Chapter 90
Quantum Computing Approach for Alignment-Free Sequence Search and Classification

Rao M. Kotamarti
Southern Methodist University, USA

Mitchell A. Thornton
Southern Methodist University, USA

Margaret H. Dunham
Southern Methodist University, USA

ABSTRACT

Many classes of algorithms that suffer from large complexities when implemented on conventional computers may be reformulated resulting in greatly reduced complexity when implemented on quantum computers. The dramatic reductions in complexity for certain types of quantum algorithms coupled with the computationally challenging problems in some bioinformatics problems motivates researchers to devise efficient quantum algorithms for sequence (DNA, RNA, protein) analysis. This chapter shows that the important sequence classification problem in bioinformatics is suitable for formulation as a quantum algorithm. This chapter leverages earlier research for sequence classification based on Extensible Markov Model (EMM) and proposes a quantum computing alternative. The authors utilize sequence family profiles built using EMM methodology which is based on using pre-counted word data for each sequence. Then a new method termed quantum seeding is proposed for generating a key based on high frequency words. The key is applied in a quantum search based on Grover algorithm to determine a candidate set of models resulting in a significantly reduced search space. Given Z as a function of M models of size N, the quantum version of the seeding algorithm has a time complexity in the order of $O(\sqrt{Z})$ as opposed to O(Z) for the standard classic version for large values of Z.

DOI: 10.4018/978-1-4666-3604-0.ch090

INTRODUCTION

Reformulation of algorithms with large complexities from conventional computers could result in greatly reduced complexity when implemented on quantum computers (Nielsen & Chuang, 2000)(Marinescu & Marinescu, 2005)(Yanofsky & Mannucci, 2008). The resulting reduction in complexity is due to the underlying quantum computational model that is no longer constrained by the limitations of a Turing machine model of the present day computing. While commercially available quantum computers are not yet available, important algorithms have been formulated and run on experimental quantum computers. As new quantum devices and manufacturing technologies mature, the availability of commercial quantum computers continues to increase.

The dramatic reductions in complexity for quantum algorithms coupled with the computationally challenging problems in some bioinformatics problems motivates us to devise efficient quantum algorithms for sequence (DNA, RNA, protein) analysis. This class of problems results in large complexities with respect to Turing machine computational models due to the long lengths and potentially large number of sequences involved. Additionally, in contrast to the traditional and generic string matching problem, sequence matching problems tend to be fuzzy in nature. These classes of problems are particularly appropriate for reformulation into quantum computer algorithms.

One bioinformatics sequence application is that of classifying a sequence, such as a string of nucleotides, based on similarity to known classes of sequences. Basic Local Alignment Search known as BLAST (Karlin & Altschul, 1990) and BLAST PSI (Altschul, 1997) are in popular use for searching across genomic databases. BLAST PSI builds a characteristic profile from an initial set of search results. The results are used to fine tune the initial profile of the related sequences and the search is retried thus successively improving the relevance and diversity of related sequences. The process is repeated until the researcher is convinced with the resulting profile and the improved set of related sequences. Another approach - Profile Hidden Markov Model, also referred to as ProfileHMM (Eddy, 1998) uses a probabilistic profile for search across a growing database of profileHMMs representing characteristic domains (small stretches of significance) such as protein families (PFAM) (Finn, et al., 2008). Though BLAST is by far more used due to its ability to work with raw sequence formats of the genomic data, ProfileHMM is steadily gaining recognition among researchers for its probabilistic basis as more and more profileHMMs are built and added. To improve performance further, in order to handle the steadily increasing sizes of genomic databases, parallel processing versions have been proposed for BLAST and profileHMM. Specialized hardware solutions have also been proposed for the latter (Oliver, Yeow, & Schmidt, 2008). Much of the growth in genomic databases is due to the advent of the next generation sequencing technology. Much more is expected, thus resulting in a significant data overload in the future as reported by several studies (Eddy, 1998)(Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2006).

The promise of quantum computing allows for drastically reduced complexities for certain classes of algorithms. Some of the more well-known quantum algorithms are a) the searching problem of Grover that offers a quadratic reduction of temporal complexity (Grover, 1996) and b) the large integer prime factoring algorithm of Shor that reduces the factoring algorithm complexity to $O(N)$ where N is the number of digits comprising the integer. The integer factoring quantum algorithm garnered a significant amount of public attention since the large Turing model complexity of large integer factoring is responsible for the security of the widely used public key cryptography system known as RSA encryption (Rivest, Shamir, & Adleman, 1978). The formulation of a quantum algorithm is not mere-