THE CHINESE UNIVERSITY OF HONG KONG

SCHOOL OF LIFE SCIENCES

LIFE SCIENCES SEMINAR SERIES

2013 - 2014

Life Sciences Seminar is a seminar series aiming to provide up-to-date research ideas and experimental approaches to graduate students in the School

The Long and Short of Improving Distant Homology Detection

by

Professor Lee Ming Ming Marianne
School of Life Sciences
The Chinese University of Hong Kong

on

21 January 2014
(Tuesday)

at

12:30 – 1:15 pm

at

Room G18 (Ching Kai Hall)
Basic Medical Sciences Building
The Chinese University of Hong Kong

ALL ARE WELCOME

The Long and Short of Improving Distant Homology Detection

The deluge of biological information from different genomic initiatives and the rapid advancement in biotechnologies have made bioinformatics tools an integral part of modern biology. One of the most widely used techniques is sequence alignment, which BLAST and PSI-BLAST are probably the most popular tools in the field. PSI BLAST, which uses a profile (PSSM)-based and iterative search strategy, is more sensitive than BLAST in detecting weak homologies, thus making it suitable for searching remote homologs. Despite PSI-BLAST's much touted computational efficiency and high specificity, the identification of distant homologs having low sequence identity (<25%) and model corruption due to the incorporation of false positive sequence remain major challenges to the field of sequence alignment.

We have established two approaches to address these challenges. In one approach, we have developed a new machine learning algorithm, LESTAT (LEngth and STructure-based sequence Alignment Tool) that uses an iterative profile-based strategy together with several novel features to enhance the ability to identify more remote sequences. To overcome the inherent bias associated with a single starting model, LESTAT utilizes three structural homologs to create a profile consisting of structurally conserved positions and block separation distances. Subsequent profiles are refined iteratively using sequence information obtained from previous cycles. Additionally, the refinement process incorporates a "lock-in" feature to retain the high scoring sequences involved in previous alignments for subsequent model building and an enhancement factor to complement the weighting scheme used to build the position specific scoring matrix. A comparison of the performance of LESTAT against PSI-BLAST reveals that LESTAT exhibits increased sensitivity and specificity over PSI-BLAST in most systems. Notably, many of the hits identified are unique to each method, suggesting that LESTAT is a useful complementary method to PSI-BLAST in the detection of distant homologs.

In our second approach, we have developed a simple and elegant strategy to resolve the problem of model corruption. We hypothesize that combining results from the first (least-corrupted) profile with results from later (most sensitive) iterations of PSI-BLAST provides a better discriminator for true and false hits. Accordingly, we have derived a formula that utilizes the *E*-values from these two PSI-BLAST iterations to obtain a figure of merit for rank-ordering the hits. Our verification results based on a "gold-standard" test set of 103 queries indicate that this figure of merit does indeed delineate true positives from false positives better than PSI-BLAST *E*-values. Perhaps what is most notable about this strategy is that it is simple and straightforward to implement.