UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Extraction of phylogenomic information for the development of new approaches for geotraceability

Pedro Rafael Vila Cerqueira

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Versão Pública

Dissertação orientada por:
Prof. Doutor Ricardo Pedro Moreira Dias

2018

Resumo

Atualmente os consumidores têm ganho uma maior sensibilização face à autenticidade e origem dos produtos alimentares, apesar de a preocupação com a segurança alimentar também aumentar. A aplicação de novas metodologias como a implementação de sistemas de rastreabilidade, são o caminho a seguir para assegurar a autenticidade dos produtos alimentares. Os sistemas de rastreabilidade têm como objetivo proporcionar a capacidade de rastrear um produto desde a sua origem até às mãos do consumidor. Estes sistemas podem ser melhorados, integrando dados convencionais de rastreabilidade com outros domínios de dados, por exemplo, com dados geográficos ou topográficos. Portanto, a integração de dados convencionais de rastreabilidade com dados geográficos é referida como georrastreabilidade. Ao longo dos anos, várias ferramentas têm sido desenvolvidas para alcançar a integração dos domínios de dados referidos anteriormente, usando informações geradas por métodos microbiológicos convencionais e também pelas tecnologias de sequenciação de alto débito (vulgarmente referenciadas como *Next Generation Sequencing* ou NGS). Desde o seu desenvolvimento, as NGS têm gerado grandes volumes de dados, que se encontram disponíveis em repositórios públicos, permitindo aos utilizadores acederem e obterem dados genómicos curados necessários para o desenvolvimento de novas metodologias. Sendo assim, o objetivo principal do presente projeto trata-se do desenvolvimento de uma pipeline capaz de produzir informação filogenómica de microorganismos, a partir de dados obtidos por NGS, de modo a permitir a sua integração com dados geográficos. Esta integração permite gerar novos conhecimentos geofilogenómicos, permitindo novas abordagens para assegurar a rastreabilidade e a autenticidade. A pipeline desenvolvida obtém domínios de dados genómicos, geográficos e temporais, interrogando as bases de dados do *National Center for Biotechnology* (NCBI). Os dados genómicos são anotados e utilizados para efetuar a identificação do core genome e, em seguida, é efetuado *core genome multi-locus sequence typing* (cgMLST) para cada isolado. Após cgMLST, uma matriz de distâncias é gerada, seguida do cálculo da *minimum spanning tree* para a reconstrução das relações filogenómicas entre os isolados. A pipeline foi testada com 2 datasets distintos, contendo diferentes escalas geográficas e espécies de microorganismos. Os resultados sugerem que o método utilizado para a identificação de rotas de transmissão epidemiológicas putativas tem capacidade para discriminar casos epidemiologicamente relacionados, com diferentes escalas geográficas e espécies de microorganismos.

Palavras Chave: Georrastreabilidade; Sequenciação de Alto Débito; Transmissão Epidemiológica; Segurança Alimentar

Abstract

The increasing awareness of consumers over food products' authenticity has been steadily increasing throughout the years, despite the increased effort to improve food safety. These facts require the application of new techniques or methodologies to ease the rising concerns over authenticity. The application of traceability over food products in a distribution chain is one of said methodologies. The aim of traceability systems is to provide the ability to trace a product from its origin to the hands of the consumer. These systems can be further improved with the integration of other data domains such as geographical data or topographical data. Thus, the integration of conventional traceability data with geographical data is referred to as geotraceability. Over the years, tools have been developed to achieve the integration of the aforementioned data domains, using data generated by common microbiology methodologies and by Next Generation Sequencing (NGS) technologies, even though most of them only allow the visualization of multi-domain data. Since the development of NGS technologies, great volumes of data have been produced which are available for open use from several public repositories, allowing users to obtain curated genomic data necessary to develop new methodologies. Therefore, the main objective of this project is to develop a pipeline to produce phylogenomic information about food-borne pathogens, using NGS data, to integrate it with geographic data for the creation of novel geophylogenomic knowledge, useful for novel approaches for product traceability and authenticity. The developed pipeline performs user defined queries to databases of the National Center for Biotechnology (NCBI), retrieving genomic data (whole-genomes), geographical data (locations and coordinates) and temporal data (collection dates). The genomic data is annotated and used to perform the identification of the core genome, which will in turn, be used to perform core genome multi-locus sequence typing for every isolate. Afterwards, a distance matrix is generated, followed by the computation of a minimum spanning tree to reconstruct the phylogenomic relationships of the isolates. The pipeline was tested using 2 distinct datasets, differing in the geographic scale and pathogen species. The results suggest that the method used to identify putative transmission routes is able to discriminate epidemiologically connected episodes, with different geographical scale and pathogen.

