

Scriptieonderwerpen



2003-2004

Prof.Dr.ir Wim Van Criekinge

<http://biobix.rug.ac.be/>

FAQ: Frequently asked questions

Waarom dit document ?

Verschillende mensen vonden de “one-liners” te kryptisch om een goed beeld te krijgen wat er juist beschikbaar was als thesisonderwerp in de bioinformatica. Vandaar een meer extensief overzicht van de verschillende projecten. De verschillende projecten zullen in verkorte vorm online gezet worden op <http://fltbwww.rug.ac.be/kco/scriptieonderwerpen/student.php> en dit document is als pdf terug te vinden op <http://biobix.rug.ac.be/>

Labo voor “Bioinformatics and Computational Genomics” ?

Bestaat sinds oktober 2002. Voornaamste activiteit tot nu was het doceren van het vak Bioinformatica een verplicht vak voor 2de proef cel- en gentechnologie. De cursus wordt eveneens deels gegeven aan de faculteit wetenschappen te Gent en toegepaste wetenschappen te Leuven. De in oprichting zijnde onderzoeksgroep, bioinformatica en computationele genomica, heeft een sterke focus op "junk mining" en "systems biology/comparative genomics". Onder "junk mining" verstaan we exploratie van het genoom in het geheel en de niet-eiwit-coderende stukken in het bijzonder. Hiervoor werd reeds 18 maanden samengewerkt met de KUL voor de identificatie van nieuwe ncRNA. Als co-promoter van Stefan Van Yper, IWT doctoraatstudent die werkt aan phylogenetic footprinting, werden reeds verassende resultaten behaald. Onder het luik "systems biology & comparative genomics" werden er verschillende SBO en FWO voorstellen ingediend en wordt er een samenwerking opgezet met het ITG (Instituut voor Tropische Geneeskunde) voor de ontwikkeling van innovatieve behandelingsmethoden van malaria. Hiervoor zal er ook een europees FP6 voorstel worden ingediend. Belangrijk in alle lopende of komende projecten is dat de resultaten gevalideerd worden door samenwerkingen met verschillende labo's in zowel binnen- als buitenland.

Voorkennis ?

Een steeds weerkerende vraag heeft betrekking op de nodige voorkennis. Tool of programmeertaal-specifieke kennis is zeker GEEN vereiste. Tools en programmeromgevingen zijn enkel een middel om ergens te geraken. De belangrijkste vereiste is motivatie. Voor alle voorgestelde projecten is de nodige ervaring aanwezig en zal de nodige tijd worden uitgetrokken om U de nodige tools eigen te maken. Sommige onderwerpen kunnen perfect met verschillende studenten worden uitgewerkt. Andere kunnen dienen als doctoraatsonderwerp.

Interesse of extra informatie nodig ?

Aarzel niet om contact op te nemen, bv voor een oriënterend gesprek om te kijken hoe we interesses, voorkennis en beschikbare onderwerpen kunnen verzoenen.

Prof.Dr.ir. Wim Van Criekinge
Bioinformatics and Computational Genomics
Department of Molecular Biotechnology
Faculty of Agricultural and Applied Biological Sciences
Coupure Links 653, 9000 Gent, Belgium
Tel +32 9 264 59 69
Wim.vancriekinge@rug.ac.be

Overzicht van de verschillende projecten:

Functioneel peptidomonderzoek 4
 Innovatieve behandelingsmethoden voor Malaria 4
 Een nieuw concept: Genoomtopologieën..... 6
 Detectie, karakterisatie en validatie van ncRNA's (non-coding RNAs)..... 7
 Functionizing: onderzoek naar nieuwe genfuncties via integratie van bestaande
 genoomdatasets (bijvoorbeeld chipdata en yeast-two-hybrid data) 10
 Geautomatiseerde representatie van geconserveerde aminozuurresiduen op 3D
 structuurmodellen..... 13
 Mining the DNA repair genes 16
 BioCardridge - BioSQL 18
 ChemCardridge en toepassing in “structural genomics”..... 19
 Functionele analyse van microarray genexpressie data 20
 Machine Learning voor Tumor-Classificatie 22
 Multiplexed Blast 23
 Biolinux 23
 Dynamisch Programmeren met Python & GATO 24
 Toepassen van “Singular Value Decomposition” op biologische problemen..... 24
 Nutrigenomics & Functional Foods 25
 Analyse van biologische netwerk topologieën..... 25
 NMD (Nonsense Mediated Decay) in Viruses..... 26
 COP, cluster of orthologous Promoters..... 26
 Virtual Northern in SVG 27
 Bioinformatics Web Services..... 27
 Bioinformatics Portal 28

Functioneel peptidomonderzoek

In het postgenoomtijdperk blijft de belangrijkste vraag: wat is de functie van al de (onbekende) genen? Verschillende parallele strategieën worden hierbij gevolgd, waaronder het verstoren van de werking van genen (bv. door systematische knock-out of knock-down experimenten), het in kaart brengen van gen- of eiwit-expressie, het opstellen van genetische of eiwit-interactiekaarten, enzovoort.

Peptiden glippen hierbij vaak tussen de mazen van het net. Ze zijn te klein om opgemerkt te worden bij de meeste proteoomstudies (die 2-dimensionale gel electroforese gebruiken). Ze zijn ook moeilijk te voorspellen op basis van gensequenties omdat ze verschillende bewerkingsstappen (klievingen, post-translationele modificaties) moeten ondergaan om bio-actief te worden.

Recente ontwikkelingen in chromatografie en massaspectrometrie laten toe peptiden (met hun post-translationele modificaties) rechtstreeks te identificeren in complexe mengsels. Daarom willen we alle peptiden in extracten van een aantal muisweefsels identificeren, en -waar onbekend- hun functie zoveel mogelijk ophelderen.

Het gen voor het precursoreiwit zal worden opgespoord, alsmede een zo compleet mogelijke staalkaart van homologen en orthologen. Prototypische eigenschappen van reeds gekende peptiden zoals genstructuur, aanwezigheid van eiwit domeinen en maturatieplaatsen, zijn cruciaal en kunnen geëtrapoleerd worden op de rest van het genoom. Op basis van het voorkomen van ESTs in cDNA banken (muis, rat en mens) zal getracht worden een eerste inzicht te krijgen in de cellulaire en weefseldistributie ("virtuele Northern blot", zie verder).

Om de verdere functionele informatie in kaart te brengen zal gebruik gemaakt worden van de zogenaamde "Rosetta stone" techniek. Hierbij worden in detail alle mogelijke facetten van de genen coderend voor peptiden, en hun respectievelijke orthologen, in kaart gebracht. Dit omvat zowel eigenschappen op genoom topologisch vlak, genetische data, als bijvoorbeeld voorhanden zijnde eiwit interactiegegevens, DNA chip gegevens, literatuurverwijzingen, enzovoort.

Gezien de specifieke maturatiemechanismen zal zorg besteed worden aan de identificatie van de verschillende peptiden en hun overeenkomstige proteïnen. Hiervoor zal onder meer gebruik gemaakt worden van bestaande protease gegevensbanken (Merops) en de bestaande diversiteit aan peptide genen zal gebruikt worden om een "classifier" te trainen die dan nadien kan gebruikt worden als predictor op de nieuw geïdentificeerde genen.

Innovatieve behandelingsmethoden voor Malaria

Malaria, wat Mal-Aria of slechte lucht betekent, gezien de verspreiding door de Anopheles mug optimaal verloopt bij stilstaand water of moeras, is een ziekte waarmee wereldwijd 500 miljoen mensen geïnfecteerd zijn. Eén vierde van de totale wereldbevolking loopt kans op infectie. Malaria doodt meer dan 100 miljoen mensen per jaar waaronder meer dan 1 miljoen kinderen. Het grootste aantal geïnfecteerde

mensen leeft in Afrika. De meeste die de ziekte overwonnen ervaren levenslang de gevolgen van anemie en immuno suppressie. Malaria kent momenteel een sterke heropleving door ontwikkeling van resistentie aan bestaande middelen en globale klimaatwijzigingen. De nood aan innovatieve behandelingsmethoden zou vlugger acuut kunnen worden dan voorzien. Het WHO wil tegen 2010 het aantal doden ten gevolge van malaria met de helft verminderen via het Roll Back Malaria investeringsbudget.

Plasmodium beschikt over een uiterst complexe levenscyclus, waarbij het verschillende malen van vorm en gedrag verandert. De parasiet bleek zich in de loop der jaren met een steeds groeiend succes aan te passen aan de bestaande bestrijdingsmiddelen. Slechts drie types geneesmiddelen zijn beschikbaar: quinolines, anti-foliaten en artemesines. Zowel tegen quinolines als anti-foliaten bestaat al resistentie.

