# STATEMENT OF PURPOSE

Obtaining the highest academic accolade, Ph.D., has been a long-cherished dream for me, as my long-term goal is to contribute to human civilization through prolific research and fruitful teaching. To make my dream a reality, I have already started doing research in my area of interest, *Bioinformatics and Computational Biology,* which combine topics from various disciplines, including life sciences and computer science. Novel contributions to these areas by continuing an advanced degree will be the first step towards meeting my long-term goal. Consequently, I consider the Department of Computer Science (CS) at the University of Texas at Austin as the right place for meeting the goal, as the department is continuously making a number of unique contributions in the aforementioned fields.

From my early childhood, I have always been fascinated and passionate about mathematics and I found computer science as a great opportunity to apply my mathematical skills and way of thinking. Therefore, I did my major in Computer Science and Engineering (CSE) from the Department of CSE in Bangladesh University of Engineering and Technology (BUET), which is one of the topmost departments in the most competitive university in Bangladesh. Even though I experienced a tremendous competitive environment during my undergraduate study, my passion for this discipline, determination to face challenges and strong desire to establish my career in academia, made my way to secure the 3rd position (with CGPA 3.97/4.00) in the department of CSE. As a recognition of my outstanding academic achievements, I have been appointed as a Lecturer in my department since my graduation. While serving my department, I am doing research through pursuing my Masters in CSE. I have also been involved with the Algorithm Engineering and Design Group (AℓEDA) since my senior year. My hard work and enthusiasm in Computational Biology helped me accomplish cutting edge research which led to notable publications in internationally reputable conferences and journals, even as an undergraduate student.

I started experiencing the fascinating world of research through my undergraduate thesis work. I always feel an urge to exploit my knowledge and expertise to do something for the betterment of human being. Therefore, I became interested to work in the field of bioinformatics when I came to know about its various life-saving applications such as uncovering new knowledge of the molecular mechanisms of disease enabling better treatments, cures and even preventative tests to be developed. Keeping that motivation in mind, I, along with one of my classmates, started exploring a popular topic in this field, "Protein Folding or Protein Structure Prediction", under Dr. Masud Hasan. We explored all the existing approximation algorithms for solving this problem both in 2D and 3D Hydrophobic-Polar (HP) lattice models. We noticed the fact that HP proteins are naturally represented as binary strings and the folding problem is combinatorially equivalent to folding a string of 0's and 1's so that the string forms a self-avoiding walk on the lattice and the number of adjacent pairs of 1's is maximized. To find a solution for this problem from stringological point of view, we started exploring various concepts of Stringology. At this time we were supervised by Dr. M. Sohel Rahman, one of the most prominent researchers in Stringology.

While working with Dr. Rahman we came to know about the research problem of reverse engineering of strings from given repetition information. Such reverse engineering is quite similar to the problem of predicting DNA/RNA or protein structures from repetitions, as the repetitions commonly found in DNA, RNA and protein sequences can be modeled with some famous data structures in Stringology (e.g., border array, cover array, suffix array etc.). We then started focusing on a specific problem. We addressed the problem of *inferring strings using least sized alphabet from cover array in linear time*. By exploring the properties of the border array and relating it to the cover array, we became successful to devise a new linear time algorithm that infers a string from a cover array using *binary alphabet*. Our work not only provides a linear time algorithm but also provides the least size bound on alphabet for this particular problem of string inference. This work has been well appreciated by the community who were working for a long time to provide the least size bound on alphabet. Our work has been published under the title "Linear Time Inference of Strings from Cover Arrays using a Binary Alphabet" in the *International Workshop on Algorithms and Computation* (WALCOM) held in 2012. Getting inspired by our first success, we extended our work to more generalized *indeterminate strings* which are often used to represent DNA/RNA sequences. In addition, we formulated three new algorithms for reverse engineering indeterminate strings from border arrays, suffix arrays and LCP arrays. These algorithms were not only the first of their kind but also had good expected running times. It is worth mentioning that, this work has been published under the title "Indeterminate String Inference Algorithms" in the StringMasters special issue of *Journal of Discrete Algorithms* (Elsevier).

