Preservação de Privacidade entre Indivíduos com Semelhança Genômica

Manuel Edvar B. Filho[1], **Javam C. Machado**[1]

[1]LSBD – Departamento de Computação – Universidade Federal do Ceará

{edvar.filho, javam.machado}@lsbd.ufc.br

Resumo. *O crescimento na produção de tecnologias que auxiliam no sequenciamento genético tem sido acompanhado do aumento na produção de dados genômicos das pessoas. Ao analisar esses dados é possível identificar informações pessoais e familiares acerca dos indivíduos, muitas delas informações sensíveis. Coloca-se assim a necessidade de se preservar a privacidade dos indivíduos quando da análise desse tipo de dado. É comum nas instituições de saúde executar o processo de comparar o dado genômico de um indivíduo com um conjunto de dados de outros pacientes, buscando encontrar semelhanças entre eles a fim de realizar análises e tratamentos similares. Este trabalho estuda a preservação de privacidade dos indivíduos neste processo. Investigamos perturbar o dado genômico por meio da privacidade diferencial com o objetivo de permitir análises úteis e ao mesmo tempo dificultar a reidentificação do titular do genoma.*

Nível: Mestrado

Programa: Programa de Pós-Graduação em Ciência da Computação

Ingresso: Março de 2020

Previsão de término: Março de 2022

Etapas concluídas: Revisão bibliográfica e Definição do problema

Defesa da qualificação: 11 de Agosto de 2021

Lista de publicações:

- Publicação Diferencialmente Privada de Dados de Pacientes de COVID-19 - Filho, Manuel E. B.; Neto, Eduardo R. D.; Machado, Javam C. - ***SBBD 2020 - Short and Vision and Industrial Papers***
- Detectando Doença de Parkinson - Uma Comparação de Modelos de Aprendizagem de Máquina com Redução de Dimensionalidade Diferencialmente Privada - Filho, Manuel E. B.; Silva, Maria de Lourdes M.; Barros, Patrícia V. S.; Mattos, César L. C.; Machado, Javam C. - ***SBBD 2020 - Short and Vision and Industrial Papers***
- Privacy-preserving of patients with Differential Privacy: an experimental evaluation in COVID-19 dataset - Filho, Manuel E. B.; Neto, Eduardo R. D.; Machado, Javam C. - ***JIDM Vol.12 2021*** - Aceito para publicação

1. Introdução

As melhorias tecnológicas no sequenciamento de genoma tem gerado uma enorme quantidade de dados genéticos que são de suma importância para muitas pesquisas. Os dados genômicos são coletados para fornecer serviços relacionados à saúde. Além disso, carregam informações diferentes de outros dados de saúde, por exemplo, informações que podem ser utilizadas para verificar a suscetibilidade de um indivíduo adquirir alguma doença futuramente e auxiliar a prevenção da mesma doença. Outra propriedade importante dos dados genômicos é a presença de semelhanças entre indivíduos que são parentes de sangue, pois a formação de um indivíduo se dá a partir da combinação de material genético dos seus pais. Isso leva indivíduos de uma mesma árvore genealógica a possivelmente apresentar as mesmas doenças hereditárias.

Quando dados genômicos são mal utilizados, uma variedade de abusos e ameaças podem acontecer. O aumento do uso e compartilhamento dessas informações levanta questões éticas e de privacidade. Em geral os problemas de privacidade relacionados a dados genômicos são: o genoma pode ser utilizado na re-identificação de um indivíduo; pode revelar informações acerca de doenças genéticas; contém informações de familiares, e o compartilhamento pode revelar problemas de saúde de uma família completa; o genoma não muda com o tempo; acarreta inúmeras questões éticas. Além disso, a integração de dados genômicos com outros dados sensíveis à privacidade, como por exemplo a localização, pode aumentar o risco de violação de privacidade.

No ambiente de estudos da genética, as misturas de DNA são caracterizadas por um DNA resultante da combinação do material genético de diversos indivíduos. Elas são muito utilizadas na investigação forense, com o intuito de identificar indivíduos maliciosos em situações de risco. Além desses indivíduos, também há a presença de indivíduos que não são ameaças em cenas de crime.

