução (1,30Å), cobertura (61%), e identidade (100%) de sequência.

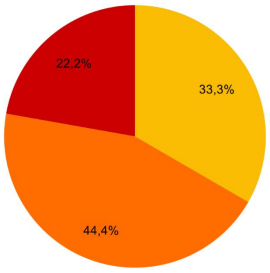


Figura 2. Análise de predição funcional das mutações da Frataxina humana. Taxa de predições deletérias para as mutações estudadas.

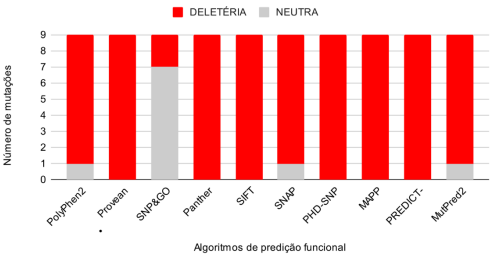


Figura 1. Análise de predição funcional das mutações da FXN humana. Predição funcional das mutações, de acordo com cada algoritmo de usado, sendo em vermelho o número de mutações preditas deletérias, e em cinza, o número de mutações preditas neutras.

O modelo gerado pelo servidor Robetta apresentou a melhor qualidade estrutural dentre os modelos analisados, obtendo um RMSD=1.53 e TM-Score=0.90. Além disso, o modelo foi considerado validado por algoritmos como ERRAT, PROCHECK, VERIFY-3D, ProSA-Web, QMEAN.
O modelo final validado, bem como seu alinhamento estrutural com o fragmento experimental da 3S4M, são mostrados na Figura 4. A inspeção visual do alinhamento reafirma a similaridade estrutural entre o modelo validado e o fragmento cristalográfico da FXN humana.

A análise de conservação evolutiva no ConSurf indicou que todas as mutações da FXN humana ocorrem em sítios conservados da molécula, como observado na Figura 3.

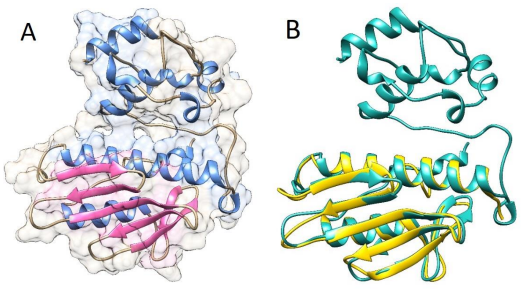


Figura 4. **(A)** Modelo final validado da FXN humana. Regiões de folhas-beta estão representados por setas rosas, ao passo que as regiões de alfa-hélices estão representadas por hélices azuis. A superfície da proteína também é mostrada na figura. **(B)** Alinhamento estrutural no TM-Align do modelo do Robetta e o fragmento experimental 3S4M. A estrutura tridimensional do modelo validado está representada em azul, ao passo que a estrutura do fragmento cristalográfico está representada em amarelo.

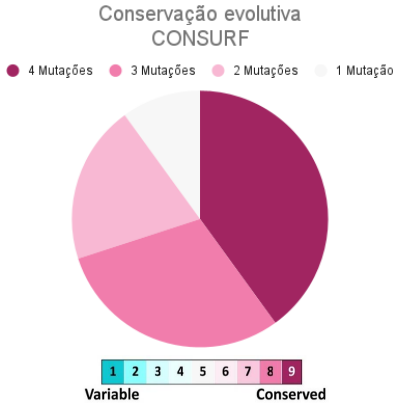


Figura 3. Análise de conservação evolutiva da proteína FXN humana no ConSurf. Representação do grau de conservação evolutiva dos resíduos afetados por mutações da FXN conforme escala colorimétrica do ConSurf.

Conclusão

O presente estudo gerou um modelo acurado, completo e inédito da proteína FXN humana. Onde a maioria das mutações foi predita como deletérias. A análise de conservação evolutiva indicou que as mutações conhecidas da frataxina humana ocorreram em posições conservadas, e possivelmente, importantes para a proteína. O modelo tridimensional e as predições realizadas nesse estudo poderiam possivelmente fornecer informações relevantes sobre as bases estruturais da FRDA e ainda contribuir de base para o desenho de novos fármacos e servir de ponto de partida para outros estudos.

Projeto e Implementação de *Workflows* Científicos Reprodutíveis de Alto Desempenho: ParslRNA-Seq

Lucas Cruz[1,2], Micaella Coelho[1], Carla Osthoff[1], Luiz Gadelha[1], Kary Ocaña[1]

[1]Laboratório Nacional de Computação Científica (LNCC)

[2]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)

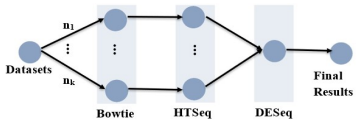
{lucruz,micaella,osthoff,lgadelha,karyann}@lncc.br

Introdução

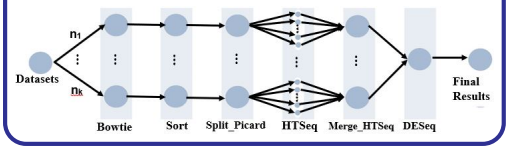
A técnica de sequenciamento RNA (RNA-Seq) é utilizada na área de bioinformática e trata da análise de Expressão Diferencial de Genes e sequenciamento genômico, ou seja, permite estudar o comportamento de um conjunto de transcritos de uma célula em uma dada condição fisiológica ou de desenvolvimento, tal como o câncer. Em geral, esses experimentos são compostos por **atividades extensivas, complexas e computacionalmente custosas**. O desenvolvimento deste trabalho parte de um projeto multidisciplinar em colaboração entre LABINFO e CENAPAD no LNCC com o propósito de acoplar ao portal de bioinformática BioinfoPortal um *workflow* científico de RNA-Seq com Processamento de Alto Desempenho, para facilitar o gerenciamento dessas atividades. O presente trabalho tem, portanto, o objetivo de **apresentar a modelagem** desse *workflow*, suas **fases de otimização** e uma **análise de seu comportamento** dentro de um ambiente de Alto Desempenho.

Modelagem e Otimização

A primeira versão do ParslRNA-Seq se compõe por três atividades: Bowtie, HTSeq e DESeq [3].

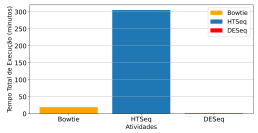


Na versão atual do ParslRNA-Seq, o *workflow* passa a ser composto por seis atividades, as quais foram adicionadas: Sort, Split_Picard e Merge_HTSeq [2].

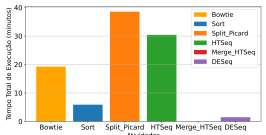


Desempenho das atividades

Considerando execuções seriais do ParslRNA-Seq com um nó computacional do SDumont, a primeira versão tem Tempo Total de Execução (TTE) de cerca de 326 minutos, sendo que **mais de 90% do TTE é gasto na execução da atividade HTSeq**.



