FEBS Workshop
THE BIOLOGY OF MODULAR PROTEIN DOMAINS
Seefeld in Tirol (AUSTRIA) September 8 –13, 2007

ABSTRACT FORM

A Motif Analysis Pipeline for Predicting Domain Binding Targets

Hugo Y.K. Lam[1*], Philip M. Kim[2*], Raffi Tonikian[3,4], Janine Mok[5], Ben Turk[6], Gary Bader[5], Charlie Boone[3,4], Michael Snyder[5] and Mark Gerstein[1,2]

[1] Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America; [2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America; [3] Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, M5S 1A8; [4] Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, M5G 1L6; [5] Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, United States of America; [6] Department of Pharmacology, Yale University, New Haven, Connecticut, United States of America; [7] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, M5S 3E1

*These authors contributed equally

A very important type of protein-protein interactions is mediated by modular protein domains, as exemplified by the SH3 domain and serine/threonine kinase domain which have been shown to mediate interactions important for signal transduction. Earlier studies have used consensus sequences from phage-display experiments to predict targets of these peptide binding domains. Also, many studies have shown various ways to improve prediction performance using genomic information. For instance, comparative genomics and secondary structure information have been used to increase prediction performance of SH3 targets. Furthermore, modern peptide library screening approaches have led to higher accuracy in determining the binding specificity of these domains.

However, thus far the prediction of biologically relevant targets of peptide binding domains has not been addressed in an integrated fashion. Therefore, we have developed a motif analysis pipeline which identifies target binding peptides by integrating comparative genomic, structural genomic and genomic data. The pipeline inputs are peptides identified from screen or pre-made PWMs. The pipeline makes use of an efficient search algorithm to scan the target proteome for potential motif hits and assign a PWM match score to each hit. It complements the motif score with a variety of pre-computed features that have been previously shown to determine biologically relevant targets, such as conservation, protein surface propensity, and disorder. Furthermore, it integrates genomic features such as interaction data, localization data, and others to further improve prediction performance. Finally, it uses a Bayesian learning algorithm that integrates all scores to give an optimal target prediction based upon a validated training data set. By taking into consideration different approaches, it aims to provide a comprehensive platform for researchers to predict biologically significant targets that are potentially recognized and bound by a particular domain of interest. The pipeline is very versatile and can predict binding targets for a wide variety of peptide binding domains.

(300 words, Times New Roman, 12 point)