De rationale van het voorgestelde project is eenvoudig. Momenteel bevinden we ons in de unieke situatie dat zowel het genoom van de betrokken parasiet Plasmodium falciparum (Nature, 30oct2002) als die van de vector, de mug Anopheles gambiae (Science 298: 291:2002) gekend zijn. Daarenboven beschikken we reeds een tweetal jaren over een mooi geannoteerde versie van het menselijke genoom, de uiteindelijke "derde man" in het verhaal.

Met deze informatie en nieuw te ontwikkelen methodologiën zullen we nieuwe genen en signaaltransductiewegen identificeren die betrokken zijn in de transmissie, pathogeniciteit of drug reactie van Plasmodium. Op deze manier denken wij de ambitieuze titel van ons project te kunnen waarmaken. Tevens zullen we proberen betere inzichten te verschaffen in de complexe gastheer-parasiet interacties

Een exhaustieve vergelijkende analyse van de drie volledig genomen, waarbij we alle mogelijke vergelijkingen uitvoeren (zowel DNA/DNA als DNA/Eiwit als Eiwit/DNA, en dit in beide richtingen (reciprook voor query en database) levert al snel meer dan 500 genoomvergelijkingen op. Gezien de computationele cost (ongeveer 1 cpujaar per genoomvergelijking), zal er een optimaal pad gekozen worden. Hierbij wordt rekening gehouden om de kans het vinden van nieuwe homologen (door bijvoorbeeld niet enkel te zoeken met reeds voorspelde genen) af te wegen tegen de kost om bestaande kennis te dupliceren. Aangezien het niet eenvoudig is beide parameters (precisie en recall) in te schatten zal er vooral gekeken worden naar reeds gekende targets. Nadien zullen de talrijke Blast resultaten, uitgevoerd op een beschikbare linux cluster, nadien netjes gestructureerd worden in een databank. Deze laat dan verdere "slicing and dicing" van de data toe. Dat deze benadering nog niet exhaustief werd uitgevoerd heeft niet enkele te maken met het ontbreken van de complete sequentie maar ook het gebrek aan datamodellen en visualisatie tools. Voor de ontwikkeling hiervan zullen we vooral gebruikmaken van open-source technologie. Voor genoomtopologieën zullen we zelf aan een algoritmische weg timmeren.

Een kort historisch overzicht bevestigt het vermoeden dat vergelijkingen tot zinvolle resultaten kunnen leiden:

Zelfs voordat het genoom volledig gekend was, werden nieuwe drug targets geïdentificeerd door te zoeken naar "unieke" genen. Unieke genen komen slechts voor in één van de onderzochte species. Bijvoorbeeld alle plasmodium genen die niet terug te vinden zijn in mens. Deze disjuncte sets leveren species specifieke genen op, die kunnen wijzen op species specifieke processen, inclusief mogelijke zwakheden. Men zou ook kunnen kijken naar de gemeenschappelijke stukken, die ofwel zeer "oud" zijn, maar die vinden we dan ook terug in bv Yeast of die recent verworven zijn door plasmodium om het humaan immuunsysteem te kunnen verschalken,

Het Plasmodium genoom is 80% AT rijk en maakte van het sequencen een praktische en computationele nachtmerrie. Gen-predictie, althans de klassieke manieren, lijken te profiteren van die codon bias. In totaal zijn ongeveer 5279 genen voorspeld, waarvan meer dan 60% geen enkele homoloog vertoont buiten Plasmodium. Dit genaantal wordt ondersteund door mass spec data van endogene authentieke peptides.

Naast de primaire vergelijkingen is het "inlinken" van hoogwaardige biologische kennis essentieel.

Een van de meest in het oog springende resultaten was wel de afwezigheid van enkele sleutelenzymen voor de generatie van ATP en NADH. Dit doet vermoeden dat Plasmodium energie genereert en stockeert op een alternatieve wijze, wat mogelijk interessant kan zijn als drug target. Met andere woorden door een combinatie met gekende signaaltransductiewegen kunnen we achterhalen voor welke nutriënten Plasmodium obligeert is. Deze informatie is eveneens voorhanden in de literatuur of meer gestructureerd in de vorm van ontologieën.

Uitzonderlijk zijn eveneens het aantal genen voorspeld in de apicoplast (10%). Dit essentieel "quadruple membraned" subcellulair compartiment is betrokken in de biosynthese van vrije vetzuren en isoprenoïden, componenten van membraaneiwitten en ijzer metabolisme. Een goede "classifier" kan nieuwe inzichten verschaffen in de functie van dit organel, dat o.a. minstens 4 van de bestaande anti-malaria targets (Waller PNAS, 1998) bevat.

De valorisatie van de resultaten zal geschieden via samenwerkingen met externe labos en zal initieel bestaan uit RNAi knockdowns, een technologie te recentelijk succesvol werd gebruikt in Plasmodium cellijnen.

Een nieuw concept: Genoomtopologieën

Er zijn tal van beperkingen van aan het vergelijken van genomene op een "gene by gene" basis. Alhoewel het uitermate pragmatisch en succesvol gebleken is om het gen als moleculaire eenheid te gebruiken blijken er op een "hoger" genoom niveau talrijke elementen van belang te zijn. Initele inzichten op dit gebied komen uit de analyse van het Anopheles genoom. Zo blijken er verschillende species te bestaan, die niettegenstaande alle genen reciproque gerepresenteerd zijn, toch niet in staat zijn succesvol te kruisen. Hier spelen topologische kenmerken van het genoom een rol die een grotere genoombasis hebben en dus niet terug te brengen zijn tot het nucleotide,

exon of gen niveau. Aan de hand van moderne evoluties in de wiskundige compressietechniek, meer bepaald wavelets, zouden we nagaan in hoeverre we abstractie kunnen maken van ons "gen" niveau. Als positieve controle zouden we in staat moeten zijn om van soortgelijk-muggen-DNA te onderscheiden van soortverschillend-muggen-DNA. hiervoor zijn de genomsequenties voor handen. Het unieke aan het voorstel zoals het er nu voorligt is dat het de klassiekere gene-by-gene approach kan aanvullen. Misschien laat het zelfs toe een unieke blik te werpen op de hogere orde organisatie in vb Plasmodium, die gezien zijn hoog AT gehalte door de natuur geselecteerd is om hierin uit te blinken !

Een andere opmerkelijke toepassing van genomtopologieën, dan wel op een iets bescheidener niveau, dat aansluit met de andere werkzaamheden en expertise in het labo voor Moleculaire Biotechnologie, heeft te maken met telomeerorganisatie. Het is namelijk in de telomeren van Plasmodium dat we veel van de oppervlakte eiwitten (var genes) terugvinden. Het zijn deze eiwitten die het humane immuunsysteem kunnen activeren (immuun evasief gedrag). Het blijkt nu dat er ongewoon veel variatie optreedt ter hoogte van de telomeren, preferentieel met telomeren van andere chromosomen ! Uiterst boeiend is de afwezigheid van deze eigenschappen in Plasmodium yoeli yoeli en Plasmodium vivax.

In parallel met deze nieuwe data-compressie technieken zal er een topologie van repeats worden geconstrueerd. Hiervoor zal gebruik worden gemaakt van distance arrays en MUMmer (van TIGR). Het corrent in beeld brengen van repeats, die in sommige gevallen het grootste deel van het genoom uitmaken, is onontbeerlijk om een correct beeld te krijgen van de informatieinhoud van een genoom.

Detectie, karakterisatie en validatie van ncRNA's (non-coding RNAs)

Kleine ncRNA's zijn uiterst belangrijk in regulatorische processen, toch werd hun bestaan en mogelijk functie slechts heel recent aangetoond. Dit is vooral te wijten aan het feit dat ncRNA genen coderen voor heel kleine RNA transcripten die slecht in lage concentraties voorkomen in de cel. ncRNA's worden dan ook weinig of niet geïsoleerd in klassieke cDNA banken maar alleen in cDNA banken die verrijkt werden voor deze kleine weinig voorkomende transcripten. Verder coderen ncRNA genen niet voor open lees ramen (ORFs), maar voor kleine niet vertaalde transcripten. ORF's kunnen relatief gemakkelijk opgespoord worden door middel van de bestaande bioinformatica technieken, maar informatica technologie voor het opsporen van ncRNA's waren tot voor kort niet ontwikkeld.