Com a nova estratégia adotada na nova modelagem houve uma redução no TTE dessa atividade em cerca de 305 a, aproximadamente, 30 minutos, o que representa **cerca de 90% de melhora**.



Referências

[1] L. Cruz. et al. Parallel Performance and I/O Profiling of HPC RNA-Seq Applications. Computación y Sistemas, 2022. (submetido)

[2] L. Cruz. et al. Workflows científicos de rna-seq em ambientes distribuídos de alto desempenho: Otimização de desempenho e análises de dados de expressão diferencial de genes. BRESCI, 2021.

[3] L. Cruz. et al. Avaliação de Desempenho de um Workflow Científico para Experimentos de RNA-Seq no Supercomputador Santos Dumont. WSCAD-WIC, 2020.

Análise de desempenho em um único nó

Fazendo utilização de um nó computacional, juntamente com a utilização do Parsl/Python foi possível realizar uma paralelização de tarefas dentro do nó. Na primeira versão, há somente paralelização *multithreading* dos processos do *software* Bowtie e o HTSeq está utilizando somente um núcleo de CPU para executar cada amostra. Então, apesar da execução paralelizada de tarefas realizadas pelo Parsl ainda **existe um número considerável de CPUs ociosas na maior parte da execução do *workflow***.

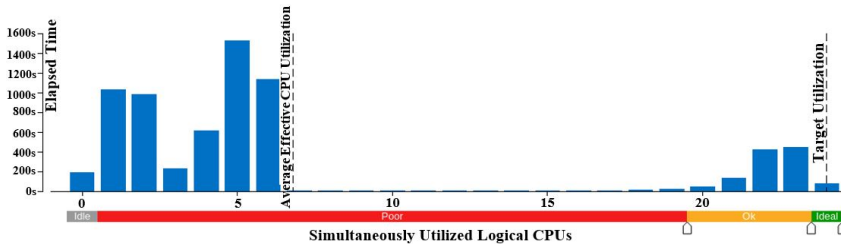


Figura: Histograma da primeira versão.

Na versão atual é possível observar uma **maior distribuição na utilização dos núcleos de CPUs disponíveis**. Observa-se um aumento considerável no número de CPUs utilizadas na maior parte do tempo de execução e que isso leva a um fator ideal de utilização dos recursos computacionais [1]. Isso se deve ao fato de ter mais uma atividade fazendo execução *multithreading*, a atividade Sort, e principalmente pela execução paralelizada em múltiplos *cores* do HTSeq.

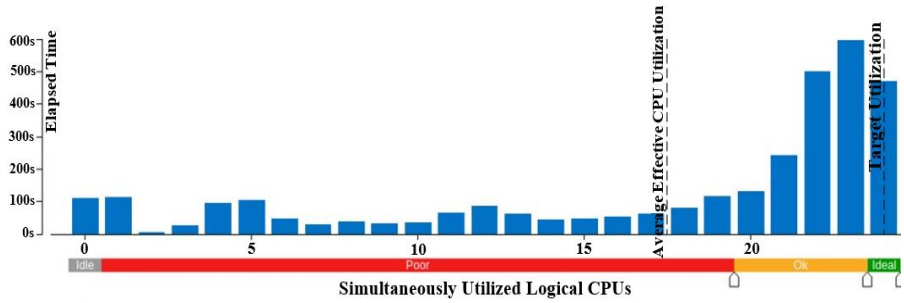
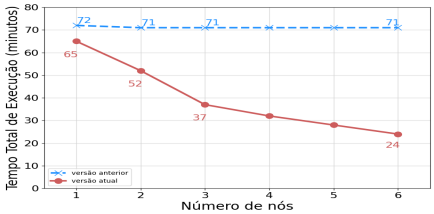


Figura: Histograma da versão atual.

Análise de desempenho em múltiplos nós



Em um cenário de execução de único nó computacional a versão atual apresenta uma redução no TTE, de 72 minutos para 65, em relação a primeira versão, que se deve à redução de núcleos de CPUs ociosas. Não é uma redução tão significativa pois há também nesse cenário de execução muita competição na utilização dos recursos computacionais. O melhor TTE dentro do cenário multinó pode ser observado pela versão atual, em 6 nós, onde o TTE é de 24 minutos.

Conclusão

Este trabalho apresentou as duas modelagens do *workflow* de RNA-Seq que foram desenvolvidas bem como uma análise comparativa sobre os desempenhos computacionais de cada uma versão do *workflow*, concluindo que **a segunda versão, de seis atividades, é a que hoje mais se adequa ao cenário paralelo e distribuído**, pois nela há uma melhor utilização dos recursos computacionais disponíveis. A nova versão apresenta uma **melhora de 66,19% no tempo computacional**, em relação a versão anterior para uma execução paralela e distribuída com um **ganho de até 90% do tempo de execução da atividade HTSeq**.

XV EAMC - Towards Provenance Support in the BioinfoPortal Gateway

Marco Cabral[1,2], Antônio Tadeu Azevedo Gomes[1], Marcelo Galheigo[1], Kary Ocaña[1]
[1] Laboratório Nacional de Computação Científica (LNCC)
[2] Universidade Federal do Rio de Janeiro (UFRJ)

{macabral, atagomes, galheigo, karyann}@lncc.br

introdução

O *gateway* Bioinfo-Portal (https://bioinfo.lncc.br/) visa a execução de aplicações de bioinformática em larga escala, no apoio às pesquisas da comunidade científica de bioinformática. Bioinfo-Portal está acoplado a recursos de computação de alto desempenho (CAD) e do supercomputador Santos Dumont a fim de diminuir o tempo de processamento de execuções. Bioinfo-Portal gerencia a execução automática de aplicações, ferramentas e coleções de dados científicos através de uma interface web amigável e iterativa e das diversas camadas de software do *gateway*. Bioinfo-Portal utiliza, via serviços Web RESTful, o *middleware* CSGrid como *framework* de integração à arquitetura do SINAPAD. Atualizações e otimizações do Bioinfo-Portal na camada de banco de dados e de gerência de execuções irão fornecer uma melhor funcionalidade e escalabilidade de processos de execuções e armazenamento de dados de proveniência, tal que auxiliem na tomada de decisões inteligentes no uso de recursos computacionais.

Objetivos

- Atualização das camadas de banco de dados e de gerência de execuções da arquitetura do Bioinfo-Portal, por meio do desenvolvimento de serviços específicos para integrar dados contidos nessas camadas.
- Análise, extração e gerência de informações de dados científicos e de proveniência extraídas das camadas da arquitetura do Bioinfo-Portal e das aplicações de bioinformática.
- Implementação e validação de um banco de dados que centralize informações do Bioinfo-Portal e do ambiente computacional.
- Desenvolvimento de sistemas para criar inteligência em análise de coleta de dados e tomada de decisão, tal que melhore a eficiência do *gateway* em termos de velocidade, execução e armazenamento.

