Universidade de Lisboa

Faculdade de Ciências
Departamento de Informática

NGSOnto - Proposta de uma ontologia para descrever o processo de Sequenciação de Alto Desempenho

Mickael Santos da Silva

2014

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Universidade de Lisboa

Faculdade de Ciências
Departamento de Informática

Dissertação

Orientado pelo Professor Doutor João André Nogueira Custódio Carriço e pelo Professor Doutor Francisco José Moreira Couto

2014

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Agradecimentos

Gostaria de agradecer ao meu orientador Prof. João Carriço pela sua ajuda, orientação no projeto e pela sua disponibilidade. Um agradecimento também ao meu co-orientador, Prof. Francisco Couto e aos professores Alexandre Francisco e Cátia Vaz, cujos conselhos e sugestões foram bastante úteis e valiosos para concluir este trabalho.

Um agradecimento também a toda a unidade de microbiologia molecular e infeção do Instituto de Medicina Molecular pelo bom acolhimento e apoio.

Ultimo agradecimento para os pais, irmãos, restante família e amigos sempre presentes.

Resumo em Inglês

With the appearance of new high throughput sequencing tecnologies, there has been a significant decrease in large scale sequencing costs through technologies known as "Next Generation Sequencing"(NGS), resulting in an increasing genomic information production. One of the successful application fields of this new sequencing technologies has been the Molecular Epidemology, where the main aim is at detecting and following bacterian outbreaks. Recent and known cases of such outbreaks where NGS technologies have proven their capacity, comparing with previous typing methods, are the cholera outbreak in 2010 at Haiti and the E.coli O104:H4 at Germany in 2011. However, to be able to compare and reproduce this data, it's necessary to keep the information of all processes, starting at the DNA Extraction process until the last final results analysis.

At this moment, the main NGS data search and insertion services, such as o Sequence Read Archive (SRA) and European Nucleotide Archive (ENA), present some limitations, namely at the annotation of performed laboratorial processes and consequent in silico data analysis processes.

Considering the previous facts, an ontology about the new sequencing generation was developed, the NGSOnto. This ontology was developed in order to describe the full workflow of a NGS sequencing process, reusing concepts of the Ontology for Biomedical Investigation (OBI) and others. The Web Ontology Language (OWL) was used to develop the NGSOnto, using the Basic Formal Ontology (BFO) 1.1 high level structure and saving the data trough the Resource Description Framework (RDF). In order to perform a concept proof of case of the NGSOnto, a REST Application Programming Interface (API) was developed, providing a mean to insert and access the data in a machine readable format, and a web interface, that uses the REST API developed, for users with less programatic skills.

Using NGSOnto, the capture of all the workflow process information provides a mean to ensure the reproducibility of the NGS process, through a controled and domain specific vocabulary, providing obvious benefits for scientific investigation areas using NGS technologies and certification of clinical results.

Keywords: Ontology, Whole Genome Sequencing, RESTFul Webservice, Microbial Typing

Resumo

Com o aparecimento dos novos métodos de sequenciação de alto desempenho, tem-se verificado uma diminuição de custos na sequenciação em larga escala de genomas através da tecnologia denominada de "Next Generation Sequencing" (NGS), resultado numa cada vez maior produção de informação genómica. Um dos campos onde a aplicação destas novas tecnologias de sequenciação, têm provas dadas de sucesso é em Epidemiologia Molecular, cujo objectivo é detectar e seguir surtos bacterianos. Casos recentes e mediáticos de surtos de estirpes bacterianas perigosas para a saúde pública, como o surto de cólera em 2010 no Haiti e E.coli O104:H4 na Alemanha em 2011, têm revelado as competências das tecnologias NGS relativamente às tecnicas de tipagem até então utilizadas. No entanto para os dados serem comparáveis e reprodutíveis, todo o processo desde a extração do DNA até há analise final de resultados necessitam de ser documentados.

Até ao momento, os principais serviços de pesquisa e inserção de dados de NGS, como o *Sequence Read Archive* (SRA) e *European Nucleotide Archive* (ENA), apresentam algumas limitações, nomeadamente no que se refere à anotação dos processos laboratoriais e em processos de analise *in silico* dos dados.

Neste trabalho foi desenvolvida uma ontologia relacionado com a sequenciação de nova geração, a NGSOnto. Esta ontologia foi construida de forma a descrever o fluxo de trabalho de um processo de sequenciação por NGS, sendo que esta ontologia reutiliza conceitos da Ontology for Biomedical Investigation (OBI) entre outras. Para construir a ontologia foi utilizada a Web Ontology Language (OWL), utilizando a estrutura da Basic Formal Ontology (BFO) 1.1 e guardando a informação através da Resource Description Framework (RDF). Foi também criada, como prova de conceito de aplicação da ontologia, uma interface programática REST de modo a possibilitar a inserção e consulta de dados num formato que sejam possíveis de ser lidos por máquinas , e uma interface web de fácil utilização para clientes com menos conhecimentos programáticos, que utiliza a REST API desenvolvida.

Com a anotação dos dados usando a NGSOnto, a captura do fluxo de trabalho de

todos os processos envolvidos permite assegurar a reprodutibilidade de todo o processo através da utilização um vocabulário controlado e especifico para o campo, com benefícios óbvios para investigação em diversas áreas que usam NGS e para validação e certificação de resultados em aplicações clinicas.

Keywords: Ontologias, Sequenciação do genoma completo, Serviço Web RESTFul, Métodos de tipagem microbiana

Índice

1 Introdução

As novas tecnologias de sequenciação de alto débito, Next Generation Sequencing (NGS) são hoje uma ferramenta que, apesar de algumas limitações, demonstram a sua utilidade em produzir grandes quantidades de dados de forma relativamente barata[1] em comparação a métodos anteriormente utilizados, como o método de Sanger[2]. Projetos como o "Human Genome Project"[3]com despesas a rondar os 3 mil milhões de dólares podem hoje, com as NGS, ser realizados com apenas alguns milhares de dólares, deixando bem clara a simetria acentuada entre a queda do preço do custo por base sequenciada e o aumento da qualidade/quantidade dos dados produzidos.

No campo de epidemiologia molecular esta tecnologia tem apresentado uma crescente importância na distinção de diferentes estirpes dentro da mesma espécie. A tipagem microbiana era inicialmente efetuada através de métodos baseados no fenótipo, como os antibiogramas ou a serotipagem. Mais recentemente começaram a ser usados alguns métodos baseados na genotipagem, como "*MultiLocus Sequence Typing*" (MLST) [4], que por sua vez começam a ser substituídos pela sequenciação completa do genoma, "*Whole Genome Sequencing*" (WGS), através da utilização das tecnologias NGS [5]. Estes estudos de WGS fornecem hoje a melhor resolução em estudos de epidemiologia, sendo esta demonstrada aquando da sua aplicação em recentes surtos, nomeadamente o surto de cólera no Haiti[6] e do caso mediático de *E.coli* O104:H4 na Alemanha [7].

Atualmente os dados obtidos através de processos de NGS são introduzidos em bases de dados internacionais como o "*Sequence Read Archive"* (SRA)[8]. Apesar do grande volume de dados, a informação existente além de não permitir uma reprodutibilidade dos processos é por vezes informação duplicada e desconexa entre si.

O facto de existirem diferentes tecnologias de NGS e múltiplas combinações de processos de preparação de DNA e de processos de análise, torna crucial a necessidade de guardar a informação referente a todos os passos do processamento de cada estirpe de modo a permitir a sua comparação e análise. Esta necessidade é um problema transversal em todas as áreas da ciência, constatável pela variedade de sistemas desenvolvidos e utilizados com base nesta temática [9 - 12] em diferentes áreas da ciência. O desenvolvimento destes sistemas

permite documentar detalhadamente o método utilizado (os seus inputs, os processos,etc) assim como a sua reprodutibilidade. A questão da reprodutibilidade é bastante referida pois é de uma importância inquestionável. Estes sistemas também têm uma enorme relevância a nível da microbiologia clínica, pois, normalmente, apenas alguns tipos dentro de uma dada espécie/subespécie bacteriana são responsáveis por doença ou quadros clínicos relevantes como por exemplo resistência a antibióticos.

Em Biologia, novo conhecimento é gerado a partir de conhecimento já adquirido, estando este conhecimento, muito rico em dados, armazenado em diversas bases de dados. A informação nestas bases de dados precisa presentemente de ser integrada de forma a derivar novos conhecimentos e, presentemente, a integração dos dados permanece, na maioria dos casos, um processo moroso e manual.

Um meio de captura de informação/conhecimento sobre determinado domínio e a sua disponibilização, tanto para leitura humana como para máquina, é conseguida utilizando ontologias [13]. Uma ontologia representa um determinado domínio através das entidades que constituem esse domínio e as relações que desenvolvem entre si. O conceito de ontologia permite também criar ligações entre diferentes ontologias, fator importante na utilização combinada de dados de domínios diferentes. Existem, neste momento, um grande número de ontologias aplicadas a ciências biológicas, sendo apontado como melhor exemplo de uma ontologia útil, com um grande número de utilizadores, com um grande alcance de espécies e granularidade, a "*Gene Ontology*" (GO) [14]. O crescimento deste número de ontologias biomédicas levou ao surgimento de ferramentas de pesquisa e navegação em ontologias como o BioPortal[15] e de iniciativas com princípios bem definidos de construção de ontologias, de forma a maximizar a interoperabilidade entre elas, como a *The Open Biological and Biomedical Ontologies Foundry* (OBO)[16].

1.1 Motivação

Até à data os principais serviços existentes de pesquisa e inserção de dados de NGS, *Sequence Read Archive* (SRA) e *European Nucleotide Archive* (ENA) [17] apresentam limitações consideráveis, apesar do enorme volume de dados neles depositados. Estas limitações prendem-se principalmente com o facto de os dados presentemente disponíveis não

permitirem uma reprodutividade dos resultados, não só em relação aos processos laboratoriais mas também processos realizados por uso de determinados softwares, nomeadamente o processamento das "reads" não tratadas até obter-se um conjunto de "contigs", no caso da assemblagem ser de-novo, ou de SNPs, no caso de ser realizado o mapeamento contra uma referÊncia. Recentemente a plataforma "EBI RDF Platform"[18], desenvolvida pelo grupo responsável pelo serviço ENA, tem iniciado um processo de desenvolvimento de ferramentas para uma migração do formato relacional dos dados para dados representados no formato RDF(abordado na secção 4.1.1), permitindo um enriquecimento semântico dos mesmos. É o caso do serviço Biosamples [19] que agrega a informação de amostras presentes em várias bases de dados pertencentes ao mesmo grupo [17; 20;21].

Apesar dos novos desenvolvimentos na organização e no formato dos dados destes serviços, a falta de informação sobre os processos realizados continua a ser uma limitação. Sem esta informação não será possível uma verdadeira comparação dos resultados, facto muito relevante no campo da epidemiologia molecular, onde pequenas variações do genoma reportado podem levar a interpretações diferentes das relações entre outras estirpes bacterianas. Todos os processos envolvidos num estudo NGS possuem parâmetros que terão influência do resultado final e que podem levar à introdução de erros ao longo de todo o processo, sendo que esses parâmetros terão de ser guardados para permitir comparações e a reprodutibilidade de dados.

A necessidade de guardar dados de NGS é uma realidade aceite, pelo que iniciativas como o "*Global Microbial Identifier*" [22] (GMI) começam a surgir, neste caso com o objetivo de criar uma plataforma para guardar dados WGS de microorganismos e permitir a sua comparação à escala global. Esta plataforma iria comparar dados existentes na plataforma com dados de amostras de pacientes por métodos BLAST, obtendo-se de forma rápida dados de extrema importância como a localização de estirpes idênticas e as melhores formas de tratamento. Esta plataforma iria possibilitar o início de tratamentos personalizados, assim como a criação de um sistema de vigilância de surtos, à escala global, de doenças infecciosas e de microorganismos em geral.

Na base da criação de uma plataforma estará sempre presente um ou mais repositórios de dados e, no caso de este ser criado de raiz, a modelação do repositório de dados é um passo crucial para o sucesso de toda a plataforma daí em diante.

O sucesso das tecnologias de armazenamento de dados tem levado a um crescimento exponencial na quantidade de bases de dados existentes para os mais variados problemas/situações. Esta explosão de bases de dados levou a que uma grande parte delas não chegasse a conhecer sucesso, ficando perdidas na imensidão de dados que compõem a world wide web, assim como os dados nelas presentes. Uma forma de manter os dados utilizáveis, mesmo sem interferência humana, é tornar os dados possíveis de ser lidos por máquinas. Formatos como RDF, XML e JSON apresentam estas características, sendo que o formato RDF apresenta vantagens na representação de dados com informação semântica e na gestão de conhecimento. A comunicação entre os dados e o utilizador (máquina ou humano) deverá ficar a cargo de um serviço web. A arquitetura REST(Representational state transfer)[23], que tem emergido como o modelo predominante de serviços web, confirmado pelo seu uso pelas gigantes tecnológicas Google [24], Ebay [25] e Paypal [26].