Dit project heeft dan ook als doelstelling om ncRNA's te identificeren. Meer in het bijzonder zal het project zich richten op ncRNA's in het humane en muis genoom, daar hier het grootste valorisatie potentieel aanwezig is. Daarnaast zal een zelfde oefening worden uitgevoerd in *C.elegans/C.briggsae*. *C. elegans* is het model organisme dat de vlugste mogelijkheden biedt op het gebied op het uitschakelen van genen door middel van RNAi en omdat in dit organisme de verdere ncRNA en RNAi

pathway op de meest efficiënte manier kan worden bestudeerd. Verder is de beschikbaarheid van het *C. briggsae* genoom een enorme troef, daar het organisme morfologisch gelijk is aan *C. elegans*, zich onderscheidt van dit organisme door een evolutionaire afstand van meer dan 200 miljoen jaar. De humane en muis ncRNA's zullen worden gevalideerd in muis, vooral gebruikmakend van morpholinos en KO muizen. Daar in de zeer nabije toekomst verwacht wordt dat de DNA sequenties van andere genomen vrij ter beschikking zullen zijn, zal het onderzoek in de loop van het project zich ook uitbreiden naar andere organismen. Als eerst volgend target organisme wordt vooral gedacht aan plasmodium, chimpansee en zebrafish. Voor plasmodium heeft het Instituut voor tropische geneeskunde reeds zijn interesse getoond.

Project Beschrijving

1) Isolatie van kandidaat ncRNA

In het initiële programma wordt op een systematische manier ncRNA's geïsoleerd en geannoteerd. Met het vorderen van het project zullen andere strategieën worden toegepast om een zo breed mogelijk spectrum aan ncRNA's te isoleren, en wel op volgende wijze:

a) Isolatie van de intergenische gebieden van het Humane en Muis genoom.

Gebruikmakend van gekende en te ontwikkelen bioinformatica middelen worden uit het genoom van mens en muis alle DNA segmenten die coderen voor eiwitten (ORFs) verwijderd. Een identieke oefening wordt uitgevoerd voor *C.elegans* en *C.briggsae*, andere organismen volgen later.

b) Detectie van geconserveerde intergenische gebieden tussen mens en muis

De intergenische gebieden van het mens genoom en het muis genoom worden vergeleken op homologie basis. Dit resulteert in een databank die alle geconserveerde intergenische gebieden bevat van mens en muis alsook voor *C. elegans/C.briggsae*, en in een later stadium voor andere organismen.

c) Verwijderen van gekende elementen

De databank met homologe intergenische gebieden bevat natuurlijk nog elementen die dienen verwijderd te worden zoals onder andere tRNA's, rRNA's, transposons, onvoorspelde genen, gekende regulatorische elementen, etc. (idem *C.elegans/C.briggsae*, andere organismen)

d) Detectie en isolatie van ncRNA's

De databank bevat nu in principe alleen homologe DNA segmenten waarvan de functie niet gekend is, of althans sequenties die niet coderen voor eiwitten, gekende RNA's transposons, etc. De overblijvende homologe gebieden worden onderworpen aan een bioinformatica programma dat selecteert voor specifieke transities. Enkel transities die mogelijke aanwezige secundaire structuren in de homologe fragmenten bewaren worden uitgeselecteerd. Een uitstekend bioinformatica instrument, Q-RNA werd daarvoor ontwikkeld. (Rivas en Eddy, BMC Bioinformatics. 2:8 (2001)).

e) Detecteren van secundaire structuren en mogelijk target genen

De geïsoleerde ncRNA's worden verder geklasseerd op basis van hun secundaire structuur. Reeds gekende ncRNA genen van de *lin-4/let-7* familie vertonen een overeenkomstige secundaire structuur (stem-loop structuur) (Zuker and Stiegler, NAR:10:133-148 (1981)). In analogie met de *lin-4/let-7* genen kan hieruit het



verwachte actieve 21nt RNA fragment worden afgeleid. Genen waar de geïsoleerde ncRNA's op binden kunnen dus worden geïsoleerd door middel van homologie vergelijkingen

De hierboven voorgestelde initiële strategie zal leiden tot de isolatie van hoog geconserveerde ncRNA's uit intergenisch regio's in de genomen van mens, muis en C.elegans/C.briggsae. Vergelijking van deze twee orthologe datasets zal de evolutionair geconserveerde ncRNA's opleveren. Herhaling van deze strategie, gebruikmakend van plasmodium, mens en chimpansee genomen, van deze laatste mag verondersteld worden dat het genoom zal in kaart worden gebracht gedurende het project, zal hoogst waarschijnlijk toelaten verder te filteren voor relevant ncRNA.

In de loop van het project zullen dan ook andere strategieën worden ontwikkeld en toegepast om ncRNA's te detecteren. Deze nieuw te ontwikkelen zoekstrategieën zullen ontwikkeld worden op basis van de vergaarde kennis.

Functionizing: onderzoek naar nieuwe genfuncties via integratie van bestaande genomdatasets (bijvoorbeeld chipdata en yeast-two-hybrid data)

Samenvatting:

Dankzij grootschalige “genoom sequencing projecten” is de voorbije jaren de genetische informatie (“het genoom”) van verscheidene organismen bekend geworden. Deze hoeveelheid data zal in de nabije toekomst nog veel sterker toenemen. Een eerste uitdaging was het identificeren van afzonderlijke genen in deze DNA sequenties. Elk gen “codeert” voor een bepaald eiwit, dat op zijn beurt een welbepaalde functie heeft binnen in de cel. Voor enkele belangrijke organismen, waaronder ook de mens, zijn de meeste genen reeds benoemd (mens: 30000 tot 35000 genen). Een volgende uitdaging is het toekennen van specifieke functies aan elk van de duizenden genen. Er moet een antwoord gegeven worden op de vraag wat de rol is van elk gen en hoe verschillende genen samenwerken tijdens bepaalde cellulaire processen.

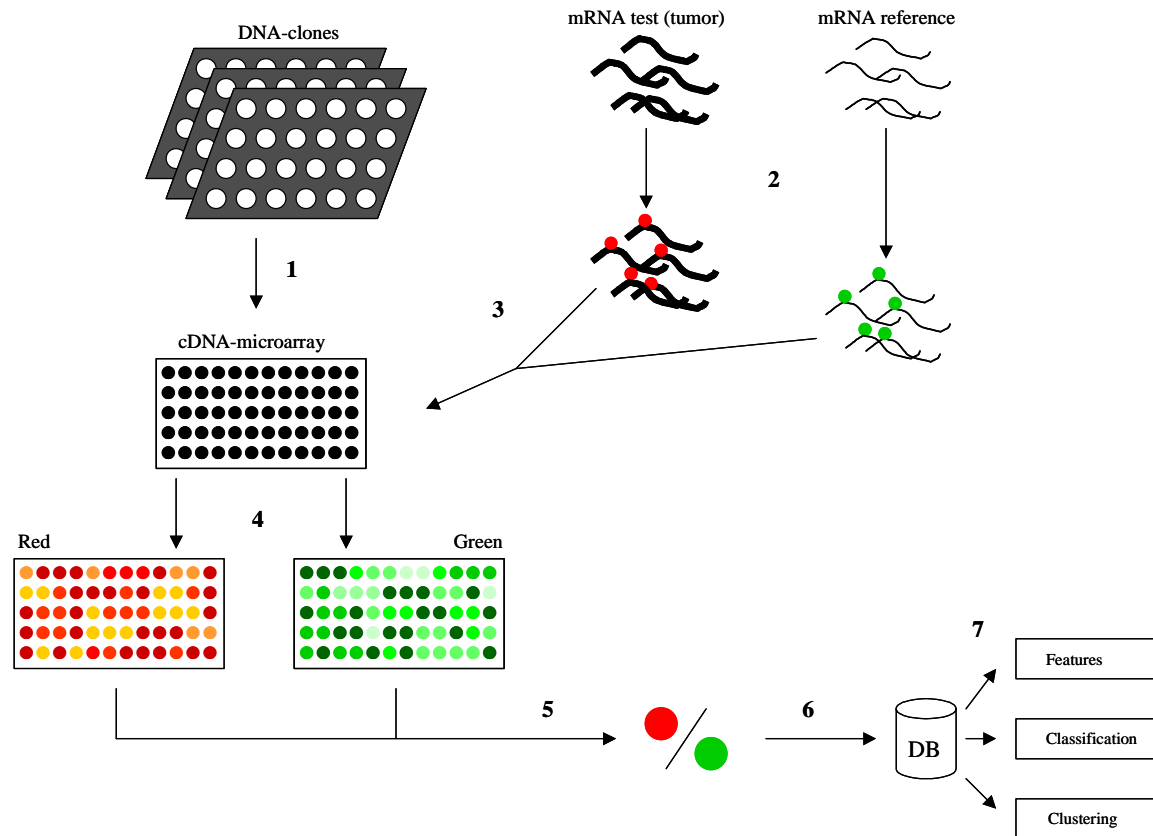
Met behulp van genexpressie-analyse (microarrays, zie Figuur 1) kan het tijdstip bepaald worden waarop de verschillende genen worden vertaald tot eiwitten. Genen die betrokken zijn bij verwante processen of behoren tot dezelfde regulatorische pathways, zullen meestal tegelijk tot expressie komen.