Metodologia

- Na primeira etapa, o projeto físico utilizou o PostgreSQL v10 como Sistema de Gerência de Banco de Dados (SGBD) relacional *Open Source* e *pgAdmin* v5.2 como plataforma de desenvolvimento e gerência.
- A segunda etapa envolve a utilização de serviços *RESTful* para o desenvolvimento dos sistemas de tomada de decisão inteligentes. A linguagem de programação utilizada é a PHP (*Hypertext Preprocessor*). Visual Studio Code é o editor de código-fonte usado para o desenvolvimento dos sistemas.

Referência

[1] Ocaña, K.A.C.S., et al. (2020). BioinfoPortal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. *In Future Generation Computer Systems*, Rio de Janeiro, v. 107, p. 23, Janeiro 2020.

[2] KIM, S.-H. et al. (2017). Science Gateway Cloud With Cost-Adaptive VM Management for Computational Science and Applications. *IEEE Systems Journal*, v. 11, n. 1, p. 173-185, Março 2017. ISSN 1932-8184.

[3] LESK, A. M (2019). Bioinformatics,*Britannica*, Pennsylvania, Fevereiro 2019.

[4] Gesing S, Krüger J, Grunzke R, Herres-Pawlis S, Hoffmann A. (2016). Using Science Gateways for Bridging the Differences between Research Infrastructures,*Journal of Grid Computing*, 2016;14:545–57.

Resultados I: Banco de dados

O modelo conceitual de banco de dados do Bioinfo-Portal foi implementado, como apresentado na Figura 1. Iniciou-se a o mapeamento dos dados na arquitetura do *gateway* para a implementação do modelo lógico.

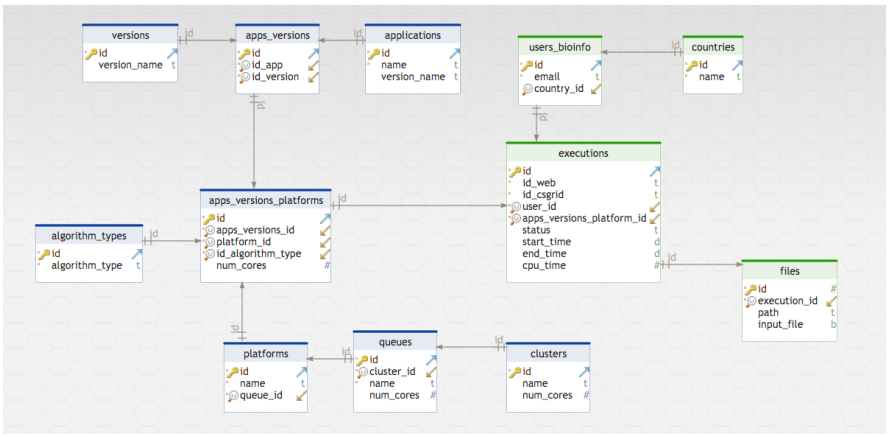


Figura 1. Esquema Conceitual Entidade-Relacionamento do Banco de Dados do Bioinfo.

Dentre as entidades do modelo conceitual ER do Bioinfo-Portal (Figura 1), *Files* e *Executions* são entidades originais, as demais entidades pertencem à nova versão do banco de dados.

Resultados II: Sistemas inteligentes

Em desenvolvimento, os sistemas utilizando serviços web *RESTful* visam interagir dinamicamente com o *middleware* CSGrid do SINAPAD. A Figura 2 apresenta o Sistema de Autenticação, por meio do método *LDAP* (Figura 2A) e *RSA* (Figura 2B). Esses sistemas extraem, tratam e armazenam dados de proveniência de usuários, como nome e identificação.

```php
<?php
    $url = ('http://tenia.sinapad.lncc.br:8080/rest/ngw/authentication/login-ldap');

    $data = array(
     'username'=> 'marco.antonio',
     'password'=> 'MYPWD',
     'service'=> 'MYSERV',
     'uuid' => $uuid

    );
    $headers = array(
     #'Accept: application/xml'
     'Accept: application/json'
    );

    $handle = curl_init();
    curl_setopt($handle, CURLOPT_URL, $url);
    curl_setopt($handle, CURLOPT_HTTPHEADER, $headers);
    curl_setopt($handle, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($handle, CURLOPT_SSL_VERIFYHOST, false);
    curl_setopt($handle, CURLOPT_SSL_VERIFYPEER, false);

    curl_setopt($handle, CURLOPT_POST, true);
    curl_setopt($handle, CURLOPT_POSTFIELDS, http_build_query($data));

    $response = curl_exec($handle);
    $obj = json_decode ( $response );
    $b = $obj->{'uuid'};
    //echo $response;
?>
```

Figura 2A: Sistema de Autenticação *LDAP*

```php
<?php
$url = "https://tenia.sinapad.lncc.br:8080/rest/ngw/authentication/login-rsa";
    $data = array (
                'service' => "CSGrid",
                'username' => "BioInfoDiscoveryService",
                'file' => new CurlFile("/Users/marcoantonio/BioInfoDiscoveryServices.key", 'multipart/form-data')
    );
    $headers = array (
                'Accept: application/json'
    );
    $handle = curl_init ();

    curl_setopt ( $handle, CURLOPT_URL, $url );
    curl_setopt ( $handle, CURLOPT_HTTPHEADER, $headers );
    curl_setopt ( $handle, CURLOPT_RETURNTRANSFER, true );
    curl_setopt ( $handle, CURLOPT_SSL_VERIFYHOST, false );
    curl_setopt ( $handle, CURLOPT_SSL_VERIFYPEER, false );
    curl_setopt ( $handle, CURLOPT_CONNECTTIMEOUT, 120 );
    curl_setopt ( $handle, CURLOPT_TIMEOUT, 600 );
    curl_setopt ( $handle, CURLOPT_POST, true );
    curl_setopt ( $handle, CURLOPT_POSTFIELDS, $data );
    curl_setopt ( $handle, CURLINFO_HEADER_OUT, true);

    $response = curl_exec ( $handle );

$obj = json_decode ( $response );
$b = $obj->{'uuid'};

    //echo $response;
?>
```

Figura 2B: Sistema de autenticação *RSA*

Conclusão

A atualização do banco de dados do Bioinfo-Portal permitirá a manipulação e armazenamento de dados científicos e de proveniência. Sobre esses dados, sistemas de tratamento e autenticação de dados estão em desenvolvimento visando tornar a coleta e a gerência de dados do Bioinfo-Portal mais eficiente.
Continuaremos na implementação e validação da nova versão do banco de dados, acoplados aos sistemas de extração de dados. Pretendemos sobre eles acoplar sistemas de tomada de decisão para tornar o Bioinfo-Portal um sistema inteligente e energeticamente eficiente.

