INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Visualização e Algoritmos de Clustering para Análise Filogenética

Projeto e Seminário

Licenciatura em Engenharia Informática e Computadores

Sousa, Adriano
sousa1990@hotmail.com
913790201

Nascimento, Marta
marta.nascimento02@gmail.com
968645077

Orientadores:
Vaz, Cátia (ISEL)
Carriço, João (IMM-FMUL)

21 de Março de 2015

1 Introdução

As sequências biológicas, nomeadamente o **DNA**, **RNA** e **proteínas**, têm um papel fundamental na biologia molecular porque definem quase todas as atividades celulares que ocorrem em cada organismo. A chave para decifrar estes processos recai em compreender como estas sequências interagem umas com as outras e com o seu ambiente envolvente. O **DNA** (ácido desoxirribonucleico) é uma macromolécula fundamental que contém o código genético da célula. É composto por duas cadeias de nucleótidos, entrelaçadas entre si, formando assim uma dupla hélice. Esta por sua vez é formada por diferentes bases unidas por pontes de hidrogénio. As quatro bases encontradas no DNA são a **adenina** (A), **citosina** (C), **guanina** (G) e **timina** (T). Encontram-se ligadas a um açúcar que por sua vez se liga a um grupo fosfato, formando assim um nucleótido completo. Os segmentos de DNA que contêm a informação genética são denominados de **genes**. Estes são a unidade molecular fundamental da hereditariedade de um organismo e apresentam variantes com sequências específicas de DNA denominadas **alelos**. A restante sequência de DNA tem importância estrutural ou está envolvida na regulação do uso da informação genética.

A sequenciação de DNA é o processo (método, tecnologia, etc) que permite determinar a ordem precisa dos nucleótidos numa molécula DNA. Quando se obtém uma amostra microbial ou viral, seja de um indivíduo infectado ou do ambiente, uma cultura pura (*i.e.*

de uma única espécie) é denominada de **estirpe isolada**. Desta é então extraído o DNA para a sua caracterização.

As tecnologias mais recentes permitem a sequenciação do DNA e RNA muito mais rapidamente e de forma mais barata. As máquinas que implementam esta tecnologia denominada de **"High Througput Sequencing"** ou sequenciação de alto débito, produzem grandes quantidades de informação em forma de milhões de "short reads", isto é, pequenos pedaços de sequências da amostra original de DNA. Estes não possuem qualquer ordem entre si nem uma localização conhecida. Também podem ter comprimentos arbitrários, de cadeia indeterminada e com um número arbitrário de cópias sobrepostas, e por vezes com erros de sequenciação.

Após a reconstrução de um genoma parcial através da montagem das "reads"(*i.e.* avaliação da sobreposição de nucleótidos entre elas em sequências de grande dimensão denominadas "contigs"), é necessário identificar os genes que compõem o genoma. Para tal, utiliza-se uma base de dados de genes conhecidos para a mesma espécie do organismo que se sequenciou e tenta-se extrair os genes que ocorrem nesse genoma. A correspondência é feita com base num alelo específico que já foi previamente identificado. Assim, através da comparação entre vários genes, é possível identificar a estirpe do organismo, isto é, a variante genética ou subtipo do organismo dentro da sua espécie. A identificação da estirpe do organismo é, em termos microbiológicos, designada por **tipagem**. O objetivo dos estudos de tipagem é permitir a avaliação de relações de descendência entre as estirpes em estudo. Os métodos de tipagem baseados em sequências representam as estirpes por sequências de caracteres.

A escolha do método de tipagem a utilizar depende do problema a resolver e do contexto epidemiológico no qual se vai utilizar o método. Estes métodos baseados em sequências permitem a análise ao nível da estirpe, fornecendo conhecimento importante para a vigilância das doenças infecciosas, investigação de surtos e a história natural de uma infecção. Além disso, com a recente introdução de tecnologias HTS e a possibilidade de ter acesso ao sequenciamento do genoma de uma estirpe microbiana em poucos dias, têm sido desenvolvidos novos métodos de tipagem, como por exemplo o **Multilocus Sequence Typing ribossomal** (MLST-ribossomal)[4] que usa **Single Nucleotide Polymorphism** (SNP) [5]. Estes avanços criaram a necessidade de algoritmos e ferramentas de processamento, análise e visualização desses dados, no contexto da epidemiologia, genética populacional e evolução.

Ao analisar um conjunto de isolados através de um método de tipagem, as relações inferidas entre os vários isolados é realizada recorrendo à utilização de algoritmos de inferência de árvores de filogenia.

