Chapter 79

A New Approach for Sequence Analysis:
Illustrating an Expanded Bioinformatics View through Exploring Properties of the Prestin Protein

Kathryn Dempsey
University of Nebraska at Omaha, USA &
University of Nebraska Medical Center, USA

Benjamin Currall
Creighton University, USA

Richard Hallworth
Creighton University, USA

Hesham Ali
University of Nebraska at Omaha, USA &
University of Nebraska Medical Center, USA

ABSTRACT

Understanding the structure-function relationship of proteins offers the key to biological processes, and can offer knowledge for better investigation of matters with widespread impact, such as pathological disease and drug intervention. This relationship is dictated at the simplest level by the primary protein sequence. Since useful structures and functions are conserved within biology, a sequence with known structure-function relationship can be compared to related sequences to aid in novel structure-function prediction. Sequence analysis provides a means for suggesting evolutionary relationships, and inferring structural or functional similarity. It is crucial to consider these parameters while comparing sequences as they influence both the algorithms used and the implications of the results. For example, proteins that are closely related on an evolutionary time scale may have very similar structure, but entirely different functions. In contrast, proteins which have undergone convergent evolution may have dissimilar primary structure, but perform similar functions. This chapter details how the aspects of evolution, structure, and function can be taken into account when performing sequence analysis, and proposes an expansion on traditional approaches resulting in direct improvement of said analysis. This model is applied to a case study in the prestin protein and shows that the proposed approach provides a better understanding of input and output and can improve the performance of sequence analysis by means of motif detection software.

INTRODUCTION

Computational methods have simplified the analysis of the massive genetic code. Sequence analysis, as it is known, takes advantage of the inherent conservation of genomes by comparing set of nucleotide or amino acid sequences to infer relationships. One of the first described methods for sequence comparison was published in 1970 by Gibbs and McIntyre *et al.* in defining a means for comparing two biopolymer sequences with the allowance of gaps to discern homology using the dot method. In 1990, Altschul *et al.* described a method for the approximation of sequence alignments that would later become known as NCBI's BLAST. Since then, various methods for sequence comparison have been proposed reducing computational cost, runtimes, and improving accuracy. Currently, a variety of reputable methods exist for creating a multiple sequence alignment and one may choose the tool appropriate for their domain with the best speed, the least computational burden, or the best accuracy. This has only highlighted that sequence comparison, by means of alignment or pattern search, remains a consistent genre of tools for discerning basic characteristics about a set of sequences.

Even with the many advances in sequence comparison since 1970, one may beg the question of improvement, especially with the recent explosion of DNA and protein sequence data gathered from high-throughput methods. Sequence comparison methods have come under fire due to their over-simplifying the analysis process. The call for 'intelligent' sequence analysis has been proposed by Wagner *et al.*(2008) by describing alignment methods that incorporate domain expertise with sophisticated methods to achieve the best alignment possible. The idea of intelligent sequence analysis implies that the user can apply their expertise to appropriately prepare the input data and thus better understand and interpret the output, and even adjust software parameters to best suit their focus. This intelligent sequence analysis requires that users have knowledge in both biological and informatics fields. By breaking down sequence analysis, these ideas can be better applied.

In our analysis of nucleotide and amino acid sequences, we first divide methods into experimental analysis or computational analysis. Experimental analysis includes all experimental methods (e.g. sequencing and mass spectrometry) used to determine the sequence of nucleotides and amino acid residues pertaining to genomes as well as experimental methods that determine the evolution, structure, and/or function of those nucleotide and amino acid sequences. Computational analysis includes methods for preparing, accessing, and analyzing sequence data, with capabilities for examination of massive volumes of data using the parallel architectures of supercomputing.

Traditional Approach to Sequence Analysis

Computational sequence analysis can be divided into three stages: input, method, and output. The input of the traditional approach may contain one to several nucleotide or amino acid sequences and are chosen based upon user discretion. Several methods exist, but typically corresponding nucleotides or amino acid residues are evaluated to determine if they match (*i.e.* have identical or similar properties) or mismatch (*i.e.* gaps in sequence or dissimilar properties). Various outputs also exist, but all somehow provide a measure of comparison depending on scoring technique, providing an estimate based on the cellular level of conservation.

When computational analysis is coupled with experimental analysis, a canonical pathway for genomic analysis is formed. The standard flow of information is that some biological observation or query based on experimental analysis becomes a means for providing data with the intention of identifying common and unique sequence charac-

Related Content

3D Structural Bioinformatics of Proteins and Antibodies: State of the Art Perspectives and Challenges
Arianna Filntisi, Dimitrios Vlachakis, George Matsopoulosand Sophia Kossida (2013). *International Journal of Systems Biology and Biomedical Technologies (pp. 67-74).*
www.irma-international.org/article/3d-structural-bioinformatics-of-proteins-and-antibodies-state-of-the-art-perspectives-and-challenges/97743

Spam Detection on Social Media Using Semantic Convolutional Neural Network
Gauri Jain, Manisha Sharmaand Basant Agarwal (2018). *International Journal of Knowledge Discovery in Bioinformatics (pp. 12-26).*
www.irma-international.org/article/spam-detection-on-social-media-using-semantic-convolutional-neural-network/202361

Recognition of Translation Initiation Sites in Arabidopsis Thaliana
Haitham Ashoor, Arturo M. Mora, Karim Awara, Boris R. Jankovic, Rajesh Chowdhary, John A.C. Archerand Vladimir B. Bajic (2012). *Systemic Approaches in Bioinformatics and Computational Systems Biology: Recent Advances (pp. 105-116).*
www.irma-international.org/chapter/recognition-translation-initiation-sites-arabidopsis/60830

Chemosensitivity Prediction of Tumours based on Expression, miRNA, and Proteomics Data
I. Tsamardinos, G. Borboudakis, E. G. Christodoulouand O. D. Røe (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications (pp. 586-604).*
www.irma-international.org/chapter/chemosensitivity-prediction-tumours-based-expression/76085

Web Services for Bioinformatics
Abad Shah, Zafar Singheraand Syed Ahsan (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications (pp. 345-363).*
www.irma-international.org/chapter/web-services-bioinformatics/76072