Uso de métodos computacionais na detecção automática da doença de Alzheimer

Mário L. Vicchietti[1], Fernando M. Ramos[2], Andriana S. L. O. Campanharo[3]

[1] Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia, Instituto de Biociências
[2] Laboratório de Computação e Matemática Aplicada, Instituto Nacional de Pesquisas Espaciais
[3] Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia, Instituto de Biociências

mario.lucas@unesp.br, fernando.ramos@inpe.br, andriana.campanharo@unesp.br

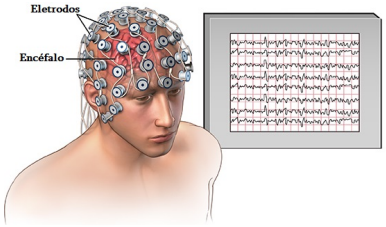
Introdução

A doença de Alzheimer (DA) é uma disfunção progressiva que acarreta na degeneração das células neuronais, levando à perda de memória, à dificuldade no aprendizado, à desorientação no espaço e no tempo, a alterações no humor e na personalidade e até à falta de capacidade motora. Uma vez que a DA não tem cura, o seu diagnóstico é a melhor alternativa para que os tratamentos que desaceleram o avanço da doença sejam iniciados. Nesse cenário, a eletroencefalografia (EEG) tem ganhado atenção, já que a mesma consiste em uma técnica não invasiva, capaz de mensurar o potencial elétrico proveniente das atividades neuronais. Nas últimas décadas, diversos grupos de cientistas têm proposto o uso de métodos computacionais de análise de séries temporais em sinais de EEG de pacientes com e sem a DA, visando identificar automaticamente a doença. O objetivo deste trabalho é classificar pacientes com e sem a DA, comparando a robustez de diferentes métodos computacionais.

Materiais e métodos

O presente trabalho foi desenvolvido de acordo com as seguintes etapas:

- Aquisição da base de dados fornecida por pesquisadores da Universidade Estadual da Flórida, a qual contém sinais de EEG referentes a 12 pacientes sadios e 80 pacientes doentes, todos com os olhos fechados (Fig. 1);



- Busca literária pelos métodos computacionais mais utilizados na detecção da DA via sinais de EEG;
- Separação das componentes dos sinais que se referem às ondas β, α, θ e δ;
- Aplicação dos métodos selecionados nos sinais de EEG originais e nos sinais de EEG filtrados;
- Cálculo da área abaixo da curva ROC (AUC) e do p-valor do teste ANOVA;
- Classificação dos indivíduos por meio de *Support Vector Machine* (SVM), usando a técnica de validação cruzada K-fold;
- Mensuração do tempo gasto nos cálculos realizados por cada um dos métodos utilizados.

Referências

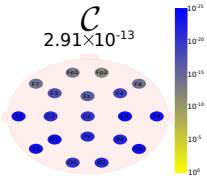
[1] Dauwels, J., Vialatte, F., and Cichocki, A. 2010. Diagnosis of *Alzheimer's disease from EEG signals: where are we standing?*. *Current Alzheimer Research*, 7(6):487-505.

[2] Kulkarni, N., and Bairagi, V. 2018. EEG-based diagnosis of alzheimer disease: a review and novel approaches for feature extraction and classification techniques.

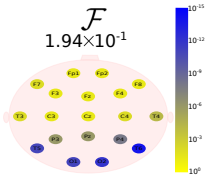
[3] Patients, C. 2019. Alzheimer's association 2019. Alzheimer's disease facts and figures. *Alzheimers Dementia*, 15(3):321-87.

[4] Pineda, A. M., Ramos, F. M., Betting, L. E., and Campanharo, A. S. 2020. Quantile graphs for EEG-based diagnosis of Alzheimer's disease. *Plos one*, 15(6):e0231169.

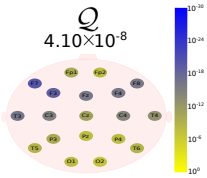
Resultados



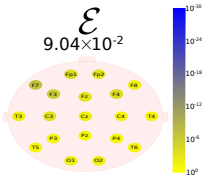
(a) Coerência Wavelet no sinal original.



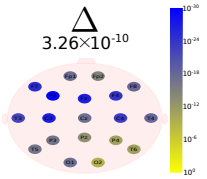
(b) Dimensão Fractal nas ondas θ.



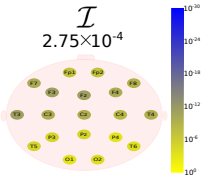
(c) Entropia Quadrática nas ondas δ.



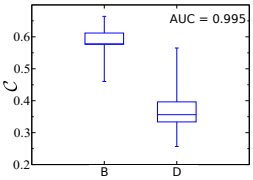
(d) Energia Wavelet nas ondas δ.



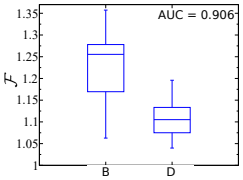
(e) Grafo de Quantis nas ondas δ.



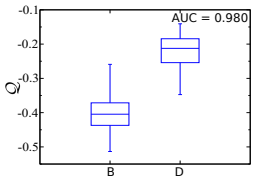
(f) Grafo de Visibilidade nas ondas δ.



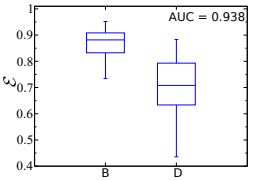
(g) Coerência Wavelet (eletrodos T3-C4).



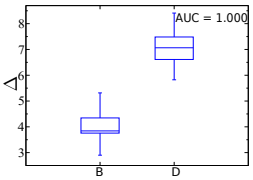
(h) Dimensão Fractal (eletrodo T6).



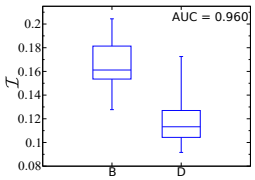
(i) Entropia Quadrática (eletrodo F7).



(j) Energia Wavelet (eletrodo F7).



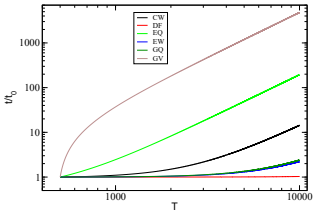
(k) Grafo de Quantis (eletrodo F3).



(l) Grafo de Visibilidade (eletrodo F3).

Medida	Sen	Esp	Acu
$\mathcal{C}$	0.978	0.987	0.917
$\mathcal{F}$	0.946	1.000	0.583
$\mathcal{Q}$	0.946	0.987	0.667
$\mathcal{E}$	0.913	0.987	0.417
Δ	1.000	1.000	1.000
$\mathcal{I}$	0.902	0.962	0.500

(m) Classificação com SVM e K-fold.



(n) Custo computacional de cada técnica.

Conclusões

Todos os métodos de detecção da DA via sinais de EEG foram eficientes, especialmente quando as ondas δ foram analisadas. Todos os lobos do encéfalo foram afetados e apresentaram diferentes padrões para diferentes técnicas. A técnica GQ foi a mais robusta, atingindo valores ótimos de AUC, especificidade, sensibilidade e acurácia. As técnicas GQ, EW e DF apresentaram-se menos custosas em relação às técnicas GV, CW e EQ. A redução da complexidade, o aumento do salto entre os quantis, a perda de coerência e o aumento da regularidade dos sinais de EEG foram as principais propriedades encontradas que discriminam os pacientes dos diferentes grupos. Em trabalhos futuros, as técnicas mais robustas serão utilizadas com o objetivo de detectar os primeiros sinais da DA e identificar a doença ainda no seu estágio incial.

