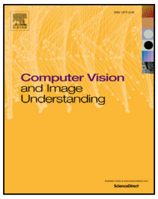
Computer Vision and Image Understanding 145 (2016) 59–69

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



A bioinformatics approach to 2D shape classification[☆]

Manuele Bicego[*], Pietro Lovato

Dipartimento di Informatica, Università di Verona, Strada le grazie 15, 37134 Verona, Italy

ARTICLE INFO

Article history:
Received 10 June 2015
Accepted 21 November 2015
Available online 27 November 2015

Keywords:
2D shape classification
Bioinformatics
Sequence alignment
Visualization

ABSTRACT

In the past, the huge and profitable interaction between Pattern Recognition and biology/bioinformatics was mainly unidirectional, namely targeted at applying PR tools and ideas to analyse biological data. In this paper we investigate an alternative approach, which exploits bioinformatics solutions to solve PR problems: in particular, we address the 2D shape classification problem using classical biological sequence analysis approaches – for which a vast amount of tools and solutions have been developed and improved in more than 40 years of research. First, we highlight the similarities between 2D shapes and biological sequences, then we propose three methods to encode a shape as a biological sequence. Given the encoding, we can employ standard biological sequence analysis tools to derive a similarity, which can be exploited in a nearest neighbor framework. Classification results, obtained on 5 standard datasets, confirm the potentials of the proposed unconventional interaction between PR and bioinformatics. Moreover, we provide some evidences of how it is possible to exploit other bioinformatics concepts and tools to interpret data and results, confirming the flexibility of the proposed framework.

1. Introduction

Research in Computational Biology and Bioinformatics experienced an unprecedented growth in the last years, mainly due to the fruitful interaction with many disciplines and fields of computer science. Among others, Pattern Recognition/Machine Learning techniques have been successfully exploited in this context [1], for many different reasons: it is possible to "learn from examples", derive quantitative models, handle non vectorial data, and deal with many classification, clustering and detection problems commonly encountered in life sciences. In many cases the particular Pattern Recognition model has not been applied "as is", but has been adapted and modified to take into account biological constraints and needs. Sometimes, this produced approaches that are very different from original methodology – a clear example is the profile-HMMs [2].

To some extent, it can be stated that this tight interaction has been mainly unidirectional, with biology/life science gaining the largest benefit[1]. In this paper, we explore an alternative direction, trying to answer the following question: *can we reverse the typical direction of interaction between Pattern Recognition and Bioinformatics?* Or, in other words, *can we exploit advanced bioinformatics models and solutions to solve pattern recognition tasks?*.

To the best of our knowledge, this perspective is rather new in the literature – the only relevant example is the video-genome project[2] [4] – and it seems a promising direction for two different reasons. First, if we are able to encode the Pattern Recognition problem in *biological terms* then we can exploit the huge range of effective, optimized, and interpretable bioinformatics tools developed by more than 40 years of research. These tools heavily rely on the solution of general pattern recognition tasks such as matching, classification, retrieval, clustering, distance computation and so on. For example, in the video-genome project [4], authors established an analogy between biological sequences and videos, defining the so called "video-DNA", a way to map features extracted from video frames into nucleotidic biological sequences. Having encoded the problem in *biological terms*, authors were then able to address the video retrieval task by using the famous BLAST [5] – an extremely fast and effective heuristic-driven algorithm for biological sequence retrieval. Second, and more important, the main goal in bioinformatics research is to derive knowledge from biological data: therefore, the interpretability of methods and solutions is a key feature, and many visualization, inspection and interpretation tools are available in the literature. These tools may be very useful also in the Pattern recognition scenarios, to better understand the different aspects of the data for a given

[☆] This paper has been recommended for acceptance by Sven Dickinson.

[*] Corresponding author. Fax: +390458027068.
E-mail address: manuele.bicego@univr.it (M. Bicego).

[1] In different cases bioinformatics issues have led to novel pattern recognition methodological challenges – the most famous example being the biclustering problem [3].

[2] See http://v-nome.org/about.html

http://dx.doi.org/10.1016/j.cviu.2015.11.011
1077-3142/© 2015 Elsevier Inc. All rights reserved.

problem: actually, in recent years interpretability has become a stringent need in Pattern Recognition [6].

This paper makes another step in this direction, providing some further evidence on the effectiveness and interpretability of bioinformatics approaches for Pattern Recognition problems. In particular, in this paper, we propose and discuss a bioinformatics approach to 2D shape classification. Analysis of 2D shapes represents an important and vibrant research area (often paving the way for 3D object classification). Many approaches appeared in the literature (see for example the reviews [7,8]): very often, the 2D shape is encoded by the contour, which proved to be an effective and natural choice in many applications. Here we propose some methods to encode the shape contour as a biological sequence, employing tailored bioinformatics tools to perform classification. In the huge literature related to 2D shape analysis, many approaches exploit sequence alignments tools to perform shape matching ([9–13], just to cite a few) – some sequence matching-based approaches which start from shape-skeletons have also been proposed [14–16]. Focusing on our main target, i.e. to use *biological* sequence alignment tools, it should be noted that few approaches exist that employ techniques developed for biological sequences to perform shape classification or matching [17,18]. Nevertheless, these approaches propose a very different perspective with respect to our approach (and the video genome project), where the main goal is to encode the PR problem in biological terms, hence exploiting tools developed for biological sequence analysis. In other words, to exploit Bioinformatics tools for Pattern Recognition, one can consider two main steps: *(i)* encoding the PR problem in biological terms; *(ii)* applying bioinformatics tools to solve the problem. From this point of view, the approaches in [17,18] are rather poor, employing one particular technique for one particular purpose, and not considering a biological encoding which would allow the use of a wide class of algorithms for sequence analysis.

