## Multi-Scale Investigation of Protein-Protein Interactions

Hou, Q.

2017

***document version***
Publisher's PDF, also known as Version of record

# Summary

Protein-protein interactions are essential for most biological activity. Investigating protein-protein interaction therefore is a leading task in structural bioinformatics. Nowadays, there are rapidly increasing amounts of protein structural data produced by high-throughput measurement techniques. However, protein sequence data even grows much faster as a result of Next-Generation Sequencing (NGS) techniques. Detecting the hidden messages in these data and uncovering the details of protein-protein interactions, and thereby the mechanisms of protein function, is brought increasingly within reach as a result of the aforementioned data explosion.

In this thesis, we implement and develop multiple tools to fetch signals from various kinds of protein information to effectively identify protein-protein interactions. The thesis is structured as follows:

Chapter 1 presents a general introduction to the research questions, providing necessary background information.

Identification of native binding poses is one of the most important tasks in structural bioinformatics. Measuring interacting free energy is the method of choice to estimate the interaction strength. In Chapter 2, we provide proof of concept that binding free energy from coarse-grained MD simulation can be used to detect the native binding poses out of thousands of docking solutions produced by docking programs. Compared with the benchmark dataset, our approach can enrich the percentage of near-native binding orientations for most of the target protein complexes. The coarse-grained force field allows us to interpret the interactions on a physical basis. Moreover, our method is precise for many targets and can select near-native binding poses in our Top 1 and Top 5 selection. This indicates that our method has a high potential to be used to scan prior to experiments, thereby reducing potentially heavy costs and might even be used for in silico prediction of interactomes

In Chapter 3, we show that Sequence Specificity from interacting and non-interacting homologs carries some signal to predict interaction sites from sequences. Many protein families involved in protein-protein interactions have several sub-families that bind to different partners or lose interaction. Specificity in these interactions is often decisive to the functions of those proteins. In this chapter, we show that the specificity between interacting versus non-interacting groups at the sequence level can be used for recognizing interaction sites. The results indicate that it is possible to predict interaction sites using nothing more than the sequence and group specificity information. To the best of our knowledge, this is the first time that sequence speci-

ficity between interacting and non-interacting homologs was used to predict interacting sites at a large scale.

In Chapter 4, we develop a Random Forest predictor to identify interaction sites from sequence information using evolutionary information, predicted structural properties and the length of proteins as features. Predicting protein-protein interacting sites from sequence information is becoming more and more important with the increasing amount of genome data available. Conservation from multiple sequence alignment, predicted surface accessibility and predicted secondary structure were used as features by machine learning methods to identify protein interfaces. Other than the 'old' features, we show that backbone flexibility derived from sequence and the length of sequence can be used to pinpoint interacting positions. Moreover, after combining all features, we have developed a sequence-based interface predictor, which achieved an excellent performance on both homodimeric and heterodimeric proteins. The success of the prediction indicates that our method captured captures the common properties of both homodimeric and heterodimeric interfaces.

In Chapter 5, we build upon the approach described in Chapter 4 and create a webserver named SeRenDIP: SEquence-based Random forest predictor with lENgth and Dynamics for Interacting Proteins. This is a user-friendly web interface allowing flexible prediction of protein-protein binding interfaces.

Finally, in Chapter 6 we draw some conclusions and discuss possible future directions.

Summarising, in this thesis we investigate protein-protein interaction through biophysical and bioinformatics approaches and validate the progress achieved using different types of protein data. The resulting methods can be applied widely and help draw a full landscape of protein-protein interaction.