UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

José Bartolomeu Alheiros Dias Neto

Análise de dados de Coinfecção Tuberculose/HIV disponíveis no SINAN utilizando o banco de dados Neo4J

Recife

2023

José Bartolomeu Alheiros Dias Neto

Análise de dados de Coinfecção Tuberculose/HIV disponíveis no SINAN utilizando o banco de dados Neo4J

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Jeane Cecília Bezerra de Melo

Coorientador: Nara Suzy Aguiar de Freitas

Recife

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

A397a Alheiros Dias Neto, José Bartolomeu
Análise de dados de Coinfecção Tuberculose/HIV disponíveis no SINAN utilizando o banco de dados Neo4J / José Bartolomeu Alheiros Dias Neto. - 2023.
66 f. : il.

Orientadora: Jeane Cecilia Bezerra de Melo.
Coorientadora: Nara Suzy Aguiar de Freitas.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2023.

1. Coinfecção. 2. Análise Exploratória. 3. Bancos de Dados Orientados a Grafos. I. Melo, Jeane Cecilia Bezerra de, orient. II. Freitas, Nara Suzy Aguiar de, coorient. III. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por José Bartolomeu Alheiros Dias Neto às 9 horas do dia 27 de abril de 2023, no link https://meet.google.com/qgn-irtp-vff, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado "Análise de dados de Coinfecção Tuberculose/HIV disponíveis no SINAN utilizando o banco de dados Neo4J", orientado por Jeane Cecília Bezerra de Melo e aprovado pela seguinte banca examinadora:

Jeane Cecília Bezerra de Melo
DC/UFRPE

Rodrigo Nonamor Pereira Mariano de Souza
DC/UFRPE

À minha avó, Norma de Souza Alheiros Dias

Agradecimentos

Ao meus pais, pela colaboração e paciência. À minha avó, por ter me dado toda a melhor educação de que foi capaz. Às minhas orientadoras, Jeane Cecília Bezerra de Melo e Nara Suzy Aguiar de Freitas, pelas inúmeras(e maravilhosas) reuniões, cheias de conteúdo. Aos meus amigos e companheiros de curso. Aos meus professores. A meu bom deus.

“A persistência é o caminho do êxito.”
(Charles Chaplin)

Resumo

Pesquisas realizadas nas últimas décadas, indicam a necessidade de investigação de processos de infecção por múltiplos patógenos, denominados processos de coinfecção. Algumas coinfecções têm alcance mundial, envolvendo doenças tais como: HIV, malária, hepatite, dengue e, mais recentemente, COVID-19. Em um estudo realizado com 500 voluntários portadores do vírus HIV (Human Immunodeficiency Virus), observou-se que a coinfecção entre o vírus HIV e a MTB (Mycobacterium tuberculosis), bactéria causadora da tuberculose, produziu um aumento de chance de haver morte em 4.07 vezes, quando comparada com outros tipos de coinfecção.

O panorama apresentado indica a necessidade de realização de estudos que permitam identificar ocorrências, mapear sua incidência em termos geográficos, e mesmo incluir aspectos que favoreçam a compreensão dos mecanismos biológicos envolvidos em processos de coinfecção, quer seja para prevenção, diagnóstico ou tratamento. No Brasil, um instrumento que auxilia no planejamento da saúde, definindo e avaliando o impacto das intervenções, é o Sistema de Informação de Agravos de Notificação – SINAN, disponibilizado pelo Departamento de Informática do SUS (DATASUS). A utilização efetiva destes bancos possibilita uma identificação da realidade epidemiológica de determinada área geográfica. O livre acesso a todos os profissionais da área de saúde, corrobora com a democratização de acesso à informação, permitindo que estas sejam disponibilizadas para a comunidade.

Neste trabalho foi realizada uma análise exploratória dos dados relativos a processos de coinfecção de TB e HIV, advindos do SINAN, com o objetivo de propor métodos que facilitem a utilização dos dados desse sistema por profissionais da área de saúde, que não tenham formação técnica em computação. Considerando que tal aplicação é fortemente embasada no relacionamento de dados, optou-se por propor um mapeamento dos dados em bancos não convencionais, orientados a grafos, como o Neo4J. Assim, além de simplificar a modelagem, as aplicações desse tipo costumam ser mais rápidas, quando comparadas a aplicações tradicionais (utilizando bancos de dados relacionais). Portanto, o mapeamento de dados disponíveis no SINAN para o Neo4J, permitiu uma visualização mais perceptível das correlações, possibilitando uma análise de múltiplos fatores e características de processos de coinfecção, potencializando as informações obtidas a partir das bases do SINAN e do Sistema de Tabulação de Dados disponibilizada pelo órgão, o TABNET.

Palavras-chave: Coinfecção, Análise Exploratória, Bancos de Dados Orientados a Grafos.

Abstract

Research carried out in recent decades indicates the need to investigate infection processes by multiple pathogens, called co-infection processes. Some coinfections have a worldwide reach, involving diseases such as: HIV, malaria, hepatitis, dengue and, more recently, COVID-19. In a study carried out with 500 volunteers carrying the HIV virus (Human Immunodeficiency Virus), it was observed that the coinfection between the HIV virus and MTB (Mycobacterium tuberculosis), the bacterium that causes tuberculosis, produced an increase in the chance of death by 4.07 times when compared to other types of co-infection.

The panorama presented indicates the need for studies to identify occurrences, map their incidence in geographic terms, and even include aspects that favor the understanding of the biological mechanisms involved in co-infection processes, whether for prevention, diagnosis or treatment. In Brazil, an instrument that helps in health planning, defining and evaluating the impact of interventions, is the Information System for Notifiable Diseases – SINAN, made available by the Department of Informatics of the SUS (DATASUS). The effective use of these databases makes it possible to identify the epidemiological reality of a given geographic area. Free access to all health professionals corroborates the democratization of access to information, allowing it to be made available to the community.

In this work, an exploratory analysis was carried out on data relating to TB and HIV co-infection processes, coming from SINAN, with the objective of proposing methods that facilitate the use of data from this system by health professionals who do not have technical training in computing. Considering that such an application is strongly based on data relationships, it was decided to propose a mapping of data in unconventional databases, oriented to graphs, such as Neo4J. Thus, in addition to simplifying modeling, applications of this type tend to be faster when compared to traditional applications (using relational databases). Therefore, the mapping of data available at SINAN to Neo4J allowed a more perceptible visualization of correlations, enabling an analysis of multiple factors and characteristics of co-infection processes, enhancing the information obtained from the bases of SINAN and the Tabulation System of Data made available by the agency, TABNET.

Keywords: Coinfection, Exploratory Analysis, Graph Oriented Databases.

Lista de ilustrações

Lista de Códigos

Lista de tabelas

Lista de abreviaturas e siglas

CSV	*Comma-separated values*(Arquivo separado por vírgulas)
DATASUS	Departamento de Informática do Sistema Único de Saúde
FT	Fator de Transcrição
HIV	*Human Immunodeficiency Virus*
IDE	*Integrated Development Environment*(Ambiente de Desenvolvimento Integrado)
MTB	*Mycobacterium tuberculosis*
NCBI	*National Center for Biotechnology Information*
NoSQL	*Not Only SQL* (Não apenas relacional)
OMS	Organização Mundial de Saúde
PVHIV	Pessoas Vivendo com HIV
SINAN	Sistema de Informação de Agravos de Notificação
SQL	*Structured Query Language* (Linguagem de Consulta Estruturada)
TB	Tuberculose
TBDR	Tuberculose Drogarresistente

Sumário

1 Introdução

1.1 Apresentação

Pesquisas realizadas nas últimas décadas, indicam a necessidade de investigação de processos de infecção por múltiplos patógenos, denominados processos de coinfecção. Algumas coinfecções têm alcance mundial, envolvendo doenças tais como: HIV, malária, hepatite, dengue e, mais recentemente, COVID-19 (LANSBURY et al., 2020). Adicionalmente, resultados mostram que a ocorrência das duas infecções no mesmo organismo potencializa a agressividade de ambas, a frequência de ocorrências em humanos, juntamente com as possíveis relações benéficas mútuas entre os patógenos envolvidos, supera a frequência de infecção única em muitas comunidades (PIETRO et al., 2018); (PICANÇO-JUNIOR et al., 2022); (GRIFFITHS et al., 2011); (JÚNIOR et al., 2010) (WATERS et al., 2020).

Em um estudo realizado com 500 voluntários portadores do vírus HIV(*Human Immunodeficiency Virus*), no Cazaquistão(antiga república soviética), (MUKHATAYEVA et al., 2021) observaram que a coinfecção entre o vírus HIV e a *Mycobacterium tuberculosis(MTB)* - bactéria causadora da tuberculose - produziu uma chance de haver morte 4.07 vezes maior, quando comparada com outros tipos de coinfecção: como HIV/Hepatite B, HIV/Hepatite C e HIV/Doenças Sexualmente Transmissíveis(DSTs) (MUKHATAYEVA et al., 2021). Ainda, conforme relatório da Organização Mundial de Saúde (GLOBAL…, 2020), em 2019, das aproximadamente 1,4 milhão de mortes relacionadas à tuberculose, 208.000 eram pessoas HIV positivas. As taxas mais altas de coinfecção por HIV em pacientes com MTB estão na Região Africana, onde 44% dos pacientes com TB com resultado de teste de HIV em 2010 eram HIV-positivos (intervalo entre os países com alta carga de TB/HIV, 8%–82%), seguidos pela Região das Américas (17%) (MONTALES et al., 2015).

O presente trabalho, busca abordar o tema da coinfecção, entre o vírus HIV e a *Mycobacterium tuberculosis(MTB)*, que desponta na literatura recente (RAWSON et al., 2020), (GHAZNAVI et al., 2022), (PETRELLIS et al., 2023), através de uma análise exploratória de uma base de dados pública, visando facilitar a análise multivariada destas para pesquisadores não especialistas em computação, através de um mapeamento em um banco de dados orientado a grafos.

1.1.1 Motivação

O panorama descrito na apresentação desse capítulo, indica a necessidade de realização de estudos que permitam identificar ocorrências, mapear sua incidência em termos geográficos, e mesmo incluir aspectos que favoreçam a compreensão dos mecanismos biológicos envolvidos em processos de coinfecção, quer seja para prevenção, diagnóstico ou tratamento, elementos que podem ser decisivos no controle dessas doenças (WATERS et al., 2020).

No Brasil, um instrumento que auxilia no planejamento da saúde, definindo e avaliando o impacto das intervenções, é o Sistema de Informação de Agravos de Notificação – SINAN, disponibilizado pelo Departamento de Informática do SUS (DATASUS). Este, engloba notificação e investigação de casos de doenças e agravos que constam da lista nacional de doenças de notificação compulsória, podendo, estados e munícipios incluir informações relevantes sobre problemas de saúde específicos de sua região. Assim, sua utilização efetiva permite uma identificação da realidade epidemiológica de determinada área geográfica. O livre acesso a todos os profissionais da área de saúde, corrobora com a democratização de acesso à informação, permitindo que estas sejam disponibilizadas para a comunidade.

Neste trabalho foi realizada uma análise exploratória dos dados relativos a processos de coinfecção de TB e HIV, advindos do SINAN, com o objetivo de propor métodos que facilitem a utilização dos dados desse sistema por profissionais da área de saúde, que não tenham formação técnica em computação. Considerando que tal aplicação é fortemente embasada no relacionamento de dados, optou-se por propor um mapeamento dos dados em bancos não convencionais, orientados a grafos, como o Neo4J. Em bancos de dados orientados a grafos, os dados são modelados em vértices e seus relacionamentos representados através de arestas. Assim, além de simplificar a modelagem, as aplicações desse tipo costumam ser mais rápidas, quando comparadas a aplicações tradicionais (utilizando bancos de dados relacionais).

Portanto, o mapeamento de dados disponíveis no SINAN para o Neo4J, permitiu uma visualização mais perceptível das correlações, possibilitando uma análise de múltiplos fatores, características de processos de coinfecção, potencializando as informações fornecidas pelas bases do SINAN e do Sistema de Tabulação de Dados disponibilizada pelo órgão, o TABNET. O presente trabalho traz alguns métodos computacionais que corroboram com tais estudos, sendo estes apresentados nos capítulos subsequentes.

1.1.2 Problema de Pesquisa

Uma maneira de analisar o processo de coinfecção é identificar ocorrências, mapear sua distribuição geográfica, estudar tratamentos e evolução dos casos, demandando a realização de uma análise multivariada. O profissional que efetua tais análises não necessariamente tem conhecimentos de computação, assim, a análise fica restrita aos recursos disponíveis na base de dados utilizada e, a visualização da informação gerada também depende destes. Assim, a pesquisa aqui descrita busca ampliar as formas de análise dos dados disponíveis nas bases públicas, objetivando simplificar e tornar a análise mais rápida e intuitiva.

1.1.2.1 Pergunta de Pesquisa

Como recursos computacionais podem ser utilizados para facilitar a análise multivariada de dados relativos à coinfecção de MTB e HIV, disponíveis em bases de dados de referência?

1.1.3 Justificativa

Nos países extremamente populosos e em desenvolvimento, como o Brasil, a tuberculose (TB) permanece uma séria questão de saúde pública. Na última década, a incidência de TB e sua taxa de mortalidade relacionada têm decaído firmemente no Brasil. No entanto, esta queda não é observada em pacientes portadores de HIV-1. TB é a mais frequente infecção oportunista e causa líder de morte entre portadores do vírus HIV em países de baixa e média-renda, particularmente entre indivíduos com HIV avançado (ESCADA et al., 2017). De 1.5 milhão de mortes atribuídas à TB em 2013, 24% foram entre pessoas portadoras do vírus HIV (MUKHATAYEVA et al., 2021); (MONTALES et al., 2015). A principal contribuição pretendida pelo presente trabalho é de propor um modelo que facilite a análise multivariada de dados para obtenção de informações relevantes, visando o tratamento e definições de políticas públicas para a coinfecção de HIV/Mycobacterium tuberculosis.

1.1.4 Objetivos

1.1.4.1 Objetivo Geral

Propor um modelo que facilite a análise multivariada de dados relativos à coinfecção de MTB e HIV, disponíveis em bases de dados de referência.

1.1.4.2 Objetivos Específicos

Apresentar um perfil epidemiológico atualizado da coinfecção entre tuberculose e o vírus da imunodeficiência, na região nordeste. Efetuar uma análise descritiva simples

deste perfil para o estado de Pernambuco. Mapear um cojunto de dados referentes a MTB e HIV disponíveis no SINAN, no Banco de Dados Neo4J. Ilustrar os benefícios do mapeamento através de exemplos de análises multivariadas.

1.1.5 Conteúdo do Documento

O presente Trabalho de Conclusão de Curso (TCC) está estruturado da seguinte forma: O Capítulo 1 é composto da apresentação do problema a ser tratado, o qual está relacionado a abordagens computacionais para estudos relacionados a processos de coinfecção, com ênfase naqueles que envolvem os patógenos HIV(Human Immunodeficiency Virus) e Mycobacterium tuberculosis(MTB). As seções que compõem o primeiro capítulo apresentam a Motivação para a realização de tais estudos, considerando aspectos mundiais e locais, assim como o Problema de Pesquisa abordado nesse trabalho, a Justificativa e Objetivos que guiaram o presente estudo. Adicionalmente, o Capítulo 1 descreve a forma como esse trabalho está estruturado. Considerando o aspecto multidisciplinar desse estudo, o Capítulo 2 traz a fundamentação teórica das principais áreas de conhecimento englobadas nesta pesquisa: Biologia Molecular e Ciência da Computação. Conceitos básicos, relacionados às diferentes áreas do conhecimento envolvidas, e apontamentos para estudos mais abrangentes são tratados neste capítulo. Uma revisão narrativa da literatura relacionada ao tema da pesquisa, realizada durante o desenvolvimento do presente trabalho, compõe o Capítulo 3. O Capítulo 4 consiste da apresentação do perfil epidemiológico da coinfecção da tuberculose e o vírus da imunodeficiência na região nordeste, e, uma análise descritiva simples deste perfil para o estado de Pernambuco, buscando descrever como o problema abordado nessa pesquisa se apresenta em nossa região. O estudo de caso para mapeamento dos dados do SINAN no banco de dados orientado a grafos, Neo4J é apresentado no Capítulo 5. As conclusões e trabalhos futuros advindos dessa pesquisa são discorridos no Capítulo 6.

