Artificial Intelligence Techniques on Biological Structures

Alexiou Athanasios, Vlamos Panayiotis

Department of Informatics, Ionian University,
Plateia Tsirigoti 7, 49100 Corfu, Greece

{alexiou, vlamos}@ionio.gr

Abstract: We propose a Case-Based Reasoning System in order to response and manage the information concerning sequential terms of biological structures and the implementation of pattern recognition on secondary structures. In this paper, we will exclusively concentrate on the theoretical approaches and the fundamental principles for the design of a bio machine-learning system.

Keywords: Case-Based Reasoning, Motzkin Words, LCS, Secondary Structures

1 Introduction

Case-based reasoning is concerned with problems that are open-ended and solutions of new problems are mainly derived by adapting previously succesful solutions for similar problems. Several CBR models have been already integrated and succesfully applied to a wide range of scientific and technological applications in the health sciences. We can assume that there are four containers in which knowledge could be stored: the vocabulary used, the similarity measure, the solution transformation and the case-base [1]. In spite of the importance of similarity measures, clear methodologies for defining them efficiently and accurately are still missing. Instead, similarity measures are often defined in an ad hoc manner or one simply applies quite general distance metrics. When defining more complex measures that take account of domain knowledge, this is often done in an unstructured and not in a goal-directed fashion and often only experienced and skilled knowledge engineers are able to produce satisfactory results [2]. CBR systems produce satisfying solutions in weak theory domains, such as molecular biology, where the number and the complexity of the rules affecting the problem are very high and there is not enough knowledge for formal representation [3].
On the other hand, several approximation algorithms and techniques had been constructed, mainly with exponential complexity, concerning the enumeration of sequential terms of biological secondary structures, through the bijection with alternative representations like energy models, plane trees and Motzkin numbers, non-crossing set partitions, Motzkin paths and Dyck paths. In our case we will concentrate mostly to the Motzkin words and the identification of patterns, through common subwords and subdomains. The existance of building blocks in any case of combinatorial lexicographic word leads to important conclusions concerning relatives biological properties. A machine-learning approach seems to express an essential solution in order to merge the above combinatorial interpretations with techniques of biological pattern recognition. The proposed CBR model attempts to converge a

bioinformatics case-based knowledge with the LCS method in order to identify special motifs and cases in secondary structures' representations.

2 Combinatorics in Secondary Structures

We list some basic definitions and relations that will be used in the next sections of this paper. Two sets X and Y have the same cardinality if and only if there is a bijection from X to Y. In the case of Dyck words, these are words in the letters *x* and *y* with as many *x's* as *y's*, and with the property that any initial segment contains at least as many *x's* as *y's*. If we assume $d \in \{x, y\}^*$ is a Dyck word then $|d|_x = |d|_y$ and if d is factored as $d = mn$ then $|m|_x \geq |m|_y$. In a relative manner, a word $m \in \{x, y, a, b\}^*$ is called a Motzkin word if $|m|_x = |m|_y$ and if *m* is factored as $m = vw$ then $|v|_x \geq |v|_y$, or equivalently if the word obtained by deleting every occurrence of a, b from m is a Dyck word of $\{x, y\}^*$. It is important to define also the *n*-th Catalan number: $C_n, n \geq 0$ is defined by $C_n = \frac{1}{n+1}\binom{2n}{n}$. Let us now, refer to the set N_{2n} of nested pairs on *[2n]*, the set D_{2n} of all Dyck words of length $2n$, the set $\widehat{W}_{2n}$ of all *m*-Motzkin words, the set M_{2n} of all closed meanders order-*n* and the set $T^{(l)}(n)$ of all plane trees with exactly *n* leaves. The various bijections between those cardinalities have been already well defined and studied [4], in addition with the conclusion that secondary structures are in a simple bijection with Motzkin paths without peaks [5].

