2D Shapes Classification Using BLAST

Pietro Lovato and Manuele Bicego

Computer Science Department - University of Verona, Italy

Abstract. This paper presents a novel 2D shape classification approach, which exploits in this context the huge amount of work carried out by bioinformaticians in the biological sequence analysis research field. In particular, in the approach presented here, we propose to encode shapes as biological sequences, employing the widely known sequence alignment tool called BLAST (Basic Local Alignment Search Tool) to devise a similarity score, used in a nearest neighbour scenario. Obtained results on standard datasets show the feasibility of the proposed approach.

Keywords: 2D shape classification, sequence alignment, biological sequences.

1 Introduction

The classification of 2D shapes represent an old and widely investigated research field in computer vision and pattern recognition. Many approaches have been proposed in the past (see e.g. the reviews [1–3]), many of them based on the analysis of the boundary: actually, object contours have shown to be very effective in many applications, with several different approaches presented over the past years, exhibiting different characteristics: robustness to noise and occlusions, invariance to translation, rotation, and scale, computational requirements, and accuracy.

In this paper, a novel approach for contour-based 2D shape classification is proposed, which exploits techniques and solutions coming from the biological sequence alignment context [4]. 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 at studying and applying PR tools and ideas to the analysi of biological data [5][1]. In this paper a somehow unexplored alternative way of interaction is investigated: the idea is to employ 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

[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 to analyse expression microarray data in order to discover subsets of genes with a coherent behaviour in subsets of samples [6].

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 273–281, 2012.

for many years by bioinformaticians. 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 [7].

In this paper we follow this line of investigation by exploiting the huge amount of work carried out in the field of biological sequence analysis [4] to face the 2D shape classification problem. In particular, we propose to transform a sequence contour into an aminoacid sequence, employing the most famous biological sequence alignment tool – the BLAST (Basic Local Alignment Search Tool [8]), – to devise a similarity measure between sequences. Such similarity is then used in a standard nearest neighbour classification scenario. The proposed approach has been tested with two standard datasets, the Chicken Pieces Database [9] and the Vehicle Shape dataset [10]; even if we applied a very simple "shape to biological sequence" mapping, obtained results were very promising, also in comparison with the state of the art.

2 Background: Sequence Alignment with BLAST

Research in biology is very often based 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 22 letters alphabet. Many different kinds of biological analyses are based on a preliminary sequence alignment step. 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); the registration is done by taking into account the biological nature of the input sequence, so that biological (usually evolutionary) events, such as mutations and rearrangements, can be clearly expressed [4].

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-wise similarity between them – see Fig. 1.

In the past, a huge amount of approaches have been proposed to deal with this task (see [11–13] for recent reviews and perspectives on the topic), with already effective methods aged in the seventies or early eighties [14, 15]. A thorough treatment of this topic is of course out of the scope of this paper. Two distinctions are important from our perspective: the former distinguishes between pairwise and multiple alignment approaches, with the former devoted at finding the best registration of two sequences and the latter aimed ad finding a simultaneous alignment of more than two sequences. The latter subdivides the approaches in global and local alignment methods: the global ones try to find the best overall alignment between sequences, whereas the local ones aim at finding short regions of high similarity.

