Abstract

Biological sequence analysis is one of the main bioinformatics sub-disciplines, bringing together several fields, from computer science to probability and statistics. Its purpose is to computationally process and decode the information stored in biological macromolecules involved in all cell mechanisms of living organisms – such as DNA and proteins – and provide prediction tools to reveal their structure, function and complex relationship networks.

This thesis addresses sequence analysis by vector maps, which are functions that transform sequences onto n-dimensional vectors in $\mathbb{R}^n$. These techniques do not depend on sequence alignment algorithms, which are ubiquitously used in bioinformatics applications, such as the BLAST procedure. The vector maps considered define a category, named "alignment-free", that although less explored in the literature, constitutes an important subject with significant contributions in the past years, given their natural formulation, elegant formalism and low computational cost.

Two types of functions are exploited in this work: the first one maps sequences onto their sub-string or L-tuple frequency vectors and the second one, chaos game representation (CGR), is anchored on iterative function systems (IFS) and fractal geometry theory, mapping symbols onto points with applicable topological and stochastic properties.

Following a bibliographic review of alignment-free methods, an extensive quantitative analysis of these word-composition distances is performed, along with the introduction of a new dissimilarity measure between proteins. The W-metric bridges alignment metrics and those based solely in L-tuple composition, by combining, in quadratic forms, aminoacid composition and mutational information given by substitution matrices. The evaluation of the dissimilarity measures previously reviewed is applied to the recognition of protein relationships specified by the SCOP database, a benchmark for protein hierarchical secondary structure classification.

In the study of CGR maps, the method is first extended to accommodate higher-length alphabets, named Universal Sequence Map (USM), allowing the representation of proteins and natural languages texts. CGR/USM generalizes any order Markov chain transition probability tables and is related to binary representation of numbers. In addition it holds noteworthy context properties, with suffixes far apart in the original sequence mapped onto contiguous regions and the ability of recovering all the sequence from just one point. They constitute the foundation of a new entropy measure of DNA sequences here presented. The Rényi continuous entropy of DNA sequences is based on CGR/USM and in non-parametric kernel density estimation with Parzen's window method. This entropy measure is tested on artificial and real DNA and its asymptotical behavior is deduced, along with Monte Carlo simulations performed to estimate the variability of this quantity. All the computer code described was developed in MATLAB$^{\text{TM}}$language and is made available online.

This work helps systematize alignment-free techniques by presenting an extensive review of these methods and applications, with a strong emphasis on uniform nomenclature and formalism that will support future developments in

this area. Additionally, a full quantitative analysis of dissimilarity measures obtained through these vector maps showed that although less sensitive and specific than alignment algorithms, they perform reasonably well which, associated with their extremely low computational cost, make them potentially important for data pre-filtering or heuristics improvement. A precise protocol for classification accuracy assessment was established which might be used to study other dissimilarity measures in the future. The vector maps (USM) generalized in this work motivated a novel measure of sequence entropy, which is in agreement with information theory and simulation studies and allows the study of uncertainty and predictability of biological sequences. It might be further applied to the computation of sequence entropic profiles and convey useful local information for prediction and classification problems.

The thesis, based on published papers, is organized in the following structure: *Chapter 1 – Introduction* – presents background information on molecular biology, sequence analysis and mathematical and computational methods used, such as information theory, vector maps, iterative function systems (IFS) and chaos game representation (CGR).

The following *Chapter 2 – Alignment-free sequence comparison – a review* – constitutes a bibliographic review of the main techniques for measuring sequence dissimilarity not requiring their pre-alignment. Moreover, it provides additional background information on words in sequences and strengthens the motivation for all the subsequent work. In *Chapter 3 – Universal sequence map (USM) of arbitrary discrete sequences* – a natural extension of CGR maps is identified, allowing the representation of higher-order alphabet sequences. The representation for backward sequences is explored and a dissimilarity measure between symbol mappings is proposed.

The next two chapters are devoted to applications of these methods to biological sequences. The work presented in *Chapter 4 – Comparative evaluation of word composition distances for the recognition of SCOP relationships* – refers to the quantitative assessment of classification accuracy of the dissimilarity measures previously reviewed. It also proposes a new word composition measure, the W-metric, which bridges alignment-free and alignment-based concepts. *Chapter 5 – Rényi continuous entropy of DNA sequences* – presents a CGR/USM-driven entropy definition, based on Rényi formalism, which constitutes a novel application of iterative maps for measure the uncertainty of DNA.

Chapter 6 – Final discussion – finalizes by bringing together the conclusions of previous chapters and summarizing the main contributions of this work for the analysis of biological sequences. This closing chapter also describes open problems and future developments in this area.

This report presents and expands on work described in the following publications: 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.