2 Fundamentação Teórica

2.1 Biologia Molecular

Segundo (MEIDANIS; SETUBAL, 1997), ”os principais atores na química da vida são moléculas chamadas de proteínas e ácidos nucleicos. Grosseiramente falando, proteínas são responsáveis pelo que um ser vivo é e faz no sentido físico.”Toda a vida em nosso planeta depende, basicamente, de três tipos de macro-moléculas bioquímicas: **DNA**(*deoxyribonucleic acid*), **RNA**(*ribonucleic acid*) e **Proteínas**. ”Na biosfera atual, o DNA é a macromolécula-padrão para o armazenamento de longo prazo e para a transmissão da informação genética.”(COX; DOUDNA; O’DONNELL, 2012) Em outras palavras: todo ser vivo é formado por tais compostos. Assim, vegetais, animais (de qualquer espécie) e humanos não ficam fora dessa regra. DNA e RNA são tipos de **Ácidos Nucleicos**. ”(...) moléculas capazes de amarmazenar e transmitir a informação genética”(NELSON DAVID L.; COX, 2014). São moléculas formadas por outros compostos denominados **Nucleotídeos**. ”Uma molécula de nucleotídeo possui três componentes: uma base nitrogenada, um açúcar de cinco átomos de carbono (pentose) e um grupamento fosfato.”(NELSON DAVID L.; COX, 2014) - ver Figura 1. Os nucleotídeos que compõem os polímeros de DNA são desoxirribonucleotídeos. Estes, por sua vez, são representados por letras, são elas: A (Adenina), T (Timina), C (Citosina), G (Guanina), U (Uracila). Milhares ou, algumas vezes, milhões de ligações covalentes entre essas pequenas estruturas dão forma às Proteínas. Na Figura 2 é possível ver uma representação do DNA. Antes de dar origem às proteínas, a infromação contida no DNA é transcrita numa fita simples, o RNA. A este processo é dado o nome de transcrição. O processo em que o RNA dá origem às proteínas é chamado de tradução.

Existem sub-regiões (sub sequências) de DNA ou RNA, com uma estrutura específica, candidatos a áreas com alguma importância funcional - muitas vezes ainda desconhecida. Essas regiões são conhecidas como ‘Motif’. A presença de determinados motifs pode ser utilizada como base na classificação dos papéis de certas proteínas no organismo, por exemplo (PRZYTYCKA,).

Segundo (D’HAESELEER, 2006), os motifs são pequenas sub-regiões do código genético de um indivíduo, que costumam se repetir diversas vezes. Essas regiões podem servir ainda como sítio de ligação para os chamados ‘Fatores de Transcrição’ (FTs), quer seja para produzir funções reguladoras dos organismos ou para participar no processo de transcrição, no qual o DNA é replicado numa fita de RNA, para, em seguida, ser traduzido em uma proteína. Tal conceito, permeia o Dogma Central da Biologia.

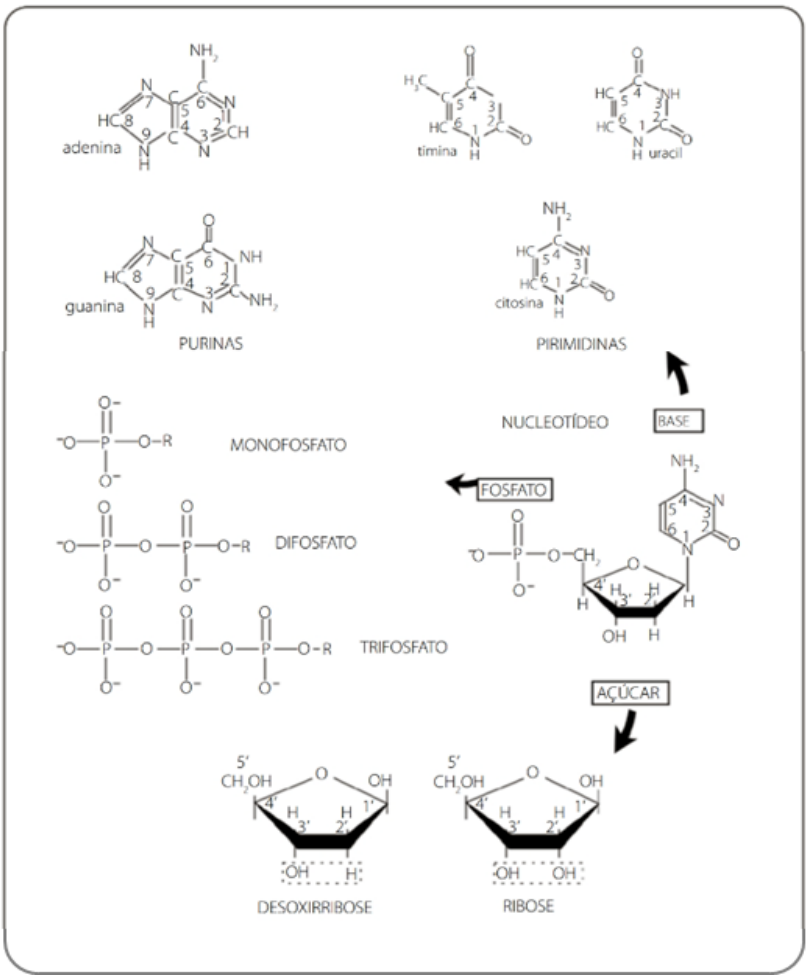


Figura 1 – Composição química de um nucleotídeo: base (purinas e pirimidinas), grupo fosfato (mono, di ou trifosfatado) e açúcar (pentose do tipo desoxirribose ou ribose).
(GIRARDI CAROLINA S.; SUBTIL, 2018)

Por serem estruturas muito pequenas, os motifs são de difícil detecção pelas ferramentas disponíveis atualmente. Assim, se faz necessário considerar as especificidades dos problemas para localização e análise dessas estruturas. Considerando, ainda, que os códons de aminoácido podem ser reconhecidos por diferentes trincas de DNA, conforme ilustrado na Figura 4, torna sua especificidade maior ainda.

Uma ilustração da representação de entrada e saída de uma abordagem computacional utilizando deep learning para a localização de motifs é exibida na Figura 8. Um outro elemento importante no problema da busca por motifs, também encontra-se ilustrado na Figura 5: FTs são responsáveis em levar a uma expressão coordenada entre genes, sendo que a interação entre fatores conduz a um padrão de resposta específico, indicando que são regulados pela expressão gênica. Modificações no padrão dessa expressão acarretam alterações de respostas celulares. Em outras palavras, os FTs são proteínas que auxiliam a tornar genes específicos em ”ligados”ou ”desligados”, ou seja, tais elementos ativam ou reprimem a transcrição de um determinado gene.

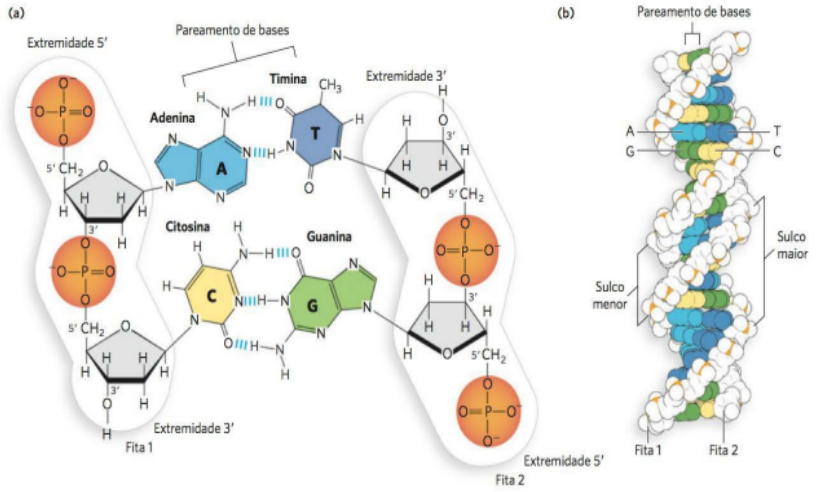


Figura 2 – Devido às suas propriedades estruturais, o DNA é muito adequado ao armazenamento de longo prazo da informação. (a) Os pares de bases G=C e A=T têm tamanho semelhante, o que lhes permite o empilhamento em qualquer sequência. O pareamento de bases complementares facilita a replicação e transmissão de uma geração para a próxima. (b) A estrutura de hélice dupla e empilhamento de bases conferem estabilidade. Os sulcos helicoidais maiores e menores na estrutura proporcionam o acesso da informação genética a uma ampla série de proteínas ligadoras de DNA. A estrutura uniforme do esqueleto do DNA permite a síntese de polímeros muito longos.
(COX; DOUDNA; O'DONNELL, 2012)

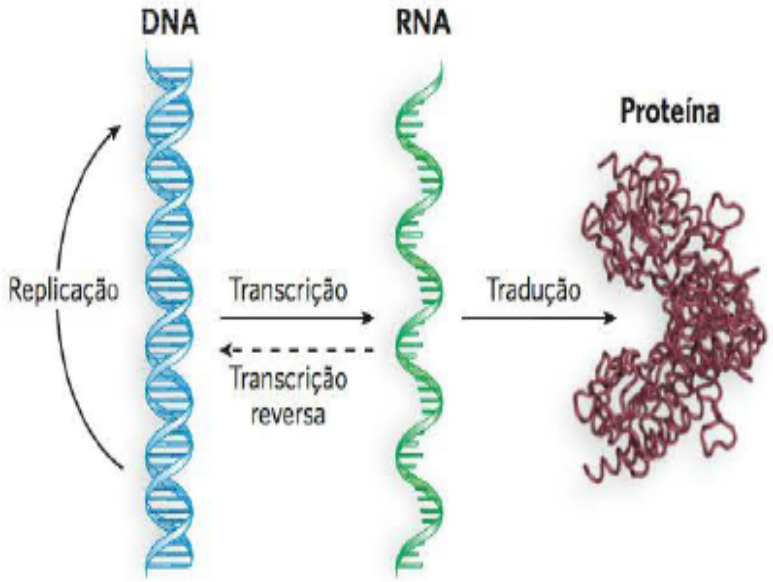


Figura 3 – **Dogma central do fluxo de informação: DNA->RNA->proteína.** A informação flui do DNA para o RNA por transcrição. A informação flui do RNA para a proteína por tradução. Em alguns casos, a informação também pode fluir de volta, do RNA para o DNA (transcrição reversa).
(COX; DOUDNA; O'DONNELL, 2012)

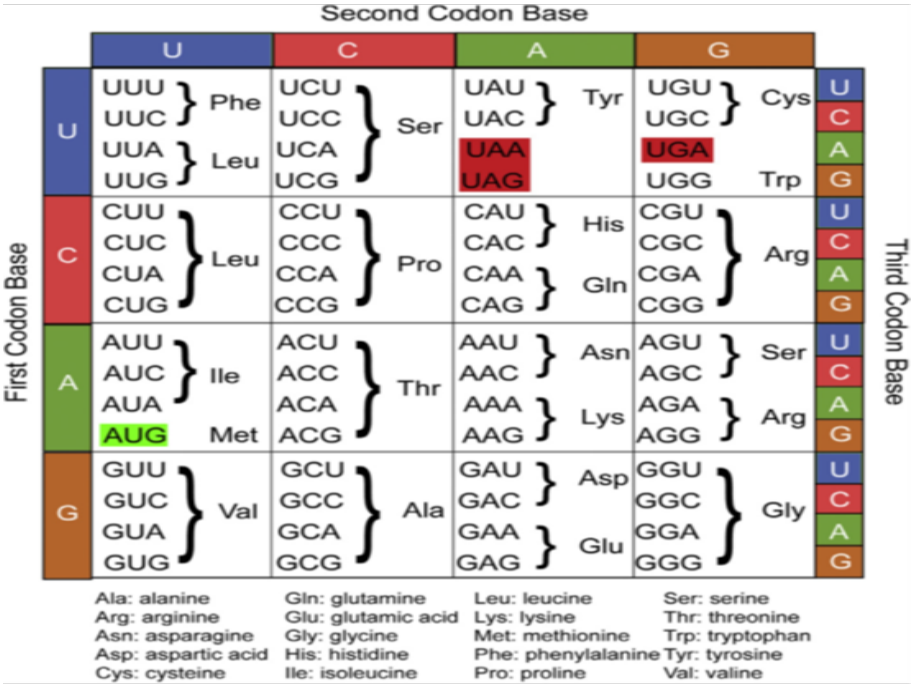


Figura 4 – Tabela do código genético. Códons de RNA e seus respectivos Aminoácidos. (LOUTEN, 2016)

Com base em um conjunto de informações, incluindo as dos fatores de transcrição, a célula irá expressar ou não um determinado gene.

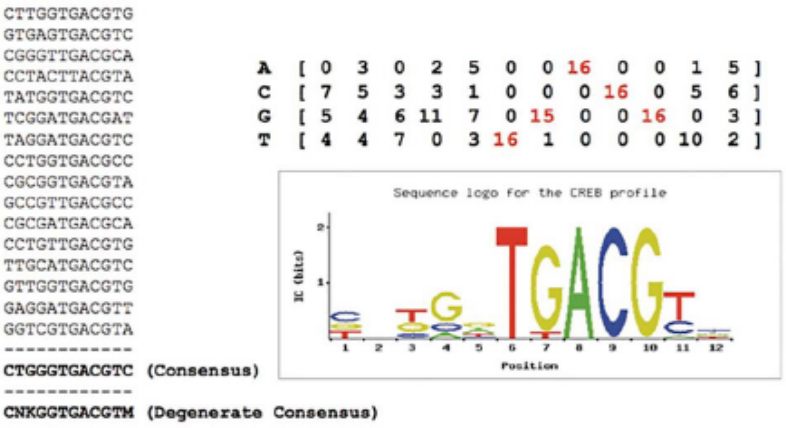


Figura 5 – Um conjunto de Sítios de ligação reconhecidos pelo mesmo fator de transcrição (FT). Representação de sequências degeneradas. Padrões de motifs e suas variações. Fonte:(HE et al., 2020)

2.2 Fundamentos da Ciência da Computação

2.2.1 Bancos de dados orientados a grafos

Segundo (ROBINSON; WEBBER; EIFREM, 2015), um grafo é uma coleção de vértices e arestas - ou, em linguagem informal, um conjunto de nós e os relacionamentos

que os conectam, Figura 6.

> Em contraste com bancos de dados relacionais, onde o desempenho de consulta intensiva se deteriora à medida que o conjunto de dados aumenta, com um grafo o desempenho do banco de dados tende a permanecer relativamente constante, mesmo quando o conjunto de dados cresce.(ROBINSON; WEBBER; EIFREM, 2015)

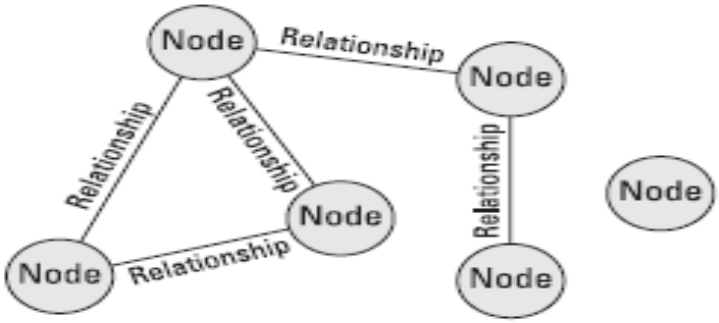


Figura 6 – Um exemplo simples de uma estrutura de grafos(nós e suas conexões). Fonte:(FRAME; BLUMENFELD, 2022)

2.2.2 Neo4J

De acordo com (WHAT..., a), o Neo4j é um banco de dados gráfico nativo NoSQL de código aberto que fornece um *back-end* transacional compatível com ACID. Dizer que o Neo4j é um banco de dados gráfico nativo significa que ele implementa um verdadeiro modelo gráfico até o nível de armazenamento, ou seja, os dados não são armazenados como uma ”abstração de gráfico”em cima de outra tecnologia.

Na Tabela 1 é possível ver o resultado de uma experimentação, em que o tempo de execução das consultas para um banco de dados relacional aumenta consideravelmente, com o aumento da quantidade de linhas retornadas, enquanto que para um banco de dados em grafo do Neo4J, esse tempo permanece praticamente constante.

Profundidade	Tempo de execução(s) no RDBMS	Tempo de execução(s) no Neo4J	Registros retornados
2	0,016	0,01	$\approx$ 2500
3	30,267	0,168	$\approx$ 110.000
4	1543,505	1,359	$\approx$ 600.000
5	Não finalizada	2,132	$\approx$ 800.000

Tabela 1 – Comparação da performance de um sistema de RDMBS - Banco de dados relacional - e as mesmas consultas, executadas no Neo4J. Fonte: adaptado de (ROBINSON; WEBBER; EIFREM, 2015)

2.2.3 Modelagem em grafos

Além dos nós e relacionamentos, uma das formas mais comuns de se modelar em grafos apresenta os nós, com a possibilidade de contar com propriedades para cada nó. Na Figura 7 é possível ver um grafo representando os relacionamentos de usuários no *Twitter*, onde temos a propriedade do *nome de usuário* em cada um deles e relacionamentos, representando quem 'Segue' quem. *Billy*, por exemplo, segue *Harry*, que, por sua vez, também o segue.

