2014 22nd International Conference on Pattern Recognition

S-BLOSUM: classification of 2D shapes with biological sequence alignment

Pietro Lovato*, Alessio Milanese†, Cesare Centomo†, Alejandro Giorgetti†, Manuele Bicego*
*Department of Computer Science, University of Verona
†Department of Biotechnology, University of Verona
Strada le Grazie 15, 37134 Verona (ITALY)
Corresponding author: Pietro Lovato; Email: pietro.lovato@univr.it

***Abstract*—Recent works investigated the possibility to design solutions for pattern recognition problems by exploiting the huge amount of work done in bioinformatics. If the pattern recognition problem is cast in biological terms, then a huge range of algorithms, exploitable for classification, detection, visualization, etc. can be effectively borrowed. In this paper, we exploit biological sequence alignment tools to classify 2D shapes, tailoring the biological parameters of these tools to account for the different semantic of the 2D shape scenario. In particular, we propose a novel substitution matrix, which is the crucial parameter determining the sequence alignment solution. The new matrix, called S-BLOSUM, learns the rates of matches/mismatches in conserved portions of shapes belonging to the same category, and incorporates prior knowledge on the chosen representation for the 2D shape. On one hand, the experimental evaluation showed that the S-BLOSUM provides a significant improvement over the biological counterpart (BLOSUM); on the other hand, classification results prove that our approach is competitive with respect to the state of the art.**

I. Introduction

2D shape analysis is an important and still open research area in computer vision, often representing the basis for recognition of 3D real-world objects. The goal in a classification / recognition task is to assign a category to an unknown 2D shape on the basis of a set of classifiers, learned from a training set which contains examples of the different categories. Several approaches have been proposed in the past (see for example the reviews [1], [2]), and many of them are based on the analysis of the boundary: very often, the 2D shape is encoded by its contour, which proved to be an effective and perceptually reasonable choice in many applications. Different techniques exhibit different characteristics: robustness to noise and occlusions, invariance to translation, rotation, and scale, computational requirements, and accuracy (see [3], [4], [5], [6] and references therein).

Recently, an alternative class of approaches have been proposed by some of the authors: in particular, we presented a parallelism between the 2D shape recognition problem and the biological sequence alignment [7], [8], exploring the idea of borrowing bioinformatics tools to solve pattern recognition problems such as the shape classification one. The main observation was that, in the past, the huge and profitable interaction between pattern recognition and biology has been mainly unidirectional, namely devoted to study how to apply PR tools, algorithms and ideas to analyse biological data [9]. An alternative, complementary way of interaction may be to translate advanced bioinformatics solutions into ideas and methodologies useful to solve a pattern recognition task[1]. For example, sequence analysis is a problem encountered every day in the life sciences: there has been a vast amount of tools and solutions, improved in more than 40 years of research, developed to analyse biological sequences (sequence alignment, motif discovery, phylogenesis are just few examples). Thus, encoding a 2D contour as a string (like the simple chaincode descriptor [11]), and mapping it into a biological sequence opened to the possibility of exploiting a wide class of techniques coming from the biological sequence alignment community, where specialized algorithms for string matching, visualization, and interpretation have been developed.

Indeed, results obtained in [7], [8] proved the suitability of the approach, even in its simplest scheme – employing the biological tools "as are". Clearly, there is room for many improvements and refinements. In this sense, we can observe that parameters in sequence alignment techniques are finely tuned to take into account the *biological* nature of the input sequences so that evolutionary events, such as mutations or rearrangements can be clearly expressed. If we use biological tools as they are we do not take into account the fact that symbols in the shape alphabet (for example, chaincodes) have a very different semantics than aminoacids in nature.

