Support Vector Learning and Rule Induction
for Knowledge Discovery in Biological Data

PhD candidate: Tho Hoan Pham Supervisor: Kenji Satou

Abstract

Proteins constitute most of a cell's dry mass. Each protein consists of a sequence of amino acids that allows it to fold up into a particular three-dimensional shape, or conformation. This three-dimensional structure allows the protein to interact with other biomolecules (such as DNAs, RNAs, and other proteins) to perform specific functions. Our research focuses on two fundamental issues of structure and interaction of proteins: (1) prediction and analysis of structure of proteins from their sequence, and (2) analysis of DNA-protein interactions.

Since predicting 3D structure of a protein is still difficult, many researchers have focused on predicting the secondary structure or recognizing structural motifs. The first part of this thesis will be devoted to the latter approach and will specially address *turn motifs*, which make up random coil areas in proteins. Turns make the protein fold into a specific three-dimensional shape. They play an important role in globular proteins from structural, interactional and functional points of view. We have developed a support vector machine (SVM)-based method to predict turn structures in a protein from its sequence of amino acids. When compared with previous methods, our approach exhibits a superior performance. Moreover, our method can estimate the relevance of amino acids for the formation of turn structures depending on their position in a protein. This information is specially useful for defining template structures when designing new molecules with certain desired characteristics.

Functionality of proteins is expressed through their interactions with other biomolecules. In the second part of this thesis, we address issues concerning interactions between proteins and DNAs, which represent vital information to uncover gene regulatory mechanisms. Both experimental and computational approaches have been proposed to establish mappings of DNA-binding locations of transcription factors, but while the former produces noisy results due to imperfect measuring methods, the latter often suffers from over-prediction problems. Also, interactions between transcription factors and DNA-binding sites are usually environment-dependent, with many regulators binding only under certain conditions. Even more, the presence of regulators at a promoter region indicates binding but not necessarily function: the regulator may act positively, negatively, or not act at all. Identifying true and functional DNA-protein interactions and discovering transcriptional regulatory patterns are therefore open and important problems in biology.

We have developed computational methods that combine DNA-protein interactions with expression profiles data to discover: (1) *relevant transcription factors* of a target gene from the set of potential ones; (2) *transcriptional regulatory rules* that qualitatively describe relationships between the expression behavior of a gene and its transcription factors; (3) *regulatory circuits* that describe how a group of transcription factors regulates target genes; (4) *transcriptional regulatory modules* that represent expression patterns of a group of genes commonly bound by the same set of transcription factors.