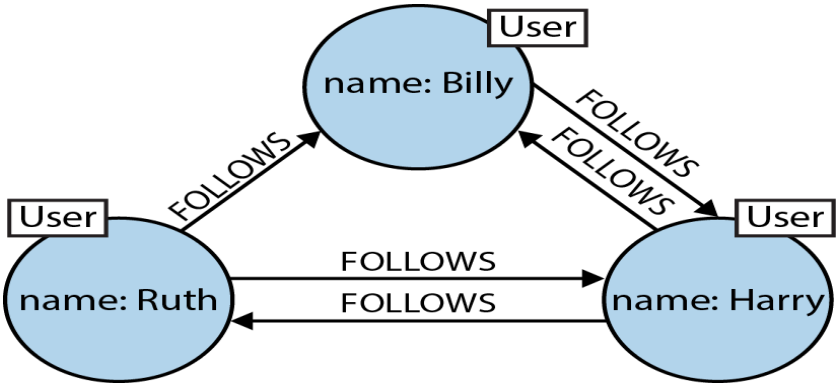


Figura 7 – Um exemplo de grafo entre Usuários do Twitter e relacionamentos (Segue), entre eles. Fonte:(ROBINSON; WEBBER; EIFREM, 2015)

Os relacionamentos 'Seguir'(*'Follows'*) são direcionados, ou seja: partem de um nó, em direção a outro nó no grafo. Sendo assim, é necessário dois relacionamentos para representar que Billy e Harry se seguem mutuamente. A Figura 8 apresenta um grafo baseado no grafo da Figura 7, tendo a vantagem de que é possível, além de saber, percorrendo caminhos no grafo(*'Paths'*), qual dos usuários é amigo ou não de outro, também descobrir quais as mensagens (ou *Tweets*) um usuário(nesse caso *Ruth*) acabou de enviar para a plataforma e suas mensagens anteriores.

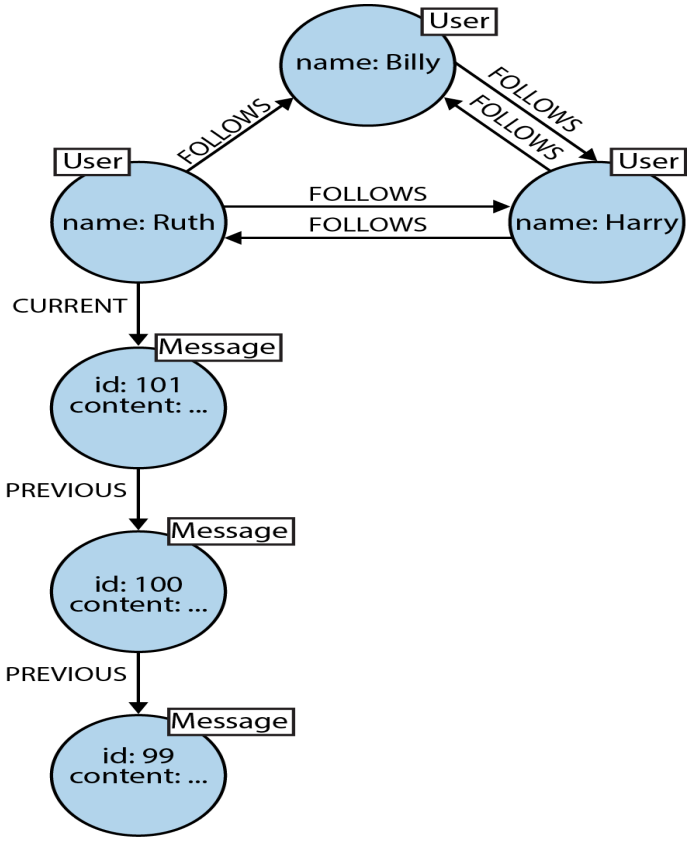


Figura 8 – Base do Twitter, com relacionamentos(Segue) entre Usuários e mensagens enviadas para a plataforma, onde é possível caminhar no grafo e montar um histórico, por exemplo, dos *tweets* enviados por *Ruth* ou por qualquer outro usuário que possua *tweets*. Fonte:(ROBINSON; WEBBER; EIFREM, 2015)

Dessa forma é possível ilustrar facilmente a base do Twitter, através de uma modelagem em grafos.

2.2.4 Cypher como uma linguagem de buscas em grafos

Para fazer consultas ao grafo é necessário uma linguagem de busca. No caso do Neo4J, uma das mais utilizadas é o *Cypher*.

Segundo (ROBINSON; WEBBER; EIFREM, 2015), o Cypher é:

> uma linguagem projetada para ser facilmente lida e compreendida por desenvolvedores, profissionais de banco de dados, e partes interessadas do negócio(*stakeholders*). Sua facilidade de uso deriva do fato de estar em acordo com a forma como intuitivamente descrevemos grafos usando diagramas.

Na Figura 9, por exemplo, temos um grafo de pessoas que conhecem pessoas. A base de dados à qual esse grafo pertence pode ser muito maior, contanto com uma rede bastante ampla de nós. Porém, a busca em Cypher que obtém esse subgrafo

está representada no Código 2.1. É possível notar que a consultar em Cypher é uma string em padrão ASCII, que conta com símbolos como setas ('<-' e '->') para indicar as arestas de chegada em cada nó e que, para indicar as arestas de saída é utilizado apenas o traço('-'). Um relacionamento é representado na forma [:NOME_DO_RELACIONAMENTO] e um nó (nome_variavel_nó).

No código 2.1, é possível notar como o Cypher é uma linguagem de consulta simples e direta, tornando o seu aprendizado bastante facilitado.

```
1 (emil)<-[:KNOWS]-(jim)-[:KNOWS]-(ian)->[:KNOWS]->(emil)
```

Código 2.1 – Código Cypher que retorna um conjunto de pessoas que se conhecem entre si. Fonte: adaptado de (ROBINSON; WEBBER; EIFREM, 2015)

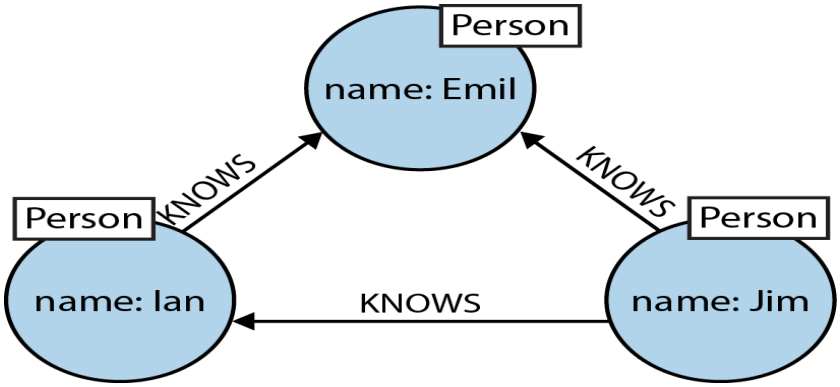


Figura 9 – Grafo que representa o resultado da busca na Figura 2.1. Fonte:(ROBINSON; WEBBER; EIFREM, 2015)

2.2.5 A utilização de Bancos NoSQL para Dados Biológicos

Tradicionalmente, os bancos de dados biológicos utilizam o modelo relacional, ou seja, o banco de dados é um conjunto de relações, cuja linguagem padrão para realização de *queries* é a SQL (do inglês *Structured Query Language*). De simples utilização, esta linguagem é utilizada, com poucas variações, em diferentes bancos de dados relacionais, dentre os quais podemos citar o **MySQL** e o **PostgreSQL**, os quais possuem versões gratuitas e de código aberto.

Os bancos de dados NoSQL (do inglês *Not Only SQL*), por sua vez, utilizam um padrão diferente de armazenamento ao compararmos com o SQL, propiciando uma melhor performance no gerenciamento de dados, além de tornar as buscas mais simples e rápidas. Como exemplos de bancos NoSQL, temos o **MongoDB**, de código aberto, e aceito por diferentes sistemas operacionais, este tem como carcterística a modelagem de dados orientada a documentos. Assim, as informações relevantes são armazenadas em documentos *JSON*, e são manipuladas através de técnicas de

agrupamento e filtragem. Um outro exemplo de banco NoSQL, utilizado nos estudos aqui reportado, é o **Neo4J**, cuja modelagem se baseia em grafos, sendo a informação armazenada em nós e as relações indicadas por arestas. Também de código aberto, este banco vem sendo bastante utilizado na análise de dados de redes sociais (NIKAM; BHOITE; SHENOY, 2020), bem como no reconhecimento de padrões (GONG et al., 2018).

Conforme destacado anteriormente, análise de bancos biológicos envolve, por vezes, buscar um número considerável de correlações entre seus elementos, em um grande volume de dados. Tais características motivam estudos que verifiquem a viabilidade de utilizar bancos NoSQL visando eficiência e simplicidade nessas buscas.

Um estudo realizado por Stothers e Nguyen (STOTHERS; NGUYEN, 2020), efetuou uma comparação entre a utilização dos tradicionais bancos SQL, e um banco de dados orientado a grafos, o Neo4j, na análise de dados biológicos. O estudo de caso foi realizado com dados de 100 pacientes, que permaneceram em unidades de terapia intensiva do Beth Israel Deaconess Medical Center, diponíveis em um banco de dados aberto, onde estes não são identificados. O trabalho destaca as vantagens dos bancos de dados relacionais, dentre as quais destacamos as já citadas segurança de dados e flexibilidade de consulta, além de serem suportados por diversas ferramentas que permitem uma análise eficaz de dados. Ao utilizar o banco de dados NoSQL, Neo4J, os orientados a grafos, os autores buscam identificar ganhos ao utilizá-lo em bancos de dados biológicos. O estudo de caso buscou comparar os bancos envolvidos em termos de complexidade, eficiência de comando, facilidade de importação e análise de dados, objetivando investigar qual deles seria mais adequado para pesquisadores que investigam dados clínicos de pacientes, buscando garantir ambientes computacionais semelhantes para ambos os experimentos. Um dos resultados observados foi que, após a implementação do banco de dados, as operações foram substancialmente mais rápidas no Neo4j do que no PostgreSQL. Mesmo com um tempo de implementação mais rápido para o banco de dados relacional, os autores destacam que Neo4j apresentou vantagens em relação à simplicidade operacional, principalmente no uso de relacionamentos para evitar a junção de tabelas. A conclusão do estudo recomenda o Neo4j para bancos de dados clínicos, podendo futuramente ser testados em outros tipos de dados biológicos.

Os bancos de dados NoSQL, Neo4J e MongoDB, foram utilizados na análise de dados clínicos, diponíveis no Sistema de Informação de Agravos de Notificação (SINAN) (CARVALHO, 2019), (SAMPAIO, 2021). Em (CARVALHO, 2019), o Neo4J é utilizado para entender as particularidades das trajetórias dos pacientes, diagnosticados com tuberculose, no SUS, do início ao encerramento do tratamento. Nesse estudo é realizada a Extração, Transformação e Carregamento (ETL) dos dados do SINAN,

relativos à cidade do Rio de Janeiro, para bancos de dados relacionais (RDMBS) e para o banco de dados em grafos.

O presente trabalho também propõe um mapeamento de dados do SINAN para o Banco Neo4J, buscando correlacionar dados de relativos a coinfecção de TB/HIV, no estado de Pernambuco. A metodologia utilizada e resultados obtidos são discorridos no capítulo 5.

3 Trabalhos Relacionados

3.1 Considerações iniciais

Neste capítulo apresentamos uma revisão exploratória, descrevendo os principais trabalhos científicos utilizados como base para a construção de nossa pesquisa. Ao final deste capítulo, discorremos sobre a influência de tais trabalhos na determinação das diretrizes e métodos usados no desenvolvimento do estudo aqui reportado.

3.2 Análise de coinfecção por TB/HIV

Nos últimos cinco anos, pesquisas relacionadas ao tema coinfecção aumentaram de forma significativa, conforme podemos observar no gráfico de número de publicações, ao lançarmos o termo *Coinfection* na PubMed, motor de busca mantido pelo *National Center for Biotechnology Information* (NCBI), numa base com mais de 35 milhões de citações e resumos da literatura biomédica, Figura 10. Os processos de infecção por múltiplos patógenos, com alcance mundial, envolvem doenças tais como: HIV, malária, hepatite, dengue e, mais recentemente, COVID-19 (LANSBURY et al., 2020). As possíveis relações benéficas mútuas entre os patógenos envolvidos, auxilia na compreensão da frequência de ocorrência maior que aquelas de infecção por um único patógeno (GRIFFITHS et al., 2011), (PIETRO et al., 2018),(PICANÇO-JUNIOR et al., 2022), (WATERS et al., 2020).

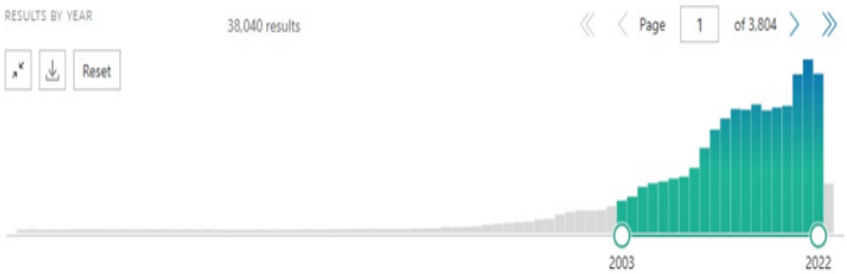


Figura 10 – Gráfico representando as publicações na plataforma pubmed para o termo coinfecção até o ano de 2022. (Fonte: PUBMED)

Em (MUKHATAYEVA et al., 2021), os autores realizam uma análise descritiva de dados, obtidos a partir de um conjunto de 500 pacientes convivendo com o HIV, relacionados à coinfecção de vários patógenos com o HIV, dentre os quais podemos citar: Hepatite C(HCV)/HIV, Hepatite B(HBV)/HIV, Tuberculose(TB)/HIV e Doenças Sexualmente Transmissíveis(STI)/HIV - incluindo gonorréia, sífilis e tricomoníase. Os autores

traçam um panorama geral sobre as coinfecções, e concluem, propondo soluções para as questões levantadas, como:

> A prevalência da coinfecção TB-HIV é maior nos países africanos e asiáticos. Além disso, a incidência de coinfecção TB-HIV está aumentando no Cazaquistão. Implementação de protocolos de vacinação vigilantes contra TB e HIV é uma forma eficaz de prevenir essas coinfecções entre PVHS.(MUKHATAYEVA et al., 2021)

Assim, dentre os diversos processos de coinfecção, as ocorrências de HIV(*Human Immunodeficiency Virus*) e a MTB (*Mycobacterium tuberculosis(MTB)* - bactéria causadora da tuberculose, estão entre os estudos de monitoramento da OMS (Organização Mundial de Saúde). O relatório de 2020 da OMS (GLOBAL…, 2020), indica, em 2019, a tuberculose (TB) como a principal causa de morte entre pessoas vivendo com HIV (PVHIV), responsável por cerca de 30% das 690.000 mortes relacionadas à AIDS no mundo. Adicionalmente, de acordo com o Relatório Global de TB de 2020, as PVHIV têm, em média, 18 vezes mais chances de desenvolver a doença ativa da TB do que as pessoas sem HIV.

Diante deste panorama, o estudo descrito no presente trabalho, pesquisou formas de contribuição da computação para mapear, quantificar e relacionar dados referentes a processos de coinfecção de HIV/TB, realizando estudos de caso em bases de dados nacionais.

No Brasil, uma base de dados que auxilia no planejamento da saúde é o Sistema de Informação de Agravos de Notificação – SINAN, o qual é disponibilizado pelo Departamento de Informática do SUS (DATASUS). O SINAN engloba notificação e investigação de casos de doenças e agravos que constam da lista nacional de doenças de notificação compulsória, incluindo dados referentes a coinfecção de TB/HIV. De livre acesso para os profissionais da área de saúde, os dados do SINAN contribuem para identificar e monitorar a realidade epidemiológica de determinada área geográfica, auxiliando na definição de políticas públicas (ROCHA et al., 2020).

O artigo intitulado ”Análise Epidemiológica da Coinfecção Tuberculose/HIV”(OLIVEIRA et al., 2018), com o objetivo de analisar o perfil epidemiológico da coinfecção por tuberculose e vírus da imunodeficiência, realizam um estudo descritivo, com base em dados obtidos do Sistema de Informação de Notificação de Agravos, para o estado do Piauí, com dados relativos a atendimentos a pacientes, no período de 2007 a 2016. O estudo se aprofunda nos dados, apresentando graficamente, na forma de tabelas, informações do perfil da coinfecção HIV/TB, como: Faixa etária dos indivíduos, Perfil étnico, Forma clínica da Tuberculose (Pulmonar ou Extra-pulmonar), e situações de encerramento, ou seja, se houve cura, abandono do tratamento ou morte(e se a causa

foi a Tuberculose ou outras causas). Esse tipo de análise traz informações mais específicas relativas a uma aréa geográfica de interesse, sendo uma forma de avaliar a evolução de casos naquela região, usando variáveis de interesse.