This paper is aimed at investigating this aspect, trying to understand to which extent this is crucial. In particular, more than exploiting biological alignment tools to classify 2D shapes, we also tune the biological parameters of these tools to account for the change in the applicative scenario. We start from the observation that biological knowledge (in the alignment process) is encoded in the form of a *substitution matrix* (the most famous one called BLOSUM [12]): an entry (i, j) in such matrix indicates the score to assign for a match/mismatch (the lower, the more penalized it is) between symbols i and j, and this models the fact that in nature substitutions between aminoacids are not all equally likely. In this paper, we propose a novel substitution matrix, which we will refer to as Shape BLOSUM (S-BLOSUM), learned

[1] The same alternative way of thinking was also pioneered in the Video Genome Project [10] – see http://v-nome.org/about.html – where internet videos were encoded as "video DNA sequences" and analysed with phylogenetic related tools.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.405

IEEE computer society

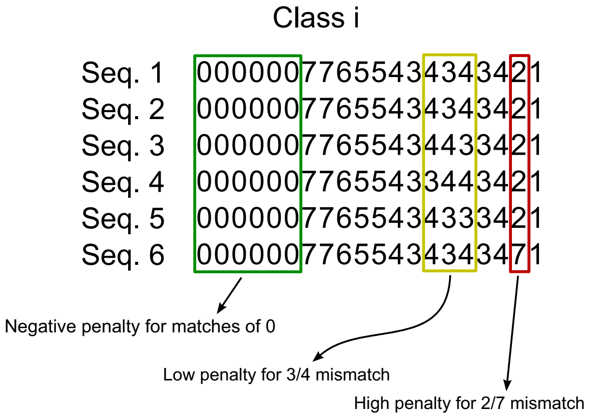

Class i
Seq. 1 00000077655434343421
Seq. 2 00000077655434343421
Seq. 3 00000077655434433421
Seq. 4 00000077655433443421
Seq. 5 00000077655434333421
Seq. 6 00000077655434343471
Negative penalty for matches of 0
Low penalty for 3/4 mismatch
High penalty for 2/7 mismatch

Fig. 1. A substitution matrix can be learned by observing the matches/mismatches frequencies that occur in chaincode strings composing a category of shapes. In the example, the symbol 0 is highly conserved, so it will have a strong negative penalty. On the other hand, substitutions between 3 and 4 happen very often, resulting in a low penalty value for the mismatch between the two.

from our shape data in a similar fashion the BLOSUM is built from biological data [12]. In this way, each element models the variations that are likely to occur within shapes of the same category, due for example to local deformations or random noise (a graphical explanation is depicted in Fig. 1). Furthermore, we introduced some *a priori* assumptions on the chaincode representation, leading to a matrix S-BLOSUM$_{\text{full}}$ which accounts for prior knowledge and is also driven by the data we actually observe; as expected, combining these aspects proves to be the best choice in terms of classification accuracy.

Once a proper matrix is chosen, it can be plugged into any biological sequence alignment program, which gives for any pair of sequences an alignment score, which reflects "how well" those sequences are aligned under the substitution matrix. Such quantity is the similarity measure exploited for classification in a nearest-neighbor setting. To quantitatively test the proposed approach, we performed experiments on three benchmark shapes datasets, demonstrating the suitability of the proposed scheme: results showed that our matrices improve the results obtained with the biological BLOSUM, and are also very competitive with recent state-of-the art techniques we evaluated for comparison.

II. Background: BLOSUM matrices and sequence alignment

Understanding and modeling living cell behavior is strongly based on the analysis of sequences, both nucleotide sequences – i.e. strings made with the 4 symbols of DNA (A, T, C, and G) – and aminoacid sequences, i.e. strings with symbols coming from a 20 letters alphabet. Sequence alignment is aimed at finding the best registration between two sequences, namely a superposition between the two where identical or similar characters are aligned in successive columns. The alignment takes into account biological – usually evolutionary – events such as mutations, insertions, deletions or rearrangements. In practice, the alignment works by inserting gaps in either of the sequences, in order to maximize their point-wise similarity.

A huge amount of algorithms for sequence alignment exist in the literature [13], [14], [15], [16], and they can be classified in several different categories. The main taxonomy divides the approaches in three categories: global alignment methods, which are aimed at finding the best overall alignment between two sequences; local alignments, which detect related segments in a pair of sequences, and multiple alignments, which are aimed at simultaneously align more than two sequences.

