**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO
EM BIOINFORMÁTICA**

JOSÉ DENEY ALVES DE ARAÚJO

**INTEGRAÇÃO DE BASES DE DADOS ADMINISTRATIVOS COM FERRAMENTAS
GENÔMICAS**

Versão Resumida, conforme Resolução CoPGr nº 7569, de 03 de outubro de 2018, a
qual regulamenta a disponibilização das dissertações e teses no Portal da
Universidade de São Paulo

São Paulo
2021

JOSÉ DENEY ALVES DE ARAÚJO

**INTEGRAÇÃO DE BASES DE DADOS ADMINISTRATIVOS COM FERRAMENTAS
GENÔMICAS**

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo (USP), como parte dos requisitos para obtenção do título de doutor em Bioinformática.

Orientador: Dr. HELDER TAKASHI IMOTO NAKAYA

São Paulo

2021

FICHA CATALOGRÁFICA

A663 Araújo, José Deney Alves de
Integração de bases de dados administrativos com ferramentas genômicas / José Deney Alves de Araújo, orientador [Prof.] Dr. Helder Takashi Imoto Nakaya.
São Paulo : 2021.
47 p.

Tese (Doutorado) - Universidade de São Paulo
Orientador: [Prof.] Dr. Helder Takashi Imoto Nakaya
Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Codificação em DNA. 2. Record linkage. 3. Ferramentas genômica. 4. Epidemiologia. 5. BLAST. I. Nakaya, Helder Takashi Imoto, orientador II. Universidade de São Paulo. III. Instituto de Matemática e Estatística. IV. Título

CDD: 572.8

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela Bibliotecária Maria Lucia Ribeiro CRB-8/2766

**À minha mãe Deuza (*in memorian*), que mesmo não tendo a
oportunidade de acesso à educação, teve o empenho
em me educar sempre em primeiro lugar.
Aqui estão os resultados dos seus esforços.
Com muita gratidão!**

Agradecimentos

Ao meu Pai Arleison e minha Mãe Deuza (*in memoriam*). Minha mãe sempre me dizia: “você vai ter a oportunidade de estudo que nunca tive”. Essa oportunidade me dada foi árdua, juntos muitas vezes transportavam cargas pesadas nos ombros para me manter nos estudos. Minha base de sustentação diária!

Meu irmão Denis e minhas irmãs: Deuzenilde, Deuzenir, Damízia e Débora que sempre me apoiaram e a resposta em comum sempre foi: o que você decidir, estamos juntos. Minha gratidão!

Ao meu orientador Helder Nakaya, pela oportunidade e a incrível liberdade que nos dá à pesquisa. Mentor fundamental para impulsionar ideias malucas e incríveis como a proposta aqui. Obrigado por acreditar e compartilhar sua visão de mundo!

Aos meus amigos e colegas do CSBL: Allan, Alysson, Amanda, André Nicolau, Bruna, Durão, Edson, Fred, Ícaro, Jeevan, Leandro, Mauro, Rodrigo, Felipe, Thomaz, Tiago, Vanessa, Vitor, Vivi e Youvika. Um especial agradecimento ao Thiago e Pati pelas correções e sugestões.

Ao André Guilherme (com sua mega experiência em genômica) e Juan Silva (doutorando) que fazem parte deste trabalho. Sem a força que me deram em longas horas de discussão este trabalho não teria avançado na mesma proporção.

Aos ex-CSBLers: César, Pedro, Diógenes, Mindy, Mary, Naty, Lucas, Fernando, Gustavo e Fábio pelos rolês e contribuição no início da ideia do trabalho.

Ao meu amigo/irmão Willian Lira, pelas trocas de ideias e força tarefa nos primeiros protótipos gerados.

À população de São Paulo, que por meio da contribuição dos impostos estaduais, tive a oportunidade de ter acesso à Universidade de São Paulo (USP), que fornece uma estrutura completa de ensino e pesquisa.

Finalmente, um agradecimento mais que especial a minha noiva Joana Azevedo por toda paciência e amor dedicado, inclusive em muitos feriados e fins de semanas sacrificados à programação. Você faz parte dessa caminhada!

*“A tarefa não é apenas ver aquilo que ninguém viu, mas pensar o que ninguém
ainda pensou sobre aquilo que todo mundo viu.”*
(Arthur Schopenhauer)

RESUMO

Araújo, JDA. **Integração de Bases de Dados Administrativos com Ferramentas Genômicas**. 2021. Tese (Doutorado em Ciências) - Universidade de São Paulo, São Paulo, 2021.

A pesquisa em saúde pública frequentemente requer a integração de informações de diferentes fontes de dados. As metodologias de *record linkage* (RL) utilizam os campos de identificação de cada registro para vincular indivíduos de diferentes bancos de dados. No entanto, erros nos registros e o alto custo computacional tornam o RL um grande desafio para integrar grandes bancos de dados administrativos. Apresentamos Tucuxi-BLAST, uma ferramenta versátil para RL que utiliza uma abordagem de codificação e análise *in silico* de DNA para grandes bancos de dados administrativos. Pela reproposição de ferramentas genômicas, fomos capazes de integrar três bases de dados de saúde brasileiras com alta sensibilidade e especificidade e rastrear indivíduos em vários bancos de dados epidemiológicos. Comparado com cinco ferramentas de RL existentes, nosso método obteve a mais alta precisão e velocidade. Além disso, a etapa de validação independente usando 300 milhões de registros simulados, mostrou um consumo de memória RAM de apenas ~4GB e 23h de processamento em um *desktop* comum, sem necessidade do uso de plataformas de processamento de alto desempenho. Tucuxi-BLAST pode melhorar a pesquisa médica baseada em dados e fornece uma maneira rápida e precisa de integrar informações individuais em vários bancos de dados administrativos.

Palavras-chave: Codificação em DNA, *record linkage*, ferramentas genômica, epidemiologia, BLAST

ABSTRACT

Araújo, JDA. **Record Linkage of Administrative Databases with Genomic Tools.**

2021. Thesis (Doctorate in Sciences) - University of São Paulo, São Paulo, 2021.

Public health research frequently requires integrating information from different data sources. Record linkage (RL) methodologies utilize the identification fields of each record to link individuals from different databases. However, errors in the records and the high computational costs involved make RL a major challenge for linking large administrative databases. We present Tucuxi-BLAST, a versatile tool for RL that utilizes a DNA-encoded approach to analyze massive administrative databases. By repurposing genomic tools, we were able to integrate three Brazilian health databases with great sensitivity and specificity, and to perform subject tracing across multiple epidemiological databases. Compared to five existing RL tools, our method obtained the highest accuracy and speed. Furthermore, in an independent validation step using 300 million simulated records. On a desktop, the RAM memory consumption was only ~4GB and 23 hours of processing. Tucuxi-BLAST can improve data-driven medical research and provides a rapid and accurate way for linking individual information across several administrative databases.

Keywords: DNA-encoded, *record linkage*, genomic tools, epidemiological, BLAST