Keywords: Geotraceability; High Throughput Sequencing; Epidemiologic Transmission; Food Security

Resumo Alargado

Num mundo cada vez mais tecnológico e industrializado, a questão da origem e autenticidade dos produtos alimentares é cada vez mais ponderada pelos consumidores. O aumento da sensibilização sobre este assunto resulta dos complexos canais de distribuição de alimentos, a produção em massa destes na indústria alimentar e a ocorrência de surtos de agentes patogénicos como, por exemplo, *Escherichia coli*. De modo a assegurar a autenticidade dos produtos alimentares, tornou-se necessário o desenvolvimento e aplicação de tecnologias capazes de certificar a veracidade das informações relacionadas com a autenticidade dos mesmos.

Uma das metodologias utilizadas para este efeito, é a implementação de sistemas de rastreabilidade. Estes sistemas conferem a capacidade para aceder a toda a informação relacionada com um certo produto em qualquer fase do seu ciclo de vida. Neste caso, considera-se que o ciclo de vida de um produto compreende o momento em que é gerado até chegar às mãos do consumidor. A rastreabilidade é um aspeto essencial da segurança alimentar numa cadeia de distribuição, devido á sua complexidade e por permitirem o consumo de produtos alimentares em todo o mundo. Permite também, reduzir os custos associados com a troca de informações e logística, uma maior precisão da informação e consequentemente uma maior satisfação do consumidor. Os sistemas de rastreabilidade podem ser complementados com a integração de dados convencionais de rastreabilidade com dados geográficos ou dados topográficos. Portanto, a sua integração é referida como rastreabilidade geográfica ou georrastreabilidade. A integração destes domínios de dados permite que os sistemas de georrastreabilidade funcionem como sistemas de alerta que possuem a capacidade de identificar produtos contaminados na sua origem ou, como sistemas de certificação que, ao provar a sua origem, aumenta o seu valor.

Para utilizar estes sistemas, é necessário a identificação do microorganismo, para tal, é necessário obter dados genómicos associados a este. Os dados genómicos podem ser obtidos através de métodos microbiológicos convencionais como a cultura de células em meio enriquecido, métodos baseados em ácidos nucleicos como a *polymerase chain reaction* (PCR) ou *multi-locus sequence typing (*MLST). O último método destaca-se devido á alta resolução e reprodutibilidade dos resultados, pois baseia-se na amplificação de um reduzido número de genes *housekeeping* (normalmente menos de 15), por PCR, para criar um perfil alélico único para cada isolado. Na última década, o desenvolvimento e melhoramento das tecnologias de sequenciação de alto débito (vulgarmente referidas como *Next Generation Sequencing* ou NGS), permitiu a sequenciação de genomas completos de microorganismos, gerando grandes volumes de dados devido ao baixo custo de sequenciação por base nucleotídica e elevada cobertura e qualidade. A sequenciação de genomas completos (frequentemente referida como *whole-genome sequencing* ou, simplesmente, WGS) tornou-se então a abordagem mais utilizada para obter dados genómicos com alta qualidade num curto espaço de tempo, de modo que é a abordagem mais adotada para a identificação de microorganismos durante surtos epidemiológicos. O volume de dados produzido por WGS são armazenados em repositórios públicos, como o *National Center for Biotechnology*, disponibilizando o seu uso à comunidade científica. Existem várias ferramentas desenvolvidas para aplicar a georrastreabilidade em diferentes áreas, utilizando dados disponibilizados nos referidos repositórios, mas apenas permitem a visualização dos domínios de dados referidos anteriormente. Tendo isto em conta, este projeto tem como objetivo desenvolver uma pipeline com capacidade para integrar dados genómicos, geográficos e temporais, permitindo a inferência de rotas de transmissão epidemiológica.

