Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning

Wenjun Zheng[1]

[1]University at Buffalo Department of Physics

August 31, 2023

Abstract

To enable personalized genetics and medicine, it is important yet highly challenging to accurately predict disease-causing mutations from the sequences alone at high throughput. To meet this challenge, we build upon recent progress in machine learning, network analysis, and protein language models, and develop a sequences-based variant site prediction workflow based on the protein residue contact networks: 1. We employ and integrate various methods of building protein residue networks using state-of-the-art coevolution analysis tools (e.g., RaptorX, DeepMetaPSICOV, and SPOT-Contact) powered by deep learning. 2. We use machine learning algorithms (e.g., Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 13 network centrality scores (calculated by NetworkX) with 7 other network scores calculated from the contact probability matrices to jointly predict key residues as hot spots for disease mutations. 3. Using a dataset of 107 proteins rich in disease mutations, we rigorously evaluate the network scores individually and collectively in comparison with alternative structures-based network scores (using predicted structures by AlphaFold). By optimally combing three coevolution analysis methods and the resulting network scores by machine learning, we are able to discriminate deleterious and neutral mutation sites accurately (AUC of ROC ~ 0.84). Furthermore, by combining our method with a state-of-the-art predictor of the functional effects of sequence variations based on large protein language models, we have significantly improved the prediction of disease variant sites (AUC ~ 0.89). This work supports a promising strategy of combining an ensemble of network scores based on different coevolution analysis methods via machine learning to predict candidate sites of disease mutations, which will inform downstream applications of disease diagnosis and targeted drug design.

Hosted file