Summary Report of My Scientific Activities during ERCIM Postdoc fellowship at NTNU

Osman Abul
Department of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology, Trondheim, Norway
osman.abul@ntnu.no

May 18, 2006

1 Introduction

This report is prepared to describe my scientific activities during the ERCIM postdoc fellowship at Norwegian University of Science and Technology (NTNU), between September 5, 2005 and June 4, 2006.

I have been affiliated with the Bioinformatics group of Department of Cancer Research and Molecular Medicine (IKM) of NTNU. The group is headed by Professor Finn Drabløs and consisted of 2 Postdocs, 2 PhD students and 1 Programmer. The group has several collaborations, including Department of Computer and Information Sciences (IDI) of NTNU, Interagon AS., and 2 Norwegian universities (Bergen and Oslo) through National Program for Bioinformatics (FUGE) project. The research interest of the group includes

- Motif and motif modules discovery from biosequences,
- Biological networks,
- Ligand design,
- DNA repair,
- Fold recognition of proteins, and
- Homology modeling of proteins.

Before coming to NTNU, my research interest was in induction and control of regulatory networks (a subclass of more general biological networks). Since motif and motif modules discovery are essential to regulatory networks, I have mainly focused on the motif and motif modules discovery from DNA sequences at NTNU.

In collaboration to IDI department, a motif discovery group, where I was a member, has been formed. We have started several research projects within the group. Although almost all are mature enough for the objectives we set they are still open to extensions, thus offering basis for future collaborations between me and the group. In addition, there are other directions for further collaborations in related fields which require gained expertise in motif discovery, such as regulatory networks, post-transcriptional issues, and in general systems biology. Besides, motif discovery in itself is still active research area especially for higher eukaryotic genomes.

The document is organized as follows. Section 2 gives a summary of research activities I have been involved. In Section 3, the scientific conferences/seminars and meetings to which I attended as audience or presenter are given. Section 4 gives abstracts and co-author list for accepted and submitted publications.

2 Summary of Research

2.1 Methodology for Motif Discovery

There are over one hundred methods (and respective tools) proposed for motif discovery from biosequences. The success of these methods depends on several factors, counting a few 1) how well their assumption of biological motifs coincide with the true biological motifs, 2) search strategies, 3) motif representation model, 4) properties of data set (size, homogeneity, etc), 5) utilization of auxiliary data (gene expression, foot-printing, gene ontology, etc). In this work, we have proposed that methodology employed is also matters and proposed a tool independent methodology for motif discovery. Our results indicated that the methodology boost motif discovery. See Section 4.1 for publication information.

2.2 A New Method for Motif Discovery

Current motif discovery algorithms find motifs using a custom motif scoring function to rank alternatives. However, there are many scoring functions to choose from each covering some notion of biological motifs and none of them are able to cover all notions. Noting this, we have defined a new two-step paradigm for motif discovery in which each step scores motif candidates with independent scoring functions. First step is supposed to dramatically reduce number of motif candidates by filtering out non-promising ones and second step decides motifs by scoring remaining ones. We also have proposed a word-counting based motif discovery method named *TScan* taking this paradigm. *TScan* is inspired by *MDScan* method, but it is aimed applicable even when ChIP-chip data are unavailable. See Section 4.2 for publication information.

2.3 False Discovery Rates in Identifying Functional DNA Motifs

Previous library based motif discovery methods are motivated for finding over and under represented motifs against a set of promoter sequences. In case we consider a library of motifs (like TRANSFAC and Jaspar), one can test over and under representations for large number of motifs simultaneously. Even though the issue is addressed in the literature from multiple-hypothesis testing perspective, the issue of assigning confidence to over and under represented motifs are not studied. This work is primarily motivated for assigning confidences (equivalently, false discovery rates) and identifying over and under represented motifs within given confidence. It is considered that the issue is important as transcription factors targeting to specific motifs are further experimented *in vivo* for the set of promoters. Clearly, bounding confidence directly corresponds to bounding labor.

We have written a draft paper describing the method and initial results. After elaboration and more experimentation we will disseminate it.

2.4 Development of a Web-service for Running Motif Discovery Tools

Some motif discovery tools are developed as form-based web applications. The difficulty in using them is requiring data and parameter values manually entered. This is usually not a problem if they are going to be run once, but a serious problem if hundreds of separate runs are desired. We have developed a web service facilitating automatic invocation and result parsing of 3 motif discovery tools, MEME, BioProspector and MDScan. The novelty of our approach is allowing automatic accesses to these tools programmatically as subroutines while developing large scale systems biology applications.