In dit project zal gebruik gemaakt worden van bestaande, publieke genexpressie-data om aan niet-gekaracteriseerde genen een functie toe te schrijven (“Functionizing”). Hiertoe wordt de expressie van de genen geanalyseerd en in kaart gebracht door middel van “self-organizing maps” (SOMs), een wiskundige clustermethode die bijzonder geschikt is voor het herkennen en classificeren van kenmerken in complexe, multidimensionale datasets. (Zie Figuur 2) Er zal getracht worden om een functie toe te kennen aan genen met een tot op heden ongekende rol, door na te gaan met welke gekarakteriseerde genen ze tot co-expressie komen. Deze methode zal worden geïntegreerd met twee andere technieken. Ten eerste zullen de SOMs gecombineerd worden met data van “yeast two-hybrid” experimenten, dewelke informatie verschaffen over welke eiwitten fysisch met elkaar kunnen interageren doch niet noodzakelijk op hetzelfde moment tot expressie komen Bovendien zal worden gebruik gemaakt van algoritmen die verwante domeinstructuren opsporen bij verschillende eiwitten (bv. InterPro). Voor verscheidene genen zal immers een functie kunnen worden voorspeld op basis van structurele overeenkomsten met goed beschreven genen.

Dit project combineert drie recente technieken binnen de bioinformatica met als doel een fundamentele bijdrage te leveren aan de genomanalyse, een “hot topic” in het biotechnologisch onderzoek. Enerzijds zullen verdienstelijke analytische technieken geoptimaliseerd en geïntegreerd worden, anderzijds zal de biologische rol van meerdere genen aan het licht worden gebracht. (wreed schoon gezegd ;-)

Bijlage: Figuren

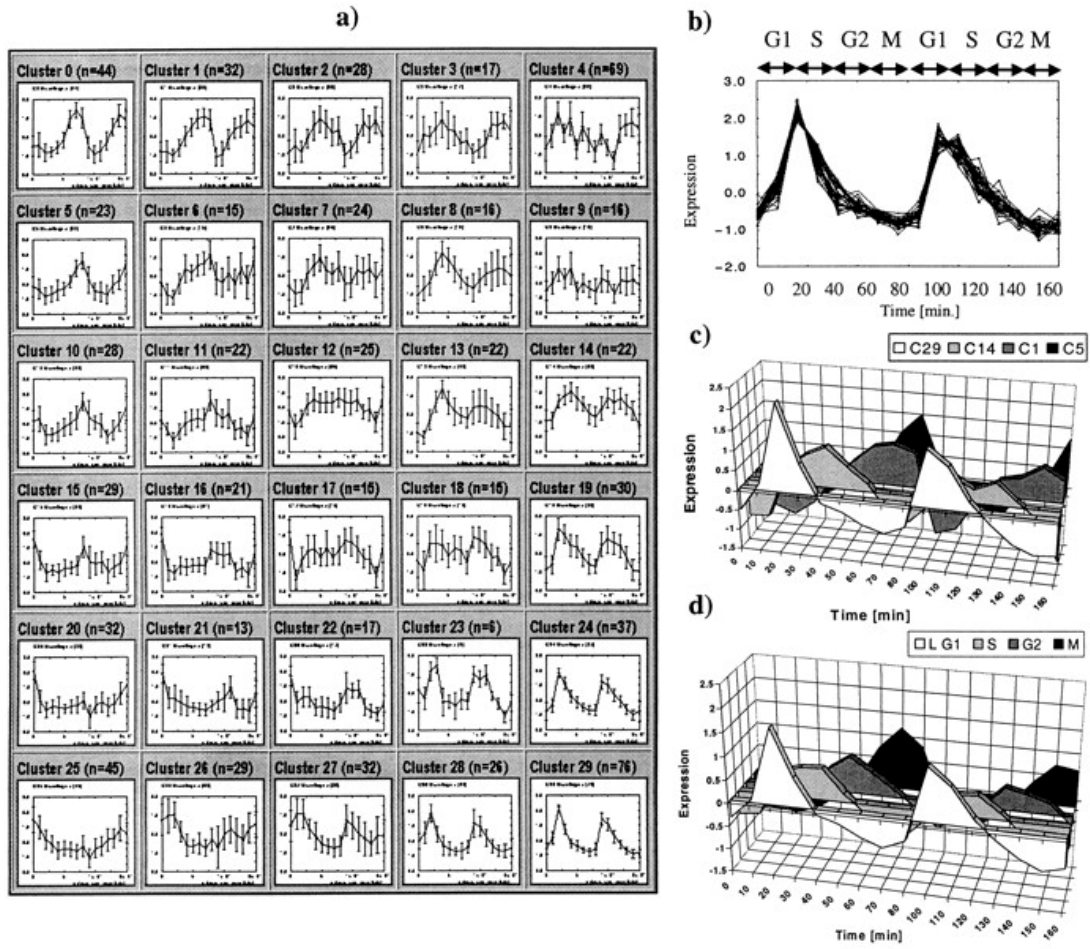
Figuur 1



Schematisch overzicht van een microarray experiment. (1) Spotten van de DNA-probes (gebaseerd op de bestudeerde genen) op een glasplaatje. (2) Labelen van het totale mRNA van het test-staal (bv. Een tumor; rood) en een referentie-staal (groen). (3) Poolen van de twee stalen en hybridisatie (4) Aflezen van de afzonderlijke signaal-intensiteiten voor elke probe. (5) Berekening van de respectievelijke expressieniveau's. (6) Bewaren van de resultaten in een databank. (7) Data exploratie.



Figuur 2



Geautomatiseerde representatie van geconserveerde aminozuurrestiduen op 3D structuurmodellen

De voorbije jaren hebben grootschalige projecten de genetische informatie (genoomsequenties) van talrijke organismen ontrafeld. Tegelijkertijd werd de nodige software ontworpen om in deze gigantische hoeveelheid informatie functioneel relevante structuren te herkennen. Zo werden algoritmes opgesteld die het mogelijk maken om genen te identificeren. Genen zijn relatief kleine gedeeltes van het genoom die “coderen” voor eiwitten, moleculen met een specifieke functie binnen de cel.

Vervolgens ontstonden algoritmes om deze DNA- of eiwitstructuren te vergelijken op het sequentieniveau. Via zogenaamde “sequentie-aligneringen” kon de variatie tussen meerdere verschillende moleculen of verschillende organismen worden nagegaan. Op die manier konden geconserveerde restiduen in kaart worden gebracht (Zie Figuur 1) en konden hiërarchische gestructureerde boomstructuren worden opgesteld die de verwantschap tussen meerdere moleculen weergeven (“phylogenetic trees”, zie Figuur 2). Het feit dat specifieke restiduen bewaard zijn gebleven tijdens de evolutie, bleek in vele gevallen te duiden op een belangrijke rol van de restiduen bij de functie van het eiwit: katalytisch domein, eiwit-eiwit interactie, vouwpatroon van het eiwit, DNA-bindingsplaats, ...

Later werden ook meer gespecialiseerde programma's gemaakt voor de representatie van de 3-dimensionele structuur van zowel DNA-, RNA- als eiwitmoleculen (bv. RasMol; zie Figuur 3). Deze ontwikkeling was een zeer belangrijke stap inzake molecuul-modellering en speelt bijvoorbeeld een essentiële rol in de ontwikkeling van medicijnen.

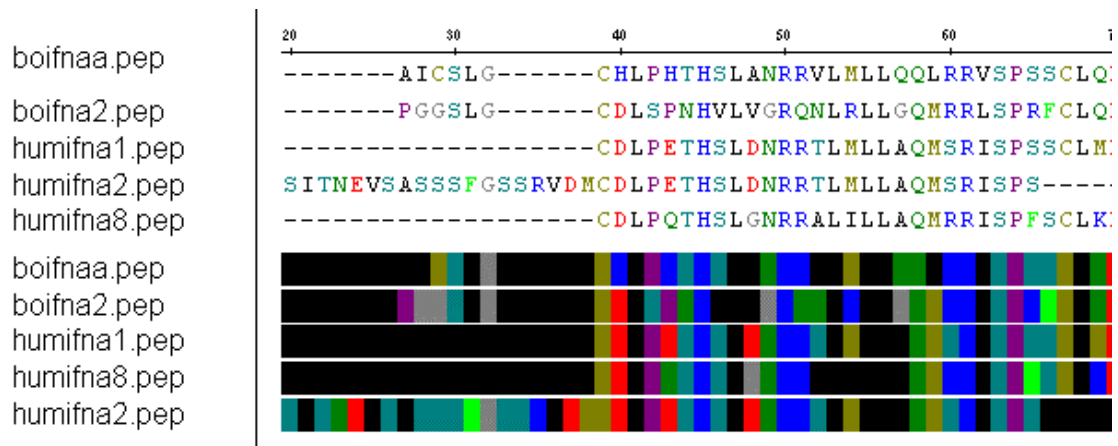
Tijdens dit project zal een link worden gelegd tussen de drie hoger beschreven technieken. Uitgaande van sequentie-aligneringen zullen geconserveerde restiduen worden geïdentificeerd binnen verwante moleculen en wordt hun onderlinge verwantschap gevisualiseerd in een manipuleerbare phylogenetische boomstructuur. Voor elk gen wordt bovendien een grafische voorstelling voorzien in een tweede panel zodat secundaire structuren (helix-structuren, hydrofobe/hydrofiele domeinen, ...) kunnen worden zichtbaar gemaakt. Hiervoor zullen we uitgaan van bestaande publieke algoritmen (bv. SMART). We zullen vervolgens een applicatie op punt stellen die deze conserveringen automatisch grafisch voorstelt in de 3-dimensionele representatie, rekening houdend met de voorwaarden die via een grafische user-interface worden verkregen van de gebruiker (vereist % homologie, representatiekleuren, ...). Hiervoor zal onoverkomelijk gebruik worden gemaakt van sterk gespecialiseerde programmeertechnieken (bv. Java).