All of these techniques heavily rely on a fundamental parameter, called the *substitution matrix*, which assigns a score for matches/mismatches based on the rate at which one character in a sequence is likely to mutate into another one (the higher, the more likely it is). Another important parameter, the *gap penalties*, is specified by a pair of values representing the cost for inserting a gap and extending an existing one.

A fundamental issue is therefore how to choose and design properly the substitution matrix B. Intuitively, B should have the highest values on the diagonal: if two symbols match, a high score should be assigned. For mismatches, B should reflect the fact that there are some that are highly improbable, due for example to physical or chemical properties of aminoacids. In the protein domain, the most employed substitution matrix is the one called BLOSUM [12], being the default choice of many bioinformatics tools available on the web. Note that many alternatives exist, the oldest one being the PAM [17].

III. The proposed approach

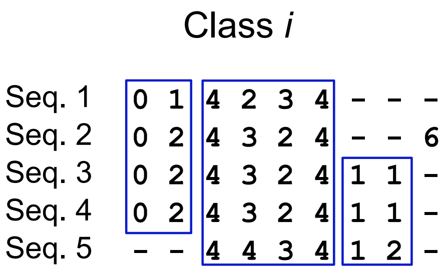
Our proposal is to build a substitution matrix able to deal with our peculiar scenario. We assume that a shape is encoded with the 8-directional chaincode, thus strings are defined over an alphabet of 8 symbols. Note however that our approach can be employed with any descriptor that is able to map a 2D shape into a string.

A. S-BLOSUM construction

In the following, we explain how the S-BLOSUM is built: starting from the description of the BLOSUM matrix (found in [12]), we will detail in every step how we account for the change in representation.

Block extraction: The starting concept in building a BLOSUM is that of a *block* of related sequences. In biological terms, related sequences are the ones which belong to the same evolutionary family – namely, they share the same biological function. Blocks then are ungapped, highly similar portions of sequences, which have been previously aligned and extracted using a multiple sequence alignment (simultaneous alignment between more than two sequences), using as substitution matrix the identity. Without entering too much into details, large databases of blocks exist, and even if the original paper dates back to the 90s, the BLOSUM matrix still used nowadays has been built using $>$ 2000 blocks.

In our case, we can set a parallelism between families (evolutionary related sequences) and classes (semantically



Class *i*

```
Seq. 1  0 1  4 2 3 4  - - -
Seq. 2  0 2  4 3 2 4  - - 6
Seq. 3  0 2  4 3 2 4  1 1 -
Seq. 4  0 2  4 3 2 4  1 1 -
Seq. 5  - -  4 4 3 4  1 2 -
```

Fig. 2. Three blocks can be extracted from the sequences depicted in the figure. The initial alignment was determined using an identity substitution matrix. Note that the last column is not considered because it is not conserved, with too many gaps in the alignment.

related shapes). It seems reasonable to consider shapes belonging to the same category as a group which shares a relationship at a high level, in the same way biological sequences are classified on the basis of their molecular functions. Similarly to the biological case, we define a block as an ungapped region of sequences inside a class of shapes. We computed an initial alignment by performing a multiple alignment (i.e. a simultaneous alignment between all sequences) using a unitary substitution matrix (matches = 1, mismatches = 0). From this initial alignment, we extracted only the columns where only few gaps appeared: even if in the biological case only columns with 0 gaps are considered, we relax this assumption retaining also the columns that contained some (up to a given threshold), as depicted in Fig. 2. Furthermore, we did not consider a block as composed by *consecutive* columns (as opposite to the biological case), and all columns with the same number of sequences contributing to the alignment have been merged together.

Log odd computation: Given a block, all possible pairs of aminoacids (in the biological context) or chaincode symbols (in our shape scenario) are counted for each column. In both cases, these counts are used to compute a matrix Q where an entry q_{ij} $(1 \leq j \leq i \leq 8)$ represents how frequently we observe symbols i and j spanning the different columns of the block. From this quantity, the probability of occurrence of the i-th chaincode symbol is

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

Then, we can estimate the *expected* probability of occurrence e_{ij} for each i, j pair in the same way of [12], namely

$$e_{ij} = \begin{cases} p_i p_j & i = j \\ 2 p_i p_j & i \neq j \end{cases}$$

Finally, we compute a log odd ratio in half-bits units:

$$b_{ij} = 2 \log_2 \left(\frac{q_{ij}}{e_{ij}} \right)$$

This value is the one appearing in a generic entry of our S-BLOSUM, and has a very intuitive meaning: if the observed frequencies are as expected, $b_{ij} = 0$, if they are less than expected (i.e. a rare mutation) $b_{ij} < 0$, if more than expected $b_{ij} > 0$.