Neste trabalho desejamos dificultar a identificação de informações sensíveis ao comparar o material genético de um indivíduo com uma mistura de DNA pública. Neste cenário o atacante faz consulta para obter informações sobre o DNA de um indivíduo alvo. Nossa proposta visa adicionar aleatoriedade garantida pela privacidade diferencial para dificultar as análises posteriores realizadas pelo atacante garantindo a privacidade do indivíduo alvo.

2. Fundamentação Teórica

2.1. Privacidade diferencial

A Privacidade Diferencial é uma técnica que fornece garantias sólidas de privacidade de indivíduos [Dwork et al. 2006, McSherry and Talwar 2007, Lecuyer et al. 2019]. Os mecanismos diferencialmente privados retornam uma resposta adicionada de ruído, diminuindo o risco de um atacante inferir algo a partir das respostas. A ideia por trás da Privacidade Diferencial é que a presença ou ausência de qualquer indivíduo no conjunto de dados não mudará a probabilidade de saída da consulta.

Definição 2.1 (Privacidade Diferencial) *Um mecanismo M é ϵ-diferencialmente privado (PD) se para quaisquer conjunto de dados D_1 e D_2 que diferem no máximo em um elemento, e para qualquer conjunto S de todas as saídas possíveis de M,*

$$Pr[M(D_1) \in S] \leq \exp(\epsilon) \times Pr[M(D_2) \in S] \quad (1)$$

O *budget* ϵ limita o impacto da adição ou remoção de qualquer indivíduo em um conjunto de dados. Um pequeno valor de ϵ indica que qualquer indivíduo irá introduzir uma mudança mínima na distribuição do mecanismo, dando maior proteção.

O mecanismo exponencial [McSherry and Talwar 2007] é utilizado para consultas sobre dados categóricos que retornam alguma das possibilidades das categorias. Ele foi proposto para situações em que adicionar ruído a uma quantidade pode afetar drasticamente a resposta. Dada uma função f e que O seja o conjunto de todas as saídas possíveis. Para satisfazer ϵ-PD, um mecanismo deve gerar os valores de O seguindo uma distribuição de probabilidade.

O mecanismo exponencial é definido com respeito a uma função de utilidade dada por $u : (\mathbb{D} \times O) \rightarrow \mathbb{R}$, que mapeia os pares de conjuntos de dados e saídas possíveis para valores de utilidade. Em relação a função de utilidade, intuitivamente, espera-se que para o valor real da consulta e para possibilidades próxima a este valor, a função de utilidade resulte em valores maiores do que quando temos possibilidades de saída com baixa probabilidade de ocorrer. Não há um padrão para funções de utilidade, e dependendo da função escolhida, esta escolha impactará diretamente na privacidade alcançada.

Definição 2.2 (Mecanismo Exponencial [McSherry and Talwar 2007]) *Dada qualquer função $u : (\mathbb{D} \times O) \rightarrow \mathbb{R}$, e um orçamento de privacidade ϵ o mecanismo de exponencial $M_f(D)$ gera saída $o \in O$ com probabilidade proporcional a $exp(\frac{\epsilon u(D,o)}{2\Delta u})$, onde:*

$$\Delta u = \max_{\forall o, D \simeq D'} |u(D, o) - u(D', o)| \tag{2}$$

é a sensibilidade da função de utilidade. Isso é,

$$Pr[M_f(D) = o] = \frac{exp(\frac{\epsilon u(D,o)}{2\Delta u})}{\sum_{o' \in O} exp(\frac{\epsilon u(D,o')}{2\Delta u})} \tag{3}$$

Teorema 2.1 (Mecanismo Exponencial) *O mecanismo exponencial satisfaz ϵ-privacidade diferencial. [Li et al. 2016]*

Uma propriedade da privacidade diferencial que é de suma importância para o nosso trabalho é a garantia do pós-processamento, em que a composição de um mecanismo ϵ-PD e qualquer pós-processamento ainda satisfazem ϵ-PD.