Deze applicatie zal toelaten om geconserveerde restiduen te localiseren binnen de 3D structuur van het molecuul. Dankzij de visualisering van de conserveringen ten op zichte van elkaar, zullen we een beter inzicht krijgen op de mogelijke actieve regio's van het molecuul aangezien deze regio's een verminderde variatie zullen vertoond hebben tijdens de evolutie.

Belangrijke informatie rond dit project is te vinden in de tutorial “Proteins: a tools approach” gegeven in tijdens het recent O’Reilly congres te San Diego. Copies hiervan zullen ter beschikking gesteld worden.

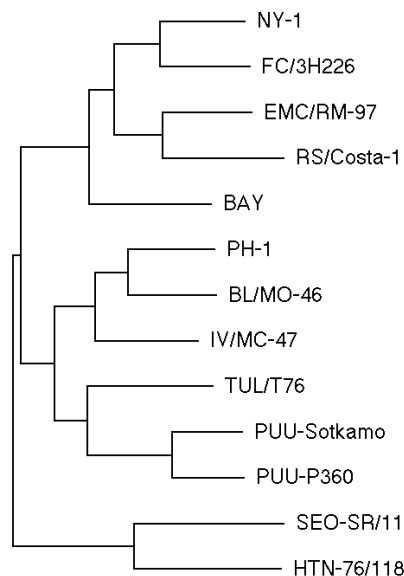
Figuren

Figuur 1



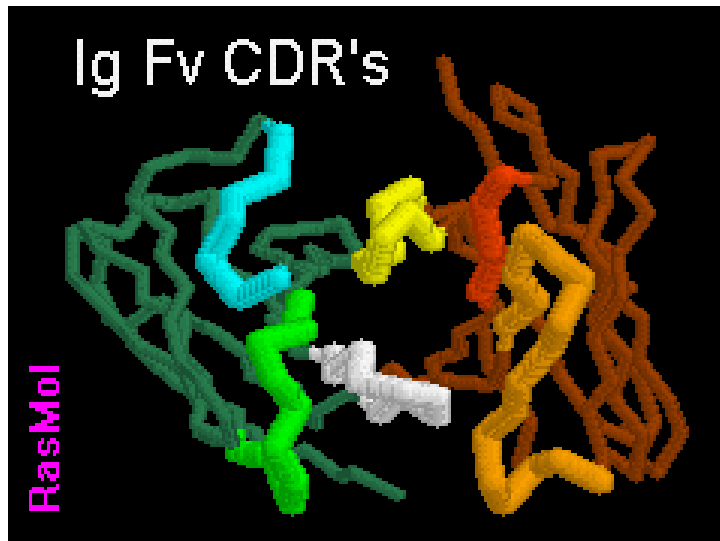
Voorbeeld van een sequentie-aligering van 5 verwante eiwitten. De graad van conservering kan visueel worden weergegeven met behulp van verschillende kleuren.

Figuur 2



Voorbeeld van een hiërarchisch gebouwde, phylogenetische boomstructuur voor de representatie van de onderlinge verwantschap tussen meerdere moleculen.

Figuur 3



3-dimensionele structuur van een immunoglobuline eiwit met behulp het 3D-visualisatieprogramma "RasMol"

Gene

- NT sequence
- GC-content
- enhancer
- promoter length
- promoter binding sites
- coding strand (transplicing !)
- transcription start site
- introns (number, size, location, frame)
- exons (number, size, location, frame)

mRNA

- translational start site(s)
- location stop codon
- 5'UTR
- 3'UTR
- codon useage
- secondary structure
- editing
- splice variants

Protein

- AA sequence
- AA composition (also for types, pairs?)
- Under/overrepresented AA
- AA conservation
- sequence length?
- (calculated, experimetal) molecular mass
- (calculated, experimental) isoelectric point
- atomic composition?
- keywords
- secondary structure (X-ray, NMR)
- predicted secondary structure
- 1. % alpha
- 2. %beta
- 3. %turn
- 4. %coil
- 5. pattern (alpha, beta, turn, coil)
- 6. length each alpha
- 7. length each beta
- 8. length each turn
- 9. length each coil
- 10. neighbour of each alpha/beta/turn/coil

Overall

- family
- sub-family
- place in ontology
- role in physiology

disease association
 phenotype(s)

References

- D.W.Wood,M.Mitchell,Sgouros J, Lindahl T, Science, 291,1284 (2001)
- J. A. Eisen, P. C. Hanawalt, Mutat. Res. DNA Repair 435, 171 (1999).
- L. Aravind, D. R. Walker, E. V. Koonin, Nucleic Acids Res. 27, 1223 (1999)

BioCardridge - BioSQL

Bijna alle bioinformatica projecten maken uitvoerig gebruik van databanken. SQL is de afkorting van "Structured Query Language". Dit is een algemene taal die gebruikt wordt om databases aan te spreken. Met behulp van SQL kunnen gegevens uit tabellen worden gehaald, gegevens worden toegevoegd en gegevens worden verwijderd. Ook kunnen er tabellen gemaakt, gewijzigd of verwijderd worden.

We zullen eerst een voorbeeld laten zien van het selecteren van gegeven. Dit is het belangrijkste deel van SQL We doen dit aan de hand van de volgende tabel:

Tabel Genes

ID	GeneName	Cancer	Sequence	# ESTs
1	ACTIN	0	MJSHBCLG ..	5
2	APC	1	MBNCVSHGF..	N/A
3	P53	1	MKHSDGG...	7
4	GFP	0	MJHGH...	2

We gaan nu de genen laten zien betrokken in kanker:

```
SELECT ID, GeneName FROM Genes WHERE Cancer = 1
```

Dit wordt een "Query" genoemd. De bedoeling is nu om de SQL taal aan te rijken met biologische "methoden". Technisch kan dit gereliseerd worden door bv C routines te incorporeren in de database "kernel". Praktisch heeft dit grote gevolgen voor de gebruiken. Zo zou het mogelijk zijn om bv

```
SELECT GeneName FROM Genes WHERE Sequence SIMILAR(1E-5,100,Actin.Sequence)
```

Het nieuwe woord "SIMILAR" initieert aan blast zoekopdracht met de Actine sequentie als query. Tevens kunnen parameters worden meegegeven in dit geval een p-waarde en een maximal aantal terug te geven sequenties. Deze manier heeft talrijke voordelen:

1. Het grote rekenwerk gebeurt op de “server” kant
2. Gebruiker hoeft zich geen zorgen te maken over specifieke implementatie van similariteit: blast, fasta, blat. Eventueel kan de keuze van programma worden ingevoerd als parameter ...
3. De resultaten kunnen direct gekoppeld in complexere (sub)queries (bv geeft enkel diegene met een voldoende aantal ESTs)
4. Deze nieuwe functies worden gebundeld in de zogenaamde BioCardridge. Men kan daardoor de code eenvoudig distribueren en beschikbaar stellen voor zowel gebruikers als developers.
5. Verschillende functies zijn te combineren en eenvoudig te gebruiken in web-applicaties
6. BioSQL is een generiek database schema dat momenteel als standard wordt gezien voor het stockeren van uiterst diverser bioinformatica informatie. De bedoeling is de extra functionaliteit van BioCardridge te integreren in BioSQL zodat er een maximale synergie tussen beiden ontstaat.

Dit project zal in samenwerking lopen met Biorebel, een bioinformatica bedrijf uit Gent. Biorebel beschikt momenteel reeds over de implementatie van een beperkte set keywords voor het berekenen van GC% en Reverse Complement.

ChemCardridge en toepassing in “structural genomics”

Terwijl de meeste bioinformatica tools “open-source” beschikbaar zijn, is dit zeker niet het geval voor chemische toepassingen. De meeste formaten om chemische structuren te stockeren zijn niet vrij beschikbaar en redelijk gedateert. Het is de bedoeling om via dit project moderne XML technologie te combineren met klassiek formaten (mol format, smiles of extended smiles) in een ChemCardridge.