From the biological point of view, primary structure of RNA composed by linear polymers of four different nucleotides. Nucleotides consist of an organic base linked to 5`-carbon sugar (ribose) that has a phosphate group attached. The nucleotides used in synthesis of RNA contain one of four different bases, adenine (A), guanine (G), cytosine (C) and uracil (U). Differences in the sizes and conformations of the various type of RNA lead to specific functions in a cell. Secondary structures can be seen in single-stranded RNAs by pairing of complementary bases within a linear sequence. The secondary structure of an RNA molecule is the collection of base pairs that occur in its 3D structure. When the 5`-end of one nucleotide fits to the 3`-end of another nucleotides forms a p-bond, while the sequence of p-bonds declares the backbone of the molecules. On the other hand certain base pairs like C-G, A-U and G-U form h-bonds, which cause folding of the molecular backbone into configuration of minimal energy. A secondary structure of size n is closed [4] if there is an h-bond connecting bases *1* and *n* and for given integers $n \geq 2, l \geq 0$, there are $S^{(l)}(n-2)$ secondary structures of size n and rank l. Also in that paper had been proved, a bijection between the above set $Z^{(l)}(n)$ and the $T^{(l)}(n)$ (the set of all closed secondary structures and the set of all plane trees with exactly n leaves respectively). A more extended definition of closed secondary structures had been given [6], through the closed regions of a secondary structure. Representing a secondary structure as an arc diagram, in which base indices are shown as vertices on a straight line, ordered form the 5`-end and arcs (always above the straight line) indicate base pairs, a region $[i; j]$ will be refered as: weakly closed if it contains at least one base pair and for all base pairs $i' \cdot j'$ of R, $i' \in [i; j]$ if and only if $j' \in [i; j]$ and closed if either $i = 1, j = n$ or if it is weakly closed and for all l with $i < 1 < j$ the regions $[i;1]$ and $[1; j]$ are not weakly closed.

In many cases, like *ncRNA*, there are not completely identified and simulated, all the basic principles that occurs, the folding into secondary and tertiary structures. The incompleteness of the corresponding theories, contribute to a high complexity problem, where data mining, statistical analysis, biological interpretation and computational techniques must incorporate in different phases, in order to achieve solution. These multidimensional principles and methods can be automated and included in a CBR system as the main components of the base knowledge. The basic combinatorial terminology will evaluate the identification of important motifs from the LCS similarity teacher and will classify users' inputs, composing experience for future cases. Similar cases and the identification of multiple alignments whose expression patterns have meaningful relationships and influence physiological bio-functionalities, consist the main objective of our Bioinformatics CBR model.

3 A Bioinformatics CBR Model

According to the explanation goals in CBR systems [7], transparency, justification, relevance, conceptualization and learning, we will base our research to the accepted model consisting of the four basic steps retrieve, reuse, revise and retain. In our case study, secondary structures' enumaration problem and the relevant representations between the cardinalities of plane trees and Motzkin numbers, non-crossing set partitions, Motzkin paths and Dyck paths seems to be well defined, therefore no further extensions have to be determined. While the traditional CBR cycle does not explicity introduce a feedback loop, it seems to have a relative situation in our bio-model, due to the fact that in most of the cases, we can get exact solutions in biological functionalities problems. Nevertheless we shall include a feedback utility among an external similarity teacher and the biological knowledge (generalizing the retain step). We have to mention that clustering and feature selection techniques have been successfully applied to CBR maintenance [8]. For instance, methods based on condensed nearest neighbor (CNN), or fuzzy decision trees. Also learning feature weights can be considered as an example of similarity maintenance. The system in these cases, is based on interactive user responses to the system's behavior and asks user to adjust feature weights for a set of cases, and applies the weights during case retrieval. There are also references [9,10] on applying weights mainly using DFS lexicographical ordering, for identifying common chemical structures in chemical datasets or in designing closed curves.

