Large-scale macromolecule classification and clustering by 2D structure bitmap image analysis

Ilkka Havukkala
Knowledge Engineering & Discovery Research Institute
Auckland University of Technology, Private Bag 92006, Auckland
1020, New Zealand
E-mail: ilkka.havukkala@aut.ac.nz

***Abstract*—The computational problems of exhaustive structural comparisons of large numbers of macromolecules by atomic coordinate data are discussed. Potential solutions for simpler proxy variables are explored, with non-coding RNA structures as an example. A new method of analyzing computed 2D structures by bitmap image processing is suggested. Implications for large-scale discovery, clustering and classification of macromolecular structures are discussed, with suggestions for future research.**

I. INTRODUCTION

The classification and comparison of protein 3-dimensional structures has engaged computational biochemists for decades, with development of sophisticated methods for alignment of corresponding amino acids and atoms in the proteins being compared, with a multitude of web servers on the internet for comparing up to a few proteins with each other [1]. Similar molecular docking methods are used in comparison and classification of other complex organic molecules as well, for which accurate molecular coordinates are available. These software applications are crucial for drug design, *e.g.* ligand to protein binding simulation, but are not easy to scale to large numbers of molecules. This has led to massive distributed computing efforts, like FightAIDS@Home on the World Community Grid, which performs AutoDock analysis of drug and HIV virus target matching on local PCs around the world [2].

Alternative quick methods have been developed, for example virtual screening of molecule candidates in the pharmaceutical industry by the well-known Lipinski rules or other proxy variables [3]. However, in general, the problem of matching, clustering and classifying large numbers of molecular structures efficiently has not been solved satisfactorily (but see one recent promising approach for proteins in [4]).

For proteins, amino acid sequence similarity of 30% is needed for reliable matching by BLAST, FASTA and other algorithms, but structural similarity (by comparison of conserved amino acid coordinates) can still be detected with only 10% conserved amino acids. Therefore sequence alignments cannot reveal all possible evolutionarily conserved structures. Also, current algorithms also have no solutions on how to handle contradicting evidence from amino acid alignment and structural alignments.

These problems are general for all macromolecules, also for DNA and RNA structures. The folding of the RNA molecules is also known to be often more conserved than their sequence. Thus the question is what kind of proxy variables, instead of the very complex (and accurate) 3D atomic coordinates, might be utilized for comparing and clustering of macromolecule structures? We approach this problem with RNA molecule structures as an example.

II. RNA STRUCTURE COMPARISON METHODS

A. Previous approaches

RNA molecules are known to fold upon themselves, so that matching nucleotide pairs A-U and C-G are formed, producing stable conformations with shapes characteristic to the type of RNA molecule, e.g. ribosomal RNAs, microRNAs, riboswitches and so on. Purely computational prediction of quite reliable secondary 2D structures from RNA sequence alone is possible, the latest method being fast, too [5]. Advanced algorithms can produce a set of most likely conformations, known as the Bolzmann ensemble [6]. Commonly the most stable structure with the lowest thermodynamic energy (ΔG) is considered to represent the biologically active form. This may not always be the case, as alternative conformations are known to exist for some special types of RNAs, and proteins or other molecules attached to the folded RNA may affect conformation, as well as changes within the cell in pH, ion concentrations, temperature etc. Very recent results [7] for mRNA Bolzmann ensembles indeed suggest relying on only the most stable conformation is often misleading.

As reviewed by [8], three traditional methods have been used to arrive at consensus structures from a limited set of RNA sequences (Figure 1, A, B, C): A, aligning sequences, followed by structure folding for the multiple sequence alignment, B, simultaneous sequence and fold alignment (Sankoff method), and C, folding sequences, followed by aligning structures.

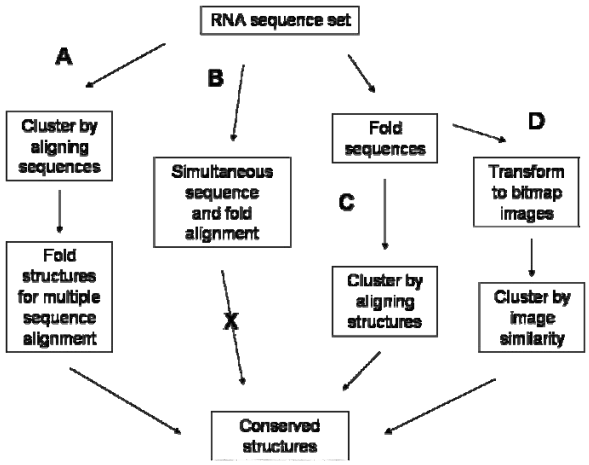

Figure 1. Methods of Clustering RNA Structures. Path D is the New Generic Method Suggested in This Paper Applicable to Other Macromolecules as Well.

