ERCIM Alain Bensoussan Fellowship Programme

ABCDE

MARIE CURIE ACTIONS

Scientific Report

First name / Family name	Avinash Achar
Nationality	India
Name of the *Host Organisation*	Norwegian University of Science and Technology
First Name / family name of the *Scientific Coordinator*	Pål Sætrom
Period of the fellowship	18/04/2012 to 17/04/2013

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Classically, RNA molecules are known to be messengers from the genome for protein synthesis. Such RNA have been referred to as mRNA in short. In the last decade or so, a number of non-coding RNA (which do not code for proteins) have been discovered and the exploration for new non-coding RNA is still on. The ncRNA have been found to play an active role in gene regulation, DNA replication and so on. An RNA is a linear chain of smaller molecules called nucleotides (which are of four types). The linear chain of nucleotides fold onto itself and forms many double-stranded regions by additional hydrogen bonds. Such a self-folded two dimensional structure is called a Secondary structure. RNA molecule in general folds into a three-dimensional structure. Given the complexity in handling the entire three dimensional structure, it is common to work with the two dimensional secondary structures of the RNA molecules. Unlike in the case of protein molecules where sequence similarity is alone an indication for a common biological function, in the case of RNA molecules, one needs to consider both sequence and secondary structure-based similarity to infer common function.

My project in the host institute broadly pertained to the discovery of common frequently occurring (Secondary structure based) local patterns from a set of RNA (Ribose Nucleic Acid) molecules or sequences. The sequences for instance could correspond to the 5' or 3' Untranslated regions (UTR) of the mRNA coming from a set of orthologous or co-expressed genes. These untranslated regions of orthologous genes are known to house a variety of regulatory motifs like Cobalmain, Lysine, glms, Iron Response Element, SECIS, Histone3 and so on. The first three examples are specifically referred to as riboswitches as they directly influence the production of proteins encoded by the mRNA on which they reside. The sequences could also correspond to a set of in vitro selected RNA like the SELEX type data. The sequences are basically randomly generated single-stranded RNA that specifically bind to a target ligand locally owing to the existence of a common motif among the sequences. The sequences could correspond to a set of ncRNAs coming from the same family or any other set of sequences where one hopes for some common secondary structure based motifs.

The RNA literature has seen a wide spectrum of computational methods which tackle the above problem. We give a detailed review of most of the existing methods that address the motif discovery problem. We discuss the underlying computational principles of each of the methods highlighting the similarties and differences. We also provide some experimental comparisons of some of the state of the art methods. To the best of our knowledge, this is the first review which gives a computational overview of the existing methods tackling RNA motif discovery.

We start off with an overview of RNA secondary structure. We explain the various basic structural motifs like loops, stems, bulges etc. displayed by RNA secondary strucutres. We describe the notion of a pseudoknot in any secondary structure. This concept is important in our context as majority of the existing algorithms can only tackle secondary structures without pseudoknots. We describe the tree representations at different resolutions for any RNA secondary structure without pseudoknots. We also briefly touch upon the dual-graph based representation for any general secondary structure (with or without pseudoknots).

The computational methods for RNA motif discovery can be broadly categorised into four categories: (1) Stochastic (2)Stem-based (3)Alignment-based and (4)Miscellaneous.

The stochastic methods pose the motif discovery as a probabilistic learning problem. We discuss three methods in this category. Each of the methods have a different way of expressing a motif. MEMERIS essentially looks for sequence motifs in single-stranded regions. CMfinder discovers local motifs based on covariance models, which are a very popular way to model families of related RNA. RNApromo is another learning method discovering short motifs based on stochastic models similar to covariance models. The stem-based methods essentially view motifs as a completely connected graph of stems. The methods essentially extract stem graph patterns which occur frequently enough in a majority of the input sequences. We describe two methods under this category. The first method, comRNA uses a slightly ad-hoc method whereas RNAmine uses a more principled approach based on a graph mining idea. The alignment based methods pose the discovery problem as a sequence and structure based local alignment problem. The methods in this category can be further divided into two categories: (a) secondary structure independent and (b) secondary structure dependent. The secondary structure dependent methods need the secondary structure information of the input sequences. Under the first category, we discuss three methods. FOLDALIGN, one of the earliest methods can only discover motifs without branching. With branching, it can only discover motifs from two input sequences. LocARNA, is another method which builds on a well-known existing secondary structure based alignment algorithm by utilizing the base pair probabilities in defining the scoring function. SCARNA_LM is another interesting method which uses Conditional Random Fields to model alignments. We also discuss a few alignment methods which need the secondary structure information and most of them can only handle a pair of sequences. The miscellaneous category includes methods which cannot be placed in any of these well-defined categories. It includes a method where motif discovery is posed as a frequent tree mining problem where each RNA input is represented by a tree. We also discuss a method which uses genetic algorithms for motif discovery.

We made experimental comparisons of some of the above algorithms, which could discover large motifs and handle more than two input sequences. For benchmarking, we used datasets from the Rfam database where sequences are grouped based on families and one knows the ground truth of the underlying common secondary structure. Since we are looking at local patterns, we tested and compared the performance of the algorithms with a gradual increase in the length of extraneous flanking regions on either side of the true motif locations. We also give performance comparisons on gradual increase of the number of noisy sequences (which do not contain the motif of interest).

In addition to my main project described above, I also had the opportunity to be a part of another project during my first REP. Algebraic Dynamic Programming (ADP) is a tool which automates the very process of writing a program to solve a combinatorial optimization problem having a dynamic programming solution. The work carried out here demonstrated how ADP could be used in automatically computing any general expectation based feature of a given RNA sequence (for eg: the number of base pairs), without having to manually reprogram with every change in the feature function.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

1. *Avinash Achar, Pål Sætrom:* "RNA Motif Discovery: A computational overview" - to be submitted soon to the journal, BMC Bioinformatics (pending).

2. *Yann Ponty, Cedric Saule and Avinash Achar :* "Efficient computation of Boltzmann ensemble features using grammar derivation and algebraic dynamic programming" – to be submitted to the journal, Bioinformatics (pending).

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

1. NTNU bioinformatics network seminar 2012
   Date: 18–19 October 2012
   Location: Laboratory Centre, NTNU/St. Olavs hospital, Trondheim

2. ABCDE seminar II
   Date: 24-25 October 2012
   Location: Sophia Antipolis, Inria

IV – RESEARCH EXCHANGE PROGRAMME (REP)

REP1:
Location: INRIA (Ecole Polytechnique)
Country: France
Department: Computer Science
Local Scientific Co-ordinator: Dr. Yann Ponty (CNRS junior research scientist),
Associate Member of INRIA AMIB team.
Dates: 15.01.2013 - 25.01.2013

My experience here was more than pleasant. As I already mentioned in the scientific activity, I got an opportunity to be a part of a project here. Also, with the kind of interactions I had, I got to learn more about the field in general. I could avail the guest house accommodation in the Ecole Polytechnique premises which made my stay extremely comfortable. It was a pleasure working with Dr. Yann.

REP2:
Location: University of Vienna
Country: Austria
Department: Theoretical Chemistry
Local Scientific Co-ordinator: Prof. Ivo Hofacker
Dates: 19.03.2013 - 25.03.2013

My experience here was also very pleasant. I learnt a lot from the interactions I had with Prof. Hofacker and some of his students. It also gave me some useful leads to my main project back home at NTNU. It was a group working on diverse aspects of RNA research.