Análise de sequências biológicas por funções vectoriais: comparação sem alinhamento de ADN e proteínas

Resumo

A análise de sequências biológicas é uma das áreas mais importantes da bioinformática que combina diversos campos científicos, desde as ciências da computação à probabilidade e estatística. Tem como objectivo o processamento computacional e a descodificação da informação armazenada nas macromoléculas biológicas, tais como o ADN e as proteínas, envolvidas nos mecanismos celulares de todos os seres vivos e, também, a criação de ferramentas para a predição da sua estrutura, função e inferência das complexas redes de interacção entre essas mesmas moléculas.

Esta tese propõe uma abordagem à análise de sequências por funções vectoriais que transformam o espaço das sequências em vectores n-dimensionais de $\mathbb{R}^n$. Estas técnicas não dependem de algoritmos de alinhamento, usados extensivamente em aplicações bioinformáticas, e.g. no programa BLAST. Estas funções definem uma categoria, denominada 'sem alinhamento' (alignment-free) que, embora menos explorada na literatura, constitui uma área com inúmeras aplicações importantes nos últimos anos, pela sua formulação natural, formalismo elegante e custo computacional reduzido.

Neste trabalho são explorados dois tipos de funções: a primeira transforma sequências nos seus vectores de composição, ou seja, nas frequências de ocorrência das palavras de tamanho L (L-tuples); a segunda função, representação por jogos de caos (chaos game representation – CGR), baseia-se em sistemas de funções iterativas (iterated function systems – IFS) e em geometria fractal, transformando símbolos em pontos com propriedades topológicas e estocásticas relevantes aplicáveis ao estudo da sequência original.

Após uma revisão bibliográfica de métodos sem alinhamento, é apresentada uma análise quantitativa dessas métricas – baseadas em composição de palavras – com a introdução de uma nova medida de dissemelhança entre proteínas. A métrica-W (W-metric) combina métodos com e sem alinhamento através da utilização de formas quadráticas de frequência de aminoácidos associadas a matrizes de substituição com informação evolutiva. A avaliação das medidas de dissemelhança anteriormente revistas é aplicada para o reconhecimento de relações entre proteínas especificadas pela SCOP, uma base de dados de referência para a classificação hierárquica da estrutura secundária de proteínas.

No estudo de mapas CGR, este método é inicialmente generalizado de forma a acomodar alfabetos com maior cardinalidade através de mapas de sequências universais (universal sequence maps – USM), permitindo, deste modo, a representação de proteínas e de textos em linguagem natural. CGR/USM generalizam tabelas de transição de cadeias de Markov de qualquer ordem, estão relacionadas com a representação binária de números e possuem propriedades de contexto importantes; por exemplo, os sufixos, mesmo se separados na sequência original, são aplicados em regiões contíguas, sendo também possível recuperar

toda a sequência a partir de apenas uma única coordenada. Este método constitui os fundamentos de uma nova medida de entropia de sequências, também apresentada. A entropia contínua de Rényi de sequências ADN é baseada em mapas CGR/USM e na estimação não paramétrica de densidades pelo método das janelas de Parzen. Esta medida de entropia é testada em sequências de ADN artificiais e reais e é deduzido o seu comportamento assimptótico. São efectuadas, também, simulações Monte Carlo, com o intuito de estimar a variabilidade desta medida. Todos os algoritmos descritos foram implementados em MATLAB$^{\mathrm{TM}}$e estão disponíveis online.

Este trabalho permite sistematizar o estudo de técnicas sem alinhamento, ao apresentar uma revisão extensiva destes métodos e da sua respectiva aplicação, com especial ênfase dado às uniformizações da nomenclatura e formalismo que irão auxiliar o desenvolvimento futuro desta área. Adicionalmente, a análise quantitativa exaustiva desses mesmos métodos e respectivas medidas de dissemelhança obtidas através das funções vectoriais a eles associadas, comprovam que, embora com menos sensibilidade e especificidade do que algoritmos baseados em alinhamento, se obtêm resultados com custo computacional reduzido, o que os torna potencialmente importantes para pré-processamento ou filtragem de sequências e melhoria de heurísticas existentes. Foi, também, estabelecido um protocolo para avaliação dos classificadores que poderá ser aplicado facilmente no futuro ao estudo de outras medidas de dissemelhança. A representação USM, generalizada neste trabalho, motivou a criação de uma nova medida de entropia de sequências que revelou estar concordante quer com a teoria de informação, quer com estudos de simulação, permitindo o estudo da incerteza e previsibilidade de sequências biológicas. Poderá ser aplicada, no futuro, ao cálculo de perfis entrópicos e fornecer informação local para problemas de previsão e classificação.

Esta tese é baseada em artigos publicados e tem a seguinte estrutura: o Capítulo 1 – *Introduction* – apresenta uma breve introdução à biologia molecular, análise de sequências e a diversos métodos matemáticos e computacionais, tais como teoria da informação, funções vectoriais, sistemas de funções iterativas (IFS) e jogos de caos (CGR).

O Capítulo 2 – *Alignment-free sequence comparison – a review* – é uma revisão bibliográfica das principais medidas de dissemelhança que não requerem técnicas de alinhamento. Adicionalmente, apresenta material suplementar teórico em sequências, fortalecendo, também, a motivação geral deste trabalho. No Capítulo 3 – *Universal sequence map (USM) of arbitrary discrete sequences* – é proposta uma extensão natural dos mapas CGR permitindo, assim, a representação de sequências com alfabetos de maior dimensão. É desenvolvida a representação da sequência invertida e propõe-se, também, uma medida de distância entre as imagens de símbolos.

Os capítulos seguintes apresentam aplicações destes métodos ao estudo de sequências biológicas. O trabalho apresentado no Capítulo 4 – *Comparative evaluation of word composition distances for the recognition of SCOP relationships* – refere-se à avaliação quantitativa da precisão dos classificadores baseados nas mediadas de dissemelhança revistas anteriormente. Também é proposta uma nova medida, a métrica-W (W-metric), que combina conceitos de algorit-

mos com e sem alinhamento. O Capítulo 5 – *Rényi continuous entropy of DNA sequences* – apresenta uma medida de entropia baseada em mapas CGR/USM e no formalismo de Rényi, constituindo uma nova aplicação de mapas iterativos para o estudo da incerteza de sequências de ADN.

O Capítulo 6 – *Final discussion* – conjuga as conclusões dos capítulos anteriores e recapitula as principais realizações deste trabalho para o estudo de sequências biológicas. Este capítulo final também descreve alguns problemas em aberto nesta área e algumas previsões acerca do seu desenvolvimento futuro.

Esta tese apresenta e desenvolve trabalho descrito nas seguintes publicações: Vinga, S. & Almeida, J. (2003) *Bioinformatics* 19, 513–523; Almeida, J. S. & Vinga, S. (2002) *BMC Bioinformatics* 3, 6; Vinga, S., Gouveia-Oliveira, R. & Almeida, J. S. (2004) *Bioinformatics* 20, 206–215; Vinga, S. & Almeida, J. S. (2004) *J. Theor. Biol.* 231, 377–388.