De bedoeling is via deze Cardridge structuren te kunnen stockeren in een onderliggende databank. Vervolgens moet de mogelijkheid voorzien worden om exact of substructure zoekopdrachten te lanceren. Dit moet voldoende zijn om een niet redundante databank te construeren. Hiervoor is integratie met het LSID project, dat beoogt unieke ID te voorzien zeker wenselijk (*A Life Science Identifier (LSID) is a persistent, location-independent Uniform Resource Name (URN) for any biologically significant datum or data set that can usefully be stored or accessed via a networked computer system*).

In tweede instantie kan de functionaliteit uitgebreid worden naar de stoking van eiwitten (Data Uniformity Project, mmCIF format), eiwit-eiwit interacties (BIND,DIP,PIM,<http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html>) en eiwit-compound interacties. In parallel zullen modules voorzien worden die zogenaamde “descriptors” kunnen berekenen van de compounds maar ook van het “pharmacophoor”, die kan van het eiwit betrokken in de interactie.

Eens de informatie consistent gebundeld is in een databank kunnen er verschillende analyse technieken gebruikt worden die toelaten informatie van gekende interacties te extrapoleren naar nieuwe targets of compounds.

Functionele analyse van microarray genexpressie data

Gen expressie studies met behulp van microarray experimenten is geïntegreerd in een groot aantal lopende en geplande projecten. Voor een grondige analyse van deze microarray data zijn een aantal verschillende ontwikkelingen noodzakelijk.

De thesistudent zal zich primair bezighouden met de ontwikkeling van een database waarin de diverse velden (denk aan de afzonderlijke expressie niveaus voor de beide kanalen, spot grootte en vorm, voor de spot gebruikte sequentie en de bijbehorende gen informatie snel toegankelijk is en via standaard interfaces (XML DTD's) beschikbaar is voor analyse tools. De ontwikkeling zal worden gebaseerd op de MIAME specificatie van de MGED groep terwijl het data interfaces gebruik zal maken de MAGE-ML specificaties en de daarvoor ontwikkelde tools (zie verder <http://www.mged.org>).

Voor de analyse zal vooral gebruik gemaakt worden van non-supervised vormen van clustering, waarbij de mathematische formulering van bestaande verbanden tussen de data (zoals bijvoorbeeld afkomstig uit in-silico hybridisatie experimenten, zie verder) centraal zal staan. De numerieke vertaling van differentiële expressie data zodat die beter geschikt zijn voor clustering moet naast individualisering van datapunten ook aspecten als afstand in grootte van expressieveranderingen (van +1.5 naar -1.7 is bijvoorbeeld veelal belangrijker dan van +70 naar +50, ook al ligt dat numeriek niet door de hand) en aanpassing van de te gebruiken algoritmes aan de aantallen datapunten en de specifieke aard van als ratio gedefinieerde relatieve datapunten omvatten.

1. Basale databehandeling en statistiek. Een aantal basale databehandelingen zijn nog niet volledig uit ontwikkeld. Het betreft bijvoorbeeld de statistische behandeling van (al dan niet flipped) replicate arrays, en de normalisatie van expressie patronen die niet over het hele expressie bereik lineair zijn (een veelvoorkomend probleem). Keuze voor standaard methodieken voor de behandeling van replicates zal waarschijnlijk neerkomen op een van de spreiding afhankelijke keuze voor middelen dan wel kiezen voor de minimale absolute differentiële expressie, een en ander in samenwerking met onderzoekers bij Incyte, statistici en de betrokken biomedische onderzoekers. Voor normalisatie zal gebruik gemaakt kunnen worden van polynomen in plaats van de nu nog gangbare eenvoudige lineaire schaling. Beide data behandelingsmethoden zullen bovendien onderdeel moeten worden van standaard software bibliotheken en als zodanig ook gepubliceerd worden.
2. Expressie evaluatie door middel van in-silico hybridisaties. Voor verschillende cDNA arrays zijn de op de microarrays gebruikte sequenties bekend. Een eenvoudige BLAST search is de standaard manier om te beschrijven welk gen door een bepaalde spot het waarschijnlijkst gerapporteerd wordt. Daarmee zijn echter een aantal problemen. In een aantal gevallen worden vooralsnog alleen ESTs gevonden en is een regelmatige automatische re-BLAST noodzakelijk. Van de beste hit is niet meer bekend dan dat het de beste was en dat hij boven een zekere drempelwaarde lag. Het kan dus om meer of minder goede hits

gaan en er kunnen ook nog andere genen zijn met een hoge overeenkomst. Dat betekent dan weer dat het opkomen van het signaal niet noodzakelijkerwijs betekent dat inderdaad het gen op het etiket een verhoogde expressie heeft. Via sequence alignment t.o.v. diverse beschikbare databases (EMBL Incyte en mogelijk Celera, telkens opgesplitst naar species en in gen versus EST libraries) zullen in-silico hybridisaties worden uitgevoerd voor alle reporter sequenties waaruit een matrix zal worden gegenereerd met voor elke reporter voor alle genen met een significante alignment de op basis van de complementaire nucleotiden berekende bindingsterkte. Complementair aan deze benadering zal voor elk primair gerapporteerd gen worden berekend welke andere spots een gecorreleerde toename in signaal zouden moeten vertonen. Deze benaderingen zijn essentieel niet allen omdat zij toelaten meer informatie uit hetzelfde array te halen maar ook omdat zo de kans op onjuiste conclusies met betrekking tot de tot expressie komende genen wordt verkleind. Het onderkennen van correlaties tussen spots door cross-hybridisaties kan bovendien helpen voorkomen dat ten onrechte clusters van gecorreleerd tot expressie komende genen, bijvoorbeeld uit een familie van iso-enzymen, worden gerapporteerd.

3. Ontwikkeling van op familie, functie en pathway gebaseerde clusteringsmethodieken. Clustering, het vinden van genen die onder diverse experimenteel gerelateerde condities of op diverse tijdstippen overeenkomsten in expressie vertonen vormt een belangrijke voorwaarde voor het krijgen van een beter begrip van de resultaten van microarray analyse. Puur mathematische clustering zal echter – zelfs wanneer rekening gehouden wordt met de specifieke eigenschappen van microarray experimenten – niet eenvoudig tot een dergelijk toegenomen inzicht leiden. We willen immers meer dan een lijstje van een paar honderd gerelateerde genen. Een automatische verkregen signaal dat aangeeft dat onder de onderzochte condities bijvoorbeeld apoptose genen verhoogd tot expressie komen is veel waardevoller. Dit veronderstelt echter twee dingen. Ten eerste moet bekend zijn welke genen behoren tot overeenkomstige functionele families behoren, betrokken zijn bij overeenkomstige pathways of bijvoorbeeld beschikken over dezelfde regulerende elementen. Ten tweede is een operationalisering van dit type groeperingen in supervised clusteringsmethodieken nodig. De eerste stap kan bereikt worden door een specifieke implementatie van in ontwikkeling zijnde ontologiën voor cardiovasculair onderzoek. Zo ontwikkelde beslisbomen voor onderlinge samenhang kunnen vervolgens gevuld worden met behulp van informatie die aanwezig is in bijvoorbeeld pathway databases. Gezien de enorme omvang van die databases en het feit dat we tienduizenden genen op die manier willen documenteren kan dat alleen middels automatische procedures. Daarom zullen methoden moeten worden ontwikkeld om die alsnog te operationaliseren. Hetgeen dan zowel kan leiden tot aanpassing van de gebruikte ontologieën zelf, als tot een betere invulling van de velden. Gebruik van de verkregen hiërarchische indelingen kan vervolgens zowel door beïnvloeding van de clusterings algorithmen zelf (bijvoorbeeld door het creëren van een zekere stickyness tussen de tot een groep behorende genen) als door het toevoegen van uitvoer filters die de gevormde clusters beoordelen op het voorkomen van informatie die wijst op een opduiken van bekende samenhangen.

Machine Learning voor Tumor-Classificatie

De idee achter dit project is dat de PCRanalyse van welbepaalde genen predictief kan zijn voor het gedrag van een welbepaalde tumor. Hiervoor is het de bedoeling een "classifier" te construeren die toelaat de predictieve informatie uit de data te extraheren. Hiervoor kan gebruik gemaakt worden van verschillende standard technieken (bv. neurale netwerken) die echter aangepast dienen te worden aan de specifieke aard van de data alsmede geïntegreerd met de database van primaire gegevens. Nadien is het de bedoeling dat er een (web)applicatie ontwikkelt wordt dit toelaat de software eenvoudig ter beschikking te stellen. Voor dit project worden er verschillende externe samenwerking opgezet.

