Computers and drug discovery : construction and data mining of chemical and biological databases

Kazius, J.

Samenvatting

Bioinformatica en cheminformatica zijn onderzoeksvelden die grotendeels gescheiden zijn, ondanks de overlap hiertussen. Zoals omschreven in Hoofdstuk 1 zijn in beide disciplines de constructie van biologische en chemische databases van belang. Opgebouwde databases kunnen door middel van *data mining* methoden geanalyseerd worden om hun inhoud compact samen te vatten en om nieuwe wetenschappelijke inzichten te ontdekken. In beide disciplines kan het data minen van dergelijke databases op een objectieve manier nieuwe causale relaties tussen chemische en/of biologische eigenschappen onthullen, zoals chemische en biologische oorzaken voor nadelige biologisch effecten in de mens. Onthulde, betrouwbare kennis kan dan gebruikt worden om de biologische effecten in te schatten van nieuwe verbindingen, genetische varianten, eiwitten en andere elementen zodat tijd en geld efficiënter geïnvesteerd kunnen worden. Hoofdstuk 1 beschrijft ook de algemene doelstelling en indeling van dit proefschrift.

Verschillende genetische varianten die in de natuur voorkomen, zowel zeldzame mutaties als veel-voorkomende polymorfismes, produceren zeer nadelige fenotypes. Een noemenswaardige verscheidenheid aan schadelijke en onschadelijke genetische varianten komt voor in G-eiwitgekoppelde receptoren in de mens (GPCRs, van G Protein-Coupled Receptors). De superfamilie van GPCRs is groot en reguleert een variëteit aan belangrijke fysiologische processen doordat deze receptoreiwitten extracellulaire signalen over celmembranen transporteren. GPCRs dienen bovendien als prominente doelwitten voor geneeesmiddelen. Om deze redenen dragen natuurlijke varianten in het menselijke DNA waarschijnlijk bij aan inter-individuele diversiteit. Hoofdstuk 2 bediscussïeert daarom de mogelijkheden en onmogelijkheden van GPCR varianten als oorzaken voor ziekte. Een observatie was dat, in vergelijking met zeldzame GPCR mutaties, weinig GPCR polymorfismes overtuigend geassociëerd zijn met ziekte. “Aantal kopieën”-polymorfismes (CNPs, van Copy Number Polymorphisms) zijn grote DNA segmenten die varieren in hoe

vaak ze herhaald worden in het menselijke genoom. Omdat CNPs complete GPCR genen kunnen bevatten, kunnen CNPs sterkere factoren zijn voor klinische fenotypes dan eenvoudige nucleotidepolymorfismes (SNPs, van Single Nucleotide Polymorphisms).

Hoofdstuk 3 omschrijft het ontwerp en de constructie van een database van GPCR varianten. Deze database kan een waardevolle bron van informatie zijn voor een verklaring dat sommige varianten ziektes veroorzaken terwijl andere onschadelijk blijken. Omdat mutatie-gerelateerde informatie sterk versnipperd is over verschillende bronnen zouden onderzoekers van GPCRs geholpen zijn met een eenduidige bron die alle informatie over GPCR mutaties integreert. Daarom hebben wij de GPCR NaVa database opgebouwd (http://nava.liacs.nl). Waar beschikbaar, bevatten varianten in deze database informatie over hun locatie in het DNA (en in de eiwitsequentie), de betrokken nucleotides (en aminozuren), de gemiddelde frequentie van ieder allel in de populatie, gepubliceerde ziekte-associaties en verwijzingen naar publieke databases en de wetenschappelijke literatuur.

Als conclusie op de Hoofdstukken 2 en 3, probeert Hoofdstuk 4 verschillende berekenbare eigenschappen van de varianten te relateren aan nadelige klinische effecten. Zoals verwacht, komen GPCRs die betrokken zijn bij reuk vaker voor in CNPs dan andere, belangrijkere GPCRs. Deze laatste groep GPCRs bevat ook minder 'stille' polymorfismes (welke geen aminozuurmutatie veroorzaken) en meer polymorfismes die wel aminozuurmutaties veroorzaken. Onze gegevens suggereren dat varianten die grovere aminozuurmutaties veroorzaken sterkere risicofactoren zijn voor ziekte. Bovendien zijn aminozuurmuterende varianten sterkere risicofactoren voor ziekte als ze in transmembraandomeinen van GPCRs voorkomen, vooral als deze aminozuren in de kern van de receptor zitten volgens de drie-dimensionale structuur van deze GPCR transmembraandomeinen. Sterk geconserveerde transmembraanposities bleken ook de meeste ziekte-mutaties te bevatten. Uit de analyse van de toegankelijke data van de GPCR NaVa database bleek dat het risico op ziekte van varianten afhing van verschillende structurele eigenschappen. Tegelijkertijd bevestigden deze bevindingen het relatieve belang van enkele domeinen en vitale aminozuurposities voor de normale werking van GPCRs.