1.2 Objetivos

Os objetivos deste trabalho são:

- desenvolver uma ontologia que permita mapear "pipelines" de todos os processos envolvidos numa sequenciação por NGS desde a extração de DNA até à montagem do genoma parcial, de modo a permitir a sua futura reprodutibilidade;
- desenvolver ligações com a ontologia TypOn e reutilizar ao máximo entidades já utilizadas em ontologias da OBO (e.g. OBI);
- permitir a integração de dados provenientes de fontes diversas como o Sequence Read Archive (SRA)/European Nucleotide Archive (ENA);
- desenvolver um web service RESTful e uma plataforma web de fácil utilização que permita aos utilizadores mapear e obter informação sobre os dados introduzidos.

1.2.1 Contribuições

As contribuições deste trabalho são as seguintes:

- Uma ontologia, NGSOnto, que permite mapear um processo completo de NGS, que até ao presente não existe. Esta ontologia foi construída com o máximo de conceitos externos possíveis e encontra-se estruturada segundo a BFO 1.1 ;
- Uma REST API que permite a ligação com um repositório de triplos onde foi aplicada a NGSOnto;
- Uma interface de fácil utilização para o cliente que utiliza a ontologia desenvolvida e a REST API, demonstrando a utilidade de ambas as contribuições aplicadas num só sistema.

1.3 Organização do documento

Este documento está organizado da seguinte forma:

- Capítulo 2 – Neste capítulo são apresentados todos os processos realizados, com algum detalhe, num processo de NGS genérico, a importância das tecnologias NGS em epidemiologia, a forma como as ontologias estão a ser utilizadas no domínio biomédico e um sub-capítulo sobre dois dos mais conhecidos sistemas de gestão de "workflows" no domínio biomédico;
- Capítulo 3 – Descrição da principal parte do trabalho desta dissertação. Os métodos e tecnologias utilizadas para o desenvolvimento da ontologia, assim como uma descrição pormenorizada da ontologia desenvolvida.
- Capítulo 4 – Descrição da segunda parte do trabalho desta dissertação. Os métodos e tecnologias utilizados para o desenvolvimento do serviço web e uma descrição resumida da interface para o cliente.
- Capítulo 5 e 6 – Discussão final sobre o trabalho desenvolvido e conclusões do mesmo.

2 Trabalhos Relacionados

2.1 Processo de "Next Generation Sequencing"

A primeira geração de métodos de sequenciação iniciou-se com o aparecimento do método de Sanger, realizado manualmente através de geles. Este método foi dominante até ao surgimento das primeiras tecnologias de sequenciação automática de 2ª geração, baseados em métodos de capilaridade. O desenvolvimento de novas tecnologias veio substituir as tecnologias de 2ª geração, ao contrário dos métodos anteriores que têm permanecido praticamente inalterados [27;28], surgindo a 3ª geração de tecnologias de sequenciação, também conhecida por *Next Generation Sequencing* (NGS) .

As tecnologias de NGS são usadas através de estratégias variadas que são constituídas por combinações de métodos de extração de DNA das amostras, preparações de bibliotecas, sequenciação das bibliotecas e posterior análise das "reads" produzidas com métodos de alinhamento a referência ou assemblagem "de-novo".

Cada tecnologia de sequenciação tem as suas características próprias, dependentes da companhia pelas quais foram desenvolvidas[29].

Diferentes combinações de protocolos específicos distinguem as diferentes tecnologias entre si, o que leva a uma importância na definição de parâmetros de controlo de qualidade final, como a "qualidade por base" pós sequenciação [1].

Nas seguintes secções são apresentados os passos seguidos num processo genérico de NGS. A sua execução numa ordem sequencial representa um pipeline, esquematizado na seguinte figura.

Figura 1. Esquematização dos processos envolvidos num pipeline de NGS.

2.1.1 Extração de DNA e Preparação de Bibliotecas

Um protocolo de NGS é geralmente iniciado com a extração de material genómico de uma amostra. Esta extração é geralmente efetuada através de kits especializados ou protocolos pré-definidos, cuja função é quebrar as células, libertando o DNA total e posteriormente purificá-lo. Neste passo de purificação o objetivo é remover todas as proteínas e pedaços de componentes celulares, de modo a obter DNA estável para o passo seguinte.

A partir do material é efetuada a preparação de bibliotecas, atualmente obtidos através de fragmentação aleatória do DNA em fragmentos pequenos. Este passo é executado devido à necessidade de obter uma representação de uma fonte de material nucleico, do genoma em estudo, que não tenha sofrido qualquer viés. Existem comercialmente disponíveis que permitem a realização de alguns destes passos em conjunto e de forma rápida, por exemplo o "Nextera DNA Sample Preparation Kit" [30].

Na maioria das tecnologias encontra-se a necessidade de prender estes fragmentos num suporte ou numa superfície sólida, permitindo deste modo um número astronómico de reações de sequenciação simultâneas relativamente a cada DNA molde. No caso de ser usado um suporte, o método mais comum é o reação em cadeia de polimerase, *Polimerase Chain Reation*(PCR), por emulsão (emPCR) [31], utilizado em 454 e Ion Torrent, enquanto que no caso de uma superfície sólida será o método de amplificação em fase sólida [32](Illumina).

O emPCR é usado para preparar DNA moldes (template) num sistema fora da célula, com a vantagem de evitar a perda de sequências genómicas aleatoriamente. Neste método é criada uma biblioteca e na sua extremidade vão ligar-se adaptadores com primers universais de modo a permitir a amplificação de genomas complexos com primers comuns de PCR. Estando os adaptadores nas extremidades dos fragmentos da biblioteca, a molécula de DNA separa-se em cadeia simples e é capturado em esferas envoltas cada uma numa única emulsão independente que contém todos os reagentes necessários para ocorrer um PCR. As esferas emPCR vão então apresentar a sua superfície coberta por fragmentos de cadeia simples, amplificadas e isoladas entre si, sofrendo processos químicos diferentes dependendo de onde serão imobilizados [1] *Polonator* [33;34] (Life/APG) ou *PicoTilterPlate* [35] (Roche/454), representativas de duas das companhias que operam no mercado comercial de sequenciação, a Life Technologies e a Roche). No caso do método utilizado ser o de amplificação em fase

sólida as bibliotecas são amplificadas numa placa de vidro, dando-se a fixação de primers direto e inverso na placa onde se vão ligar os fragmentos, fornecendo desta forma extremidades livres onde os primers de sequenciação poderão hibridizar para iniciar a reação química de NGS. Este método foi desenvolvido e utilizado pela companhia Illumina. Os métodos emPCR e amplificação em fase sólida têm como base a amplificação clonal dos fragmentos. Como todos os métodos, apresenta vantagens e desvantagens, sendo que a desvantagem principal se prende com a necessidade da realização de reações de PCR o que, por si, poderá levar a criação de mutações assim como a dificuldade em lidar com sequências ricas em AT e GC, mudando a frequência e abundância de alguns fragmentos de DNA que existiam antes da amplificação [36].

2.1.2 Métodos de Sequenciação de Alto Débito

O processo de sequenciação consiste na determinação da sequência de DNA de uma amostra. A maior fatia do mercado comercial de sequenciação, presentemente dominada pelas companhias Illumina e Roche, usa tecnologias que têm como base uma sequenciação com métodos óticos (Illumina e Roche).

No caso da tecnologia Illumina o método base é designado por terminação cíclica reversível (CRT) que, como indica o nome, utiliza de forma cíclica terminações reversíveis a cada incorporação de nucleótidos. Este processo é iniciado pela adição de uma base modificada com fluorescência à molécula DNA molde, seguida da adição de um terminador reversível que impede a ligação de novas bases, captação da fluorescência da base adicionada, remoção do terminador reversível e repetição do ciclo.

O método de pirosequenciação é utilizado pelas máquinas da Roche e tem como base a deteção de luz visível, obtida pela libertação de pirofosfato inorgânico e que é convertido proporcionalmente em luz visível através de reações enzimáticas [37] [38]. Ao contrário de outros métodos que utilizam nucleótidos modificados , a pirosequenciação manipula a DNA polimerase, sendo que a cada ciclo é introduzida uma quantidade limitada de uma única base, que, ao ser incorporada, pausa a polimerase e fará com que, através de uma série de reações enzimáticas, seja detetada a luz visível convertida da libertação de pirofosfato inorgânico. Uma vez que se adiciona apenas uma base única a cada ciclo é possível saber-se desta forma

se a base foi ou não incorporada, qual a base e em qual poço, registando os resultados numa série de picos chamado fluxograma [36].

A tecnologia de sequenciação de uma única molécula em tempo real (SMRT) [39], foi desenvolvida pela Pacific Biosciences e apresenta um método semelhante aos anteriores mas apenas focado numa única molécula, sem a necessidade de preparação de bibliotecas. Esta tecnologia permite a leitura da fluorescência única por cada base a cada momento da sua incorporação pela DNA polimerase, usando a cadeia a ser sequenciada como DNA molde.

Os métodos anteriormente referidos têm como base processos óticos, no entanto aproximações que usem processos não óticos já se encontram disponíveis e em desenvolvimento. O caso da tecnologia Ion Torrent [40], com o seu método que utiliza circuitos integrados e sensores ion-sensitive field-effect transistor (ISFET), é o que apresenta maior sucesso. Neste sistema são também usadas esferas cobertas de fragmentos, que vão ser depositadas em poços, onde irão ser fornecidos de forma iterativa um determinado nucleótido. Este nucleótido é incorporado, quando complementar, o que resulta na hidrólise do grupo trifosfato do nucleótido incorporado que, por sua vez, resulta na libertação de um protão por cada nucleótido. Esta libertação de protões leva a uma mudança de pH da solução envolvente, detetada no fundo de cada poço, convertida em voltagem e digitalizada eletronicamente [40].

Os processos de sequenciação produzem uma grande quantidade de “reads”, pedaços de sequências de caracteres, representativas do DNA da amostra .

A seguinte tabela resumida [41] permite a comparação das tecnologias referidas anteriormente.

Fabricante (Máquina)	Método Sequenciação	Média de tamanho de reads (high-end/benchtop)	Tempo de uma corrida(high-end/benchtop)	Gb por corrida (high-end/benchtop)
Roche (454GSFLX+/454 GSJunior)	Pirosequenciação	700/450 bases	23h/10h	0.7/0.45
Life Technologies (Ion Personal Genome 314)	Deteção de protões	-/100 a 400	-/3h	(dependente do chip utilizado) 0.1 até 2
Illumina (HiSeq 2500/MiSeq)	Terminador Reversível	2x150/2x250	7h-40h/5h-65h	10-180/0.3-15gb
Pacific Biosciences (PacBio RS)	Sequenciação em tempo real	3000/-	20min/-	3(dia)/-

Tabela.1 Resumo das características das diferentes tecnologias de sequenciação atualizados com valores de 2014.

A tecnologia 454 apresenta como vantagem um maior comprimento das reads, relativamente às duas outras tecnologias que utilizam métodos óticos, no entanto para organismos com genomas relativamente pequenos a tecnologia Ion Torrent e Illumina apresenta vantagens, sobretudo em termos de custo. A tecnologia Illumina permite um menor esforço relativamente a trabalho de bancada comparativamente aos restantes, assim como uma menor taxa de erro no tratamento de sequencias de homopolímeros. Por utilizar uma abordagem de molécula única a tecnologia PacBio apresenta os melhores resultados em termos de tamanho de reads, mas com a desvantagem de ter uma taxa de erro algo elevada.

Estas diferenças tornam indispensável uma reflexão antes de serem usadas, de forma a perceber qual a que terá maior vantagem relativamente ao tipo de amostra inserida e das características dos resultados que a máquina produz.

2.1.3 Análise de “reads” (Baseado em Referência/Montagem genómica “de-novo”)

Os processos de alinhamento e montagem consistem na fusão de fragmentos (reads), que se sobrepõem entre si pela comparação das sequências, em fragmentos maiores, chamados “contigs”, de modo a tentar reconstruir a sequência de DNA original. Estes processos de comparação podem utilizar um genoma de referência, de forma a facilitar a comparação dos fragmentos, conhecida por alinhamento baseado em referência ou mapeamento. No caso de não ser utilizada qualquer genoma de referência para a comparação é chamado de alinhamento “de novo”[42] [43]. Para ambos existe uma considerável variedade de algoritmos disponíveis, (e.g. BWA-SW [44] para um alinhamento com referência) cada uma com as suas vantagens e pequenas modificações adaptados ao estudo em questão. Existem também uma série de diferentes softwares, por exemplo BWA [45] para alinhamentos com referência e SPAdes [46] e Velvet [47] para alinhamentos “de-novo”, cada um com diferentes parâmetros. Neste processo as decisões serão feitas tendo em conta a aplicação biológica a que se destina, o organismo e o tempo, sendo que o alinhamento em relação ao genoma de referência é relativamente mais rápido podendo ou não devolver melhores resultados que um alinhamento “de-novo”.