Avaliação da utilização do Cálculo Fracionário, associado à Homogeneização Assintótica, na modelagem de meios micro-heterogêneos

XIII Encontro Acadêmico de Modelagem Computacional

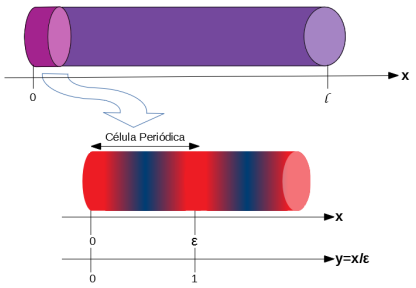
Décio Jr, R.[1], Cezaro, A.[2], Pérez-Fernández, L.[3]
[1] PPG Modelagem Computacional, FURG. [2] IMEF, FURG. [3] IFM, UFPel
roberto.decio.jr@gmail.com, adrianocezaro@furg.br, leslie.fernadez@ufpel.edu.br

Introdução

A modelagem de fenômenos que ocorrem em meios micro-heterogêneos são de interesse pois tornam possível analisar o comportamento físico de (micro)estruturas complexas. O Método de Homogeneização Assintótica (MHA) [1] tem se mostrado eficaz nesse ramo, visto que a micro-heterogeneidade implica em maior custo computacional na aplicação direta de métodos numéricos [3]. Enquanto isso, versões fracionárias de problemas clássicos têm permitido modelar comportamentos não desvendados pelos operadores de ordem inteira [2]. Nesse sentido, este trabalho visa associar derivadas fracionárias (não locais e locais) em um problema estacionário, com coeficientes continuamente diferenciáveis e rapidamente oscilantes, a fim de desvendar diferentes comportamentos no fenômeno a ser modelado, não alcançados pela aplicação do MHA no mesmo problema. Além disso, é objetivo apontar caminhos para uma abordagem híbrida dessas ferramentas na modelagem em meios micro-heterogêneos.

Meios micro-heterogêneos

Neste trabalho serão considerados meios cuja escala de distribuição das fases é muito maior que a escala atômica, e muito menor que a escala macroscópica, caracterizando assim, uma micro-heterogeneidade. Ademais, com propriedades cuja variação é gradual, caracterizando um material funcionalmente graduado.



MHA

A micro-heterogeneidade desses materiais dificulta a aplicação direta de métodos numéricos na sua modelagem, dado que a malha considerada deverá ser tão pequena quanto o parâmetro da sua , ε.
Entretanto, a separação de escalas satisfaz a hipótese da homogeneidade equivalente, tornando possível aplicar métodos de Homogeneização. Entre eles, está o MHA, no qual considera-se uma solução assintótica formal para o problema considerado, na forma de uma série de potências ε:

$$u^{\varepsilon}(x) \sim u^{(2)}(x,y) = v_0(x) + \varepsilon u_1(x,y) + \varepsilon^2 u_2(x,y).$$

De aplicá-la no problema, obtém-se uma sequência recorrente de problemas para as potências de ε cujo problema limite tomando $\varepsilon \to 0^+$ é um problema com coeficientes constantes, em relação ao material homogêneo equivalentem, cuja solução é suficientemente próxima da solução original.

$$\begin{array}{c} \frac{d}{dx}\left[K^{\epsilon}(x)\frac{du^{\varepsilon}}{dx}\right] = f(x) \\ u^{\varepsilon}|_{x=0} = g_1, u^{\varepsilon}|_{x=l} = g_2 \end{array} \xrightarrow{\varepsilon \to 0^+} \begin{array}{c} \frac{d}{dx}\left[\hat{K}\frac{dv_0}{dx}\right] = f(x) \\ v_0(0) = g_1 v_0(l) = g_2 \end{array}$$

Derivadas fracionárias

Os principais operadores fracionários são os do tipo Riemann-Liouvile e Caputo. A seguir, a definição de uma derivada do tipo Caputo para uma função $f(x) \in AC^n[a,b]$, de ordem $\mu \in \mathbb{C}$, com $Re\{\mu\} \geq 0$:

$${}_a^C D_x^{\mu}(f(x)) = \frac{1}{\Gamma(n-\mu)} \int_a^x \frac{f^{(n)}(t)dt}{(x-t)^{\mu-n+1}}, n = [Re\{\mu\}] + 1.$$

Esse operadores são não locais, e muitas das propriedades do Cálculo usual (como regra da cadeia, espaços de funções) são perdidas. Uma alternativa para seu uso é a Transformada de Laplace destes operadores [2]. No sentido de manter propriedades interessantes do cálculo usual e ser de fácil manipulação, em relação aos operadores de Riemann-Liouville e Caputo, define-se a derivada compatível de ordem α de f, denotada por $T_{\alpha}(f)$:

$$T_{\alpha}(f)(t) = \lim_{\delta \to 0} \frac{f(t + \delta t^{1-\delta}) - f(t)}{\delta}.$$

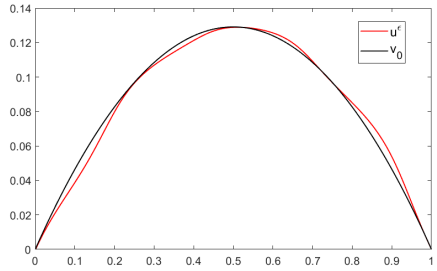
De forma literal, esses operadores não podem sem considerados fracionários pois são locais, mas têm se mostrado válido considerá-lo neste contexto, pelos resultados obtidos na sua aplicação.

Exemplo numérico

O exemplo a ser resolvido será a seguinte equação:

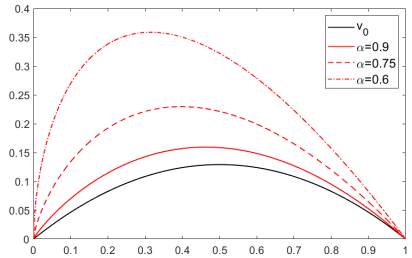
$$\frac{d}{dx}\left[K^{\varepsilon}(x)\frac{du^{\varepsilon}}{dx}\right] = -1, \; x \in (0,1)$$

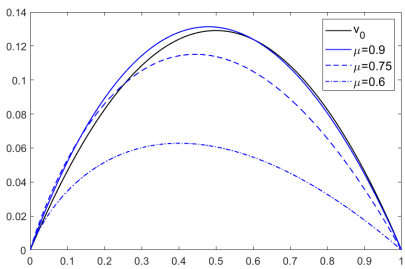
com condições de contorno nulas e $K^{\varepsilon}(x) = 1 + \frac{1}{4}\sin\left(2\pi\frac{x}{\varepsilon}\right)$. Seguem os resultados obtidos:



A comparação entre a solução original $u^{\varepsilon}(x)$ (com $\varepsilon = 1/4$) e a solução $v_0(x)$ obtida com o MHA.