Sequence clustering: Finally, one may want to reduce multiple contributions deriving from the most closely related sequences, as they can mask the contributions of the rare mutations in the computed frequencies. This is done by specifying a clustering percentage, which we will refer to as B number, in which sequences within a block that are identical for at least that percentage are grouped together. In other words, if we set this threshold to 62 (leading to the BLOSUM62 matrix, which is the best practice in the biological case), any two sequences inside a block which share more than 62% of aminoacids in the same positions are clustered together, and their contributions are averaged in calculating pair frequencies. Incrementing this number results in lower score for mismatches, which will be then more penalized. Our approach replicates this step in a verbatim way, and we provided an evaluation of the sensitivity of this parameter in the experimental section.

B. *Incorporating prior knowledge*

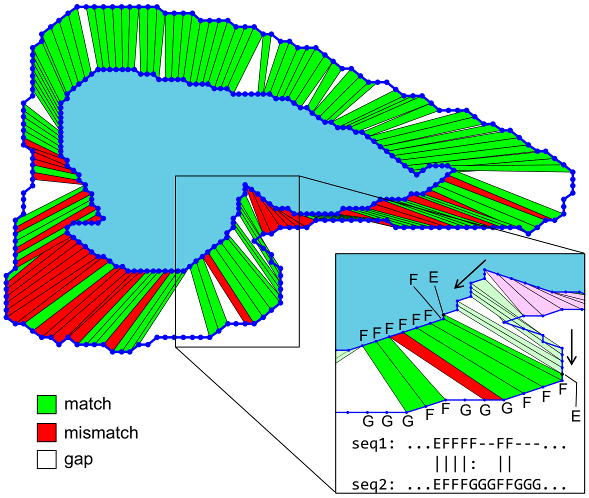
When building a pattern recognition system, the best way to estimate an unknown quantity is to incorporate prior knowledge on the learning process. Motivated by this, we can further refine the proposed S-BLOSUM by incorporating prior knowledge on the shape contour: intuitively, it is more likely that the chaincode symbol coding for the direction "south" is swapped with the symbol "south-east", rather than with the symbol "north". This information is encoded in a new matrix, which we called S-BLOSUM$_{\text{full}}$, obtained by opportunely weighing the entries of the S-BLOSUM. Such matrix is essentially a symmetric Toeplitz matrix, where the first row (thus the various weight for each substitution of the 8 chaincodes) is composed by the values $[8, -4, -7, -11, -13, -11, -7, -4]$. We will show in our experimental evaluation that this straightforward refinement proves to be a very reasonable choice.

C. *Shapes' alignment*

Once a shape is encoded as a chaincode string, and a substitution matrix has been chosen, we can align two sequences in our dataset using any biological sequence alignment tool, with the alignment score giving the similarity between those two sequences. Since we are mainly interested in evaluating the potentialities BLOSUM and S-BLOSUM substitution matrices, we employed the Smith-Waterman [14] algorithm, as it is one of the most employed tool in the computational biology community. The Smith-Waterman is a dynamic programming method for local alignment, which identifies homologous regions (i.e. regions of high similarity) between sequences by searching for optimal local alignments. The alignment score is then exploited as a similarity measure in a nearest-neighbor classifier, used for classification. A graphical representation of an alignment in our shape scenario is shown in Fig. 3.