Na realização destes processos de alinhamento e montagem, são normalmente utilizados “reads” em formato FASTQ[48]. Este formato é amplamente usado e contém não só a informação sobre o código da sequência, ou seja a base de cada nucleótido, mas também a informação sobre o “quality score” de cada base, ou seja, o grau de confiança de cada base.

A realização destes processos possibilita uma comparação de sequências entre diferentes amostras, permitindo análises de “Single Polimorfism Nucleotide” (SNP), no caso de mapeamentos com referência, ou comparações de genes, no caso de alinhamentos “de-novo”.

2.2 NGS Aplicado na Tipagem Bacteriana Molecular

Atualmente existem vários métodos de tipagem, ou seja de identificação de diferentes

tipos de estirpes dentro de uma mesma espécie bacteriana. Infelizmente não está definido nenhum método único que seja o ideal, sendo que cada método apresenta as suas vantagens e desvantagens que deverão ser ponderadas consoante resultado que se pretende obter [5]. Atualmente começa a ser reconhecido que, para melhor caracterizar o surto e a estirpe bem como reforçar os sistemas de vigilância de saúde publica, são necessários métodos com maior resolução que utilizem "Whole genome Sequencing*"* (WGS) das estirpes bacterianas, usando as tecnologias NGS.

O surto de E.coli (EHEC) O104:H4 entero-hemorrágica multi-resistente, ocorrido entre Maio e Junho em 2011na Alemanha [49;50], é o caso mais discutido onde se prova a importância da aplicação de uma abordagem WGS em epidemiologia. A infeção por este agente patogénico provocou um surgimento de síndrome hemolítico-urêmico nos pacientes, resultando em 46 mortes e mais de 4000 doentes [51], situação registada apenas uma vez em 2001 na Alemanha [52]. Utilizando o método convencional de tipagem MLST, que tem como base as sequências de fragmentos internos de 7 "housekeeping genes", a comparação dos isolados dos dois casos detetou o mesmo SPAtype 678, o que indica que as estirpes isoladas deveriam ser idênticas. Realizando uma sequenciação do genoma de ambas as estirpes por NGS e uma comparação de "*Single Polimorfism Nucleotide*" (SNP) para um genoma de referência, foi revelada uma diferença substancial entre os conteúdos cromossomal e plasmídico [50]. Um estudo [53] confirmou estes resultados, mostrando que o método tradicional MLST não consegue encontrar relações com descriminação suficiente entre isolados geneticamente relacionados mas que diferem em potencial patogénico, sendo que usando sequenciação por NGS encontraram, em 167 genes, provas de recombinação homologa entre isolados pouco aparentados. O grande número de genes obtido por estudos WGS permitem a realização de novos métodos de MLST que englobem um maior número de genes a comparar, começando a surgir uma série de métodos de análise como o rMLST[54] ou o extended MLST [55].

Outro exemplo de utilização de WGS em epidemiologia é o caso do surto de cólera no Haiti em 2010 e o processo para a identificação da origem do surto [56]. Neste estudo foi utilizada a tecnologia de sequenciação desenvolvida pela companhia Pacific Biosciences, sequenciando-se amostras recolhidas localmente e estirpes referentes a um surto de cólera na América Latina bem como duas estirpes referentes a surtos no sul da Ásia. Foram comparados

os resultados não apenas entre si mas com mais sequencias genómicas parciais resultantes de 23 diferentes estirpes. Por fim foi concluído que a estirpe responsável pelo surto de cólera no Haiti, era praticamente idêntica a uma variante de uma estirpe característica do sudeste asiático, levantando a suspeita de que este surto teria tido como causa de dispersão a chegada de equipas de auxílio daquela zona do globo.

O uso de NGS em epidemiologia molecular é uma realidade que já mostrou os seus benefícios e potencialidades, sendo de esperar que o desenvolvimento das técnicas continue, com melhores resultados e a preços mais acessíveis a qualquer laboratório, generalizando o uso destas técnicas possivelmente em conjunto com os métodos mais antigos de tipagem, de forma a complementarem-se.

No entanto, a capacidade de estandardizar e anotar as análises efetuadas tem-se revelado como o principal fator limitante para a sua aplicação ser mais difundida, à medida que os preços de sequenciação por NGS vão diminuindo.

2.3 Ontologias e a sua Aplicação na Biologia

Ontologia é um conceito filosófico originário da Grécia antiga e é etimologicamente composto por duas palavras gregas, *onto-* que significa "ser" ou "aquele que é" e *-logia* significado de "ciência, estudo, teoria", ou seja, é o estudo filosófico do ser, como algo que existe, e das suas relações. Mais recentemente, Gruber definiu uma ontologia como sendo uma "especificação de uma conceptualização, usados para ajudar a partilhar conhecimento entre programas e humanos" [57], definição bastante aceite e referida presentemente para o conceito de ontologia. Conceptualismo é a organização de conhecimento sobre um determinado universo em termos de entidades, mais propriamente as coisas, as relações entre si e as "regras" que definem essas relações. A especificação trata da representação concreta dessa conceptualização, assim como a sua codificação numa linguagem representativa de conhecimento, de forma a criar um vocabulário uniforme com estrutura semântica para por sua vez permitir ser trocada informação sobre o domínio pré-conceptualizado.

Com o aumento da quantidade e dimensionalidade de dados biológicos gerados, a necessidade de encontrar novas estratégias para integração de dados, novas linguagens de

procura assim como novas ferramentas de acesso e inserção de dados, tornou-se essencial para a investigação em biologia. A utilização de ontologias disponibilizou não só termos e definições textuais mas também uma estrutura básica, estrutura esta que é geralmente expressa numa linguagem formal com base lógica, a "*Web Ontology Language*" (OWL) [58]. Através do uso desta linguagem, o conhecimento sobre o domínio é expresso consoante o modelo axiomático [59], segundo o qual os axiomas são declarados e as consequências desses axiomas são inferidas através de regras de inferência [60] .

A aplicação destes fundamentos proporciona uma documentação expressiva e possível de ser lido por máquinas. Permite também a verificação da consistência do modelo de dados [61], a captura de dados complexos e procura por inferência automática [62], a integração de várias ontologias [63] e a diminuição do custo de desenvolvimento e manutenção da ontologia [64;65].

A estrutura em grafo das ontologias biomédicas tem também permitido o desenvolvimento de ferramentas de análise posteriores, como por exemplo o "*Gene Set Enrichment Analysis*" (GSEA) [66], que utiliza a estrutura em grafo da GO e a anotação de genes com termos GO. O uso de análise semântica também é possível em estruturas em grafo, através da aplicação de métricas numa ontologia de forma a comparar a semelhança entre dados anotados e as classes da ontologia [67].

Por todos os motivos anteriormente enunciados, as ontologias nas áreas Biomédicas têm registado um contínuo crescimento, assim como a sua utilização e desenvolvimento de ferramentas para produzir informação e ciência. Estas ontologias deverão ser alvo de avaliações constantes, através de critérios de avaliação quantitativos e qualitativos, sendo esta avaliação feita não sobre a ontologia como entidade isolada, mas aquando da sua aplicação, avaliando-se assim um sistema no seu todo [68].

2.4 Sistemas de gestão de "Workflows"

Um "workflow" consiste num padrão de atividades, produzido por uma organização de recursos em processos que transformam materiais e/ou informação, e a necessidade de

guardar esta informação é um problema transversal em todas as áreas da ciência. Por esse motivo tem-se verificado uma proliferação de sistemas de gestão de "Workflows", que têm como objetivo guardar um determinado "workflow" de forma a torná-lo reprodutível, como o projeto Galaxy [69] e o projeto Taverna [70]. O projeto Taverna tem-se tornado popular no domínio da bioinformática nas áreas de proteómica e transcriptómica, assim como "text data mining", no entanto o projeto Galaxy foi especificamente desenvolvido tendo o domínio biológico como alvo.

2.4.1 Projeto Taverna

O projeto Taverna consiste num sistema "drag and drop" desenhado de forma a combinar serviços web e/ou ferramentas locais para formar workflows de análise complexa. Uma vez construídos estes "workflows", podem ser guardados, partilhados, reutilizados e executados. Repositórios públicos de "workflows" como o myExperiment (http://www.myexperiment.org) [71] estão disponíveis para guardar e partilhar workflows de variados sistemas, incluindo o Taverna. Na prática, a maioria dos "workflows" mapeados no Taverna, são compostos por uma mistura de serviços web distribuídos e scripts locais, o que permite a execução destes "workflows" seja processada maioritariamente de forma remota pelos fornecedores dos serviços web. Esta característica é uma grande vantagem, pois permite uma execução independente das infraestruturas técnicas à disposição, diminuindo custos de equipamento/manutenção e possíveis problemas de instalação de software ou de descarregamento de bases de dados. O facto de usar serviços web remotos torna-o dependente desses mesmos serviços, o que poderá ser um problema, no entanto o facto de haver uma certa redundância nestes serviços web poderá minimizar este problema. Serviços como o BioCatalogue (http://www.biocatalogue.org) [72] são importantes na medida em que fornecem informação um leque de informações sobre os serviços, serviços semelhantes, tipo de serviço, licenças, entre outros metadados.

A maioria dos projetos que ambicionam guardar um "workflow", bem como os metadados associados, são orientados mais especificamente aos passos do "workflow" onde são utilizadas ferramentas informáticas. Este é também o caso do projeto Taverna que, na área

da bioinformática e mais especificamente sobre o tema da sequenciação por NGS, tem como alvo apenas o pós processamento de dados produzidos em qualquer momento pós-sequenciação. Isto poderá levar a que o produto final seja pouco atrativo/intuitivo para utilizadores alvo, muitas vezes na área das ciências da vida e com algumas limitações técnicas a nível informático, que poderiam apresentar alguma dificuldade em interpretar "workflows", como por exemplo o apresentado no guia de iniciação rápida (http://www.myexperiment.org/workflows/3369.html) . Ao tornar o projeto Taverna genérico para qualquer tipo de "workflow" em qualquer área, é provocada uma diminuição da sua especificidade. Por outro lado, uma vez que um workflow de sequenciação completo é composto também por uma grande parte de processos efetuados manualmente em laboratório e que irão refletir-se na qualidade dos dados produzidos pela sequenciação pela máquina NGS, o mapeamento destes passos será também de extrema importância, não só para consulta e reprodutibilidade, mas também para possíveis análises estatísticas de relação entre os passos do "workflow" e a qualidade dos dados obtidos.

2.4.2 Projeto Galaxy

O projeto Galaxy é uma plataforma web aberta que permite capturar informação de uma análise computacional completa de uma forma fácil para utilizadores sem conhecimentos de programação . Esta captura de informação permite posteriormente a sua reprodutibilidade e a sua compreensão. Por ser uma plataforma aberta é permitido aos utilizadores partilhar as suas análises e informação complementar com outros utilizadores.

Apesar de ter sido desenvolvido para um público da área Biomédica, este apenas captura a informação sobre a análise computacional, deixando de fora a parte dos processos realizados em laboratório. Esta particularidade torna o projeto Galaxy bastante semelhante com outros sistemas de gestão de "workflows", como por exemplo o Taverna, uma vez que este irá guardar a informação sobre as ferramentas informáticas utilizadas e os dados resultantes entre cada processo realizado. Apesar de tudo, a sua especificidade permite uma boa definição dos tipos de dados, as suas transformações e das ferramentas informáticas utilizadas.

3 Criação da NGSOnto

Neste capítulo são descritos os métodos utilizadas para construir o modelo que constitui a ontologia NGSOnto, o modelo final obtido utilizando os métodos previamente apresentados e uma exemplificação de como o conhecimento é representado utilizando o modelo obtido.

3.1 Métodos

A construção do modelo seguiu os seguintes passos:

- Identificação do software com os seguintes requisitos: grátis e código aberto, com capacidade produzir uma ontologia em formato OWL [58], para ser aplicada num servidor com repositório de triplos;
- Pesquisa, na bibliografia científica, dos processos envolvidos num processo de sequenciação por NGS e a forma de os representar de modo a permitir as variações de procedimentos que cada utilizador poderá ou não decidir executar;
- Organização da ontologia com base numa ontologia de alto nível, a Basic Formal Ontology 1.1 (BFO) [73];
- Pesquisa, utilizando a plataforma "BioPortal"[74], dos possíveis conceitos já definidos em outras ontologias relevantes, de preferência que façam parte da The Open Biological and Biomedical Ontology Foundry (OBO)[75], entre outras possíveis ligações a fontes de dados importantes (EBI/SRA);
- Avaliação da consistência do modelo utilizando "reasoners";
- COnstrução de um pequeno exemplo representativo das capacidades de representação de informação com a ontologia desenvolvida.

3.2 Tecnologias Utilizadas

A NGSOnto foi desenvolvido na linguagem Web Ontology Language (OWL) [58], que permite a introdução de axiomas através dos quais se pode inferir nova informação. A OWL é uma linguagem formal estendida da especificação RDF, sendo que a NGSOnto está representada em RDF/XML. Por não ter sido criado nenhum axioma para a NGSOnto, ou seja apenas os axiomas já existentes nas ontologias importadas são utilizados, não é incluído nenhum sub-capítulo sobre a OWL mas antes um sub-capítulo sobre RDF.

