Connecting Protein Structural Knowledge to Sequence Analysis and Design: Implications for Links between Sequences and Structures

by

Yao-ming Huang

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

DOCTOR OF PHILOSOPHY

Major Subject: Biology

Examining Committee:

Dr. Christopher Bystroff, Thesis Adviser

Dr. Blanca Barquera, Member

Dr. Chunyu Wang, Member

Dr. Mohammed Zaki, Member

Rensselaer Polytechnic Institute

Troy, New York

May, 2008

ABSTRACT

Proteins fold spontaneously and autonomously to acquire their functions. Blueprints of three-dimensional structures are inherent in primary sequences, and folding of proteins solely depends on inter- and intra-molecular interactions of polypeptides. The concept of interaction-driven protein folding extends abilities to predict protein structures and to design protein sequences by energetically knowing atomic interactions. Predictions seek to find the lowest energy structure given a fixed sequence, while designs search for the lowest energy sequence for a desired structure. Studies shown insight into the sequence-structure relationship are presented here.

The first study addresses the improvement of the evolutionary model in sequence pairwise alignments by protein structure predictions. Protein sequences in the "Twilight Zone" are often aligned inaccurately when global amino acid substitution matrices are used, even though there is structural homology. Since structural contexts contribute to the selective pressure of amino acid substitutions, a model, HMMSUM (HMMSTR-based SUbstitution Matrices), has newly developed to take local structures of proteins into consideration in substitution matrices. HMMSUM outperforms the alignment carried out using the BLOSUM or other structure-based substitution matrices when validated against remote homolog alignments from BAliBASE.

In following studies, several biological and computational strategies have been developed towards the goal of designing programmable peptide biosensors, which are biomolecules designed to report the presence of desired peptides. The leave-one-out design of green fluorescence protein (GFP) that has a β strand removed produces a reporting biosensor for the left-out β strand. Biochemical studies of the leave-one-out GFP show a great potential of having a real-time, self-sufficient peptide biosensor in respect of the affinity and reusability. However, to make biosensors programmable, mutations that change the specificity to selected peptides need to be identified. A computational approach is then proposed to facilitate the design of protein sequences that have desired specificities. The protein design *in silico* allows a rapid search of the enormous sequence space that usually cannot be screened efficiently through *in vitro* evolution methods.