A análise de bancos biológicos envolve, por vezes, buscar um número considerável de correlações entre seus elementos, além de apresentar um grande volume de dados. Os profisionais que necessitam das informações geradas a partir desse tipo de análise não necessariamente têm conhecimentos na área de Computação, assim usam as ferramentas e formas de entrada e saída fornecidas pelas bases utilizadas, as quais, em sua maioria, são relacionais, também referenciados como convencionais. Porém, há outros modelos de bancos que são mais adequados quando se lida com um grande número de correlações entre os dados, por exemplo, os **Bancos de Dados Orientados a Grafos**. Este tipo de banco de dados não convencional, aumenta a velocidade nas consultas, além de torná-las mais simples e intuitivas, podendo gerar *insights* mais significativos, em face aos banco de dados relacionais, onde a complexidade computacional desse tipo de consulta é muito maior.

Recentemente, Stothers e Nguyen (STOTHERS; NGUYEN, 2020) realizaram um estudo comparativo entre a utilização de bancos SQL, tradicionalmente utilizados para dados biológicos e um banco de dados orientado a grafos, o Neo4j, comumente utilizado para análise de redes sociais e transportes. Os autores destacam que os bancos de dados relacionais apresentam vantagens tais como segurança de dados e flexibilidade de consulta, além de serem suportados por diversas ferramentas que permitem uma análise eficaz de dados. A partir do estudo realiazado, os autores concluiram que embora a familiaridade, e a possibilidade de aplicar outros recursos contem a favor de bancos SQL, o Neo4j oferece vantagens operacionais significativas, em relação aos formatos possíveis de apresentar os resultados da consulta, além de favorecer a interpretação de dados,e tornar a experiência do usuário mais intuitiva, tendo se mostrado vantajoso para a implementação de bancos de dados clínicos.

Um trabalho reportado pela UFRJ, em 2021, também propõe o uso do Neo4J para análise de dados clínicos, aqueles diponíveis no Sistema de Informação de Agravos de Notificação (SINAN) (CARVALHO, 2019). A proposta é entender as particularidades das trajetórias dos pacientes, diagnosticados com tuberculose, no SUS, do início ao encerramento do tratamento. Para isto, o estudo realizou a Extração, Transformação e Carregamento (ETL) dos dados do SINAN, relativos à cidade do Rio de Janeiro, para bancos de dados relacionais (RDMBS) e para o banco de dados em grafos (Neo4J). Utilizando dados locais, contando com informações como: bairro de residência dos pacientes e da localização das unidades de saúde (tais quais Hospitais e Unidades Básicas de Saúde) foi possível executar um mapeamento geográfico e da sequência de consultas dos pacientes, além da finalização dos tratamentos destes, sendo possível

aplicar a metodologia desenvolvida para outro eventos na área de saúde, bem como em outras áreas. Um outro estudo realizado na UFRJ (SAMPAIO, 2021), propôs uma ETL de dados de saúde, anônimos, disponibilizados pelo SUS, em uma banco NoSQL, neste caso, orientado a documento, o MongoDB. O objetivo foi produzir ambiente analítico que permita a aplicação de aprendizagem de máquina e mineração de dados, visando também, o melhoramento da saúde pública. No trabalho é processado uma grande massa de dados, utilizando o MongoDB, e no final é proposta uma base com consultas que podem ser incorporadas a ferramentas como Tableau ou D3.js, com a finalidade de geração de visualizações estatísticas.

3.3 Considerações finais

Os trabalhos descritos nesse capítulo, foram a base para a definição dos direcionamentos dos estudos e resultados aqui reportados. Abordar um tema atual, coinfecção de patógenos, com ênfase em HIV/TB, cuja a incidência é signiticativa em diversas partes do mundo. A partir da definição do tema, optamos por realizar análise descritiva de dados e processos de transformação de dados para Extração, Tratamento e Carga (ETL), em bancos de dados não relacionais, mais especificamente, bancos de dados orientados a grafos, utilizando dados mais atuais, referentes à região nordeste do Brasil e ao estado de Pernambuco.

4 Perfil epidemiológico da coinfecção de MTB e HIV - Aspectos regionais

4.1 Considerações Iniciais

No presente capítulo, é apresentado o perfil epidemiológico de coinfecção da tuberculose e o vírus da imunodeficiência, fornecendo indicativos sobre como o problema tratado no presente trabalho está distribuído na região nordeste do Brasil, através de uma análise descritiva simples.

4.2 Ferramentas de *Software* utilizadas

4.2.1 R

De acordo com (WHAT…, b), R é uma linguagem e ambiente para computação estatística que fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, …) e técnicas gráficas, e é altamente extensível, significando que conta com uma ampla comunidade que desenvolve os mais diversos pacotes de extensão para as mais diversas aplicações.

R está disponível como Software Livre sob os termos da Licença Pública Geral GNU da Free Software Foundation em forma de código-fonte. Ele compila e roda em uma ampla variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS.

4.2.2 RStudio

De acordo com (RSTUDIO,), RStudio é um ambiente de desenvolvimento integrado (IDE) projetado para suportar vários idiomas, incluindo R e Python. Ele inclui um console, editor de realce de sintaxe que oferece suporte à execução direta de código e uma variedade de ferramentas robustas para plotagem, visualização de histórico, depuração e gerenciamento de seu espaço de trabalho.

O RStudio está disponível em código aberto e edições comerciais e é executado no desktop (Windows 10+, macOS 11+ e Linux) ou em um navegador conectado ao RStudio Server ou Posit Workbench. O desenvolvimento das variantes do RStudio IDE de código aberto e comercial é suportado pela Posit Software, PBC (anteriormente chamado RStudio, PBC).

4.2.3 Tabnet

O aplicativo TABNET é um tabulador genérico de domínio público que permite organizar dados de forma rápida, conforme a consulta que se deseja tabular.

Foi desenvolvido pelo DATASUS para gerar informações das bases de dados do Sistema Único de Saúde – SUS.

No Tabnet estão disponíveis informações tais como: Mortalidade; Nascidos Vivos; Informações Epidemiológicas; Morbidade; Indicadores de Saúde; Assistência à Saúde; Informações Demográficas e Socioeconômicas; Inquéritos e Pesquisas e Cadastros da Rede Assistencial,

4.3 Introdução

De acordo com o relatório da Organização Mundial da Saúde (OMS) sobre Tuberculose, publicado em 2022 <https://www.who.int/publications/i/item/9789240061729>, houve um aumento do número de casos de TB e tuberculose resistente (TBDR), bem como um crescimento no número de óbitos devido a esta doença durante a pandemia de Covid-19. O relatório da OMS estima que 10,6 milhões de pessoas contraíram TB em 2021, significando um aumento de 3,6% na taxa de incidência. Ainda, segundo a OMS, 1,6 milhões de pessoas morreram devido a tuberculose, sendo 8,6% (187.000) HIV positivas.

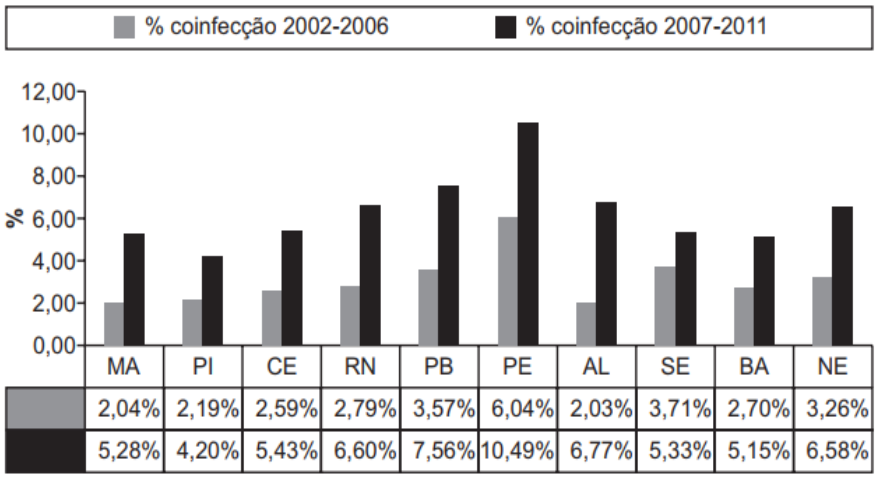
Observou-se que a tuberculose é a segunda principal causa de mortes por doenças infecciosas, precedida apenas pela Covid-19. A infecção por HIV desponta como um dos cinco fatores de risco para a tuberculose, ao lado da desnutrição, alcoolismo, tabagismo e diabetes, sendo a coinfecção TB/HIV, segundo (CAVALIN et al., 2020), principal causa de morte deste grupo de pessoas.

4.4 Estudos prévios para estados do Nordeste com pacientes vítimas de coinfecção por HIV/TB

Segundo estudos prévios, efetuados por (BARBOSA; COSTA, 2014), sobre coinfecção dos patógenos HIV(Human Immunodeficiency Virus) e Mycobacterium tuberculosis(MTB) no Nordeste do Brasil, houve um aumento significativo em todos os estados da região, quando comparados os períodos entre 2002 - 2006 e entre 2007 - 2011.

Como pode ser observado na Figura 11, o percentual de coinfecção ultrapassou(do primeiro para o segundo período) o dobro do valor nos estados do Maranhão (2,04% - 5,28%), do Ceará (2,59% - 5,43%), do Rio Grande do Norte (2,79% - 6,60%) e

da Paraíba (3,57% - 7,56%) , e triplicou no estado de Alagoas (de 2,03% para 6,77%). No estado de Pernambuco, houve um aumento de 6,04%, para 10,49%.



	MA	PI	CE	RN	PB	PE	AL	SE	BA	NE
	2,04%	2,19%	2,59%	2,79%	3,57%	6,04%	2,03%	3,71%	2,70%	3,26%
	5,28%	4,20%	5,43%	6,60%	7,56%	10,49%	6,77%	5,33%	5,15%	6,58%

Fonte: Sistema de Informação de Agravos de Notificação. Datasus, 2013.

Figura 11 – Comparação entre os percentuais de coinfecção tuberculose/HIV nos estados do Nordeste do Brasil entre os períodos de 2002 a 2006 e 2007 a 2011. Fonte: (BARBOSA; COSTA, 2014)

4.5 Perfil epidemiológico do estado de Pernambuco para HIV/TB de 2012 a 2021

Baseado em (OLIVEIRA et al., 2018), que desenvolveu perfil epidemiológico do estado do Piauí, para o período de 2007 a 2016, com os dados extraídos do SINAN, foi efetuado um levantamento dos dados de casos de coinfecção por Tuberculose e HIV para o estado de Pernambuco e efetuado processo semelhante, para o período de 2012 a 2021, com a mesma fonte dos dados.

Os passos exibidos na Figura 12 foram executados para a construção do perfil.

Figura 12 – Sequência de passos para a construção do Perfil Epidemiológico. (Fonte: o autor)

A etapa de extração dos dados é feita acessando a página web disponível em: TABNET.

A Figura 14 mostra a primeira etapa do processo, que é rolar a página para baixo, até chegar ao menu de opções do Tabnet, onde existem as opções de bases de dados disponíveis que podem ser acessadas. Depois é acessada a opção Epidemiológicas e Morbidade Casos de Tuberculose - Desde 2001(SINAN), como pode ser visto na Figura 14. A etapa seguinde é a seleção da abrangência geográfica - Unidade Federativa(UF), onde é selecionado o estado de Pernambuco. A Figura 15 ilustra essa etapa. E, finalmente, chega-se à etapa dos filtros de seleção dos dados do Tabnet (Figura 16).

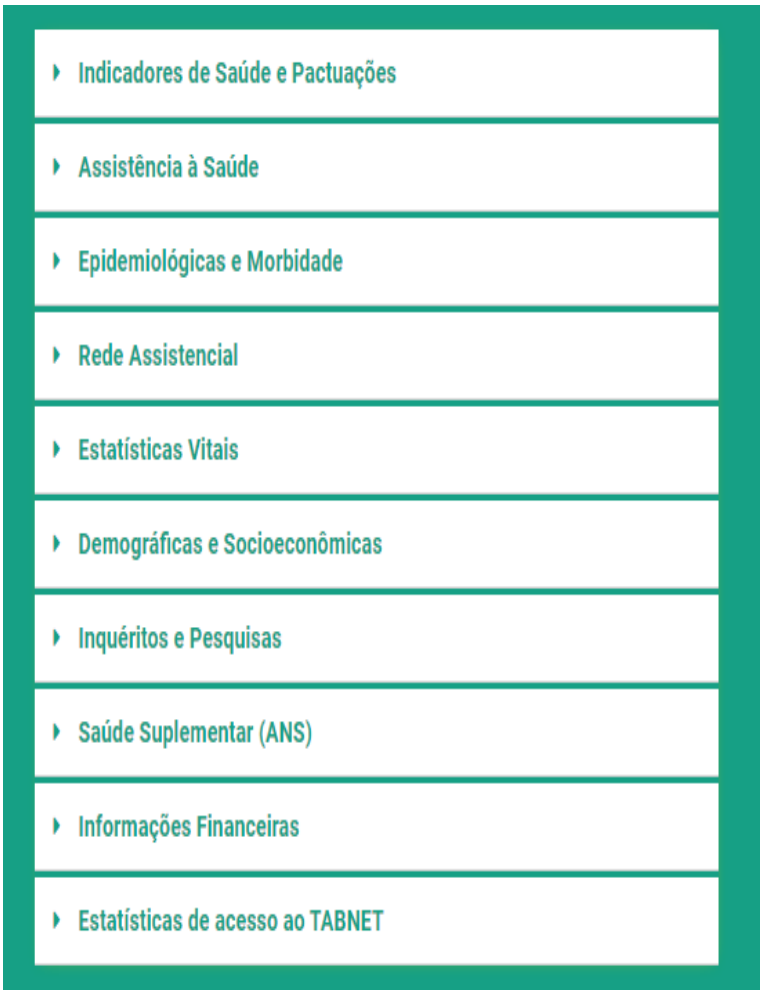


Figura 13 – Página do Tabnet - escolha da opção 'Tuberculose'. (Fonte: TABNET)

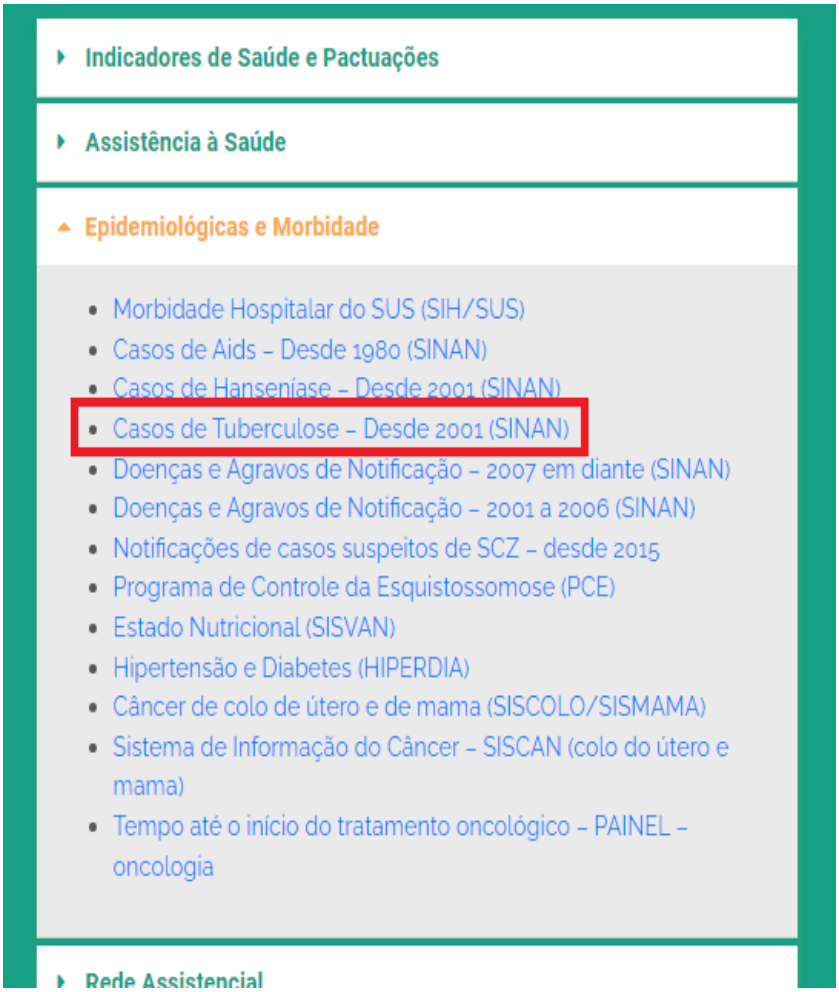


Figura 14 – Página do Tabnet - escolha da opção 'Tuberculose'. (Fonte: TABNET)