3.2.1 RDF

Resource Description Framework (RDF) é uma representação/especificação definida pelo World Wide Web Consortium(W3C) [76] , desenvolvida inicialmente como um modelo de dados para metadados. No entanto, generalizando o conceito de "Web resource", este tem sido utilizado principalmente para representar informação sobre recursos disponibilizados na Web. O alvo principal do uso da RDF está nas situações onde é necessário que a informação seja processada por aplicações, uma vez que providencia uma "framework" comum para expressar a informação entre as aplicações, sem perda de significado, ao invés de servir apenas para apresentação e leitura do utilizador humano. Uma vez que os dados são representados utilizando o padrão estandarte do RDF , as aplicações podem tirar grande vantagem da utilização de ferramentas de processamento de RDF existentes, permitindo no fim que a informação seja universalmente disponível a todas as aplicações e não só às aplicações para as quais foi criada originalmente.

A RDF possui uma sintaxe abstrata, refletindo o seu modelo de dados simples baseado em grafos, e com base numa semântica formal que providencia deduções bem fundamentadas nos dados em RDF. O modelo de data em grafo tem como base a coleção de triplos, um grafo RDF, sendo que cada triplo consiste num sujeito, um predicado e um objeto. A declaração de um triplo RDF contém uma determinada relação, indicada como predicado, que forma uma ligação entre um sujeito e um objeto, sendo que, o predicado, está sempre direcionado para o objeto. Para identificar os nós, os recursos declaradas nos triplos, são utilizados Uniform

Resource Identificator's (URI), definidos por T.Berners-Lee [77] como uma sequência de caracteres com uma sintaxe restringida, que pode atuar como uma referencia a algo que tem identidade, neste caso para identificar os recursos Web. De forma a disponibilizar uma sintaxe para escrever e trocar grafos RDF, é disponibilizada uma sintaxe normativa em XML [78] chamada RDF/XML [79]. Em RDF/XML, as sequências "sujeito predicado objeto, sujeito predicado objeto,..." serão declaradas como elementos dentro de elementos, onde o sujeito (nó) inicial, será o elemento mais externo, e a próxima relação "predicado objeto" será o elemento descendente.

```
<rdf:Description
rdf:about="http://www.NGSOnto.org/study/StudyTest">
        <ngs:title>Study Test</ngs:title>
        <ngs:owner>Jose</ngs:owner>
</rdf:Description>
```

O seguinte pequeno exemplo de RDF permite uma melhor explicação de como é representada a informação. Neste exemplo é descrita informação sobre o recurso "http://www.NGSOnto.org/study/StudyTest". Este é o sujeito, e o predicado está indicado no elemento interior seguinte, neste caso "title", enquanto que o objeto está representado pela string "Study Test". Desta forma representamos a informação de que o estudo definido com um URI único tem um título e esse título é "Study Test". O elemento seguinte, uma vez que se encontra dentro apenas do mesmo sujeito, refere que a informação é sobre o mesmo sujeito. A Tabela 2 representa os triplos de uma forma esquematizada.

Sujeito	**Predicado**	**Objeto**
http://www.NGSOnto.org/study/StudyTest	ngs:title	Study Test
	ngs:owner	Jose

Tabela 2. Representação esquematizada dos triplos do excerto de RDF apresentado anteriormente.

3.2.2 Protégé

O software Protégé [80] é um software grátis e de código aberto, que permite a edição e criação de ontologias em formato OWL, OBO, entre outros. Além de uma vasta rede de utilizadores e uma forte comunidade ativa, este software baseado na linguagem JAVA permite também estender as suas funcionalidades através de uma grande variedade de "plug-ins" que poderão ser ou não construídos por terceiros. Recentemente foi disponibilizado e encontra-se ainda em desenvolvimento um serviço online que tem como objetivo capturar as funcionalidades do software original e disponibilizá-lo online sem qualquer instalação.

Tendo em conta os requisitos inicialmente definidos como necessários e as restantes funcionalidades deste software, foi decidido a utilização do mesmo para a construção e edição da ontologia.

3.3 Descrição do Desenvolvimento da Ontologia

A ontologia tem como base a linguagem inglesa devido à sua globalidade permitir uma melhor base semântica. De notar que o nome das entidades apresentados nos parágrafos seguintes foram, quando originárias de outras ontologias, obtidos após a construção da ontologia. No entanto, para facilitar a compreensão, são mencionadas com o seu nome atribuído até ao momento pelas ontologias a que pertencem.

Escolhidas as ferramentas informáticas a utilizar, o próximo passo consistiu na recolha de informação de um processo genérico de uma sequenciação por NGS de forma a organizar os processos envolvidos num "workflow".

Qualquer processo NGS tem de partir de uma amostra biológica recolhida de um organismo, por isso será natural que qualquer "workflow" seja iniciado por uma extração de material genómico (DNA) que será designado por "DNA Extraction". A partir desta extração de material genómico procede-se um passo de grande importância, se não crucial, para um bom resultado da sequenciação NGS, a preparação de bibliotecas, indicada na ontologia por

"Library Preparation". Preparada a biblioteca é iniciado o processo de sequenciação, "DNA sequencing". Esta irá ter uma óbvia relação com um "DNA sequencer", entidade representativa de uma máquina NGS, devido à variedade existente entre o output e diferentes máquinas, softwares e versões. Após a sequenciação é obtido um output que será processado em duas fases, que foram consideradas distintas para facilitar a sua compreensão, um processo de corte, "Sequence cutting" (também conhecido por trimming) e um processo de filtragem, "Filtering", agrupados numa entidade "DataProcessing". O processo de "Trimming" consiste na remoção de dados que não são relevantes de serem considerados, como possíveis erros ou remoções de adaptadores. Este processo providencia novos dados a serem filtrados no processo de "Filtering", que serão por exemplo filtrados por tamanhos mínimos das "reads" ou qualidade das "reads". Por fim, após a realização ou não de processos de corte e filtragem, é realizado a assemblagem, representada pela entidade "Sequence assembly", que está subdividida em assemblagem "de-novo" ou "mapping" consoante a assemblagem seja feita sem qualquer genoma de referência ou com referência. Estes processos são geralmente executados em sequência, formando um "pipeline", onde o resultado de um processo é dado como "input" ao processo seguinte.

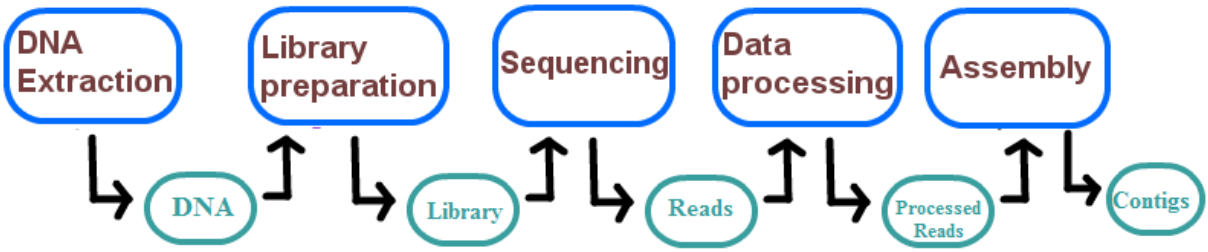

Figura 2. Representação esquematizada de uma pipeline, onde o resultado de um processo realizado alimenta o próximo processo.

Através da criação da entidade "Message" definimos o conceito que representa o input ou output da execução de um processo de uma pipeline. Cada processo terá um único output, pois a execução de um processo é única no espaço e no tempo, pelo que, foi adicionada como propriedade de cada processo uma "date" e uma relação com um "Agent", que representa quem executou o processo.

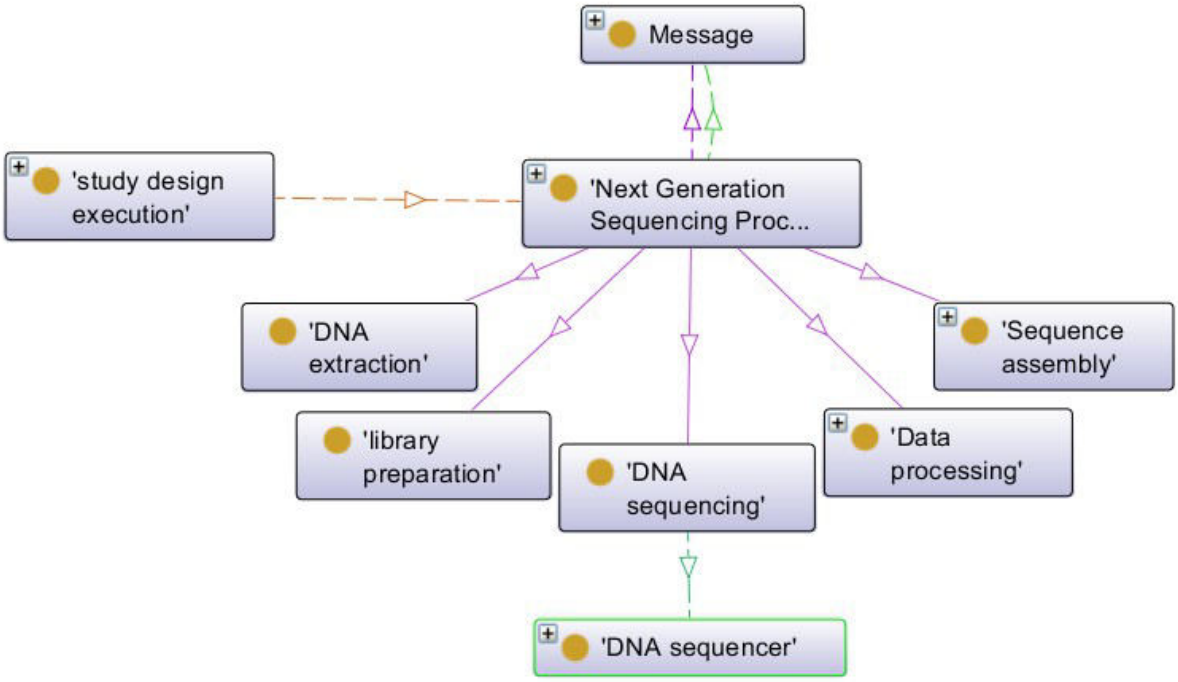

Figura 3. Representação gráfica da estrutura que modela um pipeline. Um "study design execution" será constituído, relação "has part", por uma série de "Next Generation Sequencing Process", ligados entre si por relações "has input" "has output" a uma "Message". Ligado a um "DNA sequencing" encontra-se um "DNA sequencer", através da relação "has performer".

Tendo em conta a organização dos dados dos repositórios SRA/ENA foi decidido manter a estrutura study/experiment. Nestes repositórios um "study" contem vários "experiment", no nosso caso, um "investigation" terá várias "study design execution", entidades importadas da ontologia Ontology for Biomedical Investigations (OBI).

Até este ponto temos a possibilidade de criar um estudo, que poderá conter um ou mais "pipelines", representantes de uma sequência de processos realizados, assim como a informação de quem a realizou e a data da sua realização. No entanto é notável a ausência de dados essenciais à reprodutibilidade dos processos executados, nomeadamente protocolos dos mesmos processos. Foram por esse motivo criadas entidades protocolo para cada tipo de processo possível de ser executado e referidos anteriormente. Estes protocolos ordenados formam um workflow, por intermédio da criação de uma entidade "workflow step", que

contem um número inteiro, “index”, representante da posição do protocolo na entidade “ NGS workflow”. Um protocolo com utilização de ferramentas informáticas será normalmente constituído por uma relação com um determinado “software” e os “parameter” nele utilizados.

