PROBABILISTISCHE METHODEN OM TE ZOEKEN NAAR REGULATORISCHE ELEMENTEN IN SETS VAN CO-GEREGULEERDE GENEN

In de laatste jaren heeft het onstaan van nieuwe technologiën, zoals microroosters en volledig geautomatiseerde sequentiebepaling, de hoeveelheid beschikbare biologische gegevens sterk doen toe nemen. Deze overvloed aan beschikbare gegevens doet de vraag stijgen naar nieuwe algoritmen die relevante informatie kunnen vinden in deze gegevens. Hier doet de bioinformatica zijn intrede. Op dit kruispunt tussen moleculaire biologie en informatica kunnen we ons onderzoek situeren. In dit proefschrift stellen we een set van algoritmen voor om te zoeken naar potentiële bindingsplaatsen voor transcriptiefactoren vertrekkend van een probabilistisch sequentiemodel en sets van co-reguleerde genen. Het probleem wordt benaderd vanuit twee invalshoeken: gesuperviseerd of niet-gesuperviseerd.

Om het niet-gesuperviseerd probleem aan te pakken, hebben we het originele *Gibbs sampling* algoritme om motieven te zoeken aangepast. Vooreerst hebben we een hogere-orde achtergrondmodel geïntroduceerd om beter het onderscheid te kunnen maken tussen echte motieven en achtergrondruis. Ten tweede hebben we het probabilistische raamwerk gebruikt om het aantal instanties van een motief te schatten in een sequentie. Deze uitbreidingen hebben geleid tot de implementatie van **MotifSampler**. Dankzij een doorgedreven studie van de invloed van de parameters op de performantie hebben we het nodige inzicht verworven in het gedrag van ons algoritme. Deze analyse heeft ons dan toegelaten om een uitgekiende strategie voor te stellen om motieven te zoeken in biologische voorbeelden. In deze thesis gebruiken we vier grote voorbeelden voor een gedetailleerde studie: 33 sequenties met de G-box transcriptiefactor in planten, 10 regulons in gist, de σ^{54} factor in prokaryoten en vier clusters van co-gereguleerde genen uit de cellcyclus in gist. De grote verscheidenheid van deze voorbeelden illustreert duidelijk de mogelijkheden en beperkingen van ons algoritme. Het belangrijkste resultaat is dat een goed ontworpen organisme-specifiek achtergrondmodel de performantie singificant verbetert vooral wanneer een grote hoeveelheid ruis aanwezig is in de dataset.

Vertrekkend van het probabilistisch sequentiemodel hebben we ook een gesuperviseerde methode, **MotifScanner**, geïmplementeerd om instanties van gekende motieven te detecteren. Een gedetailleerde analyse van de invloed van de parameters op de performantie toont aan dat onze methode robuuster is dan een klassiek schema om met een gewichtsmatrix te scoren. Als een set van bekende matrices voorhanden is, kunnen we een set van co-gereguleerde genen screenen en de statistische significantie berekenen van het aantal gevonden instanties. Als referentie nemen we het verwachte aantal instanties gevonden in alle promoters in het genoom. Voorbeelden in gist tonen aan dat deze methode toepasbaar is maar dat de grootste beperking de kwaliteit van de matrices is.

Tenslotte bespreken we de implementatie van **INCLUSive**, een geïntegreerd web-gebaseerd platform voor de analyse van microroostergegevens. De analyseprocedure start met Adaptive Quality-Based Clustering, wat resulteert in een aantal clusters met een gelijkaardig expressieprofiel. In de volgende stap willen we de promotersequenties van deze genen bekijken. Daartoe hebben we een systeem ontworpen dat de promoters probeert te lokaliseren in de genoomsequentie. De geselecteerde sequenties kunnen dan verwerkt worden door MotifSampler en/of MotifScanner. Om de toepasbaarheid te illustreren, kunnen we verwijzen naar het werk van anderen die onze algoritmen gebruikt hebben om specifieke motieven te detecteren binnen hun projecten.

Promotor: Prof. dr. ir. Bart De Moor

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

UDC: 681.3*F2:575.113