IV. Experimental evaluation

To quantitatively assess the suitability of the proposed approach, we performed experiments on three benchmark



```
seq1: ...EFFFF--FF---...
          ||||:  ||
seq2: ...EFFFGGGFFGGG...
```

Fig. 3. Two shape contours, encoded by the 8-directional chaincode, are represented in the figure. For clarity, one shape has been circumscribed in the other. The two have been aligned with a biological algorithm: a light (green) stripe connecting two symbols indicates a match, a dark (red) stripe a mismatch, and no connection means that the alignment procedure inserts a gap in one of the two sequences (as detailed in the zoomed portion).

datasets used in the literature for 2D shape classification. Our aim is two-fold: on one hand, we want to show that employing the proposed substitution matrices is generally better than adopting the biological one as done in the basic evaluation of [7], [8]. On the other hand, we want to see how our results compare with several state-of-the-art descriptors and methodologies.

A. Parameters setting and experimental details

Validating the proposed approach requires two parameters to be defined: the B number and the penalties for gaps. As explained in the previous sections, the B number is a percentage which defines the contribution of closely related sequences: setting this parameter with a low value induces mismatches to be more tolerated. We performed a comprehensive evaluation to assess the sensitivity of the method w.r.t. this parameter: in general, we found that the value assigned does not alter drastically classification accuracies.

The second parameter is the pair gap opening/extending penalties: they represent the penalty (which contributes to the overall similarity) to open or extend a gap region respectively. In our evaluation, we performed two sets of experiments: in the former, we left such quantities as set by default in the majority of biological tools, namely 11 for the gap opening and 1 for the gap extension; in the latter, we tried to adjust these parameters and tune them considering that we are dealing with chaincode strings, not aminoacid strings. In particular, as already observed in [7], [8], the cost for opening a gap is high: it is not desirable to break a biological sequence. In the shape case, nevertheless, such a strong constraint may not hold: actually, gaps can help in dealing with occlusions and – mainly – scale changes. For this reason, in our second set of experiments, we reduced such penalties to the values (6,2), which is the lowest pair allowed by many tools publicly available[2].

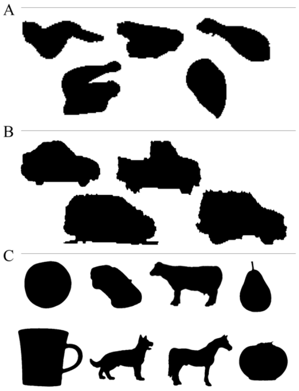


Fig. 4. Some examples of the shapes composing the Chicken pieces dataset (A), the Vehicle dataset (B), and the ETH-80 dataset (C). One shape per class is shown.

The shapes' datasets we considered for the experimental evaluation are: *i)* the *Chicken pieces*[3] [18], composed by 446 shapes of chicken parts, divided in 5 classes; *ii)* the *Vehicle*[4] [19], which contains 120 vehicle shapes classified in 4 classes; *iii)* the ETH80[5] [20], which contains 80 high-resolution color images of 3D objects from 8 categories, with each object represented by 41 images taken from different points of view, leading to a total of 3280 images (some examples can be seen in Fig. 4). Using the nearest-neighbor rule, we computed classification accuracies varying the protocols according to the state-of-the-art references we used for comparison: in particular, we employed the leave-one-out on the Chicken dataset, 10-fold crossvalidation on the Vehicle dataset, and the leave-one-object-out protocol on the ETH80, as detailed in the original paper [20]. To evaluate the proposed framework using the biological BLOSUM, we had to establish a 1:1 correspondence between a chaincode symbol and an aminoacid: we replicated the choice made in [7], [8], namely we arbitrarily assigned to each chaincode symbol one of the first 8 aminoacids as given in the IUPAC coding[6]: A, R, N, D, C, Q, E, and G.

B. Results

To give an immediate insight into the difference between the BLOSUM and S-BLOSUM matrices, we portrayed in Fig. 5 an example of two chicken shapes belonging to the same class, aligned using the biological BLOSUM80 (top row) and the proposed S-BLOSUM$_{\text{full}}$80 (bottom row). In the alignments, the symbols A and R correspond to the directions “north” and “north-east” respectively. The figure shows that the bottom alignment is more accurate, both in terms of mismatches and in terms of gap insertions – actually, in the first case the first shape is incorrectly classified. A possible explanation can be