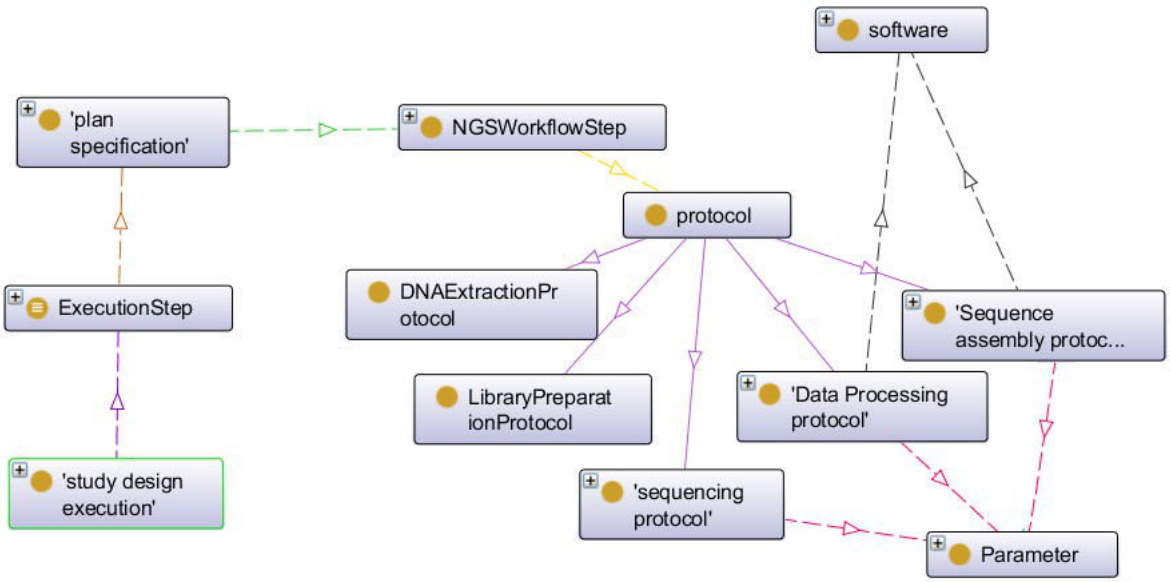

Figura 4. Representação gráfica da estrutura que modela um “workflow” e a sua ligação a um “pipeline” (study design execution). Um “pipeline”e terá um ou mais “workflows”(plan specification) ordenados pelo “ExecutionStep”, sendo que um “workflow” está organizado através do “NGSWorkflowStep”.

O desenvolvimento de um “NGS workflow” é um processo muitas vezes efetuado por diferentes pessoas/grupos em localizações geográficas diferentes e de forma não contínua no tempo, especialmente entre os protocolos laboratoriais e protocolos relacionados com tarefas bioinformáticas. Por esta razão foi decidido criar também uma entidade que permita ordenar uma sequência de “NGS workflow” utilizados para executar uma pipeline. Deste modo teremos a possibilidade de dois “pipelines” utilizarem o mesmo primeiro “NGS workflow” para representar os protocolos executados em laboratório e um segundo “NGS workflow” diferente em ambos os “pipelines”, representantes dos diferentes protocolos utilizados para a execução das ferramentas informáticas. Na Figura 5 está representada esta situação, onde ambos os “pipelines” utilizam o mesmo “workflow1” num primeiro passo, mas no segundo passo cada um utiliza um novo “workflow” diferente.

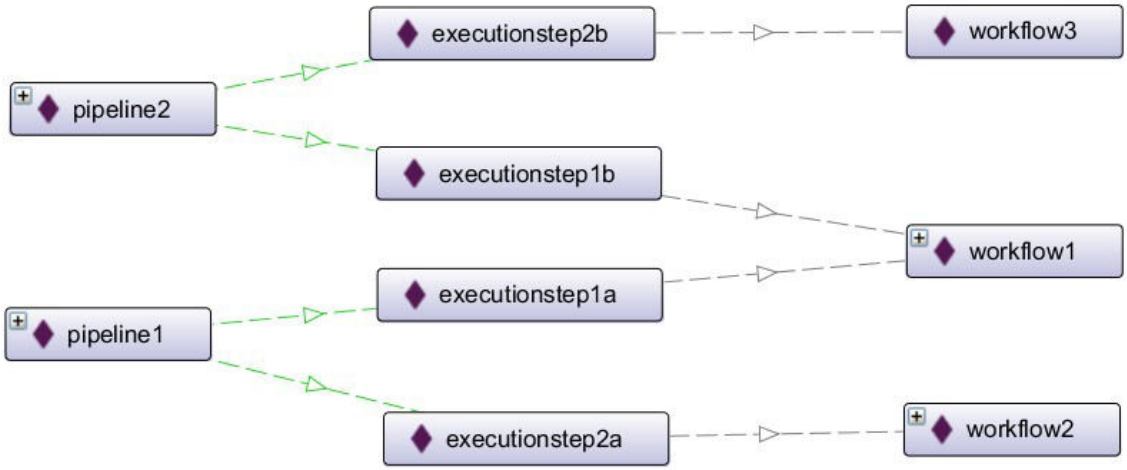

Figura 5. Representação de dois "pipelines" que partilham um mesmo "workflow".

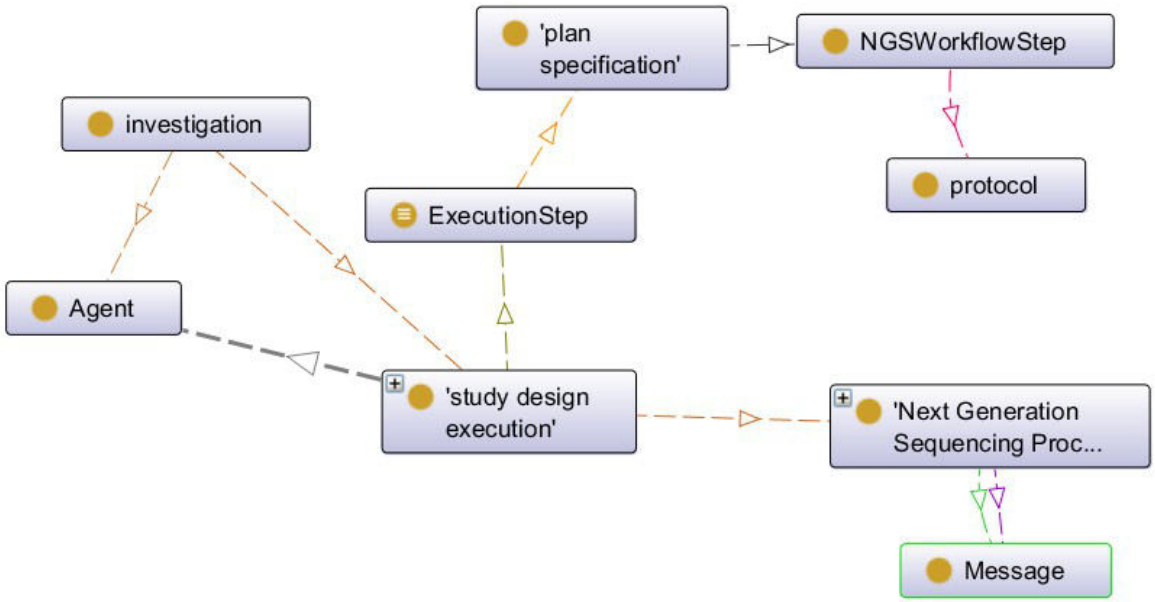

Figura 6. Representação da estrutura geral que modela por completo um estudo de NGS (investigation).

A Figura 6 representa a estrutura geral da ontologia. Como observado na figura, iniciamos com a criação de uma "investigation" (estudo) que pertence a um "Agent" (relação

"Belong to"). Um estudo é constituído (relação "has part") por uma série de "study design execution" (pipelines). Por sua vez um "pipeline" é constituído (relação "has part") por um ou mais "Next Generation Sequencing Process" (processos realizados num processo de NGS), sequencialmente ligados entre si por "Message" e relações "has input" "has output". Cada "study design execution" (pipeline) irá ter um ou mais "ExecutionStep", e um "Execution Step" terá um "index" e uma única relação com um "plan specification" (workflow). Desta forma uma "pipeline" poderá utilizar vários "workflows" de uma forma ordenada. Este sistema é novamente utilizado para representar os variados protocolos ("protocol") ordenados dentro de um workflow ("plan specification").

As propriedades que definem cada entidade estão apresentadas no anexo A. Estas propriedades são o resultado de uma seleção que poderá não representar a opinião de todos os possíveis intervenientes, pelo que a adição/remoção de novas propriedades será um processo dinâmico e alvo de uma constante discussão de todos os interessados.

Foram por fim aplicados os "plugins" de inferência automática HermiT 1.3.8 [81] e FaCT ++ 1.6.2 [82] sem ser obtido qualquer tipo de conflito e com a inferência de toda a ontologia a ser realizada de forma relativamente rápida.

3.4 Ligação com Ontologias Externas/ Conceitos Comuns

Como mencionado anteriormente, existe a necessidade de tornar a ontologia o mais interoperável com as ontologias disponíveis, nomeadamente as ontologias pertencentes à OBO foundry. A OBO foundry é um consórcio que pretende agrupar as ontologias, do domínio biomédico, da forma mais interoperável possível e utilizando uma série de princípios. As ontologias pertencentes à OBO Foundry estão construidas sobre uma ontologia de alto nível, a BFO, utilizada por esse motivo como base da NGSOnto.

A BFO agrupa as classes em duas grandes entidades, "continuant" e "occurrent". Os "continuants" são entidades que existem por si próprias ou como parte de outra entidade, os "ocurrents" são algo que possui uma parte temporal, como por exemplo uma ação (sorrir, abrir os olhos, etc). Foram consideradas como "material entity" as classes pertencentes à

subclasse “Message”, sendo que as que foram importadas da OBI já lhe pertenciam. A execução de processos, classes pertencentes à entidade “Next Generation Sequencing” foram agrupadas como uma “processual entity”.

Pertencente à OBO encontra-se a OBI, ontologia que modela todo um processo de investigação, desde os protocolos, ferramentas utilizadas, dados gerados e análises executadas sobre estes. A OBI revelou-se a maior fonte de entidades reutilizadas, facto importante uma vez que a OBI já se encontra estruturada segundo a BFO. A tabela 3 apresenta as entidades reutilizadas da ontologia OBI e o conceito pelo qual elas foram utilizadas e são definidas na OBI.

Entidade	conceito
Investigation	Sinónimo de “Estudo”
Plan specification	Informação diretiva, que quando concretizada, é realizada num processo segundo o qual, quem o realiza, pretende atingir um determinado objetivo (workflow)
Study design execution	Concretização de um protocolo (pipeline)
Protocol	Plano suficientemente detalhado que permita a reprodutibilidade de um processo (protocolo)
DNA extraction	Realização de uma extração de DNA
DNA extract	Resultado de um processo de extração de DNA
Library preparation	Realização de uma preparação de bibliotecas
DNA Sequencing	Realização de uma sequenciação de DNA
DNA sequencer	Uma máquina de sequenciação
DNA sequence data	Dados da sequência da estrutura primária do DNA
Material Sample	Sinónimo de “Amostra”
Software	software
Software method	Sinónimo de função de um Software
Software script	Instruções que podem ser executadas por um software que as interpreta

Tabela 3. Conceitos reutilizados da OBI e os conceitos pelos quais foram utilizados na NGSOnto.

Nem todos os conceitos pretendidos foram encontrados na OBI, pelo que foi

necessário recorrer a outras ontologias da OBO, nomeadamente a Sequence Ontology(SO) [83] e a Software Ontology (SWO) [84]. A ontologia SO foi especialmente útil na definição de dois tipos de conceitos importantes que incluímos dentro da classe "Message", a entidade "contig" e de "read". Da ontologia SWO foram importadas as entidades "sequence assembly" (assim como duas das suas subclasses), a entidade "sequence cutting", previamente descrita, e a entidade "Parameter", importante para definir os parâmetros utilizados num protocolo aquando da utilização de um "Software".

Por fim, foram utilizadas duas fontes de dados referentes a "material sample". Uma delas corresponde à entidade "isolate" da TypOn, referente a amostras epidemiológicas, e outra corresponde à entidade "sample", proveniente do serviço biosamples do EBI.

A entidade que representa pessoas/institutos foi importada da ontologia FOAF, a entidade "Agent".

Relativamente às relações entre entidades, foram reutilizadas algumas da Relation Ontology (RO) [85], também pertencente à OBO, e também relações da OBI e da BFO.

Relação	Ontologia
has input	RO
has output	RO
has performer	OBI
has part	BFO

Tabela 4. Relações reutilizadas na NGSOnto e a ontologia à qual pertencem.

As relações "has input" e "has output" relacionam uma entidade pertencente à classe "Message" e a uma entidade pertencente a um "Next Generation Sequencing process", ou seja, representa a relação entre um processo executado e o fluxo input/output.

A relação "has performer" representa uma pessoa ou máquina que realizou uma determinada acção. Foi utilizada nesta ontologia para definir a relação entre um "Agent" que realiza um estudo, que poderá ou não ser diferente do que realiza os processos, uma relação entre um "Agent e um "Next Generation Sequencing Process" e uma relação entre o processo de "DNA Sequencing" e a máquina que executa o processo, um "DNA sequencer".

3.5 Exemplo de um Mapeamento na NGSOnto

Considere-se o início de um estudo no instituto de medicina molecular (IMM), onde se realiza um processo genérico de NGS.