Aqui, o efeito de considerar os dois tipos de operadores fracionários na solução $v_0(x)$. Na esquerda, o que ocorre ao considerar a derivada compatível, e na direita, a derivada do tipo Caputo:





Conclusões preliminares

1. Mostrou-se algumas possibilidades de associação do MHA e do CF para resolver problemas associados a meios micro-heterogêneos.
2. Percebeu-se que cada método reproduziu detalhes diferentes no exemplo: enquanto o MHA se ocupou de aproximar a solução original de forma macroscópica, os operadores fracionários alteraram a simetria da solução obtida.
3. Esta associação de métodos se mostrou bastante produtiva, e merece melhor atenção para se obter resultados mais profundos e unificados.

Referências

[1] Bakhvalov, N., Panasenko, G. 1989. *Homogenisation: Averaging Processes in Periodic Media* Dordrecht: Kluwer Academic Publishers.

[2] Sousa, J., Vaz Jr, J., Oliveira, E. 2020. Cálculo de ordem não inteira para iniciantes. In: *Notas em Matemática Aplicada - SBMAC, v.90.*

[3] Panasenko, G. 2008. Homogenization for Periodic Media: from Microscale to Macroscale. In: *Physics of Atomic Nuclei, v.71, n.4*, pp. 681–694.

Estudo da Serine Arginine Protein Kinase de *Leishmania infantum* (LiSRPK) como alvo de ligação de análogos do SRPIN340

Sara Andrade Machado[1], Débora Cristina Pimentel[1], Giovanna Ladeira Marques[2], Marcel Arruda Diogo[3], Christiane Mariotini Vasconcellos[2], Raphael de Souza Vasconcellos[1]

[1] Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Viçosa
[2] Departamento de Biomedicina, Centro Universitário Faminas
[3] Departamento de Produtos Farmacêuticos, Universidade Federal de Minas Gerais

sara.andrade@ufv.br, debora.cristina@ufv.br, giobiomed@outlook.com, marcel.arruda@farmacia.ufjf.br, chrisbiomed@gmail.com, raphael.vasconcellos@ufv.br

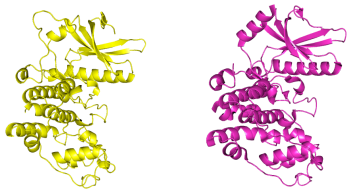
XIII Encontro Acadêmico de Modelagem Computacional

Introdução

De acordo com a Organização Mundial de Saúde (OMS), a Leishmaniose é uma doença tropical negligenciada de grande importância entre as enfermidades acometidas por protozoários, sendo endêmica em algumas regiões do mundo inclusive, no Brasil. Em território nacional, essa doença pode se manifestar por três formas clínicas, sendo a Leishmaniose Visceral (LV), causada por Leishmania infantum, a forma sistêmica no hospedeiro humano, podendo atingir órgãos internos, como baço, fígado e medula óssea, gerando grandes impactos na saúde pública. Hoje, o tratamento de pacientes com a doença é dificultado devido a presença de problemáticas relacionadas aos seus métodos, como: efeitos adversos, forma de administração incômoda e também alta resistência do parasito. Com base nisso, faz-se necessário a busca por novos alvos farmacológicos para o tratamento dessa patologia.

Pesquisa

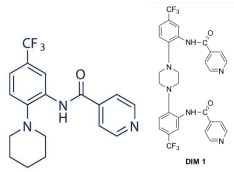
Devido ao papel significativo na regulação de processos transcricionais, de progressão e diferenciação celular as Serine-arginine Protein Kinases (SRPKs) são uma classe de enzimas estudadas como alvo para tratamento de algumas doenças em mamíferos, dentre eles, os humanos. Apesar do pouco número de informações, essas enzimas já foram previamente descritas em algumas espécies de tripanosomatídeos, como em T. cruzi e *Leishmania infantum* (LiSRPK).



Em amarelo, LiSRPK. Em rosa, HsSRPK1.

Objetivo

O SRPIN340 é um inibidor conhecido de SRPK em células de humanos, porém, estudos realizados demonstraram baixa atividade inibitória quando testada na SRPK de L. infantum (LiSRPK). Logo, o objetivo deste trabalho é realizar a triagem de substâncias análogas do SRPIN340, como os compostos chamados DIM, frente a atividade da LiSRPK e propor o desenho de novas estruturas.



SRPIN340 e DIM1, análogo da molécula.

Referências

[1] Portal D, Lobo GS, Kadener S, Prasad J, Espinosa JM, Pereira CA, Tang Z, Lin RJ, Manley JL, Kornblihtt AR, Flawiá MM, Torres HN. Trypanosoma cruzi TcSRPK, the first protozoan member of the SRPK family, is biochemically and functionally conserved with metazoan SR protein-specific kinases. Mol Biochem Parasitol. 2003 Mar;127(1):9-21.

[2] Jacky Chi Ki Ngo, Justin Gullingsrud, Kayla Giang, Melinda Jean Yeh, Xiang-Dong Fu, Joseph A. Adams, J. Andrew McCammon, Gourisankar Ghosh, SR Protein Kinase 1 Is Resilient to Inactivation, Structure, Volume 15, Issue 1, 2007,Pages 123-133.

Metodologia

1. Modelagem da LiSRPK

TriTrypDB — Kinetoplastid Informatics Resources

2. Minimização de Energia e Alinhamento estrutural

YASARA

Alinhamento estrutural entre LiSRPK e HsSRPK1 com RMSD no valor de 0,590

3. Validação do Modelo

ERRAT — Fator de qualidade - 96,033

PROCHECK — 91,5 dos resíduos em regiões favoráveis

ProSA-web Protein Structure Analysis — Z-Score = -8,15

Verify 3D — 90,38 dos resíduos com pontuação 0,2

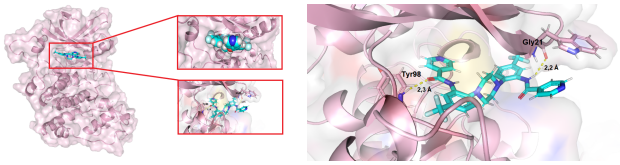
4. Ancoragem Molecular

AutoDock 4

BIOVIA

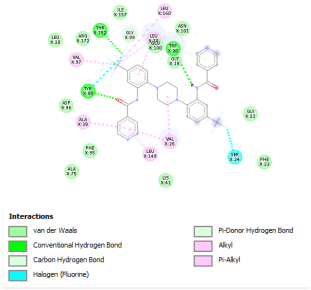
Resultados

- Ancoragem molecular com o análogo DIM1.



Interações de Hidrogênio válidas entre resíduos de Tirosina e Glicina presentes no sítio ativo e o composto DIM1

- Diagrama 2D das interações do sítio ativo da LiSRPK com o análogo do SRPIN340, DIM1.