Proposição 2.1 (Pós-processamento) *[Dwork and Roth 2014] eja M_1 um mecanismo que satisfaz ϵ-PD, então para qualquer algoritmo M_2, a composição de M_1 e M_2, i.e., $M_2(M_1(\cdot))$ satisfaz ϵ-PD.*

2.2. Dados Genômicos

Um DNA ou ácido desoxirribonucleico é composto de quatro bases nitrogenadas, denominadas A - adenina, T - timina, C - citosina e G - guanina. A sequência dessas bases indica as informações disponíveis para a construção de um organismo. Um genoma - o conjunto completo de DNA de um indivíduo - é a representação das características biológicas pessoais de um indivíduo e, portanto, é um dado sensível e deve ter uma garantia de preservação de privacidade. O genoma humano é composto por mais de 3,2 bilhões de pares de bases nitrogenadas distribuídos ao longo de 23 pares de cromossomos. Cerca

de 0,5%, apenas, dos pares de bases diferem entre quaisquer dois indivíduos, e esta quantidade pode ser menor em caso de pessoas relacionadas, que possuem algum parentesco [Wang and Moult 2001].

As diferenças mais comuns no genoma humano ocorrem em apenas um único par de bases nitrogenadas. Essas diferenças são chamadas de polimorfismos de nucleotídeo único (SNPs), e são responsáveis pelas diferenças no fenótipo e no genótipo, que são genes traduzidos em proteínas e causam nossos fenótipos. Os valores utilizados para representar um SNP é o número de alelos menores que ele possuí: 0, 1 ou 2. Então, podemos reafirmar que o genoma é considerado informação privada , contendo informações sobre a família e os ancestrais de um indivíduo. O genoma de um indivíduo contem informações únicas de indivíduos e seus familiares, portanto dados sensíveis que os seus titulares muitas vezes desejam mantê-los preservados de liberação.

A análise de dados genômicos permite a identificação de características específicas de um indivíduo a ponto de permitir ataques de identificação ou inferência do fenótipo que irão a devastar a privacidade do indivíduo. Tratando-se de ataques de identificação, um terceiro malicioso tem acesso aos dados genômicos "anônimos" e consegue recuperar com sucesso a identidade dos indivíduos. Em ataques de inferência de fenótipo, o atacante que tem acesso a informações genéticas parciais de um indivíduo conhecido deseja inferir dados sensíveis acerca deste, como doenças, por exemplo. O atacante pode descobrir informações genômicas por meio de imputação de genótipo, explorando, por exemplo, os parentes do indivíduo alvo. A perda de privacidade também ocorre quando estatísticas agregadas são divulgadas [Wang et al. 2009, Shringarpure and Bustamante 2015]. O alto acesso aos dados genômicos é de extrema importância para fins de pesquisa. Mas, é um desafio compartilhar estes dados de modo que se garanta o equilíbrio entre privacidade e utilidade.

3. Trabalhos Relacionados

Em [Asharov et al. 2018] temos a busca de k pacientes similares utilizando os dados genômicos de diversos pacientes. Trata-se de um processo a ser realizado por médicos que ao realizarem uma consulta de um paciente, busca encontrar similares em uma base de dados para realizar tratamentos e dar diagnósticos com base nos k pacientes similares e os tratamentos atribuídos aos mesmos. A relação de proximidade é utilizada por meio da distância de edição aproximada, em que a comparação não é realizada por todas as sequências genéticas completas, mas cada sequência é dividade em blocos e para cada bloco de todas as sequências, realizamos os cálculos de distâncias somente para as sequências únicas. Segundo os autores, o protocolo proposto é semi-honesto. A privacidade se dá quando uma das partes envolvidas no protocolo, cliente ou servidor, estão corrompidas. Quando isso ocorre, é possível calcular as variáveis envolvidas tendo somente uma das partes garantidas. Acontece, assim, a reconstrução dos valores necessários da parte corrompida a partir dos valores disponíveis. Sendo que os dados reconstruídos, não são os originais que são obtidos quando as duas partes estão em perfeito funcionamento.