Representando esta informação na ontologia teremos um estudo, indivíduo “Study1” da classe “investigation”, descrito pelo seu título e pela sua descrição . Este estudo irá conter um pipeline “pipeline1”, da classe “ study design execution” e será pertencente ao “Agent” “IMM”. Esta informação está presente no seguinte excerto em RDF:

```
<owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#Study1">
      <rdf:type rdf:resource="&ns3;OBI_0000066"/>
      <ns3:NGS_0000055 rdf:datatype="&xsd;string">study example</ns3:NGS_0000055>
      <ns3:NGS_0000054 rdf:datatype="&xsd;string">study to demonstrate ngsonto
mapping</ns3:NGS_0000054>
      <ns3:NGS_0000015 rdf:resource="&ns3;ngsonto.owl#IMM"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#pipeline1"/>
   </owl:NamedIndividual>
```

O prefixo “ns3:” é referente a “http://purl.obolibrary.org/obo/”. Na 1ª linha temos o sujeito e o URI pelo qual é identificado. Este sujeito tem o seu tipo definido na 2ª linha, e as propriedades “study title” e “study description” descritas como uma cadeia de caracteres (string) na 3ª e 4ª/5ª linha respectivamente. Na 6ª linha encontramos a relação “Belong to” referente à relação de posse do estudo pelo instituto identificado pelo uri “http://purl.obolibrary.org/obo/ngsonto.owl#IMM”. Por fim na 7ª linha é referida a relação “has part” entre o estudo e o pipeline “pipeline1.

Considere-se neste momento, que o processo de NGS realizado no IMM será realizado na unidade de microbiologia molecular e infeção (UMMI). Neste processo é realizado uma primeira série de passos laboratoriais por um conjunto de pessoas, e é realizado posteriormente uma nova série de passos informáticos por outro conjunto de pessoas.

O “pipeline1” irá ter uma relação com o “Agent” que a executou, neste caso a unidade “UMMI”. O pipeline será constituída por uma série de processos realizados por um determinado “Agent” e a relação com um ou mais “Execution step”.

```
<owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#pipeline1">
      <rdf:type rdf:resource="&ns3;OBI_0000471"/>
      <ns3:OBI_0001950 rdf:resource="&ns3;ngsonto.owl#UMMI"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#DNAExtraction1"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#libraryPreparation1"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#sequencing1"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#filtering1"/>
      <ns3:BFO_0000051 rdf:resource="&ns3;ngsonto.owl#deNovoAssembly1"/>
      <ns3:NGS_0000076 rdf:resource="&ns3;ngsonto.owl#executionStep1"/>
      <ns3:NGS_0000076 rdf:resource="&ns3;ngsonto.owl#executionStep2"/>
   </owl:NamedIndividual>
```

Cada processo terá uma relação “has input”, “has output” e “has performer”. No caso do processo “DNAExtraction1” terá uma relação “has input” com a “sample” “SAMEA1430591”, “has output” com o “DNAExtract1” e “has performer” com o indivíduo “Jose”. O “index” de cada “workflow” executado dentro de um “pipeline” encontra-se nas propriedades do indivíduo “executionStep1” e “executionStep2” .

```
<owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#DNAExtraction1">
      <rdf:type rdf:resource="&ns3;OBI_0000257"/>
      <ns3:RO_0002234 rdf:resource="&ns3;ngsonto.owl#dnaextract1"/>
      <ns3:OBI_0001950 rdf:resource="&ns3;ngsonto.owl#jose"/>
      <ns3:RO_0002233 rdf:resource="&ns33;SAMEA1430591"/>
   </owl:NamedIndividual>
```

Por sua vez, o “DNAExtract1” será o alvo da relação “has input” do processo seguinte, formando a ordem de execução dos processos, representados na Figura 7.

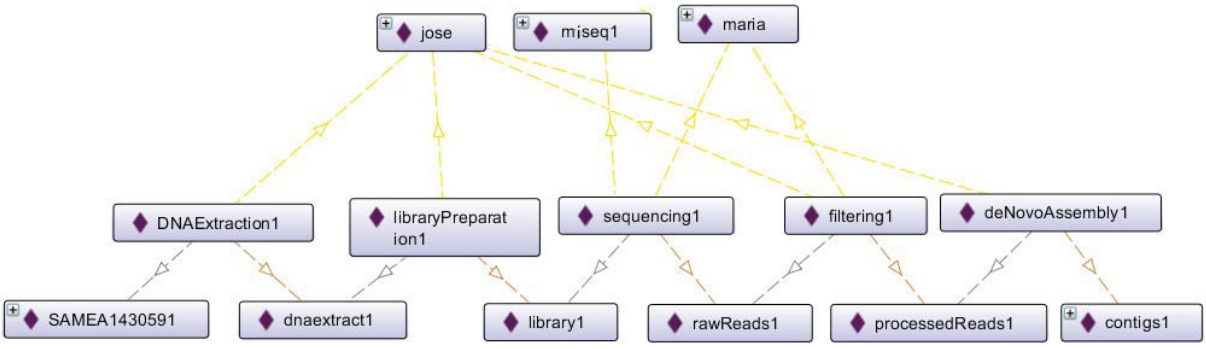

Figura 7. Representação de um pipeline em três níveis. No nível do meio estão representados os processos realizados e em baixo o fluxo output/input entre eles que forma a ordenação da

pipeline. No nível mais em cima temos os indivíduos que realizaram os processos.

Até este ponto temos representado um estudo, que contem um pipeline e os processos nele realizados. Falta-nos portanto a informação sobre os protocolos. Como anteriormente referido, um pipeline terá relações "execute" com um ou mais "Execution step", neste caso uma relação com o " executionStep1" e " executionStep2". Por sua vez, cada um destes estará relacionado com um workflow através da relação "has workflow". Desta forma conseguimos adicionar uma ordem pela qual um pipeline executou mais do que um "workflow".

```
<owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#executionstep1">
      <rdf:type rdf:resource="&ns3;NGS_0000074"/>
      <ns3:NGS_0000081 rdf:datatype="&xsd;int">1</ns3:NGS_0000081>
      <ns3:NGS_0000079 rdf:resource="&ns3;ngsonto.owl#workflow1"/>
   </owl:NamedIndividual>
```

Neste exemplo de mapeamento o nosso pipeline executou dois workflows. O workflow1 consistiu nos protocolos laboratoriais enquanto que o "workflow2" consistiu em protocolos de utilização de ferramentas informáticas. A ligação entre um workflow e um protocolo é efectuada à semelhança da ligação entre um pipeline e um workflow. O "workflow1" irá ter várias relações com vários "NGS workflow step", que terão cada um apenas uma relação com um protocolo.

```
<owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#workflow1">
      <rdf:type rdf:resource="&ns3;IAO_0000104"/>
      <ns3:NGS_0000078 rdf:resource="&ns3;ngsonto.owl#workflowStep1a"/>
      <ns3:NGS_0000078 rdf:resource="&ns3;ngsonto.owl#workflowStep2a"/>
      <ns3:NGS_0000078 rdf:resource="&ns3;ngsonto.owl#workflowStep3a"/>
   </owl:NamedIndividual>

 <owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#workflowStep1a">
      <rdf:type rdf:resource="&ns3;NGS_0000075"/>
      <ns3:NGS_0000081 rdf:datatype="&xsd;int">1</ns3:NGS_0000081>
      <ns3:NGS_0000077 rdf:resource="&ns3;ngsonto.owl#extractionProtocol1"/>
   </owl:NamedIndividual>

 <owl:NamedIndividual rdf:about="&ns3;ngsonto.owl#extractionProtocol1">
```

<rdf:type rdf:resource="&ns3;NGS_0000067"/>
<ns3:NGS_0000084 rdf:datatype="&xsd;string">NEXTprep™-Bacteria DNA Isolation Kit</ns3:NGS_0000084>
</owl:NamedIndividual>

Nestes três blocos de RDF é apresentado o 1º passo de um "workflow". No 1º bloco é representado o "workflow1" e os três passos que são efetuados neste "workflow", " workflowStep1a", "workflowStep2a" e " workflowStep3a". O 2º bloco apresenta a informação relativa ao " workflowStep1a", o seu index "1", ou seja que é a 1ª posição no "workflow", e uma relação "has protocol" com o protocolo "extractionProtocol1". No 3º e último bloco é declarado que este protocolo é executado através da utilização de um kit conhecido por "NEXTprep™-Bacteria DNA Isolation Kit " [86].

A repetição destes passos para os seguintes protocolos permite a construção de um "workflow", como o representado na Figura 8.

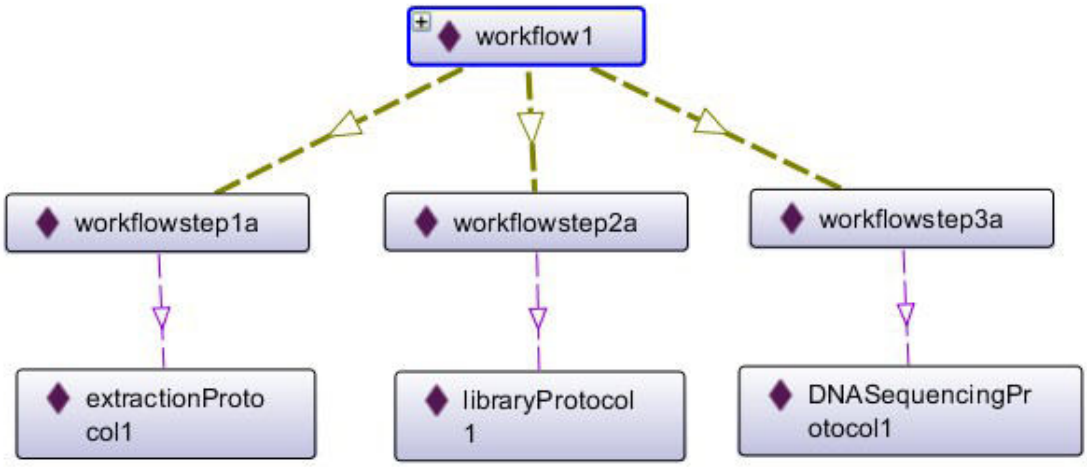

Figura 8. Representação esquemática de um "workflow" mapeado com a NGSOnto.

A construção ou reutilização de mais do que um "workflow" permite que estes sejam executados segundo uma ordem definida, como o representado na Figura 9.

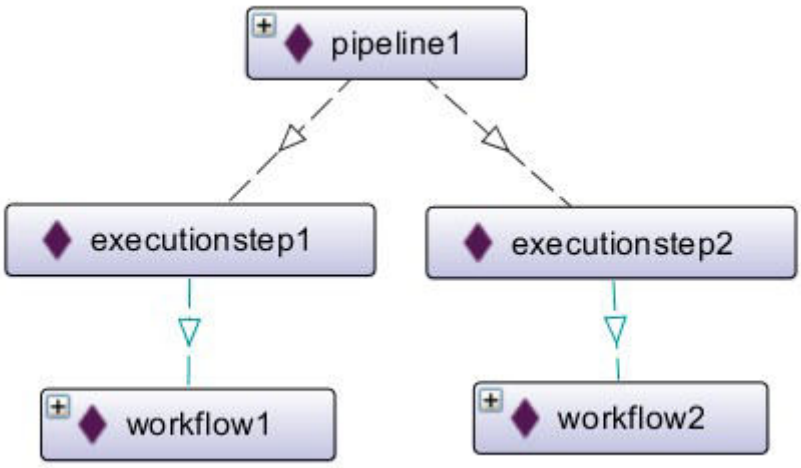

Figura 9. Representação esquemática dos "workflows" utilizados para realizar o pipeline e a ordem pela qual foram executados.

O exemplo apresentado pretende demonstrar como a informação se estrutura na NGSOnto quando mapeada. O foco está, neste exemplo, incidente sobre a relação entre os diferentes "pipelines", "workflows", "processos" e intervenientes e não sobre as propriedades em detalhe dos indivíduos mapeados.

4 Serviço Web

Neste capítulo são descritos os métodos utilizadas para construir o serviço web, com base na ontologia desenvolvida no capítulo 4 deste documento.

4.1 Métodos

A construção do serviço web seguiu os seguintes passos:

- Identificação dos softwares/linguagens a utilizar para a construção do serviço web;
- Aplicação da ontologia como modelo de base de dados num repositório de triplos;
- Desenvolvimento de uma arquitetura REST para permitir a interação entre o repositório de triplos e o exterior;
- Disponibilização de uma REST application programming interface (API);
- Desenvolvimento de uma interface de fácil utilização para o cliente.

4.2 Tecnologias Utilizadas

4.2.1 Virtuoso

O software Virtuoso[87] é um pacote de aplicações e serviços num só servidor universal. Apesar de um grande número de funcionalidades, este foi utilizado como base de dados de triplos em RDF. O Virtuoso expõe as suas funcionalidades a serviços web e permite uma fácil ligação com aplicações que utilizem drivers de acesso a dados universais (ODBC, OLE DB e JDBC, etc). Para realizar procuras (queries) é disponibilizado a tradicional linguagem SQL e a linguagem SPARQL, especializada a devolver e manipular dados em formato RDF.

4.2.2 Apache

O software Apache (versão 2.2.22) é um software de livre acesso que providencia um servidor seguro, eficiente e extensível, disponibilizando serviços HTTP atualizados com as normas. Através deste software é-nos permitido contactar com o repositório de triplos através do protocolo HTTP.

4.2.3 PHP

A linguagem PHP (versão 5.3.10) é generalizadamente utilizada e especialmente útil para o desenvolvimento de serviços web. Esta pode ser inserida dentro de HTML, no entanto ela é executada num servidor e não do lado do cliente, sendo o resultado enviado posteriormente pelo servidor. Esta linguagem é suportada pela maioria dos navegadores web e sistemas operativos e possibilita uma fácil ligação através do driver de acesso de dados universais ODBC.