[2] for example, in the BLAST webserver at http://blast.ncbi.nlm.nih.gov

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

[4] http://visionlab.uta.edu/shape-data.htm

[5] http://www.d2.mpi-inf.mpg.de/Datasets/ETH80

[6] http://www.iupac.org/publications/pac/1984/pdf/5605x0595.pdf

TABLE I
RESULTS ON THE CHICKEN DATASET, VARYING THE PARAMETERS B NUMBER AND GAP PENALTIES

B type	B(45)	B(62)	B(80)	B(90)	Avg
BLOSUM	0.79	0.81	0.77	0.81	0.76
S-BLOSUM	0.72	0.72	0.80	0.80	0.76
S-BLOSUM$_{\text{full}}$	0.85	0.85	0.85	0.85	**0.85**
BLOSUM - reduced gap	0.83	0.83	0.82	0.85	0.83
S-BLOSUM - reduced gap	0.72	0.73	0.79	0.79	0.76
S-BLOSUM$_{\text{full}}$- reduced gap	0.90	0.90	0.89	0.89	**0.89**

TABLE II
RESULTS ON THE VEHICLE DATASET, VARYING THE B NUMBER AND THE GAP PENALTIES

B type	B(45)	B(62)	B(80)	B(90)	Avg
BLOSUM	0.78	0.81	0.77	0.78	0.78
S-BLOSUM	0.58	0.82	0.83	0.78	0.75
S-BLOSUM$_{\text{full}}$	0.83	0.86	0.85	0.84	**0.84**
BLOSUM - reduced gap	0.81	0.78	0.85	0.81	0.81
S-BLOSUM - reduced gap	0.60	0.87	0.84	0.83	0.78
S-BLOSUM$_{\text{full}}$- reduced gap	0.88	0.88	0.89	0.87	**0.88**

TABLE III
RESULTS ON THE ETH-80 DATASET

Method	BLOSUM	S-BLOSUM	S-BLOSUM$_{\text{full}}$
Accuracy	0.85	0.88	**0.91**

that in the S-BLOSUM matrix the mismatch between the A and R symbols is scored with a value of -0.5, whereas their substitution is scored -3 using the biological BLOSUM. In fact, from the biological point of view, the substitution of a big positively-charged aminoacid (R) into a small non-polar one (A) is a drastic change and may heavily affect the function of the protein.

Classification results for the Chicken and Vehicle datasets are shown in Tables I and II. In the top part of the table, we employed as gap penalty values the default (11,1): on average, the most evident improvement is reached when employing our substitution matrix S-BLOSUM$_{\text{full}}$, the one accounting for both prior knowledge and observation of chaincodes' data. In the bottom part of the table we reported results obtained by reducing the gap penalties. As a final consideration, it seems that the alignment of chaincode strings is not that sensitive to the B number, neither with the standard biological BLOSUMs nor with our S-BLOSUM$_{\text{full}}$. However, the learned S-BLOSUM alone shows a different trend, as lowering too much the B number degrades performances: this result can reflect the fact that whereas in biology there are somehow "equivalent" aminoacids (which can likely be exchanged), in the 2D shapes context an exact matching can be preferred. Following these considerations, we evaluated our approach on the ETH-80 using the reduced (6,2) gap penalties and setting the B number to 80. Results are depicted in Table III, where again using the proposed methodology provides an improvement in terms of classification accuracies.

TABLE IV
COMPARATIVE RESULTS ON THE CHICKEN DATASET

Approach	Reference	Accuracy
1-NN + Levenshtein edit dist	[21]	0.67
1-NN + approx. cyclic dist	[21]	0.78
K-NN + cyclic string edit dist	[22]	0.74
1-NN + mBm-based features	[23]	0.77
1-NN + HMM-based distance	[23]	0.74
1-NN + IT kernels on n-grams	[24]	0.81
1-NN + BLOSUM (local alignment)	[8]	0.83
Our best	-	**0.89**

TABLE V
COMPARATIVE RESULTS ON THE VEHICLE DATASET

Approach	Reference	Accuracy
Ergodic HMM + max lik.	[19]	0.63
Left-right HMM + max lik.	[19]	0.71
Circular HMM + max lik.	[19]	0.73
SVM + Zernike moments	[19]	0.79
SVM + Fourier descriptor	[19]	0.83
HMM + weighted lik.	[19]	0.84
1-NN + BLOSUM (global alignment)	[8]	0.86
Our best	-	**0.88**

