21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

2D shape recognition using biological sequence alignment tools

Manuele Bicego and Pietro Lovato
Computer Science Department - University of Verona (Italy)

Abstract

In this paper a novel 2D shape recognition approach is proposed. The main idea is to exploit in this context the huge amount of work carried out by bioinformaticians in the biological sequence analysis research field. In the proposed approach, we encode shapes as biological sequences, employing standard and well established sequence alignment tools to devise a similarity score, finally used in a nearest neighbour scenario. Despite its simplicity, obtained results on standard datasets are really encouraging.

1. Introduction

Recognition of 2D shapes is without doubts an important and still open research area in computer vision and pattern recognition, very often representing the basis for 3D object classification. Many approaches have been proposed in the past [10, 16, 12], many of them based only on features extracted from the boundary: actually, object contours have shown to be very expressive in many contexts.

In this paper, a novel method for contour-based 2D shape recognition is proposed, which exploits techniques and solutions coming from the biological sequence alignment field [6]. From a very general point of view, the proposed approach starts from the observation that, in the past, the huge and profitable interaction between pattern recognition and biology/bioinformatics was mainly unidirectional, namely devoted to study how to apply PR tools and ideas to analyse biological data[1]. Here we would like to investigate a somehow unexplored alternative way of interaction, which consists in employing advanced bioinformatics solutions to solve pattern recognition problems. Actually, there are application scenarios in the bioinformatics field – like sequence modelling, phylogeny, database searches – which have been deeply and successfully investigated for many years. We are convinced that such fields can offer interesting solutions to pattern recognition problems, if we are able to encode our problem in biological terms. A very recent and interesting example of such an alternative way of thinking is the Video Genome Project[2], where internet videos were encoded as "video DNA sequences" and analysed with phylogenetic related tools [3].

In this paper we pursue this idea by exploiting the huge amount of work carried out in the field of biological sequence analysis [6] to face the 2D shape classification problem. In particular, we propose to transform a sequence contour into an aminoacid sequence, employing standard sequence alignment tools (like the Smith-Waterman [14] and the Needleman-Wunch [13] algorithms) to devise a sequence similarity measure. Such similarity is finally used in a standard nearest neighbour classification scenario. Even if there have been some recent attempts in using in the 2D shapes context algorithms originally proposed for aligning biological sequences (see e.g. [4, 7]), our point of view is completely different, trying to define the 2D shape classification problem in "biological" terms. We tested our approach with two standard datasets; even if we applied a very simple "shape to biological sequence" mapping as well as the basic standard bioinformatics solutions to this problem, we obtained very promising results, also in comparison with the state of the art.

2. Background: sequence alignment

Understanding and modelling living cell behaviour is strongly based on the analysis of 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 22 letters alphabet. Sequence alignment represents for sure

[1] In some other cases, biological/bioinformatics problems have led to the definition of novel methodological pattern recognition issues – a clear example is the biclustering problem (simultaneous clustering of features and patterns), which was initially introduced for the analysis of expression microarray data [11].

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

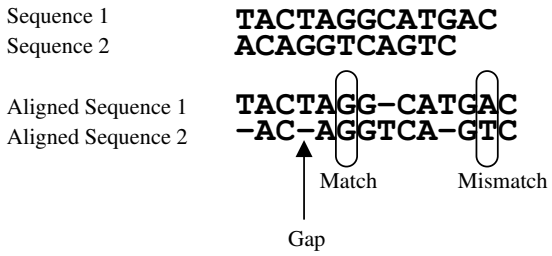

Figure 1. Alignment of two sequences.

an important basic operation, crucial in many computational biology and bioinformatics analyses. As can be intuitively understood, the alignment of two sequences is aimed at finding the best registration between them (namely the best way of superimposing one sequence on the other). From a practical point of view, alignment is obtained by inserting spaces inside the sequences (the so called gaps) in order to maximize the point to point similarity between them – see Fig. 1. A huge amount of approaches have been proposed in the past to face this problem (see [9, 8] for recent reviews and perspectives on the topic), with already effective methods aged in the seventies or early eighties [13, 14]. A thorough treatment of this topic is of course out of the scope of this paper. Here, since we are interested in investigating the basic potentialities of our ideas, we chose two very basic pairwise alignment tools (namely the Needleman-Wunsch [13] and the Smith-Waterman [14] algorithms), representing the reference in this field – being extensively employed since their proposal in the seventies/eighties.