A informação inserida no repositório é processada pelo servidor internamente, construindo ficheiros em RDF através da biblioteca para PHP, EasyRdf [88].

4.2.4 Javascript

A linguagem Javascript é uma linguagem de scrip orientada a objetos especialmente popular nos navegadores web, executada do lado do cliente e com alta utilidade para manipular dados e os objetos que constituem a página HTML.

4.2.5 REST

A arquitetura REST[23] é uma arquitetura simples que geralmente corre usando o protocolo HTTP. Esta arquitetura considera que as interações entre o cliente e o serviço é mais vantajoso se o limitarmos a operações com significado especifico, GET, POST, PUT e

DELETE. É também importante para o conceito REST a existência de identificadores únicos (URI), que podem ser manipulados pelos operadores identificados anteriormente. Um serviço REST separa os clientes dos servidores através de uma interface uniforme e não guarda qualquer informação do utilizador que não seja especificada no pedido.

4.3 Interface para o Cliente

O interface para o cliente deste serviço web foi construído em HTML e Javascript, utilizando o template Bootstrap 2.3.2 [89], sendo todos os pedidos feitos através do protocolo HTTP e utilizando os métodos implementados em PHP, listados na REST API apresentada no anexo B. Os métodos implementados estão disponíveis a qualquer pessoa/serviço através dos protocolos HTTP.

A interface para o cliente, disponível em http://darwin.phyloviz.net/~msilva/NGSonto/study , consiste numa série de formulários em HTML, e na manipulação dos dados e das páginas HTML utilizando Javascript. Os envios e os pedidos de dados são todos efetuados por pedidos HTTP, utilizando o objeto XMLHttpRequest já implementado de raiz. O cliente tem à sua disposição quatro páginas diferentes, três delas permitem a construção da pipeline e uma que permite a visualização dos dados já inseridos.

A primeira página de inserção de dados consiste na construção do estudo e de pipelines vazias. O estudo será identificado pelo seu título enquanto que as "pipelines" irão ser construídas ao ser fornecido, pelo utilizador, um autor dessa pipeline. No caso do estudo já estar previamente criado é possibilitada a adição de novas "pipelines" vazias. Uma funcionalidade disponibilizada permite que, ao ser iniciada a digitado o nome de um utilizador, ser efetuado um pedido assíncrono ao servidor para autores com nomes que contenham essas letras. Esta funcionalidade utiliza "*Asynchronous Javascript and XML"* (AJAX) e retorna uma lista de sugestões. Através desta técnica é possível fazer pedidos ao servidor sem a página congelar à espera da resposta do servidor, apresentando o resultado assim que a resposta seja recebida.

A segunda página para inserção de dados é relativa à construção da pipeline da parte laboratorial. São utilizados os dados inseridos na 1ª página de forma a localizar o estudo pretendido e são disponibilizados uma série de formulários referentes à preparação de bibliotecas e à sequenciação, de forma a construir o "workflow" dos protocolos. Uma vez que estes protocolos são muitas vezes utilizados para várias amostras ao mesmo tempo, faz sentido que a escolha das amostras que utilizaram estes protocolos seja disponibilizado no fim. Desta forma podemos construir uma lista de amostras que utilizaram o mesmo "workflow" previamente definido nos formulários. Uma particularidade do formulário das amostras prende-se com o facto de utilizar o SPARQL endpoint da base de dados biosamples pertencente ao serviço RDF do EBI. Utilizando este endpoint e AJAX é possível, à medida que o utilizador digita o nome de uma amostra, fazer perguntas assíncronas ao biosamples e obter uma lista de sugestões, de forma a saber se contém amostras que contenham as letras digitadas. Uma vez que a base de dados biosamples é bastante rica em dados, a ligação para amostras nela incluídas permite um enorme aumento de informação e integração de dados.

A terceira e última página de inserção de dados tem como objetivo construir um workflow de ferramentas informáticas utilizadas e os parâmetros nela utilizadas para, posteriormente, aplicar sobre uma pipeline. É oferecido ao utilizador a possibilidade de introduzir, em cada passo, um tipo de protocolo, um software usado, um a função do software e os parâmetros/valores usados nesse software.

Por fim foi criada uma página que possibilita uma visualização de alguns dados já inseridos no repositório. Esta página está estruturada de forma a ser navegável a partir do estudo, até aos processos, com um aumento gradual da granularidade. Foi também incluída uma funcionalidade gráfica, utilizando a biblioteca Cytoscape [90] para Javascript, que apresenta uma representação gráfica interativa da pipeline.

5 Discussão

Como qualquer ontologia, a NGSOnto apresentada estará sujeita a modificações dependentes da discussão que deverá surgir por parte dos utilizadores e entidades interessadas. É precisamente esta dinâmica que torna o uso de ontologias útil, permitindo que ocorra a discussão utilizador/criador e a modificação da ontologia paralelamente e que permitirá que a NGSOnto se torne uma importante base para futuras ferramentas informáticas.

Ao contrário do projeto Taverna e Galaxy, a NGSOnto aponta à captação de um workflow de sequenciação no seu todo e não apenas da sua componente informática, por isso o seu desenvolvimento não foi pensado com o intuito de substituir nenhum dos dois projetos, que tê as suas próprias vantagens e desvantagens. A NGSOnto oferece um vocabulário controlado e específico, destinado ao público alvo das ciências da vida, sendo que a distância entre a ontologia e o utilizador é encurtada e facilitada ao ser implementado com uma interface para o cliente de fácil utilização.

A utilização combinada do uso da ontologia com um serviço de arquitetura REST permitiu que, mesmo com as várias alterações no modelo da ontologia após a conclusão do serviço web, a modificação seja maioritariamente ao nível dos métodos disponibilizados. Estes métodos podem depois ser combinados entre si, ou serem utilizados diretamente por outros serviços web de forma fácil.

A NGSOnto cumpre o seu objetivo de mapear um workflow de sequenciação, sendo que no meu entender qualquer futura abordagem de raiz ao tema deverá chegar a um esquema com base semelhante, pelo que mesmo que não este não tenha uma grande adesão do publico alvo futuramente, será sempre um importante passo na medida em que se aborda um problema amplamente aceite e se apresenta uma solução.

Num contexto de epidemiologia molecular, este trabalho poderá ser importante para permitir uma comparação de dados e métodos utilizados em trabalhos de NGS de forma fácil e com possibilidade de ligação a entidades externas como o ENA. Apesar de este projeto ter tido uma base epidemiológica, esta ontologia poderá ser utilizada em outros trabalhos de NGS.

6 Conclusão

A ontologia NGSOnto fornece um meio de mapeamento de um pipeline de um processo de sequenciação por NGS, utilizando um vocabulário controlado e específico. A utilização desta ontologia possibilita a reprodutibilidade de um processo de NGS no seu todo, desde a extração de DNA até à montagem do genoma parcial.

Apesar de não ter sido possível a construção da NGSOnto apenas com a reutilização de conceitos em ontologias pertencentes à OBO, a grande parte dos conceitos e relações foram reutilizações. A organização da ontologia segundo a BFO e a confirmação por meio de inferência automática (reasoner) de que a ontologia não apresentam inconsistências aumenta a confiança na ontologia desenvolvida e na correta utilização dos conceitos reutilizados.

A integração de dados de serviços como o SRA e ENA ficou apenas reduzida às amostras que iniciam um processo de NGS. No entanto, a demonstração da possibilidade de ligação destes dados pelo serviço RDF fornecido pelo EBI fortalece a justificação do uso das tecnologias anteriormente referidas, uma vez que permite a ligação entre duas fontes de informação de forma fácil e sem qualquer replicação de informação.

O serviço web criado permite uma fácil introdução e pesquisa de dados. Os métodos deste serviço poderão ser modificados/melhorados ou criados novos métodos, nomeadamente métodos de edição (PUT) e de remoção (DELETE). A interface para o cliente não contém por enquanto nenhuma implementação de um sistema de autorização, no entanto este deverá ser futuramente implementado através da utilização das ligações com os indivíduos da entidade "Agent".

Por fim, esta ontologia será sempre alvo de avaliações quantitativas e qualitativas por parte dos utilizadores, sendo o processo do desenvolvimento da ontologia até este ponto apenas o início do sistema que deverá ser melhorado ao longo do tempo.

Este trabalho foi mencionado no "meeting report" [91] da conferência "International Meetings on Microbial Epidemiological Markers (IMMEM -10), realizada no instituto Pasteur em 2013.

8 Referências

1 Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31-46.

2 Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441-8.

3 Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, *et al*. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 2010;327:78-81.

4 Maiden, MC.; Bygraves, JA.; Feil, E.; Morelli, G.; Russell, JE.; Urwin, R.; Zhang, Q.; Zhou, J. et al. (Mar 1998). "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms."

5 Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM). Overview of molecular typing methods for outbreak detection and epidemiological surveillance.Euro Surveill. 2013;

6 Centers for Disease Control and Prevention Update: cholera outbreak—Haiti, 2010. MMWR Morb Mortal Wkly Rep. 2010;59:1473–9.

7 Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, et al. (2012) Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci U S A 109: 3065–3070.

8 Sequence Read Archive [online] Disponível em: http://www.ncbi.nlm.nih.gov/sra [Consultado em Março, 2014].

9 Paolo Missier, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Alex Nenadic, Ian Dunlop, Alan Williams, Tom Oinn, and Carole Goble. Taverna, reloaded. In M Gertz, T Hey, and B Ludaescher, editors, Procs. SSDBM 2010, Heidelberg, Germany, 2010.

10 I Foster, J Vockler, M Wilde, and Yong Zhao. Chimera: a virtual data system for representing, querying, and automating data derivation, 2002.

11 Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro A. Gonz´alez-Calero, Paul T. Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. IEEE Intelligent Systems, 26(1):62–72, 2011.

12 Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A. Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience, 18(10):1039–1065, 2006.

13 Robert Stevens, Carole A. Goble and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. Brief Bioinform. November 2000 .

14 Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006.

15 oy NF, Shah NH, Whetzel PL, et al. “BioPortal: ontologies and integrated data resources at the click of a mouse”. Nucleic Acids Res 2009;

16 Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg L J, Eilbeck K, Ireland A, Mungall C J, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann R H, Shah N, Whetzel P L and Lewis S (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", Nature Biotechnology 25, 1251 - 1255.

17 European Nucleotide Archive[online] Disponível em: http://www.ebi.ac.uk/ena/ [Consultado em Março, 2014].

18 EBI RDF Platform [online] Disponível em: http://www.ebi.ac.uk/rdf/ [Consultado em Março, 2014].

19 EBI Biosamples Database [online] Disponível em: http://www.ebi.ac.uk/rdf/services/biosamples/ [Consultado em Março, 2014].

20 EBI ArrayExpress Database[online] Disponível em: http://www.ebi.ac.uk/arrayexpress/ [Consultado em Março, 2014].

21 EBI Proteomics Identifications database[online] Disponível em: http://www.ebi.ac.uk/pride/archive/ [Consultado em Março, 2014].

22 Global Microbial Identifier [online] Disponível em: http://www.globalmicrobialidentifier.org/ [Consultado em Março, 2014].

23 Representational State Transfer (REST) [online] Disponível em: https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm [Consultado em Março, 2014].

24 [online] Disponível em: https://developers.google.com/custom-search/json-api/v1/overview [Consultado em Março, 2014].

25 [online] Disponível em: https://go.developer.ebay.com/developers/ebay/products/shopping-api [Consultado em Março, 2014].

26 [online] Disponível em: https://developer.paypal.com/webapps/developer/docs/api/ [Consultado em Março, 2014].

27 Metzker, M. L. Emerging technologies in DNA sequencing. Genome Res. 15, 1767–1776 (2005).

28 Hutchison, C. A. III. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 35, 6227–6237 (2007).

29 Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing

platforms. Nature Biotechnology, 30(5), pp.434–439.

30 http://www.illumina.com/products/nextera_dna_sample_prep_kit.ilmn

31 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc. Natl. Acad. Sci. USA 100, 8817–8822 (2003).

32 Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34, e22 (2006).

33 Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728–1732 (2005).

34 Kim, J. B. et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316, 1481–1484 (2007).

35 Leamon, J. H. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis 24, 3769–3777 (2003).

36 Wilhelm J. Ansorge . Next-generation DNA sequencing techniques. New Biotechnology, Volume 25, Number 4. April 2009.

37 Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. Science 281, 363–365 (1998).

38 Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242, 84–89 (1996).

39 Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;11:133–138.

40 Jonathan M. Rothberg et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul 20

41 Loman, N.J. et al., 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature Reviews Microbiology, 10(9), pp.599–606.

42 Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. Nature Biotech. 27, 455–457 (2009).

43 Chaisson, M. J., Brinza, D. & Pevzner, P. A. De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 19, 336–346 (2009).

44 Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub.