Conclusões

1. A estrutura predita teve um bom RMSD em relação ao template, no valor de 0,590
2. Várias posições foram obtidas para os complexos entre a LiSRPK e os análogos, as quais precisam passar por outros tipos de análise.

Construção de candidatos a modelo para Arginina quinase de *Trypanosoma cruzi*

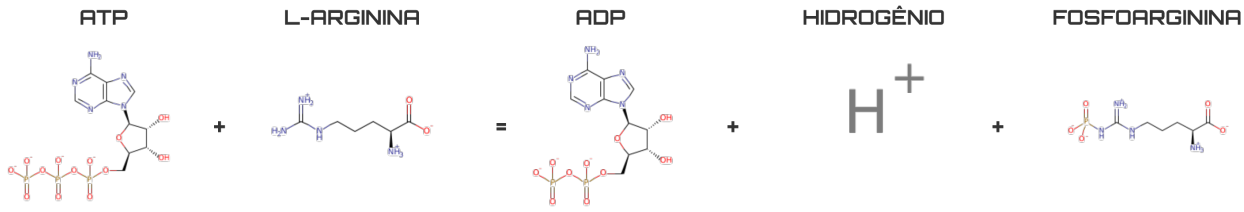
Tamara Lima da Silva[1,2], Ana Carolina Silva Bulla[3], Manuela Leal da Silva[2,3,4]

[1] Universidade Católica de Petrópolis, Petrópolis/RJ, Brasil
[2] Instituto Nacional de Metrologia, Qualidade e Tecnologia, Duque de Caxias/Rj, Brasil
[3] Instituto Oswaldo Cruz, Rio de Janeiro/RJ, Brasil
[4] Instituto de Biodiversidade e Sustentabilidade - UFRJ Macaé/RJ, Brasil

tamara.lima7@hotmail.com, anabulla@aluno.fiocruz.br, manuela@macae.ufrj.br

Introdução

Novos alvos terapêuticos para doença de Chagas vêm sendo explorados uma vez que o tratamento atual é baseado nos antiparasitários nifurtimox e benznidazol que apresentam alta toxicidade e diversos efeitos colaterais. A arginina quinase (AQ), uma fosfotransferase pertencente à família das guanidino quinases, mostrou-se um alvo promissor por ser uma enzima chave à viabilidade celular.

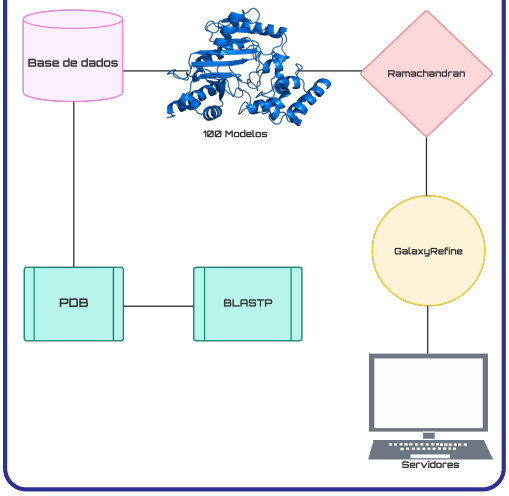


Objetivo

Diante disso, o trabalho teve como objetivo a construção de modelos tridimensionais para AQ de *T.cruzi* através de ferramentas *in silico*.

Metodologia

Os modelos foram construídos por modelagem comparativa com Modeller (vs 24) onde para seleção do molde foi feita uma busca no *Protein Data Bank* (PDB) com *BLASTP*. Os 100 candidatos gerados foram ranqueados a partir das informações do gráfico de *Ramachandran* e do desvio médio quadrático (RMSD) sendo o modelo final submetido ao refinamento no *GalaxyRefine* e aos servidores *MolProbity*, *Verify3D*, *PROSA* e *ERRAT* para análise.



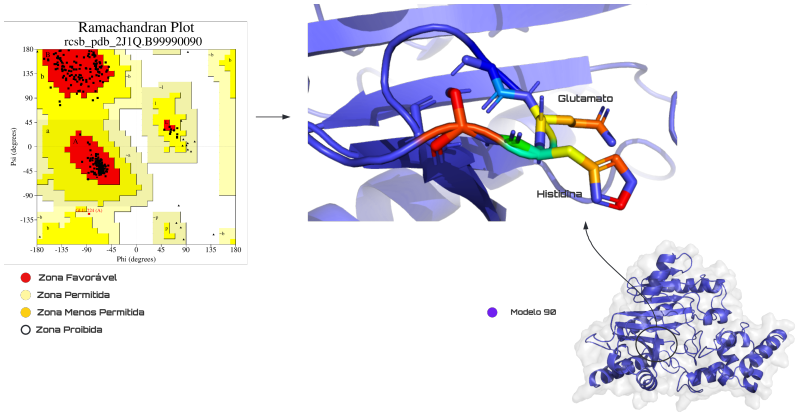
Referências

[1] FERNANDEZ, Pablo et al. The crystal structure of Trypanosoma cruzi arginine kinase. Proteins: Structure, Function, and Bioinformatics, v. 69, n. 1, p. 209-212, 2007.rlag.

[2] PEREIRA, Claudio A. et al. Trypanosoma cruzi arginine kinase characterization and cloning: a novel energetic pathway in protozoan parasites. Journal of Biological Chemistry, v. 275, n. 2, p. 1495-1501, 2000.

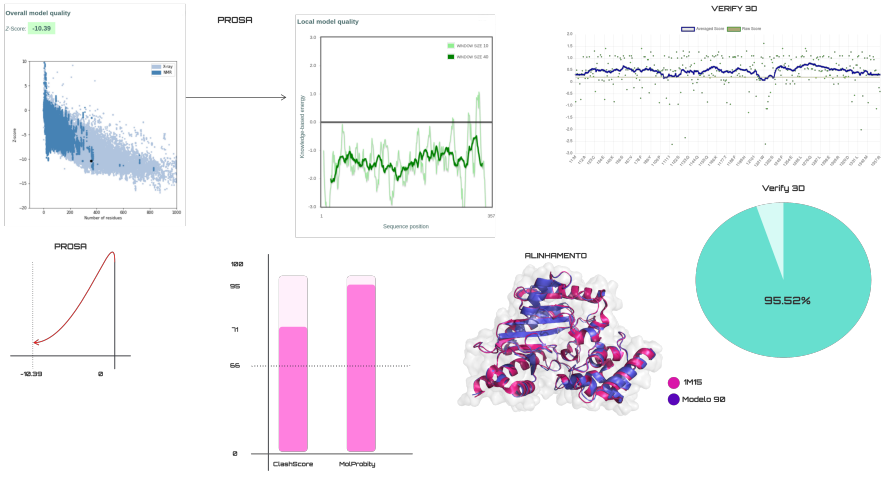
Resultados 1

As proteínas AQ de *Limulus polyphemus* (PDBid: 1M15) e do próprio *T.cruzi* (PDBid: 2J1Q) foram usadas como molde onde o modelo 90 apresentou 95,2 porcento de resíduos na região permitida e menor RMSD (0,488 Å).