In particular, the NeedlemanWunsch algorithm [13] is a dynamic programming method for finding the best *global* alignment between two sequences – it represents the first application of dynamic programming to biological sequence comparison. The basic idea is to maximize the similarity between two sequences by *i)* making use of a similarity matrix (also called Scoring Matrix) which defines the similarity between every pair of symbols in the alphabet and *ii)* by taking into account penalty values for gap opening and extension. There are many possible scoring matrices, which are typically built on the basis of biological knowledge [3].

On the other side, the Smith-Waterman algorithm [14] is a dynamic programming method for *local* alignment, which identifies homologous regions (i.e., (roughly speaking, similar regions) between sequences by searching for optimal local alignments. To find the optimal local alignment, again a scoring system is used, which includes a set of specified gap penalties.

3. The proposed approach

In order to apply the biological sequence alignment tools to the 2D shape recognition problem we have to transform 2D shapes into biological sequences. Many different transformations can be adopted, involving complicate shape descriptors as well complicated mappings from them to aminoacids[4]. Here, in order to really investigate the basic potentialities of our approach, we adopt a definitely simple scheme, describing the shape with one of the simplest descriptors (the 8 directional chain code). Subsequently we mapped each chaincode value into one of the following eight aminoacids, in a one to one fashion: A, R, N, D, C, Q, E, and G – namely the first 8 as given in Matlab ordering – in this way no information loss is present.

Given the encoding, the similarity between two shapes is computed via the alignment similarity score of the corresponding biological sequences: such quantity, which is a by-product of the alignment process, measures how "well aligned" the two shapes are. A nearest neighbour classifier can be finally used for the classification. An example of shapes (and the relative alignment obtained with Needleman Wunsch algorithm) is shown in Fig. 2. It can be noticed that the alignment between the first two shapes – Fig. 2(b) –, which belong to the same class, is definitely better (the number of matches is higher) than the alignment between the first and the third shape.

As a final comment, we have to say that this scheme is very simple, and in some cases approximated: for example the closeness of the boundary in 2D shapes does not have a clear biological counterpart in biological sequences; moreover, it can be enhanced in many different ways – as learning the mapping from a dataset, using quantized continuous shape descriptors to cover all the 22 aminoacids, defining a proper shape specific scoring matrix and so on. In any case, the results we obtained were very promising.

4. Results

The proposed idea has been tested on two different datasets, namely the *Chicken Pieces* dataset[5] [1] and the Vehicle Shape dataset[6] [15]. The first dataset con-

[3] For example, in the nucleotide case, it is known from the chemical composition of DNA basis that it is more difficult to have a change from an Adedine to a Thymine rather than to a Guanine.

[4] Reasonably, we decided to encode shapes into aminoacid sequences, these allowing more sophisticated description if compared with nucleotide sequences (alphabet of 22 symbols rather than 4).

[5] http://algoval.essex.ac.uk:8080/data/sequence/chicken/.

[6] http://visionlab.uta.edu/shape_data.htm.