Subsequently, in my Masters program, I have started working on a very interesting and highly investigated research topic – *Computational Phylogenetics*. My ongoing work covers a variant of maximum quartet consistency problem having an objective of devising an approach to infer highly accurate phylogenetic trees in highly scalable manners. Accurate phylogenetic reconstruction methods are inherently computationally extensive and therefore are limited to small number of taxa (species). Here, attaining high accuracy without losing scalability is my main research focus. My M.Sc. thesis proposal

# STATEMENT OF PURPOSE

titled "Constructing Phylogenetic Trees Using Quartet-Based Methods", having initial algorithms and experimental results, has been approved by the Board of Post Graduate Studies, Dept. of CSE. In addition to the thesis work, I have already done several research-based course projects in my masters, which are available at my online portfolio.

I am fascinated by the enormous potential for research in Bioinformatics and Computational Biology, which has been revealed to me during my undergraduate and graduate thesis. In my doctoral research, I would like to delve deeper into these areas. I have taken various graduate level courses including Meta-heuristics, Bioinformatics Algorithms, Computational Geometry, High Dimensional Data Management etc. for strengthening my base. During these courses, I worked on several projects. In *Meta-heuristics*, I became familiar with several guiding strategies in designing underlying heuristics to solve specific optimization problems for which computing optimal solutions is intractable. As a course project, I along with my four other classmates, have worked on *DNA fragment assembly using genetic algorithm*. We have devised a hybrid genetic algorithm, *Modified Scatter Search*, to solve the problem. We have evaluated our algorithm on noiseless datasets, and our algorithm outperformed the most accurate existing methods. As a part of the course *Computational Geometry*, I have done a survey on geometric analysis of the 3D structures of protein. Here, I have explored the works that use voronoi diagram based approach to analyze protein's 3D structure (e.g., volume, packing, voids of protein molecule) and derive some functionality. In *High Dimensional Data Management*, I have explored several approaches that extend the Resource Description Framework (RDF) data model for representing the uncertain information in the Semantic Web. In addition, I have done a short survey on "Alignment Free Methods for Distance-Based Phylogeny", as a part of the course *Bioinformatics Algorithms*. All these research experiences have boosted up my confidence in pursuing a Ph.D. in my field of interest.

The CS department at the University of Texas at Austin is continuously carrying out cutting edge research in Bioinformatics and Computational Biology. I would be thrilled to be a part of this great department, and to have the opportunity to get involved in fruitful research. In particular, I envision to work on different approaches of phylogenetic tree reconstruction. I would like to develop accurate and highly scalable supertree methods, with a view to eventually construct a tree of life. I am also interested in estimating species trees from gene trees considering various reasons of discordance among the gene trees. In addition, I would like to work on large-scale multiple sequence alignments. I am also open to work on other projects related to Bioinformatics and Computational Biology. I believe that, my background knowledge, experience to solve NP-hard optimization problems and programming skill will help me accomplish my doctoral research flawlessly in the fields of Bioinformatics and Computational Biology.

I started my study on phylogenetics with the tutorial "An overview of phylogeny reconstruction" by C. Randal Linder and Tandy Warnow. Getting interested by the field, I started to follow their works and the works of Prof. Tandy Warnow helped me to find a track in my current ongoing research activity. I would love to work with Prof. Tandy Warnow on phylogenetics as the research work of The *Warnow Lab* meets my interest most. I am willing to work with Prof. Warnow to develop novel methods on estimating large-scale sequence alignments and evolutionary trees. I am also interested to work with Prof. Daniel P. Miranker. The *Research in Bioinformatics and Semantic Web* (RiBS) Lab led by Prof. Miranker offers an environment full of opportunities to do breakthrough research in the field of bioinformatics. I am interested to work with Prof. Miranker on comparative RNA sequence analysis, and search and mining methods for large-scale biological databases. I have also been fascinated by the project *Ultrawrap* that uses relational database infrastructure for the data management aspects of the Semantic Web. I am also open and would be happy to work with others as well, who have an interest in the fields of Bioinformatics and Computational Biology, if such opportunity occurs.

I foresee that pursuing a Ph.D. in Computer Science is the best way to fulfill my goal of having a research career in academia. I am confident of meeting the high standards of University of Texas at Austin. I am aware of the kind of dedication, perseverance and resolve I need to have for pursuing the degree. I believe that my academic background has qualified me technically as well as has given me the right mind-set for pursuing the degree. I look forward to joining as a graduate student in your highly reputed department and having a long and mutually profitable association with University of Texas at Austin.

Please feel free to visit my online portfolio at https://sites.google.com/site/rezwanareazrimpi/ to view my research, publications and other scholastic achievements.