2 Requisitos

O PHYLOViZ [1] é um *software* desenvolvido em Java e está acessível a todas as plataformas. Permite a análise e manipulação de diferentes conjuntos de dados baseados em diferentes métodos de tipagem, com o objetivo de aprofundar os estudos epidemiológicos e de populações de bactérias. Permite ainda inferir padrões prováveis de descendência evolutiva entre perfis alélicos através do algoritmo **goeBURST** ou da sua expansão

Minimum Spanning Tree (MST) baseados em diferentes matrizes distância. Posteriormente, é feita a visualização da árvore de filogenia correspondente.

Assim, este projeto terá como requisitos obrigatórios os seguintes:

- Criação de dois módulos para a implementação de dois algoritmos de clustering, um **aglomerativo** (Neighbor-Joining) [2] e outro **hierárquico** (UPGMA) [3]. O algoritmo Neighbor-Joining permite a construção de árvores de filogenia sem raíz a partir de uma matriz de distâncias e de acordo com o princípio da evolução mínima[1]. Ou seja, procura criar uma árvore de tamanho reduzido (mínimo) através da procura de vizinhos (neighbors) que minimizem o tamanho total da árvore, *i.e.* procurando e escolhendo os pares com distância mínima para serem vizinhos e posteriormente unidos. O algoritmo UPGMA[2] permite a construção de árvores de filogenia a partir de uma raíz, baseando-se também em matrizes de distância. São unidos os dois clusters (pares) que apresentem uma distância mínima e é formado um novo cluster num nível hierárquico superior. Assim a distância entre dois clusters resulta da média da distância entre os elementos de cada cluster.

- Criação de um módulo de visualização dos outputs gerados para cada um dos algoritmos mencionados anteriormente. Nomeadamente para o algoritmo Neighbor-Joining será feita uma visualização em árvore em que os arcos de ligação tenham dimensão proporcional aos pesos e para o UPGMA uma representação no formato de um dendrograma.

- Serialização do output dos algoritmos. Ou seja, uma vez calculados e para evitar a repetição do estudo, permitir que o resultado gerado possa ser novamente visualizado sem qualquer custo computacional adicional.

O requisito seguinte é opcional uma vez que apenas será realizado caso sejam concluídos com sucesso os obrigatórios:

- Salvar o estado do processamento de um DataSet.

[1] Princípio da Evolução Mínima - baseia-se na escolha de uma árvore de tamanho mínimo. Isto é, que tenha sido criada sempre através das distâncias mais curtas.

[2] U.P.G.M.A. - Unweighted Pair-Group Method with Arithmetic Mean.

3 Calendarização

Início	Duração (semanas)	Descrição
02 de Março	1	Estudo sobre os algoritmos Neighbor-Joining e UPGMA.
02 de Março	1	Análise de alguns módulos mais relevantes da aplicação PHYLOViZ.
09 de Março	2	Escrita da Proposta.
23 de Março	3	Implementação dos algoritmos Neighbor-Joining e UPGMA.
06 de Abril	1	Criação de testes que validem os algoritmos implementados.
13 de Abril	2	Divisão dos diferentes tipos de abstração suportados para cada tipo de visualização.
27 de Abril	1	Relatório de progresso e apresentação individual.
04 de Maio	4	Implementação da visualização do output dos algoritmos.
01 de Junho	2	Criação e desenvolvimento do cartaz e entrega da versão beta.
15 de Junho	2	Serialização do output dos algoritmos.
29 de Junho	2	Optimização dos algoritmos e criação de testes de escalabilidade.
13 de Julho	2	Finalização do relatório e entrega da versão final.

Tabela 1: Calendarização do Projeto.

O relatório irá ser desenvolvido ao longo de todas as semanas.

Referências

[1] A. Francisco, C. Vaz, P. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carriço. PHYLOViZ. 2014. URL `http://www.phyloviz.net/`. Consultado em 2015-03-18.

[2] Wikipedia. Neighbor joining, 2014. URL `http://en.wikipedia.org/wiki/Neighbor_joining`. Consultado em 2015-03-18.

[3] Wikipedia. UPGMA, 2014. URL `http://en.wikipedia.org/wiki/UPGMA`. Consultado em 2015-03-18.

[4] Wikipedia. Multilocus Sequence Typing, 2015. URL `http://en.wikipedia.org/wiki/Multilocus_sequence_typing`. Consultado em 2015-03-17.

[5] Wikipedia. Single Nucleotide Polymorphism, 2015. URL `http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism`. Consultado em 2015-03-17.