In this paper we do explicitly consider this aspect: first, we establish an analogy between 2D shapes and biological sequences, this motivating the employment of bioinformatics tools. Then we propose three ways for transforming a silhouette, encoded with the 8-directional chain code [19], into an aminoacidic sequence; given that, we can compute the similarity between shapes by using established biological sequence alignment tools. Such similarity is then exploited for classification in a K-nearest-neighbor setting. Finally, we show that other biological tools and concepts (such as multiple sequence alignment, conserved domains and locality and quality of alignment) can be used for a deeper analysis of the results. We performed different experiments with five standard shape datasets; on one hand, we show that classification results are very competitive with the state-of-the art. On the other hand, we show that poor results we obtained on a retrieval case can be analysed in a deeper way by exploiting other biological sequence mining tools.

2. Background

This section briefly summarizes the bioinformatics tools exploited in our analysis. First, we present a preliminary overview of biological sequence alignment, so to clarify notations and terminology. Then, we present the tools employed for pairwise sequence alignment and multiple sequence alignment, trying to highlight specific aspects which are useful for our task.

2.1. Biological sequence alignment

Understanding and modelling the behavior of living cells is strongly dependent on the analysis of biological sequences, both nucleotide sequences – i.e. strings made with the 4 symbols of DNA, namely ATCG – and aminoacid sequences – i.e. strings with symbols coming from a 20 letter alphabet. The most important basic operation is *sequence alignment*, which is a crucial step in many computational biology and bioinformatics analyses. The alignment of a pair of sequences aims at finding the best registration between them. This is done by taking into account the biological nature of the input sequences, so that biological (usually evolutionary) events, such as mutations and rearrangements, are clearly expressed [20].

From a practical point of view, alignment is obtained by inserting spaces inside the sequences (the so called gaps) so to maximise the point-wise similarity between them – a graphical example can be seen in Fig. 1. Such maximization relies on two important parameters. The first one is the so-called *substitution matrix* $B(i, j)$ which indicates the penalty to be paid for a mismatch between symbols i and j. This encodes the fact that in nature substitutions between aminoacids/nucleotides are not all equally likely. Different alternatives exist (such as the PAM [21] and the BLOSUM [22] matrices), each one exploiting biological a priori knowledge such as chemical properties of aminoacids. The second parameter is called the *gap penalty pair*, which is a pair of numbers specifying the cost of *inserting* and the cost of *extending* a gap in one of the sequences (in biology, inserting a new gap has a different impact with respect to extending an existing one).

2.2. Pairwise sequence alignment

The simplest instance of sequence alignment aims at finding the best registration between two sequences. In this case the approaches can be divided into global and local: global methods try to find an alignment between the entire strings, whereas local approaches aim at finding short regions of high similarity. Historically, the most famous pairwise sequence alignment algorithms are the Needleman-Wunsch [23] (which operates globally) and the Smith-Waterman [24] (which is local); both methods rely on dynamic programming to solve the problem efficiently. In particular, they both have a time complexity of $O(MN)$, with M and N being the lengths of the two sequences. We chose these two established tools, dating back to 70s/80s, in order to be as basic as possible; however large margins of improvements exist, since many advanced algorithms appeared in the last 30 years; one clear example is the popular BLAST (Basic Local Alignment Search Tool) [5], which implements a set of simple but effective heuristics to drastically reduce the time complexity of the alignment.

A by-product of the alignment process is the alignment similarity score: such quantity measures how “well aligned” the two sequences are. This score can be reasonably intended as a similarity measure between two sequences.

2.3. Multiple sequence alignment

When the goal of sequence analysis is to infer evolutionary events from a set of sequences, rather than reasoning in terms of pairwise alignments, the best option is to simultaneously align all the sequences, performing the so called multiple sequence alignment (MSA - [25]). In this context, the most widely used approach employs a heuristic search known as progressive technique, which builds up the final alignment by combining pairwise alignments – starting from the most similar pair and progressing to the most unrelated. In this scenario, the most famous tool employed by researchers is ClustalW[3] [26].

Given a multiple alignment, different information can be inferred. For our scope, we will exploit two aspects:

1. The quality of the multiple alignment, which can be used to understand the local reliability of the sequence alignment (i.e. where

[3] http://www.ebi.ac.uk/Tools/msa/clustalw2/