A *pipeline* desenvolvida neste projecto tem como passo inicial a obtenção dos domínios de dados, que se encontram presentes nas bases de dados *Nucleotide* e *BioSample* do *National Center for Biotechnology* (NCBI). A ferramenta *Entrez Programming Utilities* (E-utilities) desenvolvida pelo mesmo, foi utilizada para a obtenção dos dados, facilitando a interrogação das bases de dados, referidas anteriormente, bem como o processamento dos mesmos. Após a obtenção dos dados, numa fase de pré-processamento, estes são armazenados numa base de dados MySQL, de modo a consolidar a informação obtida de cada um dos domínios de dados e melhorar a sua organização.

A próxima fase da *pipeline* inicia-se com o processamento dos domínios de dados, nomeadamente o domínio de dados geográficos, obtendo as coordenadas geográficas, latitude e longitude, de cada um dos isolados através da geocodificação das localizações geográficas. Segue-se o processamento do domínio de dados genómicos, com a identificação do *core genome*. Para o processamento dos dados genómicos, são utilizadas ferramentas *open source*, a primeira, Prokka, realiza a anotação dos genomas dos isolados, cujos resultados são utilizados para a identificação do *core genome* utilizando Roary, a segunda ferramenta. Por fim, com a identificação do *core genome*, é efetuado o *core genome multi-locus sequence typing* (cgMLST), através do LOCUST. Esta ferramenta cria um ficheiro que contém a sequência de todos os alelos de cada isolado, gerando um perfil alélico único, utilizando um esquema de cgMLST personalizado. Neste caso, os genes presentes no *core genome* identificado anteriormente, foram utilizados para determinar o perfil alélico dos isolados. A fase final da pipeline consiste no cálculo de distâncias filogenómicas e na criação de uma *minimum spanning tree* que permitiu a visualização de rotas de transmissão epidemiológicas putativas. Dois *datasets* foram utilizados para testar a *pipeline*. O primeiro *dataset*, contém dados de *Escherichia coli* recolhidos em vários continentes (*dataset* internacional). O segundo *dataset*, contém dados de *Listeria monocytogenes* recolhidos numa única região geográfica (*dataset* nacional).

A análise de dispersão dos domínios de dados mostrou que existia uma alta variância dos dados, sugerindo a existência de casos epidemiologicamente relacionados, bem como casos isolados. Para averiguar a existência destes casos, efetuou-se uma análise de agrupamentos de dados (vulgarmente referido como *clustering*). Os resultados sugerem que o processamento de dados referido anteriormente, apresenta capacidade para discriminar internacionalmente eventos epidemiológicos. Tendo em conta estes resultados, foi escolhido um segundo *dataset*, contendo dados recolhidos numa única região geográfica, neste caso, no Canadá. Os resultados obtidos com este segundo *dataset* sugerem que o método também apresenta sensibilidade para discriminar eventos epidemiológicos regionais, portanto, tem capacidade não só para discriminar eventos epidemiológicos com escalas geográficas diferentes, mas também com dados de diferentes microorganismos.

Os resultados sugerem também que deve ser efectuada uma otimização da parametrização do processamento de dados e do cálculo da *minimum spanning tree*, de modo a melhorar a deteção de eventos epidemiológicos. Também se sugere uma otimização dos parâmetros do *clustering*, através da implemetação de algoritmos como o *two-step clustering*. Por fim, é necessário definir um conjunto de regras para identificar o tipo de evento de epidemiológico ocorrido. Este conjunto de regras irá permitir, futuramente, o desenvolvimento de modelos supervisionados para a classificação do tipo de eventos epidemiólogicos.