TABLE VI
COMPARATIVE RESULTS ON THE ETH80 DATASET

Approach	Reference	Accuracy
1-NN + PCA Masks	[20]	0.83
1-NN + Cont. DynProg	[20]	0.86
SVM + Kernel-edit distance	[5]	0.91
Kernel LDA	[25]	**0.92**
Our best	-	0.91

In Tables IV, V and VI we reported some other recent results from the state of the art on the same datasets. On the Chicken dataset, many different approaches have been tested in the literature, using simple as well as complicated classifiers: in Table IV we reported only those based on nearest neighbor rules. From the table, 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 Tables V and VI, our approach also performs comparably with other techniques which employ more sophisticated classifiers (such as SVMs).

V. CONCLUSIONS

This paper is built upon a novel and recent idea aimed at translating advanced bioinformatics solutions into methodologies useful to solve a pattern recognition problem. If the pattern recognition problem is cast in biological terms, then a huge range of algorithms, exploitable for classification, detection, visualization, etc. can be effectively borrowed. A preliminary evaluation, applied in the particular context of 2D shape classification, has already appeared in [7], [8], where the problem has been cast into the biological sequence alignment one. In this paper, we improved and refined this way of thinking, tuning the parameters of the biological tools for our very different applicative scenario. We propose a novel substi-

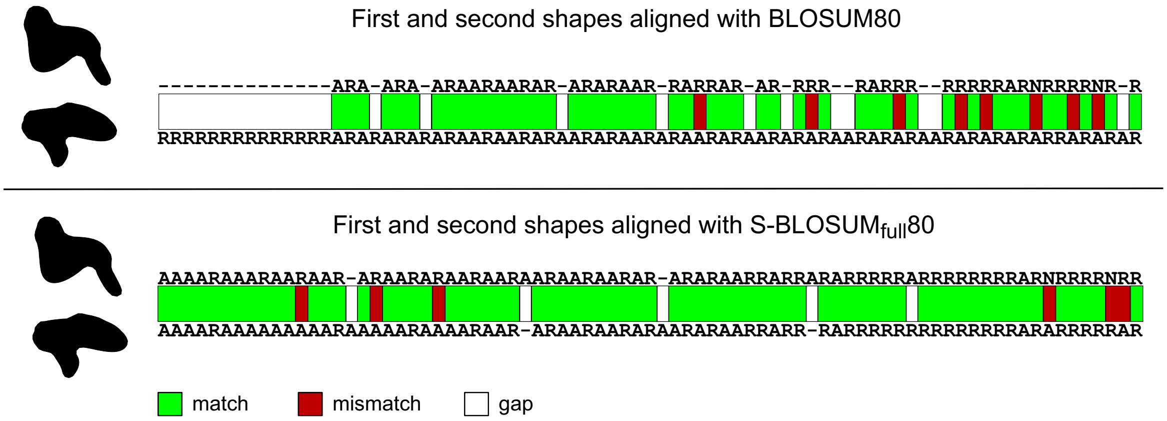

First and second shapes aligned with BLOSUM80
-------------ARA-ARA-ARAARAARAR-ARARAAR-RARRAR-AR-RRR--RARRR--RRRRRARNRRRRNR-R
RRRRRRRRRRRRRRARARARARARAARAARARAARARAARARAARARAARARAARARARARARARARARRARARAR
First and second shapes aligned with S-BLOSUMfull80
AAAARAAARAARAAR-ARAARARAARAARAARAARAARAR-ARARAARRARRARARRRRARRRRRRRRARNRRRRNRR
AAAARAAAAAAAAARAAAAAARAAAARAAR-ARAARAARARAARARAARRARR-RARRRRRRRRRRRRRARARRRRRAR
🟩 match
🟥 mismatch
□ gap