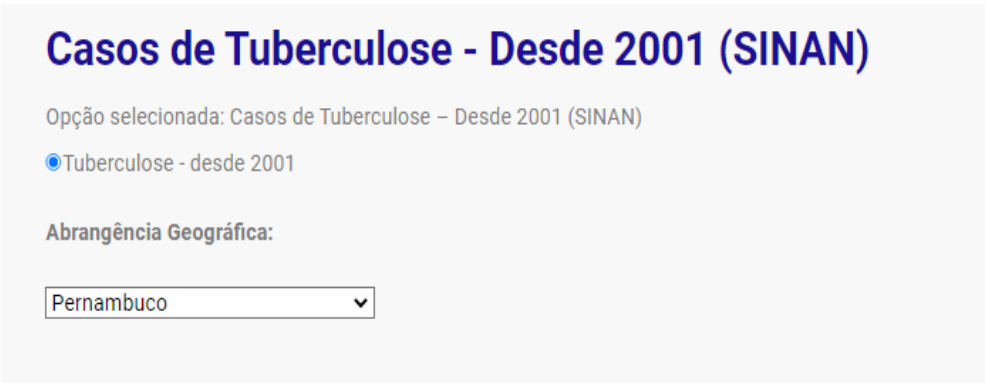


Figura 15 – Página do Tabnet - escolha da abrangência geográfica(UF). (Fonte: TABNET)

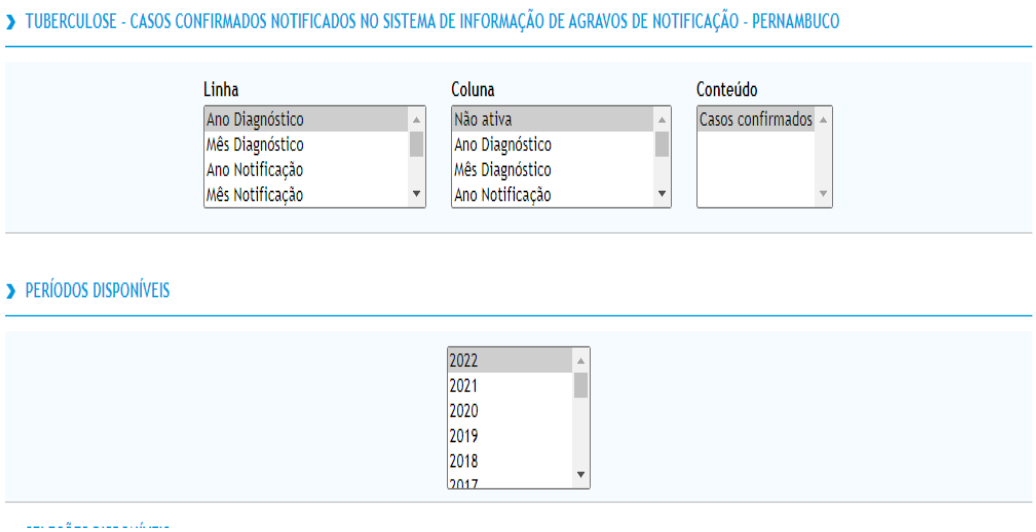


Figura 16 – Página do Tabnet - seleção dos filtros para obtenção dos dados desejados. (Fonte: TABNET)

Na Figura 19, é possível observar uma tabela que representa a visão geral dos casos de Tuberculose/HIV para o estado de Pernambuco. Essa tabela pode ser obtida nos filtros da Figura 16, fazendo a seguinte configuração: no filtro 'Linha', seleciona-se HIV, no filtro 'Coluna', seleciona-se Ano Diagnóstico e no filtro dos períodos disponíveis, seleciona-se, presionando-se a tecla *shift*, os anos de 2012 a 2022. Na Figura 17 é exibido como fica o filtro.

A etapa seguinte é rolar a página até o final e escolher a opção desejada de obtenção dos dados. As opções são 'Tabela com bordas', Texto pré-formatado e colunas separadas por vírgulas. Como a tabela gerada conta com muitas colunas, o Tabnet as exibiu de forma que ficaria muitro extensa para ser exibida em um documento A4, dessa maneira, resolveu-se carregar esses dados no formato .csv na ferramenta RStudio e fazer os tratamentos necessários via script R.

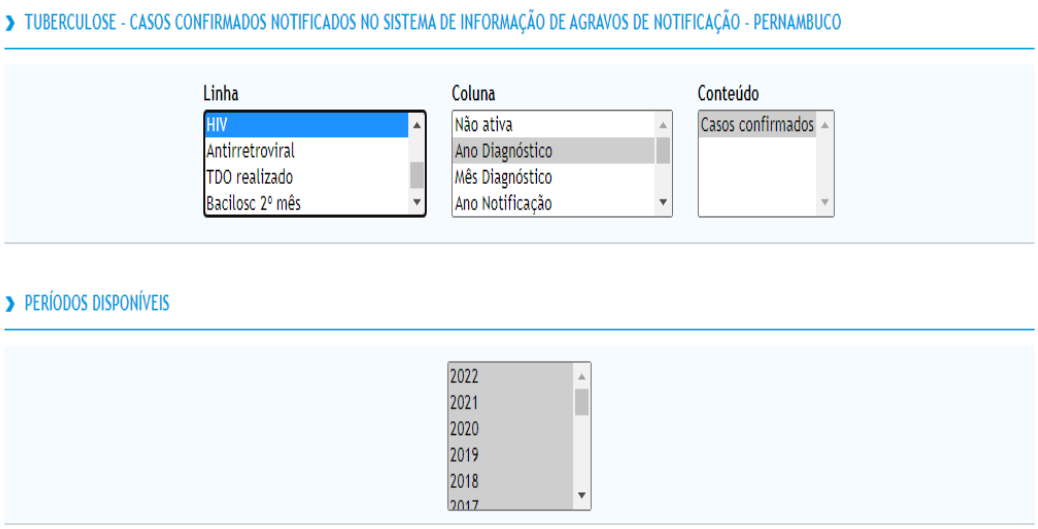


Figura 17 – Página do Tabnet - seleção dos filtros para obtenção dos dados para a tabela da Figura 19. (Fonte: TABNET)

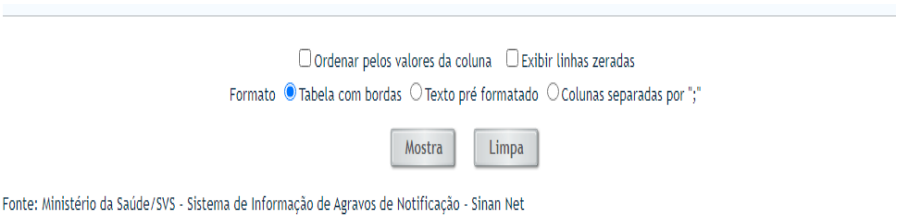


Figura 18 – Página do Tabnet - botão 'Mostrar'. (Fonte: TABNET)

```r
# carrega o banco de dados
df <- read.csv2('dados-pernambuco-tuberculose-hiv-2012-2022.csv', sep = ';', dec = ',')
# remove a coluna ('2022') do dataframe, onde houve subnotificação dos casos
df <- df[,-12]
View(df)
df <- rename(df,
     '2012'='X2012','2013'='X2013','2014'='X2014','2015'='X2015','2016'='X2016','2017'='X2017','2018'='X2018','2019'='X2019','2020'='X2020','2021'='X2021')
#criar tabela
df %>%
  gt()
```

Código 4.1 – Código em R para criação da tabela de casos de Tuberculose/HIV em Pernambuco no período de 2012 a 2021. (Fonte: o autor)

No período de 2012 a 2021, foram contabilizados 60.757, casos de Tuberculose para o estado de Pernambuco, dos quais 7.123 (11,72%), eram casos confirmados HIV positivo.

HIV	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Ign/Branco	1	-	2	2	-	-	-	-	1	1	9
Positivo	730	695	692	649	693	682	674	671	586	684	7123
Negativo	2627	2456	2756	3266	3242	3582	3790	3880	3365	3485	34217
Em andamento	511	611	268	146	109	104	81	88	97	270	2632
Não realizado	1833	1664	1730	1592	1480	1617	1409	1489	1488	1612	16776
Total	5702	5426	5448	5655	5524	5985	5954	6128	5537	6052	60757

Figura 19 – Visão geral dos casos, no período de 2012 a 2021, dos casos de coinfecção de TB/HIV, no estado de Pernambuco. (Fonte:Sistema de Informação de Agravos de Notificação - SINAN)

Ainda conforme a Figura 19, analisando apenas os casos positivos no período de 2012 a 2021, é possível observar uma média de 647 casos.

A partir do Código 4.2 foi gerado o gráfico da Figura 20, que representa a evolução dos casos de Tuberculose com teste de HIV Positivo, para o período de 2012 a 2021, no estado de Pernambuco.

```
#gráfico de linha
#pega a quantidade de casos com hiv positivo no ano de 2022
hiv_positivo_2022_pos <- df[df$HIV == 'Positivo', c('2012','2013','2014','2015','2016','2017','2018','2019','2020','2021')]

hiv_positivo_2022_pos <- t(hiv_positivo_2022_pos)
colnames(hiv_positivo_2022_pos) <- "Casos hiv positivo"
View(hiv_positivo_2022_pos)
write.csv2(hiv_positivo_2022_pos, "2012-2022-hiv-tb2.csv", row.names=TRUE)

X2012_2022_hiv_tb <- read.csv("2012-2022-hiv-tb2.csv", sep = ';', dec = ',')

colnames(X2012_2021_hiv_tb) <- c("Ano","Casos_hiv_positivo")

Ano <- factor(X2012_2021_hiv_tb$Ano)
Numero_de_casos <- factor(X2012_2021_hiv_tb$Casos_hiv_positivo)

ggplot(X2012_2021_hiv_tb, aes(x = Ano, y = Numero_de_casos, group=1)) +
  geom_line()
```

Código 4.2 – Código em R para criação do gráfico de linha dos casos de Tuberculose/HIV Positivo, em Pernambuco no período de 2012 a 2021. (Fonte: o autor)

Observando o período de 2007 a 2011, é possível notar um aumento nos casos positivos, até atingir o pico de 730, em 2012. Devido a uma subnotificação de casos para o ano de 2022, encontrada nos dados, resolveu-se não inlcuí-los no estudo.

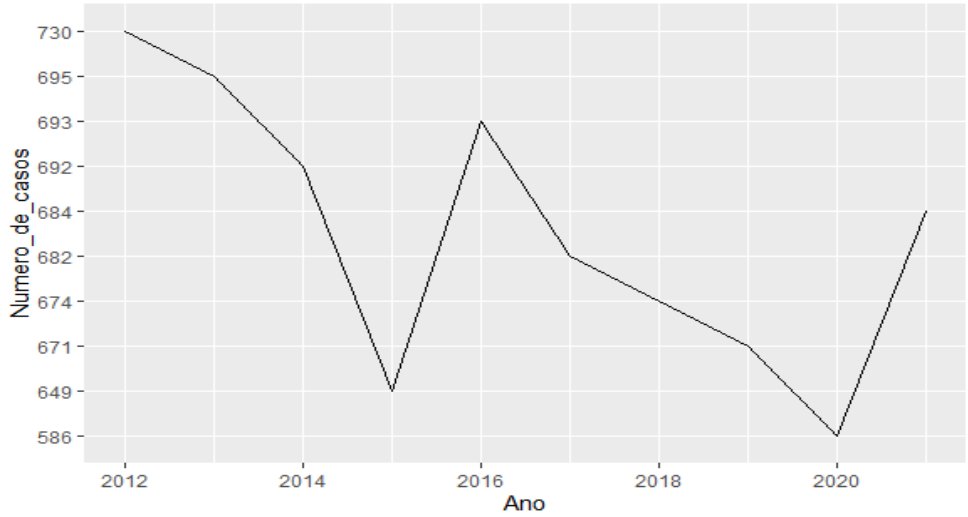


Figura 20 – Gráfico da evolução dos casos de coinfecção TB/HIV no período de 2012 a 2021, no estado de Pernambuco. (Fonte:Sistema de Informação de Agravos de Notificação - SINAN)

Na Figura 21, tem-se um panorama da coinfecção TB/HIV no estado de Pernambuco (2012-2022), onde teve-se acesso aos dados de gênero dos pacientes, onde dos 7.123 pacientes HIV positivo , com Tuberculose, 5015 (70,44%) são do sexo Masculino e 2105 (29,56%), Feminino. Houve maior incidência de casos para a faixa etária entre 30 e 39 anos, com 2.380 do total de casos, represenando um percentual de 33,43%. Em segundo lugar, no número de casos por faixa etária, temos a faixa dos 20 aos 29 anos, com 1.672 casos (23,48%) e em terceiro lugar a faixa dos 40 aos 49 anos, com 1.654 casos (23,23%). Quanto à Origem étnica é possível notar a maioria dos casos declarados Pardos (4.147 casos/58,24%). Em seguida brancos com 1010 (14,19%) e pretos com 790 casos (11,10%). Por último, com as menores porcentagens Amarelos(47 casos) 0,66% e a população indígena (30 casos) 0,42%. Quanto à forma clínica, a forma Pulmonar teve a maioria dos casos, com 4.965 casos (69,73%), seguida da forma Extrapulmonar, com 1.374 casos (19,30%) e 781 casos (10,97%) foram nas duas formas (Pulmonar e Extrapulmonar).

Coinfecção HIV/TB em PE de 2012 a 2022

Uma visão geral

Variáveis	N	%
Sexo		
Masculino	5015	70,44%
Feminino	2105	29,56%
Fx Etária		
Menor 1 ano	45	0,63%
1 a 4 anos	28	0,39%
5 a 9 anos	19	0,27%
10 a 14 anos	23	0,32%
15 a 19 anos	135	1,90%
20 a 29 anos	1672	23,48%
30 a 39 anos	2380	33,43%
40 a 49 anos	1654	23,23%
50 a 59 anos	860	12,08%
60 a 69 anos	237	3,33%
70 a 79 anos	44	0,62%
80 anos e mais	23	0,32%
Origem étnica		
Branca	1010	14,19%
Preta	790	11,10%
Amarela	47	0,66%
Parda	4147	58,24%
Indigena	30	0,42%
Não preenchido	1096	15,39%
Forma clínica		
PULMONAR	4965	69,73%
EXTRAPULMONAR	1374	19,30%
PULMONAR + EXTRAPULMONAR	781	10,97%

Figura 21 – Panorama dos casos de coinfecção TB/HIV no período de 2012 a 2022, no estado de Pernambuco. (Fonte:Sistema de Informação de Agravos de Notificação - SINAN)

A análise dos resultados da situação de encerramento dos casos de TB no estado (Figura 22) revelou uma taxa de cura de 33,79% (n=2539) e um número significativo de casos de abandono (18,75%; n=1409), óbito por tuberculose (2,37%; n =178), óbito por outras causas também bastante elevado (19,14%; n =1430) e desconhecido/em branco (5,35%; n=402).

Situacão.Encerra.	N	%
Ign/Branco	402	5.35
Cura	2539	33.79
Abandono	1409	18.75
Óbito por tuberculose	178	2.37
Óbito por outras causas	1438	19.14
Transferência	1412	18.79
TB-DR	39	0.52
Mudança de Esquema	80	1.06
Falência	2	0.03
Abandono Primário	15	0.20

Figura 22 – Situação de Encerramento para os casos de coinfecção TB/HIV no período de 2012 a 2022, no estado de Pernambuco. (Fonte:Sistema de Informação de Agravos de Notificação - SINAN)

4.6 Conclusão

A Análise Epidemiológica da Coinfecção de MTB e HIV, considerando aspectos da região Nordeste do Brasil, apresentada neste capítulo, foi desenvolvida como o objetivo de se familiarizar com os bancos de dados públicos e conhecer como este tipo de coinfecção se apresenta atualmente em nossa região, mapenado, quantificando e relacionando dados clínicos recentes, referentes a este tipo de coinfecção, disponíveis no SINAN. Nesse estudo descritivo, para o estado de Pernambuco, observou-se que a maioria de coinfectados TB/HIV em nossa região eram do sexo masculino. A faixa etária predominante foi de 30 a 39 anos, seguida, em um percentual próximo, da faixa etária dos 40 aos 49 anos. Uma outra informação presente neste tipo análise, considera a origem étinica declarada do paciente. Para esta, 58,24% dos coinfectados, se declararam pardos. Considerando a forma clínica da doença, fator que também pode conduzir a políticas públicas, 69,73% dos casos em Pernambuco eram de turberculose Pulmonar. Em relação ao encerramento dos casos de coinfecção, o estado apresentou uma taxa de 33,79% de cura da TB, porém, o índice de abandono do tratamento se mostrou bastante elevado (19,14%).