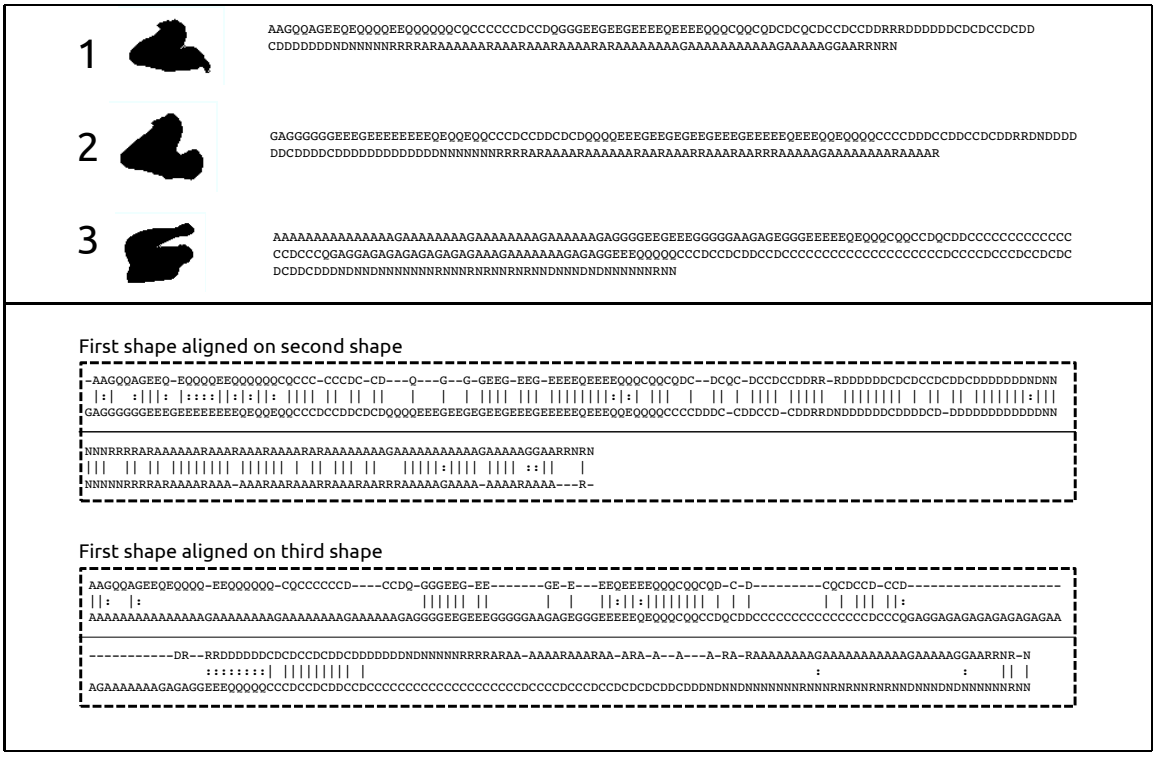

Figure 2. (Top) Three shapes and the corresponding sequences; (Bottom) Sequence alignment. Note: In the alignment, a pipe connects two matched aminoacids, while a dot connects similar residues that are not exact matches.

tains 446 binary images (silhouettes) of chicken pieces, each belonging to one of five classes representing specific chicken parts.The second dataset contains 120 vehicle shapes extracted from traffic videos using motion information – as described in [15] –, classified in four classes.Leave One Out accuracy was computed for the chicken dataset (as in many nearest neighbour approaches dealing with the chicken dataset), whereas in the vehicle shape dataset the accuracy was determined with 10-fold cross validation (as specified in [15]). The classification, in both cases, has been carried out with the nearest neighbour rule.

Two crucial parameters that should be defined when aligning two sequences are the scoring matrix and the gap opening/extending penalty. As explained in the previous Sections, the former defines the price of every substitution in the matrix, whereas the latter defines the penalty in the similarity got while opening (or extending) a gap region. These two parameters typically have a clear biological meaning, and can change drastically the final result. In this preliminary evaluation, we performed two sets of experiments: in the former (top part of Table 1) we tried to keep as easiest as possible the scheme, leaving such parameters as set by default in the Matlab implementation (Matlab bioinformatics toolbox); in the latter (bottom part of Table 1) we relaxed one biological assumption which does not hold in the 2D shape classification case – this being of course only the first step through the tailoring of the sequence alignment tools to our problem. In particular we observe that in biology the gap penalty is typically high: it's not really desirable to break a biological sequence. In the shape case, nevertheless, such a strong constraint does not hold: actually, gaps can really help in dealing with occlusions and – mainly – scale changes. From results in table 1 – where we reduced the gap opening penalty (and the gap extending penalty) – it seems evident the beneficial effect of such operation, this encouraging us to go ahead along this direction.