Em [Almadhoun et al. 2019] é apresentado uma nova versão de privacidade diferencial, esta aplicada diretamente a presença de tuplas dependentes no conjunto de dados a ser privado. Como exemplo, o trabalho traz a abordagem de tuplas dependentes ligada

diretamente a dados genéticos, mais especificamente aos SNPs. A aplicação de privacidade diferencial sobre tuplas dependentes, veio a tona quando a definição de privacidade diferencial proposta em [Dwork et al. 2006] não é aplicada a dependências dos registros no conjunto de dados. Com isso, para a nova definição das variáveis utilizadas em privacidade diferencial, um modelo de ataque foi proposto para analisar o impacto do vazamento de informações quando o adversário assume ou não a existência de tuplas dependentes no conjunto de dado a ser atacado.

Nosso problema se trata da identificação de informações sensíveis de um indivíduo após análise e comparação com uma mistura de DNA. Muitas soluções propostas para problemas similares a esse, utilizam métodos sintáticos de garantia de privacidade.Diferentemente destes trabalhos e dos trabalhos apresentados anteriormente, desejamos propor a geração de um material genético garantindo a similaridade com o material original, mas adicionando ruído decorrente da aleatoriedade da privacidade diferencial, de modo a preservar a privacidade do indivíduo dono do DNA e mantendo a utilidade dos dados. Além de analisar a construção das misturas de DNA e verificar a possibilidade de adicionar privacidade diferencial neste processo.

4. Metodologia

O modelo de nossa proposta é apresentado na Figura 1. Um atacante para realizar ataques de inferência a um indivíduo e desejar descobrir informações a cerca deste, deve possuir seu dado genético (Passo A). Com posse sobre esse dado o atacante consulta a mistura de DNA, disponível publicamente, e por fim verifica se o indivíduo está presente na mistura de DNA ou não para analisar e inferir as informações sensíveis (Passo C).

Nossa proposta está presente no fluxo entre o atacante e o DNA de um indivíduo alvo. No Passo A, o atacante solicita a um sistema específico o material genético de um indivíduo para verificar a sua presença na mistura de DNA. Com isso no Passo B, o sistema acessa uma base de dados contendo diversos DNAs, esta base disponibiliza para o sistema o DNA do indivíduo que o atacante deseja inferir informações. A garantia de privacidade está presente no Passo C, em que, para disponibilizar o material genético do indivíduo o sistema irá enviar para o atacante o DNA do indivíduo adicionado de ruído para garantir a privacidade. Por fim, no Passo D, sobre posse do DNA adicionado de ruído, o atacante compara este dado com a mistura de DNA disponibilizada publicamente para inferir informações sensíveis a cerca do indivíduo alvo. Esta inferência passa a ser dificultada devido a presença de ruído no dado genético, mas garantindo a utilidade do dado.

A privacidade será garantida no Passo C, em que há a disponibilização do material genético do indivíduo alvo para o atacante, para que este realize a comparação com a

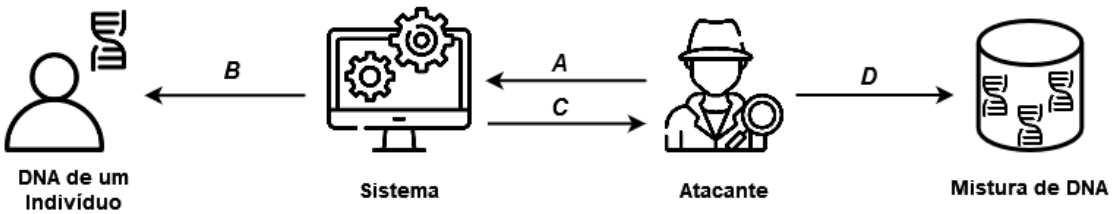

Figura 1. Modelo da proposta.

mistura de DNA. Utilizando os SNPs do DNA do indivíduo, utilizaremos o mecanismo exponencial apresentado anteriormente para preservação de privacidade da informação sensível que o indivíduo deseja manter segura. Como os SNPs são valores categóricos, a partir da quantidade de alelos menores presentes, este mecanismo é mais aplicável nestas situações.