Method A suffers from the problem of not clustering together all related sequences, as RNA structure is moreconserved than its sequence, as in proteins. Method B does not allow clustering of sequences/structures (marked with X in Figure 1) and has prohibitive time complexity of computation. Method C depends crucially on the method to align structures, and is still a young developing field. The algorithms used are based on the concept of RNA as a topological graphs [9] or trees. The main algorithms developed so far are RNAFORESTER and MARNA (reviewed in [8]) and the recent TREEMINER [10] and RSMATCH [11]. Their performance in analysing and clustering very large RNA sets is not yet known.

B. A new method

Method D in Figure 1 is the new method proposed in [12] and applied there to microRNA classification by Gabor filter features. In essence, it is based on transforming the simulated 2D structure to a sufficiently high-resolution bitmap image, and then analysing it with a suitable image clustering method. In principle the whole gamut of image analysis methods used in *e.g.* face recognition, fingerprint identification and other fields could be used, opening up a whole new toolbox to attack the problem. The simulated structure could be encoded in various ways into a bitmap image to ease pattern recognition in the images, *e.g.* one could show all adenines and uracils (As, Us) as circles, Cs and Gs as squares, highlight the individual bases or hybridizing basepair linkages with specific colours, and so on.

III. APPLICATIONS TO OTHER MACROMOLECULES AND FUTURE DIRECTIONS

The new method of using bitmap images of 2D molecular structures for clustering and classification applies to any macromolecules, for which a visualization of structure can be obtained, either by computational means, or by direct experimentation using microscopes, x-rays, etc. The approach is attractive in that the image implicitly contains a large number of possible features that modern sophisticated image analysis programs can extract. Ideally, all likely alternative structures (like Bolzmann ensemble for RNA) for each molecule should be included in the analysis.

The information obtained from images can then of course be used in conjunction with other computed proxy variables from the 2D structure, like the area, number and size of loops, A/T ratio, ΔG etc. for RNA structures, or molecular weight, hydrophobicity, number of serines etc. for proteins and so on. This new method is complementary to any previous existing methods, and could be used simultaneously to extract any informative classifiers. Thus the problem moves to the domain of artificial intelligence and data mining of large, high-dimensional datasets, where many molecules have large numbers of features for clustering. This augurs well for the further integration of the bioinformatics and artificial intelligence communities to join forces in data mining and knowledge discovery in large biological databases of organic molecules, proteins, and RNAs, which are coming from the current tsunami of information from metabolomics, proteomics and genomics.

REFERENCES

[1] K. Vlahovicek, A. Pintar, L. Parthasarathi, O. Carugo & S. Pongor S. CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures. *Nucleic Acids Research,* vol. 1:33, pp.W252-4. (2005)
[2] FightAIDS@Home, http://fightaidsathome.scripps.edu/index.html
[3] T. I. Oprea. Virtual Screening in Lead Discovery. *Molecules*, vol. 7, pp. 51–62 (2002)
[4] D. Lupyan D, A. Leo-Macias & A.R. Ortiz, A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics, vol.*;21(15), pp. 3255-63 (2005)
[5] S. Washietl, Hofacker I.L., Stadler P.F.P. & Tino, P. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA*, vol. 15:102(7), pp. 2454-24599.(2005)
[6] Y. Ding, & C.E. Lawrence, A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research,* vol. 31, pp. 7280-7301 (2003)
[7] Y. Ding, C. Chan & C.E. Lawrence, Clustering of RNA secondary structures with application to messenger RNAs. *J. Mol. Biol.* (in press)
[8] P.P. Gardner & R. Giegerich, A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics, vol. 5:140, 18 pp. (2004)
[9] H.H Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim & T. Schlick, RAG: RNA-As-Graphs database--concepts, analysis, and features. *Bioinformatics,* vol. 20, pp.1285-91 (2004)
[10] M. J. Zaki: Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. *IEEE Trans. Knowl. Data Eng.,* vol. 17(8), pp. 1021-1035 (2005).
[11] J. Liu, T.L. Jason, T.L. Wang, J. Hu & B. Tian, A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics.* vol. 6:89, 20 pp. (2005).
[12] I. Havukkala, S.N. Pang, V. Jain & Kasabov, N. Classifying microRNAs by Gabor filter features from 2D structure bitmap images on a case study of human microRNAs. *J. Comput. Theor. Nanosci.,* vol. 2(4), pp. 506-513 (2005).