Eigenschappen van genetische varianten die ziektes veroorzaken kunnen inzichten geven in de werking van receptoren, maar deze inzichten kunnen niet gebruikt worden ter verbetering van de klinische situate van de patiënten die deze mutaties hebben. Echter, chemische eigenschappen die correleren met biologische activiteiten kunnen wel meegenomen worden in het moleculaire ontwerp van potentiële geneesmiddelen; kennis hiervan kan zo een positieve impact hebben op patiënten. In dat opzicht heeft het data minen van chemische descriptoren in plaats van biologische descriptoren toegevoegde waarde. Hoofdstuk 5 behandelt basistechnieken die veel gebruikt worden in data mining, evenals de voordelen en beperkingen van deze technieken. Voorbeelden uit de cheminformatica zijn gebruikt om de moleculaire structuur te koppelen aan biologische activiteit, met name

toxiciteit, door het afleiden van structuur-activiteitsrelaties. De methoden van data mining die besproken worden in Hoofdstuk 5 zijn bovendien onafhankelijk van de geselecteerde chemische of biologische descriptoren en kunnen daarom ook nuttig zijn voor bioinformatica onderzoek. Tijdens data mining wordt de beschikbare dataset vaak gesplitst in een training set, om (cor)relaties af te leiden in de vorm van een model, en een test set, om te valideren of deze relaties voorspellende waarde hebben. Als relaties gevonden worden is het belangrijk om te weten of ze het bestudeerde biologische effect kunnen voorspellen. De statistische resultaten die de betrouwbaarheid van de geëxtraheerde relaties moeten weerspiegelen kunnen drastisch beïnvloed worden door de methode waarmee de dataset wordt gesplitst en door de verdeling van de beschikbare data. Richtlijnen voor robuuste validatie bestaan uit een volledige scheiding van testdata en trainingdata, herhaalde kruisvalidaties met grote willekeurige testsets en de definitie van een toepasbaarheidsdomein, ofwel 'applicability domain'.

Hoofdstuk 6 behandelt de substructuren die voortvloeiden uit een semi-computationele studie om structuur-activiteitsrelaties af te leiden voor mutageniciteit, een type toxiciteit. Verbindingen zijn vaak minder aantrekkelijk als geneesmiddelen wanneer ze mutageen zijn, dat wil zeggen dat ze mutaties in DNA kunnen introduceren die schadelijk kunnen zijn. Het is daarom van belang om de betrouwbaarheid en de precisie van mutageniciteitsvoospellingen te verhogen door de identificatie van nieuwe toxicoforen, ofwel toxiciteit-gerelateerde chemische substructuren. Een grote mutageniciteitsdataset werd opgebouwd voor dit doeleinde en vervolgens geanalyseerd om toxicoforen te vinden en te valideren. Een initiële substructuurzoekactie van deze dataset liet zien dat de meeste mutagene stoffen gedetecteerd werden door slechts acht algemene toxicoforen toe te passen. Van deze acht werden specifiekere toxicoforen afgeleid en geaccepteerd op basis van statistische criteria in combinatie met chemische en mechanistische kennis verkregen uit de wetenschappelijke literatuur. De resulterende negenentwintig toxicoforen voor mutageniciteit zijn nuttig voor het afschatten van de risico's van potentiële geneesmiddelen en het ontwerpen van collecties hiervan. Voorbeelden van een nieuwe toxicofoor en toxiciteit-verminderende groep zijn respectievelijk een polycyclisch, vlak ringsysteem en de trifluoromethyl substituent.

Substructuurmining-algoritmes zijn belangrijke technieken in de ontdekking van geneesmiddelen omdat ze biologisch relevante substructuren kunnen afleiden op een snelle en automatische manier. Methoden tot dusver beschouwden echter slechts een deel van alle chemische informatie die aanwezig is in een dataset van verbindingen. Daarom is het algemene doel van Hoofdstuk 7 om meer complete data mining mogelijk te maken binnen de cheminformatica door het ontwerpen van een methode die substructuren beschouwt van elke grootte, vorm en mate van chemisch detail. Een methode van chemische representatie werd ontwikkeld die zowel algemene en zeer specifieke eigenschappen bevat. Ten bewijze van het principe, leerde het efficiënte, veelzijdige *graafmining*-systeem Gaston alle substructuren van

elke grootte en vorm van de mutageniciteitsdataset die op deze uitgebreide wijze was gerepresenteerd. Van deze substructuren leidden we automatisch een set af van slechts zes onafhankelijke, onderscheidende toxicoforen die relevante biochemische kennis beschreven, zoals in Hoofdstuk 6 is uitgelegd. Deze resultaten demonstreerden het individuele en synergistische belang van zowel de uitgebreide chemische representatie als het minen van nonlineare substructuren. De combinatie van de uitgebreide chemische representatie en Gaston geeft een uitmuntende methode voor twee-dimensionale substructuurmining omdat dit recept systematische alle substructuren verkent in verschillende mate van chemisch detail.

Algemene conclusies over het onderzoek dat in dit proefschrift is beschreven worden getrokken in Hoofdstuk 8. In het kort beschrijft dit proefschrift verschillende oorzaken van nadelige biologische effects. In het geval van GPCR varianten, onthullen bioinformatica analyses het belang van hun positie in de receptorstructuur: varianten in transmembraan-domeinen veroorzaken vaker ziekte, vooral als ze in de kern van de receptor voorkomen. In het geval van chemische verbindingen werden toxicoforen geïdentificeerd als oorzaken door middel van een veelbelovende cheminformatica methode. Naast dergelijke conclusies worden ook vooruitzichten gegeven op interessante onderzoekslijnen.