Para utilização deste mecanismo é requerido uma função de utilidade que será necessária no cálculo da sensibilidade da consulta para que assim o mecanismo seja aplicado. A fim de garantir também a utilidade do dado, a aleatoriedade proveniente da privacidade diferencial, em especial do mecanismo exponencial, será aplicada no local do DNA em que há a presença da informação sensível a ser protegida. Com isso, o trecho que pode ser disponibilizado referente a informação sensível serão as possibilidades que podem ser apresentadas, dentre elas o trecho original do DNA do indivíduo.

A função de utilidade utilizada será a distância entre o DNA original do indivíduo alvo e o DNA com a modificação no trecho da informação sensível, para cada uma das possibilidades hábeis. A sensibilidade da função de utilidade será dada pela maior diferença entre os valores obtidos na função de utilidade dada pela seguinte equação:

$$D(Y, Y') = \sum_{i=1}^{n} |x_i - x'_i|, \tag{4}$$

em que Y e Y' são respectivamente o material genético original do indivíduo e o modificado no trecho da informação sensível. n é a quantidade de SNPs, x_i e x'_i são os valores do i-ésimo SNP no DNA original e no DNA modificado, respectivamente. O valor do SNP será de acordo com a porcentagem da presença dos alelos menores no par de base nitrogenadas que é 0, 1 ou 2, assim o valor utilizado na função de utilidade será 0, 0.5 ou 1, respectivamente.

Nossa abordagem será comparada com os trabalhos relacionados apresentados. Portanto, planejamos nos posicionar contra uma abordagem que utiliza circuitos truncados de Yao, assim como contra uma abordagem que aplica privacidade diferencial para tuplas dependentes. Um dos maiores desafios de nossa proposta é a semântica dos dados, pois com a utilização de dados genômicos, qualquer tipo de tratamento indevido impactará negativamente na utilidade alcançada. Um outro desafio diz respeito à dimensionalidade dos dados pois são dados de grande dimensão, o que impacta negativamente no desempenho para qualquer algoritmo que se dispuser a tratá-los.

A metodologia para implantação da proposta apresentada está ancorada nos estudos do estado da arte do problema apresentado e na análise das soluções propostas por outros autores. Estamos investigando a construção de uma estratégia para introduzir o ruído aleatório ao disponibilizar o DNA de um indivíduo para estudo e experimentar para consultas em datasets de DNAs. Pretendemos analisar os resultados experimentais a partir do volume de erro introduzido pelo ruído e da acurácia alcançada em relação ao nível de privacidade desejado.

5. Conclusão

Com o crescimento na quantidade de dados genômicos foi visto a necessidade de garantia de privacidade dos mesmos. Ao analisá-los é possível descobrir informações pessoais e

familiares do indivíduo. No intuito de propor algo que garanta a privacidade de diversos dados genômicos, realizaremos um algoritmo para disponibilização do DNA de um indivíduo a partir da utilização do mecanismo exponencial para garantir a privacidade deste ao compará-lo com uma mistura de DNA, que acarretará na descoberta de informações sensíveis. Após a construção do algoritmo, uma análise dos resultados será executada, e a compreensão dos mesmos, também. Além de que outras abordagens de privacidade diferencial também poderão ser questionadas ou propostas a partir de uma análise posterior.

Referências

Almadhoun, N., Ayday, E., and Ulusoy, O. (2019). Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics*, 36(6):1696–1703.

Asharov, G., Halevi, S., Lindell, Y., and Rabin, T. (2018). Privacy-preserving search of similar patients in genomic data. *Proc. Priv. Enhancing Technol.*, 2018(4):104–124.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.

Li, N., Lyu, M., Su, D., and Yang, W. (2016). *Differential Privacy: From Theory to Practice*. Synthesis Lectures on Information Security, Privacy, & Trust. Morgan & Claypool Publishers.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 94–103. IEEE Computer Society.

Shringarpure, S. S. and Bustamante, C. D. (2015). Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646.

Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. (2009). Learning your identity and disease from research papers: information leaks in genome wide association study. In Al-Shaer, E., Jha, S., and Keromytis, A. D., editors, *Proceedings of the 2009 ACM Conference on Computer and Communications Security, CCS 2009, Chicago, Illinois, USA, November 9-13, 2009*, pages 534–544. ACM.

Wang, Z. and Moult, J. (2001). Snps, protein structure, and disease. *Human mutation*, 17(4):263–270.