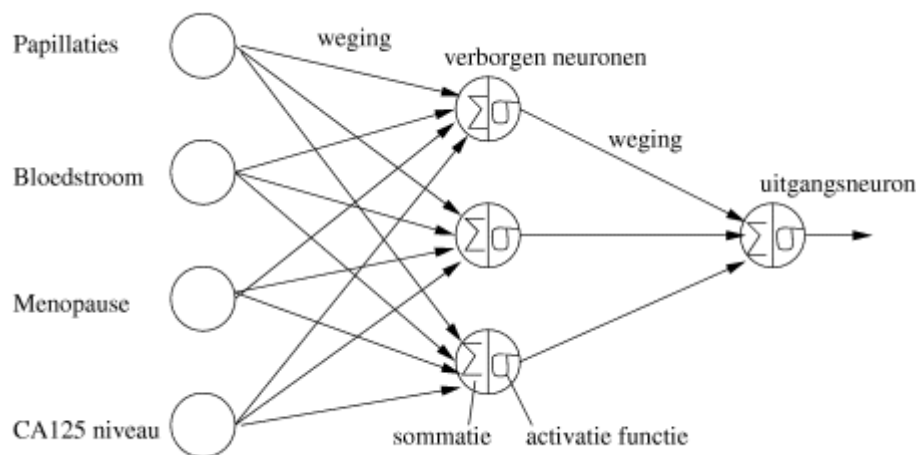
Ten einde de predictie beter te kunnen voorspellen werden reeds logistische regressie modellen en neurale netwerken toegepast. De interesse van de geneesheren gaat nu vooral uit naar de moeilijk te classificeren ovariumtumoren (een 50-tal patienten) en de betrouwbaarheid van de preoperatieve voorspelling in deze gevallen.

De student(e)(n) zullen zich moeten inwerken in:

- Artificiële neurale netwerken, meer in het bijzonder op meerlaagsperceptrons of MLP's, en verschillende trainingsmethodes.
- Bootstrapmethodes: Dit zijn computer intensieve simulatie-methodes die toelaten de betrouwbaarheid van de geschatte parameters (in ons geval : de kwaadaardigheid van de tumor) in probabilistische zin te berekenen.

De student(e)(n) zullen neurale netwerken ontwikkelen van diverse complexiteit en de performantie ervan onderzoeken en vergelijken met deze van bestaande technieken (bvb logistische regressie). De oppervlakte onder de Receiver operating characteristic (ROC) curves worden gebruikt om de performantie van algoritmen te beschrijven. Een ontwikkeling van een niet-parametrische bootstrapmethode is nodig om een vergelijk te maken in performantie tussen de onderlinge modellen. De bestaande statistische methoden kunnen hier niet gebruikt worden daar het aantal patiënten zeer klein is.

Voorbeeld van een neural netwerk:



Multiplexed Blast

Een van de meest gebruikte toepassing in de bioinformatica is Blast. Altschul, de auteur van het original Blast artikel, is hiermee de meest geciteerde wetenschapper aller tijden. Het wijd gebruik staat in schril contrast tot de kennis van het achterliggende algoritme. Zo zijn er verschillende parameters in te stellen bij het lanceren van een blast zoekopdracht. De bedoeling van de project bestaat erin de gebruiken te beschermen tegen de “default” values, alsmede te onderzoeken welke van de parameters de meeste invloed heeft. Een cruciale factor is de keuze van een scoring matrix (PAM tabel). Zo is het de bedoeling dat wanneer een blast opdracht gegeven wordt, er eigenlijk minstens 3 zoekopdrachten starten met zeer uiteenlopende PAM tabellen. Nadien worden de resultaten niet-redundant gesorteert en gepresenteert aan de gebruiker (bv deze hit werd enkel gevonden met een lage PAM table). Het spreekt voor zich dat de onderliggende methodologie nadien kan uitgebreid worden voor andere parameters, die eventueel door de gebruiker kunnen gespecificeerd worden. Zo kan men door de database variabel te een differentiële blast uitvoeren, door te special met gap creation en extension kan men proberen gekende hits beter te expanderen. Nadien zou men verder kunnen generaliseren door ook het eigenlijke programma variable te maken en zo de uitkomst van Blast/Ssearch en Blat te kunnen vergelijken.

Biolinux

Bioinformatics suffers from a vast amount of different software products which although available in open-source are difficult to get install. The RPM Package Manager (RPM) is a powerful command line driven package management system capable of installing, uninstalling, verifying, querying, and updating computer software packages. Each software package consists of an archive of files along with information about the package like its version, a description, and the like. In addition Biolinux is optimised for the intel architecture (<http://www.biolinux.org>)

Onderzoek in het kader van Biolinux kan verschillende wegen opgaan. Zo zijn tal van de nochtans cruciale bioinformatica applicaties niet geparalleliseerd, wat maakt dat

de meeste van de huidige implementaties enkel gebruik maken van een job scheduling. Ook de distributie zelf kan nog verbeterd worden. Vooral het automatiseren van updates en dependancies zou nuttig zijn en toepassingen vinden ver buiten de bioinformatica. Een laatste essentiële stap is de ontwikkelen van degelijke referentie-sets en benchmark tools die toelaten verschillende implementies te normaliseren en op een objectieve manier met elkaar te kunnen vergelijken (BioBenchMark). Tevens zijn referenties onontbeerlijk waar code geoptimaliseerd wordt. Dit om te vermijden dat de code ongelijk snel een totaal verkeerde uitkomst genereert !

Dynamisch Programmeren met Python & GATO

Dynamisch programmeren is het onderliggend algoritme in een aantal van de meest gebruikte bioinformatica toepassing (multiple alignemts). De bedoeling is een eenvoudig vorm te implementeren in Python, een object georiënteerde scripting taal. Momenteel bestaat er reeds een perl implementatie die eventueel kan dienen als basis op via de PyPerl module zelfs rechtstreeks zou kunnen beruikt worden.

GATO - the Graph Animation Toolbox - is een grafisch toolbox die mits aanpassing toelaat om de invloed van verschillende (genetwerkte) inputs na te gaan op de uitkomst. Dit laat toe een voeling te krijgen met het belang van de verschillende parameters en kan leiden tot nieuwe toepassingen.

In parallel hiermee zou het de bedoeling zijn de “extreme value” distributie, die aan de basis ligt van de statistiek en daarmee van het succes van blast, te simuleren en dit zowel met als zonder toegelaten gaps. Voor het geval met gaps is nog steeds één van de openstaande enigma's in de bioinformatie, sinds er nog een formeel bewijs bestaat voor de waargenomen distributie. Door gebruik te maken van een deze grafische omgeving in combinatie met een algoritme transparante Python hopen we een fundamentele bijdrage te kunnen leveren of op zijn minst een het inzicht in de problematiek te verhogen.

Toepassen van “Singular Value Decomposition” op biologische problemen

<http://www.nitle.org/biocon> geeft een excellent overzicht van het gebruik van SVD in verschillende biologische problemen. Niettegenstaande de source code beschikbaar is, is het bouwen van een zoek-applicatie niet eenvoudig.

Een eerste belangrijke toepassing van SVD is in zogenaamde “text-mining”. The objective of Text Mining is to exploit information contained in textual documents in various ways, including ...discovery of patterns and trends in data, associations among entities, predictive rules, etc.” (Grobelnik et al., 2001). The idea is to apply well established techniques to a set of patent database. This effort should allow to proactive mine the patent database for novel applications

In tweede instantie is het de bedoeling om gebruik makend van SVD, een specifiek biologisch probleem aan te pakken. Het betreft een nieuwe manier om aan de hand van een massa spectrum een sequentie te bepalen. Door simulatie kunnen de nodige test datasets gegenereerd worden. Via samenwerking met een Gents biotech bedrijf kunnen real life datasets getest worden en kan de applicatie getest worden.

Nutrigenomics & Functional Foods

In the near future, consumers will slide in their nutrigenomic card at any supermarket entrance to obtain a personalized list with recommended amounts, easy recipes and special discounts for certain foods. What kind of foods? Those that they ideally "should" consume to meet their genetically prescribed nutritional needs.

In other words, when following these customized dietary guidelines, an individual will be able to control at least one of the main modifiable risk factors that aid in preventing a wide range of diseases. Physical inactivity and overall industrial pollution are important other risk factors besides diet.

Much cited examples include selecting cholesterol-lowering margarine (such as Benecol) instead of butter by those genetically at risk of developing cardiovascular disease or using soy-based products instead of milk by those genetically lactose-intolerant.

Het vinden van relevante datasets binnen dit uiterst nieuwe veld is niet eenvoudig. Om toch concreet van start te kunnen gaan zou er een samenwerking worden opgezet met Wim Soutaert in de richting van Functional Foods. Hierbij is de bedoeling om zeer gericht potentieel interessante genen heteroloog tot expressie te brengen, die dan nieuwe eigenschappen kunnen verlenen aan welbepaalde levensmiddelen.