A manuscript describing the web service is prepared but has not been submitted for publication yet.

2.5 Assessing Motif Modules Discovery Tools

Motifs usually occur in combination with others rather than in isolation, thus forming motif modules or clusters. A number of tools has been developed for this task, but the literature lacks the assessment of these tools. Currently, we have started and still working on development of a benchmarking model for assessment. We aim this work will pinpoint the deficiencies and thus cause development of new methods or extension of existing ones

addressing these deficiencies.

2.6 Student Thesis Projects

There are several bioinformatics research projects going on at IDI department. These projects are mostly defined for master students as their master thesis. In collaboration with IDI department, I have been involved in (mainly) a couple of these projects.

The first project is about finding the best motif representation model. There are a number of models suggested in the literature, among them most renowned ones are probabilistic weight matrices, mismatch strings and IUPAC codes. To facilitate the study, we have casted the problem as model representation bias learning problem and evaluated the 3 models on selected data sets. See Section 4.3 for publication information.

The second project is about speeding up motif discovery using parallel hardware. Running motif discovery programs such as MEME on large data sets take more than a day on a single CPU. Noting this, parallel versions of these programs are developed to run them on supercomputers and large computer clusters. But run time is still long on relatively larger data sets. In this work, we have defined an abstract parallel module accelerating motif discovery and implemented it on a low cost parallel hardware called Pattern Matching Chip. See Section 4.4 for publication information.

2.7 Continuation of Previous Work

We have prepared two papers (See Section 4.5 and 4.6 for publication information) based on previous work, mostly from my PhD thesis. During the fellowship the works are further developed.

3 Conferences/Seminars and Meetings

I attended following two recurring meetings.

- Weekly IKM meetings: Scientific meeting where department members present their own research and results in turn. I gave a presentation.

- Biweekly motif discovery group meetings: Participants (4 from our bioinformatics group and 2 from IDI department) present recent research from literature in turn.

I attended/will attend following non recurring meetings and seminars.

- Innovation In Norwegian Biotechnology Seminar (Technoport), October 21, 2005, Trondheim.
- Bioinformatics Week (Seminars and PhD thesis defenses at NTNU), November 8-10, 2005, Trondheim. I gave a presentation.
- The 2005 Gene Expression Seminar Series (Applied Biosystems), November 23, 2005, Trondheim.
- Teveltunmøte: Biochemical society seminar (NTNU), November 1, 2005, Teveltunnet.
- Workshop during Martin Kuipers stay at NTNU, March 21, 2006, Trondheim.
- Norwegian Bioinformatics Forum, May 29-30, 2006, Trondheim.

4 Publications

4.1 Paper 1

Title: A Methodology for Motif Discovery Employing Iterated Cluster Re-assignment.

Authors: Osman Abul, Geir Kjetil Sandve, and Finn Drabløs

Abstract: Motif discovery is a crucial part of regulatory network identification, and therefore widely studied in the literature. Motif discovery programs search for statistically significant, well-conserved and over-represented patterns in given promoter sequences. When gene expression data is available, there are mainly three paradigms for motif discovery; *cluster-first, regression*, and *joint probabilistic*. The success of motif discovery depends highly on the homogeneity of input sequences, regardless of paradigm employed. In this work, we propose a methodology for getting homogenous subsets from input sequences for increased motif discovery performance. It is a unification of *cluster-first* and *regression* paradigms based on iterative cluster re-assignment. The experimental results show the effectiveness of the methodology.

Status: Accepted for publication in the Proceedings of Computational Systems Bioinformatics (CSB 2006) conference, August 14-18, 2006, Stanford University, California.

4.2 Paper 2

Title: TScan: A Two-step *De novo* Motif Discovery Method.

Authors: Osman Abul, Geir Kjetil Sandve, and Finn Drabløs

Abstract: Computational discovery of novel motifs in biological sequences is an important and well-studied problem. The key to motif discovery methods, either *de novo* or library based, is having well-defined scoring functions. Several different scalar valued scoring functions have been proposed that measure some notion of biological motifs; that is we lack a perfect one capable of measuring of all notions together. In this work, we propose a two-step *de novo* motif discovery paradigm employing two scoring functions measuring different notions of biological relevance. We define a word-counting based method, called *TScan*, taking this paradigm. It is mainly inspired from *MDScan*, but does not require supplementary ChIP-chip data. Our results on seven data sets from a recent study are promising, with discovered motifs agreeing well with the consensus motifs defined for the data sets.

