Chapter 87
Improving PSI-BLAST's Fold Recognition Performance through Combining Consensus Sequences and Support Vector Machine

Ren-Xiang Yan
China Agricultural University, China

Jing Liu
China Agricultural University, China

Yi-Min Tao
China Agricultural University, China

ABSTRACT

Profile-profile alignment may be the most sensitive and useful computational resource for identifying remote homologies and recognizing protein folds. However, profile-profile alignment is usually much more complex and slower than sequence-sequence or profile-sequence alignment. The profile or PSSM (position-specific scoring matrix) can be used to represent the mutational variability at each sequence position of a protein by using a vector of amino acid substitution frequencies and it is a much richer encoding of a protein sequence. Consensus sequence, which can be considered as a simplified profile, was used to improve sequence alignment accuracy in the early time. Recently, several studies were carried out to improve PSI-BLAST's fold recognition performance by using consensus sequence information. There are several ways to compute a consensus sequence. Based on these considerations, we propose a method that combines the information of different types of consensus sequences with the assistance of support vector machine learning in this chapter. Benchmark results suggest that our method can further improve PSI-BLAST's fold recognition performance.

DOI: 10.4018/978-1-4666-3604-0.ch087

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Alignment between two protein sequence profiles is a fundamental technique in bioinformatics. Profile is a richer encoding of a protein versus raw sequence and can be used to improve remote homology detection and fold recognition performance (Kelley et al., 2000; Ginalski et al., 2003; Anand et al., 2005; Jaroszewski et al., 2005; Zhang et al., 2005). Sequence-sequence alignment can detect homology with significant sequence identity (i.e., identity >40%) (Needleman & Wunsch, 1970; Smith & Waterman, 1981; Altschul et al., 1997). Profile-sequence or sequence-profile alignment is usually more effective and sensitive than sequence-sequence alignment, and it can recognize distant homology with lower sequence identity (i.e., identity > 20%) (Altschul et al., 1997). Profile-profile alignment approaches may be the most effective approach among the alignment methods and even can create accurate alignments in the twilight zone (i.e., identity <10%) (Yona & Levitt, 2002). Generally speaking, alignment methods should accurately identify and align the homologous proteins, and the results can be used to predict a query protein's three-dimensional structure or to infer its biological function. Here, we can distinguish three classes of alignment approaches: sequence-sequence alignment, profile-sequence alignment and profile-profile alignment. Sequence-sequence alignment methods, such as BLAST (Altschul et al., 1997), Smith-Waterman (Smith & Waterman, 1981) or Needleman-Wunsch (Needleman & Wunsch, 1970) dynamic programming are relatively faster while less sensitive when compared with profile-profile alignment methods. Profile-profile alignment approaches or HMM-HMM comparison algorithms (Remmert et al., 2009) are more sensitive but relatively slower. PSI-BLAST (Altschul et al., 1997), a typical profile-sequence alignment method, is more sensitive than sequence-sequence alignment and much faster than profile-profile alignment. Currently, PSI-BLAST has become one of the most popular tools in bioinformatics and it is widely used in life science research. The improvement of PSI-BLAST's performance is still required. A profile can be used to represent the mutational variability at each sequence position of a protein by using a vector of amino acid substitution frequencies, which usually provides more information than a single sequence. The vector of a profile can be simplified to a consensus sequence by picking the most frequent or the most informative amino acid at each position of a protein. Several ways can be employed to generate a consensus sequence (Przybylski & Rost, 2008). And profile-consensus alignment (i.e., align a profile to a consensus sequence) can significantly improve PSI-BLAST's performance (Przybylski & Rost, 2008). Profile-consensus alignment can be considered as a method that mimicks profile–profile alignments (Przybylski & Rost, 2008). Based on this observation, we also compared consensus sequence based methods with COMPASS (Sadreyev & Grishin, 2003) in this paper. COMPASS is a local profile–profile alignment method, which can analytically estimate e-values for the detected protein similarities. And we also tried to improve PSI-BLAST's remote homology and fold recognition performance by combining search results that are generated by alignments between a profile and different types of consensus sequences with the assistance of support vector machine learning.

COMPUTATIONAL MODELS AND METHODS

Dataset

The protein sequences were extracted from the SCOP ASTRAL Compendium (Andreeva et al., 2004) database (1.73 version) filtered by 10% sequence identity and *an* e-value threshold of

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/chapter/improving-psi-blast-fold-recognition/76140?camid=4v1

This title is available in InfoSci-Books, InfoSci-Medical, Healthcare, and Life Sciences, Communications, Social Science, and Healthcare, InfoSci-Select, InfoSci-Select, InfoSci-Select. Recommend this product to your librarian:
www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

Text Mining on Big and Complex Biomedical Literature
Boya Xie, Qin Ding and Di Wu (2015). *Big Data Analytics in Bioinformatics and Healthcare (pp. 21-45).*
www.igi-global.com/chapter/text-mining-on-big-and-complex-biomedical-literature/121451?camid=4v1a

Advanced Datamining Using RNAseq Data
Yan Guo, Shilin Zhao, Margot Bjoring and Leng Han (2015). *Big Data Analytics in Bioinformatics and Healthcare (pp. 1-20).*
www.igi-global.com/chapter/advanced-datamining-using-rnaseq-data/121450?camid=4v1a

Mining Medical Data to Develop Clinical Decision Making Tools in Hemodialysis: Prediction of Cardiovascular Events and Feature Selection using a Random Forest Approach
Jasmine Ion Titapiccolo, Manuela Ferrario, Sergio Cerutti, Carlo Barbieri, Flavio Mari, Emanuele Gatti and Maria G. Signorini (2011). *International Journal of Knowledge Discovery in Bioinformatics (pp. 1-17).*
www.igi-global.com/article/mining-medical-data-develop-clinical/73908?camid=4v1a

Derivation of Quantum Associative Network from Hopfield-Like ANN and HNeT
Mitja Peruš and Chu Kiong Loo (2011). *Biological and Quantum Computing for Human Vision: Holonomic Models and Applications (pp. 199-216).*
www.igi-global.com/chapter/derivation-quantum-associative-network-hopfield/50507?camid=4v1a