Om het concept van nutrigenomics te illustreren een stuk uit "the hitch hikers guide through the Galaxy"

It was called a NutriMatic Drinks Synthesizer, and he had encountered it before. It claimed to produce the widest possible range of drinks personally matched to the tastes and metabolism of whoever cared to use it. 'That drink,' said the machine sweetly, 'was individually tailored to meet your personal requirements for nutrition and pleasure.'

Analyse van biologische netwerk topologieën

Een steeds wederkerend fenomeen is de grote complexiteit van biologische netwerken. Vooral door de opkomst van high throughput technieken zijn er verschillende datasets op genome-schaal beschikbaar (bv interactie-data, chipdata, genome-wide RNAi ...). In sommige gevallen blijkt het hier te gaan om een zogenaamd schaalvrij netwerk wat enkele boeiende eigenschappen heeft. De bedoeling is de netwerk topologie te onderzoeken voor verschillende datasets. Dit kan bijvoorbeeld gebeuren door een wiskundige benadering waarin het netwerk wordt voorgesteld als een matrix waarvan vervolgens de eigenwaarden worden berekend. Het spectrum van deze eigenwaarden

is kenmerkend voor de onderliggende topologie. Deze topologie kan dienen als classificatie doch het is de bedoeling is hieruit praktische toepassingen te distilleren. Zo kan men de robuustheid testen en eventueel kritische punten identificeren, dit als onderliggend instrument voor “network engineering”. Anderzijds zijn verschillende van deze topologieën keerzijden van dezelfde medaille. Men kan proberen verschillende netwerken te vergelijken door ze proberen te laten samenvallen. Echt boeiend wordt het wanneer we proberen een species dimensie toe te voegen. We kunnen proberen aan netwerk inferentie te doen of eventueel het reconstrueren van grootste gemeenschappelijk delen, de zogenaamde ancestrale netwerken.

NMD (Nonsense Mediated Decay) in Viruses

NMD blijkt een belangrijk nieuw en generiek genregulatiemechanisme te zijn dat voorkomt in verschillende organismen. Het is de bedoeling om via bioinformatica technieken te zoeken in het genoom van virussen of zij mis(ge)bruik maken van deze manier van reguleren om de gastheer te infecteren. Het spreekt voor zich dat een beter begrip van de (regulatie) als gevolg van virusinfectie kan leiden naar nieuwe biotechnologische manieren om hierop in te grijpen.

Nonsense-mediated mRNA decay (NMD) is a pervasive cellular mechanism in eukaryotes that degrades mRNAs with premature termination codons. Intriguingly, the coupling of alternative splicing and NMD provides a general mechanism of gene regulation. In this regulatory system, splicing factors determine whether a pre-mRNA transcript will be spliced to form a productive isoform or an unproductive one. Thus, splicing factors can control, via splicing and NMD, whether a protein is expressed.

COP, cluster of orthologous Promoters

A major goal in the post-genome area is to reconstruct regulatory cellular networks starting from the available sequenced genomes and large-scale experimental expression data. A well-established stochastic strategy, named Gibbs sampling, looks for overrepresentation of so-called motifs against a background model. This approach is mainly used on sets of co-expressed genes and their corresponding promoter sequences.

Given the amount of different genomes available, one could reason, that orthologous genes, defined as vertical descendants from a common ancestor and therefore likely to have a similar function, have similar regulation. This is reflected in conserved motif(s) in what can be called, an orthologous promoter. A generic methodology, called phylogenetic footprinting, exploits this difference in evolutionary pressure on functionally important residues in different species.

COP, is een databank van orthologe promotors, die vervolgens als invoer kan gebruikt worden voor gespecialiseerde algoritmes. Het opstellen van deze databank is echter geen triviale zaak gezien het identificeren van orthologe genen op genoomschaal zeer rekenintensief is. Naast een data-model en interface zal het correct interpreteren, bv via Perl-scripts, van grote hoeveelheden recursieve blasts essentieel zijn.

Virtual Northern in SVG

De bedoeling is de resultaten van een TblastN search tegenover dbEST te representeren als een virtuele Northern. In essentie gaat het om het parsen en grouperen van blast hits. Om het geheel echter op een grafische manier te kunnen voorstellen zouden we gebruik maken van SVG, Scalable Vector Graphics (zie verder). Eens dit systeem loopt is het vrij eenvoudig het verder te generaliseren naar andere grouperingsmethoden zoals beschreven in het project “multiplexed blast”. Het gebruik van SVG in interfaces en data-visualisatie word gezien als één van de meeste beloftevolle gebieden in de bioinformatica.

SVG is exciting because it offers Web developers a method to create and animate images through an XML programming language. Consequently, rather than being removed from their code as is often the case with proprietary technology, developers can gain finer degrees of control over the appearance of Web pages. Animation techniques can range from a simple linear movement to 3D double helix morphing effects. Web developers, once they are more aware of the possibilities, can find unprecedented levels of control. There are many advantages of using SVG as the following short feature list demonstrates:

- Compatibility with other mediums such as wireless devices
- Scalable Server Solutions
- Small file sizes for faster Web page downloads
- Unlimited color and font choices
- Zoomable graphics and images
- Scripting control for custom interactive events and animation
- Clean, crisp, high-resolution printing from Web browsers
- Bitmap-style filter effects for high-impact graphics
- Text-based format easily integrates with other Web technologies
- Built in International Language Support
- Reduced Maintenance Costs
- Easily Updated
- Rich Multimedia Capabilities

Flash and SVG are often compared because the two have similar features. The reality is that SVG has some distinct advantages over its main competitor Flash. Perhaps chief among them is the compliance with other standards. SVG can utilize CSS and the DOM, where as Flash relies on proprietary technology that is not open source, at least not in the sense that we can right click on the page and see what is happening behind the scenes. SVG by contrast is open source and developers can readily learn from other developer's efforts in this area.

Bioinformatics Web Services

Het is de bedoeling om een groot van de toepassing in Biolinux (zie hoger) aan te bieden van Web Service. Concreet is het de bedoeling een generieke werkwijze op te

stellen voor ClustalW en Blast, die dan nadien kan ge-extrapolleert worden naar andere toepassingen.

Web Services are the way forward for bioinformatics. Many embryonic system are emerging. The main idea of this project is to explore them and select one technology for a proprietary implementation. By examining these specific services, we get a bird's eye view of the Web Service protocol stack, including WSDL and SOAP. Looking at working services also provides much food for thought. For example, the recently released Google API provides a glimpse of the future of business Web Services.

XEMBL provides complete access to the EMBL Nucleotide Sequence Database. This database is produced in collaboration with GenBank and the DNA Database of Japan, and currently provides access to over 16.8 million records, consisting of 19.6 billion nucleotides (see EMBL Database Stats.) It also provides access to completed genomes, including the human genome, the fruit fly, and *C. elegans*.

XEMBL is a recently released interface that provides easy XML access to the complete EMBL database. Access is provided via two main methods. The first is a REST-like interface whereby users specify parameters within a URL, and XEMBL returns a complete XML document. The second is a SOAP interface whereby users specify parameters within SOAP messages and XEMBL returns a complete XML document within a SOAP response. In responding to the current debate between REST and SOAP, you can see that the XEMBL group has not taken sides, and simply chosen both. This is in line with one of Lincoln's main points -- databases should provide multiple modes of access to data, from HTML, XML, and SQL, all the way to SOAP.

A number of other bioinformatic services are currently available or in the works. For example, the OmniGene project from MIT aims to create an open source Web Services platform for bioinformatics. You can currently download the OmniGene browser to get a feel for the platform -- the OmniGene SOAP API should be available shortly. Additionally, the Distributed Annotation Service (DAS) provides a distributed platform for aggregating genome annotation data from multiple sources. DAS 1.52 is currently implemented as XML over HTTP, but DAS 2.0 may move to a SOAP interface (see RFC0 and RFC11 for details.) Lastly, the BioMOBY project aims to provide distributed access to multiple bioinformatic services, and provide a centralized registry for finding new services. All of these projects are likely to see much progress in the near future.

Bioinformatics Portal

Bioinformatica maakt gebruik van tal van websites. De bedoeling is een onderhoudbare bioinformatica portal te maken die via <http://www.ncbi.be> ter beschikking kan worden gesteld. De tools waaraan gedacht wordt is de Google API en de Slash technologie.

De bedoeling is een paswoord protected discussieforum in te voeren dat kan gebruikt worden als ondersteuning voor de lessen bioinformatica maar ook als communicatietool bij practicumprobleem, thesissen of bioinformatica vragen in het



algemeen. Verder is het de bedoeling om een CVS-sever op te zetten die het moet toelaten verschillende versie van software code veilig te beheren.

De bedoeling is tevens een aantal programmes ter beschikking te stellen van de faculteit.