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

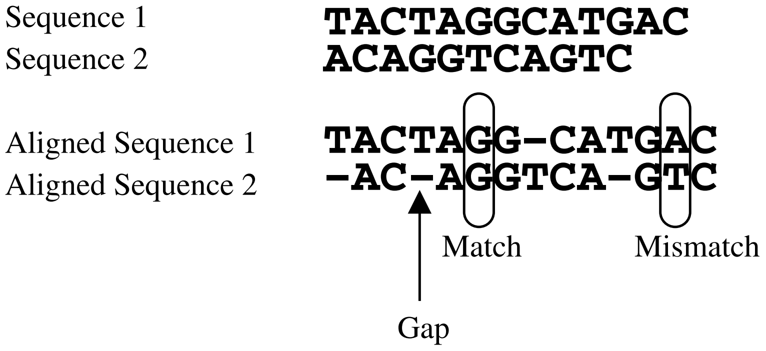

Fig. 1. Alignment of two sequences

The BLAST (Basic Local Alignment Search Tool) algorithm is for sure the most widely known alignment tool (the Scopus database indicates more than 30 thousands citations to the orignial paper, whereas for GoogleScholar they are more than 40 thousands), introduced by Altschul and colleagues in the 1990. Many different versions have been lately introduced, some of them being now very popular (e.g. psiBLAST [16]). In few words, the BLAST algorithm permits to find the sub-optimal alignment of a query sequence with respect to a dataset of other sequences, providing also a score to every pairwise alignment. BLAST is an approximate algorithm (only giving a sub-optimal yet accurate result), whose success is devoted to the simple but effective heuristics implemented inside which permit a really fast implementation (dynamic programming solutions to the same problem are nowadays absolutely not employable).

Briefly, given in input a sequence (query) to be aligned to a dataset, the algorithm performs the following steps:

1. remove low complexity regions from the query sequence
2. extract from the query sequence all the K-mers (i.e. all the possible subsequences, with overlap, of length K). These subsequences are called "words"
3. search, in the whole database, all the words having a reasonably good match with the words of the query sequence – these words are called "hits"
4. use these words as seeds, attempting to extend both forward and backward from the match to produce an alignment. The algorithm will continue this extension as long as the alignment score continues to increase or until it drops by a critical amount owing to the negative scores given by mismatches. These extended segments are called HSP (High Scoring segment Pairs), and represent the aligned part of the two sequences. In other words, the the alignment is *local*, namely is based on the alignment of a small part of the two sequences.
5. To the alignments found by BLAST during a search a statistical value is assigned, called the "Expect Value" (E-value). This number represents the number of times that an alignment as good as or better than that found by BLAST would be expected to occur by chance.

For more details about this algorithm, interested readers can refer to the book [17][3].

3 The Proposed Approach

The proposed approach is carried out in two steps: first, shapes should be transformed into biological sequences; then, the similarity score between two shapes should be extracted from the alignment of the two corresponding sequences. A nearest neighbour classifier can be finally used for the classification.

1. **From 2D shapes to biological sequences** Even if many different transformations can be adopted, involving complicate shape descriptors as well complicated mappings from them to aminoacids[4], here we adopted a rather simple scheme, in order to analyse the basic potentialities of our approach. In particular, every shape is described by encoding the contour with the 8 directional chain code [18], representing one of the simplest shape coding strategy; then, each chaincode value is directly mapped into one of eight aminoacids: A, R, N, D, C, Q, E, and G – which are the first 8 as given in Matlab ordering.
2. **From alignment to similarity** Given two shapes encoded as biological sequences, it is natural to link similarity between two shapes to the alignment similarity score: such quantity, which is a by-product of the alignment process, measures how "well aligned" the two shapes are, and is the objective function which is maximed during the alignment process. The computation of this quantity is based on the so called "scoring matrix", representing a matrix which, in a position i, j, gives a measure of the "price" we have to pay in a given alignment when substituing the aminoacid i with the aminoacid j. Different scoring matrices have been presented in the biological literature, each one starting from different biological assumptions and observations[5].

Given a testing sequence, we use the BLAST algorithm to align it to all the sequences in the training set, assigning it to the class of the most aligned training sequence. Clearly, since BLAST is a local alignment technique, multiple hits can be found of the same sequence. Nevertheless, similarly to what done in biology, we retain and consider only the first (and thus best) match. A further note: the BLAST algorithm returns a matching score (of the HSP) and the E-value. It is widely accepted in the biology to rank the aligments on the base of the E-value (the smaller the better) rather than on the alignment scores. Actually, after some

[3] Available from `http://www.ncbi.nlm.nih.gov/books/NBK1734/`

[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] The possibility of defining a scoring matrix which is specific for the shape problem is currently under investigation.

preliminary experiments, we noticed that results obtained with the E-value are substantially better than those obtained with the matching score, therefore we chose to use such value for our classification scheme.

As a final comment, we can observe that this scheme is rather 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, many enhancements can be derived – 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 obtained results are already very promising, encouraging us in going ahead along this research direction.