45 Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

46 Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Journal of Computational Biology. May 2012, 19(5): 455-477.

47 Zerbino DR. 2002. Using the Velvet de novo assembler for short-read sequencing technologies, Curr. Protoc. Bioinformatics

48 Cock PJ, Fields CJ, Goto N, Heuer ML, and Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 April.

49 Askar M, Faber MS, Frank C, Bernard H, Gilsdorf A, Fruth A, et al. Update on the ongoing outbreak of haemolytic uraemic syndrome due to Shiga toxin-producing Escherichia coli (STEC) serotype O104, Germany, May 2011. Euro Surveill. 2011;16(22):pii=19883.

50 Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PloS one. 2011;6(7):e22751.

51 Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, et al. Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. N Engl J Med. 2011;365(19):1771-80.

52 Mellmann A, Bielaszewska M, Kock R, Friedrich AW, Fruth A, Middendorf B, et al. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic Escherichia coli. Emerg Infect Dis. 2008.

53 Hao W, Allen VG, Jamieson FB, Low DE, Alexander DC. Phylogenetic incongruence in E. coli O104: understanding the evolutionary relationships of emerging pathogens in the face of homologous recombination. PloS one. 2012.

54 Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, et al. (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158: 1005–1015.

55 Miller WG, On SL, Wang G, Fontanoz S, Lastovica AJ, Mandrell RE. Extended multilocus sequence typing system for Campylobacter coli, C. lari, C. upsaliensis, and C. helveticus. J Clin Microbiol. 2005;15(5):2315–2329.

56 Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011;364:33–42

57 Gruber, T. R. (1993), 'Towards principles for the design of ontologies used for knowledge sharing', in Guarino, R. P. N., Ed., 'International Workshop on Formal Ontology, Padova, Italy, 1993'.

58 Grau B, Horrocks I, Motik B, et al. OWL 2: the next step for OWL. Web Semant Sci ServAgentWorldWideWeb 2008;6: 309–22.

59 Hilbert D. Axiomatisches Denken. Mathematische Annalen 1918;78:405–15.

60 Barwise J, Etchemendy J. Language, Proof and Logic. Center for the Study of Language and Inf, 2002.

61 Demir E, Cary MP, Paley S, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol 2010;28: 935–42.

62 Ruttenberg A, Clark T, Bug W, et al. Advancing translational research with the semantic web. BMC Bioinformatics 2007.

63 Hoehndorf R, Dumontier M, Oellrich A, et al. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. PLOS One 2011.

64 Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform 2008;41:687–93.

65 Bada M, Stevens R, Goble C, et al. A short study on the success of the Gene Ontology. Web Semant Sci ServAgents WorldWideWeb 2004;1:235–40.

66 Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. ProcNatlAcad SciUSA 2005;102:15545–50.

67 F. Couto and H. Pinto, The next generation of similarity measures that fully explore the semantics in biomedical ontologies, Journal of Bioinformatics and Computational Biology, vol. 11, no. 5 (1371001), pp. 1-12, 2013.

68 Robert Hoehndorf, Michel Dumontier, Georgios V Gkoutos. Evaluation of research in biomedical ontologies. Brief Bioinform. 2012 Sep 8.

69 Galaxy project homepage [online]. Disponível em: http://galaxyproject.org/ [Consultado em Março, 2014].

70 Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res., 34, 729–732.

71 Goble,C.A., Bhagat,J., Aleksejevs,S., Cruickshank,D., Michaelides,D., Newman,D., Borkum,M., Bechhofer,S., Roos,M., Li,P. et al. (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Res., 38, 677–682.

72 Bhagat,J., Tanoh,F., Nzuobontane,E., Laurent,T., Orlowski,J., Roos,M., Wolstencroft,K., Aleksejevs,S., Stevens,R., Pettifer,S. et al. (2010) BioCatalogue: a universal catalogue of Web Services for the life sciences. Nucleic Acids Res., 38, 689–694.

73 Basic Formal Ontology [online]. Disponível em: http://purl.obolibrary.org/obo/bfo . [Consultado em Julho, 2013]

74 Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5. Epub 2011 Jun 14.

75 Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg L J, Eilbeck K, Ireland A, Mungall C J, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann R H, Shah N, Whetzel P L and Lewis S (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", Nature Biotechnology 25, 1251 - 1255.

76 Resource Description Framework [online]. Disponível em: http://www.w3.org/RDF/ [Consultado em Março, 2014].

77 T. Berners-Lee, R. Fielding and L. Masinter, IETF. "RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax" . August 1998.

78 Extensible Markup Language [online]. Disponível em: http://www.w3.org/TR/2008/REC-xml-20081126/ [Consultado em Julho,2013].

79 RDF Syntax Specification [online]. Disponível em:http://www.w3.org/TR/REC-rdf-syntax/ [Consultado em Julho,2013].

80 Protégé home page [online]. Disponível em: http://protege.stanford.edu/ [Consultado em Março, 2014].

81 HermiT OWL Reasoner [online]. Disponível em: http://hermit-reasoner.com/ [Consultado em Março, 2014].

82 FaCT++ Reasoner [online]. Disponível em: http://owl.man.ac.uk/factplusplus/ [Consultado em Março, 2014].

83 The Sequence Ontology [online]. Disponível em: http://www.sequenceontology.org/ [Consultado em Março, 2014].

84 The Software Ontology [online]. Disponível em: http://theswo.sourceforge.net/ [Consultado em Março, 2014].

85 OBO Relations Ontology [online]. Disponível em: http://code.google.com/p/obo-relations/ [Consultado em Março, 2014].

86 NEXTprep-Bacteria DNA Isolation [online]. Disponível em: http://www.biooscientific.com/ProductsServices/NextGenSequencing/NucleicAcidIsolationforNGS/NEXTprep%E2%84%A2-BacteriaDNAIsolationKit.aspx [Consultado em Março, 2014].

87 Virtuoso Universal Server [online]. Disponível em: http://virtuoso.openlinksw.com/ [Consultado em Março, 2014].

88 EasyRdf home page [online]. Disponível em: http://www.easyrdf.org/ [Consultado em Março,

2014].

89 Bootstrap Framework [online]. Disponível em: http://getbootstrap.com/2.3.2/index.html [Consultado em Março, 2014].

90 Cytoscape.js home page [online]. Disponível em: http://cytoscape.github.io/cytoscape.js/ [Consultado em Março, 2014].

91 Brisse S, et al., Microbial molecular markers and epidemiological surveillance in the era of high throughput sequencing: an update from the IMMEM-10 conference, Research in Microbiology (2014).

Anexo A

Entidade/Propriedade	**Definição**
DNA Extraction Protocol	
Extraction method	Método usado para realizar a extração de material genómico.
DNA extraction kit name	Nome do kit de extracção de DNA utilizado
Library Preparation Protocol	
adaptor	Adaptador usado para ligar às extremidades dos fragmentos
ammountDna	Quantidade de DNA utilizado para a construção da library
rangeOfFragmentsSize	Distribuição de tamanhos dos fragmentos pós aplicação do método de fragmentação
endRepairEnzyme	Enzimas utilizadas na reparação das extremidades dos fragmentos
fragmentationMethod	Método pelo qual foi fragmentado o material genómico
libraryType	Tipo de layout da library (e.g. Mate pair, pair end, single end,...)
name	Nome da library usado caso já esteja nomeada
sizeSelector	Limites de tamanho máximo e mínimo para a seleção de fragmentos
strategy	Estratégia base utilizada para a preparação da library (e.g. WGS, Clone, ...)
tailing	Enzima ou sequencia utilizada na adição de um homopolímero nas extremidades dos fragmentos
De-novo assembly protocol	
scaffolding	Boleano que representa a existência ou não scaffolding
sequenceCoverage	Média de valores do número de nucleótidos que contribui para uma porção de uma assemblagem.
Mapping assembly protocol	
referenceGenome	Sequência que constitui o genoma que servirá de referência para o mapping.
DNA sequencer	
name	Nome da máquina usada para a sequenciação
Software version	Versão do software utilizado na máquina.
version	Versão da máquina.

Software	
name	Nome do software.
softwareVersion	Versão do software.
Software method	
name	Nome da funcção do software utilizado
Software script	
scriptBaseConcept	Finalidade e base para o uso do script.
scriptFile	Ficheiro com o script.
scriptLanguage	Linguagem utilizada para escrever o script.
Investigation	
Date	Data de início do estudo.
Study Description	Descrição do estudo.
Study Title	Titulo do estudo.
Study Type	Tipo de estudo.
Strategy	Estratégia geral do estudo
Next Generation Sequencing	
date	Data a que foi realizado o processo
Step	
index	Valor inteiro que descreve o nº do passo na ordem do workflow
Contigs	
file	Ficheiro que contem os contigs.
File extension	Tipo de ficheiro em que se encontram os contigs.
uri	Path para os contigs caso estejam depositados num repositório externo
DNA extract	
ammountDna	Quantidade de Dna extraída e mantida.
buffer	Buffer utilizado, se utilizado.
Reads	

file	Ficheiro que contem as reads.
File extension	Tipo de ficheiro em que se encontram as reads.
uri	Path para os contigs caso estejam depositados num repositório externo
Parameter	
Configuration file	Ficheiro de configuração utilizado, se utilizado.
parameter designation	Designação pelo qual o parâmetro é definido.
Parameter value	Valor do parâmetro.
DNA sequence data	
file	Ficheiro que contem as reads.
File extension	Tipo de ficheiro em que se encontram as reads.

Anexo B

REST API

study.php

POST

URI	Operation	Business Operation
/title/{title}	create	Cria estudo com título
/{id}/type/{type}	create	Adiciona um tipo a um estudo dado o ID
/{id}/sdate/{startdate}	create	Adiciona uma data a um estudo dado o ID
/{id}/description/{description}	create	Adiciona uma descrição a um estudo dado o ID
/{id}/collection/{ownerName}/ {ownerFamilyName}	create	Cria uma pipeline dentro de um estudo, pertencente a um Agent.Se não existente, é criado um novo Agent.

GET

URI	Operation	Business Operation
/id/{studyTitle}	retrieve	Devolve o id do estudo dado o título
/studyTitle/{studyTitle}	retrieve	Devolve lista títulos que contenham os caracteres dados
/collection/{studytit}/ {ownerName}/ {ownerFamilyName}	retrieve	Devolve uma lista com os pipelines dado o título do estudo e o nome do dono da pipeline

Collection.php

POST

URI	Operation	Business Operation
/{collectionid}/newWorkflow	create	Cria e adiciona um novo workflow a um pipeline
/{workflowId}/protocol/ {protocolType}	create	Cria e adiciona um novo protocol a um workflow
/{collectionid}/addWorkflow/ {workfowid}	create	Adiciona um workflow existente a um pipeline
/{workflowId}/addprotocol/ {protocolid}		Adiciona um protocolo a um workflow
/{collectionid}/execute/ {workflowid}	create	Cria, adiciona e liga a realização dos protocolos de um workflow, dado o workflow e o pipeline onde se realizam os processos
/{protocolid}/exmethod/	create	Adiciona propriedades a um protocolo de DNA

{extractionMethod}/kitname/ {kitname}		extraction
/quickLibPrepProt/{protocolid}/ {libraryType}/{strategy}/ {ammountDna}/ {fragmentationMethod}/ {sizeSelector}/{adaptor}/{tailing}/ {rangeOfFragmentsSize}	create	Adiciona propriedades a um protocolo de Library Preparation
/{protocolid}/deNovo/ {scaffolding}/{sequenceCoverage}	create	Adiciona propriedades a um protocolo de DeNovo Assembly
/collection.php/ {protocolid}/mapping/ {referenceGenome}	create	Adiciona propriedades a um protocolo de Mapping Assembly
/software/{name}/ {softwareVersion}/{protocolid}	create	Cria um novo software com as suas propriedades, caso não tenha sido já criado, e liga-o a um protocolo
/parameter/{protocolid}/ {parameterDesignation}/ {parameterValue}	create	Cria um novo parametro com as suas propriedades, caso não tenha sido já criado, e liga-o a um protocolo
/function/{protocolid}/{name}	create	Cria uma nova função e liga ao protocolo
/{processid}/output/{outputType}	create	Cria e liga um novo output a um processo
/isolated/{sampleName}	create	Pesquisa uma amostra na base de dados do Biosamples e cria como individuo do biosamples ou TypOn, se estiver ou não presente respectivamente
/collection.php/ {collectionid}/fromSample/ {sampleName}	create	Liga uma amostra ao processo de DNA extraction de um pipeline
/{processid}/input/{inputid}	create	Liga um input a um processo

GET

URI	**Operation**	**Business Operation**
/{collectionid}/process	retrieve	Devolve os processos de um pipeline
/lastProtocol/{workflowid}	retrieve	Devolve o ultimo protocolo na ordenação de um workflow
/protocol/{workflowid}/{index}	retrieve	Devolve o protocolo, dado o workflow e o passo do protocolo no workflow
LastWorkflow/{collectionid}	retrieve	Devolve o ultimo protocolo na ordenação de um pipeline