In table 2 we provide recent results from the state of the art on the same datasets. Many different approaches have been tested on the Chicken dataset, using simple as well complicated classifiers (see for example comparisons reported in [2, 5]): in Table 2(a) we reported only those based on nearest neighbour rules – taken from [2]. Even if in some cases different experimental protocols

Method	Chichen	Vehicle
SW	0.8229	0.8500
NW	0.8027	0.8500
SW (reduced gap penalty)	0.8341	0.8500
NW (reduced gap penalty)	0.8229	0.8583

Table 1. Accuracies for the proposed methods: SW (Smith-Waterman) and NW (NeedlemanWunsch)

have been employed, it seems evident that the proposed approach represents a promising alternative to classic as well as to advanced schemes. Moreover, as can be seen from Table 2(b), our approach also comparably compares with other techniques employing more sophisticated classifiers (as SVM) – here the results, all taken from [15], are fully comparable (the same validation protocol was employed).

Methodology	Accuracy
1-NN + Levenshtein edit distance	≈ 0.67
1-NN + approximated cyclic distance	≈ 0.78
K-NN + cyclic string edit distance	0.743
1-NN + mBm-based features	0.765
1-NN + HMM-based distance	0.738
1-NN + IT kernels on n-grams	0.814
Our best (SW – reduced gap penalty)	0.834

(a)

SVM + curvature	0.6250
SVM + Fourier Descriptors	0.8250
SVM + Zernike moments	0.7917
Ergodic HMM + Max Lik.	0.6250
Circular HMM + Max Lik.	0.7333
Left Right HMM + Max Lik.	0.7083
HMM + Weighted likelihood	0.8417
Our best (NW – reduced gap penalty)	0.8583

(b)

Table 2. Comparative results: (a) Chicken dataset; (b) Vehicle dataset.

5. Conclusions

In this paper we preliminary investigated the idea of exploiting bioinformatics tools to solve Pattern Recognition problems. In particular we cast the 2D shape analysis problem into the biological sequence aligment problem, for which a huge amount of approaches have been proposed in the bioinformatics community. Obtained results encourage us to go ahead along this research line.

References

[1] G. Andreu, A. Crespo, and J. Valiente. Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition. In *Proc. of IEEE ICNN97*, volume 2, pages 1341–1346, 1997.

[2] M. Bicego, A. Martins, V. Murino, P. Aguiar, and M. Figueiredo. 2d shape recognition using information theoretic kernels. In *Proc. Int. Conf on Pattern Recognition*, pages 25–28, 2010.

[3] A. Bronstein, M. Bronstein, and R. Kimmel. The video genome, 2010. arXiv:1003.5320v1.

[4] L. Chen, R. Feris, and M. Turk. Efficient partial shape matching using smith-waterman algorithm. In *Proc. Int. Conf on Computer Vision and Pattern Recognition*, 2008.

[5] M. Daliri and V. Torre. Shape recognition based on kernel-edit distance. *Computer Vision and Image Understanding*, 114(10):1097–1103, 2010.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids. cambridge univ, 1998.

[7] R. Huang, V. Pavlovic, and D. Metaxas. A profile hidden markov model framework for modeling and analysis of shape. In *Proc. Int. Conf on Image Processing*, pages 2121 –2124, 2006.

[8] C. Kemena and C. Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19), 2009.

[9] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[10] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[11] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. on Computational Biology and Bioinformatics*, 1:24–44, 2004.

[12] Y. Mingqiang, K. Kidiyo, and R.Joseph. A survey of shape feature extraction techniques. In P.-Y. Yin, editor, *Pattern Recognition Techniques, Technology and Applications*. 2008.

[13] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Modelecular Biology*, 48(3):443–453, 1970.

[14] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[15] N. Thakoor, J. Gao, and S. Jung. Hidden markov model-based weighted likelihood discriminant for 2-d shape classification. *IEEE Transactions on Image Processing*, 16(11):2707 –2719, 2007.

[16] D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37:1–19, 2004.