O conhecimento do perfil epidemiológico de TB/HIV no estado de Pernambuco corroborou também para a familiarização com os bancos de dados públicos, e motivou os estudos que permitissem uma busca mais rápida e formas de exibir a informação

gerada através dessas buscas que fossem mais legíveis para os profissionais que fazem uso cotidianos de tais bases de dados, porém, não têm formação específica na área de computação. Os resultados da pesquisa realizada nesse sentido, compõem o próximo capítulo.

5 Análise dos dados do SINAN no Neo4J

5.1 Considerações Iniciais

O presente capítulo apresenta um conjunto de processos para mapeamento da base de dados do SINAN em um bancos de dados orientado a grafos, o Neo4J. A motivação para o desenvolvimento de tais processos, surge a partir de uma necessidade inerente a análise de dados biológicos: correlacionar elementos envolveldo um grande volume de dados. Ao utilizarmos bancos orientados a grafos, o tempo de resposta para as buscas, que envolvem muitas correlações entre os itens, diminui, além de se tornarem mais simples e intuitivas.

5.2 Hardware utilizado

O Hardware utlizado foi um Notebook Samsung modelo NP300E4C, com Processador i3 2328M, vídeo integrado, 8GB de Memória RAM DDR3 1600MHz, SSD de 480GB e sistema operacional Windows 10 64 bits.

As consultas em Cypher obtiveram uma performance satisfatória, com tempos bastante baixos, com exceção da consulta de carregamento e criação dos elementos do grafo, disponível em 5.2, que levou entre 3 e 4 minutos para ser concluída, o que pode ser justificado pela quantidade de registros carregado, que foi próximo aos 6 mil. Todas as demais consultas obtiveram tempos de resposta na casa dos milisegundos.

O tratamento do Script em R também obteve um desempenho satisfatório.

Uma limitação que pôde ser notada, foi a manipulação na ferramenta gráfica do Neo4J, com o gráfico carregado por inteiro, em que o tempo de resposta para o zoom foi bastante demorado. É sugerido, então, que para esse tipo de atividade, se tenha ao menos um processador i5 e vídeo *offboard*, se possível, além de, no mínimo 8GB de RAM, de preferência DDR3, pois aplicações que envolvem gráficos costumam ter um custo computacional elevado.

Para o estudo em questão, procurou-se efetuar buscar com limite no número de nós, para evitar esse tipo de comportamento indesejado. Os resultados de tal procedimento surtiu um efeito bastante satisfatório, não atrapalhando a análise dos dados no Neo4J.

5.3 Mapeamento de Dados do SINAN para o Banco Neo4J

Na Figura 23 é possível visualizar um diagrama onde há o passo a passo para a Extração, Transformação e Carregamento(ETL - do Inglês, *'Extract Transform Load'*) dos dados do SINAN no Neo4J.

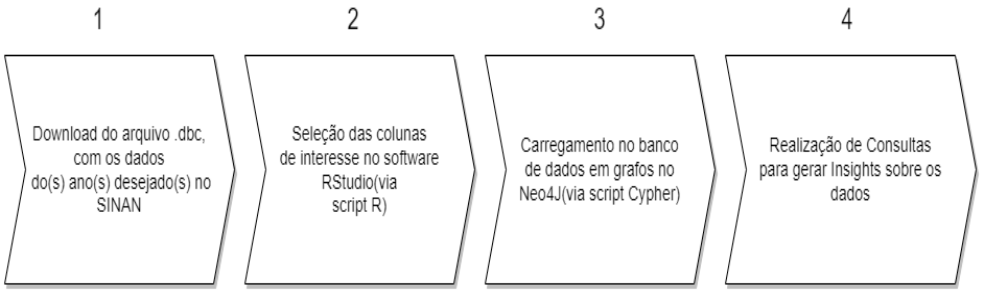


Figura 23 – Passo a passo do processo de extração dos dados do SINAN: Obtenção dos dados na fonte(SINAN), Seleção das colunas de interesse, carregamento no banco de dados em grafos e realização de consultas para gerar *insights* e produzir conhecimento sobre os dados. (Fonte: o autor)

O arquivo .dbc é o arquivo de entrada do processo de ETL do estudo, ele é obtido através da plataforma SINAN, em: Transferência de Arquivos - SINAN.

A seleção das colunas foi feita através de um script em R, e o funcionamento deste pode ser visualizado no fluxograma da Figura 26, que representa o código contido em 5.1.

5.3.1 Etapa 1 - Download do arquivo de entrada

Primeiramente foi executada a etapa de obtenção (*'download'*) do arquivo de entrada no SINAN, navegando até o endereço: Transferência de Arquivos - SINAN.

Acessada a página web (Figura 24), é possível notar que existe um conjunto de filtros, que devem ser selecionados, para a obtenção do(s) arquivo(s) desejados apra a análise.

No filtro 'Fonte', se escolhe a opção SINAN, no filtro 'Modalidade', a opção Dados, em seguida, seleciona-se o tipo de agravo no filtro 'Tipo de Arquivo'(no nosso caso TUBE - Tuberculose).

A etapa seguinte é a seleção do ano ou anos desejados, através do filtro 'Ano'e, finalmente, é preenchido o filtro UF, que possui apenas a opção 'BR'.

Para finalizar, clica-se então no botão 'Enviar' e em então obtem-se o(s) arquivo(s) para download no formato .dbc. Clicando na opção *'Download'* é possível obtê-lo(s).

A fim de realizar o carregamento para análise dos dados no Neo4J, foi necessário executar uma etapa prévia de transformação dos dados, que foi feita com o auxílio de um *script*, construído na linguagem de programação R, que pode ser visualizado no Código 5.1.

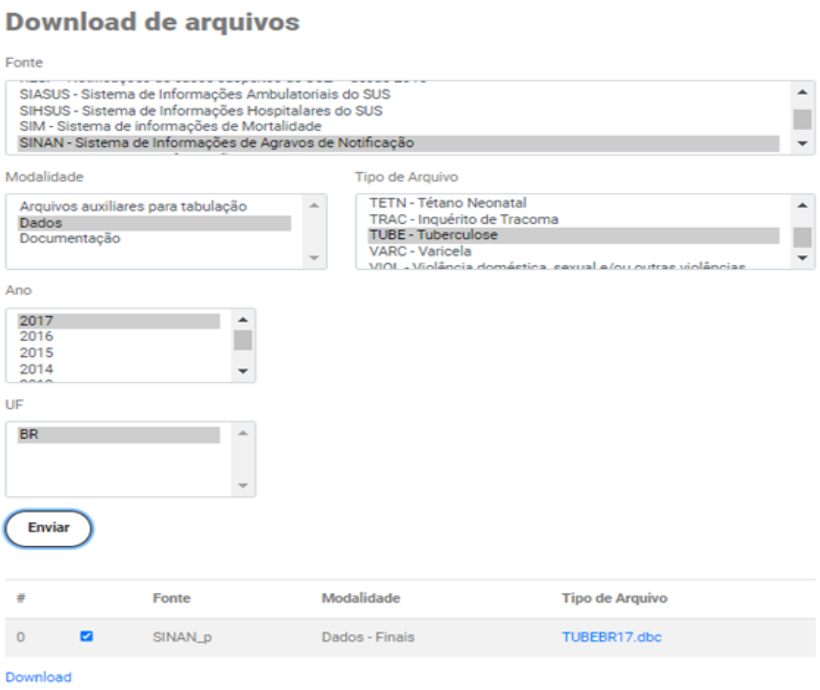


Figura 24 – Tela de download de arquivos do SINAN, onde são selecionados os arquivos de Tuberculose por ano desejado. (Fonte: DATASUS - Transferência de Arquivos)

5.3.2 Etapa 2 - Seleção das colunas de interesse

As colunas disponíveis no arquivo original .dbc foram num total de 97 e foi utilizado um dicionário de dados, disponível em: Datasus- SINAN. Porém esse dicionário conta com a descrição apenas de 33 colunas. Para o estudo aqui descrito, foram selecionadas 31 colunas, que podem ser visualizadas no diagrama da Figura 25. Vale ressaltar, também, que , dessas 31 conlunas, nem todas foram utilizadas, mas foi decidido manter algumas por sua potencial importância para processos analíticos futuros, como é o caso das colunas referentes às drogas utilizadas nos medicamentos, mas que não é sabida a razão pela qual não são preenchidas nas bases de dados verificadas do SINAN.

Efetivamente, para este trabalho, só foram utilizadas 13 colunas. As colunas utilizadas foram: ANO_NASC, ANT_RETRO, DT_DIAG, CS_SEXO, UF, DT_INIC_TR, TRATAMENTO, RAIXO_TORA, BACILOSC_E, HIV, SITUA_ENCE, AGRAVAIDS, DT_ENCERRA e estão marcadas em Laranja, na Figura 25.

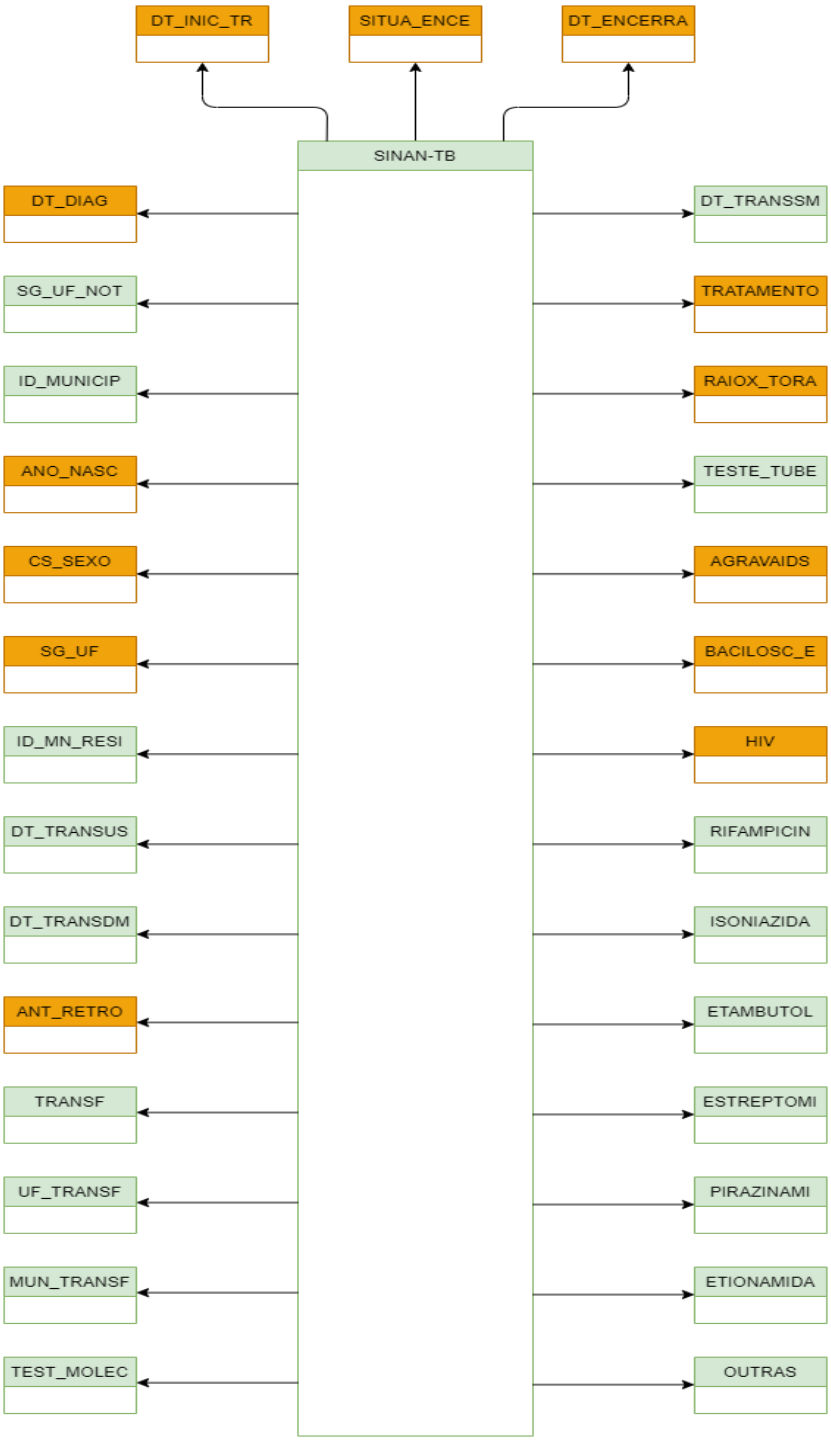


Figura 25 – Colunas selecionadas do SINAN-TB. (Fonte: o autor)

Na Figura 26, é possível visualizar um fluxograma, que representa o código contido no script de seleção das colunas, 5.1.

Como pode ser visto na linha 4, do código 5.1, as colunas são selecionadas por seus números. Foi escolhido ser feito dessa maneira, pois, na visualização prévia que é feita do arquivo, através do comando da linha 2, no RStudio, colocando o ponteiro do *mouse* sobre cada coluna, é exibido o número da mesma.

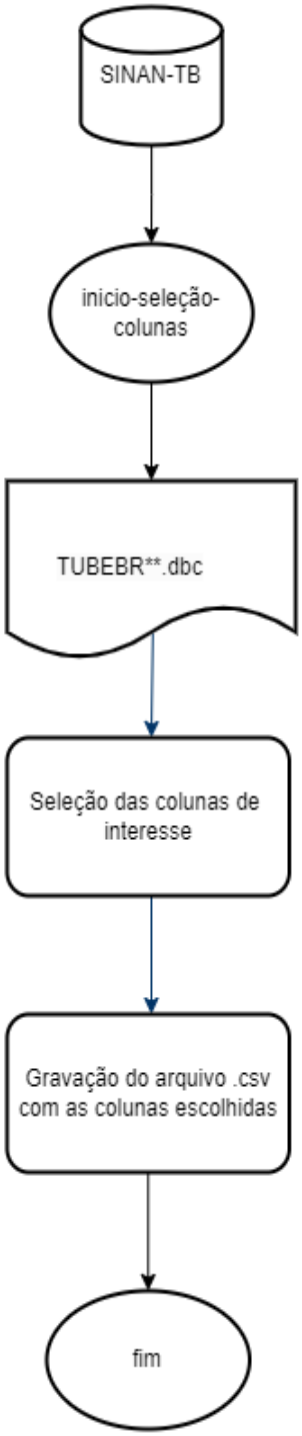


Figura 26 – Fluxograma do script em R para seleção das colunas de interesse. (Fonte: o autor)

```
1 tuber17 <- read.dbc("TUBEBR17.dbc")
2 View(tuber17) #visualizar arquivo TUBEBR17.dbc
3 #seleção das colunas
4 tuber17Inter <- tuber17[,c(5,6,8,9,11,15,16,22:24,32,34,35,40,
5             46,51,53:60,81,82,91,93,95:97)]
6
7 tuberHivPe17 = tuber17Inter[FALSE,]
```

```r
#fim seleção das colunas
c <- 1
#seleção dos pacientes do estado de Pernambuco(26)
for(i in 1:nrow(tuber17Inter)) {
  if(tuber17Inter[i, 6] == 26)
  {
    tuberHivPe17[c, ]<-tuber17Inter[i, ]
    c <- c + 1
  }
}
#fim da seleção dos pacientes do estado de Pernambuco

View(tuberHivPe17) #visualização dos dados filtrados
#gravação do arquivo em formato .csv
write.table(tuberHivPe17, "tuberHivPe17.csv", row.names=TRUE,
            quote=FALSE, sep = ",", dec = ".")
```

Código 5.1 – Código em R para filtragem das colunas de interesse dos dados do SINAN-TB. (Fonte: o autor)

Filter

	SG_UF_NOT	ID_MUNICIP	DT_DIAG	ANO_NASC	CS_SEXO	SG_UF	ID_MN_RESI	DT_TRANSUS	DT_TRANSDM	DT_TRANSSM
103	26	260190	2017-10-20	1988	F	26	260190	NA	NA	2020-12-09
194	26	261160	2017-05-15	1978	M	26	260960	NA	2017-11-10	NA
195	26	260960	2017-08-28	1978	M	26	260960	NA	NA	2019-03-25
196	26	261160	2017-03-09	1974	F	26	261160	NA	2020-03-23	NA
197	26	261160	2017-08-28	1974	F	26	261160	NA	2017-10-30	NA
198	26	261160	2017-06-26	1997	M	26	261160	NA	2018-04-16	NA
199	26	261160	2017-10-04	1997	M	26	261160	NA	2018-04-20	NA
200	26	261160	2017-11-17	1955	M	26	261160	NA	2018-01-29	NA
201	26	260345	2017-11-17	1955	F	26	260345	NA	NA	2018-08-02
202	26	260640	2017-09-12	1995	M	26	260640	NA	NA	2018-02-20
203	26	261160	2017-10-03	1985	M	26	261160	NA	2018-03-02	NA
204	26	260960	2017-08-20	1962	M	26	260960	NA	NA	2019-03-18
235	26	261160	2017-04-27	1988	F	26	261640	NA	2017-08-21	NA
245	26	260960	2017-11-08	1975	M	26	260960	NA	NA	2018-07-31
257	26	260650	2017-09-12	1963	M	26	260650	NA	NA	2019-07-08
264	26	260600	2017-10-25	1983	F	26	260600	NA	NA	2018-03-08
288	29	291840	2017-08-16	1953	F	26	260875	NA	NA	2017-08-24
297	26	260540	2017-09-19	1993	M	26	260540	NA	NA	2018-12-10
298	26	260540	2017-09-19	1993	M	26	260540	NA	NA	2019-10-01
299	26	260960	2017-06-22	1963	M	26	260960	NA	NA	2019-02-15

Showing 1 to 20 of 5,973 entries, 31 total columns

Figura 27 – Arquivo .dbc carregado no RStudio, com as 31 colunas selecionadas. (Fonte: o autor)

5.3.3 Etapa 3 - Carregamento dos dados no Neo4J

No Código 5.2, o código em *Cypher*, utilizado para carregar o arquivo .csv gerado, pode ser visualizado.