Fig. 5. Two chicken shapes of the same class (portrayed on the left) are aligned using both the biological BLOSUM (top part of the figure) and our S-BLOSUM$_{\mathrm{full}}$80 (in the bottom). The alignment done with the proposed approach is more accurate, observing that in the aligned portions there are fewer mismatches and gaps.

tution matrix, which is the crucial parameter in any biological alignment algorithms. We called our matrix S-BLOSUM: following an approach similar to the one proposed to learn the biological substitution matrix BLOSUM, our matrix learns the rates of matches and mismatches in conserved segments of shapes belonging to the same class. Furthermore, we include prior knowledge we have on the chaincode representation, accordingly weighing the elements of the S-BLOSUM. On one hand, the experimental evaluation showed that the S-BLOSUM provides a great improvement over the biological counterpart; on the other hand, classification results prove that our approach is competitive with respect to other methodologies for shape classification found in the recent literature.

References

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

[2] M. Yang, K. Kpalma, and J. Ronsin, "A survey of shape feature extraction techniques," *Pattern Recognition*, pp. 43–90, Nov 2008.

[3] S. Carlsson, "Order structure, correspondence and shape based categories," in *Shape Contour and Grouping in Computer Vision*, 1999, pp. 58–71.

[4] J. Gorman, O. Mitchell, and F. Kuhl, "Partial shape recognition using dynamic programming," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 10, no. 2, pp. 257–266, 1988.

[5] M. R. Daliri and V. Torre, "Robust symbolic representation for shape recognition and retrieval," *Pattern Recognition*, vol. 41, no. 5, pp. 1799–1815, 2008.

[6] M. Bicego, A. F. T. Martins, V. Murino, P. M. Q. Aguiar, and M. A. T. Figueiredo, "2d shape recognition using information theoretic kernels." in *ICPR*, 2010, pp. 25–28.

[7] P. Lovato and M. Bicego, "2d shapes classification using blast." in *SSPR2012*, ser. Lecture Notes in Computer Science, vol. 7626, 2012, pp. 273–281.

[8] M. Bicego and P. Lovato, "2d shape recognition using biological sequence alignment tools." in *ICPR*, 2012, pp. 1359–1362.

[9] P. Baldi and S. Brunak, *Bioinformatics - the machine learning approach (2. ed.).* MIT Press, 2001.

[10] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "The video genome," *CoRR*, vol. abs/1003.5320, 2010.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Addison-Wesley Longman Publishing Co., Inc., 2001.

[12] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *PNAS*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[13] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. of Mol. Bio.*, vol. 48, no. 3, pp. 443 – 453, 1970.

[14] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J of Mol. Bio.*, vol. 147, no. 1, pp. 195 – 197, 1981.

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipmanl, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 2, pp. 403–410, 1990.

[16] J. Thompson, D. Higgins, and T. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, pp. 4673–4680, 1994.

[17] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, 1978, vol. 5, pp. 345–352.

[18] G. Andreu, A. Crespo, and J. M. Valiente, "Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition," in *Proc. ICNN97*, vol. 2, 1997, pp. 1341–1346.

[19] N. Thakoor, J. Gao, and S. Jung, "Hidden markov model-based weighted likelihood discriminant for 2-d shape classification." *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2707–2719, 2007.

[20] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *CVPR'03*, 2003, pp. 409–415.

[21] R. A. Mollineda, E. Vidal, and F. Casacuberta, "Cyclic sequence alignments: Approximate versus optimal techniques," *IJPRAI*, vol. 16, no. 3, pp. 291–299, 2002.

[22] M. Neuhaus and H. Bunke, "Edit distance-based kernel functions for structural pattern classification," *Pattern Recognition*, vol. 39, no. 10, pp. 1852–1863, 2006.

[23] M. Bicego and A. Trudda, "2d shape classification using multifractional brownian motion," in *IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 2008, pp. 906–916.

[24] M. Bicego, A. F. T. Martins, V. Murino, P. M. Q. Aguiar, and M. A. T. Figueiredo, "2d shape recognition using information theoretic kernels." in *ICPR*, 2010, pp. 25–28.

[25] O. C. Hamsici and A. M. Martnez, "Bayes optimality in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 647–657, 2008.