Supervised learning for analyzing large-scale genome-wide DNA polymorphism data

Haipeng Li

Abstract:
Supervised learning has been extensively applied in many fields; Alpha-GO and autopilot might be two of the most well-known cases. However, its application in population and evolutionary genetics is still in childhood. Recently, we introduced the boosting, a supervised learning approach, to identify positive Darwinian selection in natural populations and estimate recombination rate along the human genome. We further analyzed the genome-wide DNA polymorphism data from nearly 10,000 human individuals (UK10K) and obtained a fine-scale genetic map for humans. The number of identified autosomal recombination hotspots is about 2.93 – 14.25 times less than that previously identified in human populations, indicating that the variance of estimated recombination rate may be underestimated when identifying recombination hotspots, especially population-specific human recombination hotspots. These results indicate that supervised learning approaches, together with deep learning and reinforced learning, could play essential roles when analyzing large-scale genome-wide DNA polymorphism data.