Samenvatting

Dit proefschrift beschrijft onderzoek op het gebied van datamining in geannoteerde grafen. Wij kozen Proteïne-Proteïne-Interactie (PPI) netwerken als domein om onze methodes op toe te passen. Het PPI-netwerk werd gemodelleerd als een graaf waarbij elke knoop een proteïne is en elke tak staat voor een fysieke interactie tussen twee proteïnes. Er zijn verschillende soorten informatie waarmee elk proteïne in het PPI-netwerk geannoteerd is. De "functionele annotatie" geeft de biologische functies van de proteïnes in het PPI-netwerk en de "ziekte/kanker gerelateerde annotatie" geeft aan of een proteïne bij een ziekte of kanker betrokken is. We probeerden deze annotaties te voorspellen met als doel de informatie over proteïnes in het PPI-netwerk te verbeteren.

De taak van het voorspellen van functies in het PPI-netwerk slaat op het voorspellen van functies van niet-geannoteerde proteïnes met behulp van de informatie in het netwerk. Hiervoor worden twee manieren voorgesteld. Bij de eerste manier gebruikten we kortste-pad afstanden tussen verschillende proteïnes als beschrijvende eigenschappen van de proteïnes, en variantie-analyse als selectiemethode om de ruis en de dimensionaliteit van de vectoren te verminderen. Daarna pasten we hier machinaal leren op toe om de eigenschappen voor de niet-geannoteerde proteïnes te voorspellen. Bij de tweede manier introduceerden we nieuwe functionele eigenschappen die de zogenaamde “Samenwerkende Functies” aangeven. Deze samenwerkende functies zijn paren functies die vaak samen voorkomen bij proteïnes waartussen een fysieke interactie bestaat. De meeste van de al bestaande methodes voorspellen de functies van de proteïnes door middel van *guilt-by-association*, waarbij er vanuit wordt gegaan dat proteïnes waartussen een interactie bestaat geneigd zijn om dezelfde functies te hebben. Wij onderzochten twee methodes om de samenwerkende functies uit het PPI-netwerk te halen. De eerste methode berekent de samenwerkingswaarde van twee functies door middel van een iteratieve versterkingsstrategie. De tweede methode gebruikt een kunstmatig neuraal netwerk. Empirisch onderzoek bevestigt dat het concept van samenwerkende functies beter werkt voor het voorspellen van functies van proteïnes, dan het concept dat proteïnes waartussen een interactie bestaat vaak dezelfde functies hebben.

Bij het voorspellen van aan kanker gerelateerde proteïnes in het PPI-netwerk wordt

gezocht naar nieuwe proteïnes die waarschijnlijk betrokken zijn bij het veroorzaken van kanker. We beschouwen reeds bestaande methodes als een tweestaps-algoritme. Bij de eerste stap worden aan de hand van de trainingsdata enkele eigenschappen geselecteerd die de proteïnes in de test data moeten beschrijven. Daarna wordt machinaal leren toegepast op deze eigenschappen om zo te voorspellen welke proteïnes aan kanker gerelateerd zijn. Empirisch onderzoek heeft aangetoond dat de kwaliteit van de voorspelling meer wordt bepaald door de beschrijvende eigenschappen waaruit geleerd wordt, dan door de gebruikte methode van machinaal leren. Als deze verschillende eigenschappen onafhankelijk van elkaar worden bekeken, lijken de biologische functies het beste te werken. Wij bedachten twee manieren om nieuwe eigenschappen te selecteren uit het PPI-netwerk. Bij de eerste manier wordt de functionele en structurele context van proteïnes bekeken met behulp van variantie-analyse en de χ-kwadraat methode. De tweede manier bestaat uit een geheel nieuwe methode, "Interactiegebaseerde chi-kwadraat", die de functionele annotaties van proteïnes combineert met de informatie die genesteld zit in de topologie van een PPI-netwerk, om zo de juiste eigenschappen te selecteren. Empirisch onderzoek heeft aangetoond dat onze manieren om eigenschappen te selecteren een duidelijke biologische betekenis hebben en ervoor zorgen dat het systeem betere voorspellingen doet.

Het voorspellen van aan ziekte gerelateerde proteïnes in het PPI-netwerk is een belangrijk onderzoeksgebied binnen de computationele biologie. Eerdere methodes gaan uit van een verzameling proteïnes waarvan al bekend is dat zij gerelateerd zijn aan deze ziekte (zogenaamde bronproteïnes), en vervolgens wordt geprobeerd deze verzameling uit te breiden door van andere proteïnes te voorspellen of deze ook aan die ziekte gerelateerd zijn. De initiële verzamelingen van bronproteïnes voor een ziekte zijn onvolledig: er zijn zo goed als zeker valse negatieven. Dit heeft een nadelige invloed op de resultaten bekomen met de vermelde methoden. Wij creëerden een nieuw "Human Disease Network" (HDN), een netwerk dat verbanden legt tussen ziekten, met behulp van zowel functionele als structurele informatie van het PPI-netwerk, dit om het aantal valse negatieven in de initiële verzameling bronproteïnes te verminderen voor 20 verschillende ziektes. Met behulp van literatuuronderzoek van nieuw voorspelde proteïnes, is bewezen dat het door ons gecreëerde HDN zeer bruikbaar is.