Resultados 2

No *MolProbity* foram obtidos valores de *ClashScore* (71 percentil) e *MolProbity* score (95 percentil) na qual os valores esperados são de um percentil 66 para ambos. Os resultados do *Verify*3D (95,52) e *ProSa* (-10.39) indicaram um modelo com energia e estrutura favoráveis.



Conclusão

Portanto, foi possível gerar um modelo tridimensional para AQ de *T. cruzi* possibilitando estudos posteriores para busca de potenciais inibidores com base em uma estrutura alvo.

Detecção de sintomas de COVID-19 em texto usando redes Transformers

Vítor Machado[1], Clécio R. Bom[1], Kary Ocaña[2], Rafael Terra[2], Miriam B. F. Chaves[2]

[1]Centro Brasileiro de Pesquisas Físicas (CBPF/MCTI)
[2] Laboratório Nacional de Computação Científica (LNCC/MCTI)

vmachado@cbpf.br, debom@cbpf.br, karyann@lncc.br, rafaelst@lncc.br, mbcm@lncc.br

Introdução

Este projeto é um passo no caminho do desenvolvimento de uma futura ferramente clínica automatizada auxiliar para identificar sintomas clínicos apresentados por pacientes. Foi utilizada uma rede neural com arquitetura BERT, versão BERTimbau pré-treinada em português brasileiro, para identificar 14 sintomas de COVID.

- Uma base de dados em português brasileiro foi construída a partir de dados extraídos do Twitter;
- Foram testadas diferentes técnicas de Aprendizado Produzido para classificação.

Redes *Transformers*

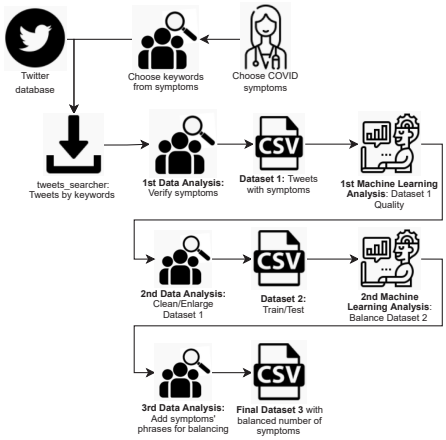
Redes Transformers são uma técnica de Aprendizado Profundo introduzido em 2017 para problemas de Processamento de Linguagem Natural. A principal ideia por trás da tecnologia Transformers é o uso de um mecanismo de atenção aprendível que escolhe quais partes do texto são mais relevantes para entender o significado de cada palavra.

Criação da base de dados

No presente projeto, foi feita uma análise de dados em tweets sobre COVID-19 no Brasil no período de 1 de fevereiro à 31 de abril de 2021. O seguinte *workflow* foi utilizado para a criação da base de dados:

- Escolha dos sintomas e palavras-chave;
- TweetSearcher para coletar dados;
- Processamento e análise dos dados;
- Geração da base de dados;
- Repetir quantas vezes for necessário.

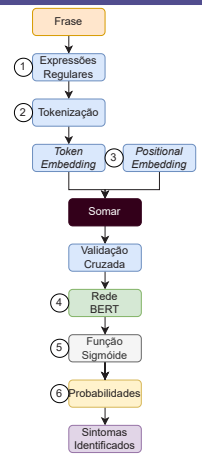
A execução do *workflow* é conduzida em uma interação humano-máquina com especialitas em análise de dados, saúde e aprendizado profundo. Além disso, o *workflow* pode ser executado repetidamente até se alcançar um consenso na qualidade dos dados.



Análise da Base de Dados de COVID-19

Foi utilizada uma rede BERT de tamanho base com pesos e tokenizador de uma versão chamada BERTimbau, pré-treinada em português brasileiro. O processo completo pela qual cada dado de nosso conjunto passou está resumido abaixo:

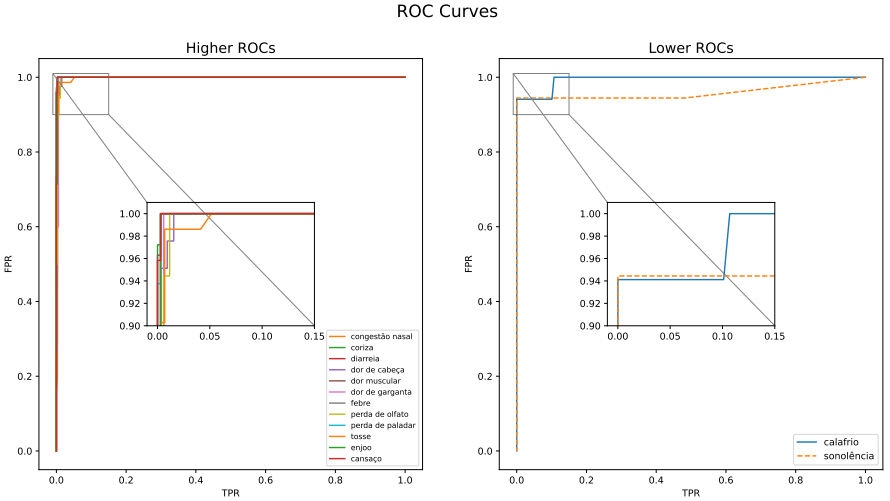
- Limpeza dos dados usando *RegEx*;
- Tokenização feita pela BERTimbau;
- Representação dos tokens e posições através de *Embeddings*;
- Validação Cruzada para avaliação de erros;
- Processamento pela rede BERTimbau;
- Geração de probabilidades com um função sigmóide;
- Lista de sintomas identificados.



Resultados

Usamos diversas métricas para validar os resultados e escolher entre os modelos usados. Entre elas, **ROC-AUC**, precisão e acurácia. Os melhores resultados vieram do uso da rede BERTimbau de tamanho base, sem pesos, para o cálculo da função perda (Entropia Cruzada Binária com Logits). Abaixo estão as Curvas ROC e suas ROC-AUC para cada sintoma.

- Mais de 97% de acurácia para cada sintoma individualmente;
- Mais de 0.95 de AUC-ROC paracada sintoma individualmente;
- Precisão acima de 84% para cada sintoma individualmente.



Conclusões

A arquitetura BERT usando a versão BERTimbau se mostrou eficiente (>0.95 AUC-ROC) ao identificar todos os 14 sintomas de COVID-19 selecionados em frases do Twitter em português. A partir destes resultados, abrem-se portas para a utilização da rede treinada para aplicações em mundo real, como em um *chatbot* para anamnese, trabalho que vêm sendo desenvolvido no momento.

References

[1] Vaswani, A. et al. 2017. Attention Is All You Need. Em: arXiv 1706.03762.

[2] Devlin, J. et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Em: arXiv 1810.04805.