3.1 The proposed model

It is obvious that the refered model consists of an application environment, combining a mixture of experts samples and a LCS similarity teacher. This mixture of experts is based on the corresponding classification of independent biological sequential samples separetely. The main task of the proposed CBR system (Fig.1) is to provide acceptable solutions either on single cases or perfoming pattern analysis on multiple biological representations. Nevertheless, it is known that RNA structure is often more conserved than the sequence during evolution. Through phylogenetic comparative analysis it is very important to remark, that all RNAs fold into a similar secondary structure, concluding that functionality is essential to structure. Therefore, the combinatorial consideration of folding proteins and the alternative description of secondary structures adapting machine-learning techniques and using case-based knowledge, leads to an accurate case.

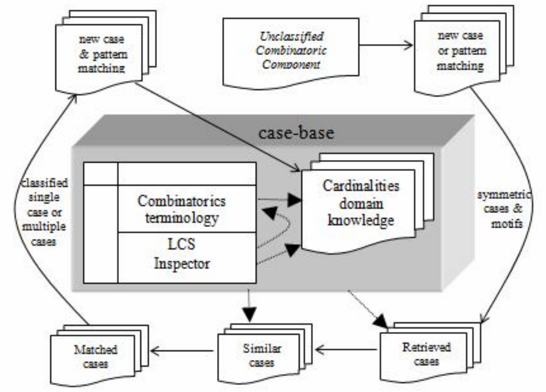

Fig.1. The proposed Bioinformatics CBR system

As we have already mention, this model uses the general CBR cycle processsing for the adaption of its structural component: *Retrieve* (the most similar case or cases), *Reuse* (information and knowledge to solve the problem), *Revise* (the proposed solution) and *Retain* (experience for future problem solving).

The problem characterizes the transaction between combinatoric elements of secondary structures, in order to establish similarity measures on multiple sequences, identify repeated motifs and classify significant patterns for future use. The initial unclassified situation-case *uc* must be determined using the various definitions of Section 2, performing an accurate and useful solution information.

In the retrieval function of the model, we shall take into consideration the assumptions of similarity values and properties between relevants and symmetric components on Motzkin words. We will adapt the approach, of acquiring training data through some similarity teacher. As we have already mentioned there is no special treatment for a specialized solution's feedback loop, in our model. An independent LCS similarity teacher will provide simulation procedures and process certain knowledge among the separetely samples of the correspondance sets (on the enumeration of biological structures ie. multiple RNA alignments). The feedback utility of the LCS part, will motivate the credibility of the CBR model through the modification of the knowledge containers (cardinalities, combinatorics terminology). A set of new cases *nc* or similar cases *sc* can be generated, corresponding to the final output, contributing in a way to the existence knowledge, regarding additional explanations about bijections between combinatorics representations and biological secondary structures. The CBR system execute and extract the final result in a friendly manner, providing information about the initial hypothesis classification. The proposed system simulates the combination of human operations and biological data mining, concluding to decisions automatically, reducing the required time and the user errors.

3.2 The LCS Inspector

A first approach for comparing RNAs secondary structure using LCS metrics was introduced under the notation *longest common subsequence for arc-annotated sequences*. This idea has been already optimized [11], considering the local normalized LCS metric for RNA sequences which measures the highest LCS scoring consecutive subsequences divided by their length. Also the idea of applying the LCS approach on RNA multiple alignment has been presented in the same paper via polynomial $O(n^2)$ time algorithm, investigating common folding patterns or secondary structures. While the number of longest common subsequences in many biological applications seems to be quite large, it is believed [12] that finding merely a longest common subsequence is not quite meaningful. In fact, finding a longest common subsequence satisfying a useful property must be the objective of any proposed technique-model. There is a study on hierarchical categorized of folding [12] refering to: maximum nested loop, maximum loop chains and maximum number of total matches.

We will extend the various techniques of LCS on RNAs sequences and the representations of secondary structures through Dyck paths and Motzkin paths, directly to Motzkin peakless words as sets of paired bases and arcs.