O arquivo .csv é carregado e todos os nós de interesse, como Paciente, Agravo_AIDS, HIV, Tratament(Tratamento), Exame, Encerramento, assim como os relacionamentos TESTOU_HIV, AGRAVO_AIDS, INICIOU, TEM, PRIM_EXAME, SEG_EXAM são criados.

```
:auto USING PERIODIC COMMIT 500
LOAD CSV WITH HEADERS from 'file:\\tuberHivPe17.csv' as row with row where
row.ANO_NASC is not null
WITH row,
(CASE row.HIV
 WHEN '1' THEN 'Positivo'
 WHEN '2' THEN 'Negativo'
 WHEN '3' THEN 'Em andamento'
 WHEN '4' THEN 'Não realizado'
 WHEN 'NA' THEN 'Não informado' END) AS hiv
MERGE (p:Paciente {ID:row.ID_PAC,UF:'PE',DT_DIAG: row.DT_DIAG, ANO_NASC:
    row.ANO_NASC,ANT_RETRO:row.ANT_RETRO,SEXO:row.CS_SEXO})
MERGE (agravo_aids:Agravo_AIDS {ID:row.ID_PAC})
MERGE (thiv:HIV {ID:row.ID_PAC})
MERGE (p)-[r:TESTOU_HIV]->(thiv)
MERGE (thiv)-[r_agravaids:AGRAVO_AIDS]->(agravo_aids)
MERGE (t:Tratament {ID_PAC:row.ID_PAC,DT_INICIO:row.DT_INIC_TR,TRATAMENTO:
    row.TRATAMENTO})
MERGE (p)-[q:INICIOU]->(t)
MERGE (t)-[:TEM]->(p)
MERGE (bac_e:Exame {ID_PAC:row.ID_PAC,TIPO:'BACILOSC_E',RESULTADO:row.
    BACILOSC_E})
MERGE (t)-[:PRIM_EXAME]->(bac_e)
MERGE (radio:Exame {ID_PAC:row.ID_PAC,TIPO:'RAIOX_TORA',RESULTADO:row.
    RAIOX_TORA})
MERGE (bac_e)-[:SEG_EXAM]->(radio)
MERGE (ence:Encerramento {ID_PAC:row.ID_PAC,DT_ENCERRAMENTO:row.DT_ENCERRA,
    SITUAC_ENCERRA:row.SITUA_ENCE})
MERGE (radio)-[:ENCERROU]->(ence)
SET thiv.RESULT=hiv
SET p.IDADE = 2017-toInteger(row.ANO_NASC)
```

Código 5.2 – Código em Cypher para carregamento das variáveis de interesse e criação do grafo inicial. (Fonte: o autor)

5.3.4 Etapa 4 - Realização de consultas no grafo

Feita a etapa 3, de carregamento dos dados na ferramenta do Neo4J e criação do grafo, com seus nós e relacionamentos, nesta seção são apresentadas algumas consultas realizadas, que puderam gerar insights sobre os dados.

A primeira consulta trata-se de um *Match*, que retorna todos os nós e relacionamentos do grafo, em 5.3 é possível visualizá-la. Na Figura 30 tem-se a representação gráfica.

```
1 MATCH (n) RETURN n
```

Código 5.3 – Código *Cypher* que retorna todos os nós do grafo carregado na consulta do Código 5.2. (Fonte: o autor)

Figura 28 – Visão geral do grafo gerado no Neo4J Browser. (Fonte: o autor)

Figura 29 – Visão aproximada do grafo - Parte 1. (Fonte: o autor)

Figura 30 – Visão aproximada do grafo - Parte 2. (Fonte: o autor)

Na Figura 31 tem-se uma visão aproximada do grafo gerado, no Neo4JBrowser. É possível notar que a ferramenta possui um recurso chamado *'Overview'* (em português Visão Geral), onde são exibidos todas as etiquetas marcadoras dos nós*'Node*

labels' e suas respectivas cores e quantidades. Cada nó possui uma cor distinta, o que possibilita diferenciá-los. Esse mesmo recurso é utilizado para exibir os tipos de relacionamentos no grafo (*'Relationship types'*).

```
1 MATCH (n) RETURN n LIMIT 1000
```

Código 5.4 – Código *Cypher* que retorna os nós e relacionamentos, com limite de 100 nós, no total, para o grafo carregado. (Fonte: o autor)

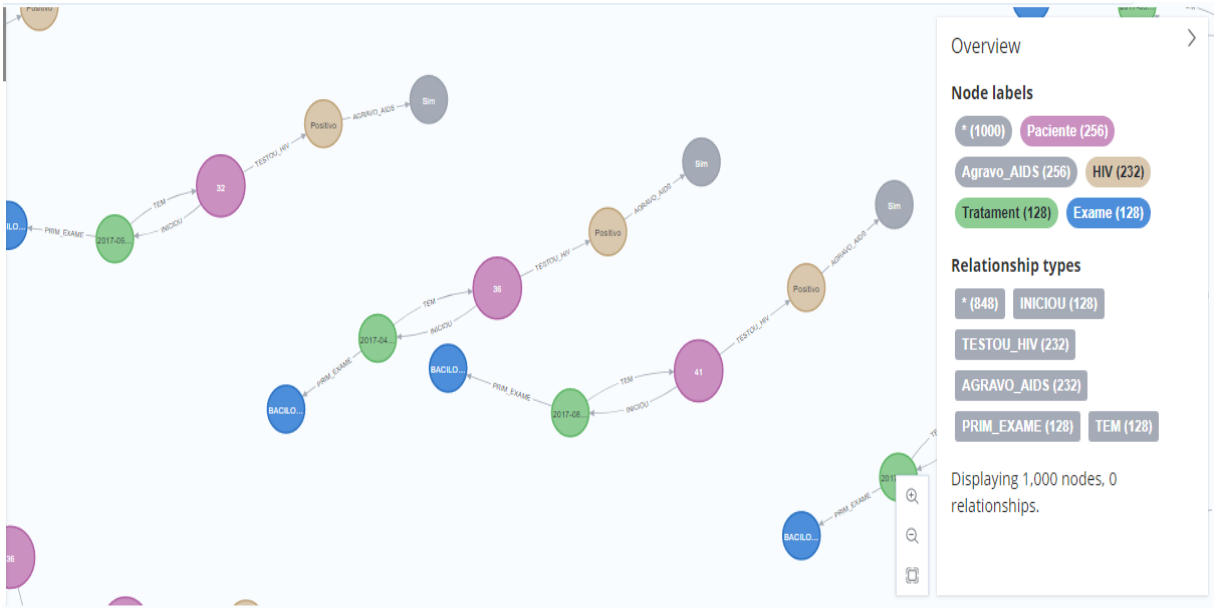


Figura 31 – Visão aproximada do grafo gerado no Neo4J Browser. (Fonte: o autor)

Segundo o dicionário de dados, fornecido pelo SINAN, o campo SITUA_ENC, carregado no Neo4J como SITUAC_ENCERRA, pode conter dados numerados de 1 a 10, em que o número 7 corresponde a pacientes que obtiveram encerramento de Tuberculose Drogarresistente (TBDR). Na Tabela 2, é possível ver essa parte do dicionário de dados.

Nome do campo	Campo	Tipo	Categorias	Descrição	Característica	DBF
62. Situação de encerramento	tp_situacao_encerramento	Varchar2(1)	1. Cura 2. Abandono 3. Óbito por TB 4. Óbito por outras causas 5. Transferência 6. Mudança de Diagnóstico 7. TB-DR 8. Mudança de Esquema 9. Falência 10. Abandono Primário	Situação de encerramento do caso notificado	Campo Obrigatório quando Campo 66 (Data de Encerramento) estiver preenchido. ...	SITUA_ENCE

Tabela 2 – Descrição do campo Situação de encerramento, no dicionário de dados do SINAN. (Fonte: Datasus - SINAN - adaptado)

No Código 5.5 é possível notar que é criado um nó de agravo 'TBDR', através do comando na linha 2: MERGE(tbdr:TBDR{AGRAVO:'TBDR'}) e, em seguida, é criado um relacionamento entre o Paciente 'n' e o nó TBDR criado, através do comando na linha 3: MERGE (n)-[:DIAGNOSTICOU]-(tbdr). Logo após, é executada uma consulta, disponível no Código 5.6, para localizar no grafo pacientes que tiveram encerramento categoria '7'. Na Figura 33 nota-se o resultado da modificação promovida pelo código 5.5, onde os nós pacientes que tiveram situação de encerramento '7', contam com um novo relacionamento [:DIAGNOSTICOU] apontando para o novo nó em comum entre eles 'TBDR', ou, como o código 5.5 especificou: (n:Paciente)-[:DIAGNOSTICOU]-(tbdr:TBDR). A essa consulta, como é possível notar no *Match* feito, foram incluídos o primeiro e o segundo exames realizados pelo paciente.

```
1 MATCH p=(thiv:HIV)<-[hiv:TESTOU_HIV]-(n:Paciente)-[i:INICIOU]->(t:Tratament
    )-[pri_ex:PRIM_EXAME]->(p_ex:Exame)-[seg_ex:SEG_EXAM]->(seg_exam:Exame)
    -[encer:ENCERROU]->(enc:Encerramento{SITUAC_ENCERRA:'7'})
2 MERGE(tbdr:TBDR{AGRAVO:'TBDR'})
3 MERGE (n)-[:DIAGNOSTICOU]-(tbdr)
```

```
4 RETURN p
```

Código 5.5 – Código *Cypher* para busca dos pacientes que tiveram encerramento de Tuberculose Drogarresistente (TBDR), com a adição do novo nó 'TBDR'. (Fonte: o autor)

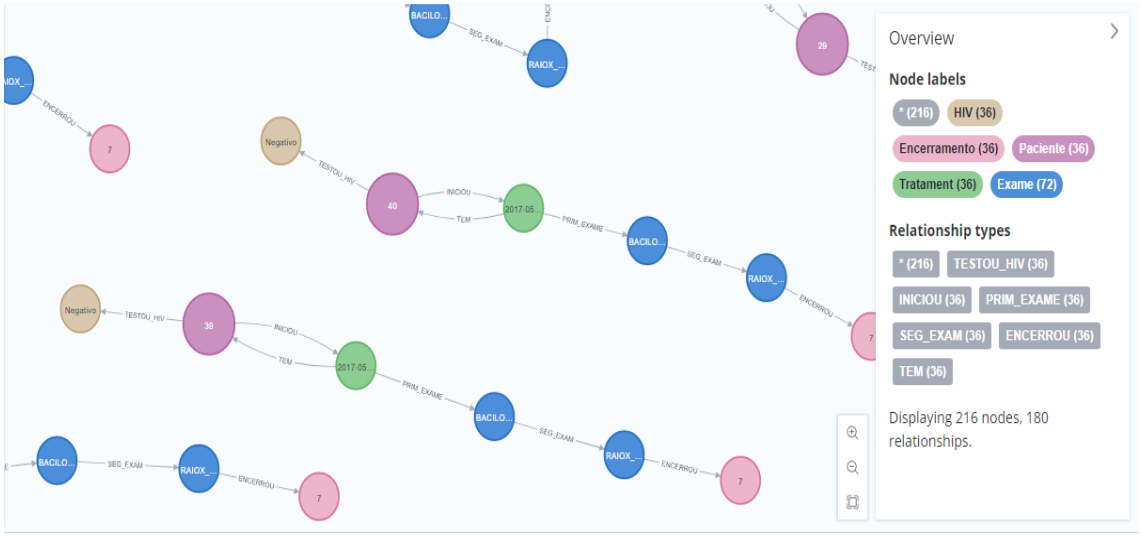


Figura 32 – Grafo resultado da busca do Código 5.5 (pacientes que tenham tido encerramento de Tuberculose Drogarresistente - TBDR). (Fonte o autor)

```
1 MATCH p=()<-[h*]-(n:Paciente)-[i:INICIOU]->(t:Tratament)-[pri_ex:PRIM_EXAME
    ]->(p_ex:Exame)-[seg_ex:SEG_EXAM]->(seg_exam:Exame)-[encer:ENCERROU]->(
    enc:Encerramento{SITUAC_ENCERRA:'7'})
2 RETURN p
```

Código 5.6 – Código *Cypher* para busca dos pacientes que tenham tido encerramento com Tuberculose Drogarresistente - TBDR. (Fonte: o autor)

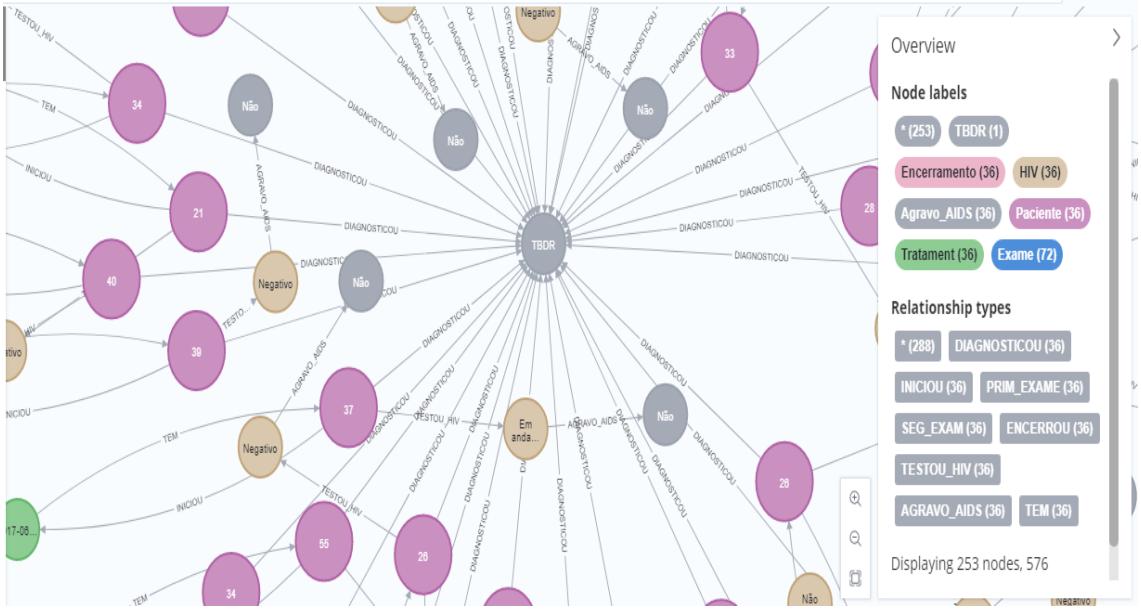
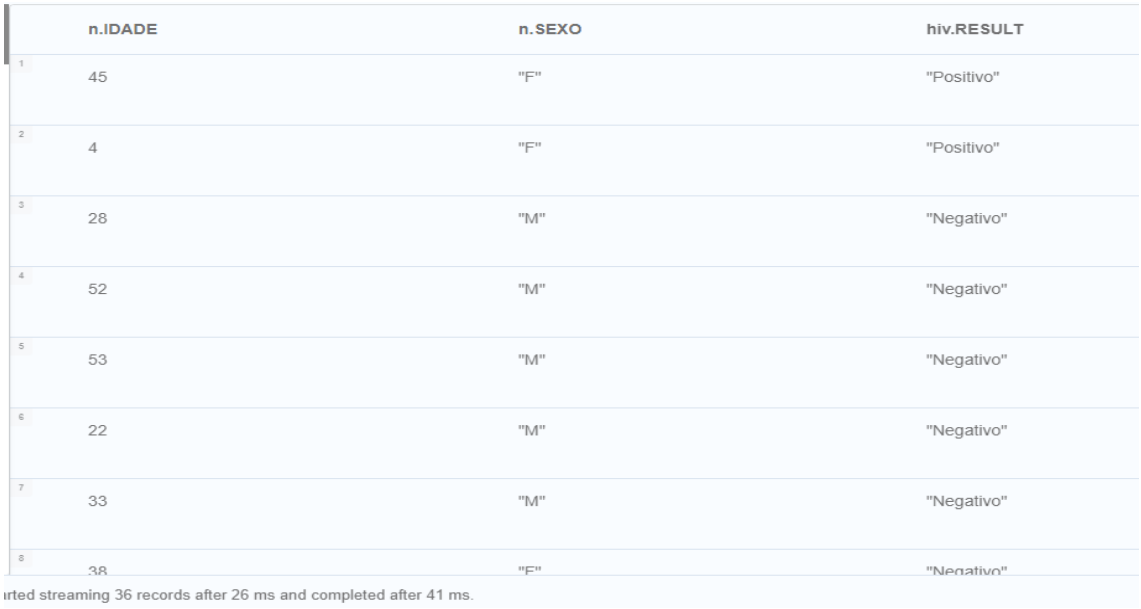