Status: Submitted for publication to The Third Annual RECOMB Satellite Workshop on Regulatory Genomics conference, July 17-18, 2006, National University of Singapore, Singapore.

4.3 Paper 3

Title: Bias Analysis of Motif Models for Biosequences.

Authors: Geir Kjetil Sandve, Osman Abul, Vegard Walseng, and Finn Drabløs

Abstract: The discovery of motifs in biosequences is an important problem and has in recent years attracted much research interest, resulting in more than hundred tools. The

algorithms differ in how motifs are modeled (motif representation language) and their search strategies. Though there are lots of motif discovery tools, there are only a few commonly used motif models, *e.g.* IUPAC expressions, mismatch strings and position weight matrices (PWMs). In the literature, several benchmark studies have been carried out for measuring the merit of motif discovery tools. However, to date no systematic analysis of the underlying motif models has been performed, and it is not clear how suitable the different motif models are for representing binding sites in DNA sequences. By casting this problem as a bias learning problem, we are able to analyze it in a theoretically sound framework. Results on benchmark data indicate that IUPAC expressions and PWMs have similar performance, and that these two models outperform mismatch strings.

Status: Submitted for publication to 6th Workshop on Algorithms in Bioinformatics (WABI 2006), September 11-13, Zurich, Switzerland.

4.4 Paper 4

Title: Accelerating Motif Discovery: Motif Matching on Parallel Hardware.

Authors: Geir Kjetil Sandve, Magnar Nedland, Øyvind Bø Syrstad, Lars Andreas Eidsheim, Osman Abul, and Finn Drabløs

Abstract: Discovery of motifs in biological sequences is an important problem, and several computational methods have been developed to date. One of the main limitations of the established motif discovery methods, is that the running time is prohibitive for very large data sets, such as upstream regions of large sets of cell-cycle regulated genes. Parallel versions have been developed for some of these methods, but this requires supercomputers or large clusters of computers. Here, we propose and define an abstract module PAMM (Parallel Acceleration of Motif Matching) with motif matching on parallel hardware in mind. As a proof-of-concept, we provide a concrete implementation of our approach called MAMA. The implementation is based on the MEME algorithm, and uses an implementation of PAMM based on specialized hardware to accelerate motif matching. Running MAMA on a standard PC with specialized hardware on a single PCI-card compares favorably to running parallel MEME on a cluster of 12 computers.

Status: Submitted for publication to 6th Workshop on Algorithms in Bioinformatics (WABI 2006), September 11-13, Zurich, Switzerland.

4.5 Paper 5

Title: An Optimal Multi-Objective Control Method for Discrete Genetic Regulatory Networks.

Authors: Osman Abul, Reda Alhajj, and Faruk Polat

Abstract: The control problem for discrete genetic regulatory networks is a major problem that has not been investigated enough yet, even though companion induction problem has been extensively studied. In this paper, we study this problem and note that the problem is actually multi-objective and thus develop an optimal multi-objective approach. Our approach includes formalization of components and identification of dimensions, resulting in a few cases for concrete problem formulation. For a selected case, namely the finite control case, a single-objective from the literature and our multi-objective solutions are presented. It is shown that multi-objective solution avoids drawbacks of single-objective solution, particularly need for defining single objective out of many. The effectiveness/applicability of the proposed approach is demonstrated on selected examples.

Status: Submitted for publication to International Conference on Computational Methods in Systems Biology, October 18-19, Trento, Italy.

4.6 Paper 6

Title: Asymptotical Lower Limits on the Required Number of Examples for Learning Boolean Networks.

Authors: Osman Abul, Reda Alhajj, and Faruk Polat

Abstract: This paper studies the asymptotical lower limits on the required number of samples for identifying Boolean Networks. It is previously given as $\Omega(logn)$ in the literature

for fully random samples. It has also been found that $O(logn)$ samples are sufficient with high probability. In this work, the main motivation is to provide lower asymptotical limits for samples obtained from time series experiments. Using the results from literature on random boolean networks, lower limits on the required number of samples from time series experiments for various cases are analytically derived using information theoretic approach.

Status: Submitted for publication to The 21st International Symposium on Computer and Information Sciences (ISCIS 2006), November 1-3, 2006, Istanbul, Turkey.