4 Results

The proposed idea has been tested on two different benchmarks, the *Chicken Pieces* dataset[6] [9] and the Vehicle Shape dataset[7] [10]. The first set is composed by 446 silhouettes of chicken pieces, each belonging to one of five classes representing specific chicken parts: wing (117 samples), back (76), drumstick (96), thigh and back (61), and breast (96). This represents a really challenging classification task, with the baseline classification accuracy of about 67% [19]. The second dataset contains 120 vehicle shapes extracted from traffic videos using motion information – as described in [10] –, classified in four classes: sedan, pickup, minivan or SUV. Some examples of shapes belonging to the two datasets are shown in Fig. 2 and 3. The classification accuracies have been computed in two different ways, in order to compare the proposed approach with the state of the art. In particular, for the chicken dataset we used Leave One Out accuracy (as in many nearest neighbour approaches dealing with the chicken dataset), whereas in the vehicle shape dataset we used 10-fold cross validation (as specified in [10]). As specified in the previous section, the classification, in both cases, has been carried out with the nearest neighbour rule.

In the alignment process of two sequences there are two crucial parameters that should be defined: the scoring matrix and the gap opening/extending penalty. As explained in the previous Sections, the former defines the price we have to pay in the alignment score for every substitution, whereas the latter defines the penalty in the similarity introduced by opening (or extending) a gap region. It is important to note that in biology these two parameters have a clear meaning, and can change drastically the final result. In this preliminary evaluation, we performed two sets of experiments: in the former (first row of Table 1) we tried to keep the easiest possible scheme, leaving such parameters as set by default in the BLAST implementation[8]. The only change we did was to remove the filter, applied within BLAST, which removes zones of low complexity (such

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

[7] `http://visionlab.uta.edu/shape_data.htm`

[8] Downloadable from `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/`

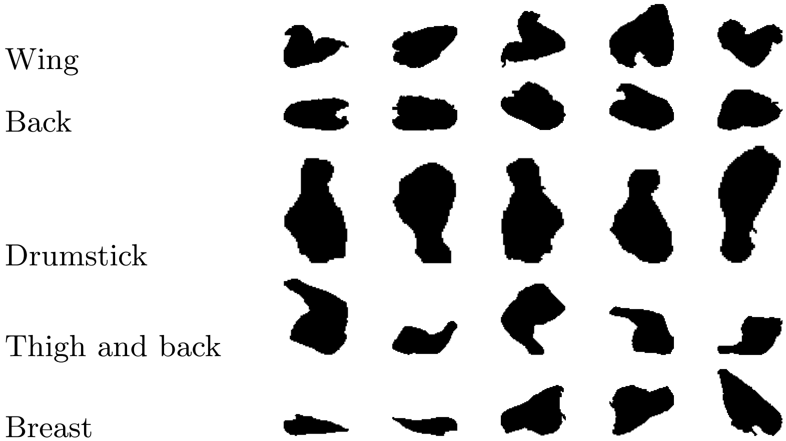

Fig. 2. Some examples from the Chicken Dataset

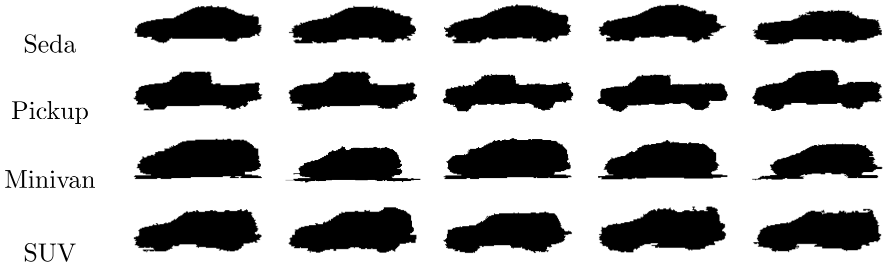

Fig. 3. Some examples from the Vehicle Dataset

as repetitions of the same symbol). Of course this has a clear meaning in biology, whereas in shapes such parts are indeed very informative (representing straight parts of the shape, for example) and should not be removed.