Let us denote two Motzkin words $m_1, m_2 \in \{x, y, a, b\}^*$ words in the set of $\overset{\frown}{W}_{2n}$. A subword m' is a common subword of m_1, m_2 by exact matching of the corresponding lexicographic sequences either by omitting any case of different adjacent paired bases in both subwords or arc deletions. The LCS of m_1, m_2 is a common subsequence of maximum cardinality.

Let us consider $m_1, m_2 \in \{x, y, a, b\}^*$, Motzkin words in the set of $\overset{\frown}{W}_{2n}$. The *LCS* of m_1, m_2 denoted by $LCS(m_1, m_2)$, are the common subsequences of m_1 and m_2 with the maximum exactly *k* matches, where $1 \le k \le 2n$.

If $m_1, m_2 \in \{x, y, a, b\}^*$ are Motzkin words in the set of $\overset{\frown}{W}_{2n}$ given the $LCS(m_1, m_2)$, then any symmetric words m_1', m_2' have the same longest common subsequence. Similar longest common subsequence can be obtained in relative representation of Dyck words.

It is very important to mention for any future research, that the graphical represenation of closed secondary structures through Motzkin paths without peaks, and under the assumption of omitting any unpaired bases and arcs, can be extended to closed meanders and system of closed meanders and vice versa.

It is quite obvious that the identification of repeated structrural motifs occuring certain graphical limitations may cause also biological unfunctionalities in the structure of the informational molecules. RNA molecules sometimes interact with proteins and other specific molecules, having common topological motifs. These identified motifs should be tested for their existance in additional sequences that could form similar structure. The proposed LCS similarity teacher in our model, provides these tools through the feedback of the identically matched cases and motifs in the biological knowledge and the part of combinatorics terminology.

4 Conclusion

As we have already mentioned, problems concerning representations of certain biological structures like secondary structures, either are characterized as NP-complete or with high complexity. In this study we proposed a theoretical combination of a machine-learning technique, with the basic combinatorics' terminology and the LCS method as a suitable and user friendly solution for accessing biological data and manage pattern recognition and mathematical modelling. Future research will proceed to the implementation and integration of this CBR model.

References

1. Richter, M.: In Case-based reasoning: experiences, lessons and future directions. Springer (1998) 1-15
2. Stahl, A.: Learning Similarity Measures: A formal view based on a generalized CBR model. Springer (2005), ICCBR 2005, LNAI 3620, 507-521
3. Jurisica, I., Glasgow, J.: Application of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special Issue on Bioinformatics 25(1) (2004) 85-95
4. Doslic, T., Veljan, D.: Secondary structures, plane trees and Motzkin numbers. Mathematical Communications 12(2007), 163-169
5. Deutsch, E., Shapiro, L.: A bijection between ordered trees and 2-Motzkin paths and its many consequences. Discrete Mathematics 256 (2002, 655-670)
6. Rastegari, B., Condon, A.: Linear time algorithm for parsing RNA secondary struvture. Springer (2005), WABI 2005, LNBI 3692, 341-352
7. Sormo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning- Perspectives and Goals. Artificial Intelligence Review (2005) 24, 109-143
8. Jurisica, I., Arshadi, N.: Maintaining case-based reasoning systems: A machine learning approach. Springer (2004), ECCBR 2004, LNAI 3155, 17-31
9. Li, K. Rahman, R., Gupta, A., Siddavatam, P., Gribskov, M.: Pattern matching in RNA structures. Springer (2008), ISBRA 2008, LNBI 4983, 317-330
10. Alexiou, A.: Meandric curves and the TSP. Masaum Journal of Basic & Applied Science (2009), Vol.I, Iss.I, 1-4
11. Backofen, R., Hermelin, D., Landau, G., Weimann, O.: Normalized Similarity of RNA Sequences. Springer (2005), SPIRE 2005, LNCS 3772, 360-369
12. Bereg, S., Zhu, B.: RNA multiple structural alignment with Longest Common Subsequences. Springer (2005), COCOON 2005, LNCS 3595, 32-41