Figura 33 – Grafo resultado da busca do Código 5.6 - pacientes que tiveram encerramento com Tuberculose Drogarresistente (TBDR). (Fonte: o autor)

A consulta do Código 5.7 retorna os dados IDADE, SEXO e RESULTADO(RESULT) do exame de HIV (Positivo, Negativo, Em andamento, Não realizado) dos pacientes. Foram retornados 36 registros no total, como é possível ver, verificando o canto esquerdo inferior da consulta.

```
1 MATCH p=(hiv:HIV)<-[h:TESTOU_HIV]-(n:Paciente)-[i:INICIOU]->(t:Tratament)-[
    pri_ex:PRIM_EXAME]->(p_ex:Exame)-[seg_ex:SEG_EXAM]->(seg_exam:Exame)-[
    encer:ENCERROU]->(enc:Encerramento{SITUAC_ENCERRA:'7'})
2 RETURN n.IDADE,n.SEXO,hiv.RESULT
```

Código 5.7 – Código *Cypher* que retorna IDADE, SEXO e Resultado do Exame de HIV, para pacientes com Tuberculose Drogarresistente (TBDR). (Fonte: o autor)



	n.IDADE	n.SEXO	hiv.RESULT
1	45	"F"	"Positivo"
2	4	"F"	"Positivo"
3	28	"M"	"Negativo"
4	52	"M"	"Negativo"
5	53	"M"	"Negativo"
6	22	"M"	"Negativo"
7	33	"M"	"Negativo"
8	38	"F"	"Negativo"

Figura 34 – Resultado da busca do Código 5.7 - pacientes que tiveram encerramento com Tuberculose Drogarresistente (TBDR). (Fonte: o autor)

```
1 MATCH p=(hiv:HIV)<-[h:TESTOU_HIV]-(n:Paciente)-[i:INICIOU]->(t:Tratament)-[
    pri_ex:PRIM_EXAME]->(p_ex:Exame)-[seg_ex:SEG_EXAM]->(seg_exam:Exame)-[
    encer:ENCERROU]->(enc:Encerramento{SITUAC_ENCERRA:'7'})
2 WHERE hiv.RESULT = "Positivo"
3 RETURN n.IDADE,n.SEXO,hiv.RESULT
```

Código 5.8 – Código *Cypher* que retorna IDADE, SEXO e Resultado do Exame de HIV Positivo, para pacientes com Tuberculose Drogarresistente (TBDR). (Fonte: o autor)

	n.IDADE	n.SEXO	hiv.RESULT
1	45	"F"	"Positivo"
2	4	"F"	"Positivo"

Figura 35 – Resultado da busca do Código 5.8 - pacientes que tiveram encerramento com Tuberculose Drogarresistente (TBDR) e HIV resultado Positivo. (Fonte: o autor)

```
MATCH p=(hiv:HIV)<-[h:TESTOU_HIV]-(n:Paciente)-[i:INICIOU]->(t:Tratament)-[
    pri_ex:PRIM_EXAME]->(p_ex:Exame)-[seg_ex:SEG_EXAM]->(seg_exam:Exame)-[
    encer:ENCERROU]->(enc:Encerramento{SITUAC_ENCERRA:'7'})
WHERE hiv.RESULT = "Negativo"
RETURN n.IDADE,hiv.RESULT,count(n)
ORDER BY n.IDADE
```

Código 5.9 – Código *Cypher* que retorna IDADE e Resultado do Exame de HIV Positivo, para pacientes com Tuberculose Drogarresistente (TBDR). (Fonte: o autor)

Como é possível ver na Figura 36, com a consulta do código 5.9 é possível obter uma medida frequencista, da quantidade de pacientes com Tuberculose e com resultado HIV Negativo, que apresentaram Encerramento com Tubeculose Droga Resistente (TBDR), uma vez que a linguagem *Cypher* faz o agrupamento nessa consulta pelo campo que varia, que é o campo IDADE. Logo, a função *count()*, apresenta um total de 2 pessoas com 21 anos, 3 com 28 anos e 2 com 29, a idade máxima apresentada foi de 80 anos e a mínima, 21 anos.

	n.IDADE	hiv.RESULT	count(n)
1	21	"Negativo"	2
2	22	"Negativo"	1
3	26	"Negativo"	1
4	28	"Negativo"	3
5	29	"Negativo"	2
6	31	"Negativo"	1
7			

rted streaming 22 records after 1 ms and completed after 12 ms.

Figura 36 – Resultado da busca do Código 5.9 - pacientes que tiveram encerramento com Tuberculose Drogarresistente (TBDR) e HIV resultado Negativo. (Fonte: o autor)

5.4 Conclusão

O presente capítulo reportou a motivação para se considerar bancos NoSQL na análise de dados biológicos. A revisão realizada, levou a implementar um mapeamento dos dados do SINAN no banco orientado a grafos Neo4J. A forma de visualização de dados e possibilidades de explorar diversos relacionamentos se mostrou promissora.

No entanto, a complexidade computacional envolvida nesse processo, pode torná-lo inacessível para profissionais que não têm domínio dessa área de conhecimento. Assim, uma proposta para a continuidade da pesquisa descrita nesse trabalho seria desenvolver uma interface que permitisse utilizar os scripts desenvolvidos para qualquer conjunto de interesse no SINAN, de modo que os pesquisadores não familiarizados com programação, realizassem esse tipo de análise de forma simples. A utilização do Neo4J em outros tipos de dados biológicos também é considerada.

6 Conclusões e Trabalhos Futuros

A pesquisa resportada no presente trabalho, buscou aplicar abordagens computacionais para dados biológicos referentes a processos de coinfecção de patógenos, com ênfase em HIV/TB. A escolha dos patógenos envolvidos, foi motivada por sua incidência mundial, a qual demanda por métodos que possam auxiliar, por exemplo, na definição de políticas públicas.

Após uma revisão exploratória da literatura, optou-se por realizar dois tipos de experimentos, utilizando dados relacionados a ocorrências de HIV/TB em Pernambuco, advindos do SINAN: Uma análise epidmiológica atual deste tipo de coinfecção e mapeamento de dados do SINAN em um bancos de dados orientados a grafos, afim de obeter insigts mais significativos relacionados a análise clínica de pacientes coninfectados por HIV/TB.

A Análise Epidemiológica da Coinfecção de MTB e HIV, além de propiciar uma familiarização com o tipo de informação disponível no SINAN, permitiu conhecer o perfil de pacientes coinfectados no estado de Pernambuco. A análise mapeou dados tais como faixa etária, gênero, origem étinica, forma clínica da doença e encerramento do caso. Dentre os resultados obtidos, descritos no Capítulo 4, destamos o alto índice de Tuberculose Pulmonar compreendendo 69,73% dos casos em Pernambuco e um índice de abandono do tratamento também bastante elevado (19,14%).

O estudo também reportou a motivação para se considerar bancos NoSQL na análise de dados biológicos, dada a natureza das pesquisas realizadas em bancos de dados clínicos. Como estudo de caso, foi realizado um mapeamento dos dados de interessse do SINAN no banco orientado a grafos Neo4J, o que permitiu explorar formas de realionamento de dados, e visualização destes, destacando as vatagens do mapeamento no Neo4J.

A complexidade computacional envolvida no processo de mapeamento, e, a ciência de que tais bancos de dados são utilizados predominantemente por profissionais que não têm domínio dessa área de conhecimento, motivou uma proposta para a continuidade da pesquisa que tornasse a utilização dos scripts mais fácil para tais profissionais.

Um outro trabalho futuro motivado pela pesquisa aqui reportada, foi a utilização do Neo4J em outros tipos de dados biológicos, mas especificamente, pretende-se realizar estudos no campo de exploração de dados de coinfecção a nível molecular, buscando regiões de interesses, como os *motifs*.

Referências

BARBOSA, I. R.; COSTA, Í. d. C. C. ESTUDO EPIDEMIOLÓGICO DA COINFEC

Cão TUBERCULOSE-HIV NO NORDESTE DO BRASIL. *Rev. Patol. Trop.*, Universidade Federal de Goias, v. 43, n. 1, abr. 2014. Citado 3 vezes nas páginas 8, 34 e 35.

CARVALHO, G. L. V. d. *Ambiente de Dados do Sistema de Informação de Agravos de Notificação com Neo4J*. 68 p. Monografia — Universidade Federal do Rio de Janeiro, 2019. Citado 2 vezes nas páginas 27 e 31.

CAVALIN, R. F. et al. Coinfecção TB-HIV. *Rev. Saude Publica*, Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), v. 54, p. e112, nov. 2020. Citado na página 34.

COX; DOUDNA; O'DONNELL. *Biologia Molecular: Princípios e Técnicas*. [S.l.]: Artmed Editora, 2012. Citado 2 vezes nas páginas 19 e 21.

D'HAESELEER, P. What are DNA sequence motifs? *Nat. Biotechnol.*, Springer Science and Business Media LLC, v. 24, n. 4, p. 423–425, abr. 2006. Citado na página 19.

ESCADA, R. O. da S. et al. Mortality in patients with HIV-1 and tuberculosis co-infection in rio de janeiro, brazil - associated factors and causes of death. *BMC Infect. Dis.*, v. 17, n. 1, p. 373, maio 2017. Citado na página 17.

FRAME, D. A.; BLUMENFELD, Z. *Graph Data Science(GDS) for Dummies*. [S.l.]: John Wiley Sons, Inc., 2022. (for Dummies). ISBN 9781119909323. Citado 2 vezes nas páginas 7 e 23.

GHAZNAVI, H. et al. SARS-CoV-2 and influenza viruses: Strategies to cope with coinfection and bioinformatics perspective. *Cell Biol. Int.*, Wiley, v. 46, n. 7, p. 1009–1020, jul. 2022. Citado na página 15.

GIRARDI CAROLINA S.; SUBTIL, F. T. R. J. O. *Biologia Molecular*. [S.l.]: Grupo A, 2018. Citado na página 20.

GLOBAL tuberculosis report 2020. Genève, Switzerland: World Health Organization, 2020. Citado 2 vezes nas páginas 15 e 30.

GONG, F. et al. Neo4j graph database realizes efficient storage performance of oilfield ontology. *PLoS One*, Public Library of Science (PLoS), v. 13, n. 11, p. e0207595, nov. 2018. Citado na página 27.

GRIFFITHS, E. C. et al. The nature and consequences of coinfection in humans. *J. Infect.*, Elsevier BV, v. 63, n. 3, p. 200–206, set. 2011. Citado 2 vezes nas páginas 15 e 29.

HE, Y. et al. A survey on deep learning in DNA/RNA motif mining. *Briefings in Bioinformatics*, v. 22, n. 4, 10 2020. ISSN 1477-4054. Bbaa229. Disponível em: <https://doi.org/10.1093/bib/bbaa229>. Citado 2 vezes nas páginas 7 e 22.

JÚNIOR, A. N. R. et al. History, current issues and future of the brazilian network for attending and studying trypanosoma cruzi/HIV coinfection. *J. Infect. Dev. Ctries.*, Journal of Infection in Developing Countries, v. 4, n. 11, p. 682–688, nov. 2010. Citado na página 15.

LANSBURY, L. et al. Co-infections in people with COVID-19: a systematic review and meta-analysis. *J. Infect.*, Elsevier BV, v. 81, n. 2, p. 266–275, ago. 2020. Citado 2 vezes nas páginas 15 e 29.

LOUTEN, J. Chapter 3 - features of host cells: Cellular and molecular biology review. In: LOUTEN, J. (Ed.). *Essential Human Virology*. Boston: Academic Press, 2016. p. 31–48. ISBN 978-0-12-800947-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012800947500003X>. Citado na página 22.

MEIDANIS, J.; SETUBAL, J. C. *Introduction to computational molecular biology*. Florence, KY: Brooks/Cole, 1997. Citado na página 19.

MONTALES, M. T. et al. Mycobacterium tuberculosis infection in a HIV-positive patient. *Respir. Med. Case Rep.*, Elsevier BV, v. 16, p. 160–162, out. 2015. Citado 2 vezes nas páginas 15 e 17.

MUKHATAYEVA, A. et al. Hepatitis b, hepatitis c, tuberculosis and sexually-transmitted infections among HIV positive patients in kazakhstan. *Sci. Rep.*, Springer Science and Business Media LLC, v. 11, n. 1, p. 13542, jun. 2021. Citado 4 vezes nas páginas 15, 17, 29 e 30.

NELSON DAVID L.; COX, M. M. *Princípios de Bioquímica de Lehninger*. Porto Alegre, BR: Artmed, 2014. Citado na página 19.

NIKAM, P.; BHOITE, S.; SHENOY, A. Neo4j graph database implementation for LinkedIn. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, Technoscience Academy, p. 339–342, dez. 2020. Citado na página 27.

OLIVEIRA, L. B. D. et al. ANÁLISE EPIDEMIOLÓGICA DA COINFEC

Cão TUBERCULOSE/HIV. *Cogitare Enferm.*, FapUNIFESP (SciELO), v. 23, n. 1, jan. 2018. Citado 2 vezes nas páginas 30 e 35.

PETRELLIS, G. et al. Parasitic worms affect virus coinfection: a mechanistic overview. *Trends Parasitol.*, Elsevier BV, v. 39, n. 5, p. 358–372, maio 2023. Citado na página 15.

PICANÇO-JUNIOR, O. M. et al. PRESEN

CA DO PAPILOMAVIRUS HUMANO TIPO 16 E expressão GÊNICA DA PROTEÍNA P16INK4A E ONCOPROTEÍNA E7 NO CARCINOMA COLORRETAL. *Arq. Bras. Cir. Dig.*, FapUNIFESP (SciELO), v. 34, n. 4, p. e1637, jan. 2022. Citado 2 vezes nas páginas 15 e 29.

PIETRO, M. D. et al. HPV/Chlamydia trachomatis co-infection: metagenomic analysis of cervical microbiota in asymptomatic women. *New Microbiol.*, v. 41, n. 1, p. 34–41, jan. 2018. Citado 2 vezes nas páginas 15 e 29.

PRZYTYCKA, T. *Lecture notes in Introduction to Computational Biology, Lecture 8: TF Biding Motifs*. Disponível em: <https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect08_Bind_Motifs.pdf>. Citado na página 19.

RAWSON, T. M. et al. Bacterial and fungal coinfection in individuals with coronavirus: A rapid review to support COVID-19 antimicrobial prescribing. *Clin. Infect. Dis.*, Oxford University Press (OUP), v. 71, n. 9, p. 2459–2468, dez. 2020. Citado na página 15.

ROBINSON, I.; WEBBER, J.; EIFREM, E. *Graph databases*. [S.l.: s.n.], 2015. Citado 8 vezes nas páginas 7, 10, 11, 22, 23, 24, 25 e 26.

ROCHA, M. S. et al. Sistema de informação de agravos de notificação (sinan): principais características da notificação e da análise de dados relacionada à tuberculose. *Epidemiol. Serv. Saude*, FapUNIFESP (SciELO), v. 29, n. 1, mar. 2020. Citado na página 30.

RSTUDIO: User guide. <https://docs.posit.co/ide/user/>. Acessado em: 05/05/2023. Citado na página 33.

SAMPAIO, R. S. *Ambiente de Dados do SIHSUS com MongoDB*. 75 p. Monografia — Universidade Federal do Rio de Janeiro, 2021. Citado 2 vezes nas páginas 27 e 32.

STOTHERS, J. A. M.; NGUYEN, A. Can neo4j replace PostgreSQL in healthcare? *AMIA Summits Transl. Sci. Proc.*, v. 2020, p. 646–653, maio 2020. Citado 2 vezes nas páginas 27 e 31.

WATERS, R. et al. The Mtb-HIV syndemic interaction: why treating m. tuberculosis infection may be crucial for HIV-1 eradication. *Future Virol.*, Future Medicine Ltd, v. 15, n. 2, p. 101–125, fev. 2020. Citado 3 vezes nas páginas 15, 16 e 29.

WHAT is Neo4j? <https://neo4j.com/developer/graph-database/#neo4j-overview>. Acessado em: 05/05/2023. Citado na página 23.

WHAT is R? <https://www.r-project.org/about.html>. Acessado em: 05/05/2023. Citado na página 33.