Table 1. Accuracies for the proposed methods

Method	**Chichen**	**Vehicle**
BLAST - Default Settings	0.7892	0.8208
BLAST - Reduced gap penalty	0.8206	0.8437
BLAST - Reduced gap penalty and BLOSUM90	0.8341	0.8542

In the second set of experiments we tried to exploit the fact that we are working with 2D shapes, using this information to properly set the two parameters. As a first trial, we relax one biological assumption which does not hold in the

2D shape classification case: 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. The second row of the Table 1 report results obtained by setting the gap opening penalty to 6 and the gap extending penalty to 2 (default values are 11 and 1, respectively[9]). It seems evident the beneficial effect of such operation.

As a second trial, we chose a substitution matrix which highly penalizes changes in the sequences (namely the algorithm is forced to try to align the sequences in the best possible way). The idea here is that whereas in biology there are somehow "equivalent" aminoacids (which can likely exchanged), in the 2D shapes context an exact matching can preferred. Results obtained by using a BLOSUM90 matrix (default is BLOSUM62, the higher the number after the word "BLOSUM" the more "conservative" the substitution matrix is) are reported in the third line of Table 1 (the gap opening/gap extending penalties were set as in previous experiment). Also in this case it can be noted the beneficial impact of such choice, even if not so evident as in the gap penalty case. We are currently continuing with further analysis of the impact of the substituion matrix on the performances and on the alignments.

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

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	0.834

(a)

Methodology	**Accuracy**
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	0.8542

(b)

[9] Unfortunately, in the BLAST implementation the choice should be made among a pre-fixed set of pair gap opening-gap extending penalties.

As a final comment, in Table 2 we reported some other 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 [20, 21]): in Table 2(a) we reported only those based on nearest neighbour rules – taken from [20]. Even if in some cases different experimental protocols have been employed, it seems evident that the proposed approach represents a promising alternative to classic as well as to advanced schemes. It is interesting to observe that the proposed approach, based on approximated matching, also outperforms exact matching techniques, as those based on edit distance. 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 [10], are fully comparable (the same validation protocol was employed).

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. Loncaric, S.: A survey of shape analysis techniques. Pattern Recognition 31(8), 983–1001 (1998)
2. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition 37, 1–19 (2004)
3. Mingqiang, Y., Kidiyo, K., Joseph, R.: A survey of shape feature extraction techniques. In: Yin, P.Y. (ed.) Pattern Recognition Techniques, Technology and Applications (2008)
4. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ. (1998)
5. Baldi, P., Brunak, S.: Bioinformatics: the Machine Learning Approach, 2nd edn. MIT Press (2001)
6. Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: a survey. IEEE Trans. on Computational Biology and Bioinformatics 1, 24–44 (2004)
7. Bronstein, A., Bronstein, M., Kimmel, R.: The video genome. arXiv:1003.5320v1 (2010)
8. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology 215, 403–410 (1990)
9. Andreu, G., Crespo, A., Valiente, J.: Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition. In: Proc. of IEEE ICNN 1997, vol. 2, pp. 1341–1346 (1997)
10. Thakoor, N., Gao, J., Jung, S.: Hidden markov model-based weighted likelihood discriminant for 2-d shape classification. IEEE Transactions on Image Processing 16(11), 2707–2719 (2007)

11. Li, H., Homer, N.: A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics 11(5), 473–483 (2010)
12. Kemena, C., Notredame, C.: Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25(19) (2009)
13. Notredame, C.: Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology 3(8) (2007)
14. Needleman, S., Wunsch, C.: 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)
15. Smith, T., Waterman, M.: Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197 (1981)
16. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997)
17. Bergman, N.: Comparative Genomics, vol. 1 and 2. Humana Press (2007)
18. Gonzalez, R., Woods, R.: Digital Image Processing, 2nd edn. Prentice Hall (2002)
19. Mollineda, R., Vidal, E., Casacuberta, F.: Cyclic sequence alignments: Approximate versus optimal techniques. Int. Journal of Pattern Recognition and Artificial Intelligence 16(3), 291–299 (2002)
20. Bicego, M., Martins, A., Murino, V., Aguiar, P., Figueiredo, M.: 2d shape recognition using information theoretic kernels. In: Proc. Int. Conf on Pattern Recognition, pp. 25–28 (2010)
21. Daliri, M., Torre, V.: Shape recognition based on kernel-edit distance. Computer Vision and Image Understanding 114(10), 1097–1103 (2010)